

TASI Lectures on Inflation

William H. Kinney

*Department of Physics, University at Buffalo, SUNY,
Buffalo, NY 14260-1500, USA**

This series of lectures gives a pedagogical review of the subject of cosmological inflation. I discuss Friedmann-Robertson-Walker cosmology and the horizon and flatness problems of the standard hot Big Bang, and introduce inflation as a solution to those problems, focusing on the simple scenario of inflation from a single scalar field. I discuss quantum modes in inflation and the generation of primordial tensor and scalar fluctuations. Finally, I provide comparison of inflationary models to the WMAP satellite measurement of the Cosmic Microwave Background, and briefly discuss future directions for inflationary physics. The majority of the lectures should be accessible to advanced undergraduates or beginning graduate students with only a background in Special Relativity, although familiarity with General Relativity and quantum field theory will be helpful for the more technical sections.

I. INTRODUCTION

Cosmology today is a vibrant scientific enterprise. New precision measurements are revealing a universe with surprising and unexpected properties, in particular the Dark Matter and Dark Energy which are now believed to be the dominant components of the cosmos. Galaxy surveys such as the Sloan Digital Sky Survey are making the first large-scale maps of the universe, and satellites such as WMAP are making exquisitely precise measurements of the Cosmic Microwave Background (CMB), the haze of relic photons left over from the Big Bang. In turn, these measurements are giving us clues which are helping to unravel one of the oldest and most profound questions people have ever asked: Where did the universe come from? In these lectures, I discuss what is currently the best motivated and most completely developed physical model for the first moments of the universe: cosmological inflation [1, 2, 3].¹ Inflation naturally explains how the universe came to be so large, so old, and so flat, and provides a compellingly elegant and predictive mechanism for generating the primordial perturbations which gave rise to the rich structure we see in the universe today [7, 8, 9, 10, 11, 12, 13]. Inflation provides a link between the Outer Space of astrophysics and the Inner Space of particle physics, and gives us a window to physics at energy scales far beyond the reach of particle accelerators. Furthermore, inflation makes *testable* predictions, which have so far proven to be an excellent match to the data.

The lectures are organized as follows:

- Section I provides an introduction and a brief overview of General Relativity.
- Section II discusses the Friedmann-Robertson-Walker spacetime and the standard hot Big Bang picture of cosmology, including the Cosmic Microwave Background.
- Section III explains unresolved issues in the standard cosmology, in particular the horizon and flatness problems.
- Section IV introduces inflation in scalar field theories.
- Section V discusses quantum fluctuations in inflation and the generation of cosmological perturbations.
- Section VI discusses the observational predictions of inflation, and current constraints from Cosmic Microwave Background measurements.
- Section VII discusses conclusions and the future outlook for inflationary physics.
- Appendix A describes in detail the generation of density perturbations during inflation.

*Electronic address: whkinney@buffalo.edu

¹ Inflation in its current form was introduced by Guth, but similar ideas had been discussed before [4, 5]. A short history of the early development of inflation can be found in Ref. [6].

The lectures are at an advanced undergraduate or beginning graduate student level. Most of the the lectures should be accessible with only a background in Special Relativity. A working knowledge of General Relativity and quantum field theory are helpful for Sections IV and V and for Appendix A. Where possible, I reference review articles for further reading on related topics. For other reviews on inflation, see Refs. [6, 14, 15, 16, 17, 18, 19, 20].

A. The Metric

The fundamental object in General Relativity is the *metric*, which encodes the shape of the spacetime. A metric is a symmetric, bilinear form which defines distances on a manifold. For example, we can express Pythagoras' theorem in a Euclidean three-dimensional space,

$$\ell^2 = x^2 + y^2 + z^2, \quad (1)$$

as a matrix product over the identity matrix $\delta_{ij} = \text{diag}(1, 1, 1)$,

$$\ell^2 = \sum_{i,j=1,3} \delta_{ij} x^i x^j. \quad (2)$$

Therefore the identity matrix δ_{ij} can be identified as the metric for the Euclidean space: if we wish to describe a non-Euclidean manifold, we replace δ_{ij} with a more complicated matrix g_{ij} , which in general can depend on the coordinates x^i . For an arbitrary path through the space, we express distances on the manifold in differential form,

$$d\ell^2 = \sum_{i,j} g_{ij} dx^i dx^j. \quad (3)$$

The distance along any path in the spacetime, or *world line*, is then given by integrating $d\ell$ along that path. A familiar example of a non-Euclidean space frequently used in physics is the Minkowski Space describing spacetime in Special Relativity. Distances along a world line in Minkowski Space are measured by the *proper time*, which is the time as measured by an observer traveling on that world line. The proper time s along a world line is given by the relation

$$\begin{aligned} ds^2 &= dt^2 - d\mathbf{x}^2 \\ &= \sum_{\mu,\nu=0,3} \eta_{\mu\nu} dx^\mu dx^\nu, \end{aligned} \quad (4)$$

where we take the speed of light $c = 1$. We express four-vectors as $\tilde{x} = (t, x, y, z) = (x^0, x^1, x^2, x^3)$, and $d\mathbf{x}^2 = dx^2 + dy^2 + dz^2$ is the Euclidean distance along a spatial interval. The metric $\eta_{\mu\nu}$ for Minkowski Space is given by

$$\eta_{\mu\nu} = \begin{pmatrix} 1 & & & \\ & -1 & & \\ & & -1 & \\ & & & -1 \end{pmatrix}. \quad (5)$$

Anything traveling the speed of light has velocity $d|\mathbf{x}|/dt = 1$. Photons therefore always travel along world lines of zero proper time, $ds^2 = dt^2 - d\mathbf{x}^2 = 0$, called *null geodesics*. Massive particles travel along world lines with real proper time, $ds^2 > 0$, called *timelike geodesics*. Causally disconnected regions of spacetime are separated by *spacelike* intervals, with $ds^2 < 0$. The set of all null geodesics passing through a given point (or *event*) in spacetime is called the *light cone* (Fig. 1) The interior of the light cone, consisting of all null and timelike geodesics, defines the region of spacetime causally related to that event.

B. General Relativity and the Einstein Field Equation

The Minkowski metric $\eta_{\mu\nu}$ of Special Relativity describes a Euclidean spacetime which is static, empty, and infinite in space and time. The addition of gravity to the picture requires General Relativity, which describes gravitational fields as curvature in the spacetime. The fundamental object in General Relativity is the metric $g_{\mu\nu}(\tilde{x})$, which describes the shape of the spacetime and in general depends on the spacetime coordinate \tilde{x} . As in Minkowski Space,

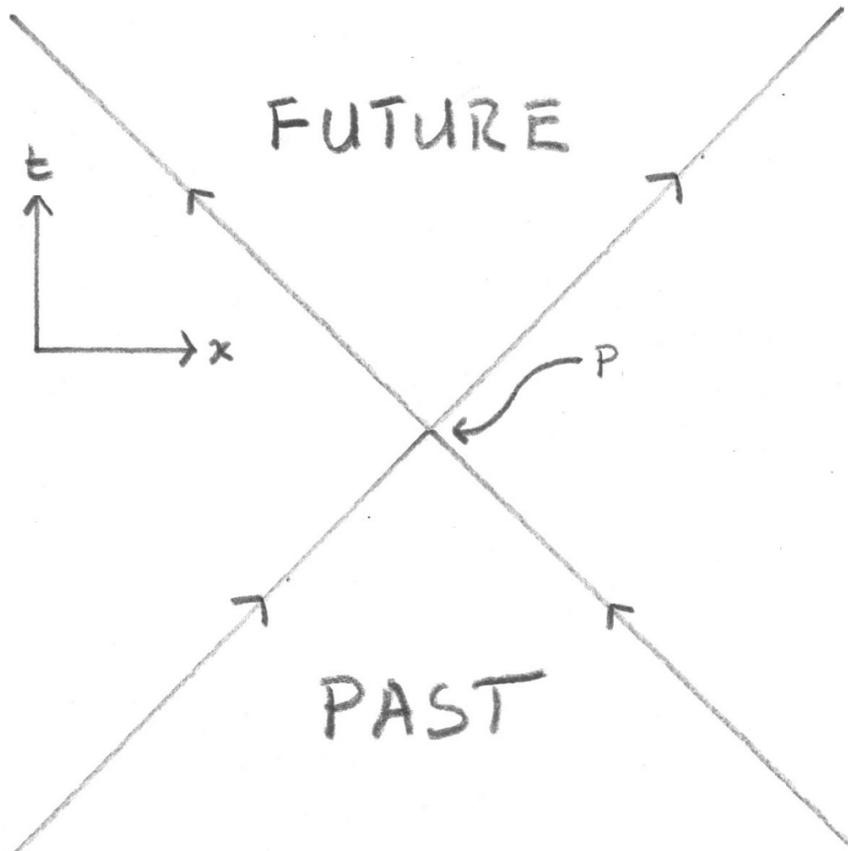


FIG. 1: Light cones in Minkowski Space. The past light cone defines the causal past of the event P , and the future light cone defines the causal future of P .

lengths in curved spacetime are measured by the proper time s , with the proper time along a world line determined by the metric

$$ds^2 = \sum_{\mu, \nu=0,3} g_{\mu\nu}(\tilde{x}) dx^\mu dx^\nu. \quad (6)$$

As in Special Relativity, photons travel along null geodesics, with $ds^2 = 0$, and massive particles travel along timelike geodesics, with $ds^2 > 0$. However, unlike Special Relativity, null geodesics need not always be 45° lines defining light cones, but can be curved by gravity.

In General Relativity, the distribution of mass/energy in the spacetime determines the shape of the metric, and the metric in turn determines the evolution of the mass/energy. Electromagnetism provides a convenient analogy: in electromagnetism, the distribution of charges and currents determines the electromagnetic field, and the electromagnetic field in turn determines the evolution of the charges and currents. Given a current four-vector J^μ , Maxwell's Equations are a set of linear, first-order partial differential equations that allow us to calculate the resulting electromagnetic field

$$\partial_\nu F^{\mu\nu} \equiv \sum_{\nu=0,3} \partial_\nu F^{\mu\nu} = \frac{4\pi}{c} J^\mu. \quad (7)$$

Here we have explicitly included the speed of light c to highlight its role as an electromagnetic coupling constant. We also adopt the typical summation convention for relativity: repeated indices are implicitly summed over. In General Relativity, we describe the distribution of mass/energy in a covariant way by specifying a symmetric rank-2 *stress-energy tensor* $T_{\mu\nu}$, which acts as a source for the gravitational field similar to the way the current four-vector J^μ sources electromagnetism. The analog of Maxwell's Equations is the Einstein Field Equation, which can be written in the deceptively simple form

$$G_{\mu\nu} = 8\pi G T_{\mu\nu}, \quad (8)$$

where the coupling constant is Newton's gravitational constant G . The tensor $G_{\mu\nu}$, called the Einstein Tensor, is a symmetric 4×4 tensor consisting of the metric $g_{\mu\nu}$ and its first and second derivatives. The Einstein Field Equation therefore represents a set of ten coupled, nonlinear, second-order partial differential equations of ten free functions, which are the elements of the metric tensor $g_{\mu\nu}$. However, only six of these equations are actually independent, leaving four degrees of freedom. The physics of gravity is independent of coordinate system, and the additional degrees of freedom correspond to a choice of a coordinate system, or *gauge* on the four-dimensional space. Gravity is *much* more complicated than electromagnetism! As with any intractably complicated problem, we simplify the job by introducing a symmetry. In General Relativity there are a number of symmetries which allow either exact or perturbative solution to the Einstein Field Equations:

- Vacuum: $T_{\mu\nu} = 0$. If we evaluate the Einstein Field Equations for small perturbations about an empty Minkowski Space, we find that they reduce at lowest order to a wave equation, and therefore General Relativity predicts the existence of gravity waves.
- Spherical Symmetry. If we assume a spherically symmetric spacetime (also empty of matter, $T_{\mu\nu} = 0$) the Einstein Field Equation can be solved exactly, resulting in the Schwarzschild solution for black holes.
- Homogeneity and Isotropy. If we assume that the stress-energy is distributed in a fashion which is homogeneous and isotropic, this is called a *Friedmann-Robertson-Walker* (FRW) space, and is the case of interest for cosmology. Since the homogeneity and isotropy remove all spatial dependence, the Einstein Field Equations reduce from a set of partial differential equations to a set of nonlinear ordinary differential equations in time. For particular types of homogeneous, isotropic matter, these equations can be solved exactly, and perturbations about those exact solutions can be handled self-consistently.

Continuing the analogy with electromagnetism, the equivalent of charge conservation,

$$\partial_\mu J^\mu = \frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{j} = 0, \quad (9)$$

in General Relativity is stress-energy conservation

$$D_\nu T^{\mu\nu} = 0, \quad (10)$$

where D_μ represents a covariant derivative, which is a generalization of the partial derivative to a curved manifold. We will also denote covariant derivatives with a semicolon, for example $T^{\mu\nu}{}_{;\nu} = 0$. Likewise, simple partial derivatives are denoted with a comma, $\partial f / \partial x^\mu \equiv \partial_\mu f \equiv f_{,\mu}$. As in the case of electromagnetism, where the charge conservation equation is not independent, but is instead a consequence of Maxwell's Equations, stress-energy conservation in General Relativity is a consequence of the Einstein Field Equations and does not independently constrain the solutions. In the next section, we discuss FRW spaces and their application to cosmology in more detail.

II. FRIEDMANN-ROBERTSON-WALKER SPACETIMES

A. The Friedmann Equation

A *homogeneous* space is one which is translationally invariant, or the same at every point. An *isotropic* space is one which is rotationally invariant, or the same in every direction. The two are not the same: a space which is everywhere isotropic is necessarily homogeneous, but a space which is homogeneous is not necessarily isotropic. (Consider, for example a space with a uniform electric field: it is translationally invariant but not rotationally invariant.) It is possible to show [21] that the most general metric consistent with homogeneity and isotropy is obtained by multiplying a static spatial geometry with a time-dependent *scale factor* $a(t)$:

$$\begin{aligned} ds^2 &= dt^2 - a^2(t) d\mathbf{x}^2 \\ &= dt^2 - a^2(t) \left[\frac{dr^2}{1 - kr^2} + r^2 d\Omega^2 \right], \end{aligned} \quad (11)$$

where we have expressed the spatial line element in terms of spherical coordinates r, θ, ϕ , and the solid angle is given by the usual $d\Omega^2 = \sin\theta d\theta d\phi$. The constant k defines the curvature of the spacetime, with $k = 0$ corresponding to flat (Euclidean) spatial sections, and $k = \pm 1$ corresponding to positive and negative curvatures, respectively. A

spacetime of this general form is called a *Friedmann-Robertson-Walker* (FRW) spacetime. Likewise, the most general homogeneous, isotropic stress-energy is diagonal, with all of its spatial components identical,

$$T^{\mu}_{\nu} = \begin{pmatrix} \rho(t) & & & \\ & -p(t) & & \\ & & -p(t) & \\ & & & -p(t) \end{pmatrix}, \quad (12)$$

where we identify the energy density ρ and the pressure p from the continuity equation arising from stress-energy conservation,

$$T^{\mu\nu}{}_{;\nu} = \dot{\rho} + 3 \left(\frac{\dot{a}}{a} \right) (\rho + p) = 0. \quad (13)$$

The Einstein field equations then reduce to a set of two coupled, non-linear ordinary differential equations,

$$\begin{aligned} \left(\frac{\dot{a}}{a} \right)^2 + \frac{k}{a^2} &= \frac{8\pi}{3m_{\text{Pl}}^2} \rho, \\ \left(\frac{\ddot{a}}{a} \right) &= -\frac{4\pi}{3m_{\text{Pl}}^2} (\rho + 3p). \end{aligned} \quad (14)$$

The first is called the *Friedmann Equation*, and the second is called the *Raychaudhuri Equation*. Note that the equations for the evolution of the scale factor depend not only on the energy density ρ , but also the pressure p : pressure gravitates! The continuity equation (13) is *not* independent of the Einstein Field Equations (14), but can be derived directly from the Friedmann and Raychaudhuri Equations. The expansion rate \dot{a}/a is called the *Hubble parameter* H :

$$H \equiv \frac{\dot{a}}{a}, \quad (15)$$

and has units of inverse time. A positive Hubble parameter $H > 0$ corresponds to an expanding universe, and a negative Hubble parameter $H < 0$ corresponds to a collapsing universe. (Since our actual universe is expanding, we will specialize to that case.) Minkowski Space can be recovered by assuming a flat geometry $k = 0$, and no expansion, $\dot{a} = 0$. The Hubble parameter sets the fundamental scale of the spacetime, *i.e.* a characteristic time is the *Hubble time* $t \sim H^{-1}$, and likewise the *Hubble length* is $d \sim H^{-1}$. We will see later that the Hubble time sets the scale for the age of the universe, and the Hubble length sets the scale for the size of the observable universe.

The coordinate system (t, \mathbf{x}) is called a *comoving* coordinate system, because observers with constant comoving coordinates are at rest relative to the expansion, *i.e.* two observers with constant separation in comoving coordinates $\Delta \mathbf{x}$ have a physical, or *proper*, separation which increases in proportion to the scale factor

$$\Delta \mathbf{x}_{\text{prop}} = a(t) \Delta \mathbf{x}_{\text{com}}. \quad (16)$$

An important kinematic effect of cosmological expansion is the phenomenon of *cosmological redshift*: we will see later that solutions to the wave equation in an FRW space have constant wavelength in *comoving* coordinates, so that the proper wavelength of (for example) a photon increases in time in proportion to the scale factor

$$\lambda \propto a(t). \quad (17)$$

For a photon emitted at time t_{em} and detected at time t_0 , the redshift z is defined by:

$$(1 + z) \equiv \frac{\lambda_0}{\lambda_{\text{em}}} = \frac{a(t_0)}{a(t_{\text{em}})}. \quad (18)$$

(Here we introduce the convention used frequently in cosmology that a subscript 0 refers to the *current* time, not an initial time.) Note that the cosmological redshift is *not* a Doppler shift caused by the relative velocity of the source and detector, but is an expansion effect: the wavelength of a photon traveling through the spacetime increases because the underlying spacetime is expanding. Another way to look at this is that a photon traveling through an FRW spacetime loses momentum with time,

$$p = h\nu \propto a^{-1}(t). \quad (19)$$

By the equivalence principle, this momentum loss must apply to massive particles as well as photons: *any* particle moving in an expanding FRW spacetime will lose momentum as $p \propto a^{-1}$. For massless particles like photons, this is manifest as a redshift in the wavelength, but it means that a massive particle will asymptotically come to rest relative to the comoving coordinate system. Thus, comoving coordinates represent a preferred reference frame reminiscent of Aristotelian physics: any free body with a “peculiar” velocity relative to the comoving frame will eventually come to rest in that frame.

There are three possibilities for the curvature of the universe: flat ($k = 0$), positively curved ($k = +1$), or negatively curved ($k = -1$). The current value of the Hubble parameter is (from the Hubble Space Telescope Key Project [22]),

$$H_0 = 72 \pm 8 \text{ km/s/Mpc.} \quad (20)$$

Therefore, we can see from the Friedmann Equation (14) that, given the expansion rate H , the curvature is determined by the density:

$$k = a^2 \left(\frac{8\pi}{3m_{\text{Pl}}^2} \rho - H^2 \right). \quad (21)$$

Note that only the *sign* of k is physically important, since any rescaling of k is equivalent to a rescaling of the scale factor a . We define the *critical density* as the density for which $k = 0$, corresponding to a geometrically flat universe,

$$\rho_c \equiv \frac{3m_{\text{Pl}}^2}{8\pi} H^2 \Rightarrow k = 0. \quad (22)$$

For $\rho > \rho_c$, the universe is positively curved and *closed*, with finite volume, and for $\rho < \rho_c$, the universe is negatively curved and *open*, with infinite volume. We express the ratio of the actual density ρ to the critical density ρ_c as the parameter Ω :

$$\Omega \equiv \left(\frac{\rho}{\rho_c} \right) = \frac{8\pi}{3m_{\text{Pl}}^2} \frac{\rho}{H^2}. \quad (23)$$

(Do not confuse the density parameter Ω with the solid angle $d\Omega$ in Eq. 11!) Table I summarizes the relation between density, curvature, and geometry. The density parameter Ω is not in general constant in time, and we can re-write

TABLE I: Cosmological density and curvature

density	curvature	geometry
$\Omega = 1$	$k = 0$	flat
$\Omega > 1$	$k = 1$	closed
$\Omega < 1$	$k = 0$	open

the Friedmann Equation as

$$\Omega(t) = 1 + \frac{k}{(aH)^2}. \quad (24)$$

Since the Hubble parameter is proportional to the inverse time $H \propto t^{-1}$, we see that the time-dependence of Ω is determined by the time dependence of the scale factor $a(t)$. In the next section, we tackle the problem of solving for $a(t)$.

B. Solving the Friedmann Equation

In the previous section, we considered the form and kinematics of FRW spaces, but not the *dynamics*, that is, how does the stress-energy of the universe determine the expansion history? The answer to this question depends on what kind of matter dominates the cosmological stress-energy. In this section, we consider three basic types of cosmological stress-energy: matter, radiation, and vacuum.

The simplest kind of cosmological stress-energy is generically referred to as *matter*. Imagine a comoving box with sides of length L . By *comoving* box, we mean a box whose corners are at rest in a comoving coordinate system, and whose proper dimension is therefore increasing proportional to the scale factor, $L_{\text{prop}} \propto a$. That is, the box is

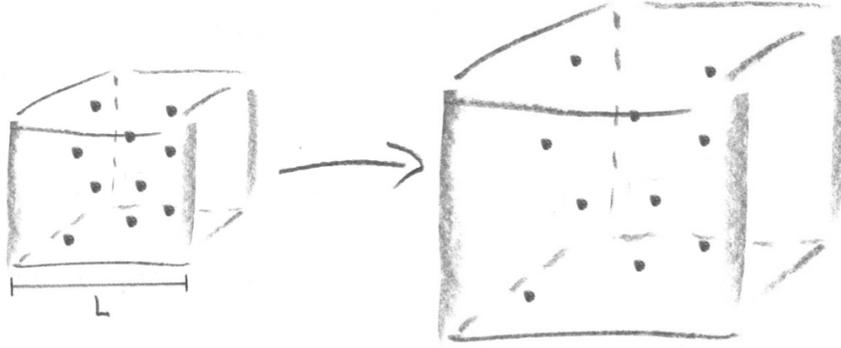


FIG. 2: A comoving box full of matter. The energy density in matter scales inversely with the volume of the box.

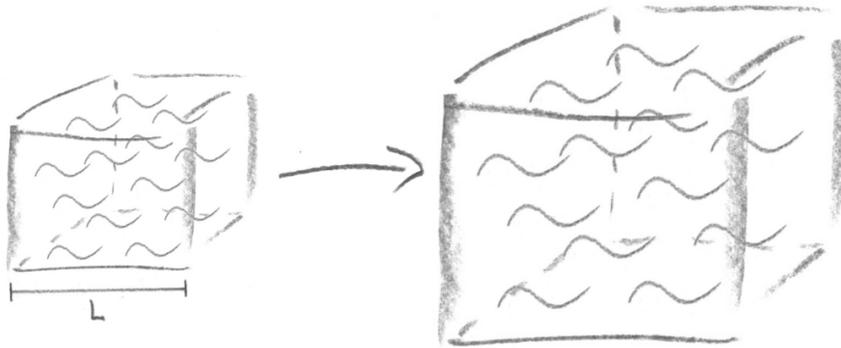


FIG. 3: A comoving box full of radiation. The number density of photons scales inversely with the volume of the box, but the photons also increase in wavelength.

growing with the expansion of the universe. Now imagine the box filled with N particles of mass m , also at rest in the comoving reference frame (Fig. 2). In units where $c = 1$, the relativistic energy density of such a system of particles is given by

$$\rho_m = \frac{MN}{V}, \quad (25)$$

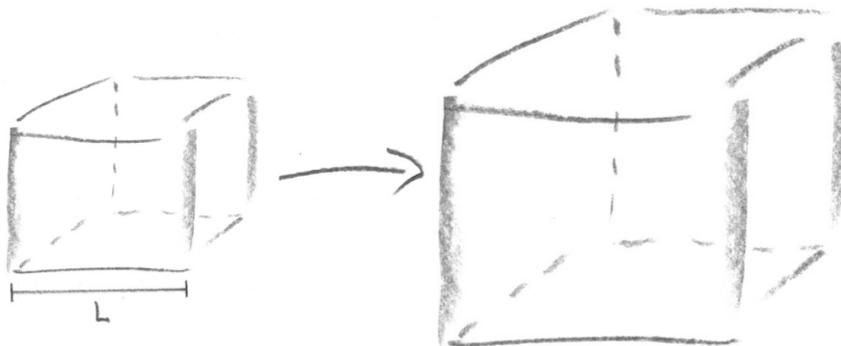


FIG. 4: A comoving box full of vacuum. The energy density of vacuum does not scale at all!

where V is the *proper* volume of the box, $V = L_{\text{prop}}^3 \propto a^3$. Since neither M nor N change with expansion, we have immediately that

$$\rho_{\text{m}} = \frac{MN}{L^3 a^3} \propto a^{-3}, \quad (26)$$

where L is the comoving size of the box. So the proper energy density of massive particles at rest in a comoving volume evolves as the inverse cube of the scale factor. Now imagine the same box filled with N photons with frequency ν (Fig. 3). The energy per photon is $h\nu$, so that the energy density in the box is then

$$\rho_{\gamma} = \frac{Nh\nu}{V}. \quad (27)$$

As in the case of massive particles, the number density of photons in the box redshifts inversely with the proper volume of the box $n = N/V \propto a^{-3}$. But each photon also loses energy through cosmological redshift, $\nu \propto a^{-1}$ (19), so that the total energy density in photons or other massless degrees of freedom, which we generically refer to as *radiation*, redshifts as

$$\rho_{\gamma} \propto a^{-4}. \quad (28)$$

Note also that cosmological redshift immediately gives us a rule for the behavior of a black-body spectrum of radiation with temperature T . Since all photons redshift at exactly the same rate, a system which starts out as a black-body *stays* a black-body, with a temperature that decreases with expansion,

$$T_{\gamma} \propto a^{-1}. \quad (29)$$

The third type of stress-energy which is important in cosmology dates back to Einstein's introduction of a "cosmological constant" to his field equations. If we take the stress-energy $T_{\mu\nu}$ and add a term proportional to the metric, the identity $D_{\nu}g^{\mu\nu} = 0$ means the stress-energy conservation equation (10) is unchanged:

$$D_{\nu}T^{\mu\nu} \rightarrow D_{\nu}(T^{\mu\nu} + \Lambda g^{\mu\nu}) = 0. \quad (30)$$

In our analogy with electromagnetism, this is like adding a constant to the electromagnetic potential, $V'(x) = V(x) + \Lambda$. The constant Λ does not affect local dynamics in any way, but it does affect the cosmology. From Eq. (12), stress-energy of the form $T^{\mu\nu} = \Lambda g^{\mu\nu}$ corresponds to an equation of state

$$p_{\Lambda} = -\rho_{\Lambda}. \quad (31)$$

The continuity equation (13) then reduces to

$$\dot{\rho} + 3 \left(\frac{\dot{a}}{a} \right) (\rho + p) = \dot{\rho} = 0, \quad (32)$$

so that vacuum has a constant energy density, $\rho_{\Lambda} = \text{const}$. A cosmological constant is also frequently referred to as *vacuum energy*, since it is as if we are assigning an energy density to empty space. With this interpretation, a comoving box full of vacuum contains a total amount of energy which *grows* with the expansion of the universe (Fig. 4). This highlights the curious property of General Relativity that, while energy is conserved in a local sense, it is *not* conserved globally. We are creating energy out of nothing!

