Syracuse University

SURFACE

Dissertations - ALL SURFACE

May 2015

Predictions and Constraints of Cosmological Correlators

Jayanth Tallakere Neelakanta
Syracuse University

Follow this and additional works at: https://surface.syr.edu/etd

 Part of the Physical Sciences and Mathematics Commons

Recommended Citation

Tallakere Neelakanta, Jayanth, "Predictions and Constraints of Cosmological Correlators" (2015). *Dissertations - ALL*. 241.
https://surface.syr.edu/etd/241

This Dissertation is brought to you for free and open access by the SURFACE at SURFACE. It has been accepted for inclusion in Dissertations - ALL by an authorized administrator of SURFACE. For more information, please contact surface@syr.edu.

Abstract

In this dissertation, we study the role of correlation functions in Cosmology. The study is bidirectional: We explore the constraints that correlation functions gathered from data impose on different theories; we also analyze the constraints that get imposed on correlation functions given symmetries of theories. For the former analysis, we use structure formation data like the CMB and matter power spectrum to set limits on the temperature of cold dark matter particles, basically only assuming that the particles were nonrelativistic when they decoupled and have interacted negligibly since. In another study, we use the same data to constrain how much Sommerfeld enhancement of dark matter annihilation could have occurred, with the analysis being insensitive to the details of the annihilation. Finally, we propose a new method to detect the so-called CMB anomalies in a more general manner than is usually considered. For the latter type of analysis, we consider the role of gauge symmetry in constraining relations between n- and $(n+1)$-point correlation functions for gravity coupled to a scalar field. Using certain assumptions, we show how novel consistency relations between fields can be derived, that arise only out of the symmetry of the action, and are independent of its particular form.

Predictions and Constraints of Cosmological Correlators

by

Jayanth T. Neelakanta

B.Sc. (Hons.) Chennai Mathematical Institute, Chennai, 2009

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Physics

Syracuse University

May 2015

Copyright 2015 Jayanth Neelakanta

All rights reserved

रोधोपघातसादृश्येभ्यो व्याभिचारादनुमानमप्रमाणम

Inference, some say, is not valid because there is irregularity due to embankment, damage and similarity—
Nyaya Sutra, Book II, Chapter 1.37

Acknowledgements

I am indebted to my advisor, Cristian Armendariz-Picon, for all that he has taught me and done for me during the course of my PhD. He has been tremendously patient, and has always encouraged me to think independently. I have never had to seek an appointment with him, and, having walked straight into his office, have more often than not asked him a question that had nothing to do with what we were working on! I don't recall a single instance of having left his office dissatisfied. The kind of research questions he asked and the manner in which he went about answering them have made a wonderful impression on me. Thanks so much, Cristian!

I would also like to thank Scott Watson for all the conversations we have had. He has been very supportive of me in all respects, and I have enjoyed several academic and non-academic conversations with him. John Laiho has taught me so much about an area of physics that I had no knowledge about at all when I began working with him. His perseverance and determination are qualities that I much admire, and hope to emulate.

Riccardo has played a big role in helping me understand several concepts and techniques. He has been very accommodating in terms of time even if I found a simple concept difficult to grasp! I have also found his non-technical advice about several things very useful, and for that I am very grateful.

I would like to thank Aarti and Rahul for all the help they rendered when I initially arrived in Syracuse. I have cherished many wonderful moments with them since, and they have been a constant source of support. I am also grateful to Pramod and Shiladitya for all the logistical help I benefited from before I even arrived in Syracuse.

Many thanks to Madhusudan Sir at the Planetarium in Bangalore for introducing me to the wonderful world of Physics, and several of the professors at CMI and IMSc for laying the Physics foundations that has got me this far.

Diane, Penny and Patty have been fantastic at making things extremely easy for

me at the administrative level. They are much the envy of several friends of mine from other Universities, who have not necessarily had it this easy!

My entry into Physics wouldn't have been possible without the support and encouragement of my parents and my brother. I owe them immensely for all that they did, which has led me this far in life.

Last, but definitely not the least, thanks Nivedita for teaching me so much about Life, the Universe and Everything.

List of Papers

Chapters 2, 3, 4 and 5 of the dissertation comprise of work carried out for the following papers, respectively:

1. *How Cold is Cold Dark Matter?*, C. Armendariz-Picon and J. T. Neelakanta, JCAP 1403, 049 (2014) [1]

2. *Structure Formation Constraints on Sommerfeld-Enhanced Dark Matter Annihilation*, C. Armendariz-Picon and J. T. Neelakanta, JCAP 1212, 009 (2012) [2]

3. *Detecting anomalies in CMB maps: a new method*, J. T. Neelakanta, arXiv:1501.03513 [3]

4. *General Covariance Constraints on Cosmological Correlators*, C. Armendariz-Picon, J. T. Neelakanta, and R. Penco, JCAP 1501, 035 (2015) [4]

Contents

List of conventions and symbols

Chapters 2, 3, 4 and 5 of the dissertation comprise of work carried out for the following papers, respectively:

1. *How Cold is Cold Dark Matter?*, C. Armendariz-Picon and J. T. Neelakanta, JCAP 1403, 049 (2014) [1]

2. *Structure Formation Constraints on Sommerfeld-Enhanced Dark Matter Annihilation*, C. Armendariz-Picon and J. T. Neelakanta, JCAP 1212, 009 (2012) [2]

3. *Detecting anomalies in CMB maps: a new method*, J. T. Neelakanta, arXiv:1501.03513 [3]

4. *General Covariance Constraints on Cosmological Correlators*, C. Armendariz-Picon, J. T. Neelakanta, and R. Penco, JCAP 1501, 035 (2015) [4]

Chapter 1

Introduction

Hydrogen is a light, odorless gas, which, given enough time, turns into people.

Edward Harrison

The question of the origin and the structure of the Universe must count as one of the oldest and most profound questions that mankind has asked. All ancient civilizations had their own theories about how the Universe was constituted. It was the invention of the telescope that made it possible to verify these theories. The telescope eventually led to the acceptance of Copernicus' heliocentric theory, which was radically at odds with what almost all ancient astronomers thought, and, indeed, what common sense suggested. Yet, it took almost three more centuries for the technology to reach a stage where one could conclude that galaxies apart from our own, so-called "island universes", existed. Moreover, it was discovered that most of these galaxies appear to be moving away from us. Interpreted within the context of General Relativity, this suggested that the universe was expanding. Some of the earliest evidence to support this view was gathered by Vesto Slipher in 1917 [5]. In 1929, Edwin Hubble [6] gathered additional evidence and also proposed an empirical relationship between the receding velocities v and distance d, $v = Hd$. Here, H is the Hubble constant and the relationship is known as Hubble's Law. (Georges Lemaître had actually derived

much of what Hubble did, with slightly less data, in 1927, but owing to the result being published [7] in a not-so-famous French journal, most of the credit continues to go to Hubble.)

Though Modern Cosmology hasn't settled the question of the origin of the Universe, it has made tremendous progress in pushing back the envelope of our understanding of its evolution by billions of years. Observational and theoretical advances over the past few decades have improved our theories regarding the Universe's composition to a remarkable degree. Interpreted within the context of Particle Physics, Cosmology still leaves unanswered several questions such as the particle nature of Dark Matter, the Cosmological Constant problem, etc. But, treated purely as a model that makes predictions given the values of a few parameters, the Concordance Model of Cosmology has been spectacularly successful in explaining the large-scale structure of the universe from the first few seconds up to today.

Arguably, the most important tool that has contributed to this improvement in our understanding is the Cosmic Microwave Background radiation (CMB). To date, Astronomy, and hence Cosmology, is predicated on using photons from outer space to better understand the constitution of the cosmos. Up until 1965, scientists were primarily using photons from celestial objects such as stars or clusters to understand what has been happening in the Universe in the relatively recent past; that is, during epochs at most since these objects had formed. There was no probe of how the Universe looked prior to that epoch, or indeed whether such an epoch existed at all. But, as we shall see, the discovery of the CMB confirms the idea that an epoch bereft of compact objects existed, and the CMB has been used to pin down the properties of the cosmos in impressive quantitative detail.

1.1 FLRW Metric

Two pillars on which much of Modern Cosmology stands are:

(i) The universe is almost exactly (spatially) homogeneous and isotropic, meaning that, on large enough scales, the universe must look the same wherever you are and in whatever direction you look. This is also called the Cosmological Principle. (This is the reason H in Hubble's Law is called Hubble's constant. Even though it changes as a function of time, it is the same everywhere in space.)

(ii) The universe is expanding. The corollary to this is that, on large enough scales, the distance between objects is increasing at a rate that is proportional to the distance between the objects. This is the same as Hubble's Law.

If the tiny deviations from homogeneity and isotropy are ignored, point (i) above implies that the line-element describing the universe is necessarily of the FLRW form:

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + a^2(t)\left(\frac{\mathrm{d}r^2}{1-kr^2} + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right)\right), \tag{1.1}$$

where $k = \{-1, 0, 1\}$ represents negative, zero and positive curvature of constant-time hypersurfaces. The metric corresponding to this line-element was discovered by Friedmann [8, 9] and Lemaître [7] independently in the context of General Relativity, and proved to be a purely geometric result by Robertson [10] and Walker [11]. In this framework, Hubble's constant $H = \dot{a}/a$ and the fact that the universe is expanding means that $H(t) > 0$, or equivalently that $a(t)$ is a monotonically increasing function of time.

In this chapter, we shall try to understand the different epochs in cosmic history in a pedagogical manner. But, because the motivations for several assumptions for a given epoch are due to what happens in other epochs, it will sometimes be difficult to follow a strictly pedagogical discussion. Hence, for reference, we have sketched the most important epochs in Figure 1.1. The meaning of these epochs should become clear in due course. For now, it must be noted that what differentiates these epochs is basically the kind of energy that dominates during the corresponding times. A scalar

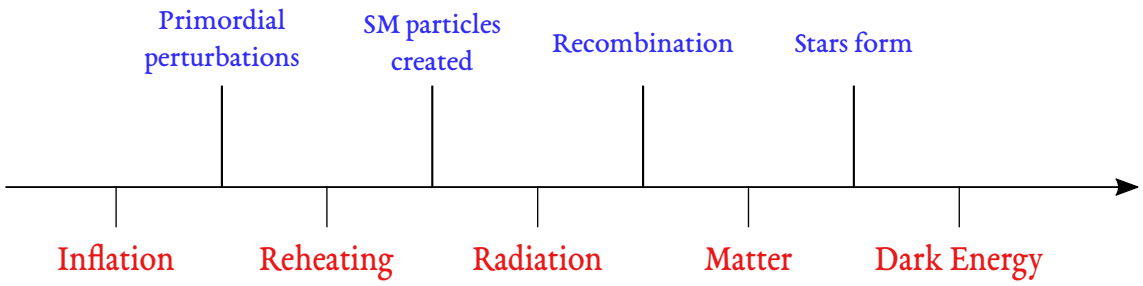


Figure 1.1: Text below, in red: The main epochs of the universe, with left to right indicating a chronological order. Text above, in blue: important events that occur during the epochs.

field dominates during inflation, photons and neutrinos during radiation-domination, dark matter and other massive particles during matter-domination and the cosmological constant dominates at the end. The interactions that are relevant during the era of reheating aren't very well-understood and hence not much is known about how long it lasts.

1.2 Hot Big Bang Model

Let us assume that the universe has been expanding for as long as it has existed. Then, the fact that the universe is made up of normal matter like photons implies, within the context of general relativity, that the energy density of the universe must increase as one goes back in time. In particular, this implies that one hits a spacetime singularity in the finite past. This isn't really an issue because quantum gravity effects are expected to dominate around these high energy densities and rescue the theory. But, it is valid to seek the implications of this increase in energy density before it becomes of the order where quantum gravity effects dominate.

For now, let us assume a universe made up only of electrons, protons and photons as they are stable particles whose properties are very well understood. Further, assume that the spectrum of the photons is that of a blackbody. We shall explain the motivation for the assumption in due course, but, the implication of a blackbody

spectrum is that, given just the temperature of the spectrum, the energy distribution function of the photons can be calculated. Now, as the universe expands, if interactions of the photons with matter are ignored, it can be shown that the temperature T_γ of this spectrum is inversely proportional to the scale factor $a(t)$ that appears in (1.1),

$$T_\gamma(t) = T_{\gamma,0}\frac{a_0}{a(t)}, \tag{1.2}$$

where $T_{\gamma,0}$ is the temperature of the photons today. So, as we go back in time, T_γ increases, which implies that the average energy of photons in the universe increases.

If photons are absent, electrons and protons combine to form Hydrogen atoms. Recall that the ionization energy of the Hydrogen atom is 13.6 eV. So, as we go back in time, we ought to reach an epoch where the average energy of the photons attains this value. Therefore, around this point of time (and during all times preceding it), neutral Hydrogen atoms cannot exist. Moreover, even though protons and photons don't interact very efficiently, Compton scattering between the electrons and the photons, and Thomson scattering between the electrons and the protons keep all three species of particles in thermal equilibrium. This is what is meant by the Hot Big Bang Model. The conclusion, if all our assumptions are true, is that the very early universe was made up of a hot plasma that was opaque due to the mean free path of the photons being very small. As the universe expanded and cooled, around the time when the average photon energy dropped to that of Hydrogen's ionization energy, the photons became free, and neutral Hydrogen formed. This phenomenon is called *recombination*.

We had started with the assumption that the universe is almost exactly homogeneous and isotropic. So, recombination must have happened everywhere in the universe. Moreover, if we assume that post recombination most photons travel to us without any interactions, as we look out in the sky, we should be able to see these photons. Say recombination occurred n billion years ago. Then, today we

should be receiving photons that have traveled n billion light-years since recombination.[1] Moreover, the photons must follow a blackbody distribution. It is one of the great triumphs of Modern Cosmology that, in 1965, this radiation was discovered by Penzias and Wilson [12]. Further investigations at different frequencies showed that the photons indeed followed a blackbody distribution, with a temperature of 2.725 K. This temperature corresponds to the microwave part of the electromagnetic spectrum and thus the radiation is called the Cosmic Microwave Background radiation. These discoveries established that the Hot Big Bang Model was the right framework to do Cosmology in.

1.3 Anisotropies

In our assumptions in Section 1.1, we said that the universe is *almost* homogeneous and isotropic. Empirically, the fact that the universe cannot be *exactly* homogeneous and isotropic is obviously true because we see structure all around us. It is also true that, at least classically, one cannot start with a system that has some spatial symmetry and, using General Relativity, break that symmetry. In other words, the current state of the universe implies that in all of its past, at least when it could be treated classically, homogeneity and isotropy were not exact. In particular, during recombination too, these symmetries must have been broken. So, the spacetime metric, even in the early universe, must have been a *perturbed* FLRW metric.

Recalling that General Relativity relates metric perturbations to density perturbations, it is natural to expect that this implies that the photon energy density wasn't exactly homogeneous either. The Stefan-Boltzmann law then implies that the temperature of the CMB photons should also depend on the region from which the photons originated. These inhomogeneities, when projected onto the sky, must correspond to

[1] Expansion of the space between where these photons initially became free and Earth's current location means that these points are further away from us today than n billion light-years, but, only by an order one number.

temperature anisotropies. In 1992, these anisotropies were seen by the COBE-DMR instrument [13], specifically designed for this purpose. (RELIKT, launched in 1983, had gathered slightly less accurate data.) The anisotropies were of the order of about 1 part in 10^5, justifying the assumption that the universe is *almost* homogeneous and isotropic.

Let $T_\gamma(\hat{n})$ denote the detected CMB temperature in direction $\hat{n}$ on the sky. Then,

$$T_\gamma(\hat{n}) = \overline{T}_\gamma + \Delta T_\gamma(\hat{n}), \tag{1.3}$$

where $\overline{T}_\gamma$ is the background CMB temperature today (2.725 K). $\Delta T_\gamma(\hat{n})$ is thus the temperature anisotropy. In terms of spherical harmonics $Y^{\ell m}(\hat{n})$, we can write

$$\frac{\Delta T_\gamma(\hat{n})}{\overline{T}_\gamma} = a_{\ell m} Y^{\ell m}(\hat{n}) \tag{1.4}$$

Thus, once a coordinate system is chosen, the multipole coefficients $a_{\ell m}$'s contain all the information about the CMB. By construction, the coefficients depend on the coordinate system chosen. On the other hand, consider the correlation function between the coefficients

$$\langle a^*_{\ell m}\, a_{\ell' m'} \rangle = C_\ell\, \delta_{ll'}\, \delta_{mm'} \tag{1.5}$$

The angular brackets in the above indicate an ensemble average over a statistically isotropic universe with random temperature fluctuations. We will motivate these properties later, but, for now, we would just like to state that it is these properties that lead to the Kronecker deltas on the right-hand side, rendering the correlation function coordinate-system independent. Therefore, if these properties are assumed, then, given temperature anisotropy data, the C_ℓ's can be estimated as

$$C_\ell = \frac{1}{2\ell + 1} \sum_{m=-\ell}^{m=\ell} |a_{\ell m}|^2 \tag{1.6}$$

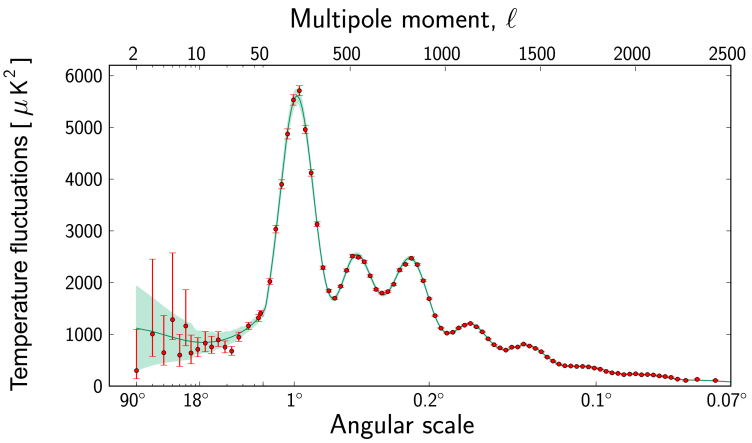


Figure 1.2: The CMB anisotropy power spectrum as measured by Planck [14], with D_ℓ being plotted along the y-axis. The red lines represent experimental error bars. The green shaded region represents cosmic variance, which is a theoretical error bar: from (1.6), we see that, for a given ℓ, there are $2\ell + 1$ modes to sample from, and so lower ℓ's will necessarily have a higher sampling variance.

Furthermore, if the fluctuations are distributed as Gaussians, as the simplest early-universe theories predict, and as the data seem to indicate, then, the C_ℓ's contain all the relevant information regarding the CMB. In Figure 1.2, we plot the observed

$$D_\ell \coloneqq \overline{T}_{\gamma,0}^2 \, \frac{\ell(\ell+1)}{2\pi} \, C_\ell$$

as a function of ℓ.

1.4 Dark Matter

As the anisotropies on the sky today are basically the evolution of perturbations from some earlier time, it is clear that their distribution depends on a number of factors like the amount of matter in the universe, the rate at which the universe is expanding, the geometry of the spatial hypersurfaces, etc. But, for a given set of these parameters, and for given initial conditions at some time before recombination, the distribution of the anisotropies as we see them on the sky today can be calculated. Conversely,

given the actual distribution of the anisotropies, the values of these parameters can be estimated.

If such an exercise is carried out, and only particles that we know from the Standard Model of Particle Physics are considered, it is seen that for no values of the parameters do the data fit the theory! So, we are led to conclude that we must either modify the theory or modify the particle content in the model. Given the success of general relativity, it is reasonable to begin by doing the latter via the addition of a new kind of particle. Indeed, as long ago as in 1933, Zwicky [15] suggested that the Coma Cluster had particles in it that had mass, but didn't interact with light. He proposed this as a solution to the observation that the mass of the galaxies that he estimated seemed to be much more than his calculations for the mass of the luminous objects in the cluster.

When we consider the distribution of the CMB anisotropies today, it is clear that matter plays two different roles—on the one hand, the phenomenon of recombination depends on the interaction between photons, electrons and protons; on the other, by having a mass, electrons and protons influence how the universe evolves, and hence how photons propagate. Physically, these are different effects—the former is more "chemical", whereas the latter is purely gravitational. Following Zwicky's suggestion, say we have two types of matter: normal matter, usually called baryons in cosmology, meaning all types of matter (heavier than neutrinos) that are part of the Standard Model; dark matter, meaning matter that interacts negligibly weakly with light. It is seen that the addition of dark matter results in a much better fit of the data with the model.

In the context of cosmological structure formation, dark matter (DM) can be modeled as a perfect fluid, just like baryons are. Thus, energy density and pressure completely characterize the dark matter fluid. As the metric perturbations are tiny[2],

[2]It should be noted that metric perturbations by themselves are *not* gauge-invariant, so their magnitude isn't a physical quantity. In subsequent chapters, we will deal with gauge-fixed metric perturbations that are physical. It is these perturbations that are being referred to in the text.

it again suggests that it is useful to decompose the DM energy density and pressure into background and linearly perturbed components.

The pressure of most fluids considered in cosmology is linearly related to the energy density, $p = w\,\rho$, where w is called the equation-of-state parameter. The role of pressure in cosmological structure formation is to suppress the growth of structure. For instance, consider the evolution of the perturbation of a pressureless fluid in two backgrounds: (i) dominated by radiation (equation of state $w = 1/3$); (ii) dominated by pressureless matter ($w = 0$). In (i), the growing mode of the perturbation evolves only logarithmically with the scale factor, whereas in (ii), it evolves linearly. This can directly be traced to the difference in w between the two background fluids.

If the particles of a fluid can be described by a phase-space distribution function, then the pressure of the fluid is related to the velocity dispersion of the particles. Further, if the particles are in kinetic equilibrium (meaning that they are described by one of Maxwell-Boltzmann, Fermi-Dirac or Bose-Einstein statistics), this velocity dispersion is related to the temperature of the system. To sum up, if particles are assumed to be in kinetic equilibrium, their pressure is related to their temperature. In particular, low temperatures correspond to low pressures. We shall assume kinetic equilibrium of dark matter particles for the rest of the chapter.

Hot and cold dark matter affect structure formation in very different ways. Fluctuations of a fluid cannot grow on scales smaller than the corresponding mean free paths of the particles simply because the random motions of these particles on those scales wash out the perturbations. Hot dark matter particles have larger velocity dispersions due to their higher temperature, and thus larger mean free paths. This implies that structure formation can only proceed in a "top-down" manner; that is, superclusters must form first, which then split away into clusters, and finally galaxies. Because, if one started with a perturbation on small (say galactic) scales, it would get suppressed with time. Whereas, in the case of cold dark matter, the structure formation proceeds in a "bottom-up" manner—structures form on the smallest scales

first, and then continue to grow into bigger and bigger-sized objects.

Evidence from the CMB and other probes prove beyond doubt that structure formation has happened in a bottom-up manner. So, the consensus in the cosmology community is that there must be a significant amount of Cold Dark Matter (CDM). In fact, data show that the energy density of CDM is four times that of baryons.

Apart from the properties of dark matter that we have discussed, little else is known about it. These properties are sufficient to explain phenomena as widely separated in scale as the anisotropies in the CMB, the clustering of galaxies and the rotation curves of stars in the Milky Way. These lines of evidence for their existence is one of the most concrete reasons to believe in physics beyond the Standard Model. In Chapter 2, we shall discuss in much greater detail what we mean by the coldness of dark matter. In Chapter 3, we shall consider one proposed form of interaction amongst CDM particles and discuss what can be said about the strength of this interaction.

1.5 Scales

We mentioned earlier that the CMB anisotropies can be used to constrain the values of parameters that are part of the Concordance Model of Cosmology. To achieve this, we obviously need to set up some initial conditions for the inhomogeneities. An ambitious plan would be to set the initial conditions at a time that corresponds to the highest energies that we have probed here on Earth. This could be, say, the TeV scale that the LHC [16] probes, or the PeV scale at which IceCube [17] has detected cosmic neutrinos. Call the time corresponding to this energy $t_{\rm init}$. In principle, to determine the distribution of the CMB anisotropies, one would have to keep track of all that happened in the universe since $t_{\rm init}$. But, due to the expansion of the universe, the task turns out to be much simpler.

To see this, let us first understand in more detail why the CMB has proved to be

so useful in understanding the universe's evolution during very early times. It has to do with the range of scales that the CMB probes. For the analysis of these scales, it turns out to be much more convenient to work in Fourier space because, in linear perturbation theory, the different modes that represent these scales decouple. This decoupling of scales is extremely crucial. For instance, measuring the temperature anisotropy on the largest scales *today* gives us a direct probe of the inhomogeneities in the gravitational potentials during the *first second* of the universe! In the absence of the decoupling of scales, such a simple relationship between the value of an observable today and its value in the very early universe wouldn't exist. It is for this reason that we shall concentrate on perturbations that can be treated linearly.

CMB and supernovae data [18, 19] indicate that the universe is spatially flat. This implies that any perturbed quantity $\delta f(\vec{x}, t) \coloneqq f(\vec{x}, t) - \bar{f}(t)$ can be written as

$$\delta f(\vec{x}, t) = \frac{1}{(2\pi)^{3/2}} \int \mathrm{d}^3 k \; e^{i\vec{k}.\vec{x}} \; \delta f(\vec{k}, t)$$

The decoupling of modes means that we could pick a mode $\delta f(\vec{k})$ and follow its evolution independently of other $\vec{k}$, at least until $\delta f(\vec{k})$ becomes non-linear. Note that large $\vec{k}$ corresponds to small scales and vice-versa.

Now, it is clear that the larger the number of the modes that we have access to, the more the information that we possess. There is an obvious lower limit to the k associated with any cosmological observation. It is related to the distance to the (imaginary) surface that produced the photons corresponding to the observation. For instance, one minute after recombination, we would only have access to the photons produced within a sphere of radius one light-minute. Today, we have access to photons from a much larger distance (hence, a much lower k). This k corresponds to the smallest value of k that the CMB probes. Call it k_{min}, and the wavelength corresponding to it λ_{max}: $\lambda_{\mathrm{max}} = \frac{2\pi}{k_{\mathrm{min}}}$. Because photons were not free before the CMB was produced, k_{min} is the smallest k value that we can *ever* hope to probe via photons.

It is for this reason that the CMB contains the best information about the very early universe.

What about the maximum value of k (smallest scale) that the CMB probes? In an era dominated by radiation, which is when recombination occurs (refer Figure 1.1), perturbations are prevented from collapsing on the smallest scales. This is because the pressure in the radiation counters the attractive nature of gravity. So, as perturbations are washed off on the smallest scales, there is no theoretical $k_{\rm max}$. But, in reality, due to the resolution and sensitivity of the detectors that measure the CMB temperature, the actual $k_{\rm max}$ probed is quite a bit smaller. Call the corresponding wavelength $\lambda_{\rm min}$. The evolution of $\lambda_{\rm min}$ and $\lambda_{\rm max}$ as a function of $a(t)$, illustrating the range of modes that the CMB power spectrum probes, is shown in Figure 1.3.

What about other scales? The Hubble constant H is a time scale that represents the rate at which the universe is expanding. Given that no signal can travel faster than the speed of light, the Hubble radius cH^{-1} can be thought of as a length scale that determines regions in space that are in causal contact. Loosely speaking, this scale determines the largest length scales that can get affected when a phenomenon occurs. For instance, the evolution of cosmological perturbations depends on the value of the Hubble radius at that point of time.[3]

The wavenumber k is just a label denoting the different modes today. At an earlier moment of time, the physical wavenumber corresponding to this label is $k_{\rm phys} = k/a$. The dimensionless ratio $k_{\rm phys} \times cH^{-1} = \frac{kc}{aH}$ is a very useful quantity for structure formation.[4] If $\frac{k}{aH} \gg 1$, it means that we are considering very small scales and the expansion of the universe can be ignored. The mode is said to be *sub-horizon*. If $\frac{k}{aH} \ll 1$, we are considering very large scales and the mode is said to be *super-horizon*.

It is useful to consider the evolution of H, which in turn yields the evolution of

[3]It must be noted that the dynamical quantity cH^{-1} is different from the kinematical quantity, the particle horizon, that determines the maximum comoving distance that a signal could have traveled since the Big Bang. In particular, the latter depends on the entire history since the Big Bang, whereas the former depends on the value of the Hubble constant at a given time.

[4]We shall set c, the speed of light, to one from now on.

the Hubble radius. Einstein's equations for the spatially-flat FLRW metric lead to two equations that are called the Friedmann equations,

$$H^2 = \frac{8\pi G}{3}\rho \tag{1.7a}$$

$$\dot{H} + H^2 = -\frac{4\pi G}{3}(\rho + 3p) \tag{1.7b}$$

If the universe contains a species that can be modeled as a non-interacting perfect fluid f, then the above equations imply that the energy density of the fluid evolves as

$$\rho_f \propto a^{-3(1+w)} \ , \tag{1.8}$$

where w is the equation-of-state parameter. From Equations 1.7a and 1.8, we see that in the particular cases of a radiation-dominated ($w = 1/3$) and a matter-dominated ($w = 0$) universe, $H \sim a^{-2}$ and $H \sim a^{-3/2}$ respectively. Thus, in both cases, and in general in cases where $(\rho + 3p) > 0$, the Hubble radius H^{-1} decreases faster than a as we go back in time.

Therefore, if the Universe is made up of energy sources such that $(\rho + 3p) > 0$, the Hubble radius decreases faster than $\lambda_{\rm phys} = \frac{2\pi}{k_{\rm phys}} = a\,\frac{2\pi}{k}$. This means that, if we go back far enough in time, the size of a causal patch (given by H^{-1}) is much smaller than the smallest length scales $\lambda_{\rm min}$ that we can probe. For times before this, any local phenomenon would not leave its imprints on the density perturbations of the CMB or any other probe of structure formation. This is shown in Figure 1.3. Thus, when we discuss the evolution of the CMB, we can ignore what happened during the earliest stages of radiation-domination.

We expect that at energies of order hundreds of MeV and higher events such as phase transitions occur and the particle content of the universe changes. But, because the interaction rates are high, all the matter in the universe is in the form of a plasma, and any phenomenon affects scales much smaller than the smallest that

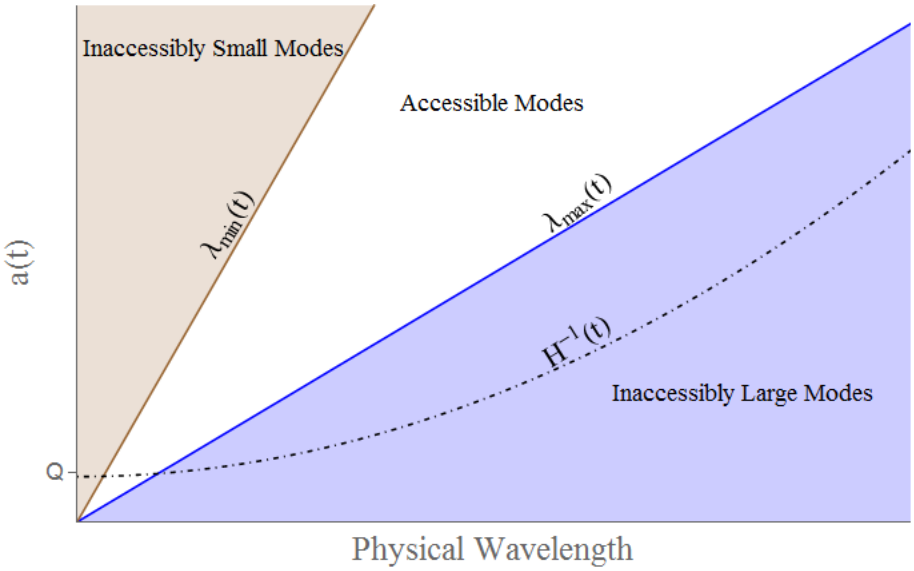


Figure 1.3: The evolution of modes in a radiation-dominated universe. The brown line represents the physical wavelength of the smallest scale that we can probe, and thus the brown region represents modes that we can't probe experimentally. The blue line represents the mode corresponding to the distance to the CMB surface and thus the blue region represents modes that are larger, and hence inaccessible. The black, dot-dashed line represents the quadratically growing Hubble radius and, early enough, like at the time corresponding to Q in the figure, $H^{-1}(t) < \lambda_{\rm min}(t)$ and hence phenomena that occur then don't leave an imprint on the CMB power spectrum, etc.

we can probe. Also, the fact that all the matter is in the form of a plasma justifies one of the assumptions that we started with—that the photons follow a blackbody spectrum. Very early in the universe's history, the interactions kept all species at the same temperature. As energy densities fell due to expansion, the interactions ceased to be effective and the species decoupled from each other. But, each species continued to be described by an equilibrium distribution, with a temperature that decreased with time.

1.6 Inflation

In light of the discussion in the previous section, it must bother us that the CMB temperature is homogeneous to 1 part in 10^5. Given that recombination occurs when

the average energy of the photons becomes of order 13.6 eV[5], we can calculate the energy density of radiation at this time. (This is a lower limit on the total energy density, but, the following argument only gets stronger if we include other contributions to the energy density.)

Given the energy density, from (1.7a) we can calculate the Hubble radius at the time of recombination. It corresponds to about a couple of degrees on the sky today. We discussed in the previous section that in the Hot Big Bang Model the Hubble radius decreases faster than a as we go back in time, whereas the physical size of a patch (the inverse of the physical wavenumber) decreases only as a. So, these patches of a couple of degrees could *never* have been in causal contact with each other. Yet, because the CMB temperature is almost exactly isotropic, we know that photons coming from even antipodal points on the sky have almost the same temperature! This feature, which can only be explained as an extremely unlikely coincidence within the Hot Big Bang Model, goes by the name of the *horizon problem*.

One way of avoiding this extreme fine-tuning is to see if the Hubble radius can be made to decrease *slower* than a as we go back in time. From (1.8), this is possible for $w < -1/3$. In this case, the universe is accelerating, meaning that $\frac{\ddot{a}}{a} > 0$. This rapid expansion of the universe is called inflation. An inflationary epoch could occur before the Universe was in the hot phase; that is, before it became dominated by normal Standard Model kind of particles. If this can be achieved, then, pre-hot phase, the causality arguments of the previous section don't apply. Thus, we have the possibility that the regions that constitute our observable universe were initially in causal contact and then fell out of causal contact as the universe entered the hot phase.

Of course, one needs a microscopic theory in which $w < -1/3$ can be achieved. It

[5]Actually, the value is about 0.25 eV. This has to do with the fact that the number density of photons is much larger than that of electrons. As the photons follow a blackbody distribution, the larger number density means that the contribution from the tail of this distribution cannot be neglected. Thus, recombination happens later than one would naïvely expect. Also, recombination is not an instantaneous process, but, can be approximated to be one for analytical considerations.

turns out that for a scalar field ϕ with potential $V(\phi)$,

$$w = \frac{\frac{1}{2}\dot{\phi}^2 - V(\phi)}{\frac{1}{2}\dot{\phi}^2 + V(\phi)} \tag{1.9}$$

If we have $\dot{\phi} \equiv 0$, then we achieve $w = -1$. Say the universe is filled with such a scalar field and that the field, called the inflaton, is homogeneous. From (1.8), we see that H is then a constant, and from (1.7), we have that the Universe is undergoing an exponential expansion, with a constant H. Therefore, this exponential expansion (called de Sitter inflation) is a possible way to get regions into causal contact *before* the universe enters the hot phase. Assuming that the scalar field is homogeneous may seem like trading the homogeneous properties of the CMB (which were the cause of the problem) into that of the scalar field. But, before inflation begins, the scalar field need only be homogeneous over the order of a few Hubble radii—not the 10^{28} that the Hot Big Bang Model requires.

The trouble with de Sitter inflation is that there is no dynamics. If $\dot{\phi}$ is exactly zero, then the field is stuck at its initial field value and absolutely nothing happens in the universe. So, we are led to considering a situation where $\dot{\phi}^2 \ll V(\phi)$. Using the equation of motion of the scalar field, this can be reduced to

$$\epsilon \coloneqq \frac{1}{16\pi G}\left(\frac{\partial V/\partial\phi}{V}\right)^2 \ll 1 \tag{1.10}$$

So, if we have a scalar field whose potential is almost flat, then the potential energy of the scalar field acts as a vacuum energy that leads to inflation. The scalar field rolls down slowly, thereby imparting dynamics to the universe. The fact that BBN (Big Bang Nucleosynthesis [20]), which occurs around the MeV scale, is so well predicted by the Hot Big Bang Model means that inflation must have definitely ended by the time of BBN, and in all likelihood, much earlier.

Figure 1.4 depicts the evolution of two different modes as a function of the scale factor during and after inflation. As shown in the figure, the Hubble radius remains

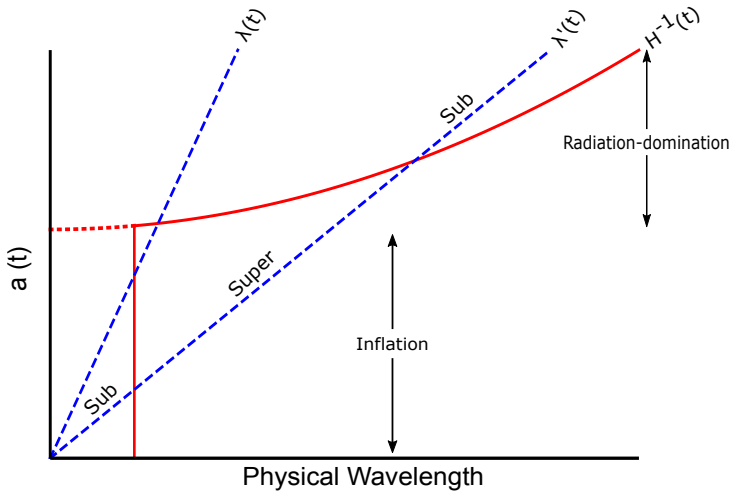


Figure 1.4: The evolution of modes during and after inflation. The blue dotted lines represent the physical wavelength. The bold, red line represents the Hubble radius, which remains constant during inflation and grows quadratically during radiation-domination. The dotted, red line represents the Hubble radius sans inflation. "Sub" and "super" refer to $\lambda'(t)$ being sub- and super-horizon respectively.

constant during inflation and then increases with time. The two dotted lines are the physical wavelengths of two modes, with $\lambda'(t) > \lambda(t)$. Firstly, note that if there was no inflation, the Hubble radius curve would have been extrapolated along the radiation-domination curve to a point on the y-axis.[6] This would imply that, at a given point of time (that is, for a given $a(t)$), the wavelength of both the modes would always be more than that of the Hubble radius. This is a pictorial depiction of the horizon problem. Also, note that the larger wavelength mode $\lambda'(t)$ exits the horizon (becomes super-horizon) earlier and re-enters it (becomes sub-horizon) later.

Data suggest that the universe would have had to have undergone at least 60 e-folds of inflation if the horizon problem is to be solved. That is, slow-roll must have lasted for at least as long as the time it took for the universe's radius to grow by a factor of e^{60}. Only then can modes with $k > k_{\min}$ have been part of a causal patch and thus lead to the CMB having almost the same temperature in all directions. The high energy density at the time, coupled with the fact that the expansion is exponential,

[6] As mentioned in the first section, quantum gravity effects are expected to dominate during the earliest times. So, the plot cannot be trusted for the earliest of times.

means that 60 e-folds can be achieved in a tiny fraction of a second. Nothing can be said about the modes that became superhorizon much before the 60 e-folds because these modes are currently outside our observable universe.

1.7 Reheating

We are faced with a different issue if we consider the exponential expansion due to inflation. During inflation, the energy density of the universe would have been dominated by the scalar field. During slow-roll, the energy density of the scalar field is almost a constant; whereas, that of normal matter gets exponentially diluted. So, if the universe expanded by a factor of least e^{60}, any matter that could have existed at those energies would have got diluted away. Recall that it is this very matter that eventually leads to the electrons, protons and photons that produce the CMB. Therefore, we need to somehow be able to convert the energy in the scalar field to that of normal matter. As inflation ends when the field starts rolling fast and reaches the minimum of the potential, it may well be that it is this kinetic energy of the scalar field that is responsible for the matter that we see around us.

We shall not discuss this creation of matter (called reheating) any further. But, simply from the fact that the energy density of matter is slightly inhomogeneous, we are led to conclude that the energy density of the scalar field must also have been very slightly inhomogeneous. (One may wonder if these perturbations could have been created after the end of inflation—at the time of reheating, say. But then, we basically encounter the horizon problem again. If the perturbations were produced after the end of inflation, these perturbations wouldn't be *coherent* on scales larger than the Hubble radius at that time—they would be completely stochastic. In particular, this would result in the CMB anisotropy not having the peaks that we observe it to.)

The problem has now been "reduced" to setting up the initial conditions for the scalar field and relating the perturbations in the scalar field to that of normal matter.

Let's deal with the latter first. A dimensionless quantity that measures the magnitude of perturbations in a fluid α is

$$\delta_\alpha(\vec{x}, t) \coloneqq \frac{\delta\rho_\alpha(\vec{x}, t)}{\bar{\rho}_\alpha(t) + \bar{p}_\alpha(t)}$$

The simplest initial condition would be to set the same δ_α for all fluids. If the total number of particles of each fluid is conserved (that is, there is no decay, annihilation, etc.), such perturbations are called adiabatic. Adiabatic perturbations have a particular signature on the CMB—they determine the location of the peaks (refer Figure 1.2). From CMB data, we can conclude that the perturbations that led to the CMB were very close to being adiabatic.

Under very general conditions, Weinberg [21] has shown that in the case of single-field inflation, the perturbations in the scalar field at the end of inflation lead to adiabatic perturbations in the plasma once the universe becomes radiation-dominated. The observed adiabaticity of the CMB anisotropies, as well as other observations such as the absence of non-Gaussianities, very strongly indicates that inflation was largely driven by a single scalar field. In this case, using Weinberg's theorem, *irrespective* of what happened between the end of inflation and the beginning of radiation-domination, the spectrum of perturbations at the end of inflation can be mapped into that of the different species at the beginning of radiation-domination.

1.8 Inflationary Perturbations

So, remarkably, all that we need for setting up the initial conditions of the fluid perturbations are the perturbations in the scalar field at the end of inflation. Chibisov and Mukhanov discovered [22] that the quantization of the scalar field itself could lead to the kind of perturbations that were being sought. As the energy scales of inflation are typically expected to be much higher than 10^{10} GeV (certainly much higher than the MeV scale of Big Bang Nucleosynthesis), it is natural to see if the high energy

densities lead to nontrivial quantum effects.

Let us start by counting the number of degrees of freedom (DOFs) in the theory. Consider the metric perturbation $\delta g_{\mu\nu}$. It is useful to classify the perturbations according to their transformations under the symmetries of the background FLRW spacetime—in particular, (spatial) rotations. Under rotations, δg_{00} transforms like a scalar. Hence, we have 1 scalar DOF. δg_{0i} transforms like a rank-1 (Cartesian) tensor, that is, $\delta g_{0i} \to \mathcal{R}_i{}^j \delta g_{0j}$, where $\mathcal{R}_i{}^j$ is the rotation matrix. The irreducible representation of such a tensor can be thought of as corresponding to that of the spherical harmonics $Y^{\ell m}$, with $\ell = 1$[7]. This corresponds to 1 scalar DOF of freedom and 2 vector DOFs (see footnote). Finally, the symmetric δg_{ij} part has as its irreducible decomposition a trace component (1 scalar DOF) and an irreducible rank-2 tensor (1 scalar DOF, 2 vector DOFs and 2 tensor DOFs). Hence, in total, in $\delta g_{\mu\nu}$, there are 4 scalar DOFs, 4 vector DOFs and 2 tensor DOFs. This decomposition is consistent with the fact that a symmetric 4×4 martix has 10 independent components. A scalar field (here, the inflaton ϕ) has 1 DOF.

The above discussion regarding the DOFs of the metric relied only on there being a metric and on the background symmetries of the metric. General relativity as a theory has additional symmetries—the predictions of the theory do not change under coordinate transformations. This property, called general covariance, can be thought of as a gauge symmetry akin to the gauge symmetries in Particle Physics. The gauge symmetry can be used to fix the values of 4 of the 11 components in the above, as a coordinate transformation is captured by a 4-vector in 4 dimensions. Further, general relativity is an example of a constrained system, in that the metric components that occur in its spatial part, that is $g_{ij}(\vec{x}, t)$, uniquely determine $g_{00}(\vec{x}, t)$ and $g_{0i}(\vec{x}, t)$. This reduces the number of independent degrees of freedom by 4 more. Thus, we are left with 3 DOFs—a scalar perturbation with one DOF; and the tensor perturbation

[7] In general, an irreducible rank-k tensor under rotations, also called a Spherical Tensor, can be thought of as being made up of spherical harmonics $Y^{\ell m}$, with $\ell = k, m \in [-\ell, \ell]$. In cosmological perturbation theory, when one refers to a scalar, a vector, or a tensor mode, one is usually referring to $m = 0, \pm 1, \pm 2$ respectively.

(in the cosmological sense), with 2 DOFs. Finally, gauge symmetry implies that for the 1 scalar DOF, we can either choose the metric perturbation ζ or the inflaton perturbation $\delta\phi$, and set the other to zero. We choose to work with ζ because it is for this quantity that Weinberg's theorem relates perturbations at the end of inflation to the beginning of radiation-domination. Recall that we are working in a flat-FLRW universe; so, constant-time slices of the universe must correspond to Euclidean space. ζ basically captures the departure of the constant-time slices of the real (perturbed) universe from a Euclidean space. The perturbed line-element, ignoring tensor modes, is

$$ds^2 = -dt^2 + e^{2\zeta(\vec{x},t)}a^2(t)dx^i dx_i$$

Because these are quantum fluctuations, the expectation value of ζ is zero. That is, if one averaged $\zeta(\vec{x},t)$ over all of space, one ought to get zero. But, the two-point correlation function $\langle\zeta(\vec{x},t)\,\zeta(\vec{x'},t)\rangle$ between fluctuations at different points in space is, in general, non-zero. Hence, the Fourier transform of the two-point correlation function, which is called the *power spectrum*, is non-zero as well. On dimensional grounds, we would expect the dimensionless power spectrum of the scalar metric perturbations to go as GH^2, as this quantity is dimensionless too. Up to order one numbers, it turns out that the scalar power spectrum actually goes as GH^2/ϵ, where ϵ was defined in (1.10).

What does it mean to say that the scalar power spectrum goes as GH^2/ϵ? Recall that during inflation modes exit the horizon (the quantity $\frac{k}{aH}$ decreases from greater than one to less than one). So, the power in each mode is given by GH^2/ϵ, where H^2 and ϵ are to be evaluated at the moment the mode exits the horizon, $k = aH$. Most inflationary models have a nearly-constant ϵ, because otherwise it is difficult to stay in the slow-roll regime. We shall henceforth assume that ϵ is constant, and by slow-roll, that it is much less than one. In that case, because from (1.7a) and (1.8), H is a decreasing function of time and because $w \geq -1$, it *must* be that the modes that exit later have less power. As the evolution of H is $O(\epsilon)$, this departure

from an equal power on all scales must be extremely small. In other words, inflation predicts an almost scale-invariant spectrum. Moreover, because the scales that leave earlier correspond to smaller k, inflation predicts that the spectrum has a red tilt; that is, if $k_2 > k_1$, then, $P(k_2) < P(k_1)$. The scalar power spectrum is thus usually parametrized as

$$P_\zeta(k) = A_s \left(\frac{k}{k_0}\right)^{n_s - 1}, \tag{1.11}$$

where k_0 is an arbitrary pivot scale and s stands for scalars. Canonical slow-roll inflationary models predict that n_s is close to and less than one. A_s corresponds to the amplitude of the power spectrum and is an observationally determined quantity. CMB data have determined n_s to be around 0.96 with extremely high significance, and this remains one of the biggest successes of inflation. Also, A_s has been determined to be close to 2×10^{-9}, the low value further justifying the usage of linear perturbation theory, as A_s is of order the square of the metric perturbation.

1.9 Consistency Relations

Throughout this chapter, we have discussed inflation being driven by a single scalar field. Arguably, this is the simplest way to realize inflation. But, from an effective field theory perspective, there are arguments that it is probably not the most natural mechanism to realize inflation. Moreover, theories of quantum gravity like String Theory postulate the existence of several scalar fields that should be relevant during the era of inflation. So, it would be nice to be able to distinguish inflation that occurs with a single field from inflation that occurs with multiple fields, because it would teach us something about the relevant degrees of freedom at high energies.

If the only observable we have is the power spectrum of the metric perturbations, then it is impossible to make this distinction. This is because the power spectrum basically constrains A_s and n_s, which can easily be matched with both single- and

multi-field inflation. Instead, if we consider the three point correlation function of the metric perturbations (called the bispectrum), then single-field inflation implies a relationship between the bispectrum and the power spectrum in the limit that one of the Fourier momenta goes to zero. Such relationships between n- and $(n+1)$-point correlation functions in the context of inflation go by the name of consistency relations. These consistency relations are quite model-independent, in that they don't depend on the interactions of the scalar field—rather, they just depend on there being only one scalar field. For instance, one of the the original consistency relations derived by Maldacena [23] is, up to order one factors and momentum-conserving delta functions,

$$\lim_{k_1 \to 0} \langle \zeta_{\vec{k_1}} \zeta_{\vec{k_2}} \zeta_{\vec{k_3}} \rangle = (n_s - 1)\, P_\zeta(k_1) P_\zeta(k_2) \tag{1.12}$$

As n_s is very close to one, this indicates that the level of non-Gaussianity must be quite low.

Usually, the non-Gaussianity parameter $f_{\rm NL}$ is given by

$$f_{{\rm NL},\zeta}(k_1, k_2, k_3) = \frac{\langle \zeta_{k_1} \zeta_{k_2} \zeta_{k_3} \rangle}{2[P_\zeta(k_1)P_\zeta(k_2) + P_\zeta(k_2)P_\zeta(k_3) + P_\zeta(k_3)P_\zeta(k_1)]}\,, \tag{1.13}$$

where P stands for the power-spectrum. So, in general, $f_{\rm NL}$ depends on the momenta of the modes involved. But, in many different classes of inflation, $f_{\rm NL}$ is only relevant for particular combinations of momenta. For instance, some models lead to a non-Gaussianity in the squeezed limit, where $k_1 \ll k_2, k_3$; some others in the equilateral limit $k_1 = k_2 = k_3$, etc. Even though these $f_{\rm NL}$'s have not been measured exactly, it is clear from the data that the level of non-Gaussianity is quite low. All the relevant $f_{\rm NL}$'s seem to be in the range 0 ± 100. From (1.13), we see that the square of the power spectrum occurs in the denominator. This squared quantity is of the same order as the value of the A_s parameter that we discussed earlier, 10^{-9}. Thus, the three-point function is a few orders of magnitude less than the two-point function,

and the data can be said to be highly Gaussian.

Recall that single-field inflation predicts a low level of non-Gaussianity. But, there also exist multi-field inflationary models that have a low level of non-Gaussianity. So, merely from the observation that the CMB anisotropies are quite Gaussian, we cannot really conclude that single-field inflation is much more favored than multi-field inflation. However, it would be extremely unnatural for multi-field inflation to satisfy the same consistency relations as single-field inflation. So, if we could measure the level of non-Gaussianity to a high-enough precision, we can actually check if consistency conditions like (1.12) are satisfied. This knowledge would lead to a much better understanding of the exact mechanism by which the universe inflated.

Maldacena derived the consistency conditions assuming slow-roll ($\epsilon \ll 1$). But, people have since generalized the result, and have derived other consistency conditions using different approaches. One of the more popular approaches is that of using symmetry—either the symmetries of the spacetime, or of the gauge theory that forms the framework of the inflationary model. In Chapter 5, we shall adopt the latter approach, and explore what the gauge symmetries of general relativity imply for the relationship between different correlation functions.

1.10 The ΛCDM Model

Somewhat belatedly, we come to the Standard Model of Cosmology. The main ingredients of the model are laid out in Figure 1.1. We have actually already discussed all but the final two epochs in the figure, the matter-domination and the dark energy epochs. The former is basically when the dominant energy density contribution is that of dark matter. As the universe expands, because the volume is increasing, the number density of particles decreases. Thus, the energy density also decreases. In addition to this, relativistic particles like the photons also lose energy due to cosmic redshift. This is why at late times one would expect non-relativistic particles like

dark matter to dominate the universe's evolution, as is borne out by Figure 1.1.

Dark energy is basically the energy density of the vacuum itself. If this vacuum energy arises due to the cosmological constant Λ of general relativity being nonzero, the energy density is a constant in space and time. Thus, a larger volume of space will have a larger amount of dark energy. So, as the universe expands, during very late times one would expect dark energy to dominate over all other forms of energy. This is again borne out by Figure 1.1.

We have discussed in earlier sections how the CMB probes the very early universe. But, the CMB photons must travel to us from the time of recombination to today. Thus, they also carry information about what has happened to the universe since recombination—in particular, during dark matter and dark energy domination.

In order to study the ΛCDM model quantitatively, one needs to choose some values for the relevant parameters. The most minimal set of these parameters contains the energy densities of baryons, dark matter, and dark energy; the value of the Hubble constant; A_s and n_s as defined earlier; and a couple of astrophysical parameters. The CMB power spectrum determines, or at least constrains, several of these parameters. When the CMB data are combined with other cosmological data, the values of these parameters get even more constrained. Thus, the Standard Model of Cosmology is also called the Concordance Model. In the next couple of chapters, we shall see how one can extend the Concordance Model to include new physics. We shall also use the CMB and other cosmological data to constrain the parameters that capture the new physics.

1.11 CMB Statistics

In (1.4), we decomposed the temperature anisotropies $\Delta T(\hat{n})$ into spherical harmonics $Y^{\ell m}(\hat{n})$ with multipole coefficients $a_{\ell m}$. We mentioned that we would treat the temperature perturbations as Gaussian distributed, with the underlying distribution

being statistically isotropic. This meant that the CMB power spectrum contained all the physical information about the CMB anisotropies. In the previous sections, we have seen that single-field inflation leads to predominantly Gaussian fluctuations in the metric perturbations. From the theorem of Weinberg that we discussed, if the perturbations are adiabatic, these fluctuations translates into predominantly Gaussian fluctuations in the CMB anisotropies. Moreover, as inflation is driven by a scalar field that is (classically) homogeneous in space, the perturbations that arise when the field is quantized will be statistically isotropic. We have thus justified the assumptions behind (1.5), as we promised.

The arguments discussed above are essentially theoretical. One needs to check if the data are actually consistent with these hypotheses. Consider the case of statistical isotropy. The index m in $a_{\ell m}$ captures the directional dependence of the quantity being decomposed, whereas ℓ captures the dependence on scale. Statistical isotropy means that the data shouldn't exhibit any non-trivial dependence on m. This is what (1.5) shows—the ensemble average of the correlation function between different directions must vanish. Assume that, for a given ℓ, the correlation function between the $a_{\ell m}$'s for two different m's is non-zero (in a statistically significant sense). Then, these two m's could be used to define a direction on the sky, thus violating statistical isotropy. An even simpler example would be that of the value of $a_{\ell m}$ being non-zero (again, in a statistically significant sense) for a particular combination of ℓ and m. This too would pick out a direction in the sky, and is thus an example of the violation of statistical isotropy.

Almost as soon as the CMB anisotropy data were available, different research groups started looking at whether the data were consistent with the hypothesis that the $a_{\ell m}$'s are distributed as independent, zero-mean Gaussians with an m-independent variance. We shall call this hypothesis the *null hypothesis*. On large scales (that, is, for small ℓ), the groups found that for some of the features they considered, the data seemed to be inconsistent with the null hypothesis at about the (3 - 3.5)σ level.

That is, the probability of the features occurring, given that they were described by the null hypothesis, was about 1 in 1000. Such features were termed CMB anomalies. For instance, one feature considered was the axis $\hat{n}$ around which the quantity $\sum_{m=-\ell}^{m=\ell} m^2 |a_{\ell m}(\hat{n})|^2$ was maximized, for different ℓ's.[8] It was found that the axes for $\ell = 2$ and $\ell = 3$ were almost exactly aligned, even though in principle they could have pointed anywhere on the two-dimensional sky. A few other such statistics were discussed.

By itself, the number 1 in 1000 isn't statistically significant. This is especially so given that the groups were reporting inconsistencies only with features that seemed maximally inconsistent with the null hypothesis. In addition to this unavoidable bias, cosmic variance, the theoretical error that we described in the text below Figure 1.2, is most pronounced for small ℓ, the very scales the groups were looking at. Finally, it is for these scales that the systematics in the CMB experiments are least well-understood. On these scales, emission of radiation from our own galaxy contaminates the CMB signal that the experiments try to measure.

Still, successive experiments, employing different systematics, have all seen roughly the same kind of features on the sky. Other groups have reported even more anomalies. Also, there is no obvious relationship between the different features, which makes it more difficult to account for them as arising out of a common fluke.

In fact, even today, there is no consensus on whether the anomalies are merely statistical features, or something more physical. Indeed, WMAP and Planck, the two largest collaborations in CMB experiments with hundreds of CMB experts, differ on the interpretation of the anomalies. In Chapter 4, we shall discuss a new method that could shed some light on whether the features are statistical or physical.

[8] Recall that the value of $a_{\ell m}$ depends on the coordinate system chosen; in particular, it depends on the axis $\hat{n}$ that acts as the conventional z-axis.

Chapter 2

The Coldness of Cold Dark Matter

2.1 Introduction

As we discussed in Section 1.4, a wide array of observations, ranging from the distribution of matter on cosmological distances, to the rotation curves of galaxies on kiloparsec scales, suggest that the universe contains a form of matter that does not interact with electromagnetic radiation and whose pressure is negligible. At present, the nature of this dark matter is unknown, but, among other hints, these phenomenological properties strongly suggest that dark matter is made of non-relativistic particles that couple very weakly to the standard model. We briefly discussed the circumstances under which the phenomenological properties lead to the latter conclusion.

There is certainly no shortage of particle dark matter models accommodating these properties, such as axions, moduli, gravitinos, Kaluza-Klein excitations, sterile neutrinos or WIMPs, just to name a few [24]. In each model, dark matter experiences a different cosmic evolution, resulting in a distribution of dark matter momenta that is often thermal at late times, although with temperatures that span many orders of magnitude in the different scenarios. Given the great variety of dark matter models and associated thermal histories, it is thus natural to ask whether we can place phenomenological limits on the dark matter temperature today, or whether in

fact there is evidence that dark matter has a non-zero temperature. Indeed, several authors have suggested that warm dark matter (see below) may help alleviate the apparent tension between the predictions of the Cold-Dark-Matter (CDM) scenario and the actual amount of clustering on sub galactic scales [25–30], although recent studies suggest that warm dark matter is disfavored by observations [31]. Similarly, the same small-structure problems may also be avoided if cold dark matter kinetically decouples rather late in cosmic history, as described in [32] and references therein. Phenomenological bottom-up limits on CDM like the one we discuss would not only constrain many of the different cold dark matter models, but also offer us a generic model-independent way to further characterize the properties of dark matter.

In this Chapter, we explore model-independent limits on the dark matter temperature to mass ratio extrapolated to the present time, T_0/m. This ratio determines the velocity dispersion of dark matter particles, which is a parameter that directly controls the growth of structure. If dark mater decouples while non-relativistic, dark matter particles travel at a root-mean-square velocity $v_{\rm rms} = \sqrt{3T/m}$, where m is the dark matter mass. As a result, anisotropies on scales much smaller than the associated free-streaming length are strongly suppressed, a phenomenon usually known as Landau damping. The absence of such suppression on observable scales thus places limits on the dark matter temperature to mass ratio. Whereas it turns out to be convenient to frame our limits in terms of the dark matter temperature, they can be equally interpreted as limits on the root-mean-square dark matter velocity. In this context, we should also point out that our limits apply even if dark matter does not consist of elementary point particles, but, instead, is made of objects of even macroscopic size, provided that their velocity dispersion is Maxwellian.

In order to reliably calculate the impact of a non-zero dark matter temperature on the formation of structure, we restrict ourselves to linear perturbation theory, and thus focus on cosmological probes applicable in this regime: the CMB and the matter power spectrum on the appropriate scales. The same suppression of structure implied

by a non-zero temperature also impacts the smallest scales that become non-linear, and, thus, the mass of the smallest proto-halos. Several authors have studied how the latter depend on the mass parameters of specific dark matter models, such as WIMPs in supersymmetric or extra-dimensional models, but in these cases the corresponding scales are too small to be probed observationally [33–35]. Interactions between dark matter particles and the thermal bath in the early universe may also imprint features on the matter power spectrum on sub-horizon scales at the time of kinetic decoupling [36, 37]. These interactions may lead to a suppression of small scale structure that is stronger than that due to free streaming on those scales, although they do not have any impact on modes outside the horizon at that time. Because the data used in our analysis only probe modes that entered the horizon at $z < z_{\rm max} \approx 5 \cdot 10^5$, these features are absent in the modes of interest if kinetic decoupling occurred at $z_{\rm dec} > z_{\rm max}$, an assumption typically satisfied in most dark matter models. In fact, a recent analysis of cosmological data does not find any evidence for features due to dark matter interactions [38], further suggesting that cosmological scales must have been outside the horizon at the time of kinetic decoupling. In that case, and in the context of our analysis, we can simply assume that cold dark matter is effectively collisionless. Signatures of dark matter interactions are then buried in small scales, beyond the reach of our cosmological data.

The ratio T_0/m, and any quantity derived from it, should be carefully interpreted. By the former we mean the temperature to mass ratio dark matter would have today in the absence of structure formation. In the real universe, however, dark matter inhomogeneities grow, become non-linear and collapse, resulting in virialized dark matter haloes whose temperature is determined just by the properties of the halo. In pure cold dark matter models the Press-Schechter mass function predicts that all of the dark matter ends up in such halos [39, 40]. If the dark matter temperature is non-zero there is a cut-off in the matter power spectrum at small scales, implying that only a fraction of dark matter collapses, but we expect this fraction to be significant

for small enough temperatures. Hence, our ratio T_0/m is not the temperature of a typical dark matter particle in today's universe, but just an extrapolation of what that ratio would be had dark matter not collapsed.

The work in the literature closest to the limits we discuss here has mainly focused on constraints on the mass of warm dark matter particles [31, 41–44]. In these models, dark matter is typically assumed to be hot, in the sense that it decouples kinetically while being relativistic, whereas in our work we assume that dark matter is cold, and thus decouples while non-relativistic. Because the parameter that actually affects the formation of dark matter structure is the velocity dispersion, in order to relate the latter to the dark matter mass, previous analyses have typically needed to introduce additional assumptions tied to the particular dark matter model being considered, such as the chemical potential of dark matter, its thermal history, or its number of degrees of freedom. As a consequence, mass limits only apply strictly in the context of the models in which they were derived, and cannot be readily extended to other scenarios. The present work focuses instead on the dimensionless ratio of temperature to dark matter mass directly, and basically relies on just two assumptions: that the distribution of dark matter momenta is Maxwellian after the time the smallest relevant scales entered the horizon, and that dark matter has been collisionless at least since that time.

2.2 Formalism

We assume that dark matter consists of collisionless particles, which for simplicity and without loss of generality we take to be spinless. As we pointed out in the introduction, these particles do not have to be point-like: As far as our analysis is concerned, the only restriction is that their size be much smaller than any other length scale in the problem.

Under these conditions, dark matter is then characterized by its distribution func-

tion f, which counts the number density of particles at coordinate time τ, comoving coordinate $\vec{x}$ and covariant momentum $\vec{p}$. Since we assume these particles to be collisionless, their distribution function $f(\tau, x^i, p_j)$ obeys the collisionless Boltzmann equation

$$\frac{p^0}{m}\left[\frac{\partial f}{\partial \tau} + \frac{\partial f}{\partial x^i}\frac{dx^i}{d\tau} + \frac{\partial f}{\partial p_i}\frac{dp_i}{d\tau}\right] = 0, \tag{2.1}$$

in which the zero on the right hand side accounts for the absence of non-gravitational interactions.

Because we assume that dark matter is non-interacting, its only observable effects involve gravitation. The energy momentum tensor of a distribution of dark matter particles characterized by f is

$$T_{\mu\nu} = \frac{1}{\sqrt{-g}}\int d^3p\,\frac{p_\mu p_\nu}{p^0}f, \tag{2.2}$$

where g is the determinant of the space-time metric.

2.2.1 Background Distribution

In a spatially flat FLRW universe with line-element

$$ds^2 = a^2(\tau)\left[-d\tau^2 + d\vec{x}^2\right], \tag{2.3}$$

the Boltzmann equation (2.1) reads

$$\frac{\partial f}{\partial \tau} = 0. \tag{2.4}$$

Hence, any homogeneous and isotropic distribution $f = f(p)$, where

$$p \equiv a\sqrt{g^{ij}p_i p_j}, \tag{2.5}$$

is a solution of the Boltzmann equation (2.4). Because we want to describe cold dark matter, we assume that the distribution function describes a gas of non-relativistic particles, and therefore choose it to be the Maxwell-Boltzmann distribution

$$f(p) = \frac{1}{(2\pi)^3} \exp\left(-\frac{m-\mu_0}{T_0} - \frac{p^2}{2mT_0}\right), \tag{2.6}$$

where m is the mass of the dark matter particles, T_0 is the temperature of dark matter today, and μ_0 is the chemical potential today, at $a_0 \equiv 1$. The temperature T_0 determines the mean kinetic energy of the dark matter particles—and hence their root mean square velocity $v_{\rm rms} = \sqrt{3T_0/m}$—and the chemical potential determines the number density of dark matter particles at present. If dark matter has $g_{\rm dm}$ degrees of freedom, the right hand side of equation (2.2) should be multiplied by $g_{\rm dm}$. Since this amounts to a change in the chemical potential μ_0, which in our approach is a free parameter anyway, we can set $g_{\rm dm} = 1$ without loss of generality.

Because in many models dark matter is in thermal equilibrium in the early universe, the distribution (2.6) is fairly generic. If $u^\mu = \delta^\mu_0/a$ is the four-velocity of a comoving observer, then $E = p_\mu u^\mu = \sqrt{m^2 + p^2/a^2}$ is the energy of a particle with four-momentum p_μ in the rest frame of the observer. Hence, the distribution function

$$f(\tau, p_i) = \frac{1}{(2\pi)^3} \exp\left(\frac{\mu - E(p_i)}{T}\right) \tag{2.7}$$

reduces to the distribution (2.6) in the non-relativistic limit $T/m \ll 1$, and remains a solution of equation (2.4), provided that the temperature and chemical potential scale appropriately,

$$T = \frac{T_0}{a^2}, \quad \mu = m + \frac{\mu_0 - m}{a^2}. \tag{2.8}$$

Indeed, in order for the distribution to remain time-independent the dark matter temperature has to be inversely proportional to the scale factor, because in the non-

relativistic limit $E(p_i) \approx m^2 + p^2/a^2$. With this temperature scaling, and in order to again preserve a time-independent distribution, the chemical potential has to be given by the equation above.

Note however that a thermal distribution of the form (2.7) only solves the collisionless Boltzmann equation (2.4) either in the relativistic or non-relativistic limits, but not in both. Hence, our choice of the distribution (2.6) is justified if dark matter particles kinetically decoupled while non-relativistic. This is indeed what happens for instance if dark matter consists of WIMPs. In this case, although dark matter typically decouples chemically from the thermal bath while mildly non-relativistic, at $T/m \approx 1/25$, interactions with standard model particles keep dark matter particles in equilibrium with the thermal bath until much later [45]. In contrast, most treatments of warm dark matter models assume that dark matter consists of fermionic particles which kinetically decouple while highly relativistic, and thus assume that the distribution function follows a (non-Gaussian) Fermi distribution with vanishing chemical potential, $f(p) = [1 + \exp(p/T_0)]^{-1}$.

Since in the ΛCDM cosmological model dark matter particles are assumed to be cold, we usually calculate their energy-momentum tensor in the strict non-relativistic limit $T/m \to 0$, in which their energy density $\rho \equiv -T^0{}_0$ becomes

$$\bar{\rho} \equiv \frac{\bar{\rho}_0}{a^3} \equiv m \exp\left(\frac{\mu_0 - m}{T_0}\right)\left(\frac{mT}{2\pi}\right)^{3/2}. \tag{2.9}$$

Here, we go beyond this non-relativistic limit and calculate the energy-momentum tensor and its perturbations to first order in T/m. Inserting equation (2.6) into (2.2) we find

$$\rho = \bar{\rho}\left(1 + \frac{3}{2}\frac{T}{m} + \cdots\right), \tag{2.10}$$

and, similarly, the pressure becomes

$$P \equiv \frac{1}{3} T^i{}_i = \bar{\rho} \frac{T}{m} + \cdots . \tag{2.11}$$

These expressions capture just what we expect from a gas of non-relativistic particles in an expanding universe. Note that $T^0{}_i$ vanishes and $T^i{}_j$ is diagonal, both because of rotational invariance. The correction factors to the conventional results arise from the thermal average of p^2, which decays as $1/a^2$ because physical momenta redshift with the scale factor. It is easy to check that the energy density and pressure above satisfy the conservation equation $\rho' + 3\mathcal{H}(\rho + P) = 0$. The equation of state parameter of this fluid is

$$w \equiv \frac{P}{\rho} \approx \frac{T}{m}, \tag{2.12}$$

which decays with the square of the scale factor, and is proportional to the small parameter T/m. Some authors have constrained the equation of state parameter w of dark matter [46–48], but they typically assume that w is a constant, rather than proportional to $1/a^2$.

The expressions above are valid only at late times, in the non-relativistic limit. As we proceed back in time the momenta of the dark matter particles increase, and thus become relativistic. As long as dark matter particles remain collisionless, the distribution function (2.6) remains a solution of the Boltzmann equation (2.4). Therefore, substitution of equation (2.6) into equation (2.2) leads, to all orders in T/m, to the energy density

$$\rho = \bar{\rho} \left(\frac{m}{4T}\right)^{1/2} \exp\left(\frac{m}{4T}\right) K_1\left(\frac{m}{4T}\right), \tag{2.13}$$

where K_1 is the corresponding modified Bessel function of the second kind and T scales as in equation (2.8). In the limit $m/T \to 0$ the energy density scales like that

of relativistic particles, whereas in the limit $m/T \to \infty$, the energy density approaches the limit (2.10).

2.2.2 Perturbations

Our next goal is to derive an equation that captures the impact of a non-zero temperature on the evolution of the dark matter density perturbations. Under different assumptions and approximations, such an equation has been derived many times in the literature, which extends as far back as to the pioneering work of Gilbert [49], the author after whom the equation is mostly named. Most of the relatively recent derivations of the Gilbert equation have focused on warm dark matter particles, which decouple while relativistic (see e.g. [50, 51]), although a few analyses have also considered particles that decouple while non-relativistic (see e.g. [52]). Our derivation here relies on the linearized and relativistic Boltzmann equation in synchronous gauge, and does not involve any approximations beyond an expansion in $\sqrt{T/m}$, a small parameter in the non-relativistic limit. Our Gilbert equation can thus be incorporated directly into existing numerical Boltzmann codes to calculate CMB anisotropy and matter power spectra in the linear regime.

In order to study the evolution of structure when dark matter has a non-zero temperature, we perturb the homogeneous and isotropic FLRW line-element (2.3),

$$ds^2 = a^2\left[-d\tau^2 + (\delta_{ij} + h_{ij})dx^i dx^j\right], \quad h_{ij} = \frac{k_i k_j}{k^2}h + 6\left(\frac{k_i k_j}{k^2} - \frac{1}{3}\delta_{ij}\right)\eta, \tag{2.14}$$

in which we have chosen synchronous gauge, and we concentrate on the perturbation caused by a single Fourier mode of wave vector $\vec{k}$. To calculate the perturbed energy momentum tensor, we need to perturb the background distribution (2.7). Let us write the perturbed distribution function as

$$f(\tau, \vec{x}, \vec{p}) = \bar{f}(p) + \delta f(\tau, \vec{x}, \vec{p}), \tag{2.15}$$

where $\bar{f}$ is the thermal distribution (2.7) and p is the magnitude of the spatial momentum,

$$p \equiv a\sqrt{g^{ij}p_i p_j} = \sqrt{\delta^{ij}p_i p_j} - \frac{1}{2}\frac{h_{ij}p_i p_j}{\sqrt{\delta^{ij}p_i p_j}}. \tag{2.16}$$

Note that p depends on the metric, so $\bar{f}$ also contributes to the perturbations of the distribution function. Then, the perturbation δf obeys the linearized Boltzmann equation

$$\frac{\partial \delta f}{\partial \tau} + \frac{\partial \delta f}{\partial x^i}\frac{1}{a^2}\frac{p_i}{p^0} - \frac{1}{2}\frac{\bar{f}'}{p}\frac{\partial h_{jk}}{\partial \tau}p_j p_k = 0, \tag{2.17}$$

where a prime denotes a derivative with respect to p in this case, and Einstein's summation convention is implied even if repeated indices are not in opposite locations. It is then easy to check that the perturbed Boltzmann equation admits the line-of-sight integral solution

$$\begin{aligned}\delta f(\tau, \vec{k}, \vec{p}) = \delta f(\tau_{\rm dec}, \vec{k}, \vec{p}) \exp\left[-i\,\Delta(\tau, \tau_{\rm dec})\vec{p}\cdot\vec{k}\right] + \\ + \frac{1}{2}\frac{\bar{f}'}{p}\int_{\tau_{\rm dec}}^{\tau} d\tau'\, p_i p_j\, h'_{ij}(\tau', \vec{k}) \exp\left[-i\,\Delta(\tau, \tau')\,\vec{p}\cdot\vec{k}\right],\end{aligned} \tag{2.18}$$

where

$$\Delta(\tau_2, \tau_1) \equiv \int_{\tau_1}^{\tau_2} d\tau \frac{1}{a^2(\tau)\,p^0(\tau)} \quad \text{and} \quad p^0 = \frac{1}{a}\sqrt{m^2 + \frac{p^2}{a^2}}. \tag{2.19}$$

Note that $\Delta(\tau_2, \tau_1)\cdot p$ is the comoving distance traveled by a dark matter particle with covariant momentum p between times τ_1 and τ_2. In particular, $p/(a^2 p^0)$ is its comoving velocity, which, in our non-relativistic approximation can be taken to be $p/(ma)$. For a thermal distribution, the magnitude of the root mean square momentum is of order

$\sqrt{mT_0}$, which leads us to define the comoving free streaming length

$$d(\tau_2, \tau_1) \equiv \sqrt{\frac{T_0}{m}} \int_{\tau_1}^{\tau_2} \frac{d\tau}{a}. \tag{2.20}$$

As we shall see shortly, the free streaming captured by the solution (2.18) leads to an exponential suppression of structure on comoving scales $k\, d \gg 1$.

In a universe dominated by matter and radiation,

$$d(\tau_2, \tau_1) = \frac{1}{2}\sqrt{\frac{T_{\rm eq}}{m}}\, \tau_{\rm eq} \log\left(\frac{\tau_2}{\tau_1}\frac{\tau_1 + 2\tau_{\rm eq}}{\tau_2 + 2\tau_{\rm eq}}\right), \tag{2.21}$$

where $T_{\rm eq}$ and $(\sqrt{2}-1)\tau_{\rm eq}$ respectively are the dark matter temperature and conformal time at matter-radiation equality. Note that during radiation domination, the product $\tau\sqrt{T/m}$ is constant, and thus roughly agrees with its value at matter-radiation equality.

We obtain the perturbed energy momentum tensor $\delta T^{\mu}{}_{\nu}$ by substituting the solution (2.18) into equation (2.2). As we describe in detail in Appendix B, the perturbed energy density becomes

$$\delta\rho = -\frac{\bar{\rho}}{2}\int_{\tau_{\rm dec}}^{\tau} d\tau' e^{-d^2k^2/2}\Big\{h' - (d\,k)^2(h' + 4\eta') + \\ + \frac{T(\tau)}{2m}\left[(5 - d^2k^2)h' - (d\,k)^2(7 - d^2\,k^2)(h' + 4\eta')\right]\Big\}. \tag{2.22}$$

In this equation, the free streaming length $d = d(\tau, \tau')$ is given by equation (2.20) and we have assumed that the perturbation δf vanishes at decoupling. Note the exponential factor inside the integrand, which suppresses the contributions of the potentials on scales $(d\,k)^2 \gg 1$. The exponential arises from the moments of spherical Bessel functions with respect to the Gaussian distribution in equation (2.6), and is thus sensitive to the precise form of the distribution function. Because the comoving free streaming length d depends on the dark matter temperature, the absence of such

suppression allows us to place quite stringent constraints on T/m. The CDM density contrast

$$\frac{\delta\rho}{\rho} \approx \frac{\delta\rho}{\bar{\rho}}\left(1 - \frac{3}{2}\frac{T}{m}\right) \tag{2.23}$$

contains an additional correction due to the non-zero dark matter temperature, but the impact of this correction is typically much smaller than that due to the free-streaming term.

As opposed to what happens in the limit $T/m \to 0$, in which the dark matter velocity can be taken to vanish in synchronous gauge, in this case the velocity potential is non-zero. With $\delta T^0{}_i \equiv (\rho + P)\partial_i v$, we find

$$(\rho + P)v = \frac{\bar{\rho}}{2}\sqrt{\frac{T}{m}}\frac{1}{k}\int_{\tau_{\rm dec}}^{\tau} d\tau' e^{-d^2k^2/2}\left[d\,k\,(3h' + 8\eta') - (d\,k)^3(h' + 4\eta')\right], \tag{2.24}$$

which shows an analogous suppression of the velocity perturbation on scales much smaller than d. Finally, following the same approach, we arrive at

$$\delta T^i{}_j = -\frac{\bar{\rho}}{2}\frac{T}{m}\int_{\tau_{\rm dec}}^{\tau} d\tau' e^{-d^2k^2/2}\Big\{h'\delta_{ij} + 2h'_{ij} + (d\,k)^2\left[(h' + 4\eta')\delta_{ij} + (5h' + 16\eta')\hat{k}_i\hat{k}_j\right] + \\ + (d\,k)^4(h' + 4\eta')\hat{k}_i\hat{k}_j\Big\},$$

from which we can immediately read off the perturbed pressure δp and the scalar anisotropic stress π, $\delta T^i{}_j \equiv \delta P\,\delta^i_j - k^i k_j\,\pi$,

$$\delta P = -\frac{\bar{\rho}}{2}\frac{T}{m}\int_{\tau_{\rm dec}}^{\tau} d\tau'\, e^{-d^2k^2/2}\left[h' - 4\eta' + (d\,k)^2(h' + 4\eta')\right], \tag{2.25}$$

$$\pi = \frac{\bar{\rho}}{2}\frac{T}{m}\frac{1}{k^2}\int_{\tau_{\rm dec}}^{\tau} d\tau'\, e^{-d^2k^2/2}\left[2h' + 12\eta' + (d\,k)^2(5h' + 16\eta') + (d\,k)^4(h' + 4\eta')\right]. \tag{2.26}$$

Note that the contribution of the anisotropic stress to the energy momentum tensor, of order $k^2\pi$, is of the same magnitude as that of the pressure perturbation. The pressure

perturbation itself is a factor T/m smaller than the energy density perturbation, as expected.

With $\delta\rho$, δP, π and δu given by the previous expressions, it is relatively straightforward, albeit tedious, to verify that the energy momentum tensor is covariantly conserved up to terms of order $(T/m)^{3/2}$. In particular, these quantities obey the perturbed hydrodynamical equations of energy conservation,

$$\delta\rho' + 3\mathcal{H}(\delta\rho + \delta P) - k^2\left[(\rho + P)v + \mathcal{H}\pi\right] + (\rho + P)\frac{h'}{2} = 0, \tag{2.27}$$

and momentum conservation,

$$[(\rho + P)v]' + 4\mathcal{H}(\rho + P)v + \delta P - k^2\pi = 0. \tag{2.28}$$

In any case, because the anisotropic stress is of the same order as the pressure perturbation, a (perfect) fluid description of dark matter breaks down on small scales. For instance, as we mention in Appendix 2.B, collisionless dark matter does not undergo acoustic oscillations, even if the dark matter particles have a non-zero velocity dispersion, and thus, a non-zero pressure.

2.3 Impact on Structure Formation

Here, we are interested in assessing the impact of a non-zero CDM temperature on the formation of structure at scales accessible to linear perturbation theory. At present, constraints on the linear power spectrum at these smallest scales rely on the Lyman-alpha forest [53], which probes comoving wave numbers of order

$$k_{\rm max} \approx 2\ h\,{\rm Mpc}^{-1} \tag{2.29}$$

at redshifts $z \approx 3$. This should be compared with the wavenumber we can probe with the ℓ'th multipole of the cosmic microwave background,

$$k_{\rm CMB} \approx 0.21 \frac{\ell}{3500} {\rm Mpc}^{-1}, \tag{2.30}$$

which is about an order of magnitude smaller than $k_{\rm max}$ even for the angular scales probed by ACT [54], SPT [55] and Planck [14].

In a ΛCDM cosmology the scale $k_{\rm max}$ enters the horizon at a redshift of about $z_{\rm max} \approx 5 \cdot 10^5$. Because our analysis assumes that at redshift $z_{\rm max}$ cold dark matter particles were already non-relativistic, the temperature today hence needs to obey

$$\frac{T_0}{m} \lesssim 2 \cdot 10^{-12}, \tag{2.31}$$

and has to be proportionally smaller if we are interested in length scales smaller than $k_{\rm max}$. In addition, because we also assume that dark matter is collisionless, it needs to decouple at redshift $z_{\rm dec} > z_{\rm max}$. This behavior should be contrasted with that of neutrinos or warm dark matter, which decouple kinetically while being relativistic and become non-relativistic after galactic scales have entered the horizon. The impact of a non-zero dark matter temperature on scales with wave numbers much larger than (2.29), as well as the imprint of dark matter decoupling on the matter power spectrum, is discussed in [33, 36].