It is straightforward to solve the Einstein Field Equations for the three basic types of stress-energy. Consider first a matter-dominated universe. We can write the time derivative of the energy density as:

$$\rho_{\text{m}} \propto a^{-3} \Rightarrow \dot{\rho}_{\text{m}} = -3 \left(\frac{\dot{a}}{a} \right) \rho. \quad (33)$$

From the continuity equation (13), we have

$$\dot{\rho} + 3 \left(\frac{\dot{a}}{a} \right) (\rho + p) = 3 \left(\frac{\dot{a}}{a} \right) p = 0. \quad (34)$$

We then have that the pressure of matter vanishes, $p_{\text{m}} = 0$. The matter-dominated Friedmann Equation becomes

$$\left(\frac{\dot{a}}{a} \right)^2 + \frac{k}{a^2} = \frac{8\pi}{3m_{\text{Pl}}^2} \rho \propto a^{-3}. \quad (35)$$

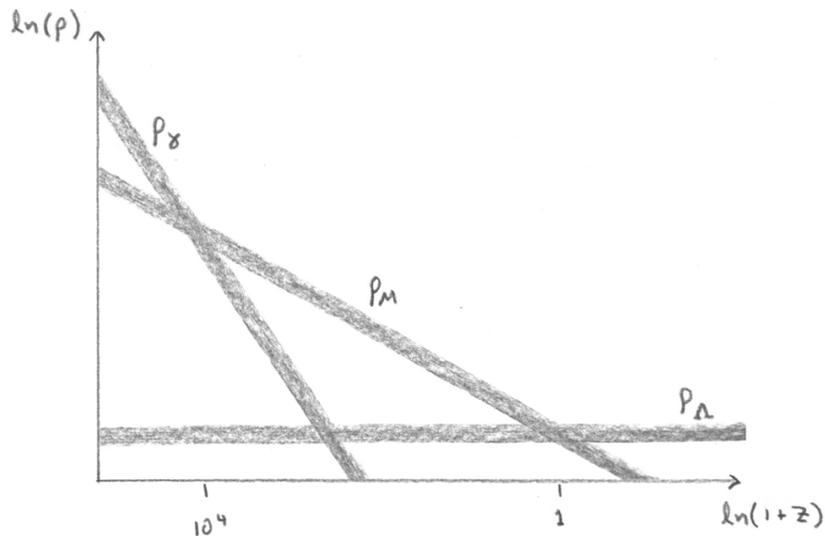


FIG. 5: Schematic diagram of how the three types of stress-energy scale with redshift $1+z \propto a$: at early time, radiation dominates, followed by matter, and finally the universe is dominated by vacuum energy.

In the case of a flat universe, $k = 0$, the solution is especially simple:

$$\left(\frac{\dot{a}}{a}\right)^2 \propto a^{-3} \Rightarrow a(t) \propto t^{2/3}. \quad (36)$$

Similarly, for a radiation dominated universe, the continuity equation implies that

$$\rho_\gamma \propto a^{-4} \Rightarrow p_\gamma = \rho_\gamma/3. \quad (37)$$

Again assuming a flat geometry,

$$\left(\frac{\dot{a}}{a}\right)^2 \propto a^{-4} \Rightarrow a(t) \propto t^{1/2}. \quad (38)$$

Finally, solving the the Friedmann Equation for the vacuum case gives

$$\left(\frac{\dot{a}}{a}\right)^2 \propto \rho_\Lambda = \text{const.} \Rightarrow a(t) \propto e^{Ht}, \quad (39)$$

so that the universe expands exponentially quickly, with a time constant given by the Hubble parameter

$$H = \sqrt{\frac{8\pi}{3m_{\text{Pl}}^2} \rho_\Lambda} = \text{const.} \quad (40)$$

Such a spacetime is called *de Sitter space*.

Note in particular that the energy density in radiation redshifts away more quickly than the energy density in matter, and vacuum energy does not redshift at all, so that a universe with a mix of radiation, matter and vacuum will be radiation-dominated at early times, matter-dominated at later times, and eventually vacuum-dominated (Fig. 5). Note also that for either matter- or radiation-domination, the universe is singular as $t \rightarrow 0$: the universe has finite age! Since the scale factor vanishes at $t = 0$, and the density scales as an inverse power of a , the initial singularity consists of infinite density. Likewise, since temperature also scales inversely with a , the initial singularity is also a point of infinite temperature. We therefore arrive at the standard hot Big Bang picture of the universe: a cosmological singularity at finite time in the past, followed by a hot, radiation-dominated expansion, during which the universe gradually cools as $T \propto a^{-1}$ and the radiation dilutes, followed by a period of matter-dominated expansion during

which galaxies and stars and planets form. Finally, if the vacuum energy is nonzero, it will inevitably dominate, and the universe will enter a state of exponential expansion. Current evidence indicates that the real universe made a transition from matter-domination to vacuum-domination at a redshift of around $z = 1$, or about a billion years ago, so that the densities of the three types of matter today are of order

$$\begin{aligned}\Omega_\Lambda &\simeq 0.7, \\ \Omega_m &\simeq 0.3, \\ \Omega_\gamma &\simeq 10^{-4}.\end{aligned}\tag{41}$$

In the next section, we discuss one important prediction of the hot Big Bang: the presence of a background of relic photons from the early universe, called the *Cosmic Microwave Background*.

C. The Hot Big Bang and the Cosmic Microwave Background

The basic picture of an expanding, cooling universe leads to a number of startling predictions: the formation of nuclei and the resulting primordial abundances of elements, and the later formation of neutral atoms and the consequent presence of a cosmic background of photons, the Cosmic Microwave Background (CMB) [23, 24, 25, 26, 27]. A rough history of the universe can be given as a time line of increasing time and decreasing temperature [28]:

- $T = \infty$, $t = 0$: Big Bang.
- $T \sim 10^{15}$ K, $t \sim 10^{-12}$ sec: Primordial soup of fundamental particles.
- $T \sim 10^{13}$ K, $t \sim 10^{-6}$ sec: Protons and neutrons form.
- $T \sim 10^{10}$ K, $t \sim 3$ min: Nucleosynthesis: nuclei form.
- $T \sim 3000$ K, $t \sim 300,000$ years: Atoms form.
- $T \sim 10$ K, $t \sim 10^9$ years: Galaxies form.
- $T \sim 3$ K, $t \sim 10^{10}$ years: Today.

The epoch at which atoms form, when the universe was at an age of 300,000 years and at a temperature of around 3000 K is oxymoronically referred to as “recombination”, despite the fact that electrons and nuclei had never before “combined” into atoms. The physics is simple: at a temperature of greater than about 3000 K, the universe consisted of an ionized plasma of mostly protons, electrons, and photons, with a few helium nuclei and a tiny trace of lithium. The important characteristic of this plasma is that it was *opaque*, or, more precisely, the mean free path of a photon was a great deal smaller than the Hubble length. As the universe cooled and expanded, the plasma “recombined” into neutral atoms, first the helium, then a little later the hydrogen.

If we consider hydrogen alone, the process of recombination can be described by the Saha equation for the equilibrium ionization fraction X_e of the hydrogen [28]:

$$\frac{1 - X_e}{X_e^2} = \frac{4\sqrt{2}\zeta(3)}{\sqrt{\pi}} \eta \left(\frac{T}{m_e}\right)^{3/2} \exp\left(\frac{13.6 \text{ eV}}{T}\right).\tag{42}$$

Here m_e is the electron mass and 13.6 eV is the ionization energy of hydrogen. The physically important parameter affecting recombination is the density of protons and electrons compared to photons. This is determined by the *baryon asymmetry*,² which is described as the ratio of baryons to photons:

$$\eta \equiv \frac{n_b - n_{\bar{b}}}{n_\gamma} = 2.68 \times 10^{-8} (\Omega_b h^2).\tag{43}$$

Here Ω_b is the baryon density and h is the Hubble constant in units of 100 km/s/Mpc,

$$h \equiv H_0/(100 \text{ km/s/Mpc}).\tag{44}$$

² If there were no excess of baryons over antibaryons, there would be no protons and electrons to recombine, and the universe would be just a gas of photons and neutrinos!

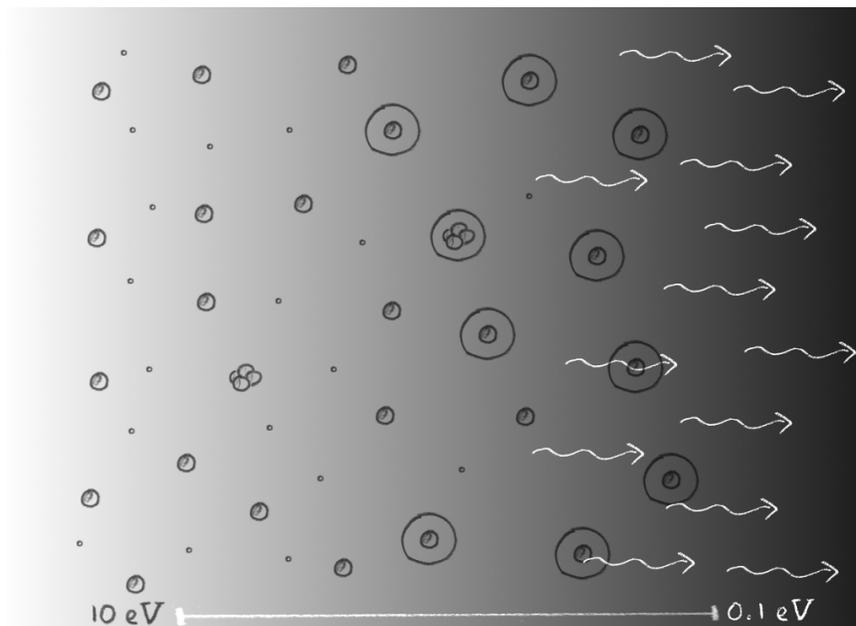


FIG. 6: Schematic diagram of recombination. At early time, the temperature of the universe is above the ionization energy of hydrogen and helium, so that the universe is full of an ionized plasma, and the mean free path for photons is short compared to the Hubble length. At late time, the temperature drops and the nuclei capture the electrons and form neutral atoms. Once this happens, the universe becomes transparent to photons, which free stream from the surface of last scattering.

The most recent result from the WMAP satellite gives $\Omega_b h^2 = 0.02273 \pm 0.00062$ [29]. Recombination happens quickly (i.e., in much less than a Hubble time $t \sim H^{-1}$), but it is not instantaneous. The universe goes from a completely ionized state to a neutral state over a range of redshifts $\Delta z \sim 200$. If we define recombination as an ionization fraction $X_e = 0.1$, we have that the temperature at recombination $T_R = 0.3$ eV.

What happens to the photons after recombination? Once the gas in the universe is in a neutral state, the mean free path for a photon becomes much larger than the Hubble length. The universe is then full of a background of freely propagating photons with a blackbody distribution of frequencies. At the time of recombination, the background radiation has a temperature of $T = T_R = 3000$ K, and as the universe expands the photons redshift, so that the temperature of the photons drops with the increase of the scale factor, $T \propto a(t)^{-1}$. We can detect these photons today. Looking at the sky, this background of photons comes to us evenly from all directions, with an observed temperature of $T_0 \simeq 2.73$ K. This allows us to determine the redshift of recombination,

$$1 + z_R = \frac{a(t_0)}{a(t_R)} = \frac{T_R}{T_0} \simeq 1100. \quad (45)$$

This is the cosmic microwave background. Since by looking at higher and higher redshift objects, we are looking further and further back in time, we can view the observation of CMB photons as imaging a uniform “surface of last scattering” at a redshift of 1100 (Fig. 7).

To the extent that recombination happens at the same time and in the same way everywhere, the CMB will be of precisely uniform temperature. While the observed CMB is highly isotropic, it is not perfectly so. The largest contribution to the anisotropy of the CMB as seen from earth is simply Doppler shift due to the earth’s motion through space. (Put more technically, the motion is the earth’s motion relative to a comoving cosmological reference frame.) CMB photons are slightly blueshifted in the direction of our motion and slightly redshifted opposite the direction of our motion. This blueshift/redshift shifts the temperature of the CMB so the effect has the characteristic form of a “dipole” temperature anisotropy (Fig. 8). The dipole anisotropy, however, is a *local* phenomenon. Any intrinsic, or primordial, anisotropy of the CMB is potentially of much greater cosmological interest. To describe the anisotropy of the CMB, we remember that the surface of last scattering appears to us as a spherical surface at a redshift of 1100. Therefore the natural parameters to use to describe the anisotropy of the CMB sky is as an expansion in spherical

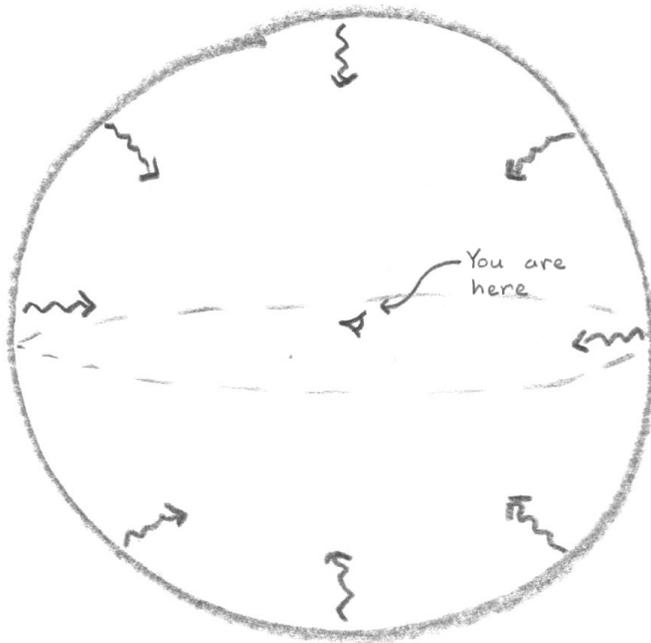


FIG. 7: Cartoon of the last scattering surface. From earth, we see blackbody radiation emitted uniformly from all directions, forming a “sphere” at redshift $z = 1100$.

harmonics $Y_{\ell m}$:

$$\frac{\Delta T}{T} = \sum_{\ell=1}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\theta, \phi). \quad (46)$$

If we assume isotropy, there is no preferred direction in the universe, and we expect the physics to be independent of the index m . We can then define

$$C_{\ell} \equiv \frac{1}{2\ell + 1} \sum_m |a_{\ell m}|^2. \quad (47)$$

The $\ell = 1$ contribution is just the dipole anisotropy,

$$\left(\frac{\Delta T}{T} \right)_{\ell=1} \sim 10^{-3}. \quad (48)$$

The dipole was first measured in the 1970’s by several groups [30, 31, 32]. It was not until more than a decade after the discovery of the dipole anisotropy that the first observation was made of anisotropy for $\ell \geq 2$, by the differential microwave radiometer aboard the Cosmic Background Explorer (COBE) satellite [33], launched in 1990. COBE observed that the anisotropy at the quadrupole and higher ℓ was two orders of magnitude smaller than the dipole:

$$\left(\frac{\Delta T}{T} \right)_{\ell>1} \simeq 10^{-5}. \quad (49)$$

Fig. 8 shows the dipole and higher-order CMB anisotropy as measured by COBE. This anisotropy represents intrinsic fluctuations in the CMB itself, due to the presence of tiny primordial density fluctuations in the cosmological matter present at the time of recombination. These density fluctuations are of great physical interest, since these are the fluctuations which later collapsed to form all of the structure in the universe, from superclusters to planets to graduate students. While the physics of recombination in the homogeneous case is quite simple, the presence of inhomogeneities in the universe makes the situation much more complicated. I describe some of the major effects in a qualitative way here, and refer the reader to the literature for a more detailed technical explanation of the relevant physics [23, 24, 25, 26, 27]. In these lectures, I primarily focus on the current status of the CMB as a probe of inflation, but there is much more to the story.

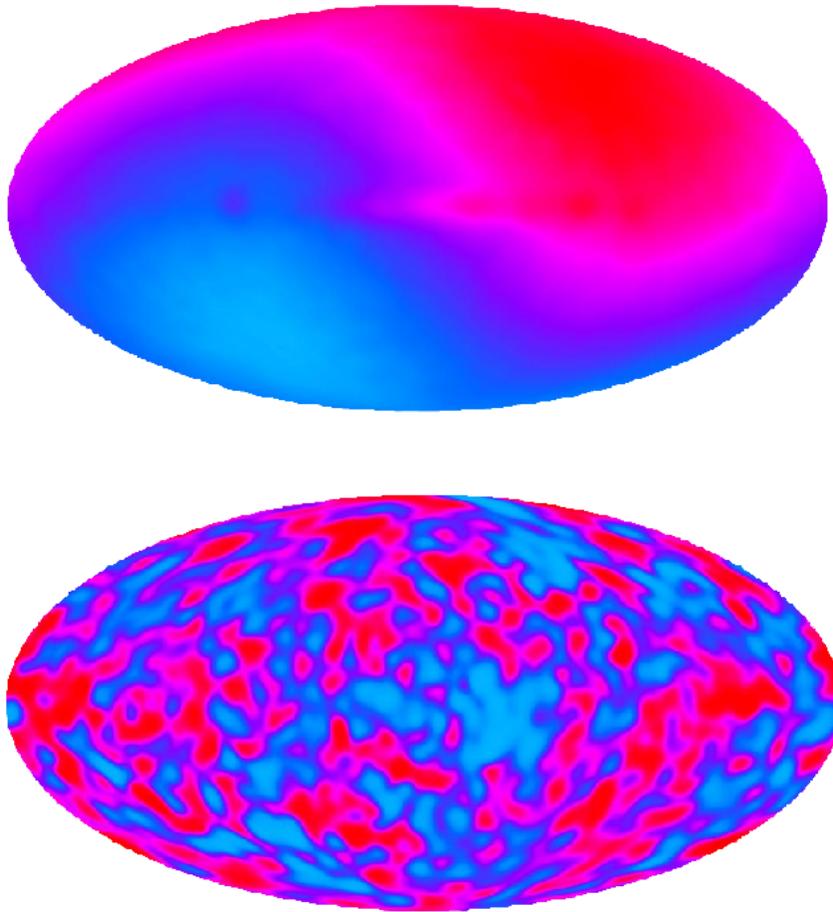


FIG. 8: The COBE measurement of the CMB anisotropy [33]. The top oval is a map of the sky showing the dipole anisotropy $\Delta T/T \sim 10^{-3}$. The bottom oval is a similar map with the dipole contribution and emission from our own galaxy subtracted, showing the anisotropy for $\ell > 1$, $\Delta T/T \sim 10^{-5}$. (Figure courtesy of the COBE Science Working Group.)

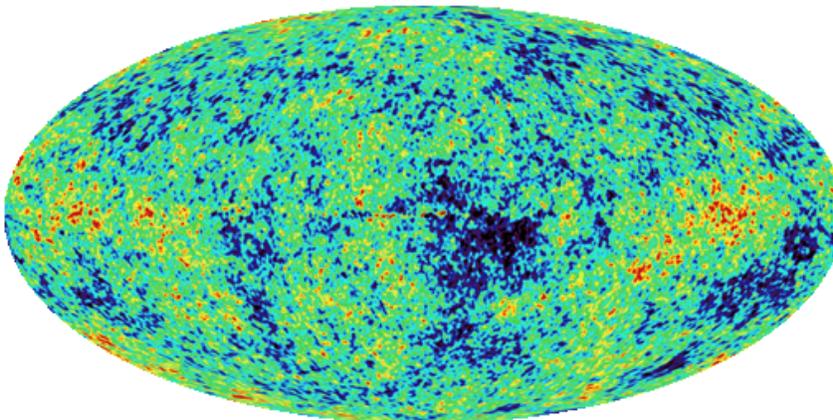


FIG. 9: The WMAP measurement of the CMB anisotropy [34]. (Figure courtesy of the WMAP Science Working Group.) WMAP measured the anisotropy with much higher sensitivity and resolution than COBE.

The simplest contribution to the CMB anisotropy from density fluctuations is just a gravitational redshift, known as the *Sachs-Wolfe effect* [35]. A photon coming from a region which is slightly denser than the average will have a slightly larger redshift due to the deeper gravitational well at the surface of last scattering. Conversely, a photon coming from an underdense region will have a slightly smaller redshift. Thus we can calculate the CMB temperature anisotropy due to the slightly varying Newtonian potential Φ from density fluctuations at the surface of last scattering:

$$\frac{\delta T}{T} = \frac{1}{3} [\Phi_{\text{em}} - \Phi_{\text{obs}}], \quad (50)$$

where Φ_{em} is the potential at the point the photon was emitted on the surface of last scattering, and Φ_{obs} is the potential at the point of observation, which can be treated as a constant. The factor $1/3$ is a General Relativistic correction. This simple kinematic contribution to the CMB anisotropy is dominant on large angular scales, corresponding to multipoles $\ell < 100$. However, the amount of information we can gain from these multipoles is limited by an intrinsic source of error called *cosmic variance*. Cosmic variance is a result of the statistical nature of the primordial power spectra: since we have only one universe to measure, we have only one realization of the random field of density perturbations, and therefore there is an inescapable $1/\sqrt{N}$ uncertainty in our ability to reconstruct the primordial power spectrum, where N is the number of independent wave modes which will fit inside the horizon of the universe! On very large angular scales, this problem becomes acute, and we can write the cosmic variance error on any given C_ℓ as

$$\frac{\Delta C_\ell}{C_\ell} = \frac{1}{\sqrt{2\ell + 1}}, \quad (51)$$

which comes from the fact that any C_ℓ is represented by $2\ell + 1$ independent amplitudes $a_{\ell m}$. Even a perfect observation of the CMB can only approximately measure the true power spectrum — the errors in the WMAP data, for example, are dominated by cosmic variance out to $\ell \sim 400$ (Fig. 10).

For fluctuation modes on smaller angular scales, more complicated physics comes into play. The dominant process that occurs on short wavelengths is *acoustic oscillations* in the baryon/photon plasma. The idea is simple: matter tends to collapse due to gravity onto regions where the density is higher than average, so the baryons “fall” into overdense regions. However, since the baryons and the photons are still strongly coupled, the photons tend to resist this collapse and push the baryons outward. The result is “ringing”, or oscillatory modes of compression and rarefaction in the gas due to density fluctuations. The gas heats as it compresses and cools as it expands, which creates fluctuations in the temperature of the CMB. This manifests itself in the C_ℓ spectrum as a series of peaks and valleys (Fig. 10). The specific shape and location of the acoustic peaks is created by complicated but well-understood physics, involving a large number of cosmological parameters. The presence of acoustic peaks in the CMB was first suggested by Sakharov [36], and later calculated by Sunyaev and Zel’dovich [37, 38] and Peebles and Yu [39]. The complete linear theory of CMB fluctuations was worked out by Ma and Bertschinger in 1995 [40]. The shape of the CMB multipole spectrum depends, for example, on the baryon density Ω_b , the Hubble constant H_0 , the densities of matter Ω_m and cosmological constant Ω_Λ , the amplitude of primordial gravitational waves, and the redshift z_{re} at which the first generation of stars ionized the intergalactic medium. This makes interpretation of the spectrum something of a complex undertaking, but it also makes it a sensitive probe of cosmological models.

In addition to anisotropy in the temperature of the CMB, the photons coming from the surface of last scattering are expected to be weakly polarized due to the presence of perturbations [41, 42]. This polarization is much less well measured than the temperature anisotropy, but it has been detected by WMAP and by a number of ground- and balloon-based measurements [43, 44, 45, 46, 47]. Measurement of polarization promises to greatly increase the amount of information it is possible to extract from the CMB. Of particular interest is the odd-parity, or *B-mode* component of the polarization, the only primordial source of which is gravitational waves, and thus provides a clean signal for detection of these perturbations. The B-mode has yet to be detected by any measurement.

III. THE FLATNESS AND HORIZON PROBLEMS

We have so far considered two types of cosmological mass-energy – matter and radiation – and solved the Friedmann Equation for the simple case of a flat universe. What about the more general case? In this section, we consider non-flat universes with general contents. We introduce two related questions which are not explained by the standard Big Bang cosmology: why is the universe so close to flat today, and why is it so large?