Our next goal is to estimate the matter density perturbation (2.22) after matter-radiation equality, at $\tau_0 > \tau_{\rm eq}$. At recombination, the density of dark matter has an impact on the structure of the CMB Doppler peaks, and at redshift zero, the dark matter perturbation is directly related to the matter power spectrum. Because dark matter particles are non-relativistic, we assume that we are in a regime in which we can drop the term proportional to $T/2m$ in equation (2.22), which in this

approximation becomes

$$\delta\rho = -\frac{\bar{\rho}(\tau_0)}{2}\left\{\int_{\theta_{\rm dec}}^{\theta_{\rm eq}} d\theta + \int_{\theta_{\rm eq}}^{\theta_0} d\theta\right\} e^{-d^2k^2/2}\left[\frac{dh}{d\theta} - (d\,k)^2\left(\frac{dh}{d\theta} + 4\frac{d\eta}{d\theta}\right)\right]. \tag{2.32}$$

Note that we have split the integral into the contribution to $\delta\rho$ during radiation domination and that during matter domination, and that we have introduced the new integration variable

$$\theta = \frac{k\tau}{\sqrt{3}}. \tag{2.33}$$

2.3.1 Radiation domination

Well in the radiation-dominated era, the free-streaming length in equation (2.21) becomes

$$d(\theta_0, \theta) \approx \frac{1}{2}\sqrt{\frac{T_{\rm eq}}{m}}\tau_{\rm eq}\log\left(2\frac{\theta_{\rm eq}}{\theta}\right), \qquad \theta \ll \theta_{\rm eq}, \tag{2.34}$$

During this era, we can neglect the impact of dark matter on the gravitational potentials, which take their standard values

$$\frac{dh}{d\theta} = \frac{12\mathcal{R}_i}{\theta}\left(\frac{2(\cos\theta - 1)}{\theta^2} + \frac{2\sin\theta}{\theta} - 1\right), \quad \frac{d\eta}{d\theta} = -\frac{4\mathcal{R}_i}{\theta}\left(\frac{\sin\theta}{2\theta} - \frac{1-\cos\theta}{\theta^2}\right), \tag{2.35}$$

where $\mathcal{R}_i$ is the initial (primordial) curvature perturbation.

There are two dimensionless ratios that determine the behavior of the perturbations: $\theta_{\rm eq}$ and the ratio of wavelength to the free-streaming length in equation (2.34), $k\,d$. In order to proceed, we focus on the short-wavelength limit $\theta_{\rm eq} \gg 1$ and discuss the limits $k\,d \ll 1$ and $k\,d \gg 1$, in this order, separately.

In the absence of free-streaming ($k\,d \equiv 0$), the dominant contributions to the integral in equation (2.32) stem from the extrema of $dh/d\theta$ and $d\eta/d\theta$ at $\theta_{\rm max} \approx 3$.

Because the logarithmic derivative of $(k\,d)^2$ is

$$\frac{d\log(k\,d)^2}{d\theta} = -\frac{2}{\theta}\log^{-1}\frac{2\theta_{\rm eq}}{\theta}, \tag{2.36}$$

in the limit of large $\theta_{\rm eq}$ both the exponential and the factor of $(kd)^2$ in the integrand are then slowly varying functions of θ around $\theta_{\rm max}$ for small $k\,d$. Therefore, taking the factors of $k\,d$ out of the integral, and evaluating at the extrema of the corresponding potential derivatives we get

$$\frac{\delta\rho}{\bar{\rho}} \approx 6\mathcal{R}_i\left(\log\theta_{\rm eq}+\gamma-\frac{1}{2}\right) - \frac{\mathcal{R}_i}{8}\frac{T_{\rm eq}}{m}\,\theta_{\rm eq}^2\log^2\frac{2\theta_{\rm eq}}{3}\times\left(18\log\theta_{\rm eq}+18\gamma-13\right), \tag{2.37}$$

where γ is Euler's constant. As expected, free streaming suppresses density perturbations at small scales, although, by assumption, the effect is small in this limit.

When $k\,d$ is large, the exponential in equation (2.32) is a rapidly varying function of θ, which strongly suppresses the contribution of a mode when $\theta \ll \theta_{\rm eq}$. Therefore, in this limit, the integral is dominated by the values of the integrand around $\theta = \theta_{\rm eq}$. Assuming constant derivatives of the gravitational potentials around that point, and taking those functions outside the integral we thus get

$$\frac{\delta\rho}{\bar{\rho}} \approx -6\,\mathcal{R}_i\log 2\exp\left[-\frac{3}{8}(T_{\rm eq}/m)\theta_{\rm eq}^2\log^2 2\right], \tag{2.38}$$

where we have only kept the contribution of the term proportional to $(d\,k)^2$. In this case, the exponential suppression of the density perturbations essentially smoothes out dark matter inhomogeneities on comoving scales with $\theta_{\rm eq}^2 \gg m/T_{\rm eq}$.

2.3.2 Matter domination

It still remains to calculate the contributions to (2.32) from the matter-dominated era. During this epoch, the comoving free-streaming length in equation (2.21) becomes

$$d(\theta_0, \theta) \approx \frac{1}{2}\sqrt{\frac{T_{\rm eq}}{m}}\tau_{\rm eq} \log\left(1 + 2\frac{\theta_{\rm eq}}{\theta}\right), \qquad \theta \gg \theta_{\rm eq}. \tag{2.39}$$

Hence, well in the matter-dominated era, the free-streaming length is suppressed by a factor of order $\theta_{\rm eq}/\theta$ relative to that of the streaming length during radiation domination, even though most of the growth in $\delta\rho/\rho$ happens during this time. As a result, in the limit in which free-streaming has a sizable impact on structure formation, we expect that impact to be largest during the radiation-dominated regime.

To illustrate the impact of free-streaming during the matter-dominated epoch, consider for instance the ratio

$$X \equiv \frac{\mathcal{T}(k, z, T_0/m)}{\mathcal{T}(k, z, 0)}, \tag{2.40}$$

where $\mathcal{T}(k, z, T_0/m)$ is the dark matter transfer function at wave number k, redshift z and present dark matter temperature T_0/m. In Figure 2.1, we plot X as a function of k for fixed $T_0/m = 10^{-7}$ at redshifts $z = 3300$ and $z = 0$. As seen in the figure, at $z = 3300$ most of the suppression of dark matter inhomogeneities (due to free streaming) is already in place. There is an additional suppression at $z = 0$, but the latter is not significantly different from that at $z = 3300$.

Equations (2.34) and (2.39) allow us to obtain a rough estimate of the comoving length scale below which free streaming leads to a suppression of dark matter anisotropies. Both equations show that the free streaming length d responsible for the exponential suppression of structure in equation (2.32) equals, modulo a logarithmic factor, $\sqrt{T_{\rm eq}/m}\,\tau_{\rm eq}/2$. We are thus led to define the comoving free-streaming

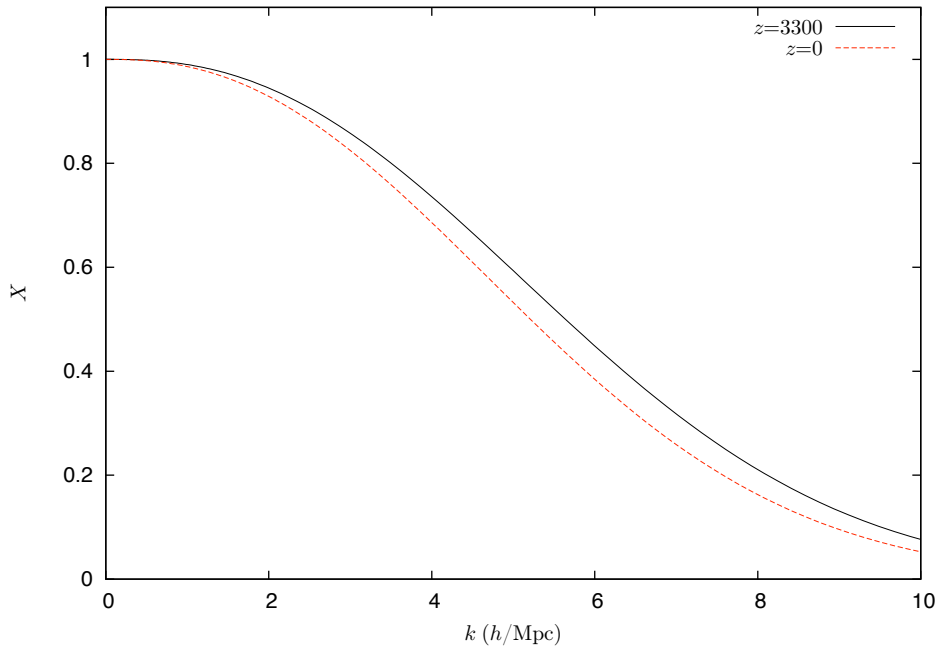


Figure 2.1: A plot of the dark matter transfer function ratio X in equation (2.40) for fixed temperature $T_0/m = 10^{-7}$ at redshifts $z = 3300$ (black continuous) and $z = 0$ (red dashed). Most of the impact of free streaming on the suppression of dark matter inhomogeneities occurs before matter-radiation equality, at $z \approx 3300$.

wave number

$$k_{\rm fs} = \sqrt{\frac{m}{T_{\rm eq}}}\frac{2}{\tau_{\rm eq}} \approx \sqrt{\frac{m}{T_{\rm eq}}}\frac{\Omega_m h^2}{16\,{\rm Mpc}}. \tag{2.41}$$

Noting that the dark matter temperature today T_0 is related to that at equality by

$$T_{\rm eq} = (1+z_{\rm eq})^2 T_0 \approx (2.3\cdot 10^4\,\Omega_m h^2)^2\,T_0, \tag{2.42}$$

we thus obtain the free-streaming wave number

$$k_{\rm fs} \approx 2.6\cdot 10^{-6}\sqrt{\frac{m}{T_0}}\,{\rm Mpc}^{-1}, \tag{2.43}$$

which does not depend on the dark matter density. This is basically the length

scale derived in references [28, 50], although the scalings with $\Omega_m h^2$ differ. The reason is that whereas in our analysis T_0/m and $\Omega_m h^2$ are independent parameters, in warm dark matter models, the current dark matter density and the dark matter mass are used to determine the dark matter temperature (we are ignoring baryons here). Modulo the logarithmic factor that we dropped, the free-streaming scale (2.41) also agrees with the scale derived for WIMPs in [33], provided that equations with the same parameters are compared. For further analytical estimates of the impact of a non-zero temperature on the matter power spectrum during matter domination, see reference [52].

2.3.3 Power Spectra

We have seen that a non-vanishing dark matter temperature generically leads to a suppression of structure on small scales. In order to determine the quantitatively precise nature of this suppression, we need to rely on a numerical solution of the perturbation equations.

In Figure 2.2, we plot the temperature anisotropy power spectrum for different values of the dark matter temperature. Although these temperatures do not satisfy the condition (2.31), they are low enough for our non-relativistic approximation to be trusted, since, according to equation (2.30) the comoving scales probed by the CMB are smaller than $k_{\rm max}$ in equation (2.29). Even at these temperatures, the impact of a non-zero temperature on the CMB power spectrum is not visible, so we have magnified the difference with respect to the the $T = 0$ power spectrum by a factor of a hundred. As seen in the figure, a non-zero temperature does not shift the location of the acoustics peaks significantly, which simply grow in amplitude, the growth being more pronounced for the even peaks. At least partially, this behavior is relatively simple to explain: in synchronous gauge, and in the approximation of instantaneous recombination, the temperature anisotropies are mostly determined by

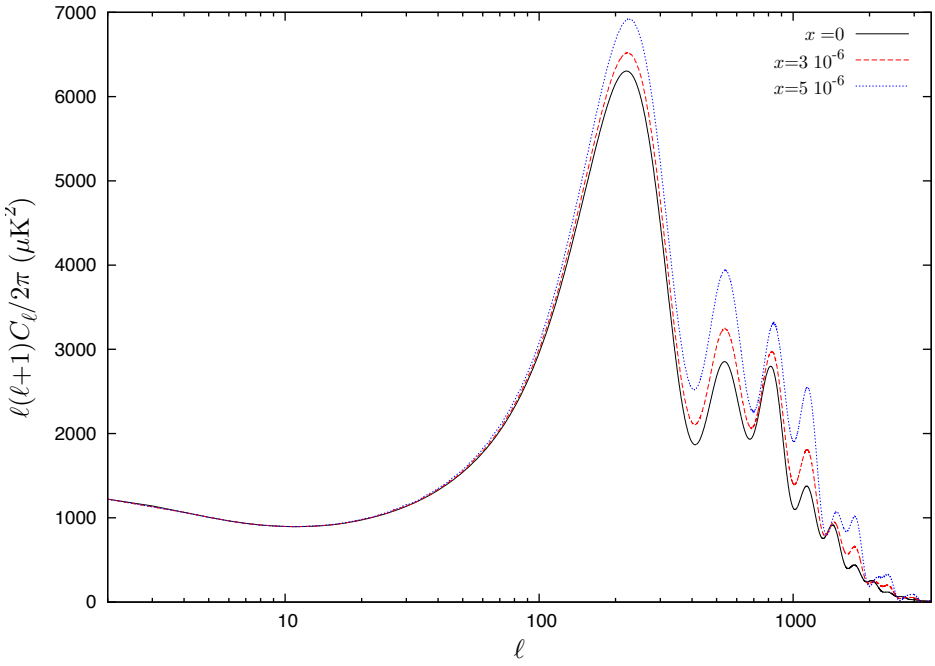


Figure 2.2: Temperature anisotropy power spectra for different values of $x \equiv \sqrt{T_0/m}$ as a function of spherical multipole ℓ. Differences in the angular power spectra with respect to the $x = 0$ case have been magnified by a factor of 10^2.

the Sachs-Wolfe term at last scattering [56]

$$F(k) = \frac{\mathcal{R}_i}{5}\left[3\mathcal{T}(k/k_{\rm eq})R_L - \frac{\mathcal{S}(k/k_{\rm eq})}{(1+R_L)^{1/4}}\exp\left(-\int_0^{\tau_L}\Gamma d\tau\right)\cos\left(k\int_0^{\tau_L}\frac{d\tau}{\sqrt{1+R}}\right)\right], \tag{2.44}$$

where $\sqrt{2}k_{\rm eq}$ is the mode that enters the horizon at recombination, R is the baryon to photon density ratio, $\mathcal{T}$ is the dark matter transfer function, and $\mathcal{S}$ is the transfer function that determines the amplitude of the photon acoustic oscillations. A non-zero dark matter temperature hardly impacts the amplitude of the acoustic oscillations $\mathcal{S}$, since the latter is determined during radiation domination, but it does suppress the transfer function $\mathcal{T}$, because free-streaming damps dark matter perturbations. As a result, the source term F increases in magnitude at the location of the even peaks (where the cosine in equation (2.44) is positive), leading to a power increase in the

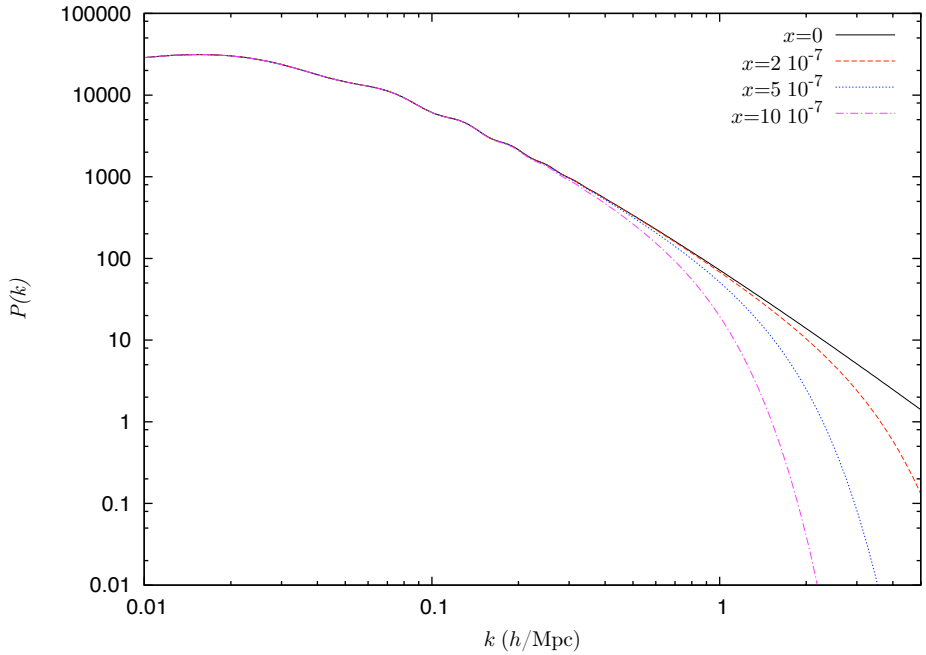


Figure 2.3: Matter power spectrum for different values of $x \equiv \sqrt{T_0/m}$ as a function of comoving wave number $k\,h\,\mathrm{Mpc}$. On length scales smaller than the free streaming length, structure is suppressed. Because the free-streaming length is proportional to the dark matter temperature, larger temperatures lead to more suppression.

even peaks of the angular power spectrum. Of course, by the same token we would expect a power decrease at the odd peaks; instead we just observe a less prominent increase in the peak amplitude.

The effects of a non-zero dark matter temperature are more pronounced in the (total) matter power spectrum at $z = 0$, which is a far more direct probe of the dark matter distribution. In Figure 2.3, we plot the matter power spectrum also for a different set of dark matter temperatures. Whereas the impact of a non-zero temperature on the CMB was hardly visible, here, departures in the matter power spectrum are very prominent on scales $k > 0.4\,h\,\mathrm{Mpc}^{-1}$, and show the expected suppression due to the free streaming of dark matter particles. Of course at these scales linear perturbation theory breaks down, so our linear calculation has to be

appropriately interpreted.[1]

Ma has found that in warm dark matter models the ratio of dark matter linear transfer functions X in equation (2.40) is well fit by

$$X \approx \frac{1}{[1 + (\alpha k)^{2\nu}]^{5/\nu}}, \tag{2.45}$$

where α depends on various cosmological parameters, such as Ω_m and the mass of the warm dark matter particle, and the exponent ν is a constant, $\nu \approx 1.2$ [28]. Although our non-relativistic approximation does not allow us to numerically explore the regime $k\,d \gg 1$ in which we expect structure to be exponentially suppressed, we find that different exponents ν provide better fits to the numerical results as we vary the dark matter temperature. Say, for $T_0/m = 10^{-16}$ the exponent $\nu \approx 1.17$ gives a squared sum of square residuals about thirty times smaller than for $\nu = 1.2$, whereas for $T_0/m = 10^{-14}$, $\nu \approx 1.25$ gives sum of square residuals about three times smaller than for $\nu = 1.2$.

We have not explored however how the parameter α depend on the dark matter temperature or the remaining cosmological parameters. If for a given non-zero CDM temperature and fixed cosmological parameters we simply determine the value of α in equation (2.45) that best fits the the cold dark matter spectrum for fixed $\nu = 1.2$, we find an excellent agreement between both. This agreement between the CDM and WDM spectra is what one would expect from the relative similarity of the exponent ν described above. On the other hand, when we compare CDM and WDM spectra with the same cosmological parameters and the same velocity dispersion at present[2] we find a significant disagreement at small scales, as shown in Figure 2.4 .

[1] If the dark matter temperature is high enough, the associated suppression of structure may keep all scales in the linear regime. Obviously, there would not be any collapsed haloes in such a universe, which would be very different from ours.

[2] We fix the mass of the warm dark matter particle by matching equation (A3) of reference [28] to the desired CDM velocity dispersion $v_{\rm rms} = \sqrt{3T_0/m}$. The resulting mass is then substituted into equation (A9) of [28], which then determines the coefficient α.

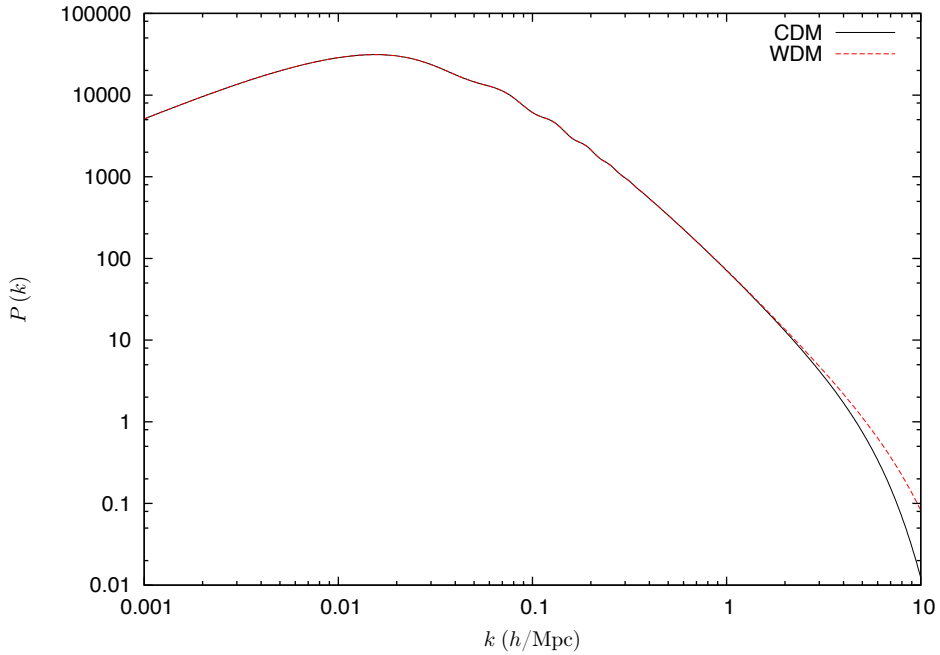


Figure 2.4: Comparison of a CDM matter power spectrum at $T_0/m = 10^{-14}$ with a WDM obtained from the fitting formulate in equation (2.45), with $\nu = 1.2$ and the value of α that matches the CDM velocity dispersion at present. The remaining cosmological parameters have the same values.

2.4 Limits

The results of the previous section allow us to place a rough but conservative limit on the dark matter temperature today. A host of cosmological measurements of small scale structure seem to be in good agreement with the standard ΛCDM cosmological model. As we saw in Section 2.3, the CMB is not very sensitive to the dark matter temperature on these small scales. On the other hand, the distribution of large scale structure is directly affected by a non-zero dark matter temperature, and can probed down to $k_{\text{max}} \approx 2\,h\,\text{Mpc}^{-1}$ by Lyman-alpha forest observations in reference [53]. Therefore, we expect the most stringent constraints on the dark matter velocity to arise from measurements of the dark matter power spectrum on these scales.

The most recent (published) analysis of how the Lyman-alpha forest constrains

the matter power spectrum is that of reference [53]. The measurement suffers from significant systematic errors, affecting the amplitude of the power spectrum at $z = 3$ and $k = 2\,h\,\mathrm{Mpc}^{-1}$ by factors of up to 25%. Demanding then that the relative correction to the dark matter overdensity in equation (2.37) be less than 50% on those scales we thus arrive at the limit $T_{\rm eq}/m \lesssim 10^{-6}$. Because the temperature is inversely proportional to a^2, and $1 + z_{\rm eq} \approx 3 \cdot 10^3$, this implies that $T_0/m \lesssim 10^{-13}$. Since this ratio is smaller that the one necessary for the validity our approximation, equation (2.31), our analysis is at the very least self-consistent. We derive a sharper numerical limit next.

Numerical Results

In order to place rigorous and precise marginalized limits on the temperature to mass ratio, we resort to the by-now standard Bayesian approach to parameter estimation based on Markov-Chain Monte Carlo methods. We have modified the publicly available Boltzmann integrator CAMB and the Markov-Chain Monte Carlo engine CosmoMC [3] [57, 58] by including the necessary modifications of the dark matter equations needed to account for a non-zero dark matter temperature, as detailed in Appendix 2.A. We sample the posterior probability for a spatially flat cosmological model with parameters H_0 (Hubble's constant today), Ω_Λ (critical density fraction of a cosmological constant), $\Omega_b h^2$ (baryon density), τ (optical depth), n_s (scalar spectral index), A_s (scalar spectral amplitude), $A_{\rm SZ}$ (amplitude of a Sunyaev-Zeldovich template) and $\sqrt{T_0/m}$ (square root of present dark matter temperature to mass ratio) with a set of four Monte Carlo Markov chains of at least 2×10^5 elements each, generated with an appropriately modified version of CosmoMC. We impose flat priors on all parameters, assume that the universe is spatially flat and neglect tensor modes. To check for the converge of our chains, we monitor the Gelman and Rubin statistic [59], which stays under 10^{-2}. Following CosmoMC output, we also estimate

[3] http://cosmologist.info/cosmomc

the statistical errors on our upper limits by exploring their changes upon split of our chains in several subsamples, which remain of the order of 1%.

In order to obtain the strictest constraints on the dark matter temperature, it is crucial to employ observations at small scales. We thus include constraints on the linear matter power spectrum at redshift $z = 3$ derived from Lyman alpha observations in reference [53] (surprisingly, this 2005 analysis is still state-of-the-art). Measurements of the linear power spectrum on the scales probed by the Lyman alpha forest are notoriously difficult, and typically require structure formation simulations for various input power spectra and cosmological parameters. Although the simulations carried in reference [53] just involved the standard ΛCDM cosmological model, their constraints on the linear power spectrum should remain valid as long as the linear matter power spectrum does not significantly deviate from that in ΛCDM. We enforce such an agreement at the corresponding scales with the temperature prior

$$\sqrt{\frac{T_0}{m}} \leq 2 \cdot 10^{-7}, \tag{2.46}$$

which also guarantees the validity of our perturbative equations.

Although at the temperatures of interest the cosmic microwave background is hardly affected, observations of the cosmic microwave background are nevertheless crucial to constrain the remaining cosmological parameters. We therefore include cosmic microwave measurements from the WMAP 9 year data release [60], as well as ACT [54] and SPT data [55], which probe the angular power spectrum on smaller scales. We also include large scale structure data from an SDSS luminous red galaxy (LRG) sample [61]. We calculate the likelihood of our angular power spectra with the numerical codes supplied by the corresponding collaboration [62], and we employ the patch[4] written by Anže Slosar to evaluate matter power spectra likelihoods on Lyman-alpha scales [53].

Proceeding as outlined above, using the afore-mentioned datasets, we obtain the

[4]http://www.slosar.com/aslosar/lya.html

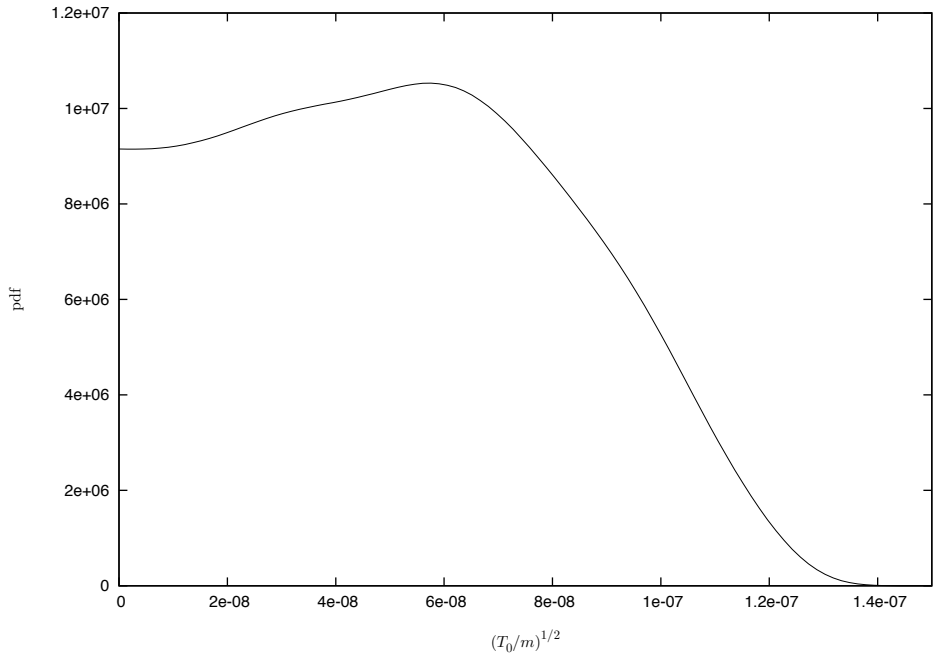


Figure 2.5: Marginalized posterior distribution of $\sqrt{T_0/m}$. Note the relatively flat plateau at low temperatures, which indicates that data cannot discriminate between temperatures in the range $\sqrt{T_0/m} \lesssim 6 \cdot 10^{-8}$.

marginalized posterior distributions for $\sqrt{T_0/m}$ shown in Figure 2.5. We also list mean, standard deviation and credible upper limits of the corresponding posterior distribution in Table 2.1. Simple inspection of the posterior distribution shows that there is no evidence for a non-zero dark matter temperature in the data. In fact, adding the temperature to mass ratio $\sqrt{T_0/m}$ to the standard cosmological parameters in ΛCDM improves the log likelihood just by 0.15. The 95% upper credible limit on the dark matter to temperature ratio then is

$$\frac{T_0}{m} \leq 1.07 \cdot 10^{-14}, \tag{2.47}$$

which translates into an upper limit on the present dark matter rms velocity $v_{\rm rms} \leq 54\,\mathrm{m/s}$. The mean of the posterior distribution of T_0/m is several standard deviations away from the edge of our prior (2.46), which therefore has no influence on the upper

Dataset	μ	σ	68%	95%
CMB+LRG+Lyα	$5.22 \cdot 10^{-8}$	$3.05 \cdot 10^{-8}$	$\leq 6.84 \cdot 10^{-8}$	$\leq 1.03 \cdot 10^{-7}$

Table 2.1: Marginalized posterior mean μ, standard deviation σ and 68% and 95% upper credible limits on $\sqrt{T_0/m}$.

limit (2.47). These results do not depend on any particular model, and only rely on the assumptions that dark matter is collisionless, and that the distribution of its momenta is Maxwellian, with a temperature $T/m \ll 1$. As we emphasized previously, equation (2.47) should not be interpreted as a constraint on the actual dark matter temperature at present (because the latter is mostly found in collapsed haloes), but as an extrapolation: The limit implies that at redshift z, where $1 \ll z \leq z_{\rm dec}$, the dark matter temperature has to obey

$$\frac{T}{m} \leq 1.07 \cdot 10^{-14}(1+z)^2. \tag{2.48}$$

Imagine, for example, that a dark matter model predicts a decoupling redshift $z_{\rm dec}$, at which dark matter matter decouples kinetically. If this decoupling redshift obeys $z_{\rm dec} > z_{\rm max} \approx 5 \cdot 10^5$ the assumptions of our analysis hold. Then, evaluating the inequality (2.48) at $z_{\rm dec}$, we obtain an actual limit on the dark matter temperature to mass ratio at decoupling. If the ratio in the model under consideration violates this limit, the model is consequently ruled out by our analysis.

It is also illustrative to compare the temperature limit (2.47) with the baryonic temperature to mass ratio at present. Although most electrons recombine with hydrogen and helium nuclei around last scattering, there is a residual ionization that keeps baryons and photons in thermal contact until a redshift of order $z \approx 140$ [63]. Therefore, ignoring reionization and any other process that may affect the hydrogen temperature on large scales, we would expect the present hydrogen temperature to mass ratio to be $T_H^0/m \approx 1.7 \times 10^{-15}$, which is comparable to the ratio in the limit (2.47). It may come as a surprise that dark matter does not have to be much colder

than baryonic matter.

2.5 Implications for dark matter models

In order to illustrate an application of our limit (2.47) to a particular class of dark matter scenarios, let us consider how it impacts the mass of an eventual fermionic dark matter candidate χ that only couples to the three species of (Dirac) neutrinos in the standard model. Dark matter couplings to photons, quarks or electrons are severely constrained by direct and indirect detection experiments [64–67], but due to their elusive nature, interactions with neutrinos are beyond the reach of most of these experiments. Neutrino telescopes do constrain direct dark matter annihilation into neutrinos, but only if dark matter is sufficiently massive [68–70]. It is in cases like this where cosmological limits like the one we derived turn out to be most powerful. Similar models and their cosmological implications have therefore been discussed in the literature: References [71–73] mostly focus, for instance, on the effects of dark matter interactions on cosmological observables, whereas references [32] and [74] explore whether late time kinetic decoupling could resolve some of the problems of CDM on small scales, and are therefore closely related to our analysis.

To proceed in a fairly model-independent way, let us assume that the coupling between χ and standard model neutrinos ν is universally described by one of the two effective four-fermion interactions

$$\mathcal{L}_{\rm int}^{S} = \frac{1}{\Lambda_S^2}\sum_i \bar{\chi}\chi\,\bar{\nu}_i\nu_i, \quad \mathcal{L}_{\rm int}^{V} = \frac{1}{\Lambda_V^2}\sum_i (\bar{\chi}\gamma^\mu\chi)(\bar{\nu}_i\gamma_\mu\nu_i), \tag{2.49}$$

where Λ_S and Λ_V are constants with dimensions of energy, and i runs over the three neutrinos species $i = e, \mu, \tau$. We expect this effective description to remain valid up to energies of order $E \sim \Lambda_{S,V}$, which, because dark matter particles are non-relativistic

leads us to impose

$$2m \lesssim \Lambda_{S,V}. \tag{2.50}$$

The coupling in $\mathcal{L}^S_{\rm int}$ is what we expect in any model in which interactions between dark matter and neutrinos are mediated by a heavy scalar of mass $m_{\rm scalar} \lesssim \Lambda_S$, whereas that in $\mathcal{L}^V_{\rm int}$ is what we expect from the mediation of a heavy gauge boson of mass $m_{\rm gauge} \lesssim \Lambda_V$. Such interactions are two of the possible couplings in the effective field theory approach to dark matter that is often used to constrain dark matter couplings. Of course, from an effective field theory approach there is no reason why dark matter should interact with neutrinos alone, but by the same token there are many properties of the standard model itself that cannot be explained in this framework.

This class of models has indeed been previously considered in the literature. Reference [32] for instance proposes a model in which dark matter decouples at late times because of the interactions between dark matter and neutrinos mediated by a heavy gauge boson, and analyzes whether the resulting suppression of structure due to free streaming could resolve some of the problems of the CDM scenario at small scales. Motivated by similar considerations, Shoemaker studies constraints on effective interactions between neutrinos and dark matter like those in equation (2.49), and how these affect the masses of the smallest proto-haloes in reference [74].

As we shall see, the limit (2.47) becomes particularly relevant for sufficiently light dark matter particles. In this case, χ decouples kinetically rather late in the history of the universe (after nucleosynthesis), which is what we shall assume in what follows. Although the neutrinos themselves decouple from the remaining standard model particles around nucleosynthesis, we assume that their interactions with dark matter particles maintain neutrinos and dark matter particles in thermal equilibrium until kinetic decoupling. Neutrino self-interactions also impact the CMB and the matter power spectrum [72, 73], but such an impact should be negligible as long as

kinetic decoupling takes place before observable scales enter the horizon.

The mass m may be related to the scales Λ_S and Λ_V if the couplings (2.49) determine the dark matter relic density. In the non-relativistic limit, the total thermally averaged dark matter annihilation cross section times relative velocity becomes, to lowest non-trivial order in the relative velocity,

$$\langle \sigma v_{\rm rel} \rangle_S = \frac{9}{4\pi} \frac{m^2}{\Lambda_S^4} \frac{T}{m}, \quad \langle \sigma v_{\rm rel} \rangle_V = \frac{3}{\pi} \frac{m^2}{\Lambda_V^4}, \tag{2.51}$$

where we have used the results in reference [75] to calculate the thermal average of the annihilation cross section times the relative velocity, and we assume that dark matter may annihilate into any of the three Dirac neutrino species. Chemical decoupling (freeze-out) then occurs at temperatures of order [76]

$$T^S_{\rm freeze} \approx m \left[41.1 + 3 \log \frac{m}{\rm GeV} - 4 \log \frac{\Lambda_S}{\rm GeV} - \frac{3}{2} \log \left(41.1 + 3 \log \frac{m}{\rm GeV} - 4 \log \frac{\Lambda_S}{\rm GeV} \right) \right]^{-1}, \tag{2.52a}$$

$$T^V_{\rm freeze} \approx m \left[40.7 + 3 \log \frac{m}{\rm GeV} - 4 \log \frac{\Lambda_V}{\rm GeV} - \frac{1}{2} \log \left(40.7 + 3 \log \frac{m}{\rm GeV} - 4 \log \frac{\Lambda_V}{\rm GeV} \right) \right]^{-1}, \tag{2.52b}$$

where we have set the effective number of relativistic degrees at freeze-out to be $g_* = 3.36$. Note that this equation only applies under the assumption that dark matter decouples while non-relativistic, and as long as our effective field theory remains valid, $2m \ll \Lambda_{V,S}$. At present, the corresponding relic density is [76]

$$\Omega^S_{\rm cdm} = 5.4 \cdot 10^{-10} \left(\frac{m}{T^S_{\rm freeze}} \right)^2 \left(\frac{\rm GeV}{m} \right)^2 \left(\frac{\Lambda_S}{\rm GeV} \right)^4, \tag{2.53a}$$

$$\Omega^V_{\rm cdm} = 2.0 \cdot 10^{-10} \left(\frac{m}{T^V_{\rm freeze}} \right) \left(\frac{\rm GeV}{m} \right)^2 \left(\frac{\Lambda_V}{\rm GeV} \right)^4. \tag{2.53b}$$

Because the dark matter fraction of the critical density $\Omega_{\rm cdm}$ is well constrained,

equations (2.53) can be used to express m in terms of Λ or vice-versa. Say, in the low-mass regime, the relations

$$\Lambda_S = 32 \left(\frac{m}{\text{GeV}}\right)^{0.48} \text{GeV}, \quad \Lambda_V = 89 \left(\frac{m}{\text{GeV}}\right)^{0.49} \text{GeV} \tag{2.54}$$

provide a good fit for the numerical solution of equations (2.53) with $\Omega_{\text{cdm}} = 0.23$. Note that several cosmic ray anomalies can be explained if dark matter couples to a gauge boson with a mass of order 10 GeV, which happens to be the scale suggested by the previous equations for sub-GeV dark matter particles. The explanation of these anomalies relies on the temperature-dependent enhancement of the annihilation cross section caused by an additional interaction mediated by the relatively light gauge boson. In the presence of such "Sommerfeld" enhancement, the dark matter annihilation cross section at present is thus decoupled from dark matter primordial abundance constraints [77]. Note however that there is no Sommerfeld enhancement as long as our effective field theory description of dark matter remains valid.

Even after chemical freeze-out, interactions between dark matter and standard model particles keep dark matter in thermal equilibrium, until they kinetically decouple later on. The kinetic decoupling temperature critically depends on the forward scattering amplitude between dark matter and standard model particles. For the interactions in (2.49), the spin-averaged square amplitudes for scattering between a non-relativistic WIMP and a relativistic neutrino are [67]

$$\frac{1}{4}\sum_{\text{spins}} |\mathcal{M}_S|^2 = \frac{16 m^2 m_\nu^2}{\Lambda_S^4}, \quad \frac{1}{4}\sum_{\text{spins}} |\mathcal{M}_V|^2 = \frac{16 m^2 E_\nu^2}{\Lambda_V^4}, \tag{2.55}$$

where E_ν is the neutrino energy in the frame in which dark matter is at rest, and, for simplicity, we assume that all neutrinos have the same mass m_ν (neutrino oscillations actually imply that the three neutrino masses are all different). Using the results of reference [45] and taking into account the fact that dark matter only couples to

neutrinos we then find that the decoupling temperature is

$$\frac{T_{\rm dec}^S}{m} = 0.23 \left(\frac{\rm GeV}{m}\right)^{1/2} \left(\frac{1\,\rm eV}{m_\nu}\right) \left(\frac{\Lambda_S}{\rm GeV}\right)^2, \tag{2.56a}$$

$$\frac{T_{\rm dec}^V}{m} = 8.3 \cdot 10^{-6} \left(\frac{\rm GeV}{m}\right)^{3/4} \left(\frac{\Lambda_V}{\rm GeV}\right). \tag{2.56b}$$

After kinetic decoupling, the dark matter temperature redshifts with the square of the scale factor. Hence, assuming adiabatic expansion, the dark matter temperature to mass ratio at present is

$$\frac{T_0}{m} = \left(\frac{4}{11}\right)^{2/3} \frac{T_\gamma^2}{m\, T_{\rm dec}}, \tag{2.57}$$

where T_γ is the current photon temperature and we have used the fact that at the time dark matter kinetically decouples from the neutrino background, its temperature is $(4/11)^{1/3}$ times smaller than that of the photons. With the decoupling temperature given by equations (2.56), equation (2.57) becomes

$$\frac{T_0^S}{m} \approx 1.2 \cdot 10^{-25} \left(\frac{\rm GeV}{m}\right)^{3/2} \left(\frac{m_\nu}{\rm eV}\right) \left(\frac{\rm GeV}{\Lambda_S}\right)^2, \tag{2.58a}$$

$$\frac{T_0^V}{m} \approx 3.4 \cdot 10^{-21} \left(\frac{\rm GeV}{m}\right)^{5/4} \left(\frac{1\,\rm GeV}{\Lambda_V}\right). \tag{2.58b}$$

Combining our limit (2.47) on the dark matter temperature today with equations (2.58) we finally obtain 95% CL lower bounds on the corresponding scale Λ,

$$\Lambda_S \geq 3.4 \cdot 10^{-6}\,\rm GeV \left(\frac{m_\nu}{\rm eV}\right)^{1/2} \left(\frac{\rm GeV}{m}\right)^{3/4}, \quad \Lambda_V \geq 3.2 \cdot 10^{-7}\,\rm GeV \left(\frac{\rm GeV}{m}\right)^{5/4}. \tag{2.59}$$

Recall however that the limit (2.47) holds only under the assumptions that dark matter decoupled while non-relativistic ($T_{\rm dec}/m \ll 1$) and before observable scales

had entered the horizon ($z_{\rm dec} > z_{\rm max}$). Because the decoupling redshifts are

$$z^S_{\rm dec} = 1.4 \cdot 10^{12} \left(\frac{m}{\rm GeV}\right)^{1/2} \left(\frac{1\,\rm eV}{m_\nu}\right) \left(\frac{\Lambda_S}{\rm GeV}\right)^2 , \tag{2.60a}$$

$$z^V_{\rm dec} = 4.9 \cdot 10^{7} \left(\frac{m}{\rm GeV}\right)^{1/4} \left(\frac{\Lambda_V}{\rm GeV}\right) , \tag{2.60b}$$

observable scales enter the horizon after decoupling if

$$\Lambda_S \geq 6.1 \cdot 10^{-4}\,\rm GeV \left(\frac{\rm GeV}{m}\right)^{1/4} \left(\frac{m_\nu}{\rm eV}\right)^{1/2}, \quad \Lambda_V \geq 1.0 \cdot 10^{-2}\,\rm GeV \left(\frac{\rm GeV}{m}\right)^{1/4}, \tag{2.61}$$

whereas the demand that dark matter decouple kinetically while non-relativistic leads to

$$\Lambda_S \leq 2.1\,\rm GeV \left(\frac{m}{\rm GeV}\right)^{1/4} \left(\frac{m_\nu}{\rm eV}\right)^{1/2}, \quad \Lambda_V \leq 1.2 \cdot 10^{5}\,\rm GeV \left(\frac{m}{\rm GeV}\right)^{3/4}. \tag{2.62}$$

Note that for our purposes it does not matter whether dark matter freezes out while relativistic or non-relativistic; in order for the dark matter distribution to be Maxwellian, it just suffices that dark matter decouples kinetically while non-relativistic.

As shown in Figures 2.6 and 2.7, equations (2.61) and (2.62), together with condition (2.50) define a wedge in parameter space in which dark matter can be considered to be cold and collisionless for observational purposes, and in which we can trust our effective field theory description. This is the region in parameter space in which our analysis holds, and for which our limit (2.47) applies. There may exist viable dark matter models beyond this shaded region, but these models must violate one of our assumptions, so, our analysis and limits do not apply to them.

Combining the region of parameter space for which our analysis holds with the lower limits on Λ, we find the fraction of parameter space excluded by observations, namely that region inside the shaded wedge that lies below the red line in the cor-

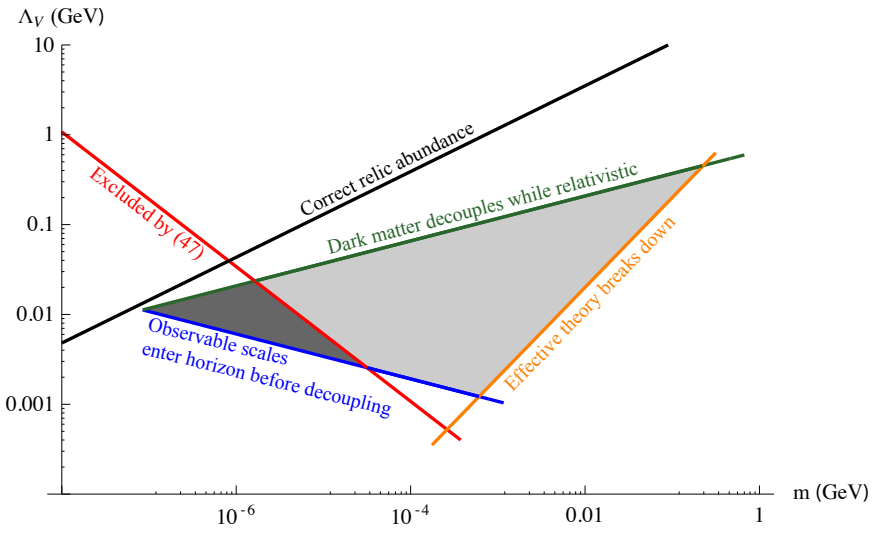


Figure 2.6: Constraints on the scalar interaction scale Λ_S in equation (2.49) as a function of the dark matter mass m (for a neutrino mass $m_\nu = 0.1\,\text{eV}$). The area under the dark green line corresponds to models in which dark matter decouples kinetically while non-relativistic [equation (2.62)], whereas the area above the blue line describes models in which dark matter decouples before observable modes have entered the horizon [equation (2.61)]. The area above the orange line corresponds to parameter choices in which we can trust the effective field theory [equation (2.50)]. Hence, the light shaded area describes models in which dark matter is cold and collisionless for practical purposes, and we can trust our calculation. Parameters under the red line [equation (2.59)] are incompatible with our limit (2.47), which excludes the dark shaded region at the 95% level. Along the black line [equation (2.54)], the scalar interaction leads to the observed dark matter relic abundance.

responding figure. We also plot in the corresponding figure equation (2.54), which determines the scale Λ required for the present dark matter density to agree with the observed one. The figures thus imply that in some of the models that explain the current dark matter density, dark matter is either not cold or collisionless, in disagreement with the cold-dark-matter paradigm. Of course, different interactions or mechanisms may be responsible for establishing the present dark matter density, which is in fact what needs to happen in the low mass regime; at $m \lesssim 10\,\text{MeV}$ equations (2.52) imply that dark matter is in chemical equilibrium while still relativistic (or nearly relativistic) during nucleosynthesis. Dark matter then significantly af-

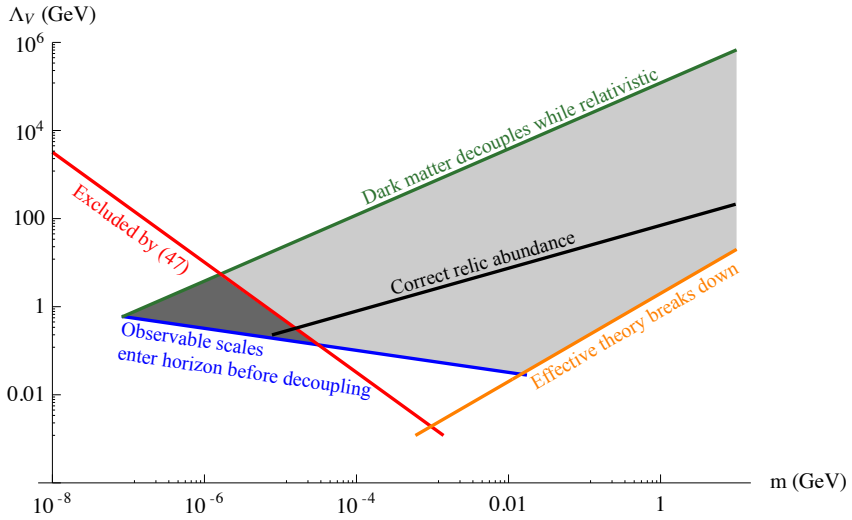


Figure 2.7: Constraints on the vector interaction scale Λ_V in equation (2.49) as a function of the dark matter mass m. See caption of Figure 2.6 for more details.

fects the expansion rate during that time, thus modifying the predicted light element abundances, in conflict with observations [78]. As we discussed above, Sommerfeld enhancement [77] can drastically alter the primordial abundance of dark matter particles. In this case, one would also need to study how Sommerfeld enhancement affects the thermal evolution of dark matter particles [79–81].

Figures 2.6 and 2.7 are the analogues of the mass vs. scattering cross section exclusion plots derived from direct dark matter search experiments (see e.g. [82]). Note however that our limit reaches down to much lighter dark matter masses, of order of a keV. Indeed, inspection of Figures 2.6 and 2.7 reveals that our constraint imposes absolute lower mass limits on cold and collisionless dark matter models that interact according to equations (2.49). In fact, equation (2.47) yields a lower limit on the dark matter mass which is essentially independent of the scattering dark matter scattering rate. In the case at hand, for instance, from equations (2.59) and (2.62) we obtain a lower mass limit

$$m \geq 1.6\,\mathrm{keV}, \tag{2.63}$$

which does not depend on the interaction type (vector or scalar), and is also neutrino mass independent. Such a mass limit is comparable to those derived in the context of warm dark matter models [31]. As constraints on the matter power spectrum tighten, we expect our lower limits on Λ (the red line in the figure) to move up in the exclusion plots, thus ruling out larger portions of parameter space, and further increasing the lower dark matter mass limit.

2.6 Summary and Conclusions

In the currently accepted cosmological model, ΛCDM, dark matter is a pressureless and non-interacting perfect fluid. Although such a model is sufficient to capture the properties of our universe on large scales, it does not really address the nature of dark matter. The simplest explanation of these dark matter properties postulates that the latter consists of sufficiently cold non-relativistic and non-interacting particles, an assumption that is often taken to be part of the ΛCDM model itself.

In this Chapter, we have addressed how cold dark matter particles would have to be in order to be compatible with large scale structure observations. In a wide variety of dark matter models, dark matter is in thermal equilibrium in the early universe, and its distribution remains thermal at least until the time when non-linear structures form. As long as dark matter particles decoupled while non-relativistic, we expect their momentum distribution to be Maxwellian, with a sufficiently low temperature T. A measure of how cold dark matter is stems from its temperature to mass ratio T/m, which for non-interacting and non-relativistic particles redshifts with the square of the scale factor. This ratio determines the root-mean-square velocity of dark matter particles, and the ratio of dark matter pressure to energy density, which is of order T/m. Hence, dark matter is cold as long as $T/m \ll 1$.

A non-zero dark matter temperature implies a non-zero velocity dispersion of its constituents. Such a non-zero velocity leads to dark matter free-streaming, which

tends to erase structure on length scales smaller than the corresponding free-streaming length. The absence of such suppression in the matter power spectrum on the smallest scales accessible to linear perturbation theory thus allows us to place limits on the dark matter temperature to mass ratio at present. Indeed, combining cosmic microwave background and large scale structure observations, down to the scales probed by Lyman alpha forest observations, we derive the 95% credible limit on the extrapolated present dark matter to temperature ratio,

$$\frac{T_0}{m} \leq 1.07 \cdot 10^{-14}. \tag{2.64}$$

This limit only applies within the cold-dark-matter paradigm, but is otherwise fairly model-independent. It assumes that since the time the smallest observable scales enter the horizon, dark matter can be described by an ensemble of non-interacting particles with a Maxwellian momentum distribution, whose temperature remains non-relativistic until today. Whether these particles are point-like or have a finite extent does not affect our limit, as long as the size of dark matter particles is much smaller than the scales probed by the cosmological observations. The limit also implies that dark matter had to be quite cold already at the time galactic scales entered the horizon, thus supporting the assumptions made in its derivation, and placing the cold dark matter scenario within quantitative boundaries.

The limit (2.64) does not constrain typical WIMP scenarios very tightly. Say, for neutralinos with $m \sim 100\,\mathrm{GeV}$ the kinetic decoupling temperature can be as low as $T_{\mathrm{dec}} \sim 10\,\mathrm{MeV}$ [35], implying $T_0/m \lesssim 10^{-24}$, far away from our limit. On the other hand, (2.64) allows us to constrain dark matter models that are otherwise unconstrained by direct or indirect dark matter searches. Say, if dark matter only couples to standard model neutrinos through a four-fermion scalar or vector interaction, the limit (2.64) implies that the dark matter mass has to be heavier than 1 keV. Improved constraints on the matter power spectrum should tighten the limit (2.64) and further rule out portions of parameter space in this and other classes of models.

Appendix

2.A Numerical Implementation

We have modified the publicly available Boltzmann integrator CAMB to take into account the effects of a non-zero dark matter temperature on the evolution of structure in the linear regime. Instead of pursuing the conventional expansion of the distribution function in multipoles (see e.g. [83]), we follow the approach described in Section 2.2. As a consequence, we simply need to evaluate the dark matter density perturbation (2.22) and the velocity perturbation (2.24) numerically (these suffice to determine the evolution of the metric potentials h and η.) Inspection of equations (2.22) and (2.24) quickly reveals that in order to calculate $\delta\rho$ and v we need the integrals

$$H_n \equiv \int_{\tau_{\rm dec}}^{\tau} d\tau' \exp\left(-\frac{k^2 d^2(\tau,\tau')}{2}\right) [d(\tau,\tau')k]^n \, h'(\tau'), \tag{2.65a}$$

$$E_n \equiv \int_{\tau_{\rm dec}}^{\tau} d\tau' \exp\left(-\frac{k^2 d^2(\tau,\tau')}{2}\right) [d(\tau,\tau')k]^n \, \eta'(\tau'), \tag{2.65b}$$

for values of n ranging from zero to four. These integrals obey the recursion relations

$$\frac{dH_n}{d\tau} = \delta_{n0} h' + k\sqrt{\frac{T}{m}}(nH_{n-1} - H_{n+1}), \tag{2.66a}$$

$$\frac{dE_n}{d\tau} = \delta_{n0} \eta' + k\sqrt{\frac{T}{m}}(nE_{n-1} - E_{n+1}), \tag{2.66b}$$

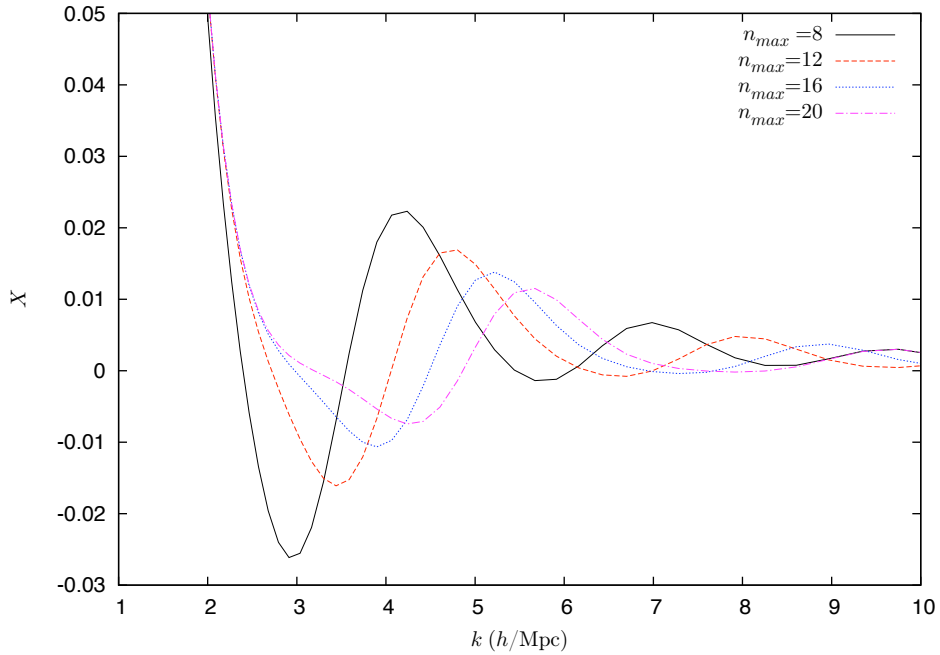


Figure 2.A.1: A plot of the transfer function ratio (2.40) at $z = 0$ obtained for different number of equations in the hierarchy $n_{\max}$ at a fixed temperature $T_0/m = 10^{-12}$.

which, unfortunately, lead to an infinite hierarchy of coupled differential equations. We choose to truncate the hierarchy at $n_{\max} = 12$. This is a good approximation if $d\,k$ remains small, but fails when $d\,k$ becomes sufficiently large (see Figure 2.A.1).

In order to determine the values of $k\,T_0/m$ where our finite $n_{\max}$ approximation works, we plot the suppression factor (2.40) at $z = 0$ as a function of comoving scale k for different values of $n_{\max}$, as in Figure 2.A.1. As seen in the figure, all the suppression factors agree at $k \leq 2h/\text{Mpc}$, but disagree at larger k. The smaller the $n_{\max}$, the earlier the corresponding curve starts to disagree from the remaining curves. Given the structure of these curves, we infer that our approximation can be trusted at $k \lesssim 2h\,\text{Mpc}^{-1}$ and $T_0/m = 10^{-12}$ for $n_{\max} \geq 12$. Conversely, since we fix $n_{\max} = 12$, and because our expansion parameter $k\,d$ is proportional to $\sqrt{T_0/m}$, we

conclude that our approximation is valid as long as

$$\sqrt{\frac{T_0}{m}} \leq 10^{-6} \times \frac{2h\,\mathrm{Mpc}^{-1}}{k_{\mathrm{max}}} \tag{2.67}$$

Because we are interested in the constraints imposed by the Lyman-α forest, we want to make sure that our numerical results are accurate up to scales of order $k_{\mathrm{max}} = 2h\mathrm{Mpc}^{-1}$. We therefore impose the conservative prior (2.46).

In order to solve the system of equations (2.66) we need to specify initial conditions during radiation domination, at a time when the corresponding mode is well outside the horizon. We calculate the appropriate value of H_n and E_n by integrating equations (2.65) analytically, using expressions (2.79) for the gravitational potentials. As we argue in Appendix 2.B, as long as the corresponding mode is outsize the horizon, the integrals do not depend on the lower integral limit τ_{dec}, which we hence set to zero. Under these assumptions we find

$$H_n(\tau) = -\mathcal{R}_i \frac{n!}{\sqrt{2^n}} \frac{m}{T} U\left(1 + \frac{n}{2}, \frac{3}{2}, \frac{m}{T}\frac{2}{k^2\tau^2}\right) + \cdots, \quad E_n(\tau) = -\frac{5+4R_\nu}{12(15+4R_\nu)} H_n(\tau) + \cdots. \tag{2.68}$$

where U is the confluent hypergeometric function. Initial conditions on H_n and E_n are thus determined by evaluation of equations (2.68) at an appropriate initial time τ_i, chosen so that $k\,\tau_i \ll 1$. Note that in the radiation dominated era, $k^2\tau^2 T/m$ is time-independent.

2.B Calculation of $\delta\rho$

Our goal is to calculate the perturbed components of the energy-momentum tensor. As noted in reference [56], because $\delta T^0{}_0$, $\delta T^0{}_i$ and $\delta T^i{}_j$ respectively transform as a scalar, vector and tensor under spatial diffeomorphisms, and because there are no non-trivial functions of the spatial metric with these transformation properties,

metric perturbations do not contribute, and we can focus on the contributions from δf. Below we shall also argue that the contributions from $\delta f(\tau_{\rm dec})$ to the energy-momentum tensor are negligible, so we shall omit them in what follows.

We begin by calculating $\delta\rho \equiv -\delta T^0{}_0$. To do so, we substitute the perturbed distribution function (2.18) into the expression for the energy momentum (2.2),

$$\delta\rho(\tau) = -\frac{1}{a^4}\int d^3p\, p_0\, \frac{\bar{f}'}{2p}\int d\tau' p_i p_j\, h'_{ij}(\tau')\exp\left(-i\Delta\,\vec{p}\cdot\vec{k}\right), \tag{2.69}$$

where Δ is defined in equation (2.19), and a sum over repeated indices (regardless of location) is implied. We are interested here in the non-relativistic limit, so we just need to calculate this expression to first order in T/m. Because the root mean square momentum is $\sqrt{3mT}$, at this order we need to expand the covariant momentum p_0 in the integrand to quadratic order in p/m.

$$p_0 = -a\sqrt{m^2+\frac{p^2}{a^2}} \approx -a\,m\left(1+\frac{p^2}{2a^2m^2}+\cdots\right). \tag{2.70}$$

For the same reason, it suffices to expand p_0 inside Δ to zeroth order, since the exponential in the integrand is already linear in p/m. We could then subsequently expand that exponential to quadratic order in p, but because the exponent includes terms that may become large on small scales (large k), we do not follow this route. Instead, we treat the exponential exactly in an expansion in powers of k.

We begin by carrying out the integrals over the angular part of the covariant momentum variable $\vec{p}$, which we compute using the formula

$$\int d^2p\; p_i p_j \exp\left[-i\vec{p}\cdot\vec{v}\right] = 4\pi p^2\left[\delta_{ij}\frac{j_1(v)}{v} - \hat{v}_i\hat{v}_j j_2(v)\right], \tag{2.71}$$

where a hat denotes unit vector in the corresponding direction and j_1 and j_2 are spherical Bessel functions of the first kind. We then proceed to evaluate the integral

over the magnitude of the momentum,

$$\int_0^\infty dp\,\frac{\bar{f}'}{2p}\left(p^4+\frac{p^6}{2a^2m^2}+\cdots\right)\left[\delta_{ij}\frac{j_1\left(\Delta\cdot p\cdot k\right)}{\Delta\cdot p\cdot k}-\hat{k}_i\hat{k}_j\,j_2\left(\Delta\cdot p\cdot k\right)\right], \tag{2.72}$$

which has a closed analytic form because for the Maxwell-Boltzmann distribution (2.6), $\bar{f}'/p$ is a Gaussian. To calculate the integral in the last equation it suffices to know the generating functions

$$\int_0^\infty dp\,p^3\exp\left(-\frac{p^2}{2mT_0}\right)\frac{j_1(\Delta\,k\,p)}{\Delta\,k}=\sqrt{\frac{\pi}{2}}(mT_0)^{5/2}\exp\left(-\frac{mT_0\,\Delta^2k^2}{2}\right), \tag{2.73a}$$

$$\int_0^\infty dp\,p^4\exp\left(-\frac{p^2}{2mT_0}\right)j_2(\Delta\,k\,p)=\sqrt{\frac{\pi}{2}}(mT_0)^{7/2}\exp\left(-\frac{mT_0\,\Delta^2k^2}{2}\right), \tag{2.73b}$$

from which integrals with higher even powers of p can be calculated by formal differentiation with respect to $\alpha\equiv 1/(mT_0)$. In particular, every additional power of p^2/m^2 in our non-relativistic expansion yields relative corrections of order T_0/m and $(T_0/m)(mT_0\Delta^2k^2)$. Hence, such an expansion is justified provided that both quantities are small.

The structure of the exponential in equations (2.73) suggests that we define $d^2\equiv mT_0\Delta^2$, which in the non-relativistic limit results in the definition of the streaming length in equation (2.20). Thus, the analysis in the last paragraph reveals that our approximations are valid as long as $T/m\ll 1$ and $(T_0/m)k\,d\ll 1$. Using equations (2.73) in (2.72) and substituting into (2.71) we finally arrive at equation (2.22). The derivation of equations (2.24), (2.25) and (2.26) is completely analogous.

It is also illustrative to check how the specific form of the distribution function affects these results. Instead of the Maxwell-Boltzmann distribution (2.6), let us assume that $\bar{f}$ is instead [51]

$$\bar{f}=\exp(-p/T_0), \tag{2.74}$$

which is the high-momentum limit of both the Fermi and Bose distributions. In this

case, the generating functions analogous to those in equation (2.73) are

$$\int_0^\infty dp\, p^2 \exp\left(-\frac{p}{T_0}\right) \frac{j_1(\Delta\, k\, p)}{\Delta\, k} = 2T_0^4 \frac{1}{(1+T_0^2\Delta^2 k^2)^2}, \tag{2.75}$$

$$\int_0^\infty dp\, p^4 \exp\left(-\frac{p}{T_0}\right) j_2(\Delta\, k\, p) = 8T_0^6 \frac{5 - T_0^2\Delta^2 k^2}{(1+T_0^2\Delta^2 k^2)^4}. \tag{2.76}$$

The structure of these integrals thus suggests identifying the free streaming length with $d = T_0\Delta$, which differs from the definition in equation (2.20). In both cases, however, the free streaming length is proportional to the root mean square velocity of the particles, namely, $v_{\rm rms} \sim \sqrt{T_0/m}$ for a Maxwell-Boltzmann distribution, and $v_{\rm rms} \sim T_0/m$ for the distribution (2.74). On length scales smaller than this free-streaming length, there is again a suppression of structure, but instead of exponential, as in (2.73), the suppression here is polynomial.

To conclude this section we still need to show that the contributions to the energy-momentum tensor of the term $\delta f(\tau_{\rm dec})$ in equation (2.18) are negligible. Consider for that purpose the energy density. Let us assume that dark matter decouples with a thermal distribution with vanishing chemical potential while (mildly) non-relativistic, as in WIMP models. Then, at or shortly before decoupling we have

$$\delta f(\tau_{\rm dec}, \vec{k}, \vec{p}) \approx \frac{1}{(2\pi)^3} \exp\left[-\frac{m}{T_{\rm dec}} - \frac{\vec{p}^{\,2}}{2mT_{\rm dec}a_{\rm dec}^2}\right] \frac{\delta T(\tau_{\rm dec}, \vec{k})}{T_{\rm dec}}. \tag{2.77}$$

Substituting this expression into equation (2.2) and integrating over momenta as before we find

$$\delta\rho = \bar{\rho}\frac{m}{T_{\rm dec}}\frac{\delta T_{\rm dec}}{T_{\rm dec}}\left[1 + \frac{1}{2}\frac{T_{\rm dec}}{m}\left(1 + \frac{a_{\rm dec}^2}{a^2}\right)\left(3 - d^2(\tau, \tau_{\rm dec})k^2\right)\right] \exp\left(-\frac{1}{2}d^2(\tau, \tau_{\rm dec})k^2\right), \tag{2.78}$$

which again shows the expected exponential suppression of structure due to free streaming.

For adiabatic perturbations we expect the temperature perturbation to be of the

same order as the density perturbation in photons $\delta T_{\rm dec}/T_{\rm dec} = \delta_\gamma/4$. On super-horizon scales the latter and the gravitational potentials are given by

$$\eta = -\zeta_i \left(1 - \frac{5 + 4R_\nu}{12(15 + 4R_\nu)} k^2\tau^2 + \cdots\right), \quad h = -\frac{\zeta_i}{2} k^2\tau^2 + \cdots, \quad \delta_\gamma = \frac{\zeta_i}{3} k^2\tau^2 + \cdots. \tag{2.79}$$

Hence, because the time derivatives of the potentials grow linearly in time, the contribution from the integral in equation (2.22) is always larger than that of equation (2.78), as long as the corresponding mode was super-horizon sized at decoupling. For the same reason, and under the same assumption, the integral in equation (2.22) is not very sensitive to the time of decoupling, which can be taken to be zero.

Equation (2.78) also illustrates an important property of an ensemble of collisionless particles. In the absence of gravity, equation (2.78) is the energy density associated with the solution of Boltzmann's equation with the corresponding initial conditions. In the case of a perfect fluid with a non-zero pressure, we would expect such a solution to describe acoustic oscillations, but equation (2.78) instead shows exponential decay on scales smaller than the free-streaming length if the gas is collisionless. If we had carried out an analogous calculation with massless particles, we would have found that the energy density is proportional to $j_0[(\tau - \tau_{\rm dec})k]$. This function does in fact oscillate in time (albeit with a decaying amplitude), but the corresponding frequency is $\omega = k$, instead of $\omega = k/\sqrt{3}$, the acoustic oscillation frequency for a fluid of relativistic particles.

Chapter 3

Structure Formation Constraints on Sommerfeld-enhanced dark matter annihilation

3.1 Introduction

We have discussed in the introductory chapter to the book, as well as in the previous chapter, that the simplest way to explain the properties of the dark matter fluid is to assume that it consists of non-interacting and non-relativistic particles. In this scenario, the amount of dark matter in our universe is a free parameter that has to be chosen to fit observations and thus remains unexplained. On the other hand, if dark matter particles are assumed to self-annihilate with an averaged cross section times relative velocity of the order of the weak scale,

$$\langle \sigma v \rangle_w \equiv 3 \cdot 10^{-26} \mathrm{cm}^3\,\mathrm{s}^{-1}, \tag{3.1}$$

dark matter particles decouple from radiation in the early universe while being non-relativistic, with an abundance that roughly fits the observed amount of dark matter,

$$\Omega_c h^2 \approx 0.1 \frac{\langle \sigma v \rangle_w}{\langle \sigma v \rangle}. \tag{3.2}$$

This equation holds regardless of the precise value of the dark matter mass and the particles dark matter annihilates into. In this scenario, we not only explain the major properties of dark matter, but also its amount. This is why weakly interacting massive particles (WIMPs) are widely believed to be the dark matter constituents.

But somewhat recently, motivated by certain anomalies in cosmic ray spectra [84–86], several authors have suggested that the dark matter self-annihilation rate today may differ from the rate suggested by equation (3.2) [77, 87, 88]. If f is the fraction of the energy deposited into standard model particles by two annihilating dark matter wimps, these models require [87, 89]

$$f \cdot \langle \sigma v \rangle \sim 10^2 \langle \sigma v \rangle_w \tag{3.3}$$

for a WIMP of mass $m \sim 1\,\mathrm{TeV}$. Therefore, in order to preserve the successful postdiction of the dark matter abundance, these authors have suggested that the dark matter annihilation rate is inversely proportional to the dark matter velocity, and thus increases as the universe expands and the velocity redshifts.

$$\langle \sigma v \rangle \propto \frac{1}{v}. \tag{3.4}$$

A simple way to accomplish such an increase involves the Sommerfeld enhancement of the annihilation cross section induced by a new, sufficiently light force carrier [90, 91].

Recombination places quite stringent constraints on the annihilation cross section of the enhanced dark matter models. If dark matter efficiently annihilates into radiation during recombination, the injection of this radiation into the plasma sig-

nificantly affects the temperature anisotropies in the cosmic microwave background radiation. Using this effect, several groups have been able to place an upper limit on the thermally averaged annihilation rate times velocity during recombination [92–94],

$$\langle \sigma v \rangle \leq 15 \frac{\langle \sigma v \rangle_w}{f} \frac{mc^2}{\text{TeV}} \quad \text{at} \quad 95\% CL. \tag{3.5}$$

On the face of this limit, models that explain cosmic ray anomalies with enhanced annihilation cross section are already ruled out.

Unfortunately, the limit on the annihilation cross section (3.5) depends on the model-dependent parameter f, which can vary by several orders of magnitude. In those (nearly ruled out) models that attempt to explain the cosmic ray anomalies mentioned above, f is of order one, whereas in models in which dark matter is part of a dark sector that interacts only gravitationally with the standard model, f vanishes. In extreme cases like the latter, the limit (3.5) is not very useful.

In this Chapter, we set limits on the dark matter annihilation cross section that do not depend on f, and thus apply to a wider class of dark matter candidates, beyond those designed to address the aforementioned cosmic ray anomalies. Our constraints are based on the impact of dark matter annihilation on the formation and growth of large-scale structure, including the cosmic microwave background anisotropies and the distribution of dark matter. Because the presence of additional force carriers in the dark sector still remains well-motivated, regardless of the dark matter annihilation channels, and because models with enhanced annihilation cross section provide distinct phenomenological signatures we focus on dark matter that self-annihilates into dark radiation with a Sommerfeld-enhanced cross section (several specific models in this class have been studied for instance in [95, 96].) Our dark radiation is assumed to not interact with standard model particles, which corresponds to the limit $f = 0$ in the class of models discussed above. Hence, any imprint of annihilation on cosmic observables must come from either the suppressed growth of dark matter structures, or from the gravitational interactions of its annihilation products, which are present

in any scenario in which dark matter self-annihilates.

For negliglible values of f, one can also derive quite stringent constraints on the self-scattering cross-section of dark matter (which should also experience Sommerfeld enhacement), because the latter would cause the central cores of gravitational bound astrophysical systems to become spherical, rather than elliptical, in conflict with observations [79, 97]. Unfortunately however, there is no model-independent relation between the scattering and annihilation cross sections, so these constraints cannot be directly applied to self-annihilation. In addition, these constraints only limit the scattering cross section at velocities of the order found in the corresponding dark matter halo. In contrast, our limits on annihilation do not depend on the dark matter velocity, and only rely on the assumption that dark matter is non-relativistic.

Our considerations of dark matter annihilation with a Sommerfeld-enhanced cross section are further motivated by two seemingly unrelated phenomenological problems. On one hand, it has been argued for some time that in the standard ΛCDM model the central densities of dark matter haloes, and the number of small subhaloes, do not appear to match observations [98, 99], although this eventual disagreement may have conventional astrophysical explanations [100, 101]. A natural way to explain the discrepancy is to assume that dark matter interacts or annihilates with a cross section that is inversely proportional to the dark matter velocity, as in Sommerfeld-enhanced models [32, 102], or simply to assume that dark matter self-annihilates with cross section larger than that required by equation (3.2) [103]. On the other hand, it has also been noticed that cosmic microwave data seem to indicate an additional relativistic dark component that interacts only gravitationally with the standard model (see for instance [104–107]). It is thus worthwhile to investigate whether this additional radiation could originate from dark matter annihilation, a circumstance that would link these two apparently unrelated problems.

In the context of the original Sommerfeld-enhancement models designed to explain the cosmic ray anomalies, our limits can be used for instance to determine the values

of f for which the effects of dark matter annihilation on structure formation have to be taken into account. In the general case, they help further constrain the properties of the yet to be identified dark matter particle, and, eventually, may explain the origin of the dark radiation hinted at by cosmic microwave data.

3.2 Annihilating Dark Matter

As we mentioned in the introduction, for our purposes dark matter is well described by a pressureless perfect fluid, with energy momentum tensor

$$T^{(c)}_{\mu\nu} = \rho_c u^{(c)}_\mu u^{(c)}_\nu, \tag{3.6}$$

where ρ_c is the energy density of dark matter, and $u^\mu_{(c)}$ its four-velocity, $g^{\mu\nu}u^{(c)}_\mu u^{(c)}_\nu = -1$. By assumption, the pressure of dark matter vanishes. In Appendix 3.A we link this perfect fluid description to a kinetic description, in which dark matter is regarded as an ensemble of non-relativistic particles. Our goal is to study the effects of dark matter annihilation on the growth of structure. For simplicity, we assume that dark matter annihilates into relativistic particles that interact only gravitationally with the standard model, but interact sufficiently rapidly with other particles in the dark sector (or themselves) to justify a perfect fluid approximation on the scales of interest. In that sense, the behavior of dark radiation mimics the behavior of photons prior to recombination. We shall thus regard the dark matter annihilation products as a perfect relativistic dark fluid, with energy-momentum tensor

$$T^{(d)}_{\mu\nu} = (\rho_d + p_d)u^{(d)}_\mu u^{(d)}_\nu + p_d\, g_{\mu\nu}, \quad \text{where} \quad p_d = \frac{\rho_d}{3}. \tag{3.7}$$

As it turns out, present cosmic microwave anisotropy data suggest the existence of such an additional relativistic species (see for instance [107]). Our dark radiation provides a natural candidate for this additional relativistic component for three reasons:

i) Since dark matter is negligible during early radiation domination, its annihilation products are unlikely to conflict with the successful predictions of big-bang nucleosynthesis. *ii*) Cosmic microwave anisotropy data probe times during which the amount of dark matter was sizable. *iii*) As we shall see, with a Sommerfeld-enhanced annihilation cross section, dark matter does not entirely freeze out at early times, but keeps annihilating until the dark-matter dominated era. Note that studies suggesting the presence of an additional dark relativistic species typically model this radiation as collisionless (neutrino-like) [104–107]. Although in this case a hydrodynamical description breaks down at small scales, this difference in description should not have much of an impact on cosmological observables, because dark radiation is not visible and never becomes the dominant component of the universe.

In the absence of particle number violating interactions, the energy-momentum tensor of dark matter is covariantly conserved, but in the presence of annihilation, dark matter particles transfer energy to its annihilation products. To determine the energy lost by the dark matter fluid, we rely on the kinetic description of Appendix 3.A, which yields

$$\nabla_\nu T^{(c)\nu}_\mu = -\frac{\langle\sigma v\rangle}{m}\rho_c^2 u^{(c)}_\mu. \tag{3.8a}$$

Here, $\langle\sigma v\rangle$ is the average dark matter annihilation cross section times relative velocity defined in equation (3.50), and m is the dark matter particle mass. Note that the rates at which energy and momentum are lost are inversely proportional to m, because the annihilation rate is proportional to the square of the number density, and the energy density is proportional to the mass m. The energy lost by the dark matter fluid due to annihilation is gained by the dark radiation fluid, so

$$\nabla_\nu T^{(d)\nu}_\mu = +\frac{\langle\sigma v\rangle}{m}\rho_c^2 u^{(c)}_\mu. \tag{3.8b}$$

In this way, the combined energy-momentum tensor of cold dark matter and dark radiation remains covariantly conserved.

3.2.1 Background Evolution

We turn our attention now to the evolution of the dark matter and dark radiation energy densities in an unperturbed, spatially flat FLRW universe,

$$ds^2 = a^2(\tau)\left[-d\tau^2 + d\vec{x}^{\,2}\right]. \tag{3.9}$$

The equations of motion of dark matter and dark radiation are given by the time components of equations (3.8a) and (3.8b), with the four velocities of dark matter and dark radiation taken to be $u^\mu = \delta^\mu{}_0/a$,

$$\rho_c' + 3\mathcal{H}\rho_c = -\frac{\langle\sigma v\rangle}{m}\rho_c^2 a, \tag{3.10a}$$

$$\rho_d' + 4\mathcal{H}\rho_d = +\frac{\langle\sigma v\rangle}{m}\rho_c^2 a. \tag{3.10b}$$

We have defined $\mathcal{H} = a'/a$, and a prime denotes a derivative with respect to conformal time τ. In a spatially flat universe we are free to choose the value of a today, which we set to one. Equations (3.10) hold after the kinetic decoupling of dark matter, which typically occurs well before nucleosynthesis [80, 108].

If the evolution of the scale factor is known, equations (3.10) can be readily integrated to give the evolution of dark matter and dark radiation,

$$\rho_c = \frac{\rho_c^i a_i^3}{a^3}\left(1 + \rho_c^i a_i^3 \int_{\tau_i}^{\tau} d\tau' \frac{\langle\sigma v\rangle}{m}\frac{1}{a^2}\right)^{-1} \tag{3.11a}$$

$$\rho_d = \frac{1}{a^4}\int_{\tau_i}^{\tau} d\tau' \frac{\langle\sigma v\rangle}{m}\frac{(\rho_c^i a_i^3)^2}{a}\left(1 + \rho_c^i a_i^3 \int_{\tau_i}^{\tau'} d\tau'' \frac{\langle\sigma v\rangle}{m}\frac{1}{a^2}\right)^{-2}, \tag{3.11b}$$

where ρ_c^i, a_i and τ_i are integration constants, and we have assumed that sufficiently early, at $\tau = \tau_i$, the amount of dark radiation is negligible, $\rho_d^i = 0$. With this choice, dark radiation only originates from dark matter annihilation; a non-zero value of ρ_d^i would lead to an additional contribution to the dark radiation density that may or may not have originated from the former.

In this Chapter, we mostly concentrate on the regime in which Sommerfeld enhancement operates, when, according to the discussion in Appendix 3.A, the averaged relative velocity between dark matter particles $v_{\rm rel}$ lies in the appropriate interval next to equation (3.43). We therefore assume that dark matter annihilates into dark radiation with a Sommerfeld-enhanced cross section, which, according to equation (3.61), is proportional to the scale factor,

$$\frac{\langle \sigma v \rangle}{m} = \Gamma a, \tag{3.12}$$

with constant Γ. Although the times at which Sommerfeld enhancement operates are strongly model-dependent, we note that dark matter particles typically decouple from the thermal bath well before big-bang nucleosynthesis, so we expect their velocities to be below the velocity v_0 introduced in the Appendix, certainly by nucleosynthesis. In our numerical solutions, we therefore assume that Sommerfeld enhancement is already operating at an initial scale factor $a_i = 10^{-10}$.