We can describe a general homogeneous, isotropic mass-energy by its equation of state

$$p = w\rho, \quad (52)$$

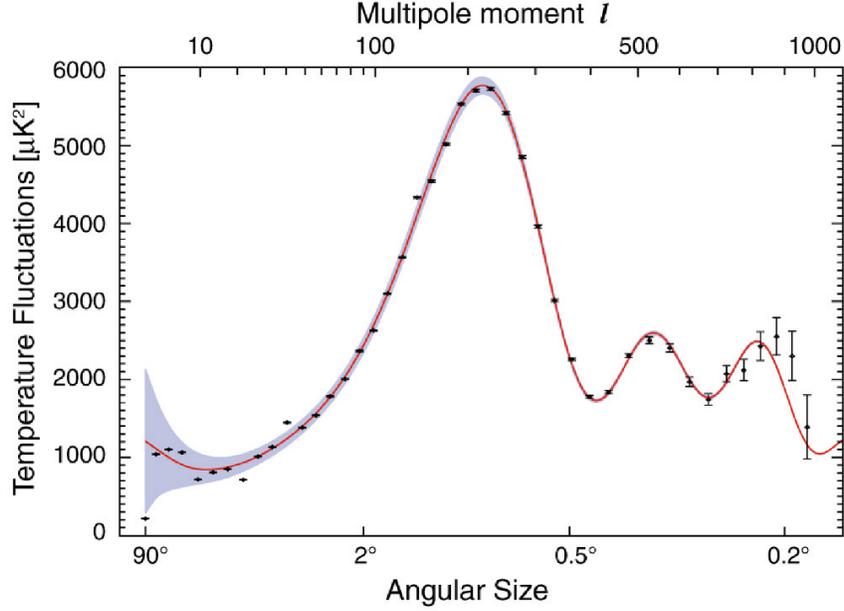


FIG. 10: The C_ℓ spectrum for the CMB as measured by WMAP, showing the peaks characteristic of acoustic oscillations. The gray shaded region represents the uncertainty due to cosmic variance. (Figure courtesy of the WMAP Science Working Group.)

so that pressureless matter corresponds to $w = 0$, and radiation corresponds to $w = 1/3$. We will consider only the case of constant equation of state, $w = \text{const}$. From the continuity equation, we have

$$\dot{\rho} + 3(1+w)\frac{\dot{a}}{a}\rho = 0, \quad (53)$$

with solution

$$\rho \propto a^{-3(1+w)}. \quad (54)$$

The Friedmann Equation for a flat universe is then

$$\left(\frac{\dot{a}}{a}\right)^2 \propto a^{-3(1+w)}, \quad (55)$$

so that the scale factor increases as a power-law in time,

$$a(t) \propto t^{2/3(1+w)}. \quad (56)$$

What about the evolution of a non-flat universe? Analytic solutions for $a(t)$ in the $k \neq 0$ case can be found in cosmology textbooks. For our purposes, it is sufficient to consider the time-dependence of the density parameter Ω . From Eqs. (14, 23, 24) it is not too difficult to show that the density parameter evolves with the scale factor a as:

$$\frac{d\Omega}{d \ln a} = (1 + 3w)\Omega(\Omega - 1). \quad (57)$$

Proof is left as an exercise for the reader. Note that a flat universe, $\Omega = 1$ remains flat at all times, but in a non-flat universe, the density parameter Ω is a time-dependent quantity, with the evolution determined by the equation of state parameter w . For matter ($w = 0$) or radiation ($w = 1/3$), the prefactor in Eq. (57) is positive,

$$1 + 3w > 0, \quad (58)$$

which means a flat universe is an *unstable* fixed point:

$$\frac{d|\Omega - 1|}{d \ln a} > 0, \quad (1 + 3w) > 0. \quad (59)$$

Any deviation from a flat geometry is amplified by the subsequent cosmological expansion, so a nearly flat universe today is a highly fine-tuned situation. The WMAP5 CMB measurement tells us the universe is flat to within a few percent, $|\Omega_0 - 1| < 0.02$ [29, 48]. If we are very conservative and take a limit on the density today as $\Omega_0 = 1 \pm 0.05$, that means that at recombination, when the CMB was emitted, $\Omega_{\text{rec}} = 1 \pm 0.0004$, and at the time of primordial nucleosynthesis, $\Omega_{\text{nuc}} = 1 \pm 10^{-12}$. Why did the universe start out so incredibly close to flat? The standard Big Bang cosmology provides no answer to this question, which we call the *flatness problem*.

There is a second, related problem with the standard Big Bang picture, arising from the finite age of the universe. Because the universe has a finite age, photons can only have traveled a finite distance in the time since the Big Bang. Therefore, the universe has a *horizon*: the further out in space we look, the further back in time we see. If we could look far enough out in any direction, past the surface of last scattering, we would be able to see the Big Bang itself, and beyond that we can see no further. Every observer in an FRW spacetime sees herself at the center of a spherical horizon which defines her observable universe. To calculate the size of our horizon, we use the fact that photons travel on paths of zero proper length:

$$ds^2 = dt^2 - a^2(t) |d\mathbf{x}|^2 = 0, \quad (60)$$

so that the comoving distance $|d\mathbf{x}|$ traversed by a photon in time dt is

$$|d\mathbf{x}| = \frac{dt}{a(t)}. \quad (61)$$

Therefore, the size of the cosmological horizon at time t after the Big Bang is

$$d_{\text{H}}(t) = \int_0^t \frac{dt'}{a(t')}. \quad (62)$$

To convert comoving length to proper length, we just multiply by $a(t)$, so that the proper horizon size is

$$d_{\text{H}}^{\text{prop}}(t) = a(t) d_{\text{H}}^{\text{com}}(t). \quad (63)$$

Normalizing $a(t_0) = 1$, the horizon size of a 14-billion year-old flat, matter-dominated universe is $d_{\text{H}} = 3t_0 \sim 13$ Gpc.

To see why the presence of a horizon is a problem for the standard Big Bang, we examine the causal structure of an FRW universe. Take the FRW metric

$$ds^2 = dt^2 - a^2(t) |d\mathbf{x}|^2, \quad (64)$$

and re-write it in terms of a redefined clock, the *conformal time* τ :

$$ds^2 = a^2(\tau) [d\tau^2 - |d\mathbf{x}|^2]. \quad (65)$$

Conformal time is a “clock” which slows down with the expansion of the universe,

$$d\tau = \frac{dt}{a(t)}, \quad (66)$$

so that the comoving horizon size is just the age of the universe in conformal time

$$d_{\text{H}}(t) = \int_0^t \frac{dt'}{a(t')} = \int_0^\tau d\tau' = \tau. \quad (67)$$

The conformal metric is useful because the expansion of the spacetime is factored into a static metric multiplied by a time-dependent conformal factor $a(\tau)$, so that photon geodesics are simply described by $d|\mathbf{x}| = d\tau$. In a diagram of τ versus $|\mathbf{x}|$, photons travel on 45° angles. (Note that this is true even for curved spacetimes!) We can draw light cones and infer causal relationships with the expansion factored out, in a manner identical to the usual case of Minkowski Space.

There is one major difference between FRW and Minkowski: an FRW spacetime has a *finite age*. Therefore, unlike the case of Minkowski Space, which has an infinite past, an FRW spacetime is “chopped off” at some finite past time $\tau = 0$ (Fig.11). The initial singularity is a surface of constant conformal time, and it is easy to see from Eq. (67) that our horizon size is the width of our past light cone projected on the surface defined by the initial singularity. This is a very different picture from the notion many people (even scientists) have of the Big Bang, which is something akin to an explosion, with the universe initially a cosmic “egg” of zero size. On the contrary, in the case of a flat or open

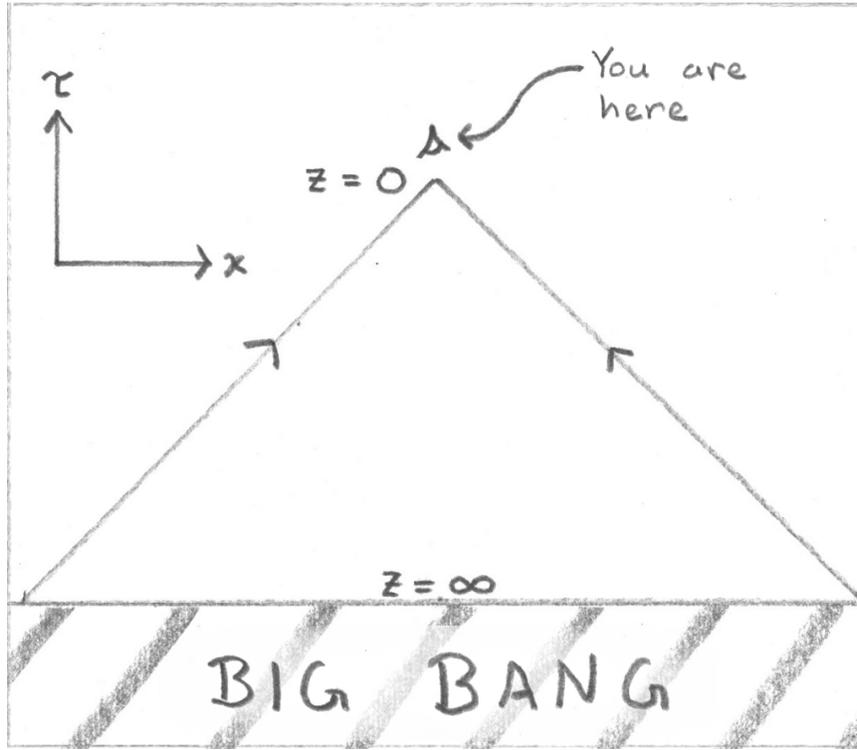


FIG. 11: A conformal diagram of a Friedmann-Robertson-Walker space. The FRW space is causally identical to Minkowski Space, except that it is not past-infinite, so that past light cones are “cut off” at the Big Bang, which is a spatially infinite surface at time $t = 0$.

universe, the universe is spatially infinite an infinitesimal amount of time after the initial singularity: the Big Bang happens everywhere at once in an infinite space! Our *observable* universe is finite because we can only see a small patch of the much larger cosmos.³ Closed universes are spatially finite, but are still much larger in extent than our observable patch. The key point is that two events on the conformal spacetime diagram are causally connected only if they share a causal past: that is, if their past light cones overlap.

Consider two points on the CMB sky 180° degrees apart (Fig. 12). Their past light cones do not overlap, and the two points are causally *disconnected*. Those two points on the surface of last scattering occupy completely separate, disconnected observable universes. How did these points reach the observed thermal equilibrium to a few parts in 10^5 if they never shared a causal past? This apparent paradox is called the *horizon problem*: the universe somehow reached nearly perfect equilibrium on scales much larger than the size of any local horizon. From the Friedmann Equation, it is easy to show that the horizon problem and the flatness problem are related: consider a comoving length scale λ . It is easy to show that for $w = \text{const.}$, the ratio of λ to the horizon size d_H is related to the curvature by a conservation law

$$\left(\frac{\lambda}{d_H}\right)^2 |\Omega - 1| = \text{const.} \quad (68)$$

Proof is left as an exercise for the reader. Therefore, for a universe evolving away from flatness,

$$\frac{d|\Omega - 1|}{d \ln a} > 0, \quad (69)$$

³ Of course, this is an idealization, and the actual universe could well have a nontrivial global topology, even if it is locally flat, as long as the scale of the overall manifold is much larger than our horizon size [49, 50].

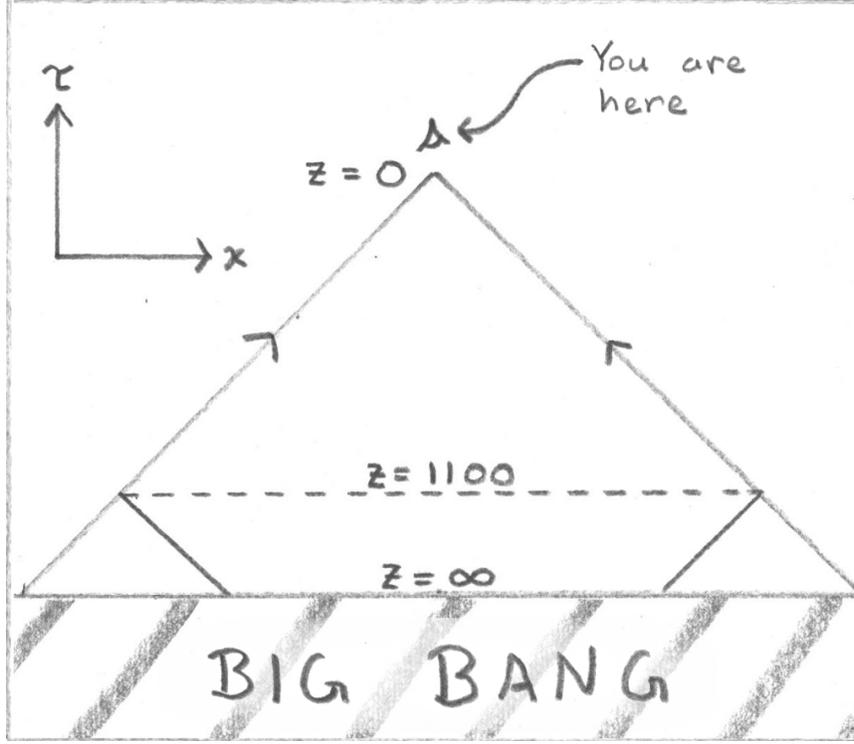


FIG. 12: A conformal diagram of the Cosmic Microwave Background. Two points on opposite sides of the sky are causally separate, since their past light cones do not intersect.

the horizon size gets bigger in comoving units

$$\frac{d}{d \ln a} \left(\frac{\lambda}{d_H} \right) < 0. \quad (70)$$

That is, more and more space “falls into” the horizon, or becomes causally connected, at late times.

What would be required to have a universe which evolves *toward* flatness, rather than away from it? From Eq. (57), we see that having $1 + 3w$ negative will do the trick,

$$\frac{d|\Omega - 1|}{d \ln a} < 0, \quad (1 + 3w) < 0. \quad (71)$$

Therefore, if the energy density of the universe is dominated not by matter or radiation, but by something with sufficiently negative pressure, $p < -\rho/3$, a curved universe will become flatter with time. From the Raychaudhuri Equation (14), we see that the case of $p < -\rho/3$ is exactly equivalent to an accelerating expansion:

$$\frac{\ddot{a}}{a} \propto -(1 + 3w) > 0, \quad (1 + 3w) < 0. \quad (72)$$

If the expansion of the universe is slowing down, as is the case for matter- or radiation-domination, the curvature evolves away from flatness. But if the expansion is speeding up, the universe gets flatter. From Eq. (68), we see that this negative pressure solution also solves the horizon problem, since accelerating expansion means that the horizon size is shrinking in comoving units:

$$\frac{d}{d \ln a} \left(\frac{\lambda}{d_H} \right) < 0, \quad (1 + 3w) < 0. \quad (73)$$

When the expansion accelerates, distances initially smaller than the horizon size are “redshifted” to scales larger than the horizon at late times. Accelerating cosmological expansion is called *inflation*.

The simplest example of an accelerating expansion from a negative pressure fluid is the case of vacuum energy we considered in Section (IIB), for which the scale factor increases exponentially,

$$a \propto e^{Ht}. \quad (74)$$

For such expansion, the universe is driven exponentially toward a flat geometry,

$$\frac{d \ln \Omega}{d \ln a} = 2(1 - \Omega). \quad (75)$$

We can see that the horizon problem is also solved by looking at the conformal time:

$$d\tau = \frac{dt}{a(t)} = e^{-Ht} dt, \quad (76)$$

so that

$$\tau = -\frac{1}{H} e^{-Ht} = -\frac{1}{aH}. \quad (77)$$

The conformal time during the inflationary period is *negative*, tending toward zero at late time. Therefore, if we have a period of inflationary expansion prior to the early epoch of radiation-dominated expansion, inflation takes place in negative conformal time, and conformal time $\tau = 0$ represents not the initial singularity but the transition from the inflationary expansion to radiation domination. The initial singularity is pushed back into negative conformal time, and can be pushed arbitrarily far depending on the duration of inflation. Figure 13 shows the causal structure of an inflationary spacetime. The past light cones of two points on the CMB sky do not intersect at $\tau = 0$, but inflation provides a “sea” of negative conformal time, which allows those points to share a causal past. In this way, inflation solves the horizon problem.

In more realistic models of inflation in the early universe, the energy density is approximately, but not exactly, constant, and the expansion is approximately, but not exactly, exponential. In such quasi-de Sitter spaces, the qualitative picture above still holds, and inflation provides a clean and compelling explanation for the peculiar boundary conditions for our universe. In the next section, we discuss how to construct more detailed models of inflation in field theory.

IV. INFLATION FROM SCALAR FIELDS

The example of de Sitter evolution we considered in Section III gives a good qualitative picture of how inflation, or accelerated expansion, solves the horizon and flatness problems of the standard Big Bang cosmology. However, this leaves open the question: what physics is responsible for the accelerated expansion at early times? It cannot be Einstein’s cosmological constant, simply because a universe dominated by vacuum energy *stays* dominated by vacuum energy for the infinite future, since in a de Sitter background matter ($\rho \propto a^{-3}$) and radiation ($\rho \propto a^{-4}$) are diluted exponentially quickly. Therefore, we will never reach a radiation-dominated phase, and we will never see a hot Big Bang. In order to transition from an inflating phase to a thermal equilibrium, radiation-dominated phase, the vacuum-like energy during inflation must be time-dependent. We model this dynamics with a scalar field ϕ , for which we assume the following action:

$$S = \int d^4x \sqrt{-g} \mathcal{L}_\phi, \quad (78)$$

where $g \equiv \text{Det}(g_{\mu\nu})$ is the determinant of the metric and the Lagrangian for the field ϕ is

$$\mathcal{L}_\phi = \frac{1}{2} g^{\mu\nu} \partial_\mu \phi \partial_\nu \phi - V(\phi). \quad (79)$$

Comparing the action (78) and the Lagrangian (79) with their Minkowski counterparts illustrates how we generalize a classical field theory to curved spacetime:

$$S_{\text{Minkowski}} = \int d^4x \left[\frac{1}{2} \eta^{\mu\nu} \partial_\mu \phi \partial_\nu \phi - V(\phi) \right]. \quad (80)$$

The metric appears in two places in the curved-spacetime action: First, it appears in the measure of volume in the four-space, d^4x , where the determinant of the metric takes the role of the Jacobian for arbitrary coordinate

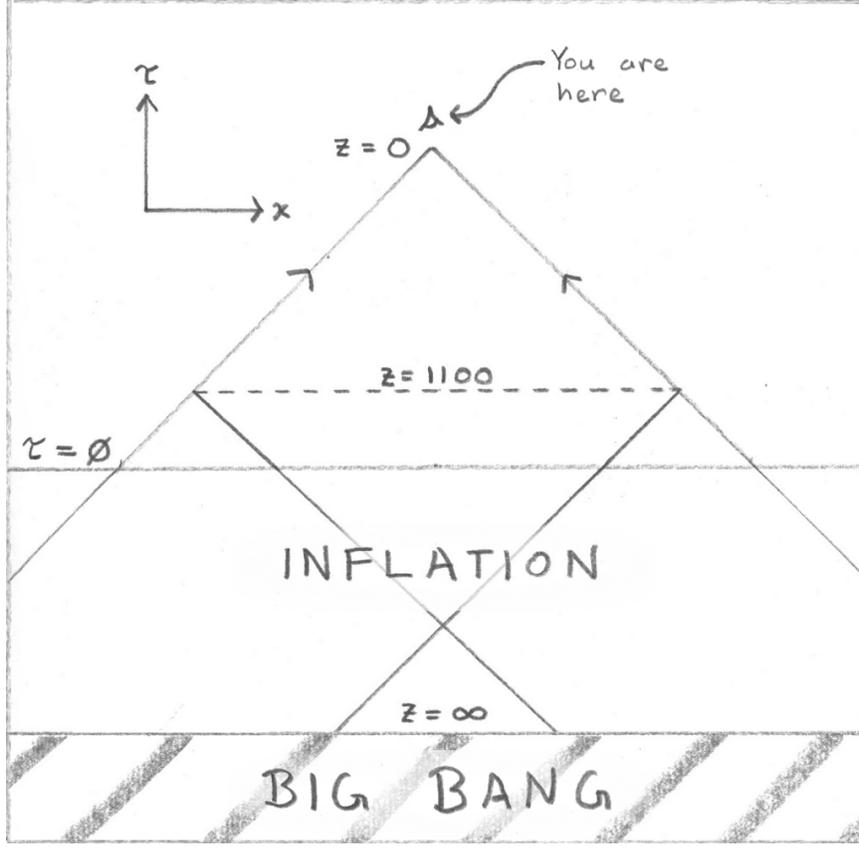


FIG. 13: A conformal diagram of light cones in an inflationary universe. Inflation ends in reheating at conformal time $\tau = 0$, which is the onset of the radiation-dominated expansion of the hot Big Bang. However, inflation provides a “sea” of negative conformal time, which allows the past light cones of events at the last scattering surface to overlap.

transformations, $x \rightarrow x'$. Second, the metric appears in the kinetic term for the scalar field, where we replace the Minkowski metric $\eta^{\mu\nu}$ with the general metric $g^{\mu\nu}$.

The action (78) is not the most general assumption we could make, as we can see by writing the full action including gravity,

$$S_{\text{tot}} = \int d^4x \sqrt{-g} \left[\frac{m_{\text{Pl}}^2}{16\pi} R + \mathcal{L}_\phi \right]. \quad (81)$$

Here R is the Ricci Scalar, composed of the metric and its derivatives. Variation of the first term in the action results in the Einstein Field Equation (8). Such a *minimally coupled* theory assumes that there is no direct coupling between the field and the metric, which would be represented in a more general action by terms which mix R and ϕ . In practice, many such non-minimally coupled theories can be transformed to a minimally coupled form by a field redefinition. We could also write a more general theory by modifying the scalar field Lagrangian (79) to contain non-canonical kinetic terms,

$$\mathcal{L}_\phi = F(\phi, g^{\mu\nu} \partial_\mu \phi \partial_\nu \phi) - V(\phi). \quad (82)$$

where $F()$ is some function of the field and its derivatives. Such Lagrangians appear frequently in models of inflation based on string theory, and are a topic of considerable current research interest. We could also complicate the gravitational sector by replacing the Ricci scalar R with a more complicated function $f(R)$. An example of such a model is the inflation model of Starobinsky [51], which can be reduced to the form (78) through a field redefinition. We could also introduce multiple scalar fields.

Here we will confine ourselves for simplicity to a canonical Lagrangian (79) of a single scalar field, for which the

only adjustable quantity is the choice of potential $V(\phi)$. For simplicity, we assume a flat spacetime,

$$g_{\mu\nu} = \begin{pmatrix} 1 & & & \\ & -a^2(t) & & \\ & & -a^2(t) & \\ & & & -a^2(t) \end{pmatrix}, \quad (83)$$

and the equation of motion for the field ϕ with a Lagrangian given by Eq. (79) is:

$$\ddot{\phi} + 3H\dot{\phi} - \nabla^2\phi + \frac{\delta V}{\delta\phi} = 0, \quad (84)$$

where an overdot indicates a derivative with respect to the coordinate time t , and $H = \dot{a}/a$ is the Hubble parameter. We will be particularly interested in the homogeneous mode of the field, for which the gradient term vanishes, $\nabla\phi = 0$, so that the functional derivative $\delta V/\delta\phi$ simplifies to an ordinary derivative, and the equation of motion simplifies to⁴

$$\ddot{\phi} + 3H\dot{\phi} + V'(\phi) = 0. \quad (85)$$

The stress-energy for a scalar field is given by

$$T_{\mu\nu} = \partial_\mu\phi\partial_\nu\phi - g_{\mu\nu}\mathcal{L}_\phi, \quad (86)$$

and, for a homogeneous field, it takes the form of a perfect fluid with energy density ρ and pressure p , with

$$\begin{aligned} \rho &= \frac{1}{2}\dot{\phi}^2 + V(\phi), \\ p &= \frac{1}{2}\dot{\phi}^2 - V(\phi). \end{aligned} \quad (87)$$

We see that the de Sitter limit, $p \simeq -\rho$, is just the limit in which the potential energy of the field dominates the kinetic energy, $\dot{\phi}^2 \ll V(\phi)$. This limit is referred to as *slow roll*, and under such conditions the universe expands quasi-exponentially,

$$a(t) \propto \exp\left(\int H dt\right) \equiv e^{-N}, \quad (88)$$

where it is conventional to define the number of e-folds N with the sign convention

$$dN \equiv -H dt, \quad (89)$$

so that N is large in the far past and decreases as we go forward in time and as the scale factor a increases.

This can be made quantitative by plugging the energy and pressure (87) into the Friedmann Equation

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi}{3m_{\text{Pl}}^2} \left[\frac{1}{2}\dot{\phi}^2 + V(\phi)\right], \quad (90)$$

and the Raychaudhuri Equation, which we write in the convenient form

$$\left(\frac{\ddot{a}}{a}\right) = -\frac{4\pi}{3m_{\text{Pl}}^2} (\rho + 3p) = H^2 (1 - \epsilon). \quad (91)$$

⁴ The astute reader may well ask: if we are claiming inflation is a solution to the problems of flatness and homogeneity in the universe, why are we assuming flatness and homogeneity from the outset? The answer is that, as long as inflation gets started *somehow* and goes on for long enough, the late-time behavior of the field ϕ will always be described by Eq. (85). We will see later that we only have observational access to the *end* of the inflationary period, and therefore a consistent theory of initial conditions is not required for investigating the observational consequences of inflation.