Clearly, in the presence of annihilation, the density of dark matter decreases faster than it otherwise would. For a cross section of the form (3.12), the density of dark matter during radiation domination is, for instance,

$$\rho_c = \rho_c^i \left(\frac{a_i}{a}\right)^3 \left(1 + \frac{\langle \sigma v \rangle_i}{m} \frac{\rho_c^i}{H_i} \log \frac{a}{a_i}\right)^{-1}, \tag{3.13}$$

where, again, the index i denotes the initial value of the corresponding quantity. In contrast to what happens in the conventional freeze-out scenarios, the correction factor proportional to $\langle \sigma v \rangle_i$ slowly varies for $a \gg a_i$, suggesting that annihilation keeps operating during radiation domination. Also note that the dark matter density diverges at a scale factor $a < a_i$. Of course, at early times our description of cold dark matter ceases to valid, because at sufficiently high densities we are not supposed to ignore inverse annihilations and other processes responsible for keeping the dark matter density in local thermal equilibrium.

To proceed with our analysis, we assume that the annihilation cross section is sufficiently small. On general grounds, we expect the quantitative effects of annihilation to be controlled by the relative change in the number of particles in a comoving volume during a Hubble time,

$$R \equiv -\frac{1}{\mathcal{H}} \frac{d \log(a^3 \rho_c)}{d\tau} = \frac{\langle \sigma v \rangle}{m} \frac{\rho_c}{H}. \tag{3.14}$$

This is, for instance, the case in equation (3.13), in which this factor appears explicitly in the correction to the energy density. Therefore, $\langle \sigma v \rangle$ is small if R remains much smaller than one throughout cosmic history. In that case, it is enough to calculate the impact of annihilation on any cosmological variable just to first order in $\langle \sigma v \rangle$. Note that to leading order in $\langle \sigma v \rangle$, R is constant during radiation domination, and proportional to $a^{-1/2}$ during matter domination.

To see how this works, consider for instance the amount of dark radiation. Neglecting the higher order correction in the denominator of the integrand in (3.11b) we find

$$\rho_d \approx \frac{\langle \sigma v \rangle}{m} \rho_c^2 a \cdot (\tau - \tau_i). \tag{3.15}$$

This equation shows that in this limit the amount of dark radiation does not depend on τ_i for $\tau \gg \tau_i$, and that ρ_d actually scales like non-relativistic matter instead of radiation. In the same limit, the fraction of the total radiation in the dark form during radiation domination is

$$\frac{\rho_d}{\rho_r} \approx \frac{\langle \sigma v \rangle}{m} \frac{\rho_c}{H} \frac{\Omega_c}{1 - \Omega_c}, \tag{3.16}$$

showing that for an R of order one, the amount of dark radiation is negligible during big-bang nucleosynthesis, but becomes sizable, about 10% at redshifts of about $z \approx 5 z_{\rm eq}$, where $z_{\rm eq}$ is the redshift of matter-radiation equality. This is relevant because scales entering the horizon at that time are probed by cosmic microwave temperature

multipoles of about $\ell \approx 700$, which roughly corresponds to the region probed by WMAP cosmic microwave anisotropy data [104].

Equations (3.11) are useful during radiation domination, when the scale factor is explicitly known. In order to determine how the energy density of dark matter evolves during matter domination, we introduce the scale factor a as a time variable in equation (3.10a). To integrate the resulting expression we use Friedmann's equation, neglecting both standard and dark radiation. The solution is

$$\rho_c \approx \rho_c^i \left(\frac{a_i}{a}\right)^3 \left[1 + \frac{\langle\sigma v\rangle_i}{m}\frac{\rho_c^i}{H_i}\left(1 - \frac{a_i^{1/2}}{a^{1/2}}\right)\right]^{-2}, \tag{3.17}$$

where the index i denotes the value of the corresponding quantity at an arbitrary scale factor a_i. Therefore, as opposed to what happens during radiation domination, dark matter freezes out at $a \gg a_i$, when its density decays as in the absence of annihilation. From equation (3.15), the amount of dark radiation is simply

$$\rho_d = 2\frac{\langle\sigma v\rangle}{m}\frac{\rho_c^2}{H}. \tag{3.18}$$

The time of matter-radiation equality depends on $\langle\sigma v\rangle$, because both the amount of dark matter and dark radiation depend on the latter. Since R is proportional to $a^{-1/2}$ during matter domination, we do not expect the values of $\langle\sigma v\rangle$ long after matter-radiation equality to significantly affect cosmological observables, even under the assumption that $\langle\sigma v\rangle$ has been growing with the scale factor since that time. This is important because, as we discuss in appendix 3.A, $\langle\sigma v\rangle$ should become constant at late times, presumably during matter domination. We do not incorporate this saturation in our model, however, so as to avoid an excessive proliferation of free parameters.

To conclude our analysis of the background evolution, let us consider the effect of

annihilation on the age of the universe,

$$t_0 = \frac{1}{H_0}\int_0^1 \frac{da}{a\sqrt{\Omega_\Lambda^0 + \Omega_b^0 a^{-3} + \rho_c(a)/\rho_{\rm crit}^0 + \Omega_r^0 a^{-4} + \rho_d(a)/\rho_{\rm crit}^0}}, \tag{3.19}$$

where $\rho_{\rm crit}^0$ is the critical density today. Clearly, for fixed values of the remaining cosmological parameters (including the dark matter density today), an increase in Γ causes an increase in ρ_d, and also induces an increase in ρ_c at earlier times. Therefore, such a change lowers the age of the universe.

3.2.2 Linear Perturbations

Our main concern here is the impact of annihilating dark matter on the formation of structure in the linear regime. We thus consider linear perturbations around the FLRW spacetime (3.9), and decompose them in Fourier modes,

$$ds^2 = a^2\left[-d\tau^2 + (\delta_{ij} + h_{ij})dx^i dx^j\right], \quad h_{ij} = \frac{k_i k_j}{k^2}h + 6\left(\frac{k_i k_j}{k^2} - \frac{1}{3}\delta_{ij}\right)\eta. \tag{3.20}$$

Here, η and h are the conventional metric potentials in synchronous gauge, which we adopt to connect our equations with the numerical results presented below.

Because the energy momentum tensors of dark matter and dark radiation still have perfect fluid form, the linearized Einstein equations retain their conventional form, and we shall not write them down here (see for instance [109] for the explicit equations.) We shall primarily address the modifications that annihilation imposes on the dynamics of both dark matter and radiation. The linearized time and spatial components of equation (3.8a) for dark matter are

$$\delta_c' + \frac{1}{2}h' - k^2 v_c + \frac{\delta\langle\sigma v\rangle}{m}\rho_c\, a + \frac{\langle\sigma v\rangle}{m}\rho_c\delta_c\, a = 0, \tag{3.21a}$$

$$v_c' + \mathcal{H}v_c = 0, \tag{3.21b}$$

where $\delta\langle\sigma v\rangle$ is given by equation (3.24), and we define velocity potentials by $u_i \equiv a\,\partial_i v$

where u_i are the spatial components of the four-velocity. Note that the annihilation cross section does not enter the equation for the velocity perturbation, which admits

$$v_c = 0 \tag{3.22}$$

as a solution. Therefore, as in the absence of annihilation, we can use the residual gauge freedom of synchronous gauge to set $v_c = 0$. In this gauge, the equations of motion for dark radiation simplify to

$$\delta_d' + \frac{2}{3}h' - \frac{4}{3}k^2 v_d - \frac{\delta\langle\sigma v\rangle}{m}\frac{\rho_c^2}{\rho_d}\,a - \frac{\langle\sigma v\rangle}{m}\frac{\rho_c^2}{\rho_d}(2\delta_c - \delta_d)\,a = 0, \tag{3.23a}$$

$$v_d' + \frac{1}{4}\delta_d + \frac{\langle\sigma v\rangle}{m}\frac{\rho_c^2}{\rho_d}v_d\,a = 0. \tag{3.23b}$$

Note that if $\langle\sigma v\rangle$ is time-dependent, it is not consistent to assume that its fluctuations $\delta\langle\sigma v\rangle$ vanish. Indeed, as we argue in appendix 3.A, in the non-relativistic limit we should set, to leading order in couplings,

$$\delta\langle\sigma v\rangle = \langle\sigma v\rangle\,\frac{h}{6}. \tag{3.24}$$

Heuristically, with $\langle\sigma v\rangle = m\Gamma a$, a perturbation in the scale factor $a \to a + \delta a$ induces a perturbation in the averaged cross section $\delta\langle\sigma v\rangle = \langle\sigma v\rangle\delta a/a$. But on large scales (in cosmic time coordinates) such a perturbation is equivalent to a metric perturbation with $h = 6\,\delta a/a$, from which equation (3.24) automatically follows.

3.2.3 Initial Conditions

In order to calculate the impact of dark matter annihilation on the temperature anisotropies and the distribution of matter, we need to specify initial conditions for the perturbations in all the components of the universe, including dark mater and dark radiation. These initial conditions are set well into the radiation-dominated era,

when all modes of cosmological interest are much larger than the Hubble radius.

At present, the angular correlations of cosmic microwave background temperature anisotropies are well-fit by a nearly scale-invariant spectrum of *adiabatic* primordial perturbations, in agreement with the predictions of the arguably simplest (single field) inflationary models. We would therefore like to impose adiabatic initial conditions on our perturbations, which we expect to be different for dark matter and dark radiation.

It turns out that in the presence of annihilation, and in synchronous gauge, the question of adiabaticity is a subtle one. Weinberg has shown for instance that the linearized perturbation equations in longitudinal gauge always admit (under rather mild assumptions) an "adiabatic" solution in the long wavelength limit $k \to 0$ [110]. The form of this adiabatic solution is explicitly known, regardless of the dynamics of the universe constituents, and this makes it straightforward to impose adiabatic initial conditions in longitudinal gauge, even in the presence of annihilation. But if one transforms this longitudinal adiabatic solution to synchronous gauge one finds that the total energy density perturbation vanishes, while η remains finite. Although this in fact solves the synchronous gauge equations for spatially constant perturbations, this solution cannot be extended to spatially varying perturbations. As argued by Weinberg, the appropriate adiabatic perturbations in synchronous gauge must come from the solution of the longitudinal gauge equations to next-to-leading order in the long-wavelength expansion. But the latter is in general unknown.

In synchronous gauge, the conventional approach to determine appropriate adiabatic initial conditions involves an expansion of the linearized solutions in powers of conformal time τ, which one can use to find appropriate initial conditions in the long-wavelength limit $k\tau \to 0$. In order to do so, one has to expand the scale factor and energy densities in powers of τ [111]. This does not pose any technical problem in the standard scenario, but it the presence of annihilation in fails because, from equation (3.11a), the dark matter density is non-analytic around $\tau = 0$. More generally, an expansion around $\tau = 0$ requires assumptions about the evolution of the

universe around the time of the big-bang, which is precisely the time around which we know the least about the universe.

In the specific case of coupled fluids, however, Malik and Wands have shown that the linearized perturbation equations in any gauge admit an adiabatic solution in the long-wavelength limit with

$$\frac{\delta\rho_\alpha}{\rho'_\alpha} = \frac{\delta\rho_\beta}{\rho'_\beta} \tag{3.25}$$

if the intrinsic non-adiabatic energy transfer of each individual fluid $\delta Q_{\mathrm{intr},\alpha}$ vanishes [112]. To check whether this is true in our case, we note that during radiation domination we can neglect the influence of dark matter and dark radiation perturbations on the metric potentials. In that limit, the adiabatic solution for the dominant constituents takes its conventional form [111],

$$\eta = -\zeta_i, \quad h = -\frac{\zeta_i}{2}(k\tau)^2, \quad \delta_\gamma = \frac{\zeta_i}{3}(k\tau)^2, \quad \delta_\nu = \delta_\gamma, \qquad \delta_b = \frac{3}{4}\delta_\gamma, \tag{3.26a}$$

$$v_\gamma = -\frac{\tau}{12}\delta_\gamma, \quad v_\nu = \frac{23+4R_\nu}{15+4R_\nu}v_\gamma, \quad v_b = v_\gamma, \tag{3.26b}$$

where the normalization has been chosen so that the curvature perturbation equals ζ_i (along this adiabatic solution ζ is conserved), and we only quote the leading terms in the long-wavelength expansion, since the subleading corrections depend on the unknown behavior of dark matter around $\tau = 0$. Then, it is simple to check using equations (3.24), (3.25) and (3.26a) that the intrinsic energy transfer of dark matter

$$\delta Q_{\mathrm{intr},c} = -\left(\frac{\delta\langle\sigma v\rangle}{m} - \frac{\langle\sigma v\rangle'}{m}\frac{\delta\rho_c}{\rho'_c}\right)\rho_c^2 \tag{3.26c}$$

vanishes, because in the Sommerfeld regime $\langle\sigma v\rangle' = \mathcal{H}\langle\sigma v\rangle$. We can therefore specify initial conditions for dark matter and dark radiation using equation (3.25),

$$\delta_c = \left(\frac{3}{4} + \frac{\langle\sigma v\rangle}{4m}\frac{\rho_c a}{\mathcal{H}}\right)\delta_\gamma, \quad \delta_d = \left(1 - \frac{\langle\sigma v\rangle}{4m}\frac{\rho_c^2 a}{\rho_d \mathcal{H}}\right)\delta_\gamma. \tag{3.26d}$$

Again, the magnitude of the impact of annihilation on the initial conditions is determined by the ratio R in equation (3.14).

On the other hand, the adiabatic solution discussed in reference [112] does not constrain the velocity perturbations. In order to determine the latter we note that equation (3.23b) has the integral solution

$$\rho_d v_d = -\frac{1}{4a^4}\int_{\tau_i}^{\tau} a^4 \delta\rho_d,$$

where we have assumed that at τ_i, $\rho_d v_d$ vanishes. For $\tau_i \ll \tau$ this reproduces for instance the conventional adiabatic solution if we replace dark radiation by standard radiation in the last equation. The dark radiation density perturbation can be found using equation (3.26d), which to first order in $\langle\sigma v\rangle$ gives

$$v_d = v_\gamma + \frac{1}{64}\frac{\langle\sigma v\rangle}{m}\frac{\rho_c a}{\mathcal{H}}\frac{\rho_c}{\rho_d}\tau\delta_\gamma. \tag{3.26e}$$

This expression reduces to the standard adiabatic solution in the limit $\langle\sigma v\rangle \to 0$, even though the second term on the right-hand-side typically dominates when ρ_d is very small. In our numerical code we use $\rho_d v_d$ as an independent variable, which, according to equation (3.26e) has a well-defined value even if ρ_d is initially zero. Recall that $v_c \equiv 0$ by gauge choice.

3.2.4 Impact on Structure Formation

Annihilations impact the growth of the perturbations on many fronts: The evolution of the background differs from the one without annihilations, the evolution of the perturbations differ from their counterparts without annihilation, and, finally, the initial conditions differ from their counterparts in the absence of annihilation.

Because annihilation enters the equations that model dark matter annihilation only through the combination $\langle\sigma v\rangle/m$, in the limit we are considering, cosmological observables are only sensitive to the combination $\langle\sigma v\rangle/m$. Here, as throughout this

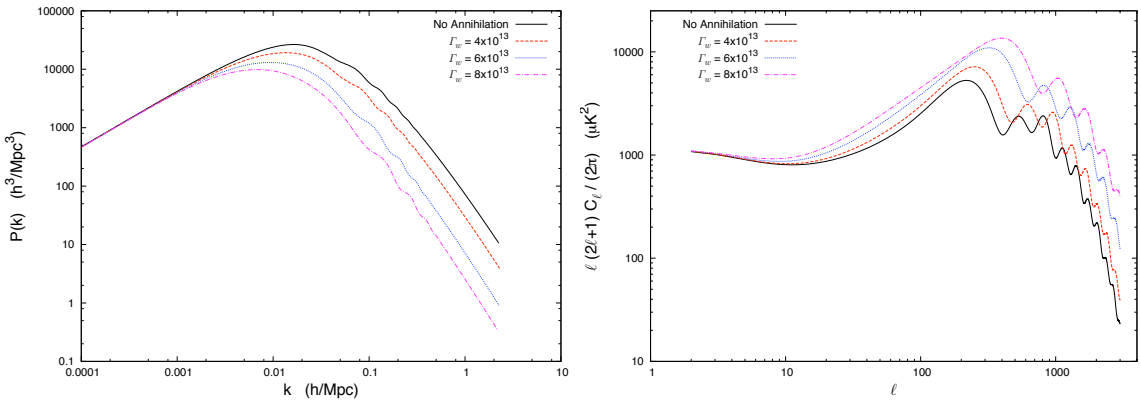


Figure 3.1: Matter and temperature anisotropy power spectra for different values of Γ_w. All the remaining cosmological parameters, including Ω_c^0, are kept fixed.

work, we focus on the case of Sommerfeld-enhanced annihilation, in which the average cross section times velocity is proportional to the scale factor. It is thus convenient to introduce an appropriately normalized constant Γ_w implicitly determined by

$$\frac{\langle \sigma v \rangle}{mc^2} \equiv \Gamma_w \frac{\langle \sigma v \rangle_w}{\mathrm{TeV}}\, a, \tag{3.27}$$

where $\langle \sigma v \rangle_w$ is defined in equation (3.1). Thus, because in our conventions a equals one today, Γ_w is the present value of $\langle \sigma v \rangle / mc^2$ in units of $\langle \sigma v \rangle_w$/TeV.

To determine the precise effects of annihilation on the CMB and matter power-spectrum we have modified the Boltzmann integrator CAMB [57] by including the three contributions mentioned above. In figure 3.1, we plot the matter and temperature anisotropy power spectra for different values of Γ_w, while keeping the remaining cosmological parameters fixed. The impact of annihilation on the CMB power spectrum is visible only for relatively large values of Γ_w, for which the amount of dark radiation is significant. This dark radiation is what drives most of the impact on the power spectra in this regime. In particular, dark radiation delays the onset of matter-domination, which shifts the wave number of the mode that enters the horizon at matter-radiation equality, k_{eq} to larger scales. On small scales ($k \ll k_{\mathrm{eq}}$) this shift

has no effect, because the transfer function approaches a constant, whereas at larger scales ($k \gg k_{\rm eq}$), the power is suppressed by the corresponding factor of $(k_{\rm eq}/k^0_{\rm eq})^2$ from the transfer function, where $k^0_{\rm eq}$ is the mode that enters at equality in the absence of annihilation. Accordingly, the maximum of the power spectrum at $k = k_{\rm eq}$ is shifted to smaller values of k, as seen in the left panel of figure 3.1.

The delay in matter-radiation equality also affects the size of the sound horizon at recombination, which becomes smaller because, with the remaining parameters fixed, the latter is a monotonically growing function of the redshift at matter-radiation equality (see for instance [56]). Hence, the angular size of the sound horizon at recombination decreases, thus shifting the cosmic microwave acoustics peaks to higher values of ℓ. Apart from Silk damping at very small scales, the amplitudes of these acoustic peaks depend on a monotonically growing function of $k/k_{\rm eq}$. Hence, a shift in $k_{\rm eq}$ to smaller values causes the anisotropy at a given angular scale (fixed value of k) to increase, as observed on the right panel of figure 3.1.

For smaller (and more realistic) values of Γ_w, the shift in matter-radiation equality is not as pronounced, and an accurate description of the impact of annihilation becomes impractical, because no single effect dominates the phenomenological signatures of annihilation.

3.3 Results

Because annihilation affects both the cosmic microwave temperature anisotropies and the matter power spectrum, measurements of the latter place constraints on how strongly dark matter annihilates. We can obtain a rough estimate of the kind of limits that we should be able to impose on Γ_w, defined in equation (3.27), by estimating the Fisher information. As we argued above, the impact of annihilation is dictated by the magnitude of R in equation (3.14), so on dimensional grounds we expect $\Delta C_\ell/C_\ell \sim R$. At leading order, R is constant during radiation domination and decays during matter

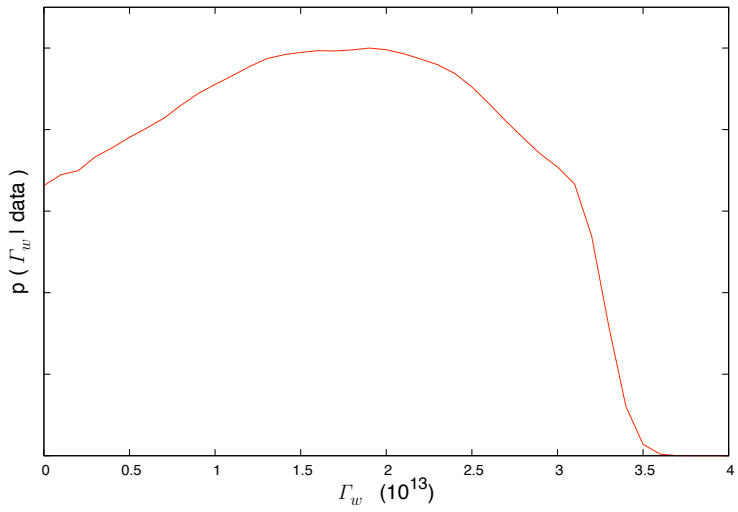


Figure 3.2: Smoothed marginalized posterior probability distribution function of Γ_w.

domination. Replacing R by $R_{\rm eq}$, and assuming that the temperature multipoles are normally distributed, the Cramer-Rao bound on the variance of an estimate of $R_{\rm eq}$ leads to

$$\Delta\Gamma_w \lesssim \frac{8\pi G}{3c^2}\frac{a_{\rm eq}^{1/2}}{H_0}\frac{\rm TeV}{\langle\sigma v\rangle_w}\frac{1}{\sqrt{\ell^{\,2}_{\rm max}}} \approx \frac{10^{14}}{\ell_{\rm max}}. \tag{3.28}$$

Here, $\ell_{\rm max}$ is the maximum multipole probed by the WMAP and ACBAR missions, $\ell_{\rm max} \approx 10^3$. As we shall see, the rough estimate in equation (3.28) is in fact not far from the actual standard deviation of Γ_w that we calculate later.

Upper Limits

To obtain upper limits on the value of Γ_w we follow the standard Bayesian approach in cosmological parameter estimation. We sample the posterior probability for a cosmological model with parameters H_0 (Hubble's constant today), Ω_Λ^0 (critical density fraction of a cosmological constant), $\Omega_b^0 h^2$ (baryon density), τ (optical depth), n_s (scalar spectral index), A_s (scalar spectral amplitude) and Γ_w in equation (3.27) with a set of four Monte Carlo Markov chains of 2.5×10^5 elements each, generated with an appropriately modified version of COSMOMC [58]. We impose flat priors on all

parameters, assume that the universe is spatially flat and neglect tensor modes. To check for the converge of our chains, we monitor the Gelman and Rubin statistic [59], which stays under 2×10^{-3} for all parameters. Following COSMOMC output, we also estimate the statistical errors on our upper limits by exploring their changes upon split of our chains in several subsamples; the corresponding relative errors remain below 1%.

To derive our first limits, we use cosmic microwave temperature anisotropy and polarization data from the seven year WMAP release [104], small angular scale temperature anisotropy data from the ACBAR experiment [113], and large scale structure data from an SDSS luminous red galaxy (LRG) sample [61]. For the WMAP and LRG data sets, the likelihood of a model is calculated using the codes supplied by the corresponding collaboration. The (smoothed) marginalized posterior probability density of Γ_w is shown in Figure 3.2. The posterior mean and standard deviation of Γ_w are

$$\langle \Gamma_w \rangle = 1.65 \cdot 10^{13}, \quad \sqrt{\langle \Gamma_w^2 \rangle^2 - \langle \Gamma_w \rangle^2} = 8.94 \cdot 10^{12}, \tag{3.29}$$

which suggests that there is no significant evidence for dark matter annihilation. In fact, the highest density set[1] with probability content $p = 95\%$ contains $\Gamma_w = 0$, which confirms that the latter is in reasonable agreement with the data.

In order to settle the question of the evidence for a non-zero value of Γ_w, we focus on the likelihood of the data under the two hypotheses

$$\begin{cases} H_0 : & \Gamma_w = 0, \\ H_1 : & \Gamma_w \neq 0. \end{cases} \tag{3.30}$$

The Bayes factor (the ratio of marginalized likelihoods under both hypotheses) is often advocated in Bayesian hypothesis testing. Unfortunately, for nested hypothesis of the form (3.30), it is ill-defined for improper priors, or very sensitive to the width of

[1] In our context, a highest density set is a credible interval of prescribed probability content and minimal length. See, for instance, 2.50 in [114].

Dataset	μ	σ	68%	95%
WMAP+ACBAR+LRG	$1.65 \cdot 10^{13}$	$8.94 \cdot 10^{12}$	$\leq 2.18 \cdot 10^{13}$	$\leq 3.09 \cdot 10^{13}$

Table 3.1: Posterior mean μ and standard deviation σ of Γ_w, and 68% and 95% upper credible limits on Γ_w.

any uninformative prior placed on the additional parameters (see e.g. 7.17 in [114]). We focus instead on the likelihood ratio

$$\lambda \equiv \frac{\max_{H_0} L(\text{data}|H_0)}{\max_{H_1} L(\text{data}|H_1)}, \tag{3.31}$$

which is a statistic that has often proved to be sensible in the classical context, and is closely connected to the Bayes factor asymptotically. Evaluating the maximum likelihoods under both hypotheses, we find

$$-2\log\lambda = 0.92. \tag{3.32}$$

Recall that under H_0, $-2\log\lambda$ is asymptotically distributed like χ^2 with 1 dof, so the evidence against the null hypothesis H_0 is weak at best.[2]

We thus proceed to set an upper limit on the value Γ_w. From the posterior distribution we finally derive the 95% credible upper limit

$$\Gamma_w \leq 3.09 \cdot 10^{13}. \tag{3.33}$$

Our results are summarized in Table 3.1.

In our conventions $a = 1$ today, so the previous limits translate for instance into $\langle\sigma v\rangle/mc^2 \lesssim 2.81 \cdot 10^{10} \langle\sigma v\rangle_w/\text{TeV}$ around recombination, at $z = 1100$. In that

[2]Both the distribution of $-2\log\lambda$ and its relation to the Bayes factor in the form of the Schwarz information criterion are often derived in the limit of a large number of independent and identically distributed variables. Although the temperature multipoles $a_{\ell m}$ are indeed independent in a statistically isotropic universe, they are however not identically distributed. Hence, care should be when quoting precise statistical predictions based on likelihood ratios.

respect, for moderately small values of f, our constrain on the value of $\langle\sigma v\rangle$ at recombination is orders of magnitude weaker than the limit (3.5) based on recombination alone. Therefore, in those cases it is safe to ignore the impact of annihilation on the evolution of structure. But in any case, our limits are fundamentally different from (3.5) because while the latter only constrains $\langle\sigma v\rangle/m$ at recombination, the former are sensitive to the evolution of $\langle\sigma v\rangle/m$ throughout cosmic history. Since WMAP is sensitive to comoving scales with $k\,\tau_0 \sim 10^3$, our limits are sensitive to the value of $\langle\sigma v\rangle$ at redshifts of about $z \sim 10^4$.

The remaining cosmological parameters $(H_0, \Omega_\Lambda^0, \Omega_b^0 h^2, \tau, n_s, A_s)$ do not differ significantly from their values with $\Gamma_w = 0$. In particular, their best fit values under H_1 fall within the posterior 95% credible limits on the corresponding parameters under H_0. The parameter Γ_w shows the strongest correlations with the amount of baryons $\Omega_b h^2$ (-80%), the Hubble parameter H_0 (76%) and the age of the universe (-89%), although the latter still remain well constrained by the data. The negative correlation between Γ_w and the age of the universe is, for instance, what we expect from our analysis of the background evolution at the end of subsection 3.2.1.

3.4 Summary and Conclusions

We have studied the impact of dark matter annihilation on the cosmic microwave background and the matter power spectrum, under the assumption that dark matter annihilates into dark radiation with an averaged cross section times velocity $\langle\sigma v\rangle$ that grows in proportion to the scale factor. This Sommerfeld enhancement is expected to occur generically in any dark matter model in which dark matter particles experience an additional attractive interaction, regardless of the dark matter annihilation channels. Most previous analyses of this scenario assumed that dark matter predominantly annihilates into standard model particles (visible radiation). Our analysis focuses on the purely gravitational impact of the annihilation, and thus holds for a

much wider class of models. In particular, within the context of our analysis we can address whether the relativistic dark matter annihilation products consitute the dark radiation that some analyses of cosmic data seem to favor.

Actual cosmic microwave anisotropy and large scale structure data do not show evidence of dark matter annihilation with such a growing cross section, so we have derived the limits on the corresponding averaged cross section times velocity listed in table 3.1 (we define Γ_w in equation (3.27)). As seen in the table, these upper limits allow $\langle \sigma v \rangle$ to be several orders of magnitude larger than a typical weak annihilation cross section. In particular, our limits indicate that if dark matter annihilation deposits a significant fraction of the annihilation energy into visible radiation, $f \gg 10^{-8}$, the effects on the cosmic microwave background that we have studied here are subdominant. On the other hand, if $f \ll 10^{-8}$ the impact of annihilation on recombination is subdominant, and the gravitational effects that we have studied here play the dominant role. Because the data do not seem to support the dark matter annihilation hypothesis, we do not find evidence supporting an additional dark relativistic species originating from such annihilations either.

At present, the nature of dark matter remains a mystery. The limits that we have derived are not only useful in further constraining the properties of dark matter itself, but also in constraining its interactions with other elements of the sector where it resides. Because large scale structure still allows for very large annihilation cross sections, the dark sector may in principle host a dark matter candidate with properties far different from the standard collisionless WIMP.

Appendix

3.A Microscopic Description

In order to determine the impact of annihilation on the dark matter density we begin with a microscopic description of annihilation. Let us consider the phase space distribution of dark matter particles in the universe, f. It is useful to resort to a formulation in which the distribution function depends on the space-time coordinates τ and x^i, and on covariant spatial momenta p_j, $f = f(\tau, x^i, p_j)$ (for simplicity we assume that dark matter particles are spinless.) In that case, the distribution function f is a scalar under diffeomorphisms, and it obeys the Boltzmann equation [115]

$$\frac{p^0}{m}\left[\frac{\partial f}{\partial \tau} + \frac{\partial f}{\partial x^i}\frac{dx^i}{d\tau} + \frac{\partial f}{\partial p_i}\frac{dp_i}{d\tau}\right] = C[f, f], \tag{3.34}$$

where m is the WIMP mass and $dp_i/d\tau$ is dictated by the geodesic equation

$$p^0 \frac{dp_j}{d\tau} = \frac{1}{2m}\frac{\partial g_{\alpha\beta}}{\partial x^j} p^\alpha p^\beta. \tag{3.35}$$

In the above, p^0 should be expressed in terms of the covariant momenta p_i and the spacetime metric.

Although this is not manifest, the left hand side of equation (3.34) is a diffeomorphism scalar. Hence, the collision term C is a scalar too, and describes the changes in the distribution function caused by collisions and annihilations. For definiteness, let us assume that the only relevant processes involve the annihilation of two dark

matter particles χ into two spinless particles ϕ of four-momenta q_1 and q_2. Then, the collision term is

$$C[f,f] = -\frac{(2\pi)^4}{m^2 m_\phi^2}\int d_*^4 p_2\, d_*^4 q_1\, d_*^4 q_2\, f(x^\mu, p_{1\nu}) f(x^\mu, p_{2\nu}) R_{\rm ann}\,\sqrt{-g}\,\delta^4(p_1+p_2-q_1-q_2), \tag{3.36}$$

where we identify $p_1 \equiv p$, and all four-momenta are covariant (as opposed to contravariant.) Note the minus sign in front of the last equation, which reflects that we are considering annihilation processes only.

The combination

$$d_*^4 p \equiv \frac{d^4 p}{\sqrt{-g}}\, 2m\,\theta(p_0)\delta(p^2+m^2) = \frac{m}{\sqrt{-g}}\frac{d^3 p}{p^0} \tag{3.37}$$

is a scalar under diffeomorphism, so $R_{\rm ann}$ has to be a scalar too. We can thus calculate $R_{\rm ann}$ using the standard rules of quantum field theory in a local Lorentz frame, in which

$$R_{\rm ann} \equiv p_1^0 p_2^0 q_1^0 q_2^0 |\mathcal{M}_{\rm ann}|^2, \tag{3.38}$$

and $\mathcal{M}_{\rm ann}$ determines the $\mathcal{S}$-matrix, $\mathcal{S} = -2\pi i \mathcal{M}\delta^4(p_1+p_2-q_1-q_2)$. Say, for an interaction of the form

$$S_{\rm int} = \int d^4x\sqrt{-g}\,\frac{\lambda}{4}\chi^2\phi^2, \tag{3.39}$$

where χ represents the dark matter field and ϕ its (relativistic) annihilation products, $R_{\rm ann} = \lambda^2/16$ at tree level (we follow the conventions of [116].)

In this Chapter, however, we are interested in annihilation processes for which the annihilation rate is boosted by a factor S from Sommerfeld enhancement,

$$R_{\rm ann} = \frac{\lambda^2}{16}\times S(v_0/v_{\rm rel}), \tag{3.40}$$

where λ is a constant (not necessarily related to the simple model in equation (3.39)), and v_0 is a constant with dimensions of velocity and $v_{\rm rel}$ is the appropriate relativistic

expression for the relative velocity [117]

$$v_{\rm rel} = \frac{\sqrt{-(p_1+p_2)^4 - 4(p_1+p_2)^2 m^2}}{-(p_1+p_2)^2 - 2m^2}. \tag{3.41}$$

(Because we are interested in the non-relativistic limit, any diffeomorphism scalar $v_{\rm rel}$ that reduces to $|\vec{v}_1 - \vec{v}_2|$ at non-relativistic momenta in a local Lorentz frame would suffice). The factor S describes the enhancement of the cross section. In models in which such an enhancement is caused by an attractive interaction mediated by a light force carrier of mass m_Y coupling to dark matter with amplitude λ_Y it has the form [77]

$$S(x) \approx \begin{cases} \dfrac{m}{m_Y \alpha}, & \dfrac{v_{\rm rel}}{c} \ll \dfrac{2m_Y}{m} \\ \dfrac{v_0}{v_{\rm rel}}, & \dfrac{2m_Y}{m} \ll \dfrac{v_{\rm rel}}{c} \ll 2\pi\alpha \\ 1, & 2\pi\alpha \ll \dfrac{v_{\rm rel}}{c} \end{cases}, \tag{3.42}$$

where $\alpha = \lambda_Y^2/(4\pi)$ and $v_0 = 2\pi\alpha$. For the rest of our analysis we restrict ourselves to the intermediate regime, in which the enhancement is inversely proportional to the relative velocity,

$$S \approx \frac{v_0}{v_{\rm rel}} \qquad \left(\frac{2m_Y}{m} \ll \frac{v_{\rm rel}}{c} \ll 2\pi\alpha\right). \tag{3.43}$$

Clearly, this range of velocities is strongly model-dependent, although typically, for light force carriers and not too weak couplings it can span several orders of magnitude in $v_{\rm rel}$.

3.A.1 Perfect Fluid Description

The energy momentum tensor of the ensemble of particles described by f is

$$T_{\mu\nu} = \int d^4_* p \, \frac{p_\mu p_\nu}{m} f, \tag{3.44}$$

which clearly transforms like a tensor. In order to determine whether this energy momentum is conserved, it is convenient to consider a local inertial frame, in which $g_{\mu\nu} = \eta_{\mu\nu}$ and $\Gamma^{\mu}{}_{\nu\rho} = 0$. Then, using the Boltzmann equation (3.34) and general covariance it is easy to show that in an arbitrary coordinate system the energy momentum tensor satisfies

$$\nabla_\mu T^{\mu\nu} = \int d^4_* p \, p^\nu \, C[f, f], \tag{3.45}$$

since the latter holds in any local inertial frame. Thus, in the absence of annihilation the energy momentum tensor is covariantly conserved, as it should.

In order to relate the kinetic to the fluid description, following Eckart [118], we define the four-velocity of the fluid to be proportional to the averaged particle velocity,

$$u^\mu \equiv \frac{\langle p^\mu \rangle}{\sqrt{-\langle p^\nu \rangle \langle p_\nu \rangle}}, \tag{3.46}$$

where the average of any function g of momentum is defined by

$$\langle g(\vec{p}) \rangle = \frac{1}{n} \int d^4_* p \, g(\vec{p}) f, \tag{3.47}$$

and n is the (scalar) particle number density,

$$n \equiv \int d^4_* p \, f. \tag{3.48}$$

It is simple to show that for *any* distribution the Boltzmann equation (3.34) implies that in the absence of annihilations the current $n\langle p^\mu \rangle$ is covariantly conserved,

$$\frac{1}{m} \nabla_\mu [n \, \langle p \rangle^\mu] = -n^2 \langle \sigma v \rangle, \tag{3.49}$$

where we have defined the averaged annihilation cross section times relative velocity,[3]

$$\langle \sigma v \rangle = -\frac{1}{n^2} \int d^4_* p \, C[f, f], \tag{3.50}$$

which of course vanishes in the absence of annihilations. In that case, the four velocity (3.46) is proportional to the current that captures the conservation of matter.

We shall assume that the distribution f is such that the energy momentum tensor is well approximated by that of a perfect fluid,

$$T_{\mu\nu} = (\rho + p) u_\mu u_\nu + p \, g_{\mu\nu}. \tag{3.51}$$

Using equations (3.44) and (3.46), the energy density thus becomes

$$\rho \equiv T_{\mu\nu} u^\mu u^\nu = -\frac{n}{m} \frac{\langle p_\mu p_\nu \rangle \langle p^\mu \rangle \langle p^\nu \rangle}{\langle p_\rho \rangle \langle p^\rho \rangle}. \tag{3.52}$$

In order to determine the pressure we note that equation (3.44) implies that $T^\mu{}_\mu = -m\,n$, whereas equation (3.51) implies that $T^\mu{}_\mu = 3p - \rho$. Therefore, the pressure of the fluid simply is

$$p = \frac{\rho - mn}{3}, \tag{3.53}$$

which clearly shows that only relativistic components, those for which the energy density ρ is larger than the "rest" energy $m\,n$, contribute to the pressure.

Pressureless fluids

By definition, the pressure of a non-relativistic fluid of particles vanishes, which implies that $\rho = mn$, as expected. Looking back at equation (3.52) and noting that $\langle p_\mu p^\mu \rangle = -m^2$ we see that this is the case if the covariance of the four momentum

[3] Recall that cross sections are rates per flux, and that the flux is proportional to the relative velocity between the annihilating particles.

vanishes,

$$\langle p_\mu p_\nu \rangle = \langle p_\mu \rangle \langle p_\nu \rangle. \tag{3.54}$$

It then follows, using (3.46), that

$$u^\mu = \frac{\langle p^\mu \rangle}{m}. \tag{3.55}$$

Given that the covariance of the momenta vanishes by assumption, it is natural to assume that we can also replace the momentum on the rhs of equation (3.45) by its average. Then, the conservation equation becomes

$$\nabla_\mu T^{\mu\nu} = -\frac{\langle \sigma v \rangle}{m} \rho^2 u^\nu. \tag{3.56}$$

Since for a pressureless fluid ρ is proportional to the number density n, equation (3.56) also expresses conservation of particle number, as can be seeing by looking at the projection of that equation onto u_ν. To conclude, we note that because $\int d^4_* q_1 d^4_* q_2 \sqrt{-g}\, \delta^{(4)}(p_1 + p_2 - q_1 - q_2)$ is a scalar, in the non-relativistic limit the averaged annihilation rate in a universe with metric (3.20) becomes

$$\langle \sigma v \rangle \approx \frac{(2\pi)^5}{m^2} \frac{\lambda^2}{16} \frac{a^2}{(-g)} \frac{1}{n^2} \int d^3 p_1 d^3 p_2 \, \frac{v_0}{v_{\rm rel}} f(\vec{p}_1) f(\vec{p}_2), \tag{3.57}$$

where we have assumed that the relative velocities are in the regime in which Sommerfeld enhancement is effective, equation (3.43), and that all the annihilation products are highly relativistic, $m \gg m_\phi$.

3.A.2 Background

Let us turn our attention now to the evolution of $\langle \sigma v \rangle$ in the unperturbed universe (3.9). Because of homogeneity and isotropy, the distribution function f can only

depend on the magnitude of the momentum $f = f(\tau, p)$, where

$$p \equiv a\sqrt{g^{ij}p_i p_j}. \tag{3.58}$$

In the WIMP scenario dark matter decoupled while being non-relativistic, so it would be natural to consider a Maxwell-Boltzmann ansatz for the distribution function, but this is problematic because it can be shown, that annihilation does not preserve this form of the distribution function [80].