Here H^2 is given in terms of ϕ by the Friedmann Equation (90), and the parameter ϵ specifies the equation of state,

$$\epsilon \equiv \frac{3}{2} \left(\frac{p}{\rho} + 1 \right) = \frac{4\pi}{m_{\text{Pl}}^2} \left(\frac{\dot{\phi}}{H} \right)^2. \quad (92)$$

It is a straightforward exercise to show that ϵ is related to the evolution of the Hubble parameter by

$$\epsilon = -\frac{d \ln H}{d \ln a} = \frac{1}{H} \frac{dH}{dN}, \quad (93)$$

where N is the number of e-folds (89). This is a useful parameterization because the condition for accelerated expansion $\ddot{a} > 0$ is simply equivalent to $\epsilon < 1$. The de Sitter limit $p \rightarrow -\rho$ is equivalent to $\epsilon \rightarrow 0$, so that the potential $V(\phi)$ dominates the energy density, and

$$H^2 \simeq \frac{8\pi}{3m_{\text{Pl}}^2} V(\phi). \quad (94)$$

We make the additional approximation that the friction term in the equation of motion (85) dominates,

$$\ddot{\phi} \ll 3H\dot{\phi}, \quad (95)$$

so that the equation of motion for the scalar field is approximately

$$3H\dot{\phi} + V'(\phi) \simeq 0. \quad (96)$$

Equation (96) together with the Friedmann Equation (94) are together referred to as the *slow roll approximation*. The condition (95) can be expressed in terms of a second dimensionless parameter, conventionally defined as

$$\eta \equiv -\frac{\ddot{\phi}}{H\dot{\phi}} = \epsilon + \frac{1}{2\epsilon} \frac{d\epsilon}{dN}. \quad (97)$$

The parameters ϵ and η are referred to as *slow roll parameters*, and the slow roll approximation is valid as long as both are small, $\epsilon, |\eta| \ll 1$. It is not obvious that this will be a valid approximation for situations of physical interest: η need *not* be small for inflation to take place. Inflation takes place when $\epsilon < 1$, regardless of the value of η . We later demonstrate explicitly that slow roll does in fact hold for interesting choices of inflationary potential. In the limit of slow roll, we can use Eqs. (94, 96) to write the parameter ϵ approximately as

$$\epsilon = \frac{4\pi}{m_{\text{Pl}}^2} \left(\frac{\dot{\phi}}{H} \right)^2 \simeq \frac{m_{\text{Pl}}^2}{16\pi} \left(\frac{V'(\phi)}{V(\phi)} \right)^2. \quad (98)$$

The inflationary limit, $\epsilon \ll 1$ is then just equivalent to a field evolving on a flat potential, $V'(\phi) \ll V(\phi)$. The second slow roll parameter η can likewise be written approximately as:

$$\begin{aligned} \eta &= -\frac{\ddot{\phi}}{H\dot{\phi}} \\ &\simeq \frac{m_{\text{Pl}}^2}{8\pi} \left[\frac{V''(\phi)}{V(\phi)} - \frac{1}{2} \left(\frac{V'(\phi)}{V(\phi)} \right)^2 \right], \end{aligned} \quad (99)$$

so that the curvature V'' of the potential must also be small for slow roll to be a valid approximation. Similarly, we can write number of e-folds as a function $N(\phi)$ of the field as:

$$\begin{aligned} N &= -\int H dt = -\int \frac{H}{\dot{\phi}} d\phi = \frac{2\sqrt{\pi}}{m_{\text{Pl}}} \int \frac{d\phi}{\sqrt{\epsilon}} \\ &\simeq \frac{8\pi}{m_{\text{Pl}}^2} \int_{\phi_e}^{\phi} \frac{V(\phi)}{V'(\phi)} d\phi, \end{aligned} \quad (100)$$

The limits on the last integral are defined such that ϕ_e is a fixed field value, which we will later take to be the end of inflation, and N increases as we go *backward* in time, representing the number of e-folds of expansion which take place between field value ϕ and ϕ_e .

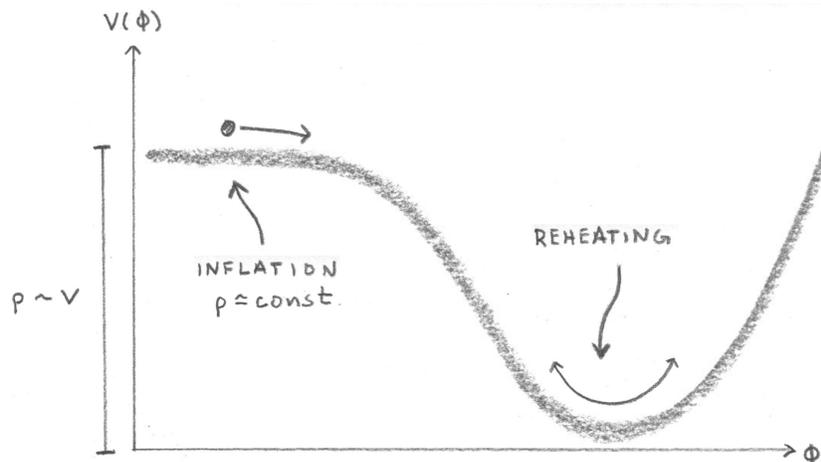


FIG. 14: A schematic of the potential for inflation. Inflation takes place on the region of the potential which is sufficiently “flat”, and reheating takes place near the true vacuum for the field.

The qualitative picture of scalar field-driven inflation is that of a phase transition with order parameter given by the field ϕ . At early times, the energy density of the universe is dominated by the field ϕ which is slowly evolving on a nearly constant potential, so that it approximates a cosmological constant (Fig. 14). During this period, the universe is exponentially driven toward flatness and homogeneity. Inflation ends as the potential steepens and the field begins to oscillate about its vacuum state at the minimum of the potential. At this point, we have an effectively zero-temperature scalar in a state of coherent oscillation about the minimum of the potential, and the universe is a huge Bose-Einstein condensate: hardly a hot Big Bang! In order to transition to a radiation-dominated hot Big Bang cosmology, the energy in the inflaton field must decay into Standard Model particles, a process generically termed *reheating*. This process is model-dependent, but it typically happens very rapidly. Note that the field ϕ need not be a fundamental field like a Higgs boson (although it could in fact be fundamental). *Any* order parameter for a phase transition will do, as long as it has the quantum numbers of vacuum, and the effective potential has the correct properties. The inflaton ϕ could well be a scalar composite of more fundamental degrees of freedom, the coordinate of a brane in a higher-dimensional compactification from string theory, a supersymmetric modulus, or something even more exotic. The simple single-field picture we discuss here is therefore an effective representation of a large variety of underlying fundamental theories. All of the physics important to inflation is contained in the shape of the potential $V(\phi)$. (The details of the underlying theory *are* important for understanding the epoch of reheating, since the reheating process depends crucially on the specific couplings of the inflaton to the other degrees of freedom in the theory.)

How long does inflation need to go on in order to solve the flatness and horizon problems? We use a thermodynamic argument, which rests on a simple fact about cosmological expansion: as long as there are no decays or annihilations of massive particles, all other interactions conserve photon number, so that the number of photons in a comoving volume is *constant*. Since the entropy of photons is proportional to the number density, that means the entropy per comoving volume is also constant. Therefore, the total entropy in the Cosmic Microwave Background (or, equivalently, the total number of photons) is a convenient measure of spatial volume in the universe. Since the entropy per photon s is (up to a few constants) given by the cube of the temperature,

$$s \sim T^3, \quad (101)$$

the total photon entropy S in our current horizon volume is of order

$$S_{\text{hor}} \sim T_{\text{CMB}}^3 d_{\text{H}}^3 \sim \left(\frac{T_{\text{CMB}}}{H_0} \right)^3 \sim 10^{88}, \quad (102)$$

where we have taken the CMB temperature to be 2.7 K and the current Hubble parameter H_0 to be 70 km/s/MpC. (The interesting unit conversion from km/s/MpC to Kelvin is left as an exercise for the reader.)

Let us consider a highly over-simplified picture of the universe, in which no particle decays or annihilations occur between the end of inflation and today. In that case, the only time when the photon number (and therefore the entropy) in the universe changes is during the reheating process itself, when the inflation ϕ decays into radiation and

sets the initial state for the hot Big Bang. Therefore, we must *at minimum* create an entropy of 10^{88} during reheating. Let us say that the energy density during inflation is

$$\rho \sim V(\phi) \sim \Lambda^4, \quad (103)$$

where Λ is some energy scale. Therefore, the horizon size during inflation is then

$$d_H \sim H^{-1} \sim \frac{m_{\text{Pl}}}{\Lambda^2}, \quad (104)$$

so that the initial volume of the inflationary “patch” which undergoes exponential expansion is

$$V_i \sim d_H^3 \sim \frac{m_{\text{Pl}}^3}{\Lambda^6}. \quad (105)$$

Suppose inflation continues for N e-folds of expansion, so that the scale factor a increases by a factor of e^N during inflation. The *proper* volume of the initial inflationary patch increases by the cube of the scale factor

$$V_f \sim e^{3N} d_H^3 \sim e^{3N} \frac{m_{\text{Pl}}^3}{\Lambda^6}. \quad (106)$$

Inflation takes a tiny patch of the universe and blows it up exponentially large, but in such a way that the energy *density* remains approximately constant: we have created an exponential amount of energy out of nothing! During reheating, this huge store of energy in the coherently oscillating field ϕ decays into radiation and the temperature and entropy of the universe undergo an explosive increase. If reheating is highly efficient, then all or most of the energy stored in the inflaton field will be transformed into radiation, and the temperature of the universe after reheating will be of order the energy density of the inflaton field,

$$T_{\text{RH}} \sim \Lambda. \quad (107)$$

The entropy per comoving volume after reheating will then be $s_{\text{RH}} \sim T_{\text{RH}}^3 \sim \Lambda^3$, and the *total* entropy in our inflating patch will be

$$S_{\text{RH}} \sim V_f T_{\text{RH}}^3 \sim e^{3N} \frac{m_{\text{Pl}}^3}{\Lambda^3}. \quad (108)$$

Since this is our only source of entropy in our toy-model universe, this entropy must be at least as large as the entropy in our current horizon volume, $S_{\text{RH}} \geq 10^{88}$. The only adjustable parameter is the number of e-folds of inflation. Taking the logarithm of both sides gives a lower bound on N ,

$$N \geq 68 + \ln \left(\frac{\Lambda}{m_{\text{Pl}}} \right). \quad (109)$$

We will see later that the amplitude of primordial density fluctuations $\delta\rho/\rho \sim 10^{-5}$ typically constrains the inflationary energy scale to be of order $\Lambda \sim 10^{-4} m_{\text{Pl}}$, so that we have a lower limit on the number of e-folds of inflation of

$$N > N_{\text{min}} \sim 60. \quad (110)$$

Most inflation models hugely oversaturate this bound, with $N_{\text{tot}} \gg N_{\text{min}}$. There is in fact no *upper* bound on the number of e-folds of inflation, an idea which is central to Linde’s idea of “eternal” inflation [52, 53, 54, 55], in which inflation, once initiated, never completely ends, with reheating occurring only in isolated patches of the cosmos. Furthermore, it is easy to see that our oversimplified toy model of the universe gives a remarkably accurate estimate of N_{min} . In the real universe, all sorts of particle decays and annihilations happen between the end of inflation and today, which create additional entropy. However, our lower bound (109) is only logarithmically sensitive to these processes. The dominant uncertainty is in the reheat temperature: it is possible that the energy scale of inflation is very low, or that the reheating process is very inefficient, and there are very few *observational* bounds on these scales. We do know that the universe has to be radiation dominated and in equilibrium by the time primordial nucleosynthesis happens at temperatures of order MeV. Furthermore, the baryon asymmetry of the universe is at least a good hint that the Big Bang was hot to at least the scale of electroweak unification. A typical assumption is that the reheat temperature is something between 1 TeV and 10^{16} GeV, which translates into a range for N_{min} of order [56, 57]

$$N_{\text{min}} \simeq [46, 60]. \quad (111)$$

A. Example: the $\lambda\phi^4$ potential

We are now in a position to apply this to a specific case. We use the simple case of a quartic potential,

$$V(\phi) = \lambda\phi^4. \quad (112)$$

The slow roll equations (96, 94) imply that the field evolves as:

$$\dot{\phi} = -\frac{V'(\phi)}{3H} = -\sqrt{\frac{m_{\text{Pl}}^2}{24\pi}} \frac{V'(\phi)}{\sqrt{V(\phi)}} \propto \phi. \quad (113)$$

Note that this potential does not much qualitatively resemble the schematic in Fig. 13: the “flatness” of the potential arises because the energy density $V(\phi) \propto \phi^4$ rises much more quickly than the kinetic energy, $\dot{\phi}^2 \propto \phi^2$, so that if the field is far enough out on the potential, the slow roll approximation is self-consistent. The field rolls down to the potential toward the vacuum at the origin, and the equation of state is determined by the parameter ϵ ,

$$\epsilon(\phi) \simeq \frac{m_{\text{Pl}}^2}{16\pi} \left(\frac{V'(\phi)}{V(\phi)} \right)^2 = \frac{1}{\pi} \left(\frac{m_{\text{Pl}}}{\phi} \right)^2. \quad (114)$$

The field value ϕ_e at the end of inflation is when $\epsilon(\phi_e) = 1$, or

$$\phi_e = \frac{m_{\text{Pl}}}{\sqrt{\pi}}. \quad (115)$$

For $\phi > \phi_e$, $\epsilon < 1$ and the universe is inflating, and for $\phi < \phi_e$, $\epsilon > 1$ and the expansion enters a decelerating phase. Therefore, even this simple potential has the necessary characteristics to support a period of early-universe inflation followed by reheating and a hot Big Bang cosmology. What about the requirement that the universe inflate for at least 60 e-folds? Using Eq. (98), we can express the number of e-folds before the end of inflation (100) as

$$N = \frac{2\sqrt{\pi}}{m_{\text{Pl}}} \int_{\phi_e}^{\phi} \frac{dx}{\sqrt{\epsilon(x)}} = \pi \left(\frac{\phi}{m_{\text{Pl}}} \right)^2 - 1, \quad (116)$$

where we integrate *backward* from ϕ_e to ϕ to be consistent with the sign convention (89). Therefore the field value N e-folds before the end of inflation is

$$\phi_N = m_{\text{Pl}} \sqrt{\frac{N+1}{\pi}}, \quad (117)$$

so that

$$\phi_{60} = 4.4m_{\text{Pl}}. \quad (118)$$

We obtain sufficient inflation, but at a price: the field must be a long way (several times the Planck scale) out on the potential. However, we do *not* necessarily have to invoke quantum gravity, since for small enough coupling λ , the energy density in the field can much less than the Planck density, and the energy density is the physically important quantity.

In this section, we have seen that the basic picture of an early epoch in the universe dominated by vacuum-like energy, leading to nearly exponential expansion, can be realized within the context of a simple scalar field theory. The equation of state for the field approximates a cosmological constant $p = -\rho$ when the energy density is dominated by the field potential $V(\phi)$, and inflation ends when the potential becomes steep enough that the kinetic energy $\dot{\phi}^2/2$ dominates over the potential. To solve the horizon and flatness problems and create a universe consistent with observation, we must have *at least* 60 or so e-folds of inflation, although in principle inflation could continue for much longer than this minimum amount. This dynamical explanation for the flatness and homogeneity of the universe is an interesting, but hardly compelling scenario. It could be that the universe started out homogeneous and flat because of initial conditions, either through some symmetry we do not yet understand, or because there are many universes, and we just happen to find ourselves in a highly unlikely realization which is homogeneous and geometrically flat. In the absence of any other observational handles on the physics of the very early universe, it is impossible to tell. However, flatness and homogeneity are not the whole story: inflation provides an elegant mechanism for explaining the *inhomogeneity* of the universe as well, which we discuss in Section V.

V. PERTURBATIONS IN INFLATION

The universe we live in today is homogeneous, but only when averaged over very large scales. On small scales, the size of people or solar systems or galaxies or even clusters of galaxies, the universe we see is highly inhomogeneous. Our world is full of complex structure, created by gravitational instability acting on tiny “seed” perturbations in the early universe. If we look as far back in time as the epoch of recombination, the universe on all scales was homogeneous to a high degree of precision, a few parts in 10^5 . Recent observational efforts such as the WMAP satellite have made exquisitely precise measurements of the first tiny inhomogeneities in the universe, which later collapsed to form the structure we see today. (We discuss the WMAP observation in more detail in Section VI.) Therefore, another mystery of Big Bang cosmology is: what created the primordial perturbations? This mystery is compounded by the fact that the perturbations we observe in the CMB exhibit correlations on scales much larger than the horizon size at the time of recombination, which corresponds to an angular multipole of $\ell \simeq 100$, or about 1° as observed on the sky today. This is another version of the horizon problem: not only is the universe homogeneous on scales larger than the horizon, but whatever created the primordial perturbations must also have been capable of generating fluctuations on scales larger than the horizon. Inflation provides just such a mechanism [7, 8, 9, 10, 11, 12, 13].

Consider a perturbation in the cosmological fluid with wavelength λ . Since the proper wavelength redshifts with expansion, $\lambda_{\text{prop}} \propto a(t)$, the *comoving* wavelength of the perturbation is a constant, $\lambda_{\text{com}} = \text{const}$. This is true of photons or density perturbations or gravitational waves or any other wave propagating in the cosmological background. Now consider this wavelength relative to the size of the horizon: We have seen that in general the horizon as measured in comoving units is proportional to the conformal time, $d_H \propto \tau$. Therefore, for matter- or radiation-dominated expansion, the horizon size *grows* in comoving units, so that a comoving length which is larger than the horizon at early times is smaller than the horizon at late times: modes “fall into” the horizon. The opposite is true during inflation, where the conformal time is negative and evolving toward zero: the comoving horizon size is still proportional to τ , but it now *shrinks* with cosmological expansion, and comoving perturbations which are initially smaller than the horizon are “redshifted” to scales larger than the horizon at late times (Fig. 15).

If the universe is inflating at early times, and radiation- or matter-dominated at late times, perturbations in the density of the universe which are initially smaller than the horizon are redshifted during inflation to superhorizon scales. Later, as the horizon begins to grow in comoving coordinates, the perturbations fall back into the horizon, where they act as a source for structure formation. In this way inflation explains the observed properties of perturbations in the universe, which exist at both super- and sub-horizon scales at the time of recombination. Furthermore, an important consequence of this process is that the last perturbations to exit the horizon are the *first* to fall back in. Therefore, the shortest wavelength perturbations are the ones which exited the horizon just at the end of inflation, $N = 0$, and longer wavelength perturbations exited the horizon earlier. Perturbations about the same size as our horizon today exited the horizon during inflation at around $N = 60$. Perturbations which exited the horizon earlier than that, $N > 60$, are still larger than our horizon today. Therefore, it is only possible to place observational constraints on the *end* of inflation, about the last 60 e-folds. Everything that happened before that, including anything that might tell us about the initial conditions which led to inflation, is most probably inaccessible to us.

This kinematic picture, however, does not itself explain the physical origin of the perturbations. Inflation driven by a scalar field provides a natural explanation for this as well. The inflaton field ϕ evolving on the potential $V(\phi)$ will not evolve completely classically, but will also be subject to small quantum fluctuations about its classical trajectory, which will in general be *inhomogeneous*. Since the energy density of the universe during inflation is dominated by the inflaton field, quantum fluctuations in ϕ couple to the spacetime curvature and result in fluctuations in the density of the universe. Therefore, in the same way that the classical behavior of the field ϕ provides a description of the background evolution of the universe, the quantum behavior of ϕ provides a description of the inhomogeneous perturbations about that background. We defer a full treatment of inflaton perturbations to Appendix A, and in the next section focus on the much simpler case of quantizing a *decoupled* scalar φ in an inflationary spacetime. In addition to its relative simplicity, this case has direct relevance to the generation of gravitational waves in inflation.

A. The Klein-Gordon Equation in Curved Spacetime

Consider an arbitrary free scalar field, which we denote φ to distinguish it from the inflaton field ϕ . The Lagrangian for the field is

$$\mathcal{L} = \frac{1}{2} g^{\mu\nu} \partial_\mu \varphi \partial_\nu \varphi, \quad (119)$$

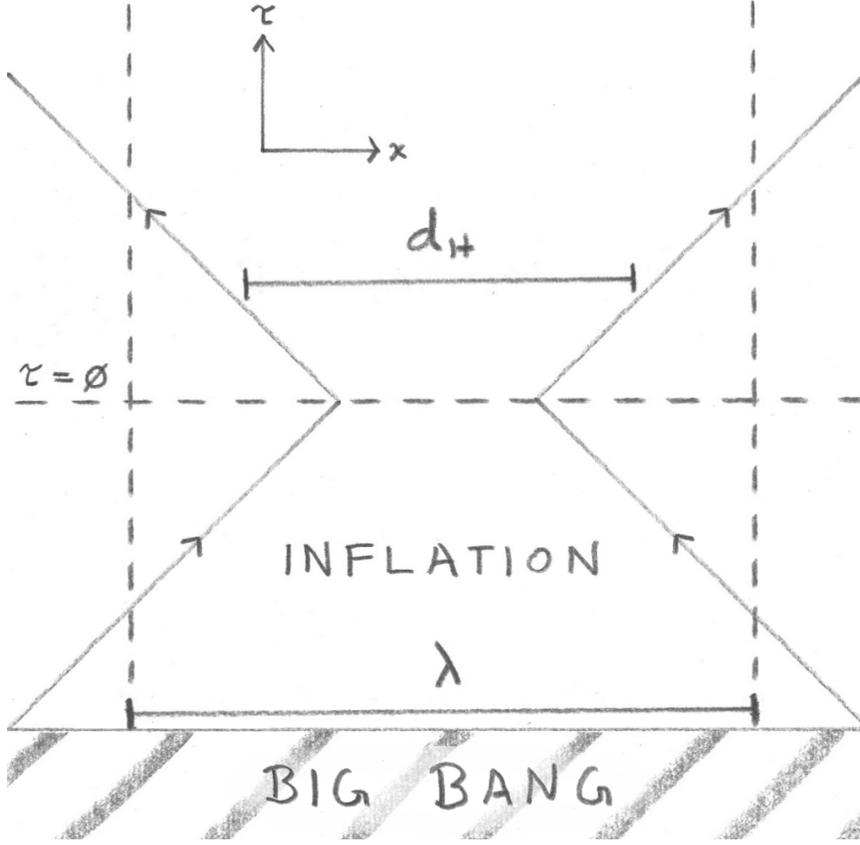


FIG. 15: A conformal diagram of the horizon in an inflationary universe. The comoving horizon shrinks during inflation, and grows during the radiation- and matter-dominated expansion, while the comoving wavelengths of perturbations remain constant. This drives comoving perturbations to “superhorizon” scales.

and varying the action (119) gives the Euler-Lagrange equation of motion

$$\frac{1}{\sqrt{-g}} \partial_\nu (g^{\mu\nu} \sqrt{-g} \partial_\mu \varphi) = 0. \quad (120)$$

It will prove convenient to express the background FRW metric in conformal coordinates

$$g_{\mu\nu} = a^2(\tau) \eta_{\mu\nu} \quad (121)$$

instead of the coordinate-time metric (83) we used in Section (IV). Here τ is the conformal time and $\eta_{\mu\nu} = \text{diag.}(1, -1, -1, -1)$ is the Minkowski metric. In conformal coordinates, the free scalar equation of motion (120) is

$$\varphi'' + 2 \left(\frac{a'}{a} \right) \varphi - \nabla^2 \varphi = 0, \quad (122)$$

where $' = d/d\tau$ is a derivative with respect to *conformal* time. Note that unlike the case of the inflaton ϕ , we are solving for perturbations and therefore retain the gradient term $\nabla^2 \varphi$. The field φ is a decoupled spectator field evolving in a *fixed* cosmological background, and does not effect the time evolution of the scale factor $a(\tau)$. An example of such a field is gravitational waves. If we express the spacetime metric as an FRW background $g_{\mu\nu}^{\text{FRW}}$ plus perturbation $\delta g_{\mu\nu}$, we can express the tensorial portion of the perturbation in general as a sum of two scalar degrees of freedom

$$\begin{aligned} \delta g_{0i} &= \delta g_{i0} = 0 \\ \delta g_{ij} &= \frac{32\pi}{m_{\text{Pl}}^2} (\varphi_+ \hat{e}_{ij}^+ + \varphi_\times \hat{e}_{ij}^\times), \end{aligned} \quad (123)$$

where $i, j = 1, 2, 3$, and $\hat{e}_{ij}^{+, \times}$ are longitudinal and transverse polarization tensors, respectively. It is left as an exercise for the reader to show that the scalars $\varphi_{+, \times}$ behave to linear order as free scalars, with equation of motion (122).

To solve the equation of motion (122), we first Fourier expand the field into momentum states φ_k ,

$$\varphi(\tau, \mathbf{x}) = \int \frac{d^3k}{(2\pi)^{3/2}} [\varphi_{\mathbf{k}}(\tau) b_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}} + \varphi_{\mathbf{k}}^*(\tau) b_{\mathbf{k}}^* e^{-i\mathbf{k}\cdot\mathbf{x}}]. \quad (124)$$

Note that the coordinates \mathbf{x} are comoving coordinates, and the wavevector \mathbf{k} is a comoving wavevector, which does not redshift with expansion. The proper wavevector is

$$\mathbf{k}_{\text{prop}} = \mathbf{k}/a(\tau). \quad (125)$$

Therefore, the comoving wavenumber \mathbf{k} is not itself dynamical, but is just a set of constants labeling a particular Fourier component. The equation of motion for a single mode $\varphi_{\mathbf{k}}$ is

$$\varphi_{\mathbf{k}}'' + 2\left(\frac{a'}{a}\right)\varphi_{\mathbf{k}} + k^2\varphi_{\mathbf{k}} = 0. \quad (126)$$

It is convenient to introduce a field redefinition

$$u_k \equiv a(\tau)\varphi_{\mathbf{k}}(\tau), \quad (127)$$

and the mode function u_k obeys a generalization of the Klein-Gordon equation to an expanding spacetime,

$$u_k'' + \left[k^2 - \frac{a''}{a}\right]u_k = 0. \quad (128)$$

(We have dropped the vector notation \mathbf{k} on the subscript, since the Klein-Gordon equation depends only on the magnitude of k .)

Any mode with a fixed comoving wavenumber k redshifts with time, so that early time corresponds to short wavelength (ultraviolet) and late time corresponds to long wavelength (infrared). The solutions to the mode equation show qualitatively different behaviors in the ultraviolet and infrared limits:

- *Short wavelength limit*, $k \gg a''/a$. In this case, the equation of motion is that for a conformally Minkowski Klein-Gordon field,

$$u_k'' + k^2 u_k = 0, \quad (129)$$

with solution

$$u_k(\tau) = \frac{1}{\sqrt{2k}} (A_k e^{-ik\tau} + B_k e^{ik\tau}). \quad (130)$$

Note that this is in terms of *conformal* time and *comoving* wavenumber, and can only be identified with an exactly Minkowski spacetime in the ultraviolet limit.

- *Long wavelength limit*, $k \ll a''/a$. In the infrared limit, the mode equation becomes

$$a'' u_k = a u_k'', \quad (131)$$

with the trivial solution

$$u_k \propto a \Rightarrow \varphi_k = \text{const.} \quad (132)$$

This illustrates the phenomenon of *mode freezing*: field modes φ_k with wavelength longer than the horizon size cease to be dynamical, and asymptote to a constant, *nonzero* amplitude.⁵ This is a quantitative expression of our earlier qualitative notion of particle creation at the cosmological horizon. The amplitude of the field at long wavelength is determined by the boundary condition on the mode, *i.e.* the integration constants A_k and B_k .

Therefore, all of the physics boils down to the question of how we set the boundary condition on field perturbations in the ultraviolet limit. This is fortunate, since in that limit the field theory describing the modes becomes approximately Minkowskian, and we know how to quantize fields in Minkowski Space. Once the integration constants are fixed, the behavior of the mode function u_k is completely determined, and the long-wavelength amplitude of the perturbation can then be calculated without ambiguity. We next discuss quantization.

⁵ The second solution to this equation is a decaying mode, which is always subdominant in the infrared limit.

B. Quantization

We have seen that the equation of motion for field perturbations approaches the usual Minkowski Space Klein-Gordon equation in the ultraviolet limit, which corresponds to the limit of early time for a mode redshifting with expansion. We determine the boundary conditions for the mode function via canonical quantization. To quantize the field φ_k , we promote the Fourier coefficients in the classical mode expansion (124) to annihilation and creation operators

$$b_{\mathbf{k}} \rightarrow \hat{b}_{\mathbf{k}}, \quad b_{\mathbf{k}}^* \rightarrow \hat{b}_{\mathbf{k}}^\dagger, \quad (133)$$

with commutation relation

$$[\hat{b}_{\mathbf{k}}, \hat{b}_{\mathbf{k}'}^\dagger] \equiv \delta^3(\mathbf{k} - \mathbf{k}'). \quad (134)$$

Note that the commutator in an FRW background is given in terms of *comoving* wavenumber, and holds whether we are in the short wavelength limit or not. In the short wavelength limit, this becomes equivalent to a Minkowski Space commutator. The quantum field φ is then given by the usual expansion in operators $\hat{b}_{\mathbf{k}}, \hat{b}_{\mathbf{k}}^\dagger$

$$\varphi(\tau, \mathbf{x}) = \int \frac{d^3k}{(2\pi)^{3/2}} [\varphi_{\mathbf{k}}(\tau) b_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}} + \text{H.C.}] \quad (135)$$

The corresponding canonical momentum is

$$\Pi(\tau, \mathbf{x}) \equiv \frac{\delta\mathcal{L}}{\delta(\partial_0\varphi)} = a^2(\tau) \frac{\partial\varphi}{\partial\tau}. \quad (136)$$

It is left as an exercise for the reader to show that the canonical commutation relation

$$[\varphi(\tau, \mathbf{x}), \Pi(\tau, \mathbf{x}')] = i\delta^3(\mathbf{x} - \mathbf{x}') \quad (137)$$

corresponds to a Wronskian condition on the mode u_k ,

$$u_k \frac{\partial u_k^*}{\partial\tau} - u_k^* \frac{\partial u_k}{\partial\tau} = i, \quad (138)$$

which for the ultraviolet mode function (130) results in a condition on the integration constants

$$|A_k|^2 - |B_k|^2 = 1. \quad (139)$$

This quantization condition corresponds to one of the two boundary conditions which are necessary to completely determine the solution. The second boundary condition comes from vacuum selection, *i.e.* our definition of which state corresponds to a zero-particle state for the system. In the next section, we discuss the issue of vacuum selection in detail.

C. Vacuum Selection

Consider a quantum field in Minkowski Space. The state space for a quantum field theory is a set of states $|n(\mathbf{k}_1), \dots, n(\mathbf{k}_i)\rangle$ representing the number of particles with momenta $\mathbf{k}_1, \dots, \mathbf{k}_i$. The creation and annihilation operators $\hat{a}_{\mathbf{k}}^\dagger$ and $\hat{a}_{\mathbf{k}}$ act on these states by adding or subtracting a particle from the state:

$$\begin{aligned} \hat{a}_{\mathbf{k}}^\dagger |n(\mathbf{k})\rangle &= \sqrt{n+1} |n(\mathbf{k}) + 1\rangle \\ \hat{a}_{\mathbf{k}} |n(\mathbf{k})\rangle &= \sqrt{n} |n(\mathbf{k}) - 1\rangle. \end{aligned} \quad (140)$$

The ground state, or vacuum state of the space, is just the zero particle state:

$$\hat{a}_{\mathbf{k}} |0\rangle = 0. \quad (141)$$

Note in particular that the vacuum state $|0\rangle$ is *not* equivalent to zero. The vacuum is not nothing:

$$|0\rangle \neq 0. \quad (142)$$

To construct a quantum field, we look at the familiar classical wave equation for a scalar field,

$$\frac{\partial^2 \phi}{\partial t^2} - \nabla^2 \phi = 0. \quad (143)$$

To solve this equation, we decompose into Fourier modes $u_{\mathbf{k}}$,

$$\phi = \int d^3k [a_{\mathbf{k}} u_{\mathbf{k}}(t) e^{i\mathbf{k}\cdot\mathbf{x}} + a_{\mathbf{k}}^* u_{\mathbf{k}}^*(t) e^{-i\mathbf{k}\cdot\mathbf{x}}], \quad (144)$$

where the mode functions $u_{\mathbf{k}}(t)$ satisfy the ordinary differential equation

$$\ddot{u}_{\mathbf{k}} + k^2 u_{\mathbf{k}} = 0. \quad (145)$$

This is a classical wave equation with a classical solution, and the Fourier coefficients $a_{\mathbf{k}}$ are just complex numbers. The solution for the mode function is

$$u_{\mathbf{k}} \propto e^{-i\omega_k t}, \quad (146)$$

where ω_k satisfies the dispersion relation

$$\omega_k^2 - \mathbf{k}^2 = 0. \quad (147)$$

To turn this into a quantum field, we identify the Fourier coefficients with creation and annihilation operators

$$a_{\mathbf{k}} \rightarrow \hat{a}_{\mathbf{k}}, \quad a_{\mathbf{k}}^* \rightarrow \hat{a}_{\mathbf{k}}^\dagger, \quad (148)$$

and enforce the commutation relations

$$[\hat{a}_{\mathbf{k}}, \hat{a}_{\mathbf{k}'}^\dagger] = \delta^3(\mathbf{k} - \mathbf{k}'). \quad (149)$$

This is the standard quantization of a scalar field in Minkowski Space, which should be familiar. But what probably is not familiar is that this solution has an interesting symmetry. Suppose we define a new mode function $u_{\mathbf{k}}$ which is a rotation of the solution (146):

$$u_{\mathbf{k}} = A(k) e^{-i\omega t + i\mathbf{k}\cdot\mathbf{x}} + B(k) e^{i\omega t - i\mathbf{k}\cdot\mathbf{x}}. \quad (150)$$

This is *also* a perfectly valid solution to the original wave equation (143), since it is just a superposition of the Fourier modes. But we can then re-write the quantum field in terms of our original Fourier modes and new *operators* $\hat{b}_{\mathbf{k}}$ and $\hat{b}_{\mathbf{k}}^\dagger$ and the original Fourier modes $e^{i\mathbf{k}\cdot\mathbf{x}}$ as:

$$\phi = \int d^3k [\hat{b}_{\mathbf{k}} e^{-i\omega t + i\mathbf{k}\cdot\mathbf{x}} + \hat{b}_{\mathbf{k}}^\dagger e^{i\omega t - i\mathbf{k}\cdot\mathbf{x}}], \quad (151)$$

where the new operators $\hat{b}_{\mathbf{k}}$ are given in terms of the old operators $\hat{a}_{\mathbf{k}}$ by

$$\hat{b}_{\mathbf{k}} = A(k) \hat{a}_{\mathbf{k}} + B^*(k) \hat{a}_{\mathbf{k}}^\dagger. \quad (152)$$

This is completely equivalent to our original solution (144) as long as the new operators satisfy the same commutation relation as the original operators,

$$[\hat{b}_{\mathbf{k}}, \hat{b}_{\mathbf{k}'}^\dagger] = \delta^3(\mathbf{k} - \mathbf{k}'). \quad (153)$$

This can be shown to place a condition on the coefficients A and B ,

$$|A|^2 - |B|^2 = 1. \quad (154)$$

Otherwise, we are free to choose A and B as we please.

This is just a standard property of linear differential equations: any linear combination of solutions is itself a solution. But what does it mean physically? In one case, we have an annihilation operator $\hat{a}_{\mathbf{k}}$ which gives zero when acting on a particular state which we call the vacuum state:

$$\hat{a}_{\mathbf{k}} |0_a\rangle = 0. \quad (155)$$

Similarly, our rotated operator $\hat{b}_{\mathbf{k}}$ gives zero when acting on some state

$$\hat{b}_{\mathbf{k}} |0_b\rangle = 0. \quad (156)$$

The point is that the two “vacuum” states are not the same

$$|0_a\rangle \neq |0_b\rangle. \quad (157)$$

From this point of view, we can define any state we wish to be the “vacuum” and build a completely consistent quantum field theory based on this assumption. From another equally valid point of view this state will contain particles. How do we tell which is the *physical* vacuum state? To define the real vacuum, we have to consider the spacetime the field is living in. For example, in regular special relativistic quantum field theory, the “true” vacuum is the zero-particle state as seen by an inertial observer. Another more formal way to state this is that we require the vacuum to be Lorentz symmetric. This fixes our choice of vacuum $|0\rangle$ and defines unambiguously our set of creation and annihilation operators \hat{a} and \hat{a}^\dagger . A consequence of this is that an *accelerated* observer in the Minkowski vacuum will think that the space is full of particles, a phenomenon known as the Unruh effect [58]. The zero-particle state for an accelerated observer is different than for an inertial observer. The case of an FRW spacetime is exactly analogous, except that the FRW equivalent of an inertial observer is an observer at rest in comoving coordinates. Since an FRW spacetime is asymptotically Minkowski in the ultraviolet limit, we choose the vacuum field which corresponds to the usual Minkowski vacuum in that limit,

$$u_k(\tau) \propto e^{-ik\tau} \Rightarrow A_k = 1, B_k = 0. \quad (158)$$

This is known as the *Bunch-Davies* vacuum. This is not the only possible choice, although it is widely believed to be the most natural. The issue of vacuum ambiguity of inflationary perturbations is a subject which is extensively discussed in the literature, and is still the subject of controversy. It is known that the choice of vacuum is potentially sensitive to quantum-gravitational physics [59, 60, 61], a subject which is referred to as *Trans-Planckian* physics [18, 62, 63]. For the remainder of our discussion, we will assume a Bunch-Davies vacuum.

The key point is that quantization and vacuum selection together *completely* specify the mode function, up to an overall phase. This means that the amplitude of the mode once it has redshifted to long wavelength and frozen out is similarly determined. In the next section, we solve the mode equation at long wavelength for an inflationary background.

D. Exact Solutions and the Primordial Power Spectrum

The exact form of the solution to Eq. (128) depends on the evolution of the background spacetime, as encoded in $a(\tau)$, which in turn depends on the equation of state of the field driving inflation. We will consider the case where the equation of state is constant, which will *not* be the case in general for scalar field-driven inflation, but will nonetheless turn out to be a good approximation in the limit of a slowly rolling field. Generalizing Eq. (77) to the case of arbitrary equation of state parameter $\epsilon = \text{const.}$, the conformal time can be written

$$\tau = - \left(\frac{1}{aH} \right) \left(\frac{1}{1-\epsilon} \right), \quad (159)$$

and the Friedmann and Raychaudhuri Equations (14) give

$$\frac{a''}{a} = a^2 H^2 (2 - \epsilon), \quad (160)$$

where a prime denotes a derivative with respect to conformal time. The conformal time, as in the case of de Sitter space, is negative and tending toward zero during inflation. (Proof of these relations is left as an exercise for the reader.) We can then write the mode equation (128) as

$$u_k'' + [k^2 - a^2 H^2 (2 - \epsilon)] u_k = 0. \quad (161)$$

Using Eq. (159) to write aH in terms of the conformal time τ , the equation of motion becomes

$$\tau^2 (1 - \epsilon)^2 u_k'' + [(k\tau)^2 (1 - \epsilon)^2 - (2 - \epsilon)] u_k = 0. \quad (162)$$

This is a Bessel equation, with solution

$$u_k \propto \sqrt{-k\tau} [J_\nu(-k\tau) \pm iY_\nu(-k\tau)], \quad (163)$$

where the index ν is given by:

$$\nu = \frac{3 - \epsilon}{2(1 - \epsilon)}. \quad (164)$$

The quantity $-k\tau$ has special physical significance, since from Eq. (159) we can write

$$(-k\tau)(1 - \epsilon) = \frac{k}{aH}, \quad (165)$$

where the quantity (k/aH) expresses the wavenumber k in units of the comoving horizon size $d_H \sim (aH)^{-1}$. Therefore, the short wavelength limit is $-k\tau \rightarrow -\infty$, or $(k/aH) \gg 1$. The long-wavelength limit is $-k\tau \rightarrow 0$, or $(k/aH) \ll 1$.

The simple case of de Sitter space ($p = -\rho$) corresponds to the limit $\epsilon = 0$, so that the Bessel index is $\nu = 3/2$ and the mode function (163) simplifies to

$$u_k \propto \left(\frac{k\tau - i}{k\tau} \right) e^{\pm ik\tau}. \quad (166)$$

In the short wavelength limit, $(-k\tau) \rightarrow -\infty$, the mode function is given, as expected, by

$$u_k \propto e^{\pm ik\tau}. \quad (167)$$

Selecting the Bunch-Davies vacuum gives $u_k \propto e^{ik\tau}$, and canonical quantization fixes the normalization,

$$u_k = \frac{1}{\sqrt{2k}} e^{-ik\tau}. \quad (168)$$

Therefore, the fully normalized exact solution is

$$u_k = \frac{1}{\sqrt{2k}} \left(\frac{k\tau - i}{k\tau} \right) e^{-ik\tau}. \quad (169)$$

This solution has no free parameters aside from an overall phase, and is valid at *all* wavelengths, including after the mode has been redshifted outside of the horizon and becomes non-dynamical, or “frozen”. In the long wavelength limit, $-k\tau \rightarrow 0$, the mode function (169) becomes

$$u_k \rightarrow \frac{1}{\sqrt{2k}} \left(\frac{i}{(-k\tau)} \right) = \frac{i}{2k} \left(\frac{aH}{k} \right) \propto a, \quad (170)$$

consistent with the qualitative result (132). Therefore the field amplitude φ_k is given by

$$|\varphi_k| = \left| \frac{u_k}{a} \right| \rightarrow \frac{H}{\sqrt{2}k^{3/2}} = \text{const}. \quad (171)$$

The quantum mode therefore displays the freezeout behavior we noted qualitatively above (Fig. 16). The amplitude of quantum fluctuations is conventionally expressed in terms of the two-point correlation function of the field φ . It is left as an exercise for the reader to show that the vacuum two-point correlation function is given by

$$\begin{aligned} \langle 0 | \varphi(\tau, \mathbf{x}) \varphi(\tau, \mathbf{x}') | 0 \rangle &= \int \frac{d^3k}{(2\pi)^3} \left| \frac{u_k}{a} \right|^2 e^{i\mathbf{k} \cdot (\mathbf{x} - \mathbf{x}')} \\ &= \int \frac{dk}{k} P(k) e^{i\mathbf{k} \cdot (\mathbf{x} - \mathbf{x}')}, \end{aligned} \quad (172)$$

where the *power spectrum* $P(k)$ is defined as

$$P(k) \equiv \left(\frac{k^3}{2\pi^2} \right) \left| \frac{u_k}{a} \right|^2 \longrightarrow \left(\frac{H}{2\pi} \right)^2, \quad -k\tau \rightarrow 0. \quad (173)$$

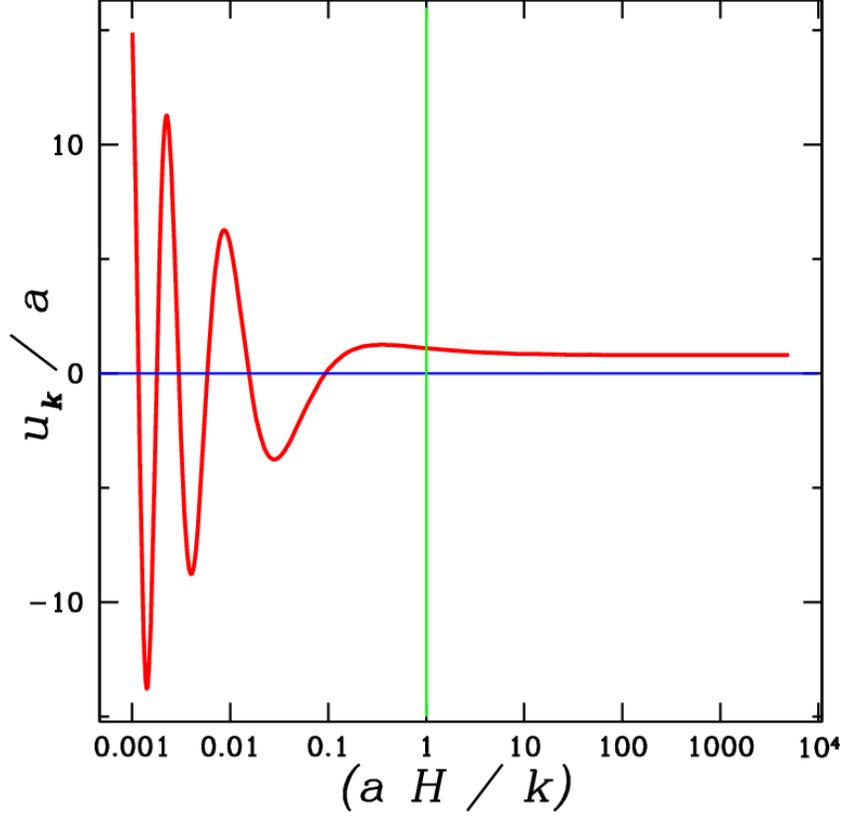


FIG. 16: The normalized mode function in de Sitter space, showing oscillatory behavior on subhorizon scales $k/aH > 1$, and mode freezing on superhorizon scales, $k/aH < 1$.

The power per logarithmic interval k in the field fluctuation is then given in the long wavelength limit by the Hubble parameter $H = \text{const}$. This property of scale invariance is exact in the de Sitter limit.

In a more general model, the spacetime is only *approximately* de Sitter, and we expect that the power spectrum of field fluctuations will only be approximately scale invariant. It is convenient to express this dynamics in terms of the equation of state parameter ϵ ,

$$\epsilon = \frac{1}{H} \frac{dH}{dN}. \quad (174)$$

We must have $\epsilon < 1$ for inflation, and for a slowly rolling field $|\eta| \ll 1$ means that ϵ will also be slowly varying, $\epsilon \simeq \text{const}$. It is straightforward to show that for $\epsilon = \text{const} \neq 0$ that:

- The Bunch-Davies vacuum corresponds to the positive mode of Eq. (163),

$$u_k \propto \sqrt{-k\tau} [J_\nu(-k\tau) + iY_\nu(-k\tau)]. \quad (175)$$

- Quantization fixes the normalization as

$$u_k = \frac{1}{2} \sqrt{\frac{\pi}{k}} \sqrt{-k\tau} [J_\nu(-k\tau) + iY_\nu(-k\tau)]. \quad (176)$$

- The power spectrum in the long-wavelength limit $k/aH \rightarrow 0$ is a power law in k :

$$[P(k)]^{1/2} \longrightarrow 2^{\nu-3/2} \frac{\Gamma(\nu)}{\Gamma(3/2)} (1-\epsilon) \left(\frac{H}{2\pi}\right) \left(\frac{k}{aH(1-\epsilon)}\right)^{3/2-\nu}, \quad (177)$$

where $\Gamma(\nu)$ is a gamma function, and

$$\nu = \frac{3 - \epsilon}{2(1 - \epsilon)}. \quad (178)$$

Proof is left as an exercise for the reader.⁶ Note that in the case $\epsilon = \text{const.}$, both the background and perturbation equations are *exactly* solvable.

We can use these solutions as approximate solutions in the more general slow roll case, where $\epsilon \ll 1 \simeq \text{const.}$, so that the dependence of the power spectrum on k is approximately a power-law,

$$P(k) \propto k^n, \quad (179)$$

with spectral index

$$n = 3 - 2\nu = 3 - \frac{3 - \epsilon}{1 - \epsilon} \simeq -2\epsilon. \quad (180)$$

Equation (177) is curious, however, because it does not obviously exhibit complete mode freezing at long wavelength, since a and H both depend on time. We can show that $P(k)$ does in fact approach a time-dependent value at long wavelength by evaluating

$$\begin{aligned} \frac{d}{dN} \left[H \left(\frac{k}{aH} \right)^{3/2-\nu} \right] &= \frac{d}{dN} \left[H \left(\frac{k}{aH} \right)^{-\epsilon/(1-\epsilon)} \right] \\ &= H\epsilon \left(\frac{k}{aH} \right)^{-\epsilon/(1-\epsilon)} - \frac{\epsilon}{1-\epsilon} \left(\frac{k}{aH} \right)^{-\epsilon/(1-\epsilon)-1} \left(\frac{k}{aH} - \frac{\epsilon k}{aH} \right) \\ &= 0, \end{aligned} \quad (181)$$

which can be easily shown using $a \propto e^{-N}$ and $H \propto e^N$. That is, the time-dependent quantities a and H in Eq. (177) are combined in such a way as to form an *exactly* conserved quantity. Since it is conserved, we are free to evaluate it at any time (or value of aH) that we wish. It is conventional to evaluate the power spectrum at *horizon crossing*, or at $aH = k$, so that

$$P^{1/2}(k) \simeq \left(\frac{H}{2\pi} \right)_{k=aH}, \quad (182)$$

where we have approximated the ν -dependent multiplicative factor as order unity.⁷

It is straightforward to calculate the spectral index (180) directly from the horizon crossing expression (182) by using

$$a \propto e^{-N}, \quad H \propto e^N, \quad (183)$$

so that we can write derivatives in k at horizon crossing as derivatives in the number of e-folds N ,

$$d \ln k|_{k=aH} = d \ln(aH) = \frac{1}{aH} \frac{d(aH)}{dN} dN = (\epsilon - 1) dN. \quad (184)$$

The spectral index is then, to lowest order in slow roll

$$\begin{aligned} n &= \frac{d \ln P(k)}{d \ln k} = \frac{k}{H^2} \frac{dH^2}{dk} \Big|_{k=aH} \\ &= \frac{1}{H^2(\epsilon - 1)} \frac{dH^2}{dN} \end{aligned}$$

⁶ Note that the quantization condition (137) can be applied to the solution (163) exactly, resulting in the normalization condition (139), without approximating the solution in the short-wavelength limit!

⁷ This is *not* the value of the scalar field power spectrum at the moment the mode is physically crossing outside the horizon, as is often stated in the literature: it is the value of the power spectrum in the asymptotic long-wavelength limit. It is easy to show from the exact solution (176) that the mode function is still evolving with time as it crosses the horizon at $k = aH$, and the asymptotic amplitude differs from the amplitude at horizon crossing by about a factor of two. See Ref. [64] for a more detailed discussion of this point.

$$\begin{aligned}
&= \frac{2\epsilon}{(\epsilon - 1)} \\
&\simeq -2\epsilon,
\end{aligned} \tag{185}$$

in agreement with (180). Note that we are rather freely changing variables from the wavenumber k to the comoving horizon size $(aH)^{-1}$ to the number of e-folds N . As long as the cosmological evolution is monotonic, these are all different ways of measuring time: the time when a mode with wavenumber k exits the horizon, the time at which the horizon is a particular size, the number of e-folds N and the field value ϕ are all effectively just different choices of a clock, and we can switch from one to another as is convenient. For example, in the slow roll approximation, the Hubble parameter H is just a function of ϕ , $H \propto \sqrt{V(\phi)}$. Because of this, it is convenient to define $N(k)$ to be the number of e-folds (100) when a mode with wavenumber k crosses outside the horizon, and $\phi_N(k)$ to be the field value $N(k)$ e-folds before the end of inflation. Then the power spectrum can be written equivalently as *either* a function of k or of ϕ :

$$P^{1/2}(k) = \left(\frac{H}{2\pi}\right)_{k=aH} = \left(\frac{H}{2\pi}\right)_{\phi=\phi_N(k)} \simeq \sqrt{\frac{2V(\phi_N)}{3\pi m_{\text{Pl}}^2}}. \tag{186}$$

Wavenumbers k are conventionally normalized in units of $h\text{Mpc}^{-1}$ as measured in the *current* universe. We can relate N to scales in the current universe by recalling that modes which are of order the horizon size in the universe today, $k \sim a_0 H_0$, exited the horizon during inflation when $N = [46, 60]$, so that we can calculate the amplitude of perturbations at the scale of the CMB quadrupole today by evaluating the power spectrum for field values between ϕ_{46} and ϕ_{60} .

One example of a free scalar in inflation is gravitational wave modes, where the transverse and longitudinal polarization states of the gravity waves evolve as independent scalar fields. Using Eq. (123), we can then calculate the power spectrum in gravity waves (or *tensors*) as the sum of the two-point correlation functions for the separate polarizations:

$$P_T = \langle \delta g_{ij}^2 \rangle = 2 \times \frac{32}{m_{\text{Pl}}^2} \langle \varphi^2 \rangle = \frac{16H^2}{\pi m_{\text{Pl}}^2} \propto k^{n_T}, \tag{187}$$

with spectral index

$$n_T = -2\epsilon. \tag{188}$$

If the amplitude is large enough, such a spectrum of primordial gravity waves will be observable in the cosmic microwave background anisotropy and polarization, or be directly detectable by proposed experiments such as Big Bang Observer [65, 66].