We can nevertheless proceed without making any assumptions about the form of f when the coupling λ is sufficiently small. Namely, because $\langle\sigma v\rangle$ in equation (3.57) is already of order λ^2, to leading order we can calculate $\langle\sigma v\rangle$ by substituting into (3.57) the solution of the Boltzmann equation (3.34) to zeroth order in λ,

$$\frac{\partial f}{\partial \tau} = 0. \tag{3.59}$$

In this case, *any* distribution function $f = f(p)$ solves equation (3.59), and the density of dark matter particles (to zeroth order) evolves as we would expect in the absence of annihilation,

$$n = \frac{1}{a^3}\int d^3p\, f(p), \tag{3.60}$$

provided that f has support for non-relativistic momenta only. Using this form of the density and equation (3.57) the averaged cross section becomes

$$\langle\sigma v\rangle = \frac{(2\pi)^5}{m^2}\frac{\lambda^2}{16}\frac{a\int d^3p_1 d^3p_2\, f(p_1)f(p_2)\, m v_0/|\vec{p}_1 - \vec{p}_2|}{\left(\int d^3p f(p)\right)^2}. \tag{3.61}$$

The crucial point is that the the thermal average is proportional to the scale factor a, simply because the relative velocity v_{rel} between dark matter particles redshifts as the universe expands.

3.A.3 Perturbations

In a perturbed universe (3.20) we also need to consider the perturbations in the annihilation cross section, $\delta\langle\sigma v\rangle$. Again, in the non-relativistic limit it is possible to do so for an arbitrary background distribution and arbitrary metric perturbations by focusing on the leading result in a small-coupling expansion. In particular, we can calculate $\delta\langle\sigma v\rangle$ to leading order in λ by solving the perturbed Boltzmann equation for δf to zeroth order and substituting the corresponding solution into equation (3.57). In doing so, we shall be able to remain in the perfect fluid approximation, without the need to include the evolution of δf into our system of perfect fluid equations.

Following [119] let us write the perturbed distribution function as

$$f(\tau, \vec{x}, \vec{p}) = \bar{f}(p) + \delta f(\tau, \vec{x}, \vec{p}), \tag{3.62}$$

where $\bar{f}$ is an arbitrary distribution with support at non-relativistic momenta, and p is the magnitude of the spatial momentum defined in (3.58),

$$p = \sqrt{p_k p_k} - \frac{1}{2}\frac{h_{ij}p_i p_j}{\sqrt{p_k p_k}}. \tag{3.63}$$

Here and in the following, Einstein's summation convention is implied even if repeated indices are not in opposite locations. Because p now depends on the metric, $\bar{f}$ also contributes to the perturbations of the distribution function. Then, the perturbation δf obeys the linearized Boltzmann equation

$$\frac{\partial \delta f}{\partial \tau} + \frac{\partial \delta f}{\partial x^i}\frac{1}{a^2}\frac{p_i}{p^0} - \frac{1}{2}\frac{\bar{f}'}{p}\frac{\partial h_{jk}}{\partial \tau}p_j p_k = 0, \tag{3.64}$$

where a prime denotes derivative with respect to the argument (p in this case). Recall that we set the collision term to zero because we are only interested in evaluating $\delta\langle\sigma v\rangle$ to zeroth order in λ.

The linearized Boltzmann equation (3.64) has the line of sight solution

$$\delta f = \frac{1}{2}\frac{\bar{f}'}{p}\int_{\tau_i}^{\tau} d\tau' p_i p_j \, h_{ij,\tau}\left(\tau', \vec{x} - \int_{\tau'}^{\tau} d\tau'' \frac{1}{a}\frac{\vec{p}}{m}\right), \tag{3.65}$$

which assumes that δf was negligible at the initial time τ_i, and that dark matter is non-relativistic. In general, this solution is a non-local functional of h_{ij}, but in the non-relativistic limit in which $p/m \ll 1$, we can set the momentum in the argument of the integral to zero, which yields a simple local expression for δf in terms of the metric perturbations h_{ij},

$$\delta f(\tau, \vec{x}, \vec{p}) = \frac{1}{2}\frac{\bar{f}'}{p} h_{ij}(\tau, \vec{x})\, p_i p_j, \tag{3.66}$$

where we have assumed again that the perturbations of h are initially negligible (as we discuss in subsection 3.2.3, this holds for adiabatic initial conditions.) Note that the first correction to this result away from the strict non-relativistic limit would be proportional to three momenta, and would therefore vanish in momentum integrals invariant under rotations like the ones involved in the calculation of $\delta\langle\sigma v\rangle$. Given the structure of the terms we have omitted, we expect this approximation to be valid on scales

$$(k\tau)^2 \ll \left(\frac{m}{p/a}\right)^2, \tag{3.67}$$

which for non-relativistic momenta encompasses modes well within the horizon. Since momenta redshift with a, this approximation becomes increasingly accurate.

With the explicit expression for δf in equation (3.66) at hand, we can calculate $\delta\langle\sigma v\rangle$ by substituting the solution (3.66) into equation (3.61). Because the latter is a function of g_{ij}, invariant under spatial diffeomorphisms, metric perturbations do not contribute to $\delta\langle\sigma v\rangle$, and we may restrict our attention directly to the contributions from δf alone. The resulting integrals can be simplified by noting that rotational invariance and linearity demand that $\delta\langle\sigma v\rangle$ be proportional to the trace of h_{ij}, and

explicit calculation shows that

$$\delta\langle\sigma v\rangle = \frac{h}{6}\langle\sigma v\rangle. \tag{3.68}$$

In this way, the system of perfect fluid equations remains closed, and there is no need to track the evolution of δf in our system of equations. Also note that the velocity perturbation associated with the solution (3.66) vanishes, and is therefore consistent with the gauge choice $v_c = 0$ in equation (3.22).

It is also instructive to explore how δf is affected by gauge transformations, and how the residual gauge symmetry allowed by synchronous gauge leads to the existence of gauge mode solutions. Under a gauge transformation

$$x^\mu \to \tilde{x}^\mu = x^\mu + \epsilon^\mu \tag{3.69}$$

the perturbation in the distribution function δf defined in equation (3.62) transforms as

$$\Delta\delta f \equiv \delta\tilde{f} - \delta f = \frac{\bar{f}'}{p}\left(\epsilon_{,ij}\, p_i p_j + \epsilon^0{}_{,i}\, p_i p_0\right) + \frac{1}{2}\frac{\bar{f}'}{p} p_i p_j \Delta h_{ij}, \tag{3.70}$$

where we have used that $\epsilon^i \equiv \partial_i \epsilon$ in the scalar sector. The additional term proportional to Δh_{ij} originates from the dependence of $\bar{f}$ on the metric perturbations. Under the same gauge transformations, the latter transform as

$$\Delta h_{00} = 2\mathcal{H}\epsilon^0 + 2\epsilon^0{}_{,\tau} \tag{3.71a}$$

$$\Delta h_{0i} = \epsilon^0{}_{,i} - \epsilon_{,i\tau} \tag{3.71b}$$

$$\Delta h_{ij} = -2\epsilon_{,ij} - 2\mathcal{H}\epsilon^0\delta_{ij}. \tag{3.71c}$$

Equations (3.71a) and (3.71b) immediately reveal that synchronous gauge contains a

residual gauge freedom. A coordinate transformation with

$$\epsilon^0 = \frac{A(x)}{a}, \quad \epsilon = B(x) + A(x) \int^\tau \frac{d\tau'}{a(\tau')} \tag{3.72}$$

preserves the synchronous conditions $h_{00} = h_{0i} = 0$, and thus leads to the existence of gauge modes. In fact, it is easy to check that equations (3.70) and (3.71), with ϵ^μ given by equations (3.72) solve the linearized Boltzmann equation (3.64). Substituting this gauge mode into expression (3.57) we find that the term proportional to $p_i p_0$ does not contribute to $\delta\langle\sigma v\rangle$ because of rotational invariance. In the remaining terms, the factors of ϵ cancel, so the corresponding $\delta\langle\sigma v\rangle$ equals what we would get from a perturbation δf of the form (3.66) with an effective metric perturbation

$$h_{ij} = -2\mathcal{H}\epsilon^0 \delta_{ij}. \tag{3.73}$$

If we substitute this effective metric perturbation into equation (3.68) we find that

$$\delta\langle\sigma v\rangle = -\mathcal{H}\langle\sigma v\rangle\epsilon^0. \tag{3.74}$$

This is precisely what we expect from a gauge transformation of a scalar proportional to the scale factor, and it also leads to a (gauge mode) solution of the perturbed equations (3.21a). This agreement thus provides a check of expression (3.68) and the consistency of our approach.

Chapter 4

Detecting anomalies in CMB maps: a new method

4.1 Introduction

A dramatic increase in the amount of observed data has, over the last couple of decades, led to a much better understanding of the Universe we inhabit. In fact, the cosmology community is so confident about the standard paradigm that the paradigm is referred to as the Standard (or Concordance) Model, after the Standard Model of Particle Physics. Seven or eight parameters, along with general relativity and elementary quantum mechanics, are sufficient to explain a host of observations on the largest scales, once initial conditions are set deep in the radiation era. Standard field quantization techniques applied to cosmic inflation have been remarkably successful in explaining these initial conditions even. The cosmology being studied today is called Precision Cosmology because parameters have been determined to percent-level precision [120, 121].

But, as is well-known, there is a difference between precision and accuracy. Questions abound over some of the postulates of the Concordance Model. Because we have access to only one universe, the usual method of testing postulates by repeating

experiments cannot be carried out. As inflation postulates that the primordial seeds of the universe's structure themselves arise out of a stochastic process, this inability to repeat experiments is an even bigger handicap.

The cosmic microwave background (CMB) has turned out to be the cosmologist's most useful aid in understanding what has happened in the universe from just a few minutes after the Big Bang, all the way up to the present. Since most CMB photons have travelled to us without any scattering, they represent a very faithful picture of the universe when it was about 400,000 years old. Moreover, at the scales relevant to us today, the density perturbations were small enough that linear perturbation theory is an excellent approximation. This implies that the statistical properties of the primordial fluctuations were preserved all the way to the surface of last scattering, and thence to us today.

In vanilla models of inflation, the Fourier modes of the primordial fluctuations have the same dynamics as harmonic oscillators in their ground state, and are thus distributed as Gaussians.[1] Moreover, statistical homogeneity and isotropy imply that the variance of this Gaussian distribution doesn't depend on the direction of the wavenumbers of the Fourier modes, and that the variance is the same for the real and the imaginary parts of the Fourier modes [122]. In 2013, Planck announced that the CMB data put very strong constraints on the amount of non-Gaussianity in the primordial power spectrum [123]. In effect, this meant that several exotic inflationary models got ruled out with very high probability. So, the accepted wisdom is that the Fourier modes of the initial density perturbations are independent and distributed as Gaussians.

The only challenge to this postulate of independent, normally distributed perturbations probably has to do with the so-called CMB anomalies. The CMB anisotropies across the sky are usually expressed in terms of $a_{\ell m}$'s, the co-efficients corresponding to the spherical harmonics $Y_{\ell m}$'s. Expressing Fourier modes in terms of the spherical

[1] We ignore non-Gaussianities of the kind calculated by Maldacena [23] as they are highly suppressed.

harmonics, and using the results from the previous paragraph, we are led to conclude that most viable inflationary models predict that the $a_{\ell m}$'s are normally distributed with zero mean and with a variance C_ℓ (hence independent of m).

When WMAP announced [124] its first set of results, the authors in [125] and [126] analysed the $a_{\ell m}$'s to test this hypothesis. (See [127] for an earlier analysis with the COBE data.) They employed a variety of tests and found weak evidence for correlation amongst the $a_{\ell m}$'s that corresponded to the largest scales (low-ℓ's). The anomalies reported dealt with the alignment of different multipoles and how planar a few of these multipoles were. Several authors [128–130] performed similar analyses and again found weak evidence. A different kind of anomaly, having to do with a low value for the variance in the CMB sky, was observed by [131] for the WMAP data. The authors in [132–135] considered the isotropy of the angular power spectrum and concluded that it appears to be anisotropic. A few more anomalies were reported in [136–138], amongst other works.

Apart from the weak evidence, two arguments were proferred questioning the "real" nature of these anomalies: one, that they arose from the systematics that WMAP employed; two, that these anomalies were checked for *a posteriori*. So, now that Planck has confirmed that most of the anomalies are present in their data too [139], one may reasonably argue that the anomalies are a *bona fide* feature of the CMB. The question remains as to whether this feature is physically relevant or not. As Planck also concluded that there is only weak evidence for these anomalies, this question has not been settled convincingly. Many authors contend [140, 141], with good reason, that given a large enough dataset, one can always find any feature that one desires. Compounding the problem of the large dataset is the fact that the anomalies have been observed for low-ℓ's—it is here that the effect of cosmic variance is most pronounced. This makes statistical inferences about the anomalies even more dubious. Also, there is the perennial question of foreground contamination—without a reliable model for galactic dust, it isn't clear how accurate the determination of the

$a_{\ell m}$'s is. (Though, with the availability of multiple probes and multiple frequency channels, this is less of an issue [142] than it used to be.)

But, the fact remains that there are many anomalies with weak evidence. Some of them are so apparently different from the rest that, at the outset, it is seems hard to believe that they all arose from a common statistical fluke. And, the anomalies seem to be present "coherently" across different ℓ's too, seemingly making it harder to believe that it is the consequence of a fluke.

This chapter tries to address the second of the arguments put forth against the existence of the anomalies—that the tests are all *a posteriori*. We propose two statistics that test the null hypothesis that the $a_{\ell m}$'s are independent, normally distributed zero-mean variables. As we shall show, these statistics are such that one cannot reasonably be accused of performing the analysis after "seeing" the anomalies in the data. The point is to perform as general an analysis of the data as possible, without worrying about whether a test statistic is physically-motivated or high-confidence-interval motivated. We shall achieve this by not arbitrarily choosing the ℓ's and the m's to analyse; instead, we consider their linear and quadratic combinations. For one, this makes the analysis more general; but, crucially, if the anomalies are physical, it is very unlikely that they arose because of a coupling between just two or three $a_{\ell m}$'s. This anomalous nature must be present for a range of modes and thus considering combinations of the modes should lead to an enhancement of the signal. Also, previous analyses of CMB anomalies have involved several Monte Carlo simulations to produce a reference set of Gaussian sky maps. And, one gets several p-values as different statistics are considered. In our case, once a maximum ℓ value is chosen, one gets a single p-value for each of the two statistics considered.

4.2 Y—A Linear Statistic

4.2.1 Motivation

In broad terms, the way a hypothesis is statistically tested is this: Assume that a given dataset is described by a known probability distribution P_1; formulate a statistic that is a function of the corresponding random variables; determine the expected distribution P_2 of this statistic, assuming the fiducial distribution P_1; see how compatible the actual (realized) value of the statistic using the given dataset is, with the distribution P_2. If the compatibility is very low, then one concludes that the data are inconsistent with the hypothesis.[2] It is clear, however, that the conclusion strongly depends on the statistic chosen. Ideally, one would like to do the analysis for several different statistics.

Let us look at linear test statistics; that is, if an n-component vector $\vec{X}$ describes n variables of a dataset, then consider $S = \vec{a} \cdot \vec{X}$, where each choice of the constant co-efficients $\vec{a}$ would correspond to one statistic. If one wants to do a blind analysis of the data, one is tempted to consider several different choices of $\vec{a}$—for instance, by making $\vec{a}$ itself a random vector. If one knows the underlying distribution of $\vec{a}$, and the null hypothesis for the distribution of $\vec{X}$, then one may hope to determine the distribution of S. In general, this distribution would be quite complicated. In the next section, we show that for a specific choice of the distribution of $\vec{a}$, and a specific null hypothesis, the distribution of S becomes very simple.

4.2.2 The Y Statistic

Let $\vec{a}$ be an N-component random variable vector, with each component being described by a zero-mean normal distribution, $\mathcal{N}(0, \alpha_i^2)$. Let $\vec{X}$ be another N-component vector with each of its components being described by $\mathcal{N}(0, \beta_i^2)$. Further, assume that

[2] This is more of a goodness-of-fit test than a hypothesis test because we are not specifying an alternative hypothesis. But, the former can be thought of as a special case of the latter, where the alternative hypothesis is that the data are *not* described by the null hypothesis.

$\alpha_i^2 = 1/\beta_i^2$. That is, the combined probability distribution function is

$$P(\vec{a}, \vec{X}) = \frac{1}{(2\pi)^N} \left(\det \boldsymbol{\Sigma_a} \det \boldsymbol{\Sigma_X}\right)^{-1/2} \exp\left(-\frac{1}{2}\vec{a}^{\,T} \boldsymbol{\Sigma_a^{-1}} \vec{a}\right) \exp\left(-\frac{1}{2}\vec{X}^{\,T} \boldsymbol{\Sigma_X^{-1}} \vec{X}\right), \tag{4.1}$$

where $\boldsymbol{\Sigma_a}$ and $\boldsymbol{\Sigma_X}$ are diagonal matrices with $(\boldsymbol{\Sigma_a})_{ij} = (\boldsymbol{\Sigma_X})_{ij}^{-1} = \alpha_i^2 \delta_{ij}$.

Consider a random variable arising out of these two random variables,

$$Y = \vec{a} \cdot \vec{X} = a_i X^i \tag{4.2}$$

In the above, Einstein's summation convention is implied. Though both $\vec{a}$ and $\vec{X}$ are random variables, we shall eventually consider the case where there is only one realization of $\vec{X}$. That is, the two random variables must not be considered to be on the same footing. We shall first treat $\vec{X}$ as a constant vector, carry out all operations with respect to $\vec{a}$ and finally promote $\vec{X}$ to a random vector and carry out operations with respect to it. This shall become more clear when we apply it to the case of Cosmology.

For a given realization of $\vec{X}$, Y is a linear combination of the independent normal variables $\vec{a}$. Hence, Y is normally distributed too:

$$Y \sim \mathcal{N}\left(0, \alpha_1^2 X_1^2 + \cdots + \alpha_N^2 X_N^2\right) \coloneqq \mathcal{N}\left(0, \sigma^2\right) \tag{4.3}$$

This is for a given realization of the X^i's. But, the X^i's themselves are random variables with an underlying distribution. Thus, we may ask how σ^2 is distributed. Because $\alpha_i^2 = 1/\beta_i^2$, that is, the reciprocal of the variance of X^i, σ^2 is the sum of squares of N normally distributed random variables with zero mean and unit variance. Hence, σ^2 follows a Chi-squared distribution with N degrees of freedom, $\sigma^2 \sim \chi^2(N)$. To calculate the probability distribution of Y, that is, $P(Y = y)$, we

need to marginalize over this $\chi^2(N)$ because the variance is now a random variable:

$$\begin{aligned} P(Y=y) &= \int_0^\infty \mathrm{d}\sigma^2 \; P(y|\sigma^2)\, P(\sigma^2) \\ &= \int_0^\infty \mathrm{d}\sigma^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-y^2}{2\sigma^2}\right] \chi^2(N,\sigma^2) \\ &= \sqrt{\frac{1}{\pi}\left(\frac{|y|}{2}\right)^{N-1}} \quad \frac{\mathrm{K}\left(\frac{N-1}{2},|y|\right)}{\Gamma\left(N/2\right)}, \end{aligned} \tag{4.4}$$

where K is the modified Bessel function of the second kind. This distribution can be considered to be the generalization of the well-known distribution of the random variable that is the product of two standard Gaussian variables. The latter corresponds to the $N = 1$ case of the former.

As the dependence of Y is only on N, one may wonder where the distributions of $\vec{a}$ and $\vec{X}$ enter. It is only because of the choice of the variances of the distribution, $\alpha_i^2 = 1/\beta_i^2$, that the dependence on the details of the distribution "cancels out". So, as promised earlier, we have shown that a specific choice of the distribution for the co-efficients ($\vec{a}$ in this case) results in a very simple form for the distribution of the statistic.

Usually, the word (test) statistic is reserved for a function of the data. In particular, for each set of data, such a statistic returns a single number. In our case, by construction, the Y statistic is not a single number because a given $\vec{X}$ is multiplied by several $\vec{a}$. We shall call such statistics *vector statistics*.

4.2.3 Realizations

In cosmology, we have only one realization of the Universe. For our purposes, this translates into one realization of the $a_{\ell m}$'s, which we take to be the *real* multipole co-efficients corresponding to the real spherical harmonics $Y_{\ell m}$'s (see, for instance, Appendix A of [143]). The index m then ranges from $[-\ell, \ell]$. But, the Concordance Model of Cosmology predicts that each $a_{\ell m}$ is a random variable, arising from a

Gaussian distribution $\mathcal{N}(0, C_\ell)$. This can be thought of as the null hypothesis H_0.

Thus, the $a_{\ell m}$'s are like our $\vec{X}$ and we shall refer to them as $\vec{X}$ in order to keep matters general. One way of testing H_0 is by considering different test statistics of $\vec{X}$ and seeing if their realized values are compatible with that predicted by the null hypothesis. A trouble with this method is that $\vec{X}$ doesn't have a basis-independent definition—its meaning depends on the coordinate system employed in the sky. Further, the p-values that one derives depend fundamentally on the test statistic chosen. So, just because one such p-value is compatible with the null hypothesis doesn't mean that the data are.

In our case, note that the Y statistic is independent of the co-ordinate system chosen in the sky. To see this, consider a (passive) rotation of the co-ordinate system. It can be shown that the transformed $\vec{X}$, say $\vec{X}'$, is related to $\vec{X}$ by a real orthogonal matrix[3], say $\mathcal{R}$; that is, $\vec{X}' = \mathcal{R} \cdot \vec{X}$. The Y statistic arising out of $\vec{X}'$, say $Y' = a_i (X')^i = a_i \mathcal{R}^i{}_k X^k = \mathcal{R}_k{}^i a_i X^k \coloneqq (a')_i X^i$. Using (4.1) and $\mathcal{R}^T \mathcal{R} = \mathbb{1}$, it is clear that the PDFs for a_i and $(a')_i$ are the same and hence the Y statistic is co-ordinate system independent.

Now that we have discussed the test statistic and its properties, let us detail our motivation for considering this statistic and what we intend to do with it. One might wonder why a linear combination of the components of $\vec{X}$ is being considered. This has to do with the kind of anomalies that are usually discussed. It is very natural to assume that these anomalies are the result of some correlation between the different components of $\vec{X}$. Indeed, many models that attempt to explain these anomalies posit precisely such a correlation (see [146] and references therein for a review of the anomalies and some proposed explanations for their origins). The only way to test these correlations is by considering functions that "mix" the different components. A linear superposition is just the simplest of these functions. We shall consider second-

[3]The transformation matrix is given by $\mathcal{C}^* \mathcal{D} \mathcal{C}^T$ [144], where $\mathcal{C}$ is a matrix that relates the complex spherical harmonics to the real ones, and $\mathcal{D}$ is the Wigner D-matrix [145] that describes how complex spherical harmonics transform under rotations. Both matrices are unitary and * denotes complex conjugation.

order statistics in due course.

We now consider a more operational definition of $\vec{X}$. We specialize to the case where $\vec{x}$, the realization of $\vec{X}$, is the set of $a_{\ell m}$'s. That is $x_1 = a_{2,-2},\ x_2 = a_{2,-1}, \ldots, x_5 = a_{2,2};\ x_6 = a_{3,-3}, \ldots; x_N = a_{\ell_{\max},-\ell_{\max}}$.[4] Here, $\ell_{\max}$ is the largest ℓ value that we go up to:

$$\ell_{\max}^2 + 2\ell_{\max} - (3 + N) = 0 \tag{4.5}$$

The strategy is the following: Under the null hypothesis H_0, we have the distribution for Y, given in (4.4). From CMB experiments such as WMAP and Planck, we have the realized values of $\vec{X}$ in the actual sky. We use these realized values of $\vec{X}$, $\vec{x}_{sky}$, and determine the distribution of Y, $P(y_{sky})$. We can compare this distribution with (4.4) and can then infer the compatibility of CMB data with H_0.

4.3 Hypothesis Testing

Usually, hypothesis testing involves calculating the probability of the realized value of a statistic, given the distribution of the statistic under the assumption of the null hypothesis. This procedure cannot be directly implemented in our approach because, by construction, our test statistic Y doesn't yield a single number for a given dataset—it is a vector statistic. So, whereas in the usual case we only have to compare one realized value of the test statistic with the expected value, in our case, by its very nature, we must compare the realized distribution $P(y_{sky})$ with that in (4.4).

Now, there is no unique way of comparing two arbitrary distributions. As we are basically looking for a measure of goodness-of-fit, we could consider a chi-squared test. But, chi-squared tests are more useful in circumstances where one is estimating the parameters in a given model. In that case, minimising chi-squared leads to the best-

[4] As is usual in CMB analyses, we ignore the monopole and the dipole ($\ell = 0$ and $\ell = 1$).

fit parameters. That is not what we are doing here. We are actually comparing data with a fiducial distribution function. Moreover, using the chi-squared test involves binning the data, and some information is lost in this process. It would be more desirable to work with tests that use the data themselves, not bins of data.

Different such tests have been proposed in the literature, and we shall adopt the Anderson-Darling (A-D) test [147], which we shall describe shortly. The reason for the choice is that studies [148] have shown that, for a variety of distributions, this test is more powerful than others such as the more commonly used Kolmogorov-Smirnov (K-S) test. A possible drawback of using the A-D test instead of, say, the K-S test is that the critical values depend on the distribution corresponding to the null hypothesis, but, because we know the form of this distribution (4.4), the critical values can be calculated. Moreover, this dependence on the distribution is reflective of the fact that the A-D test is much more sensitive to the underlying distribution than the K-S test, and hence more powerful.

4.3.1 Anderson-Darling Test

Let V be a random variable and let the null hypothesis be that the (continuous) probability distribution $F(V)$ describes this variable. Further, let the m-component vector v_i represent m samples of V, sorted in increasing order. Define $\Phi(w)$ to be the cumulative distribution function,

$$\Phi(w) = \int_{-\infty}^{w} dv\, F(v)$$

Also, define

$$S = \frac{1}{m}\sum_{i=1}^{m}(2i-1)\Big(\log[\Phi(v_i)] + \log[1-\Phi(v_{m+1-i})]\Big) \tag{4.6}$$

The A-D statistic is then given by

$$A^2 = -m - S \tag{4.7}$$

For well-known distributions, such as the normal distribution, critical values of the A^2 statistic have been calculated in the literature. Associated with each critical value is a p-value, with which the null hypothesis can be rejected at the corresponding significance. For example, a value of A^2 more than 3.857 would mean rejecting the null hypothesis that the data are described by a normal distribution with a given mean and variance at the 1% level.[5]

As our distribution (4.4) is not one of the common distributions (the earliest reference to it that we could find is in [149]), published critical values for the A-D test do not exist. But, for a given N, we can determine them simply by generating a large number of realizations drawn from (4.4), calculating the corresponding value of A^2, and repeating this procedure a sufficient number of times. This would give us the distribution of A^2 for (4.4), from which the critical values can be calculated. Call this distribution $\Psi_Y(A^2, N)$.

A peculiar feature arises out of the fact that we only have access to one realization of the $a_{\ell m}$'s. For typical PDFs, the distribution of A^2 in (4.7) asymptotes fairly quickly to a fixed distribution as the number of realizations (m in the equation) increases. But, recall that we have only one realization of $\vec{X}$. So, even if we increase the number of Y statistics generated (thereby increasing the corresponding m), this is not equivalent to an ergodic sampling of the distribution. In particular, if we choose, say, $m = 10^5$, then, it *does* matter whether we generate m realizations of Y by choosing 10^5 realizations of $\vec{a}$ and 1 realization of $\vec{X}$, or by choosing 10^3 realizations of $\vec{a}$ and 10^2 realization of $\vec{X}$. Thus, it turns out that in our case the distribution

[5] This is in the limit of infinite data, and for data that have been standardised (subtract the mean from the data, and divide by the standard deviation), though, for the case of the normal distribution, modifications for finite m exist.

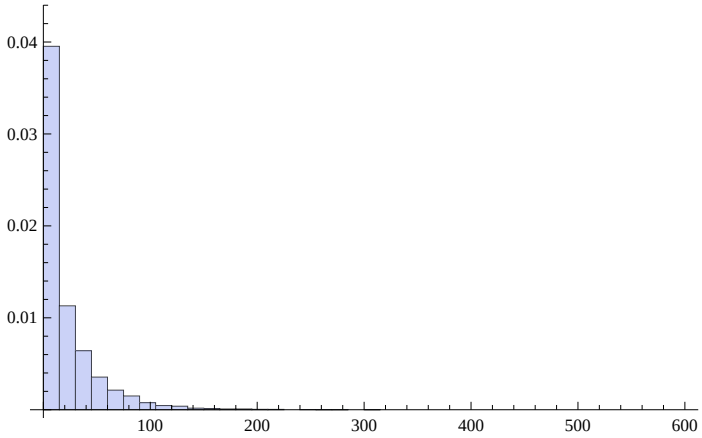


Figure 4.1: PDF for $\Psi_Y(A^2, N, m)$ for $m = 10^5$, $N = 672$.

of A^2 for a given N, what we called $\Psi_Y(A^2, N)$, depends on m. We shall denote this distribution by $\Psi_Y(A^2, N, m)$. Implicit in this notation is the fact that we are choosing only one realization of $\vec{X}$.

For a given N and m, we can determine $\Psi_Y(A^2, N, m)$ numerically by simply following the procedure outlined in (4.2): We choose both a and X to be N-dimensional normal vectors with zero mean and unit variance. We pick one realization of $\vec{X}$ and m realizations of $\vec{a}$, calculate the corresponding Y and one corresponding realization of $\Psi_Y(A^2, N, m)$. We repeat this procedure several times until we have mapped out the distribution $\Psi_Y(A^2, N, m)$ reasonably well. This distribution for $N = 672$ and $m = 10^5$ is shown in Figure 4.1.

Once we have determined $\Psi_Y(A^2, N, m)$, putting limits on how anomalous the data are, in terms of our formalism, is relatively straightforward. We have discussed in the previous section how we can generate a given number (say, m) of the Y statistic. On sorting this data vector in increasing order, we can proceed to calculate the realized value of A^2, with the distribution in (4.4) corresponding to $F(V)$ above. Call this $a^2_{\rm sky}$.[6] This value can be compared with $\Psi_Y(A^2, N, m)$ and a p-value can be calculated.

[6]Henceforth, we shall follow the standard practice of representing random variables by capital letters and realized values by small letters.

4.4 Z—A Quadratic Statistic

4.4.1 Ensemble

In the previous section, we considered a linear combination of the different $a_{\ell m}$'s. We mentioned that if the anomalies are real, then models that could produce these anomalies without correlations amongst the different $a_{\ell m}$'s would likely be very contrived. So, to probe these correlations better, it is natural to consider test statistics that are second order in the $a_{\ell m}$'s. We shall do that here and revert to using $\vec{X}$ to denote the ordered set of $a_{\ell m}$'s.

Consider the following test statistic:

$$Z = B_{12}X_1X_2 + B_{34}X_3X_4 + \cdots + B_{N-1N}X_{N-1}X_N \tag{4.8}$$

For now, assume that N is even, so that this definition always makes sense (N is even for an odd $\ell_{\max}$). We shall comment on dealing with an odd N later.

B_{ij} is a random variable distributed as $\mathcal{N}\left(0, \sigma^2_{Bij}\right)$, where $\sigma^2_{Bij} = \frac{1}{X_i^2\beta_j^2}$. Recall that β_j^2 is the variance of the normally distributed X_j. Note that we are using X_i itself as a parameter describing a distribution. (Compare this with the distribution of $\vec{a}$, which depended on the variance of $\vec{X}$, and not on $\vec{X}$ itself.) This is not an issue because $\vec{X}$ is still being treated as a fixed vector. The reason for this choice of σ^2_{Bij} will become clear momentarily, but, it must be borne in mind that it gets determined *after* a choice of $\vec{X}$ is made.

With this, Z is basically a sum of $N/2$ Gaussian random variables B_{ij}, with constant coefficients X_iX_j. Thus, we have that $Z \sim \mathcal{N}\left(0, \sigma_Z^2\right)$, where

$$\begin{aligned}
\sigma_Z^2 &= X_1^2X_2^2\sigma^2_{B12} + \cdots + X^2_{N-1}X^2_N\sigma^2_{B,N-1,N} \\
&= \frac{X_2^2}{\beta_2^2} + \cdots + \frac{X_N^2}{\beta_N^2} \\
&\sim \chi^2(N/2).
\end{aligned}$$

Here, like in the analysis in Section 4.2.2, we have used the fact that σ_Z^2 is the sum of the squares of $N/2$ normally distributed, zero-mean random variables with unit variance. Similar to what we did for the test statistic Y, we now perform an ensemble average of Z with respect to the distribution of $\vec{X}$. Repeating the calculation that led to (4.4), with half the number of terms, we have that the distribution of Z is

$$P(Z=z) = \sqrt{\frac{1}{\pi}\left(\frac{|z|}{2}\right)^{(N/2-1)}} \quad \frac{\mathrm{K}\left(\frac{N-2}{4}, |z|\right)}{\Gamma\left(N/4\right)} \tag{4.9}$$

Again, because of the choice of the distribution of the B_{ij} variables, the distribution of Z is solely a function of N. This is quite a useful feature for the following reason: Consider four random variables R_1, R_2, R_3, R_4. Let only R_1 and R_3 be correlated, and R_2 and R_4 be correlated:

$$\langle R_1 R_3 \rangle = \langle R_2 R_4 \rangle = \epsilon, \text{ where } \epsilon \ll 1 \tag{4.10}$$

Now, say you are testing the null hypothesis that all four variables are mutually independent. You come up with two test statistics, $T_1 \coloneqq R_1 R_2 + R_3 R_4$ and $T_2 \coloneqq R_1 R_3 + R_2 R_4$. From (4.10), it is clear that $\langle T_1 \rangle$ is indistinguishable from that predicted by the null hypothesis, whereas $\langle T_2 \rangle$ gives a different prediction from the null hypothesis. Of course, the distribution of both T_1 and T_2 will be different from that predicted by the null hypothesis, but, at least for non-pathological distributions, T_1 is an $\mathcal{O}(\epsilon)$ worse discriminator for testing the null hypothesis.

If the CMB anomalies are due to correlations amongst the different $a_{\ell m}$'s, from the form of (4.8), one may naïvely worry that just like with the R_i's, the order in which the $a_{\ell m}$'s appear in the equation may matter. That is, instead of the order in (4.8), one could alternatively consider

$$Z' = B_{13} X_1 X_3 + B_{24} X_2 X_4 + \cdots + B_{N-2,N} X_{N-2} X_N$$

This is a different statistic from Z. In this manner, there are $(N-1)!!$ alternatives to Z.[7] In principle, each of these combinations will have a different distribution for Z. But, because of our choice of the distribution of B_{ij}, we have that the distribution of Z depends only on N. With this motivation, let us define Perm(Z) as a permutation of the indices in Z that ensures that each index appears once and only once. Now, define $\widetilde{Z}$ as the set of all Perm(Z). It is obvious that $\widetilde{Z}$ is distributed as (4.9). It is $\widetilde{Z}$ that is the statistic that we shall consider for the rest of this chapter, though, by abuse of notation, we shall refer to it as Z. In this way, the choice of the distribution function for B_{ij} helps us overcome the difficulty of having to consider $(N-1)!!$ different distributions, while ensuring that there is no loss of generality in the sequence of indices chosen.

Earlier, we had stated that we would discuss the case with an odd N, which arises if we have an even $\ell_{\max}$. In that case, we can just consider pairs of the first $(N-1)$ of the indices of Perm($\{1,\ldots,N\}$), which occurs in $\widetilde{Z}$ anyway. This would mean that we are losing out on one mode during every permutation, but, the procedure ensures that there isn't any arbitrariness in the choice of that mode.

4.4.2 Hypothesis Testing

The procedure of testing the null hypothesis H_0 is identical to the one we employed for the linear test statistic Y. The expected distribution under H_0 is given by (4.9) and we can use actual data to determine the realized distribution. We can then calculate the statistical significance of a departure from H_0 by using the procedure outlined in the previous section. Let us denote the probability distribution function for the Anderson-Darling statistic for the Z statistic as $\Psi_Z(A^2, N, m)$. We can repeat the procedure outlined in 4.3.1 to determine this PDF—the only change will be that,

[7] Consider the sequence $\{1..N\}$. Each index has to occur once. So, there is no freedom in choosing the i in (4.8). For the j corresponding to the first term, there are $(N-1)$ possibilities. Again, the i for the second term is effectively fixed, as it must appear in the sum. For the j corresponding to this term, there are $(N-3)$ possibilities, and so on.

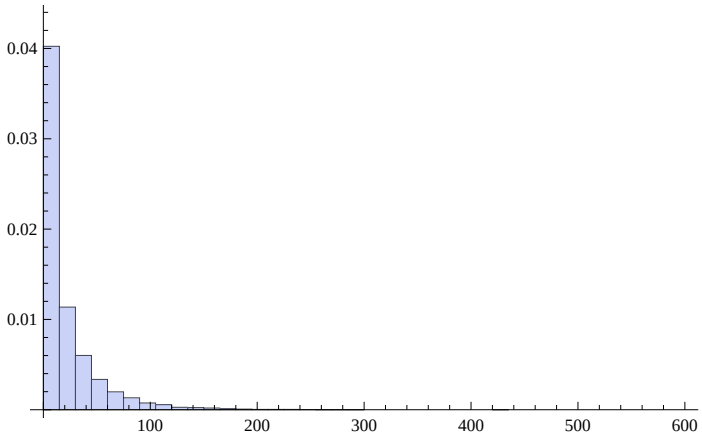


Figure 4.2: PDF for $\Psi_Z(A^2, N, m)$ for $m = 10^5$, $N = 672$.

instead of generating distributions of the Y statistic, we will generate distributions of the Z statistic, and, instead of using (4.4), we shall use (4.9). For a particular choice of N and m, this PDF is shown in Figure 4.2. Figures 4.1 and 4.2 look to be very similar, and we have confirmed this for other values of $\ell_{\max}$. That is, for a given $\ell_{\max}$, the distribution of the A-D statistic is the same for both the Y and the Z statistics. The distributions *are* different for different values of $\ell_{\max}$. For illustrative purposes, we have also plotted the PDFs for $m = 10^5$ and $N = 252$ (which corresponds to $\ell_{\max} = 15$) in Figure 4.5. This figure appears just before the concluding section of the chapter.

4.5 Results

Having discussed the method for testing for the null hypothesis in the previous sections, in this section, we demonstrate that the method actually works. To do this, we break one of the assumptions in the null hypothesis. The easiest condition to break (in the sense that the new probability distribution is easiest to describe) is that of zero-mean. Previous studies [143] have looked at relaxing this condition, though they concentrate on somewhat larger values of ℓ. They found that, at least in the range of multipoles they considered, the data seemed to be consistent with the zero-mean

hypothesis. Here, we choose to break the condition of independence and normal distribution of the $a_{\ell m}$'s, mostly because that is usually posited as the reason behind the anomalies. But, we should emphasise that a similar analysis can be performed (in fact, more easily) with a non-zero mean.

Now, there is an infinite number of ways of breaking the independent, normally distributed hypothesis [150]. We break it by deliberately masking the fiducial CMB sky about the equator. This masking breaks statistical isotropy and thus leads to a correlation between modes.[8] The resulting probability distribution of the $a_{\ell m}$'s is difficult to analytically estimate, but it is clear that a greater degree of masking leads to a "bigger" departure from the null hypothesis. Then, the strategy behind the demonstration is this:

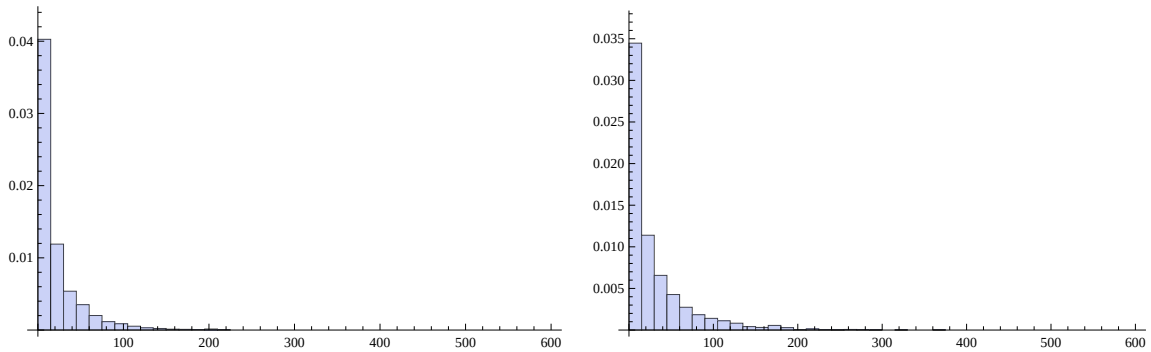
1. Generate a set of fiducial CMB sky maps from a known set of C_ℓ's.
2. Generate Y and Z statistics using the $a_{\ell m}$'s of these maps.
3. Mask these maps to varying degrees and determine the resulting $a_{\ell m}$'s, and Y and Z statistics.

The method can then be said to work if increasing the masking leads to a bigger departure from the null hypothesis (in the sense of the Anderson-Darling test applied to the Y and Z statistics). Also, for zero masking, the distribution one gets with the CMB maps must correspond to $\Psi_Y(A^2, N, m)$ and $\Psi_Z(A^2, N, m)$ respectively.

As mentioned earlier, one of the things that we need to pick is the range of ℓ's that we will be considering. Because we are concentrating on low-ℓ anomalies, we start with the lowest relevant ℓ ($\ell = 2$) and go up to an $\ell_{\rm max}$. For the rest of this section, let us choose $\ell_{\rm max} = 25$. From (4.5), this corresponds to $N = 672$.

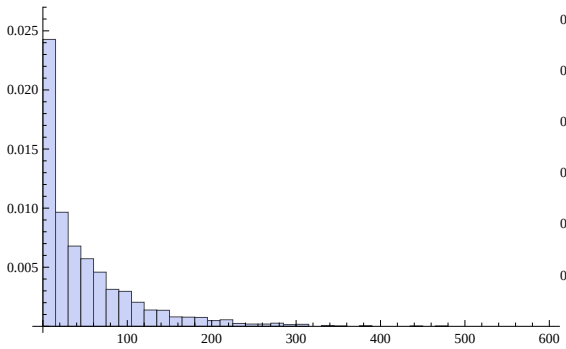
Therefore, what we now need to do is to generate m realizations of Y and Z for each of the CMB maps described in the strategy above and compare this distribution with $\Psi_Y(A^2, N, m)$ for Y and $\Psi_Z(A^2, N, m)$ for Z. We employ routines in HEALPix[9]

[8] See [151] for more discussion on breaking statistical isotropy in the context of CMB anomalies.
[9] http://healpix.sourceforge.net/

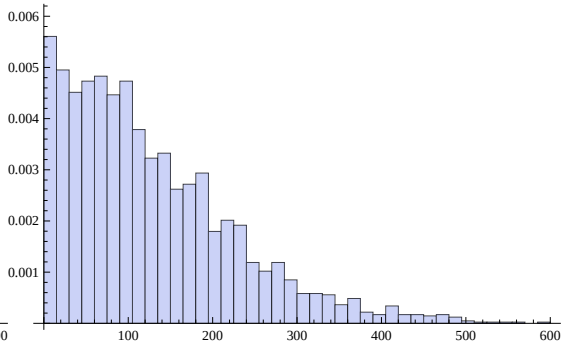


(a) No galactic mask

(b) 5% of pixels about the equator masked



(c) 10% of pixels about the equator masked

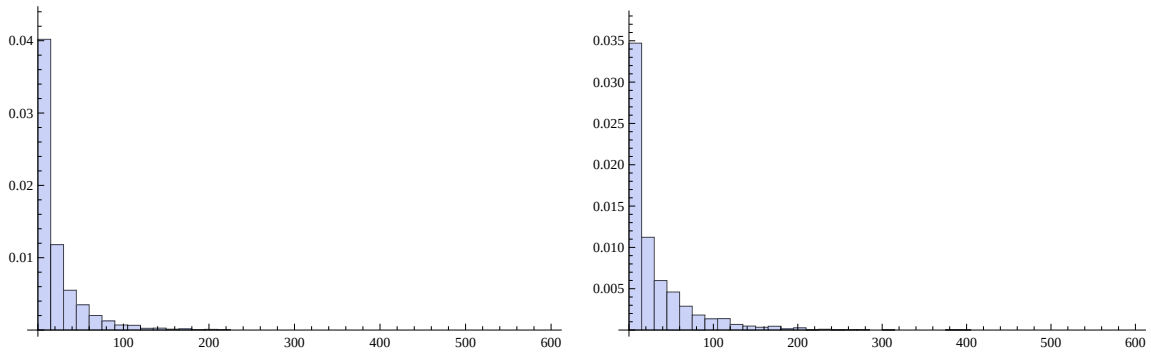


(d) 15% of pixels about the equator masked

Figure 4.3: PDF for $\psi_y(a^2, N, m)$ for $N = 672$, $m = 10^5$, and for different masks.

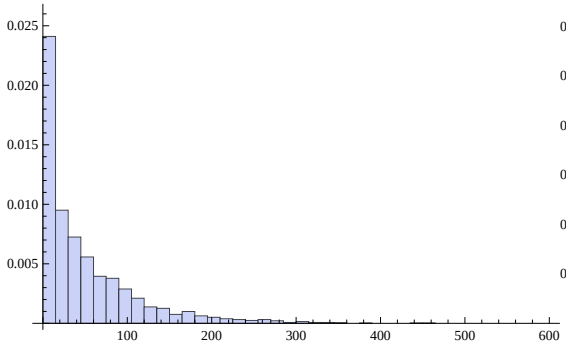
[152] to generate CMB maps from a given set of C_ℓ's, mask the maps, and then determine the corresponding $a_{\ell m}$'s. For the C_ℓ's, we use the Planck best-fit values, though, because this is for testing, any reasonable set would be sufficient. We consider four sets of maps: unmasked, and a mask of 5%, 10% and 15% of the pixels about the galactic equator. We choose $m = 10^5$ so that we can compare the distribution of the realized vector statistic with that in Figures 4.1 and 4.2. We use C++ to generate the Y and Z statistics and MATHEMATICA to calculate the A-D statistic.

For the Y statistic, the results are plotted in Figure 4.3. As expected, the distribution for the unmasked sky [Figure 4.3 (a)] resembles that in Figure 4.1 to a very high degree, and the other three to a much lesser degree. Clearly, a bigger mask, and thus a bigger departure from statistical isotropy (and the null hypothesis), leads to a bigger departure of the distribution from that in Figure 4.1. Similar results hold for

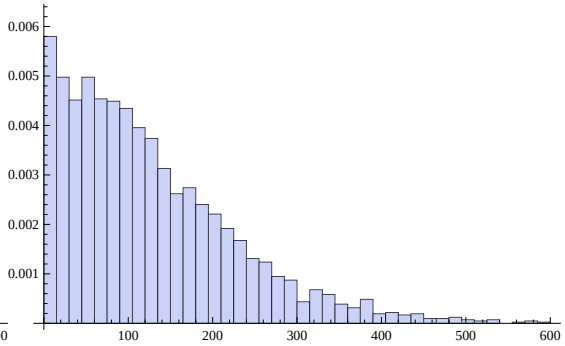


(a) No galactic mask

(b) 5% of pixels about the equator masked



(c) 10% of pixels about the equator masked

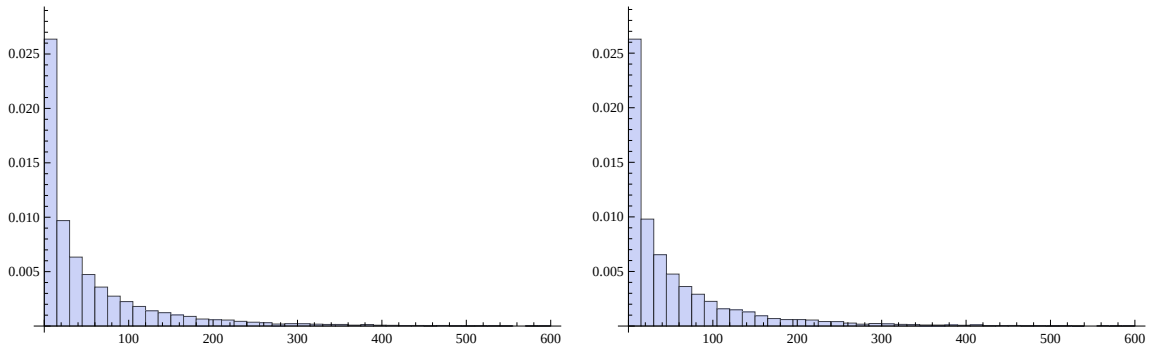


(d) 15% of pixels about the equator masked

Figure 4.4: PDF for $\psi_z(a^2, N, m)$ for $N = 672$, $m = 10^5$ and for different masks.

the Z statistic, plotted in Figure 4.4. These plots are to be compared with those in Figure 4.2.

In the calculation of the Y and Z statistics, we need the values of the C_ℓ's of the distribution in the null hypothesis. For the unmasked sky, it is clear that these variances are the same C_ℓ's with which the fiducial map was generated. For the masked skies, it isn't immediately obvious what C_ℓ's one must choose. We choose them to be the same as that for the unmasked sky for the following reason: Because the anomalies are said to occur on large scales, we have restricted our analysis to an $\ell_{\max}$ of 25. The ΛCDM hypothesis is implicit in our tests. In particular, the theoretical C_ℓ's, the ones that form part of the null hypothesis, are estimated by first determining the best-fit parameters in the ΛCDM model. That is, these parameters uniquely determine the null hypothesis. By cosmic variance, the measurements at



(a) $\Psi_Y(A^2, N, m)$ (b) $\Psi_Z(A^2, N, m)$

Figure 4.5: PDFs for $m = 10^5$, $N = 252$

very low ℓ's (of the range that we are considering) contribute negligibly towards this determination of the best-fit parameters. So, in a sense, the values of the C_ℓ's at very low ℓ's are fixed by measurements at higher ℓ's. Thus, even if the masking affects the estimation at very low ℓ's, it is safe to compare the masked and the unmasked sky with the same variances that generated the maps.

4.6 Summary and Conclusion

The last couple of decades have seen a tremendous amount of progress in the understanding of the large-scale structure of our Universe. Some parameters have been determined to several decimal places and some models have been ruled out to extremely high significance. Observationally, the only real challenge to this ΛCDM paradigm seem to be the large-scale CMB anomalies, of which many have been reported. The most important criticism levelled against the anomalies has to do with the fact that the anomalies are an *a posteriori* phenomenon—one tests for anomalies after having "looked" at the data. This is a fair criticism and in this paper we have proposed a method that addresses this very criticism. In a very general manner, we have sought to test the null hypothesis that the $a_{\ell m}$'s are independent, zero-mean, normally distributed variables with an m-independent variance.

We considered linear (Y) and quadratic (Z) combinations of the $a_{\ell m}$'s, with randomized co-efficients. The probability distribution of these co-efficients is of a very specific form, but, depends only on the C_ℓ's. This choice greatly simplifies the PDFs of Y and Z. Given a CMB map, the Y and Z distribution corresponding to the $a_{\ell m}$'s of the map can be determined. This distribution can be compared with the fiducial distribution for Y and Z (given in (4.4) and (4.9) respectively) and a high degree of incompatibility between the distributions would mean that the data are not well described by the null hypothesis.

To make this comparison between distributions, we have suggested a very slight modification of the Anderson-Darling test. Of course, other tests could also be used for this purpose. In order to demonstrate the usefulness of the test, we generated CMB maps with varying degrees of masking in them. This masking breaks statistical isotropy and thus results in a departure from the null hypothesis. We demonstrated that, firstly, for zero masking, the distribution of the Anderson-Darling test statistic is what we expect it to be. Secondly, increasing the masking did lead to distributions of the Anderson-Darling test statistic that were further and further removed from the distribution that arises out of the null hypothesis.

A few points to note regarding this method are: (i) Like most other "goodness-of-fits" tests without an alternative hypothesis, this is a frequentist analysis. In particular, because of its very general and stochastic nature, the test may be susceptible to Type II Errors; that is, a failure to reject the null hypothesis. If we do have an alternative hypothesis, we can then compute the power of the test and make a quantitative statement about the probability of Type II errors. Or, indeed, do a Bayesian analysis. In the absence of this alternative hypothesis, a p-value compatible with the null hypothesis should *not* be taken to mean that the data indicate that the null hypothesis is true. (ii) In our analysis, we have mostly assumed that the C_ℓ's are fixed numbers, but, at least from a Bayesian perspective, they themselves are random variables, with an associated variance. We don't see a way around this,

because taking into account the stochastic nature of the C_ℓ's would make the analysis extremely complicated. Also, recall that the variance of C_ℓ is proportional to C_ℓ itself. Thus, for multipoles where the random nature of C_ℓ is most pronounced (that is, the lowest of the ℓ's), the value of C_ℓ is large to begin with. This partially alleviates the problem associated with assuming that the C_ℓ's are fixed numbers. (iii) Though we have concentrated on using the method to make statements about the $a_{\ell m}$'s, it is clear that our method works in general for any set of random variables that are hypothesized to be described by H_0. So, our method could be used to test H_0 in a variety of situations, becoming particularly useful when there are only a few realizations of several independent, *non*-identically distributed Gaussian variables.

In a future publication, we hope to use our method and actual CMB data to quote p-values for the departure of the data from the null hypothesis. Planck is soon expected to release CMB polarization data, which can easily be incorporated into our analysis and should tell us more about the largest scales of the observable universe.

Chapter 5

Diffeomorphism-Invariance Constraints on Cosmological Correlators

5.1 Introduction

Most studies in Cosmology involve the analysis of correlation functions of different perturbed quantities. For example, we discussed the correlation functions of CMB temperature anisotropies in Section 1.3, and used actual data of these correlation functions in Chapters 2 and 3 to draw conclusions about dark matter interactions. We have also discussed how, in the case of adiabatic perturbations, the correlation functions of metric perturbations at the end of inflation can be related to that of the fluid perturbations at the beginning of radiation-domination. As inflation most likely occurs at energies that we will not achieve here on Earth, at least in the foreseeable future, such analyses validate Zel'dovich's statement that the universe is a "poor man's accelerator".

As a theory, even classical general relativity is quite complicated, in the sense that it involves coupled, second-order partial differential equations, and closed-form

solutions don't exist in general. But, in Cosmology, because of the high degree of symmetry in the background FLRW spacetime, the evolution equations for the perturbations become much simpler. Given the evolution equations, one can always calculate the correlation between different fields at any point of time, if their initial values are specified. But, in some cases, symmetry arguments can be used to calculate the value of these correlations, thus doing away with having to evolve the fields.

For instance, consider an initial state of metric perturbations that is parity-invariant, meaning that $\delta g_{\mu\nu}(\vec{x}, t) = \delta g_{\mu\nu}(-\vec{x}, t)$. The parity-invariance of the system implies that all correlation functions of the perturbations must be parity-invariant too. We have discussed in Section 1.8 how the metric perturbations can be decomposed into irreducible components that are equivalent to the spherical harmonics $Y^{\ell m}$. The components that correspond to $m = 0$ (for example, a scalar S) are parity invariant whereas those that correspond to $m \neq 0$ (for example, a tensor T) are not. Consider the correlation function $\langle ST \rangle$. From the discussion above, we know that this state must be parity-invariant. But, as $S \to S$ and $T \to -T$ under parity, it is clear that $\langle ST \rangle \equiv 0$ at all times. We have not had to use the evolution equations of the theory at all. In fact, we have not used any information about the "details" of the theory, apart from its parity-invariance.

While this means that an observation of a vanishing $\langle ST \rangle$ doesn't teach us anything about the details of the theory, it also means that if the correlation function is observed to be non-zero, it rules out entire classes of theories that are parity-invariant. (Indeed, similar ideas were used by Lee and Yang [153] and Wu et al [154] to demonstrate parity violation in weak interactions in Particle Physics.) In this respect, symmetry arguments are very useful because they can help constrain the classes of theories that seem to describe Nature.

We have discussed in Section 1.8 that different inflationary models lead to different values for n_s and A_s. These are the only two inflationary parameters that have been estimated to a precision high enough that they can be said to have been measured. As

we discussed in Section 1.9, measurements of these two parameters aren't sufficient to pin down the exact mechanism of inflation. Different potentials with different numbers of fields, and different interactions amongst the fields can all be made to yield the same value of n_s and A_s.

It is in this context that relations that arise out of symmetries have proved to be useful in inflationary cosmology. Say R is a relation that arises out of the symmetries of one class of theories, and that it is very unnatural to expect R to hold in other classes of theories. Then, even though one may not have the precise measurement of a variable, if the measurement is good enough to show that R is not violated, it would, in effect, rule out the latter class of theories.

Let us now discuss the different kinds of relations that one can derive from symmetries. Broadly speaking, there are two kinds of symmetries in physics—spacetime (global) symmetries, that are symmetries of the system being considered; and gauge (local) symmetries, that are symmetries that arise out of the redundancies in the language used to describe the system. As the names suggest, the parameter that describes the action of a global symmetry transformation is a constant, whereas that for a local symmetry depends on spacetime.

5.1.1 Consequences of Local Symmetries

Whether local symmetries such as diffeomorphism invariance have any physical content has been a subject of intense debate ever since the inception of general relativity. Indeed, already in 1917, Erich Kretschmann argued that the principle of general covariance is physically vacuous: Any non-covariant theory ought to be made covariant without changing any of its physical predictions [155], and, conversely, any covariant theory can be made non-covariant by gauge fixing, a process that preserves the physical implications of the covariant theory.

Yet diffeomorphism (or gauge) invariance does seem to have significant physical implications. In general relativity, for instance, diffeomorphism invariance enforces

the equivalence principle [156, 157]. More generally, the invariance of gauge theories under gauge transformations severely constrains the structure of the counter-terms, and plays a crucial role in the demonstration that these theories are renormalizable. Indeed, local symmetries such as diffeomorphism and gauge invariance are the basis of our understanding of all interactions, both in the standard model and general relativity. But whether any of the physical implications of these theories do really follow from gauge invariance, or whether they are a consequence of the field content and the residual global symmetries that gauge fixing allows us to preserve, often remains somewhat obscure.

In this Chapter we study the extent to which diffeomorphism invariance constrains the properties of the primordial perturbations. We formulate a set of identities that relate different connected correlators, and also different one-particle-irreducible (1PI) diagrams in general relativity coupled to a scalar field. These identities belong to a family of relations connected to symmetry, and have appeared under different names in different contexts. They are known as Slavnov-Taylor, Ward-Takahashi or Dyson-Schwinger equations, although they are all basically equivalent: They all express the invariance of the theory under diffeomorphisms.

These identities have to be interpreted appropriately, however. In order to quantize a theory with a local symmetry, such as diffeomorphism invariance in the case at hand, the symmetry has to be explicitly broken by gauge-fixing terms. Hence, strictly speaking, the Ward-Takahashi and Slavnov-Taylor identities we discuss actually mirror the way in which diffeomorphism invariance has been broken, and therefore often depend on the particular gauge choice. Many of the identities exist because they involve the correlators of gauge-variant fields, and hence cannot have a physically invariant meaning. Actual observables do not depend on any particular gauge choice, but in this paper we will not attempt to connect our gauge-variant correlators to any gauge-invariant observables like the statistical properties of the cosmic microwave background anisotropies. Although such a connection is relatively simple in linear

perturbation theory, it is highly non-trivial beyond the linear order.

At this point the Slavnov-Taylor identities could be viewed as useful checks on the validity of intermediate cosmological perturbation theory calculations. In one-loop calculations of cosmological correlations, for instance, one needs to regularize the theory first. In some cases the regularization procedure may unintendedly break diffeomorphism invariance, and the resulting violation of the identities we derive can help diagnose such violations.

But, perhaps, the most important application of the Slavnov-Taylor identities is the derivation of "consistency relations" between the different correlators of cosmological perturbations. To illustrate our methods, we derive consistency relations that follow essentially from diffeomorphism invariance alone, although these lack the predictive power of other relations, and just seem to reflect the underlying redundancy associated with diffeomorphism invariance. The constraining power of diffeomorphism invariance changes significantly when combined with an assumption about the analyticity of the correlators of the theory [158]. In this case, analyticity allows one to go beyond what appears to follow merely from gauge redundancies, allowing one to derive physically predictive consistency relations in specific gauges, in the limit in which one of the field momenta approaches zero. In that sense, the ensuing relations are close relatives of the constraints on the vertex function that guarantee the validity of the equivalence principle in general relativity [156, 157], which also follow from diffeomorphism invariance and analyticity around zero momentum transfer.

Some of the constraints that spatial diffeomorphism invariance imposes on the primordial perturbations have been recently discussed in [159–161], and more specifically in references [158, 162, 163] by Berezhiani, Khoury and Wang. All these papers attempted to generalize or derive relations between correlation functions of cosmological perturbations that go back to the consistency condition originally discussed by Maldacena in [23].[1] Different arguments and symmetries have been used to derive

[1]Consistency relations following from diffeomorphisms have also been derived in the context of large scale structure—for more details, see for instance [164, 165].

such relations [166–168], although it appears that an approximate conformal symmetry is the cleanest way to understand their origin [169–172]. Our work is most similar to [158], which we generalize to arbitrary gauges and also extend to time diffeomorphisms.

The plan of the paper is as follows: In Section 5.2 we explore the action of diffeomorphisms on cosmological perturbations, and establish how to calculate their expectation values. In Section 5.3 we formulate identities for the connected correlators of the theory, while in Section 5.4 we derive equivalent identities for the one-particle-irreducible diagrams. We derive consistency relations between bispectra and power spectra in Section 5.5, and finally conclude and summarize in Section 5.6.

5.2 Diffeomorphism Invariance

We are interested here in theories whose action is invariant under (infinitesimal) diffeomorphism transformations. The resulting equations of motion are then automatically covariant, so we take diffeomorphism invariance and general covariance to be synonymous. We assume that these theories describe gravity coupled to a scalar field, so their action is of the general form

$$S = S[g_{\mu\nu}, \phi] \equiv \int_{\mathcal{M}} d^4x \sqrt{-g}\, \mathcal{L}, \tag{5.1}$$

where the Lagrangian density $\mathcal{L}$ depends on the metric $g_{\mu\nu}$, the scalar ϕ, and their derivatives, and where the integral runs over the spacetime manifold $\mathcal{M}$. Although we focus on a single scalar for simplicity, our results can easily be generalized to accommodate further scalar fields.

Under passive diffeomorphisms $x^\mu \to x^\mu - \xi^\mu(x)$ any tensor field $\mathbf{T}$ transforms as follows:

$$\mathbf{T} \to \mathbf{T} + \pounds_\xi \mathbf{T}, \tag{5.2}$$

where $\pounds_\xi$ is the Lie-derivative along ξ. Hence, if the Lagrangian density transforms like a scalar, $\mathcal{L} \to \mathcal{L} + \xi^\mu \partial_\mu \mathcal{L}$, the change in the action (5.1) under infinitesimal diffeomorphisms is given by

$$\Delta S = \int_{\mathcal{M}} d^4x \sqrt{-g}\, \nabla_\mu (\mathcal{L}\, \xi^\mu) = \int_{\partial\mathcal{M}} d^3x \sqrt{\gamma}\, n_\mu\, \mathcal{L}\, \xi^\mu, \tag{5.3}$$

where n^μ is the normal to the boundary $\partial\mathcal{M}$ and γ the determinant of its metric. Hence, in the diffeomorphism invariant theories we discuss here the action is actually only invariant up to boundary terms.

5.2.1 Cosmological Background

Our goal is to constrain correlators of cosmological perturbations, that is, field fluctuations around a cosmological background. Therefore, we expand the metric and the scalar field as

$$g_{\mu\nu} \equiv \bar{g}_{\mu\nu} + h_{\mu\nu}, \quad \phi \equiv \bar{\phi} + \varphi, \tag{5.4}$$

where $\bar{g}_{\mu\nu}$ and $\bar{\phi}$ are the background values of the metric and the scalar, and $h_{\mu\nu}$ and φ are its fluctuations. We choose our cosmological background to be that of a spatially flat universe filled by a homogeneous scalar,

$$d\bar{s}^2 = a^2(\eta) \left[-d\eta^2 + d\vec{x}^{\,2}\right], \quad \bar{\phi} = \bar{\phi}(\eta). \tag{5.5}$$

Then, from equation (5.2), the perturbations around this background transform according to

$$\Delta h_{\mu\nu} = g_{\mu\alpha}\partial_\nu \xi^\alpha + g_{\alpha\nu}\partial_\mu \xi^\alpha + \xi^\alpha \partial_\alpha g_{\mu\nu}, \quad \Delta\varphi = \xi^\alpha \partial_\alpha \phi, \tag{5.6}$$

where $g_{\mu\nu}$ and ϕ are to be replaced by the corresponding expressions in equations (5.4). Note that these transformations are valid to first order in ξ (which we take to be infinitesimal), but to all orders in the fluctuations. In particular, diffeomorphisms act linearly (albeit non-homogeneously) on the field perturbations $h_{\mu\nu}$ and φ.

5.2.2 Cosmological Perturbations

The isometry group of the background, those diffeomorphisms under which the background fields are invariant, plays a particularly important role in cosmological perturbation theory. Just as it is convenient to classify fields in Minkowski space according according to their transformation properties under the isometries of the Minkowski metric, it turns out to be convenient to classify cosmological perturbations in terms of their transformation properties under spatial translations and rotations. We thus introduce a set of eleven tensors $Q_{\mu\nu}{}^{f}(\vec{x};\vec{p})$ and $Q^{\varphi}(\vec{x};\vec{p})$ that transform irreducibly under translations and rotations [173]. We list the components of these tensors in Appendix 5.A. What matters to us here is that we can expand the metric and scalar fluctuations in terms of these tensors,

$$h_{\mu\nu}(\eta,\vec{x}) = \sum_{f}\int d^3p\, Q_{\mu\nu}{}^{f}(\vec{x};\vec{p})\, f(\eta,\vec{p}), \quad \varphi(\eta,\vec{x}) = \int d^3p\, Q^{\varphi}(\vec{x};\vec{p})\varphi(\eta,\vec{p}), \tag{5.7}$$

where the sum over f runs over the ten metric perturbation fields in momentum space

$$f \in \{A, B, H_L, H_T, B_+, B_-, H_+, H_-, H_{++}, H_{--}\}. \tag{5.8}$$

The fields $f = f, \varphi$ are eigenvectors of spatial translations by $\vec{a}$ [with eigenvalues $\exp(-i\vec{p}\cdot\vec{a})$], and spatial rotations by an angle θ around the $\vec{p}$ axis [with eigenvalues $\exp(-im\theta)$, where $m = 0$ for $f \in \{A, H_L, H_T, B, \varphi\}$ (scalars), $m = \pm 1$ for $f \in \{B_\pm, H_\pm\}$ (vectors) and $m = \pm 2$ for $f \in \{H_{\pm\pm}\}$ (tensors)]. Conversely, given arbitrary metric and scalar perturbations $h_{\mu\nu}(\eta,\vec{x})$ and $\varphi(\eta,\vec{x})$ we can determine the corresponding perturbation variables with the projection operators $Q_f{}^{\mu\nu}(\vec{p};\vec{x})$ and $Q_\varphi(\vec{p};\vec{x})$, whose components we also gather in Appendix 5.A. By definition, we thus have

$$f(\eta,\vec{p}) = \int d^3x\, Q_f{}^{\mu\nu}(\vec{p};\vec{x})h_{\mu\nu}(\eta,\vec{x}), \quad \varphi(\eta,\vec{p}) = \int d^3x\, Q_\varphi(\vec{p};\vec{x})\varphi(\eta,\vec{x}). \tag{5.9}$$

Notice that the decomposition (5.7) of the metric fluctuations is equivalent to the following parametrization of the perturbed line element:

$$\begin{aligned} ds^2 &= a^2(\eta)\left\{-(1+2A)d\eta^2 + 2\left(\frac{\partial_i B}{\sqrt{\nabla^2}} + B_i\right)dx^i d\eta \right. \\ &\left. + \left[\delta_{ij}(1+2H_L) + 2\left(\frac{\delta_{ij}}{3} - \frac{\partial_i\partial_j}{\nabla^2}\right)H_T + 2\frac{\partial_{(i}H_{j)}}{\sqrt{\nabla^2}} + H_{ij}\right]dx^i dx^j\right\}, \end{aligned} \tag{5.10}$$

where B_i and H_i are two transverse vectors with polarizations $B_\pm$ and $H_\pm$ respectively, while H_{ij} is a traceless and transverse tensor with polarizations H_{++} and H_{--}.

To simplify the notation it shall be convenient to simplify our equations by switching to DeWitt notation [174], in which Latin indices $a, b, \ldots$ collectively denote the type of field and its spacetime arguments, and functional derivatives are treated as partial derivatives $\partial F/\partial f_a \equiv F_{,a}$, and also denoted simply by F^a where confusion is not likely. Along the same lines, the index α shall denote both the components and the spacetime argument of the diffeomorphism parameter $\xi^\alpha(x)$. Indices in opposite locations imply both a sum over type of fields or parameter components, and an integral over spacetime arguments.

For example, because diffeomorphisms are linear and inhomogeneous, we shall write equations (5.6) as

$$\Delta_a = \left(\mathcal{S}_{a\alpha} + \mathcal{T}_a{}^b{}_\alpha\, f_b\right)\xi^\alpha, \tag{5.11}$$

where Δ_a is the change of the field f_a under diffeomorphisms, and, in real space, the non-vanishing components of the "tensor" $\mathcal{S}_{a\alpha}$ are

$$\mathcal{S}_{h_{\mu\nu}(x)\xi^\alpha(y)} = \left[\bar{g}_{\mu\alpha}\frac{\partial}{\partial x^\nu} + \bar{g}_{\alpha\nu}\frac{\partial}{\partial x^\mu} + \frac{\partial \bar{g}_{\mu\nu}}{\partial x^\alpha}\right]\delta^{(4)}(x-y), \tag{5.12a}$$

$$\mathcal{S}_{\varphi(x)\xi^\alpha(y)} = \frac{\partial\bar{\phi}}{\partial x^\alpha}\delta^{(4)}(x-y). \tag{5.12b}$$

The free action for the perturbations is invariant under transformations with $\mathcal{T}_a{}^b{}_\alpha \equiv 0$, which is why we refer to (5.12) as the transformation of the fields under "linear dif-

feomorphisms." It readily follows from equation (5.6) that for the isometries of the background, namely, translations ($\bar{\xi}^\alpha = \delta^\alpha{}_i$) and rotations ($\bar{\xi}^\alpha = \epsilon^\alpha{}_{ij}x^j$), the corresponding linear transformations vanish, $\mathcal{S}_{a\alpha}\,\bar{\xi}^\alpha = 0$. The non-vanishing components of diffeomorphism transformations linear in the field perturbations themselves, $\mathcal{T}_a{}^b{}_\alpha$, are given by

$$\mathcal{T}_{h_{\mu\nu}(x)}{}^{h_{\rho\sigma}(y)}{}_{\xi^\alpha(z)} = -\left[\delta_{\mu\alpha}{}^{\rho\sigma}\frac{\partial}{\partial z^\nu} + \delta_{\alpha\nu}{}^{\rho\sigma}\frac{\partial}{\partial z^\mu} + \delta_{\mu\nu}{}^{\rho\sigma}\frac{\partial}{\partial y^\alpha}\right]\delta^{(4)}(x-y)\delta^{(4)}(x-z), \tag{5.13a}$$

$$\mathcal{T}_{\varphi(x)}{}^{\varphi(y)}{}_{\xi^\alpha(z)} = -\frac{\partial}{\partial y^\alpha}\delta^{(4)}(x-y)\delta^{(4)}(x-z). \tag{5.13b}$$

In the above, a Kronecker delta function with 4 indices refers to a delta function symmetrized with respect to, say, both the upper indices.

Also, instead of the standard notation for the functional derivative $\delta\Delta f_a(x)/\delta\xi^\alpha(y)$, we shall write the more compact expression

$$\Delta_{a\alpha} \equiv \frac{\partial\Delta_a}{\partial\xi^\alpha} \equiv \mathcal{S}_{a\alpha} + \mathcal{T}_a{}^b{}_\alpha f_b. \tag{5.14}$$

In this notation, the transition between metric perturbation fields in real space, and the cosmological perturbations in Fourier space that we introduce in Appendix 5.A amounts to a matrix multiplication. Denoting by $f_{\tilde{a}}$ the fields in Fourier space, and by f_a those in real space, we have

$$f_{\tilde{a}} = Q_{\tilde{a}}{}^a f_a, \quad f_a = Q_a{}^{\tilde{a}} f_{\tilde{a}}, \tag{5.15}$$

with $Q_a{}^{\tilde{a}}\,Q_{\tilde{a}}{}^b = \delta_a{}^b$ and $Q_{\tilde{a}}{}^a\,Q_a{}^{\tilde{b}} = \delta_{\tilde{a}}{}^{\tilde{b}}$. Along the same lines, we can parameterize diffeomorphism transformations ξ^α in terms of its irreducible components $\xi^{\tilde{\alpha}}$, with

$$\xi^{\tilde{\alpha}} = Q^{\tilde{\alpha}}{}_\alpha\xi^\alpha, \quad \xi^\alpha = Q^\alpha{}_{\tilde{\alpha}}\xi^{\tilde{\alpha}}, \tag{5.16}$$

where the components of the transformations Q are also listed in Appendix 5.A. In this language, under diffeomorphisms the fields $f^{\tilde{i}}$ transform according to $f_{\tilde{a}} \to f_{\tilde{a}} + \Delta_{\tilde{a}\tilde{\alpha}}\xi^{\tilde{\alpha}}$, where

$$\Delta_{\tilde{a}\tilde{\alpha}} = Q_{\tilde{a}}{}^{a}\, \Delta_{a,\alpha}\, Q^{\alpha}{}_{\tilde{\alpha}}. \tag{5.17}$$

An advantage of this formalism is that it is covariant in field space. As long as the transformations between fields are linear, all our equations retain the same form, provided that the field tensors $\mathcal{S}$ and $\mathcal{T}$ are transformed appropriately. We list the components of $\mathcal{S}$ and $\mathcal{T}$ in the basis of the irreducible components in Appendix 5.B.

5.2.3 Expectation Values

Primordial perturbations are characterized by the moments of the different metric perturbations at sufficiently early times. These moments are identified with equal time vacuum expectation values of the corresponding product of fields in the quantum theory,

$$\langle \Pi_i f_{a_i}(\eta, \vec{x}_i) \rangle \equiv \langle 0_{\rm in} | \Pi_i f_{a_i}(\eta, \vec{x}_i) | 0_{\rm in} \rangle. \tag{5.18}$$

Therefore, to make predictions about the primordial perturbations we need to quantize the theory and find a way to calculate expectation values of quantum fields. As far as the quantization is concerned, Fadeev and Popov have argued that the canonical quantization of gravity is equivalent to the covariant path-integral formulation, as long as one includes appropriate gauge-fixing and ghost terms, and as long as one appropriately modifies the functional measure in the path integral [175]. The actual form of the path integral measure, however, has been the subject of some controversy and does not appear to be settled [176]. The author of the last reference, for instance, argues that the correct measure is

$$Dg \equiv \prod_{x\in\mathcal{M}} g^{00}\cdot g^{-1}\, dg_{\mu\nu}(x), \quad D\phi \equiv \prod_{x\in\mathcal{M}} (g^{00})^{1/2}\cdot g^{1/4}\, d\phi(x), \tag{5.19}$$

and also suggests that, in spite of their appearance, both measures are invariant under diffeomorphisms. We adopt the path integral formulation of quantum gravity here because it is better suited to handle local symmetries such as diffeomorphism invariance. The actual form of the measure is not important to us, as long as it is invariant under diffeomorphisms. This is a requirement for the self-consistency of the theory, analogous to the demand that gauge theories be anomaly-free.

In order to calculate expectation values of fields at conformal time η in the path integral approach, we need to either double the number of fields [177], or introduce a time contour $\mathcal{C}$ extending from the asymptotic past to η and back to the asymptotic past, $\mathcal{C} \equiv (-\infty, \eta] \cup [\eta, -\infty)$ [178]. This last formulation is more convenient because it is formally analogous to that of the in-out formalism, and because it allows us to work with a single set of fields. In particular, the expectation value of a product of fields is simply

$$\langle 0_{\rm in}|\Pi_i f_{a_i}(\eta, \vec{x}_i)|0_{\rm in}\rangle = \int Dh\, D\varphi\, D\omega\ [\Pi_i f_{a_i}(\eta, \vec{x}_i)] \exp\left(iS_{\rm tot}[h_{\mu\nu}, \varphi, \omega]\right), \qquad (5.20)$$

where the functional integral runs over field configurations on the extended time contour $\mathcal{C}$, and we have introduced the ghost fields ω. The values of the fields at the endpoints of this contour, $\partial\mathcal{C} = \{-\infty, -\infty\}$, determine the state whose expectation value we are calculating. In the in-in formalism, the field configurations at both endpoints of the contour are identical, and hence, the boundary terms cancel and do not contribute to the change of the action under diffs. Therefore, any identity that follows from diffeomorphism invariance alone will apply to expectation values in arbitrary states. By shifting this contour by a small imaginary contribution, we can project onto the in-vacuum of the theory $|0_{\rm in}\rangle$.

Naively, one may think that it is irrelevant whether we integrate over all metric and field configurations $g_{\mu\nu}$ and ϕ, or just over its fluctuations $h_{\mu\nu}$ and φ, since they just differ by the given background values. But given the (somewhat uncertain) non-linear structure of the measure in equation (5.19), such a shift may introduce

fluctuation-dependent terms in the measure. Nevertheless, this has no impact on our analysis, as long as the measure for the new fields $h_{\mu\nu}$ and φ remains diffeomorphism invariant, which, as we argued above, is a condition for the self-consistency of the theory.

The classical action for the perturbations is simply

$$S_{\rm inv}[h_{\mu\nu},\varphi] \equiv S[\bar{g}_{\mu\nu}+h_{\mu\nu},\bar{\phi}+\varphi], \tag{5.21}$$

where the functional S on the right-hand-side is diffeomorphism invariant, that is, satisfies equation (5.3). Again, $S_{\rm inv}$ is invariant under the transformations of the perturbations (5.6) because in the in-in formalism there is no contribution from the boundary terms.

The change of variables (5.7) casts the path integral in terms of the cosmological perturbations fields f. Because the transformation (5.7) is linear in the perturbations, the functional Jacobian is field independent and has no impact on cosmological correlators. In particular, we can go back and forth between the representation of the fluctuations in terms of the fields $h_{\mu\nu}$ and φ in real space, and the perturbations f in Fourier space. On the other hand, a non-linear change of variables, $h_{\mu\nu}(x) = F(\zeta_{\mu\nu}(x))$ would force us to introduce a field-dependent Jacobian in the path integral measure, which would amount to the additional term in the action

$$S_J = -i\Omega^{-1}\int d^4x \log F'(\zeta_{\mu\nu}(x)), \tag{5.22}$$

with a divergent constant $\Omega^{-1} = \delta^{(4)}(0)$. This would affect cosmological correlators, although only beyond tree level.

5.2.4 Gauge Fixing

To render the functional integral (5.20) well-defined, we have to introduce gauge-fixing and ghost terms into the action functional,

$$S_{\rm tot} = S_{\rm inv} + S_{\rm gf} + S_{\rm gh}. \tag{5.23}$$

The actual form of the gauge-fixing terms is not particularly important, as long as they are *not* invariant under the set of gauge symmetries under consideration. In DeWitt notation, if the gauge-fixing terms are taken to be of the form

$$\exp(iS_{\rm gf}) = B[F_\beta(f_a)], \tag{5.24}$$

where B is an arbitrary functional of its arguments, and the F_β are a set of arbitrary local functions of the field perturbations f_a (one function F_β for each local symmetry), the only condition is that the matrix $F_\beta{}^a\Delta_{a\alpha}$ be invertible, which amounts to the functional F_β not being invariant under any combination of infinitesimal diffeomorphisms. We focus here on tree-level calculations, so we shall ignore the ghost fields, although they could be easily incorporated into our analysis.

Component Approach The conventional approach to gauge-fixing in cosmological perturbation theory is to impose conditions that enforce the vanishing of a subset of the fields f_g, $g \in G$

$$\exp(iS_{\rm gf}) = \prod_{g\in G} \delta(f_g). \tag{5.25}$$

Because we are using DeWitt notation, the index g here runs over the fields that have been set to zero, and all their spacetime arguments. Since diffeomorphisms are parameterized by four independent functions ξ^α, we need to specify four independent gauge fixing conditions, and we need to make sure that these conditions are not preserved by any infinitesimal diffeomorphism. Say, in longitudinal gauge we may

choose

$$G_{\text{long}} = \{B, H_T, H_+, H_-\}. \tag{5.26}$$

Equations (5.90) then suffice to check that the condition $f_g = 0\,(g \in G_{\text{long}})$ is not preserved by infinitesimal diffeomorphisms.

The change in the total action (5.23) under diffeomorphisms plays a crucial role in the identities we derive below. By assumption, the classical action S_{inv} is invariant and thus does not contribute to the total change. There is also a simple way to calculate the change of the gauge fixing terms S_{gf}, given a set of gauge fixing conditions $f_g = 0$, $g \in G$. If in the absence of gauge-fixing conditions the action is gauge invariant, by definition it must be that

$$0 = \Delta S_{\text{inv},\alpha} = S_{\text{inv}}{}^{a}\,\Delta_{a\alpha}. \tag{5.27}$$

We now split the sum over the fields f_g subject to the gauge condition $f_g = 0$, and those fields f_u which remain unconstrained, and impose the gauge-fixing conditions $f_g = 0$ on the resulting equation, $S_{\text{inv}}{}^{u}\,\Delta_{u\alpha}|_{f_g=0} = -S_{\text{inv}}{}^{g}\,\Delta_{g\alpha}|_{f_g=0}$. But this equation just states that the gauge-fixed action $S_{\text{inv}}|_{g=0} = S_{\text{inv}} + S_{\text{gf}}$ changes by

$$(S_{\text{inv}} + S_{\text{gf}})_{,\alpha} = -S^{g}_{\text{inv}}\,\Delta_{g\alpha}|_{g=0}, \tag{5.28}$$

which completes our determination of the variation of the total action under diffeomorphisms. As an example consider the gauge fixing condition $\varphi = 0$. Combination of equations (5.6) and (5.28) implies that this condition breaks time, but preserves spatial diffeomorphisms.