The second type of perturbation generated during inflation is perturbations in the density of the universe, which are the dominant component of the CMB anisotropy $\delta T/T \sim \delta\rho/\rho \sim 10^{-5}$, and are responsible for structure formation. Density, or *scalar* perturbations are more complicated than tensor perturbations because they are generated by quantum fluctuations in the inflaton field itself: since the background energy density is dominated by the inflaton, fluctuations of the inflaton up or down the potential generate perturbations in the density. The full calculation requires self-consistent General Relativistic perturbation theory, and is presented in Appendix A. Here we simply state the result: Perturbations in the inflaton field $\delta\phi \simeq H/2\pi$ generate density perturbations with power spectrum

$$P_{\mathcal{R}}(k) = \left(\frac{\delta N}{\delta\phi}\delta\phi\right)^2 = \frac{H^2}{\pi m_{\text{Pl}}^2 \epsilon} \Big|_{k=aH} \propto k^{n_S-1}, \tag{189}$$

where N is the number of e-folds. Scalar perturbations are therefore enhanced relative to tensor perturbations by a factor of $1/\epsilon$. The scalar power spectrum is also an approximate power-law, with spectral index

$$n_S - 1 = \frac{\epsilon}{H^2(\epsilon - 1)} \frac{d}{dN} \left(\frac{H^2}{\epsilon}\right) \simeq -4\epsilon + 2\eta, \tag{190}$$

where η is the second slow roll parameter (97). Therefore, for any particular choice of inflationary potential, we have four measurable quantities: the amplitudes P_T and $P_{\mathcal{R}}$ of the tensor and scalar power spectra, and their spectral indices n_T and n_S . However, not all of these parameters are independent. In particular, the ratio r between the scalar and tensor amplitudes is given by the parameter ϵ , as is the tensor spectral index n_T :

$$r \equiv \frac{P_T}{P_S} = 16\epsilon = -8n_T. \tag{191}$$

This relation is known as the *consistency condition* for single-field slow roll inflation, and is in principle testable by a sufficiently accurate measurement of the primordial perturbation spectra.

In the next section, we apply these results to our example $\lambda\phi^4$ potential and calculate the inflationary power spectra.

E. Example: $\lambda\phi^4$

For the case of our example model with $V(\phi) = \lambda\phi^4$, it is now straightforward to calculate the scalar and tensor perturbation spectra. We express the normalization of the power spectra as a function of the number of e-folds N by

$$\begin{aligned} P_{\mathcal{R}}^{1/2} &= \frac{H}{m_{\text{Pl}}\sqrt{\pi\epsilon}} \Big|_{\phi=\phi_N} \\ &= \frac{4\sqrt{24\pi} [V(\phi_N)]^{3/2}}{3m_{\text{Pl}}^3 V'(\phi_N)} \\ &= \frac{24\pi}{3} \left(\frac{N+1}{\pi} \right) \lambda^{1/2} \sim 10^{-5}, \end{aligned} \quad (192)$$

where we have used the slow roll expressions for H (94) and ϵ (98) and Eq. (117) for ϕ_N . For perturbations about the current size of our horizon, $N = 60$, and CMB normalization forces the self-coupling to be very small,

$$\lambda \sim 10^{-15}. \quad (193)$$

The presence of an extremely small parameter is not peculiar to the $\lambda\phi^4$ model, but is generic, and is referred to as the *fine tuning* problem for inflation.

We can similarly calculate the tensor amplitude

$$P_T^{1/2} = \frac{4H}{m_{\text{Pl}}\sqrt{\pi}}, \quad (194)$$

which is usually expressed in terms of the tensor/scalar ratio

$$\begin{aligned} r &= 16\epsilon(\phi_N) = \frac{m_{\text{Pl}}}{\pi} \left(\frac{V'(\phi_N)}{V(\phi_N)} \right)^2 \\ &= \frac{16}{\pi} \left(\frac{m_{\text{Pl}}}{\phi_N} \right)^2 = \frac{16}{N+1} \simeq 0.26, \end{aligned} \quad (195)$$

where we have again taken $N = 60$. For this particular model, the power in gravitational waves is large, about a quarter of the power in scalar perturbations. This is *not* generic, but is quite model-dependent. Some choices of potential predict large tensor contributions (where “large” means of order 10%), and other choices of potential predict very tiny tensor contributions, well below 1%.

The tensor spectral index n_T is fixed by the consistency condition (191), but the scalar spectral index n_S is an independent parameter because of its dependence on η :

$$n = 1 - 4\epsilon(\phi_N) + 2\eta(\phi_N), \quad (196)$$

where

$$\epsilon(\phi_N) = \frac{1}{N+1}, \quad (197)$$

and

$$\begin{aligned} \eta(\phi_N) &= \frac{m_{\text{Pl}}^2}{8\pi} \left[\frac{V''(\phi_N)}{V(\phi_N)} - \frac{1}{2} \left(\frac{V'(\phi_N)}{V(\phi_N)} \right)^2 \right] \\ &= \frac{m_{\text{Pl}}^2}{8\pi} \left[\frac{12}{\phi_N^2} - \frac{8}{\phi_N^2} \right] \\ &= \frac{1}{2\pi} \left(\frac{m_{\text{Pl}}}{\phi_N} \right)^2 = \frac{1}{2(N+1)}. \end{aligned} \quad (198)$$

The spectral index is then

$$n = 1 - \frac{3}{N+1} \simeq 0.95. \quad (199)$$

Note that we have assumed slow roll from the beginning in the calculation without *a priori* knowing that it is a good approximation for this choice of potential. However, at the end of the day it is clear that the slow roll ansatz was a good one, since ϵ and η are both of order 0.01.

Finally, we note that the energy density during inflation is characterized by a mass scale

$$\rho^{1/4} \sim \Lambda \sim \lambda^{1/4} m_{\text{Pl}} \sim 10^{15} \text{ GeV}, \quad (200)$$

about the scale for which we expect Grand Unification to be important. This interesting coincidence suggests that the physics of inflation may be found in Grand Unified Theories (GUTs). Different choices of potential $V(\phi)$ will give different values for the amplitudes and shapes of the primordial power spectra. Since the normalization is fixed by the CMB to be $P_{\mathcal{R}} \sim 10^{-5}$, the most useful observables for distinguishing among different potentials are the scalar/tensor ratio r and the scalar spectral index n_S . In single-field inflationary models, the tensor spectral index is fixed by the consistency condition (191), and is therefore not an independent parameter. The consistency condition can therefore be taken to be a *prediction* of single-field inflation, which is in principle verifiable by observation. In practice, this is very difficult, since it involves measuring not just the amplitude of the gravitational wave power spectrum, but also its *shape*. We will see in Section VI that current data place only a rough upper bound on the tensor/scalar ratio r , and it is highly unlikely that any near-future measurement of primordial gravitational waves will be accurate enough to constrain n_T well enough to test the consistency condition. In the next section, we discuss current observational constraints on the form of the inflationary potential.

VI. OBSERVATIONAL CONSTRAINTS

Our simple picture of inflation generated by single, minimally coupled scalar field makes a set of very definite predictions for the form of primordial cosmological fluctuations:

- **Gaussianity:** since the two-point correlation function of a free scalar field $\langle \varphi^2 \rangle$ is Gaussian, cosmological perturbations generated in a single-field inflation model will by necessity also form a Gaussian random distribution.
- **Adiabaticity:** since there is only one order parameter ϕ governing the generation of density perturbations, we expect the perturbations in all the components of the cosmological fluid (baryons, dark matter, neutrinos) to be *in phase* with each other. Such a case is called *adiabatic*. If one or more components fluctuates out of phase with others, these are referred to as *isocurvature* modes. Single-field inflation predicts an absence of isocurvature fluctuations.
- **Scale invariance:** In the limit of de Sitter space, fluctuations in any quantum field are exactly scale invariant, $n = 1$, as a result of the fact that the Hubble parameter is exactly constant. Since slow-roll inflation is quasi-de Sitter, we expect the perturbation spectra to be nearly, but not exactly scale invariant, with $|n_S - 1| = |2\eta - 4\epsilon| \ll 1$.
- **Scalar perturbations dominate over tensor perturbations,** $r = 16\epsilon$.

Furthermore, given a potential $V(\phi)$, we have a “recipe” for calculating the form of the primordial power spectra generated during inflation:

1. Calculate the field value at the end of inflation ϕ_e from

$$\epsilon(\phi_e) = \frac{m_{\text{Pl}}^2}{16\pi} \left(\frac{V'(\phi_e)}{V(\phi_e)} \right)^2 = 1. \quad (201)$$

2. Calculate the field value N e-folds before the end of inflation ϕ_N by integrating backward on the potential from $\phi = \phi_e$,

$$N = \frac{2\sqrt{\pi}}{m_{\text{Pl}}} \int_{\phi_e}^{\phi_N} \frac{d\phi'}{\sqrt{\epsilon(\phi')}}. \quad (202)$$

3. Calculate the normalization of the scalar power spectrum by

$$P_{\mathcal{R}}^{1/2} = \frac{H}{m_{\text{Pl}}\sqrt{\pi\epsilon}} \Big|_{\phi=\phi_N} \sim 10^{-5}, \quad (203)$$

where the CMB quadrupole corresponds to roughly $N = 60$. A more accurate calculation includes the uncertainty in the reheat temperature, which gives a range $N \simeq [46, 60]$, and a corresponding uncertainty in the observable parameters.

4. Calculate the tensor/scalar ratio r and scalar spectral index n_S at $N = [46, 60]$ by

$$r = 16\epsilon(\phi_N), \quad (204)$$

and

$$n_s = 1 - 4\epsilon(\phi_N) + 2\eta(\phi_N), \quad (205)$$

where the second slow roll parameter η is given by:

$$\eta(\phi_N) = \frac{m_{\text{Pl}}^2}{8\pi} \left[\frac{V''(\phi_N)}{V(\phi_N)} - \frac{1}{2} \left(\frac{V'(\phi_N)}{V(\phi_N)} \right)^2 \right]. \quad (206)$$

The key point is that the scalar power spectrum $P_{\mathcal{R}}$ and the tensor power spectrum $P_{\mathcal{T}}$ are both completely determined by the choice of potential $V(\phi)$.⁸ Therefore, if we measure the primordial perturbations in the universe accurately enough, we can in principle constrain the form of the inflationary potential. This is extremely exciting, because it gives us a very rare window into physics at extremely high energy, perhaps as high as the GUT scale or higher, far beyond the reach of accelerator experiments such as the Large Hadron Collider.

It is convenient to divide the set of possible single-field potentials into a few basic types [69]:

- *Large-field potentials* (Fig. 17). These are the simplest potentials one might imagine, with potentials of the form $V(\phi) = m^2\phi^2$, or our example case, $V(\phi) = \lambda\phi^4$. Another widely-noted example of this type of model is inflation on an exponential potential, $V(\phi) = \Lambda^4 \exp(\phi/\mu)$, which has the useful property that both the background evolution and the perturbation equations are exactly solvable. In the large-field case, the field is displaced from the vacuum at the origin by an amount of order $\phi \sim m_{\text{Pl}}$ and rolls down the potential toward the origin. Large-field models are typically characterized by a “red” spectral index $n_S < 1$, and a substantial gravitational wave contribution, $r \sim 0.1$.
- *Small-field potentials* (Fig. 18). These are potentials characteristic of spontaneous symmetry breaking phase transitions, where the field rolls off an unstable equilibrium with $V'(\phi) = 0$ toward a displaced vacuum. Examples of small-field inflation include a simple quadratic potential, $V(\phi) = \lambda(\phi^2 - \mu^2)^2$, inflation from a pseudo-Nambu-Goldstone boson or a shift symmetry in string theory (called *Natural Inflation*) with a potential typically of the form $V(\phi) = \Lambda^4 [1 + \cos(\phi/\mu)]$, or Coleman-Weinberg potentials, $V(\phi) = \lambda\phi^4 \ln(\phi)$. Small-field models are characterized by a red spectral index $n < 1$, and a small tensor/scalar ratio, $r \leq 0.01$.
- *Hybrid potentials* (Fig. 19). A third class of models are potentials for which there is a residual vacuum energy when the field is at the minimum of the potential, for example a potential like $V(\phi) = \lambda(\phi^2 + \mu^2)^2$. In this case, inflation will continue *forever*, so additional physics is required to end inflation and initiate reheating. The *hybrid* mechanism, introduced by Linde [70], solves this problem by adding a second field coupled to the inflaton which is stable for ϕ large, but becomes unstable at a critical field value ϕ_c near the minimum of $V(\phi)$. During inflation, however, only ϕ is dynamical, and these models are effectively single-field. Typical models of this type predict negligible tensor modes, $r \ll 0.01$ and a “blue” spectrum, $n_S > 1$, which is disfavored by the data, and we will not discuss them in more detail here. (Ref. [48] contains a good discussion of current limits on general hybrid models.) Note also that such potentials will also support large-field inflation if the field is displaced far enough from its minimum.

⁸ Strictly speaking, this is true only for scalar fields with a canonical kinetic term, where the speed of sound of perturbations is equal to the speed of light. More complicated scenarios such as DBI inflation [67] require specification of an extra free function, the speed of sound $c_S(\phi)$, to calculate the power spectra. For constraints on this more general class of models, see Ref. [68].

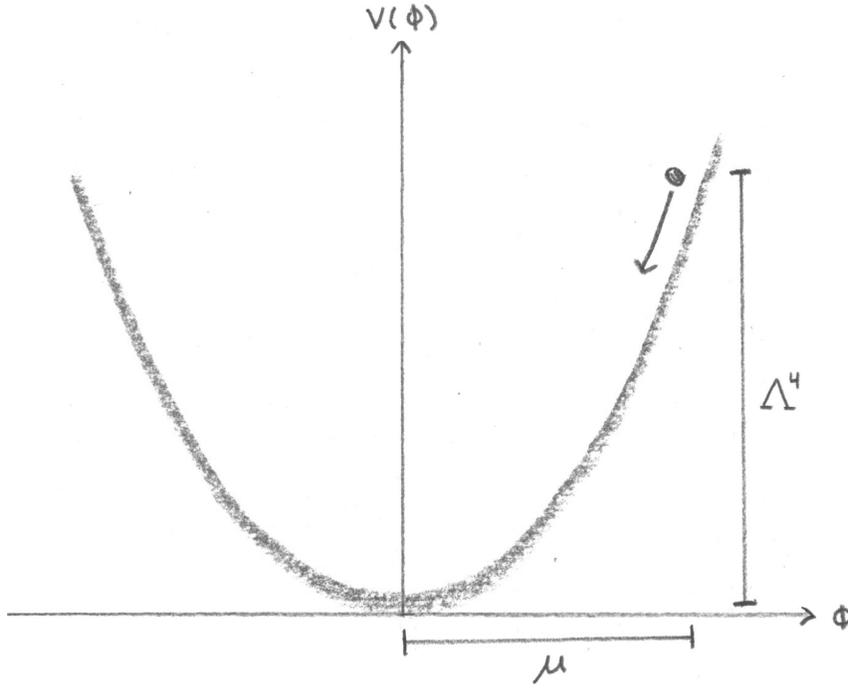


FIG. 17: A schematic of a large-field potential.

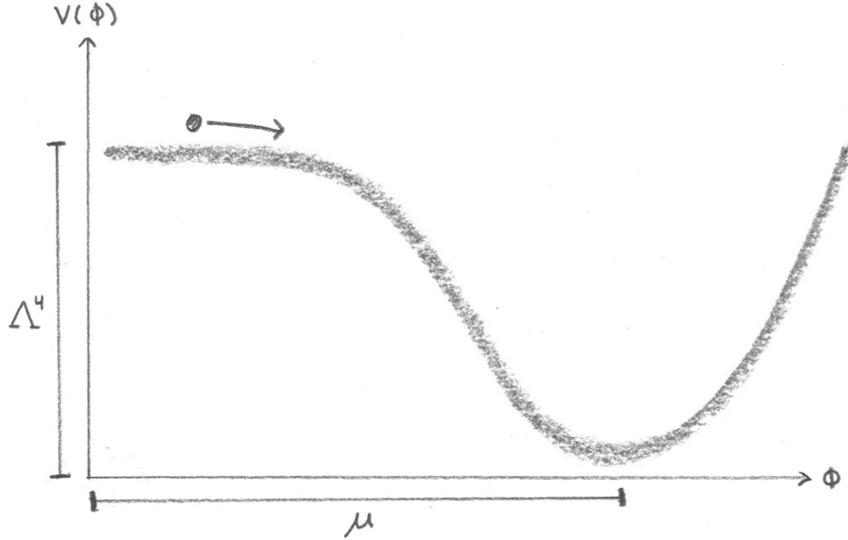


FIG. 18: A schematic of a small-field potential.

An important feature of all of these models is that each is characterized by two basic parameters, the “height” of the potential Λ^4 , which governs the energy density during inflation, and the “width” of the potential μ . (Hybrid models have a third free parameter ϕ_c which sets the end of inflation.) In order to have a flat potential and a slowly rolling field, there must be a hierarchy of scales such that the width is larger than the height, $\Lambda \ll \mu$. As we saw in the case of the $\lambda\phi^4$ large-field model, typical inflationary potentials have widths of order the Planck scale $\mu \sim m_{\text{Pl}}$ and heights of order the scale of Grand Unification $\Lambda \sim M_{\text{GUT}} \sim 10^{15}$ GeV, although models can be constructed for which inflation happens at a much lower scale [70, 71, 72].

The quantities we are interested in for constraining models of inflation are the primordial power spectra $P_{\mathcal{R}}$ and P_T , which are the underlying source of the CMB temperature anisotropy and polarization. However, the observed CMB

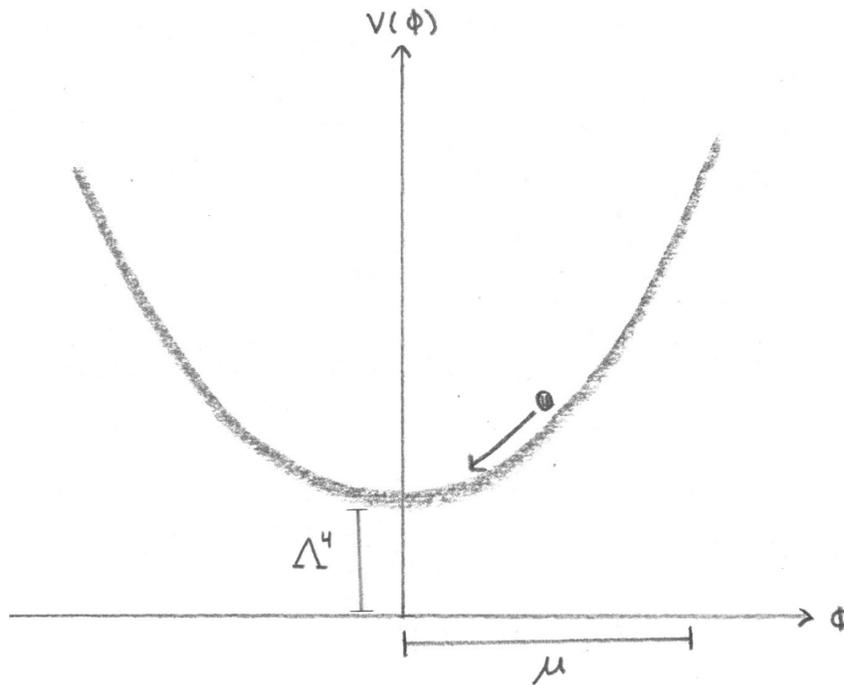


FIG. 19: A schematic of a hybrid potential.

anisotropies depend on a handful of unrelated cosmological parameters, since the primordial fluctuations are processed through the complicated physics of acoustic oscillations. This creates uncertainties due to parameter degeneracies: our best-fit values for r and n_S will depend on what values we choose for the other cosmological parameters such as the baryon density Ω_b and the redshift of reionization z_{ri} . To accurately estimate the errors on r and n_S , we must fit all the relevant parameters *simultaneously*, a process which is computationally intensive, and is typically approached using Bayesian Monte Carlo Markov Chain techniques [73]. Here we simply show the results: Figure 20 shows the regions of the r , n_S parameter space allowed by the WMAP 5-year data set [74, 75]. We have fit over the parameters Ω_{CDM} , Ω_b , Ω_{Lambda} , H_0 , $P_{\mathcal{R}}$, z_{ri} , r , and n_s , with a constraint that the universe must be flat, as predicted by inflation, $\Omega_b + \Omega_{\text{CDM}} + \Omega_{\text{Lambda}} = 1$. We see that the data favor a red spectrum, $n_S < 1$, although the scale-invariant limit $n_S = 1$ is still within the 95%-confidence region. Our example inflation model $V(\phi) = \lambda\phi^4$ is convincingly ruled out by WMAP, but the simple potential $V(\phi) = m^2\phi^2$ is nicely consistent with the data.⁹ Figure 21 shows the WMAP constraint with r on a logarithmic scale, with the prediction of several small-field models for reference. There is no evidence in the WMAP data for a nonzero tensor/scalar ratio r , with a 95%-confidence upper limit of $r < 0.5$. It is possible to improve these constraints somewhat by adding other data sets, for example the ACBAR high-resolution CMB anisotropy measurement [76] or the Sloan Digital Sky Survey [77, 78], which improve the upper limit on the tensor/scalar ratio to $r < 0.3$ or so. Current data are completely consistent with Gaussianity and adiabaticity, as expected from simple single-field inflation models. In the next section, we discuss the outlook for future observation.

VII. OUTLOOK AND CONCLUSION

The basic hot Big Bang scenario, in which the universe arises out of a hot, dense, smooth initial state and cools through expansion, is now supported by a compelling set of observations, including the existence of the Cosmic Microwave Background, the primordial abundances of the elements, and the evolution of structure in the universe, all of which are being measured with unprecedented precision. However, this scenario leaves questions unanswered:

⁹ Liddle and Leach point out that $\lambda\phi^4$ models are special because of their reheating properties, and should be more accurately evaluated at $N = 64$ [56]. However, this assumes that the potential has no other terms which might become dominant during reheating, and in any case is also ruled out by WMAP5.

WMAP 5 Limits on Inflation

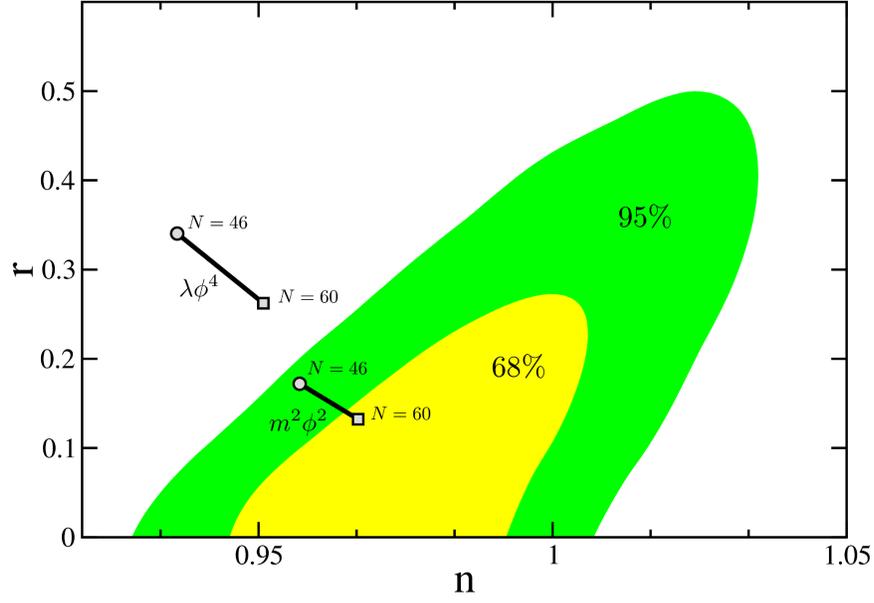


FIG. 20: Constraints on the r, n plane from Cosmic Microwave Background measurements. Shaded regions are the regions allowed by the WMAP5 measurement to 68% and 95% confidence. Models plotted are “large-field” potentials $V(\phi) \propto \phi^2$ and $V(\phi) \propto \phi^4$.

WMAP 5 Limits on Inflation

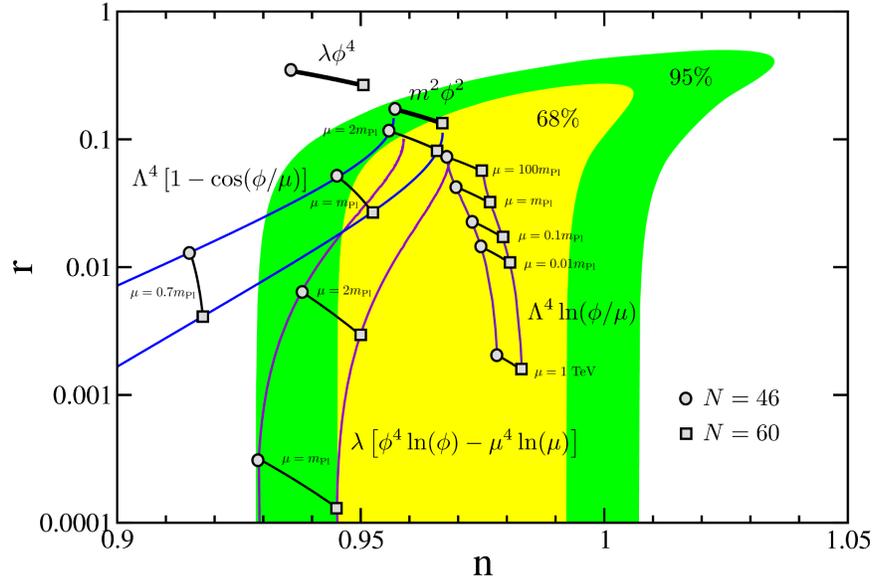


FIG. 21: Constraints on the r, n plane from Cosmic Microwave Background measurements, with the tensor/scalar ratio plotted on a log scale. In addition to the large-field models shown in Fig. 20, three small-field models are plotted against the data: “Natural Inflation” from a pseudo-Nambu-Goldstone boson [79], with potential $V(\phi) = \Lambda^4 [1 - \cos(\phi/\mu)]$, a logarithmic potential $V(\phi) \propto \ln(\phi)$ typical of supersymmetric models [80, 81, 82], and a Coleman-Weinberg potential $V(\phi) \propto \phi^4 \ln(\phi)$.

Why is the universe so big and so old? Why is the universe so close to geometrically flat? What created the initial perturbations which later collapsed to form structure in the universe? The last of these questions is particularly interesting, because recent observations of the CMB, in particular the all-sky anisotropy map made by the landmark WMAP satellite, have directly measured the form of these primordial perturbations. A striking property of these observed primordial perturbations is that they are correlated on scales larger than the cosmological horizon at the time of last scattering. Such apparently *acausal* correlations can only be produced in a few ways [83]:

- Inflation.
- Extra dimensions [84].
- A universe much older than H_0^{-1} [85, 86].
- A varying speed of light [87].

In addition, the WMAP data contain spectacular confirmation of the basic predictions of the inflationary paradigm: a geometrically flat universe with Gaussian, adiabatic, nearly scale-invariant perturbations. No other model explains these properties of the universe with such simplicity and economy, and much attention has been devoted to the implications of WMAP for inflation [29, 48, 74, 75, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98]. Inflation also makes predictions which have not been well tested by current data but *can* be by future experiments, most notably a deviation from a scale-invariant spectrum and the production of primordial gravitational waves. A non-scale-invariant spectrum is weakly favored by the existing data, but constraints on primordial gravity waves are still quite poor. The outlook for improved data is promising: over the next five to ten years, there will be a continuous stream of increasingly high-precision data made available which will allow constraint of cosmological parameters relevant for understanding the early universe. The most useful measurements for direct constraint of the inflationary parameter space are observations of the CMB, and current activity in this area is intense. The Planck satellite mission is scheduled to launch in 2009 [99, 100], and will be complemented by ground- and balloon-based measurements using a variety of technologies and strategies [43, 44, 101, 102, 103, 104, 105, 106].

At the same time, cosmological parameter estimation is a well-developed field. A set of standard cosmological parameters such as the baryon density $\Omega_b h^2$, the matter density $\Omega_m h^2$, the expansion rate $H_0 \equiv 100h$ km/sec are being measured with increasing accuracy. The observable quantities most meaningful for constraining models of inflation are the ratio r of tensor to scalar fluctuation amplitudes, and the spectral index n_S of the scalar power spectrum. This kind of simple parameterization is at the moment sufficient to describe the highest-precision cosmological data sets. Furthermore, the simplest slow-roll models of inflation predict a nearly exact power-law perturbation spectrum. In this sense, a simple concordance cosmology is well-supported by both data and by theoretical expectation. It could be that the underlying universe really is that simple. However, the simplicity of concordance cosmology is at present as much a statement about the data as about the universe itself. Only a handful of parameters are required to explain existing cosmological data. Adding more parameters to the fit does no good: any small improvement in the fit of the model to the data is offset by the statistical penalty one pays for introducing extra parameters [107, 108, 109, 110, 111, 112, 113, 114, 115]. But the optimal parameter set is a moving target: as the data get better, we will be able to probe more parameters. It may be that a “vanilla” universe [116] of a half-dozen or so parameters will continue to be sufficient to explain observation. But it is reasonable to expect that, as measurements improve in accuracy, we will see evidence of deviation from such a lowest-order expectation. This is where the interplay between theory and experiment gains the most leverage, because we must understand: (1) what deviations from a simple universe are predicted by models, and (2) how to look for those deviations in the data. It is of course impossible to predict which of the many possible signals (if any) will be realized in the universe in which we live. I discuss below four of the best motivated possibilities, in order of the quality of current constraints. (For a more detailed treatment of these issues, the reader is referred to the very comprehensive CMBPol Mission Concept Study [117].)

Features in the density power spectrum

Current data are consistent with a purely power-law spectrum of density perturbations, $P(k) \propto k^{n_S-1}$ with a “red” spectrum ($n_S < 1$) favored by the data at about a 90% confidence level, a figure which depends on the choice of parameter set and priors. Assuming it is supported by future data, the detection of a deviation from a scale-invariant ($n_S = 1$) spectrum is a significant milestone, and represents a confirmation of one of the basic predictions of inflation. In slow-roll inflation, this power-law scale dependence is nearly exact, and any additional scale dependence is strongly suppressed. Therefore, detection of a nonzero “running” $\alpha = dn_S/d \ln k$ of the spectral index would be an indication that slow roll is a poor approximation. There is currently no evidence for scale-dependence in the spectral index, but constraints on the overall shape of the power spectrum are likely to improve dramatically through measurements of the CMB anisotropy at small angular scales, improved polarization measurements, and better mapping of large-scale structure. Planck is expected to measure the shape of the spectrum with 2σ uncertainties of order $\Delta n \sim 0.01$

and $\Delta\alpha \sim 0.01$ [118, 119, 120, 121]. Over the longer term, measurements of 21cm radiation from neutral hydrogen promises to be a precise probe of the primordial power spectrum, and would improve these constraints significantly [122].

Primordial Gravitational Waves

In addition to a spectrum $P_{\mathcal{R}}$ of scalar perturbations, inflation generically predicts a spectrum P_T of tensor perturbations. The relative amplitude of the two is determined by the equation of state of the fluid driving inflation,

$$r = 16\epsilon \quad (207)$$

Since the scalar amplitude is known from the COBE normalization to be $P_{\mathcal{R}} \sim H^2/\epsilon \sim 10^{-10}$, it follows that measuring the tensor/scalar ratio r determines the inflationary expansion rate H and the associated energy density ρ . Typical inflation models take place with an energy density of around $\rho \sim (10^{15} \text{ GeV})^4$, which corresponds to a tensor/scalar ratio of $r \sim 0.1$, although this figure is highly model-dependent. Single-field inflation does not make a definite prediction for the value of r : while many choices of potential generate a substantial tensor component, other choices of potential result in an unobservably small tensor/scalar ratio, and there is no particular reason to favor one scenario over another.

There is at present no observational evidence for primordial gravitational waves: the current upper limit on the tensor/scalar ratio is around $r \leq 0.3$. Detection of even a large primordial tensor signal requires extreme sensitivity. The crucial observation is detection of the odd-parity, or B-mode, component of the CMB polarization signal, which is suppressed relative to the temperature fluctuations, themselves at the 10^{-4} level, by at least another four orders of magnitude. This signal is considerably below known foreground levels [123], severely complicating data analysis. Despite the formidable challenges, the observational community has undertaken a broad-based effort to search for the B-mode, and a detection would be a boon for inflationary cosmology. Planck will be sensitive to a tensor/scalar ratio of around $r \simeq 0.1$, and dedicated ground-based measurements can potentially reach limits of order $r \simeq 0.01$. The proposed CMBPol polarization satellite would reach r of order 10^{-3} [117, 124], and direct detection experiments such as BBO could in principle detect r of order 10^{-4} [65].

Primordial Non-Gaussianity

In addition to a power-law power spectrum, inflation predicts that the primordial perturbations will be distributed according to Gaussian statistics. Like running of the power spectrum, non-Gaussianity is suppressed in slow-roll inflation [125]. However, detection of even moderate non-Gaussianity is considerably more difficult. If the perturbations are Gaussian, the two-point correlation function completely describes the perturbations. This is not the case for non-Gaussian fluctuations: higher-order correlations contain additional information. However, higher-order correlations require more statistics and are therefore more difficult to measure, especially at large angular scales where cosmic variance errors are significant. Current limits are extremely weak [88, 126], and future high angular resolution CMB maps will still fall well short of being sensitive to a signal from slow-roll inflation or even weakly *non-slow-roll* models [127]. It will take a strong deviation from the slow-roll scenario to generate observable non-Gaussianity. However, a measurement of non-Gaussianity would in one stroke rule out virtually all slow-roll inflation models and force consideration of more exotic scenarios such as DBI inflation [67], Warm Inflation [128], or curvaton scenarios [129].

Isocurvature perturbations

In a universe where the matter consists of multiple components, there are two general classes of perturbation about a homogeneous background: adiabatic, in which the perturbations in all of the fluid components are in phase, and isocurvature, in which the perturbations have independent phases. Single-field inflation predicts purely adiabatic primordial perturbations, for the simple reason that if there is a single field ϕ responsible for inflation, then there is a single order parameter governing the generation of density perturbations. This is a nontrivial prediction, and the fact that current data are consistent with adiabatic perturbations is support for the idea of quantum generation of perturbations in inflation. However, current limits on the isocurvature fraction are quite weak [130, 131]. If isocurvature modes are detected, it would rule out *all* single-field models of inflation. Multi-field models, on the other hand, naturally contain multiple order parameters and can generate isocurvature modes. Multi-field models are naturally motivated by the string “landscape”, which is believed to contain an enormous number of degrees of freedom. Another possible mechanism for the generation of isocurvature modes is the curvaton mechanism, in which cosmological perturbations are generated by a field other than the inflaton [132, 133].

The rich interplay between theory and observation that characterizes cosmology today is likely to continue for the foreseeable future. As measurements improve, theory will need to become more precise and complete than the simple picture of inflation that we have outlined in these lectures, and single-field inflation models could yet prove to be a poor fit to the data. However, at the moment, such models provide an elegant, compelling, and (most importantly) scientifically useful picture of the very early universe.

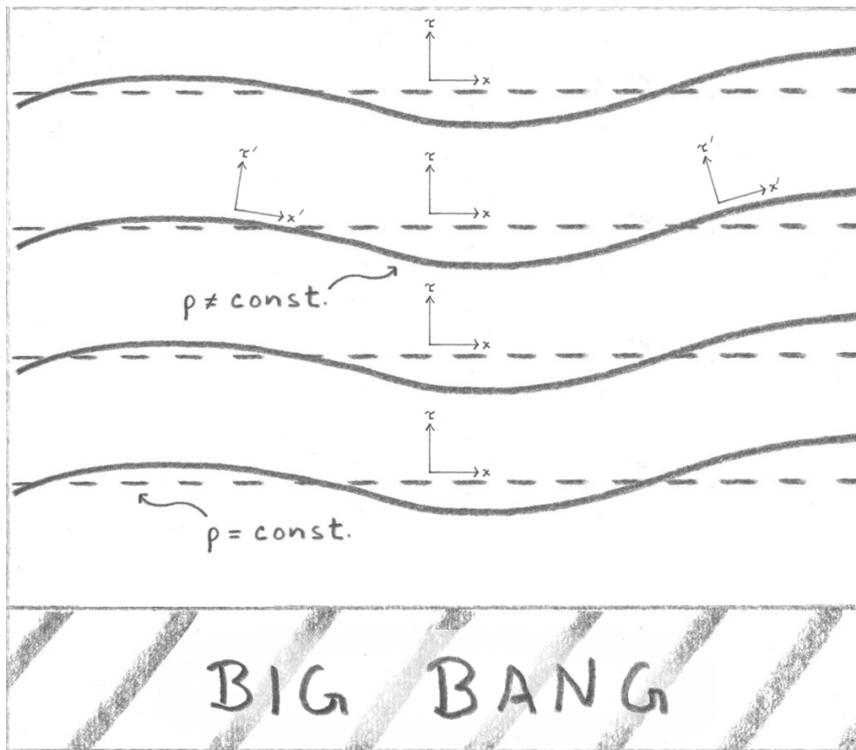


FIG. 22: Foliations of an FRW spacetime. Comoving hypersurfaces (dashed lines) have constant density, but another choice of gauge (solid lines) will have unphysical density fluctuations which are an artifact of the choice of gauge.

Acknowledgments

I would like to thank the organizers of the Theoretical Advanced Studies Institute (TASI) at Univ. of Colorado, Boulder for giving me the opportunity to return to my alma mater to lecture. Various versions of these lectures were also given at the Perimeter Institute Summer School on Particle Physics, Cosmology, and Strings in 2007, at the Second Annual Dirac Lectures at Florida State University in 2008, and at the Research Training Group at the University of Würzburg in 2008. This research is supported in part by the National Science Foundation under grants NSF-PHY-0456777 and NSF-PHY-0757693. I thank Dennis Bessada, Richard Easther, Hiranya Peiris, and Brian Powell for comments on a draft version of the manuscript.

APPENDIX A: THE CURVATURE PERTURBATION IN SINGLE-FIELD INFLATION

In this section, we discuss the generation of perturbations in the density $\delta(\mathbf{x}) \equiv \delta\rho/\rho$ generated during inflation. The process is similar to the case of a free scalar field discussed in Sec. V: the inflaton field ϕ , like any other scalar, will have quantum fluctuations which are stretched to superhorizon scales and subsequently freeze out as classical perturbations. The difference is that the energy density of the universe is dominated by the inflaton potential, so that quantum fluctuations in ϕ generate perturbations in the density ρ . Dealing with such density perturbations is complicated by the fact that in General Relativity, we are free to choose any coordinate system, or *gauge*, we wish. To see why, consider the case of an FRW spacetime evolving with scale factor $a(t)$ and uniform energy density $\rho(t, \mathbf{x}) = \bar{\rho}(t)$. What we mean here by “uniform” energy density, or homogeneity, is that the density is a constant in *comoving* coordinates. But the physics is independent of coordinate system, so we could equally well work in coordinates t', \mathbf{x}' for which constant-time hypersurfaces do *not* have constant density (Fig. 22). Such a division of spacetime into a time coordinate and a set of orthogonal spacelike hypersurfaces is called a *foliation* of the spacetime, and is an arbitrary choice equivalent to a choice of coordinate system.

For an FRW spacetime, comoving coordinates correspond to a foliation of the spacetime into spatial hypersurfaces with constant density: this is the most physically intuitive description of the spacetime. Any other choice of foliation of the spacetime would result in density “perturbations” which are entirely due to the choice of coordinate system.

Such unphysical perturbations are referred to as *gauge modes*. Another way to think of this is that the division between what we call “background” and what we call “perturbation” is itself gauge-dependent. For perturbations with wavelength smaller than the horizon, it is possible to define background and perturbation without ambiguity, since all observers can agree on a definition of time coordinate t and on an average density $\bar{\rho}(t)$. Not so for superhorizon modes: if we consider a perturbation mode with wavelength much larger than the horizon size, observers in different horizons will see themselves in independently evolving, homogeneous patches of the universe: a “perturbation” can be defined only by comparing causally disconnected observers, and there is an inherent gauge ambiguity in how we do this. The canonical paper on gauge issues in General Relativistic perturbation theory is by Bardeen [134]. A good pedagogical treatment with a focus on inflationary perturbations can be found in Ref. [135].

In practice, instead of the density perturbation δ , the quantity most directly relevant to CMB physics is the Newtonian potential Φ on the surface of last scattering. For example, this is the quantity that directly appears in Eq. (50) for the Sachs-Wolfe Effect. The Newtonian potential is related to the density perturbation δ through the Poisson Equation:

$$\nabla^2 \Phi = 4\pi G \bar{\rho} a^2 \delta, \quad (\text{A1})$$

where the factor of a^2 comes from defining the gradient ∇ relative to comoving coordinates. Like δ , the Newtonian potential Φ is a gauge-dependent quantity: its value depends on how we foliate the spacetime. For example, we are free to choose spatial hypersurfaces such that the density is constant, and the Newtonian potential vanishes everywhere: $\Phi(t, \mathbf{x}) = 0$. This foliation of the spacetime is equivalent to the qualitative picture above of different horizon volumes as independently evolving homogeneous universes. Observers in different horizons use the density ρ to synchronize their clocks with one another. Such a foliation is not very useful for computing the Sachs-Wolfe effect, however! Instead, we need to define a gauge which corresponds to the Newtonian limit in the present universe. To accomplish this, we describe the evolution of a scalar field dominated cosmology using the useful fluid flow approach [136, 137, 138, 139, 140]. (An alternate strategy involves the construction of gauge-invariant variables: see Refs. [141, 142] for reviews.)

Consider a scalar field ϕ in an arbitrary background $g_{\mu\nu}$. The stress-energy tensor of the scalar field may be written

$$T_{\mu\nu} = \phi_{,\mu} \phi_{,\nu} - g_{\mu\nu} \left[\frac{1}{2} g^{\alpha\beta} \phi_{,\alpha} \phi_{,\beta} - V(\phi) \right]. \quad (\text{A2})$$

Note that we have not yet made any assumptions about the metric $g_{\mu\nu}$ or about the scalar field ϕ . Equation (A2) is a completely general expression. We can define a fluid four-velocity for the scalar field by

$$u_\mu \equiv \frac{\phi_{,\mu}}{\sqrt{g^{\alpha\beta} \phi_{,\alpha} \phi_{,\beta}}}. \quad (\text{A3})$$

It is not immediately obvious why this should be considered a four-velocity. Consider any perfect fluid filling spacetime. Each element of the fluid has four-velocity $u^\mu(x)$ at every point in spacetime which is everywhere timelike,

$$u^\mu(x) u_\mu(x) = 1 \quad \forall x. \quad (\text{A4})$$

Such a collection of four-vectors is called a *timelike congruence*. We can draw the congruence defined by the fluid four-velocity as a set of flow lines in spacetime (Fig. 23). Each event P in spacetime has one and only one flow line passing through it. The fluid four-velocity is then a set of unit-normalized tangent vectors to the flow lines, $u^\mu u_\mu = 1$. For a scalar field, we construct a timelike congruence by Eq. (A3), which is by construction unit normalized:

$$u^\mu u_\mu = \frac{g^{\mu\nu} \phi_{,\mu} \phi_{,\nu}}{g^{\alpha\beta} \phi_{,\alpha} \phi_{,\beta}} = 1. \quad (\text{A5})$$

We then define the “time” derivative of any scalar quantity $f(x)$ by the projection of the derivative along the fluid four-velocity:

$$\dot{f} \equiv u^\mu f_{,\mu}. \quad (\text{A6})$$

In particular, the time derivative of the scalar field itself is

$$\dot{\phi} \equiv u^\mu \phi_{,\mu} = \sqrt{g^{\alpha\beta} \phi_{,\alpha} \phi_{,\beta}}. \quad (\text{A7})$$

Note that in the homogeneous case, we recover the usual time derivative,

$$\nabla \phi = 0 \Rightarrow \dot{\phi} = \sqrt{g^{00} \phi_{,0} \phi_{,0}} = \frac{d\phi}{dt}. \quad (\text{A8})$$

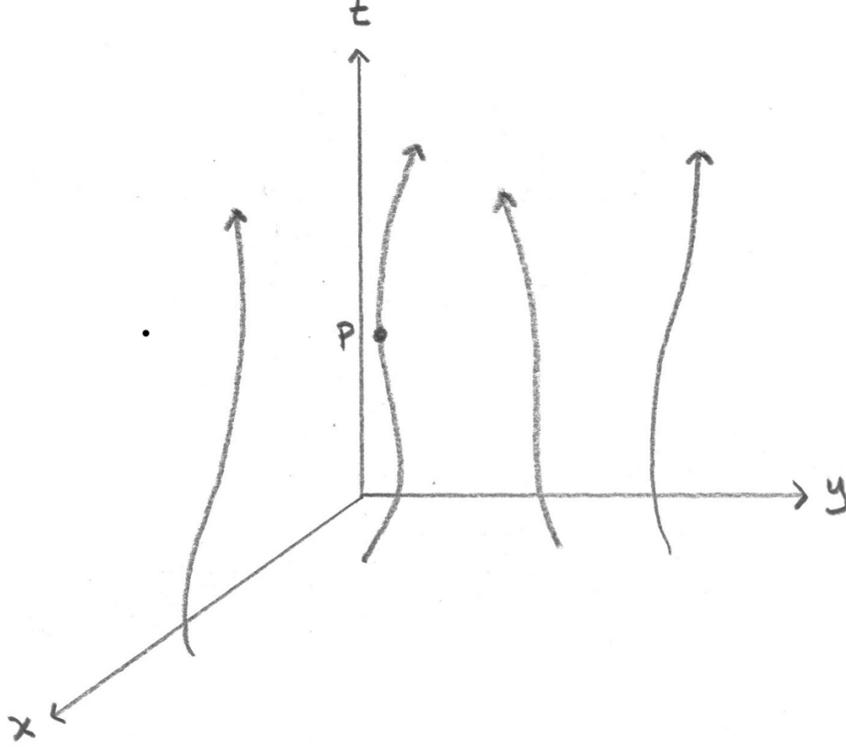


FIG. 23: A timelike congruence in spacetime. Each event P is intersected by exactly one world line in the congruence.

The stress-energy tensor (A2) in terms of $\dot{\phi}$ takes the form

$$T_{\mu\nu} = \left[\frac{1}{2} \dot{\phi}^2 + V(\phi) \right] u_\mu u_\nu + \left[\frac{1}{2} \dot{\phi}^2 - V(\phi) \right] (u_\mu u_\nu - g_{\mu\nu}). \quad (\text{A9})$$

We can then define a generalized density ρ and and pressure p by

$$\begin{aligned} \rho &\equiv \frac{1}{2} \dot{\phi}^2 + V(\phi), \\ p &\equiv \frac{1}{2} \dot{\phi}^2 - V(\phi). \end{aligned} \quad (\text{A10})$$

Note that despite the familiar form of these expressions, they are defined without any assumption of homogeneity of the scalar field or even the imposition of a particular metric.

In terms of the generalized density and pressure, the stress-energy (A2) is

$$T_{\mu\nu} = \rho u_\mu u_\nu + p h_{\mu\nu}, \quad (\text{A11})$$

where the tensor $h_{\mu\nu}$ is defined as:

$$h_{\mu\nu} \equiv u_\mu u_\nu - g_{\mu\nu}. \quad (\text{A12})$$

The tensor $h_{\mu\nu}$ can be easily seen to be a projection operator onto hypersurfaces orthogonal to the four-velocity u^μ . For any vector field A^μ , the product $h_{\mu\nu} A^\nu$ is identically orthogonal to the four-velocity:

$$(h_{\mu\nu} A^\nu) u^\mu = A^\nu (h_{\mu\nu} u^\mu) = 0. \quad (\text{A13})$$

Therefore, as in the case of the time derivative, we can define gradients by projecting the derivative onto surfaces orthogonal to the four-velocity

$$(\nabla f)^\mu \equiv h^{\mu\nu} f_{,\nu}. \quad (\text{A14})$$

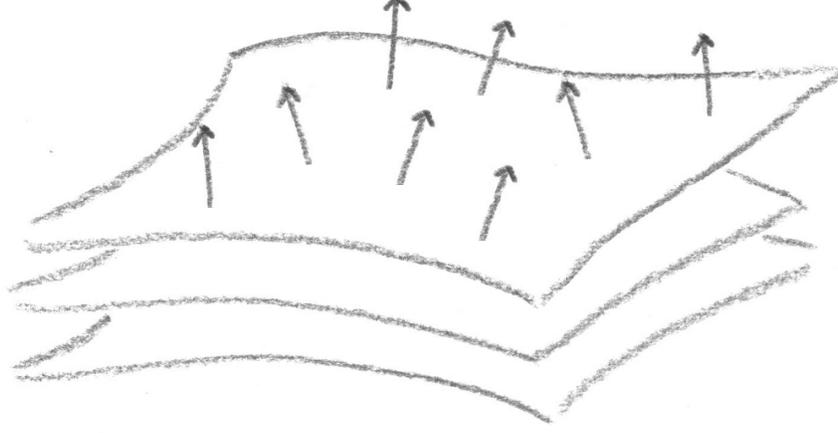


FIG. 24: A comoving foliation of spacetime. Spatial hypersurfaces are everywhere orthogonal to the fluid four-velocity u^μ .

In the case of a scalar field fluid with four-velocity given by Eq. (A3), the gradient of the field identically vanishes,

$$(\nabla\phi)^\mu = 0. \quad (\text{A15})$$

Note that despite its relation to a “spatial” gradient, ∇f is a covariant quantity, *i.e.* a four-vector.

Our fully covariant definitions of “time” derivatives and “spatial” gradients suggest a natural foliation of the spacetime into spacelike hypersurfaces, with time coordinate orthogonal to those hypersurfaces. We can define spatial hypersurfaces to be everywhere orthogonal to the fluid flow (Fig. 24). This is equivalent to choosing a coordinate system for which $u^i = 0$ everywhere. Such a gauge choice is called *comoving* gauge. In the case of a scalar field, we can equivalently define comoving gauge as a coordinate system in which spatial gradients of the scalar field $\phi_{,i}$ are defined to vanish. Therefore the time derivative (A6) is just the derivative with respect to the coordinate time in comoving gauge

$$\dot{\phi} = \left(\frac{\partial\phi}{\partial t} \right)_c. \quad (\text{A16})$$

Similarly, the generalized density and pressure (A10) are just defined to be those quantities as measured in comoving gauge.

The equations of motion for the fluid can be derived from stress-energy conservation,

$$T^{\mu\nu}{}_{;\nu} = 0 = \dot{\rho}u^\mu + (\nabla p)^\mu + (\rho + p)(\dot{u}^\mu + u^\mu\Theta), \quad (\text{A17})$$

where the quantity Θ is defined as the divergence of the four-velocity,

$$\Theta \equiv u^\mu{}_{;\mu}. \quad (\text{A18})$$

We can group the terms multiplied by u^μ separately, resulting in familiar-looking equations for the generalized density and pressure

$$\begin{aligned} \dot{\rho} + \Theta(\rho + p) &= 0, \\ (\nabla p)^\mu + (\rho + p)\dot{u}^\mu &= 0. \end{aligned} \quad (\text{A19})$$

The first of these equations, similar to the usual continuity equation in the homogeneous case, can be rewritten using the definitions of the generalized density and pressure (A10) in terms of the field as

$$\ddot{\phi} + \Theta\dot{\phi} + V'(\phi) = 0. \quad (\text{A20})$$

This suggests identifying the divergence Θ as a generalization of the Hubble parameter H in the homogeneous case. In fact, if we take $g_{\mu\nu}$ to be a flat Friedmann-Robertson-Walker (FRW) metric and take comoving gauge, $u^\mu = (1, 0, 0, 0)$, we have

$$u^\mu{}_{;\mu} = 3H, \quad (\text{A21})$$

and the generalized equation of motion (A20) becomes the familiar equation of motion for a homogeneous scalar,

$$\ddot{\phi} + 3H\dot{\phi} + V'(\phi) = 0. \quad (\text{A22})$$

Now consider perturbations $\delta g_{\mu\nu}$ about a flat FRW metric,

$$g_{\mu\nu} = a^2(\tau) [\eta_{\mu\nu} + \delta g_{\mu\nu}], \quad (\text{A23})$$

where τ is the conformal time and η is the Minkowski metric $\eta = \text{diag}(1, -1, -1, -1)$. A general metric perturbation $\delta g_{\mu\nu}$ can be separated into components which transform independently under coordinate transformations [134],

$$\delta g_{\mu\nu} = \delta g_{\mu\nu}^{\text{scalar}} + \delta g_{\mu\nu}^{\text{vector}} + \delta g_{\mu\nu}^{\text{tensor}}. \quad (\text{A24})$$

The tensor component is just the transverse-traceless gravitational wave perturbation, discussed in Section V, and vector perturbations are not sourced by single-field inflation. We therefore specialize to the case of scalar perturbations, for which the metric perturbations can be written generally in terms of four scalar functions of space and time A , B , \mathcal{R} , and H_T :

$$\begin{aligned} \delta g_{00} &= 2A \\ \delta g_{0i} &= \partial_i B \\ \delta g_{ij} &= 2[\mathcal{R}\delta_{ij} + \partial_i \partial_j H_T]. \end{aligned} \quad (\text{A25})$$

We are interested in calculating \mathcal{R} . Recall that in the Newtonian limit of General Relativity, we can write perturbations about the Minkowski metric in terms of the Newtonian potential Φ as:

$$ds^2 = (1 + 2\Phi) dt^2 - (1 - 2\Phi) \delta_{ij} dx^i dx^j. \quad (\text{A26})$$

Similarly, we can write Newtonian perturbations about a flat FRW metric as

$$ds^2 = a^2(\tau) [(1 + 2\Phi) d\tau^2 - (1 - 2\Phi) \delta_{ij} dx^i dx^j]. \quad (\text{A27})$$

We therefore expect $\Phi \propto \mathcal{R}$ in the Newtonian limit. A careful calculation [138, 141] gives

$$\Phi = -\frac{3(1+w)}{5+3w} \mathcal{R}, \quad (\text{A28})$$

so that in a matter-dominated universe,

$$\Phi = -\frac{3}{5} \mathcal{R}. \quad (\text{A29})$$

In these expressions, \mathcal{R} is the curvature perturbation measured on comoving hypersurfaces. To see qualitatively why comoving gauge corresponds correctly to the Newtonian limit in the current universe, consider the end of inflation. Since inflation ends at a particular field value $\phi = \phi_e$, comoving gauge corresponds to a foliation for which inflation ends at *constant time* at all points in space: all observers synchronize their clocks to $\tau = 0$ at the end of inflation. This means that the background, or unperturbed universe is exactly the homogeneous case diagrammed in Fig. 13, and the comoving curvature perturbation \mathcal{R} is the Newtonian potential measured relative to that background.

To calculate \mathcal{R} , we start by calculating the four-velocity u^μ in terms of the perturbed metric.¹⁰ If we specialize to comoving gauge, $u^i \equiv 0$, the norm of the four-velocity can be written

$$u^\mu u_\mu = a^2(1 + 2A) (u^0)^2 = 1, \quad (\text{A30})$$

and the timelike component of the four-velocity is, to linear order,

$$\begin{aligned} u^0 &= \frac{1}{a} (1 - A) \\ u_0 &= a (1 + A). \end{aligned} \quad (\text{A31})$$

¹⁰ This treatment closely follows that of Sasaki and Stewart [139].

The velocity divergence Θ is then

$$\begin{aligned}\Theta &= u^\mu{}_{;\mu} = u^0{}_{,0} + \Gamma^\alpha{}_{\alpha 0} u^0 \\ &= 3H \left[1 - A - \frac{1}{aH} \left(\frac{\partial \mathcal{R}}{\partial \tau} + \frac{1}{3} \partial_i \partial_i \frac{\partial H_T}{\partial \tau} \right) \right],\end{aligned}\quad (\text{A32})$$

where the unperturbed Hubble parameter is defined as

$$H \equiv \frac{1}{a^2} \frac{\partial a}{\partial \tau}.\quad (\text{A33})$$

Fourier expanding H_T ,

$$\partial_i \partial_i H_T = k^2 H_T,\quad (\text{A34})$$

we see that for long-wavelength modes $k \ll aH$, the last term in Eq. (A32) can be ignored, and the velocity divergence is

$$\Theta \simeq 3H \left[1 - A - \frac{1}{aH} \frac{\partial \mathcal{R}}{\partial \tau} \right].\quad (\text{A35})$$

Remembering the definition of the number of e-folds in the unperturbed case,

$$N \equiv \int H dt.\quad (\text{A36})$$

we can define a generalized number of e-folds as the integral of the velocity divergence along comoving world lines:

$$\mathcal{N} \equiv \frac{1}{3} \int \Theta ds = \frac{1}{3} \int \Theta [a(1+A) d\tau].\quad (\text{A37})$$

Using Eq. (A35) for Θ and evaluating to linear order in the metric perturbation results in

$$\mathcal{N} = \int H dt - \mathcal{R},\quad (\text{A38})$$

and we have a simple expression for the curvature perturbation,

$$\mathcal{R} = N - \mathcal{N}.\quad (\text{A39})$$

This requires a little physical interpretation: we defined comoving hypersurfaces such that the field has no spatial variation,

$$(\nabla \phi)^\mu = 0 \Rightarrow \phi = \text{const}.\quad (\text{A40})$$

Then \mathcal{N} is the number of e-folds measured on comoving hypersurfaces. But we can equivalently foliate the spacetime such that spatial hypersurfaces are flat, and the field exhibits spatial fluctuations:

$$A = \mathcal{R} = 0 \Rightarrow \phi \neq \text{const}.\quad (\text{A41})$$

On flat hypersurfaces, the field varies, but the curvature does not, so that the metric on these hypersurfaces is exactly of the FRW form (11) with $k = 0$. We then see immediately that

$$N = \int H dt = \text{const}.\quad (\text{A42})$$

is the number of e-folds measured on flat hypersurfaces, and has no spatial variation. The curvature perturbation \mathcal{R} is the difference in the number of e-folds between the two sets of hypersurfaces (Fig. 25). This can be expressed to linear order in terms of the field variation $\delta\phi$ on flat hypersurfaces as

$$\mathcal{R} = N - \mathcal{N} = \frac{\delta N}{\delta \phi} \delta \phi\quad (\text{A43})$$

where \mathcal{R} is measured on comoving hypersurfaces, and $\delta N/\delta\phi$ and $\delta\phi$ are measured on flat hypersurfaces. We can

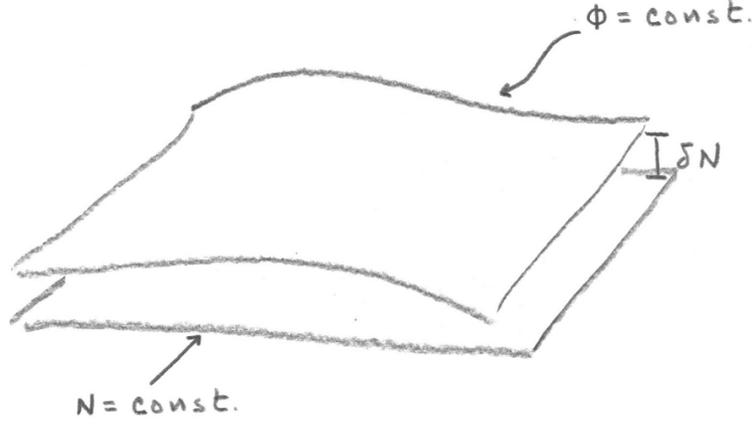


FIG. 25: Flat and comoving hypersurfaces.

express N as a function of the field ϕ :

$$N = \int H dt = \int \frac{H}{\dot{\phi}} d\phi. \quad (\text{A44})$$

For monotonic field evolution, we can express $\dot{\phi}$ as a function of ϕ , so that

$$\frac{\delta N}{\delta \phi} = \frac{H}{\dot{\phi}}, \quad (\text{A45})$$

and the curvature perturbation is given by

$$\mathcal{R} = N - \mathcal{N} = \frac{\delta N}{\delta \phi} \delta \phi = \frac{H}{\dot{\phi}} \delta \phi. \quad (\text{A46})$$

Note that this is an expression for the metric perturbation \mathcal{R} on comoving hypersurfaces, calculated in terms of quantities defined on *flat* hypersurfaces. For $\delta \phi$ produced by quantum fluctuations in inflation, the two-point correlation function is

$$\sqrt{\langle \delta \phi^2 \rangle} = \frac{H}{2\pi}, \quad (\text{A47})$$

and the two-point correlation function for curvature perturbations is

$$\sqrt{\langle \mathcal{R}^2 \rangle} = \frac{H^2}{2\pi \dot{\phi}} = \frac{H}{m_{\text{Pl}} \sqrt{\pi \epsilon}}, \quad (\text{A48})$$

which is the needed result.

-
- [1] A. H. Guth, Phys. Rev. D **23**, 347 (1981).
 - [2] A. D. Linde, Phys. Lett. B **108**, 389 (1982).
 - [3] A. Albrecht and P. J. Steinhardt, Phys. Rev. Lett. **48**, 1220 (1982).
 - [4] E. B. Gliner, Sov. Phys.-JETP **22**, 378 (1966).
 - [5] E. B. Gliner and I. G. Dymnikova, Sov. Astron. Lett. **1**, 93 (1975).
 - [6] A. Linde, Lect. Notes Phys. **738**, 1 (2008) [arXiv:0705.0164 [hep-th]].
 - [7] A. A. Starobinsky, JETP Lett. **30** (1979) 682 [Pisma Zh. Eksp. Teor. Fiz. **30** (1979) 719].
 - [8] V. F. Mukhanov and G. V. Chibisov, JETP Lett. **33** (1981) 532 [Pisma Zh. Eksp. Teor. Fiz. **33** (1981) 549].
 - [9] S. W. Hawking, Phys. Lett. B **115**, 295 (1982).

- [10] S. W. Hawking and I. G. Moss, Nucl. Phys. B **224**, 180 (1983).
- [11] A. A. Starobinsky, Phys. Lett. B **117** (1982) 175.
- [12] A. H. Guth and S. Y. Pi, Phys. Rev. Lett. **49**, 1110 (1982).
- [13] J. M. Bardeen, P. J. Steinhardt and M. S. Turner, Phys. Rev. D **28**, 679 (1983).
- [14] D. H. Lyth and A. Riotto, Phys. Rept. **314**, 1 (1999) [arXiv:hep-ph/9807278].
- [15] G. S. Watson, arXiv:astro-ph/0005003.
- [16] A. Riotto, arXiv:hep-ph/0210162.
- [17] C. H. Lineweaver, arXiv:astro-ph/0305179.
- [18] W. H. Kinney, arXiv:astro-ph/0301448.
- [19] M. Trodden and S. M. Carroll, arXiv:astro-ph/0401547.
- [20] D. Baumann and H. V. Peiris, arXiv:0810.3022 [astro-ph].
- [21] S. Weinberg, “Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity,” John Wiley & Sons (1972) Ch. 13.
- [22] W. L. Freedman *et al.* [HST Collaboration], Astrophys. J. **553**, 47 (2001) [arXiv:astro-ph/0012376].
- [23] M. J. White, D. Scott and J. Silk, Ann. Rev. Astron. Astrophys. **32**, 319 (1994).
- [24] W. Hu and S. Dodelson, Ann. Rev. Astron. Astrophys. **40**, 171 (2002) [arXiv:astro-ph/0110414].
- [25] A. Kosowsky, arXiv:astro-ph/0102402.
- [26] D. Samtleben, S. Staggs and B. Winstein, Ann. Rev. Nucl. Part. Sci. **57**, 245 (2007) [arXiv:0803.0834 [astro-ph]].
- [27] W. Hu, arXiv:0802.3688 [astro-ph].
- [28] E. W. Kolb and M. S. Turner, “The Early Universe,” Addison-Wesley (1990), Ch. 3.
- [29] J. Dunkley *et al.* [WMAP Collaboration], arXiv:0803.0586 [astro-ph].
- [30] P. S. Henry, Nature, **231**, 516 (1971).
- [31] B. E. Corey and D. T. Wilkinson, Bull. Amer. Astron. Soc., **8**, 351 (1976).
- [32] G. F. Smoot, M. V. Gorenstein, and R. A. Muller, Phys. Rev. Lett **39**, 898 (1977).
- [33] C. L. Bennett *et al.*, Astrophys. J. **464**, L1 (1996) [arXiv:astro-ph/9601067].
- [34] G. Hinshaw *et al.* [WMAP Collaboration], arXiv:0803.0732 [astro-ph].
- [35] R. K. Sachs and A. M. Wolfe, Astrophys. J. **147**, 73 (1967).
- [36] A. D. Sakharov, JETP **49**, 345 (1965).
- [37] Y. B. Zeldovich and R. A. Sunyaev, Astrophys. Space Sci. **4**, 301 (1969).
- [38] R. A. Sunyaev and Y. B. Zeldovich, Astrophys. Space Sci. **7**, 3 (1970).
- [39] P. J. E. Peebles and J. T. Yu, Astrophys. J. **162**, 815 (1970).
- [40] C. P. Ma and E. Bertschinger, Astrophys. J. **455**, 7 (1995) [arXiv:astro-ph/9506072].
- [41] A. Kosowsky, New Astron. Rev. **43**, 157 (1999) [arXiv:astro-ph/9904102].
- [42] M. Zaldarriaga, arXiv:astro-ph/0305272.
- [43] E. M. Leitch, J. M. Kovac, N. W. Halverson, J. E. Carlstrom, C. Pryke and M. W. E. Smith, Astrophys. J. **624**, 10 (2005) [arXiv:astro-ph/0409357].
- [44] J. L. Sievers *et al.*, arXiv:astro-ph/0509203.
- [45] T. E. Montroy *et al.*, Astrophys. J. **647**, 813 (2006) [arXiv:astro-ph/0507514].
- [46] J. H. Wu *et al.*, arXiv:astro-ph/0611392.
- [47] M. R. Nolta *et al.* [WMAP Collaboration], arXiv:0803.0593 [astro-ph].
- [48] E. Komatsu *et al.* [WMAP Collaboration], arXiv:0803.0547 [astro-ph].
- [49] M. Kamionkowski, Science **280**, 1397 (1998) [arXiv:astro-ph/9806347].
- [50] J. J. Levin, Phys. Rept. **365**, 251 (2002) [arXiv:gr-qc/0108043].
- [51] A. A. Starobinsky, Phys. Lett. B **91** (1980) 99.
- [52] A. D. Linde, Mod. Phys. Lett. A **1**, 81 (1986).
- [53] A. H. Guth, Phys. Rept. **333**, 555 (2000) [arXiv:astro-ph/0002156].
- [54] A. Aguirre, arXiv:0712.0571 [hep-th].
- [55] S. Winitzki, “Eternal inflation,” *Hackensack, USA: World Scientific (2008)*.
- [56] A. R. Liddle and S. M. Leach, Phys. Rev. D **68**, 103503 (2003) [arXiv:astro-ph/0305263].
- [57] W. H. Kinney and A. Riotto, JCAP **0603**, 011 (2006) [arXiv:astro-ph/0511127].
- [58] W. G. Unruh, Phys. Rev. D **14**, 870 (1976).
- [59] L. Hui and W. H. Kinney, Phys. Rev. D **65**, 103507 (2002) [arXiv:astro-ph/0109107].
- [60] U. H. Danielsson, Phys. Rev. D **66**, 023511 (2002) [arXiv:hep-th/0203198].
- [61] R. Easther, B. R. Greene, W. H. Kinney and G. Shiu, Phys. Rev. D **66**, 023518 (2002) [arXiv:hep-th/0204129].
- [62] J. Martin and R. H. Brandenberger, Phys. Rev. D **63**, 123501 (2001) [arXiv:hep-th/0005209].
- [63] J. C. Niemeyer, Phys. Rev. D **63**, 123502 (2001) [arXiv:astro-ph/0005533].
- [64] W. H. Kinney, Phys. Rev. D **72**, 023515 (2005) [arXiv:gr-qc/0503017].
- [65] T. L. Smith, M. Kamionkowski and A. Cooray, Phys. Rev. D **73**, 023504 (2006) [arXiv:astro-ph/0506422].
- [66] B. C. Friedman, A. Cooray and A. Melchiorri, Phys. Rev. D **74**, 123509 (2006) [arXiv:astro-ph/0610220].
- [67] E. Silverstein and D. Tong, Phys. Rev. D **70**, 103505 (2004) [arXiv:hep-th/0310221].
- [68] N. Agarwal and R. Bean, Phys. Rev. D **79**, 023503 (2009) [arXiv:0809.2798 [astro-ph]].
- [69] S. Dodelson, W. H. Kinney and E. W. Kolb, Phys. Rev. D **56**, 3207 (1997) [arXiv:astro-ph/9702166].
- [70] A. D. Linde, Phys. Rev. D **49**, 748 (1994) [arXiv:astro-ph/9307002].
- [71] L. Knox and M. S. Turner, Phys. Rev. Lett. **70**, 371 (1993) [arXiv:astro-ph/9209006].

- [72] W. H. Kinney and K. T. Mahanthappa, Phys. Rev. D **53**, 5455 (1996) [arXiv:hep-ph/9512241].
- [73] A. Lewis and S. Bridle, Phys. Rev. D **66**, 103511 (2002) [arXiv:astro-ph/0205436].
- [74] W. H. Kinney, E. W. Kolb, A. Melchiorri and A. Riotto, Phys. Rev. D **74**, 023502 (2006) [arXiv:astro-ph/0605338].
- [75] W. H. Kinney, E. W. Kolb, A. Melchiorri and A. Riotto, Phys. Rev. D **78**, 087302 (2008) [arXiv:0805.2966 [astro-ph]].
- [76] C. L. Reichardt *et al.*, arXiv:0801.1491 [astro-ph].
- [77] J. Loveday [the SDSS Collaboration], arXiv:astro-ph/0207189.
- [78] K. N. Abazajian *et al.* [SDSS Collaboration], arXiv:0812.0649 [astro-ph].
- [79] K. Freese, J. A. Frieman and A. V. Olinto, Phys. Rev. Lett. **65**, 3233 (1990).
- [80] G. R. Dvali, Q. Shafi and R. K. Schaefer, Phys. Rev. Lett. **73**, 1886 (1994) [arXiv:hep-ph/9406319].
- [81] E. D. Stewart, Phys. Rev. D **51**, 6847 (1995) [arXiv:hep-ph/9405389].
- [82] J. D. Barrow and P. Parsons, Phys. Rev. D **52**, 5576 (1995) [arXiv:astro-ph/9506049].
- [83] D. N. Spergel and M. Zaldarriaga, Phys. Rev. Lett. **79**, 2180 (1997) [arXiv:astro-ph/9705182].
- [84] J. Khoury, B. A. Ovrut, P. J. Steinhardt and N. Turok, Phys. Rev. D **64**, 123522 (2001) [arXiv:hep-th/0103239].
- [85] J. Khoury, P. J. Steinhardt and N. Turok, Phys. Rev. Lett. **92**, 031302 (2004) [arXiv:hep-th/0307132].
- [86] R. Brandenberger and N. Shuhmaher, JHEP **0601**, 074 (2006) [arXiv:hep-th/0511299].
- [87] A. Albrecht and J. Magueijo, Phys. Rev. D **59**, 043516 (1999) [arXiv:astro-ph/9811018].
- [88] D. N. Spergel *et al.* [WMAP Collaboration], Astrophys. J. Suppl. **170**, 377 (2007) [arXiv:astro-ph/0603449].
- [89] L. Alabidi and D. H. Lyth, JCAP **0608**, 013 (2006) [arXiv:astro-ph/0603539].
- [90] U. Seljak, A. Slosar and P. McDonald, JCAP **0610**, 014 (2006) [arXiv:astro-ph/0604335].
- [91] J. Martin and C. Ringeval, JCAP **0608**, 009 (2006) [arXiv:astro-ph/0605367].
- [92] J. Lesgourgues, A. A. Starobinsky and W. Valkenburg, JCAP **0801**, 010 (2008) [arXiv:0710.1630 [astro-ph]].
- [93] H. V. Peiris and R. Easther, JCAP **0807**, 024 (2008) [arXiv:0805.2154 [astro-ph]].
- [94] L. Alabidi and J. E. Lidsey, arXiv:0807.2181 [astro-ph].
- [95] J. Q. Xia, H. Li, G. B. Zhao and X. Zhang, Phys. Rev. D **78**, 083524 (2008) [arXiv:0807.3878 [astro-ph]].
- [96] J. Hamann, J. Lesgourgues and W. Valkenburg, JCAP **0804**, 016 (2008) [arXiv:0802.0505 [astro-ph]].
- [97] T. L. Smith, M. Kamionkowski and A. Cooray, Phys. Rev. D **78**, 083525 (2008) [arXiv:0802.1530 [astro-ph]].
- [98] H. Li *et al.*, arXiv:0812.1672 [astro-ph].
- [99] [Planck Collaboration], arXiv:astro-ph/0604069.
- [100] F. R. Bouchet [Planck Collaboration], Mod. Phys. Lett. A **22**, 1857 (2007).
- [101] C. L. Kuo *et al.*, arXiv:astro-ph/0611198.
- [102] J. E. Ruhl *et al.* [The SPT Collaboration], arXiv:astro-ph/0411122.
- [103] K. W. Yoon *et al.*, large angular scale CMB polarimeter," arXiv:astro-ph/0606278.
- [104] A. C. Taylor [the Clover Collaboration], New Astron. Rev. **50**, 993 (2006) [arXiv:astro-ph/0610716].
- [105] D. Samtleben and f. t. Q. collaboration, arXiv:0806.4334 [astro-ph].
- [106] B. P. Crill *et al.*, arXiv:0807.1548 [astro-ph].
- [107] R. Trotta, Mon. Not. Roy. Astron. Soc. **378**, 72 (2007) [arXiv:astro-ph/0504022].
- [108] J. Magueijo and R. D. Sorkin, Mon. Not. Roy. Astron. Soc. Lett. **377**, L39 (2007) [arXiv:astro-ph/0604410].
- [109] D. Parkinson, P. Mukherjee and A. R. Liddle, Phys. Rev. D **73**, 123523 (2006) [arXiv:astro-ph/0605003].
- [110] A. R. Liddle, P. Mukherjee and D. Parkinson, Astron. Geophys. **47**, 4.30-4.33 (2006) [arXiv:astro-ph/0608184].
- [111] A. R. Liddle, Mon. Not. Roy. Astron. Soc. Lett. **377**, L74 (2007) [arXiv:astro-ph/0701113].
- [112] C. Pahud, A. R. Liddle, P. Mukherjee and D. Parkinson, Mon. Not. Roy. Astron. Soc. **381**, 489 (2007) [arXiv:astro-ph/0701481].
- [113] E. V. Linder and R. Miquel, arXiv:astro-ph/0702542.
- [114] A. R. Liddle, P. S. Corasaniti, M. Kunz, P. Mukherjee, D. Parkinson and R. Trotta, arXiv:astro-ph/0703285.
- [115] G. Efstathiou, arXiv:0802.3185 [astro-ph].
- [116] R. Easther, AIP Conf. Proc. **698**, 64 (2004) [arXiv:astro-ph/0308160].
- [117] D. Baumann *et al.*, arXiv:0811.3919 [astro-ph].
- [118] W. H. Kinney, Phys. Rev. D **58**, 123506 (1998) [arXiv:astro-ph/9806259].
- [119] E. J. Copeland, I. J. Grivell and A. R. Liddle, Mon. Not. Roy. Astron. Soc. **298**, 1233 (1998) [arXiv:astro-ph/9712028].
- [120] L. P. L. Colombo, E. Pierpaoli and J. R. Pritchard, arXiv:0811.2622 [astro-ph].
- [121] P. Adshead and R. Easther, JCAP **0810**, 047 (2008) [arXiv:0802.3898 [astro-ph]].
- [122] V. Barger, Y. Gao, Y. Mao and D. Marfatia, arXiv:0810.3337 [astro-ph].
- [123] A. Kogut *et al.*, arXiv:0704.3991 [astro-ph].
- [124] J. Dunkley *et al.*, arXiv:0811.3915 [astro-ph].
- [125] J. M. Maldacena, JHEP **0305**, 013 (2003) [arXiv:astro-ph/0210603].
- [126] P. Creminelli, A. Nicolis, L. Senatore, M. Tegmark and M. Zaldarriaga, JCAP **0605**, 004 (2006) [arXiv:astro-ph/0509029].
- [127] M. Liguori, A. Yadav, F. K. Hansen, E. Komatsu, S. Matarrese and B. Wandelt, Phys. Rev. D **76**, 105016 (2007) [Erratum-ibid. D **77**, 029902 (2008)] [arXiv:0708.3786 [astro-ph]].
- [128] I. G. Moss and C. Xiong, JCAP **0704**, 007 (2007) [arXiv:astro-ph/0701302].
- [129] M. Sasaki, J. Valiviita and D. Wands, Phys. Rev. D **74**, 103003 (2006) [arXiv:astro-ph/0607627].
- [130] K. Moodley, M. Bucher, J. Dunkley, P. G. Ferreira and C. Skordis, [arXiv:astro-ph/0407304].
- [131] R. Bean, J. Dunkley and E. Pierpaoli, Phys. Rev. D **74**, 063503 (2006) [arXiv:astro-ph/0606685].
- [132] D. H. Lyth, C. Ungarelli and D. Wands, Phys. Rev. D **67**, 023503 (2003) [arXiv:astro-ph/0208055].
- [133] D. H. Lyth and D. Wands, Phys. Rev. D **68**, 103516 (2003) [arXiv:astro-ph/0306500].

- [134] J. M. Bardeen, *Phys. Rev. D* **22** (1980) 1882.
- [135] E. Komatsu, arXiv:astro-ph/0206039.
- [136] S. W. Hawking, *Astrophys. J.* **145**, 544 (1966).
- [137] G. F. R. Ellis and M. Bruni, *Phys. Rev. D* **40**, 1804 (1989).
- [138] A. R. Liddle and D. H. Lyth, *Phys. Rept.* **231**, 1 (1993) [arXiv:astro-ph/9303019].
- [139] M. Sasaki and E. D. Stewart, *Prog. Theor. Phys.* **95**, 71 (1996) [arXiv:astro-ph/9507001].
- [140] A. Challinor and A. Lasenby, *Astrophys. J.* **513**, 1 (1999) [arXiv:astro-ph/9804301].
- [141] H. Kodama and M. Sasaki, *Prog. Theor. Phys. Suppl.* **78**, 1 (1984).
- [142] V. F. Mukhanov, H. A. Feldman and R. H. Brandenberger, *Phys. Rept.* **215**, 203 (1992).