Gauge-Fixing Terms The drawback of demanding that individual components of the field perturbations vanish is that the variation of the action under broken diffeomorphisms in equation (5.28) not only depends on the particular fields gauge-fixed to zero, but also on the actual invariant action S_{inv} of the theory. In that case, it

does not appear to be possible to derive relations that only follow from diffeomorphism invariance, no matter what the specific action of the theory is.

There is however a physically equivalent way to impose a gauge-fixing condition $f_g = 0$ while preserving almost all of the symmetry of the action. Suppose that we add to our action the gauge-fixing term

$$S_{\text{gf}} = -\frac{M^2}{2} \sum_{g \in G} f_g^2, \tag{5.29}$$

where M is a constant that will be taken to infinity at the end of the calculation (the reader may think of this as a mass term for the field f_g.) This is to some extent analogous to the R_ξ gauges employed in the quantization of non-abelian gauge symmetries. It amounts to a choice of a Gaussian B in equation (5.24) and a set of linear functions $F_\beta(f_a) \equiv \delta_{a\beta}$, where β runs over the fields in G.

For sufficiently large M, the free propagator for the gauge-fixed fields f_g and the remaining "unconstrained" fields f_u becomes

$$\langle f_{g_1} f_{g_2} \rangle_{\mathcal{C}} = -\frac{1}{M^2} \delta_{g_1 g_2} + \mathcal{O}(M^{-4}), \quad \langle f_g f_u \rangle_{\mathcal{C}} = \mathcal{O}(M^{-2}). \tag{5.30}$$

Hence, in the limit $M \to \infty$, the fields g decouple. The theory still has cubic and higher vertices containing the fields f_g, but their contributions to any diagram with no external heavy fields vanish because the internal line propagators approach zero as M tends to infinity. Effectively, the theory is the same as if we had gauge-fixed $f_g \equiv 0$. If we keep M finite, (5.29) remains a valid gauge-fixing term for appropriate choices of the fields f_g, but in this case, the massive fields f_g are not decoupled from the theory.

Reduced Action Gauge symmetries enhance the invariance group of a theory, at the expense of introducing redundant degrees of freedom. In some cases, it is convenient to trade back these gauge symmetries for a description of the theory

that involves a smaller number of fields. In many diffeomorphism invariant theories, such as general relativity coupled to a canonical scalar, the metric components h_{00} and h_{0i} are auxiliary fields, and the matrix $S^{,ab}$ is non-singular for $a, b \in \{h_{00}, h_{0i}\}$. Therefore, if our gauge-fixed theory (5.23) belongs to the last class, we can integrate these variables out perturbatively,

$$\exp(iS_R[\varphi, h_{ij}]) \equiv \int Dh_{00}\, Dh_{0i} \exp(iS_{\rm tot}). \tag{5.31}$$

The resulting reduced action S_R does not depend on the redundant variables h_{00} and h_{0i} any longer, and, as a result, it appears to have lost some of the original symmetries of $S_{\rm inv}$. In fact, equation (5.6) for the spatial components of the metric

$$\Delta h_{ij} = g_{i\alpha}\partial_j\xi^\alpha + g_{\alpha j}\partial_i\xi^\alpha + \xi^\alpha\partial_\alpha g_{ij}, \tag{5.32}$$

implies that S_R is invariant under spatial diffeomorphisms ($\alpha \neq 0$), but not under the original time diffeomorphisms ($\alpha = 0$), basically because only under the former does Δh_{ij} depend on the unconstrained variables h_{ij} alone. The only exception consists of those diffeomorphisms that amount to a spatially global time shift, $\xi^0 = \xi^0(\eta)$, because in that case Δh_{ij} does not involve the variables being integrated out. But in any case, this apparent loss of invariance under diffeomorphisms is not fatal, among other things, because time diffeomorphisms have to be gauge-fixed (and hence broken) anyway. As it will become apparent below, at tree level it does not matter whether a symmetry has been broken by gauge fixing terms or otherwise.

So far, we have simply integrated out the four auxiliary fields, but we have not fixed the gauge yet. In this context, one usually fixes time-diffeomorphisms by imposing "unitary gauge" $\varphi = 0$. In unitary gauge the action is still invariant under spatial diffeomorphisms. We shall discuss different ways to gauge-fix the latter below.

5.3 Schwinger-Dyson Equations for Connected Correlators

We are now ready to investigate the constraints diffeomorphism invariance places on the correlators of cosmological perturbations. For the sake of generality, we shall derive these identities in an arbitrary gauge. Those terms in the identities that involve variation with a field that has been gauged to zero should then be ignored.

Let us define the generator of connected correlators $W(J^a)$ by

$$\exp(iW) = \int Df \exp\left[iS_{\rm tot} + iJ^a\, f_a\right], \tag{5.33}$$

where, again, fields are defined along the time contour $\mathcal{C}$ appropriate for the in-in formalism. Taking functional derivatives of iW with respect to the currents iJ^a and setting the latter to zero thus allows us to calculate contour-ordered correlators of arbitrary products of fields. Changing variables $f_a \to f_a + \Delta_a$ in equation (5.33), and assuming that the measure is invariant under such an infinitesimal transformation, results in the master identity

$$J^a\left(\mathcal{S}_{a\alpha} + \mathcal{T}_a{}^b{}_\alpha \frac{\delta W}{\delta J^b}\right) = -W_{\Delta S,\alpha}, \tag{5.34}$$

where we have introduced the generator of connected diagrams with an insertion of the change in the action under diffeomorphisms,

$$W_{\Delta S,\alpha} = \frac{\int Df\, (\delta\Delta S_{\rm tot}/\delta\xi^\alpha) \exp\left[iS_{\rm tot} + iJ^a\, f_a\right]}{\int Df\, \exp\left[iS_{\rm tot} + iJ^a\, f_a\right]}, \tag{5.35}$$

If the action $S_{\rm tot}$ is invariant under diffeomorphisms, $(\Delta S_{\rm tot})_{,\alpha} = 0$, the generator $W_{\Delta S,\alpha}$ vanishes, but typically, the action contains gauge-fixing terms that break the symmetry, leading to a non-zero $W_{\Delta S,\alpha}$. In fact, the master identity above is valid no matter what the total action $S_{\rm tot}$ is. Any eventual change of the action under

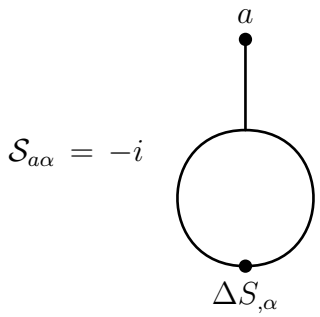


Figure 5.1: Diagrammatic representation of equation (5.38). Dots denote vertices, and circles the sum of all connected diagrams with the corresponding number of insertions. In particular, a circle stands for iW, and a circle with a vertex insertion of $\Delta S_{,\alpha}$ stands for $W_{\Delta S,\alpha}$. Each additional field a connected to a circle then amounts to a functional derivative of the corresponding generator with respect to iJ^a.

diffeomorphisms (or any transformation of the form (5.14)) is then captured by $W_{\Delta S,\alpha}$.

By taking functional derivatives of $W_{\Delta S,\alpha}$ with respect to the currents iJ^a we obtain the sum of all connected diagrams with the corresponding number of fields f_a and a single insertion of $(\Delta S_{\rm tot})_{,\alpha}$. In standard notation, letting $s(x)$ and $t^{\mu\nu}(x)$ denote the currents conjugate to φ and $h_{\mu\nu}$, the master identity for the real space fields reads

$$s\frac{\partial}{\partial x^\alpha}\left(\bar{\phi}+\frac{\delta W}{\delta s(x)}\right)+t^{\mu\nu}\frac{\partial}{\partial x^\alpha}\left(\bar{g}_{\mu\nu}+\frac{\delta W}{\delta t^{\mu\nu}(x)}\right)-2\frac{\partial}{\partial x^\mu}\left[t^{\mu\nu}\left(\bar{g}_{\alpha\nu}+\frac{\delta W}{\delta t^{\alpha\nu}}\right)\right]=-W_{\Delta S,\alpha}(x). \tag{5.36}$$

Note that if we contract equation (5.34) with the generator of a background isometry $\bar{\xi}^\alpha$, the inhomogeneous term $\mathcal{S}_{a\alpha}\,\bar{\xi}^\alpha$ drops out the equation. Typically, the gauge-fixing terms are chosen to respect the background isometries, $(\Delta S_{\rm tot})_{,\alpha}\,\bar{\xi}^\alpha = 0$, so functional derivatives of the contracted equation express then the invariance of the correlation functions under such global transformations.

Equation (5.36) (or (5.34)) captures the constraints imposed by diffeomorphism invariance on the connected diagrams of the theory and is one of the main results of this section. By taking functional derivatives of (5.36) with respect to the currents $it^{\mu\nu}$ and is, we derive relations between the correlators of cosmological perturbations. If the gauge-fixing terms set some of the field perturbations φ or $h_{\mu\nu}$ to zero, the

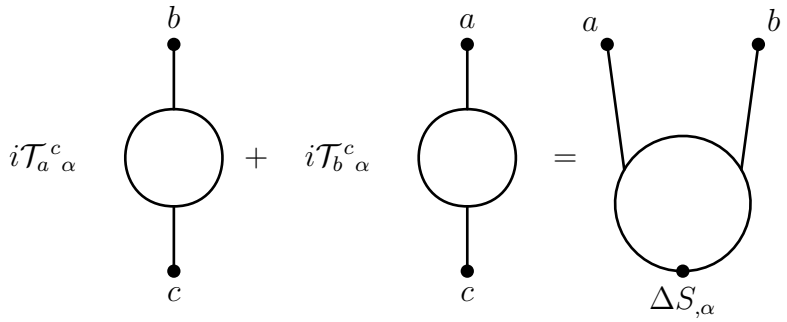


Figure 5.2: Diagrammatic illustration of equation (5.39). Same conventions as in Figure 5.1 apply. Note that there is an implied sum and integral over the repeated dummy indices. The reader should be mindful of the different factors of i that appear in the relation between connected diagrams and functional derivatives of W; for instance, the propagator is $-iW_{,ab}$.

generating functional does not depend on the associated conjugate currents, and the corresponding functional derivatives vanish. Similarly, if we are working with the reduced action (5.31), the functional derivatives with respect to t^{00} and $t^{0i} = t^{i0}$ can be set to zero, since these currents, and their conjugate fields, are not part of the theory.

For instance, the simplest relation follows by just evaluating equation (5.34) at zero currents,

$$W_{\Delta S,\alpha} = 0. \tag{5.37}$$

This just states that the sum of all vacuum diagrams with an insertion of the vertex $\Delta S_{,\alpha}$ vanishes. In some cases this would follow from translational invariance, although we have not made that assumption here. Taking one functional derivative of equation (5.34) with respect to J^a and setting the currents to zero then yields

$$\mathcal{S}_{a\alpha} = -(W_{\Delta S,\alpha})_{,a}, \tag{5.38}$$

where we have assumed that the fields have zero expectation, $W_{,a} = 0$ and we denote $\delta F/\delta J^a$ by $F_{,a}$ (again, where confusion is unlikely, we shall simply write F_a). The previous equation thus relates the sum of all connected diagrams with an insertion of $\Delta S_{,\alpha}$ and a single field f_a to the inhomogeneous component of the change of f_a

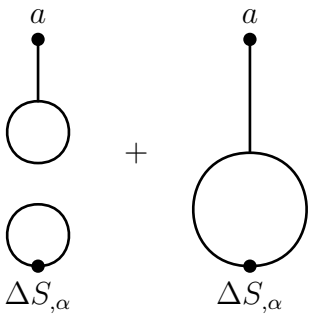


Figure 5.3: The sum of all diagrams with insertions of f_a and $\Delta S_{,\alpha}$, $\langle f_a \Delta S_{,\alpha} \rangle$, expressed in terms of sums of products of connected diagrams. We do not include vacuum-to-vacuum diagrams, which are common factors to all of these, equal to one if the quantum state is normalized. Note that vacuum diagrams with a single field insertion vanish by assumption, $\langle f_a \rangle = 0$, and that the vacuum diagrams with an insertion of $\Delta S_{,\alpha}$ vanish by equation (5.37).

under a diff transformation. We represent such a relation diagrammatically in Figure 5.1. Similarly, taking two functional derivatives of equation (5.34) with respect to the currents yields the identity

$$\mathcal{T}_a{}^c{}_\alpha W_{cb} + \mathcal{T}_b{}^c{}_\alpha W_{ca} = -(W_{\Delta S,\alpha})_{,ab}, \tag{5.39}$$

which relates the propagators of the theory to the sum of all connected diagrams with an insertion of $\Delta S_{,\alpha}$ and two fields, and is represented diagrammatically in Figure 5.2.

As we shall see, equations (5.38) and (5.39) are closely related to a family of relations known as Slavnov-Taylor or Ward-Takahashi identities. To further illustrate their meaning, let us here elaborate on their connection with the Schwinger-Dyson equations. The latter reflect the fundamental theorem of calculus, namely, that the functional integral of a functional derivative vanishes. In particular, for any functional F of the fields f_a we have, in DeWitt notation,

$$\langle F^c \rangle + i \langle F S^c \rangle = 0, \tag{5.40}$$

where S is the total action of the theory, and $\langle \cdots \rangle$ denotes the sum of *all* diagrams

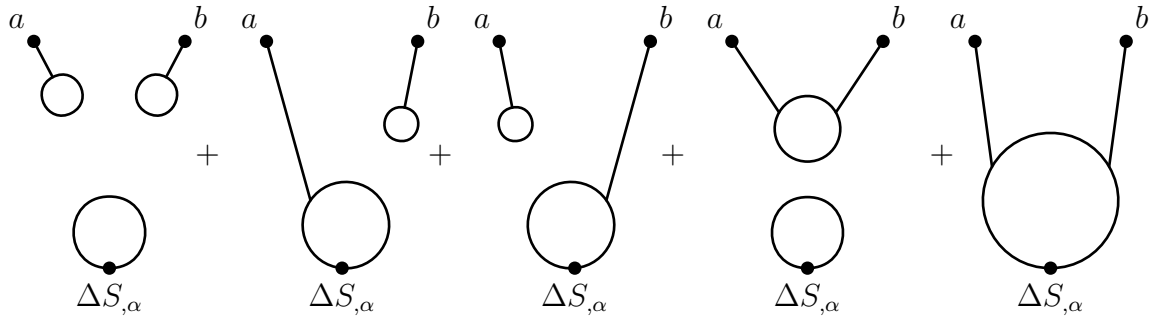


Figure 5.4: The sum of all diagrams with insertions of f_a, f_b and $\Delta S_{,\alpha}$, $\langle f_a f_b \Delta S_{,\alpha} \rangle$, expressed in terms of sums of products of connected diagrams. Same comments as those in the caption to Figure 5.3 apply. Hence, all but the last diagram on the right vanish.

(connected and disconnected) with the corresponding number of insertions. Equation (5.40) is also the statement that the equations of motion $S^c = 0$ of the classical theory hold in the quantum theory, modulo contact terms, for which the functional derivative F^c is non-vanishing. Now, setting $F \equiv f_a \, \Delta_{c\alpha}$ in equation (5.40), and summing over c results in the identity

$$\langle \Delta_{a\alpha} + f_a \Delta_{c\alpha}{}^{,c} \rangle + i \langle f_a \Delta S_{,\alpha} \rangle = 0. \tag{5.41}$$

Because $\Delta_{a\alpha}$ is linear in the fields, taking Figure 5.3 into account, and bearing in mind that $\langle f_a \rangle = 0$, this is nothing but equation (5.38). To arrive at this conclusion we also need to assume that $\Delta_{c\alpha}{}^{,c} = \mathcal{T}_c{}^c{}_\alpha = 0$. This is again the statement that the integral of a derivative vanishes. As mentioned by DeWitt in [157] it is also a condition for the internal consistency of the theory. Similarly, setting $F = f_a f_b \Delta_{c\alpha}$ in equation (5.40) and summing over c we find,

$$\langle f_b \Delta_{a\alpha} \rangle + \langle f_a \Delta_{b\alpha} \rangle + \langle f_a f_b \Delta_{c\alpha}{}^{,c} \rangle + i \langle f_a f_b \Delta S_{,\alpha} \rangle = 0. \tag{5.42}$$

Taking Figure 5.4 into account, and recalling equation (5.37), this becomes equation (5.39). We can therefore think of equations (5.38) and (5.39) as consequences of the equations of motion in the quantum theory.

5.4 Slavnov-Taylor Identities for the Effective Action

In many cases, it is more convenient to restrict the properties of the one-particle irreducible diagrams of the theory, which are generated by the effective action. The quantum effective action Γ is the Legendre transformation of the generator of connected diagrams W,

$$\Gamma(\bar{f}_a) = W(J^a_*) - \bar{f}_a J^a_*, \tag{5.43}$$

where the currents J^a_* are defined by the condition

$$\left.\frac{\delta W}{\delta J^a}\right|_{J=J_*} = \bar{f}_a, \tag{5.44}$$

and $\bar{f}_a$ is the prescribed expectation value of the field f_a (hence the bar, which we shall later drop for simplicity.) The only difference here with respect to the in-out formalism is that, once more, time integrals run over the contour $\mathcal{C}$ we introduced in Section 5.2.3.

The generating functional W does not depend on the currents conjugate to those fields that the gauge-fixing terms constrain to vanish, so the effective action does not depend on the corresponding field expectations. Therefore, Γ is a functional of the prescribed expectations of the unconstrained perturbations alone. Functional derivatives of $i\Gamma$ with respect to these fields give the sum of all one-particle-irreducible diagrams with the corresponding number of external fields. These one-particle-irreducible diagrams are then the building blocks from which one can calculate connected correlators, by summing over tree diagrams whose vertices are determined by the corresponding functional derivatives of the effective action.

If the action of a theory with fields f_a changes by $\Delta S_{\mathrm{tot},\alpha}$ under an infinitesimal transformation $f_a \to f_a + \Delta_a$, where Δ_a is *linear* in the fields like in equation (5.11),

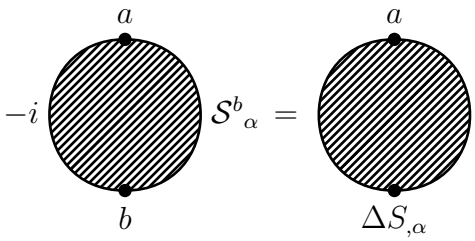


Figure 5.5: Diagrammatic representation of equation (5.48). Here, shaded circles denote sum of all one-particle-irreducible diagrams with the corresponding number of insertions. In particular, a shaded circle stands for $i\Gamma$, whereas a circle with a vertex labeled by $\Delta S_{,\alpha}$ stands for $\Gamma_{\Delta S,\alpha}$. Each additional field vertex a denotes then a functional derivative of the corresponding quantum action with respect to $\bar{f}_a$.

one can show (see e.g. [179]) that

$$\frac{\delta\Gamma}{\delta\bar{f}_a}(\mathcal{S}_{a\alpha} + \mathcal{T}_a{}^b{}_\alpha \bar{f}_b) = \Gamma_{\Delta S,\alpha}, \tag{5.45}$$

where $\Gamma_{\Delta S,\alpha}$ (note the missing factor of i) is the sum of all one-particle irreducible diagrams with an insertion of $\Delta S_{\mathrm{tot},\alpha}$. In particular, if the action is invariant, $\Delta S = 0$, the previous equation states that linear symmetries are also symmetries of the effective action. If we contract equation (5.45) with an isometry unbroken by the gauge-fixing terms, the equation just expresses again the invariance of the effective action with respect those transformations, as before. Equations relating the change of the effective action under a set of local transformations are generally known as Slavnov-Taylor identities, although they are often referred to as Ward-Takahashi identities too. Adapting equation (5.45) to the standard notation, and dropping the bar from the arguments of the effective action we obtain in real space

$$\frac{\delta\Gamma}{\delta h_{\mu\nu}(x)}\frac{\partial g_{\mu\nu}}{\partial x^\alpha} + \frac{\delta\Gamma}{\delta\varphi(x)}\frac{\partial\phi}{\partial x^\alpha} - 2\frac{\partial}{\partial x^\mu}\left(\frac{\delta\Gamma}{\delta h_{\mu\nu}(x)}g_{\alpha\nu}\right) = \Gamma_{\Delta S,\alpha}(x), \tag{5.46}$$

where $g_{\mu\nu}$ and ϕ are the fields defined in equations (5.4). This is our master identity for the effective action, which holds for $\alpha = 0$ (time diffeomorphisms) and $\alpha = i$ (spatial diffeomorphisms). Again, if a certain set of fields are constrained to vanish by the gauge-fixing conditions, or they have been integrated out from the action, the

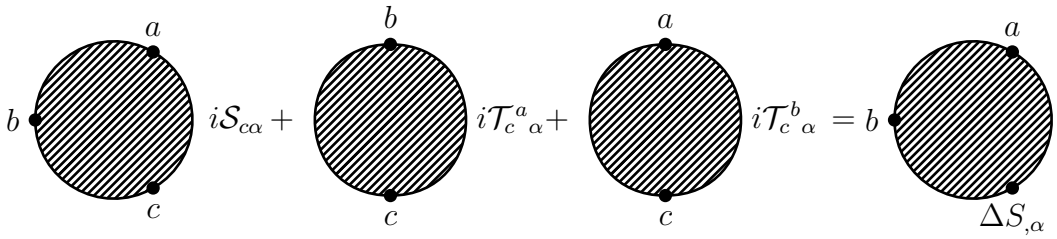


Figure 5.6: Diagrammatic representation of equation (5.49). Same conventions as in Figure 5.5 apply.

corresponding functional derivatives vanish. It is worth stressing that this master equation is valid at all orders in perturbation theory and for any gauge-fixing conditions. For spatial diffeomorphisms, and in the reduced formulation of the theory, essentially the same identity is derived in reference [158].

In some non-linear parameterizations of the metric perturbations, such as the one employed for instance in [23], diffeomorphisms act non-linearly on the cosmological perturbations. In this case, equation (5.45) still holds at all orders in perturbation theory, provided that we truncate the action of diffeomorphisms on the fields of the theory to its linear components. In that case, $\Gamma_{\Delta S}$ on the right-hand-side will include the change in the action ΔS under these truncated linear diffeomorphisms, a change that would otherwise receive contributions from the gauge fixing terms alone. At tree level this is a trivial consequence of the identity $\Gamma = S_{\rm tot}$ and invariance of the classical action under diffeomorphisms, $S^{,a}_{\rm inv}\,\Delta_{a\alpha} = 0$. In particular, at tree level the Slavnov-Taylor equation

$$\frac{\delta\Gamma}{\delta\bar{f}_a}\,\Delta_{a\alpha} = \Gamma_{\Delta S,\alpha} \tag{5.47}$$

holds even if $\Delta_{a\alpha}$ is a non-linear functional of the fields.

5.4.1 Derivation of the Identities

Taking functional derivatives of the effective action, and evaluating the latter at zero fields yields relations between the 1PI diagrams of the theory. For instance, taking

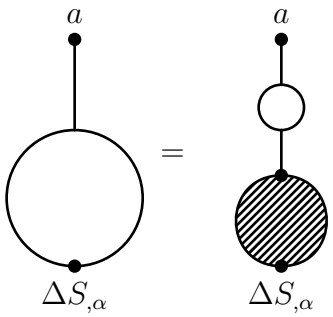


Figure 5.7: The sum of all connected diagrams with an insertion of $\Delta S_{,\alpha}$ and a field f_a. The shaded blob indicates the sum of all 1PI diagrams with the given vertices, $\Gamma_{\Delta S,\alpha}$.

one functional derivative of equation (5.45) and using that $\Gamma^a = 0$ results in

$$\Gamma^{ba}\,\mathcal{S}_{b\alpha} = (\Gamma_{\Delta S,\alpha})^{,a}, \tag{5.48}$$

whereas taking two functional derivatives gives

$$\Gamma^{cba}\mathcal{S}_{c\alpha} + \Gamma^{cb}\mathcal{T}_c{}^a{}_\alpha + \Gamma^{ca}\mathcal{T}_c{}^b{}_\alpha = (\Gamma_{\Delta S,\alpha})^{,ba}. \tag{5.49}$$

These identities are represented diagrammatically in Figures 5.5 and 5.6.

The reader may wonder whether the identities that involve Γ bear any relation to those obeyed by the generators of connected diagrams, W. In fact, it is quite easy to see that both sets of identities are essentially the same. Compare for example equations (5.38) and (5.48), and their corresponding diagrammatic representations in Figures 5.1 and 5.5. Because the sum of all connected diagrams with an insertion of $\Delta S_{,\alpha}$ and an external line is given by the diagrams in Figure 5.7, equation (5.38) just states that

$$\mathcal{S}_{a\alpha} = -i\,(\Gamma_{\Delta S,\alpha})^{,b}\,(-iW_{ba}). \tag{5.50}$$

Using the fact that the propagator $-iW_{ba}$ is just minus the inverse of the self-energy $i\Gamma^{ba}$ equation (5.48) immediately follows. Similarly, equation (5.39) and Figure 5.8

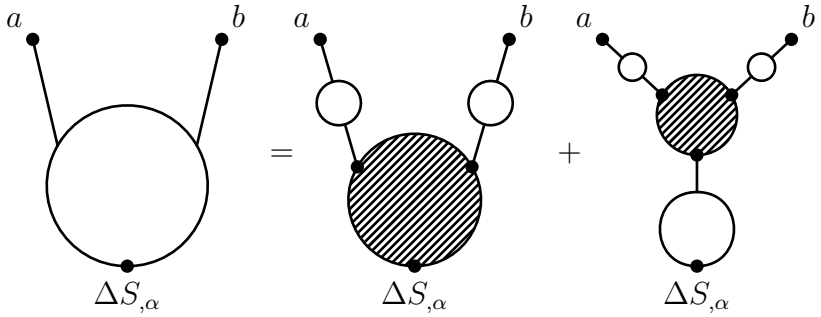


Figure 5.8: The sum of all connected diagram with an insertion of ΔS_α and two fields f_a and f_b. The shaded blobs indicates the sum of all 1PI diagrams with an insertion of $\Delta S_{,\alpha}$ and the number of fields indicated by the thick dots.

imply that

$$\mathcal{T}_a{}^c{}_\alpha W_{cb} + \mathcal{T}_b{}^c{}_\alpha W_{ca} = \Gamma^{cd}_{\Delta S,\alpha}(-iW_{ca})(-iW_{db}) + i\Gamma^{ecd}(-iW_{ca})(-iW_{db})\frac{1}{i}(W_{\Delta S,\alpha})_{,e}. \tag{5.51}$$

Contracting left and right of this equation with two factors of the self-energy, and using equation (5.38) yields equation (5.49).

Equation (5.39) relates cubic to quadratic terms in the effective action, and thus provides constraints on the possible form of the cubic terms. These equations are analogous to the identities that relate the vertex for graviton emission by matter to the matter propagator, and ultimately enforce the equivalence principle in general relativity [156, 157]. If any of the fields appearing in these equations has been gauge-fixed to vanish, the corresponding term in the equation should be set to zero. The identities also hold in the reduced theory defined by equation (5.31), provided that functional derivatives of Γ with respect to $\bar{A}, \bar{B}$ and $\bar{B}_\pm$ are also set to zero.

5.4.2 Illustration

As an application of these results, consider equation (5.50) in a case in which the gauge fixing term is of the form (5.29). Then, at tree level, the effective action with

an insertion of $\Delta S_{,\alpha}$ satisfies

$$(\Gamma_{\Delta S,\alpha})^{,a} = -M^2 \sum_g \delta_g{}^a \mathcal{S}_{g\alpha}. \tag{5.52}$$

Therefore, inserting the last equation into (5.50) we arrive at

$$\mathcal{S}_{a\alpha} = M^2 \sum_g \mathcal{S}_{g\alpha} W_{ga}. \tag{5.53}$$

Suppose now the field a is invariant under a particular linear diffeomorphism α, $\mathcal{S}_{a\alpha} = 0$, This is for instance the case for the gauge-invariant scalar perturbations introduced by Bardeen [173]. Then, if a particular gauge-fixed field g does transform under the same diffeomorphism, $\mathcal{S}_{g\alpha} \neq 0$, and also happens to be the only one appearing in the sum on the right hand side of (5.53), it follows that $W_{ga} = 0$. Therefore, at tree level, there is no correlation between gauge-invariant and gauge-fixed perturbations. Note that we expect the fields g to change under diffeomorphisms, because otherwise equation (5.29) would not be an appropriate gauge-fixing term.

5.5 Consistency Relations from Diffeomorphisms

One of the main motivations for the introduction of this formalism is the study of the extent to which diffeomorphism invariance constrains the properties of the perturbations created during a scalar field driven inflationary stage. These primordial perturbations are conveniently characterized by the equal time expectation values of their products. For two fields we speak of spectra, and for three fields we speak of bispectra; these are the only quantities that are observationally relevant at this point.

Building on the work of references [159, 160], the authors of [158] considered the implications of spatial diffeomorphism invariance by formulating identities that relate the bispectrum of cosmological perturbations to their power spectrum, in the limit of a squeezed triangle. Similar relations had been derived earlier from the requirement

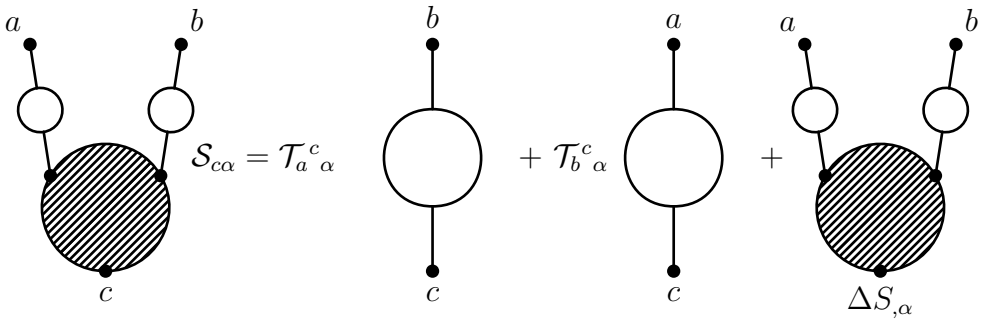


Figure 5.9: Diagrammatic representation of equation (5.55), the basis for the derivation of consistency relations in cosmological perturbation theory.

of conformal invariance [169–171]. Here, we extend the work of [158] to arbitrary diffeomorphisms and arbitrary gauges, not necessarily in the reduced formulation of the theory.

Our goal is to relate the bispectrum of cosmological perturbations to their power spectrum, with a gauge-fixing action of the form (5.29). Equation (5.39) appears to be the perfect starting point for such analysis, since its right hand side already contains almost what we are looking for: The sum of all connected diagrams with insertions of f_a, f_b and $\Delta S_{,\alpha}$, the latter being known for a gauge-fixing term (5.29). Yet, since equation (5.39) contains contact terms that become singular in the limit of equal times, it is more convenient to start with the equivalent identity (5.51), which, when combined with equation (5.38) results in

$$\Gamma^{cde} W_{da} W_{eb}\, \mathcal{S}_{c\alpha} = \mathcal{T}_a{}^c{}_\alpha W_{cb} + \mathcal{T}_b{}^c{}_\alpha W_{ca} + \Gamma^{de}_{\Delta S,\alpha} W_{da} W_{eb}. \tag{5.54}$$

Diagrammatically, this equation can be represented as in Figure 5.9. The left hand side of equation (5.54) (or the equation in Figure 5.9) is almost the sum of all connected diagrams with three external fields, since $W_{abc} = \Gamma^{def} W_{da} W_{eb} W_{fc}$. To simplify the notation, we are going to think of W_{ab} as a metric, which we can use to lower indices in field space. In this case, equation (5.54) simplifies to

$$\Gamma^c{}_{ab}\, \mathcal{S}_{c\alpha} = \mathcal{T}_{ab\alpha} + \mathcal{T}_{ba\alpha} + (\Gamma_{\Delta S,\alpha})_{ab}. \tag{5.55}$$

The reader should thus remember the natural position of the indices to determine whether an index has been lowered with the propagator.

Equation (5.55) is where we need to stop if we are not willing to make additional assumptions. In order to proceed further, we shall work at tree level, where the analysis simplifies considerably, because the effective action Γ is then just the same as the total (gauge-fixed) classical action $S_{\rm tot}$. If we further assume that the gauge-fixing term is of the form (5.29) we know exactly what $\Delta S_{,\alpha}$ is, and because at tree level $(\Gamma_{\Delta S,\alpha})^{cd} = (\Delta S_{,\alpha})^{cd}$ it follows that

$$(\Gamma_{\Delta S,\alpha})_{ab} = -M^2 \sum_{g\in G} (\mathcal{T}_g{}^c{}_\alpha W_{ga} W_{cb} + \mathcal{T}_g{}^c{}_\alpha W_{ca} W_{gb})\,. \tag{5.56}$$

Hence, if both external fields have a vanishing correlation with the massive fields f_g, the last term on the right hand side of equation (5.55) vanishes, regardless of the action of the theory and the particular diffeomorphism α involved. The latter is what happens for instance if the fields f_g are scalars or vectors, and the fields f_a, f_b are tensors. Note that this simplification occurs because of a global symmetry, namely, invariance of the background under translations and rotations. In the meantime, we concentrate on three-point functions that involve two tensor modes, for which the breaking term does not contribute under any circumstance. Later on we shall consider more general cases.

5.5.1 Diffeomorphism Invariance

With the term proportional to $\Gamma_{\Delta S,\alpha}$ gone, we can focus on the irreducible vertex $\Gamma^c{}_{ab}$, which appears in the identity contracted with $\mathcal{S}_{c\alpha}$. As seen from equation (5.90c), invariance under transverse diffeomorphisms ($\alpha = \pm$) constrains the vertices that include $B_\pm$ and $H_\pm$. Because equation (5.90c) contains a time derivative of a delta function, however, equation (5.55) thus affects the time derivative of 1PI diagrams with an external vector $B_\pm$. It is hence not possible to translate such equation into

an equation for connected correlators with a vector $B_\pm$, although such an equation would not be particularly relevant, since in the class of theories we are studying, vectors are redundant fields anyway. In the reduced formulation of the theory, for instance, the action S_R defined in equation (5.31) remains invariant under the two transverse diffeomorphisms. Hence, in order to fix the gauge we can simply impose the condition $H_+ = H_- = 0$, thus eliminating vectors from the theory altogether.

Consider instead longitudinal diffeomorphisms. In this case, the corresponding transformations in equation (5.90b) do not contain time derivatives of any field in the reduced formulation of the theory, in which h_{00} and h_{0i} have been integrated out, and the field B is therefore absent (the term proportional to ${\delta_f}^B$ can be set to zero.) At this point, instead of working with the scalars H_L and H_T, it is convenient to regard the effective action in the scalar sector as a functional of the two fields

$$\Psi \equiv H_L + \frac{H_T}{3}, \quad H_L. \tag{5.57}$$

Note that Ψ is invariant under longitudinal diffeomorphisms in the linearized theory. Hence, by introducing these fields, we are dividing the two-dimensional scalar sector into a direction in field space that changes under linear longitudinal diffeomorphisms (H_L), and one which does not (Ψ). This division is to some extent arbitrary, since we may add any multiple of Ψ to the gauge-variant direction, without changing the transformation properties under longitudinal diffs of the latter. Clearly, in order to fix longitudinal diffeomorphisms we need to give a mass to H_L, which is the field that is not invariant under the symmetry,

$$S_{\rm gf} = -M^2 \int d\eta\, d^3p\, H_L(\eta, \vec{p}) H_L(\eta, -\vec{p}). \tag{5.58}$$

In terms of Ψ and H_L, and in unitary gauge, equation (5.90c) simply reduces to

$$\mathcal{S}_{f(\eta_1,\vec{p}_1)\,\xi^L(\eta_2,\vec{p}_2)} = \frac{p_1}{3}\delta(\eta_1 - \eta_2)\delta(\vec{p}_1 - \vec{p}_2){\delta_f}^{H_L}. \tag{5.59}$$

Equation (5.55) thus constrains 1PI diagrams that contain the scalar field H_L, the gauge-variant direction in the scalar sector.

To obtain an equation that involves the connected correlators, we need to contract equation (5.55) with a scalar propagator. According to the discussion of Section 5.4.2, invariance under longitudinal diffeomorphisms ($\alpha = L$) implies that there is no correlation between the gauge-invariant field Ψ and the gauge-fixed field H_L, $W_{H_L\Psi} = 0$. Hence, the propagator in the scalar sector is diagonal in the fields H_L and Ψ. Because the propagator is diagonal, and equation (5.55) involves a vertex with a field H_L, we just multiply the left and right hand side of equation (5.55) by $W_{\tilde{H}_L H_L}$. Recall that if f_a and f_b are tensors, the term that involves $\Gamma_{\Delta S}$ does not contribute. Hence, setting a and b to be tensors with respective helicities σ_2 and σ_3, and integrating over the undisplayed time and momentum variables we finally obtain

$$\frac{p_1}{3}\frac{W_{H_L H_{\sigma_2} H_{\sigma_3}}(\eta,\vec{p}_1;\eta,\vec{p}_2;\eta,\vec{p}_3)}{\bar{W}_{H_L H_L}(\eta,\vec{p}_1)} = \Bigg\{\Bigg[2p_j^1 Q_{H_{\sigma_3}}{}^{ij}(\vec{p}_3) Q_{ik}{}^{H_{\sigma_2}}(-\vec{p}_2)\hat{p}_1^k + Q_{H_{\sigma_3}}{}^{ij}(\vec{p}_3) Q_{ij}{}^{H_{\sigma_2}}(-\vec{p}_2)\,\vec{p}_2\cdot\hat{p}_1\Bigg]\bar{W}_{H_{\sigma_2}H_{\sigma_2}}(\eta,\vec{p}_2) + 2\leftrightarrow 3\Bigg\}\frac{\delta(\vec{p}_1+\vec{p}_2+\vec{p}_3)}{(2\pi)^{3/2}}, \quad (5.60)$$

where we define the power-spectrum of an arbitrary variable f by

$$\langle f(\eta,\vec{p}) f(\eta,\vec{p}')\rangle \equiv -i\bar{W}_{ff}(\eta,\vec{p})\delta(\vec{p}+\vec{p}'), \quad (5.61)$$

and the components of the projection tensors for tensor perturbations are listed in Appendix 5.A. Equation (5.60) therefore relates the three-point function of cosmological perturbations to their power spectra. It is the consistency relation that follows from the original invariance under longitudinal diffeomorphisms. Note that it is valid for all scalar momenta, and not only in the soft limit $\vec{p}_1 \to 0$. As should be manifest from our derivation, it applies only at tree level and in the reduced formulation of the theory, with a gauge fixing term (5.57) that gives H_L an arbitrary (but finite) mass. Other than that it only relies on the invariance of the theory under spatial

diffeomorphisms and the isometries of a cosmological background. The consistency relation does not explicitly contain M^2, although the power spectra and the three-point function implicitly depend on that quantity. It is also important to realize that the gauge in which this consistency relation holds is not one of the conventional gauge choices used in cosmological perturbation theory. By giving a finite mass term to a scalar variable, we are not eliminating it from the theory. Hence, the scalar sector here consists of two fields (Ψ and H_L), rather than one, as in the standard ζ-gauge, in which $H_T \equiv 0$. If we had simply set H_T to zero, we would have lost the ability to calculate $\Gamma_{\Delta S}$.

Diffeomorphism invariance also constrains the expectation of a product of three scalar perturbations, or the product of two scalars and a tensor. In this case, however, the last term on the right hand side of equation (5.55) contributes a non-vanishing correction proportional to M^2. As a result, the ensuing consistency relation becomes explicitly gauge-dependent. Hence, we shall not write down the corresponding consistency relation here, although it can be easily derived from the previous equations. A consistency relation for time diffeomorphisms can be derived along the same lines.

5.5.2 Analyticity

As we have seen, diffeomorphism invariance alone constrains the cubic vertices of the theory only along gauge-variant directions in field space. As shown in reference [158], however, additional analyticity properties allow us to extend these constraints to the full cubic vertex itself, in the limit in which one of the momenta approaches zero.

Spatial Diffeomorphisms

Unitary Gauge To see how this works, it is going to be useful to consider the sum of all diagrams with insertions of $h_{ij}(\eta_1, \vec{p}_1)$ and two arbitrary fields $f_2(\eta, \vec{p}_2)$ and $f_3(\eta, \vec{p}_3)$, with the propagator of h_{ij} stripped off, and the overall momentum-

conserving delta function omitted,

$$\bar{\Gamma}^{ij}{}_{f_2 f_3}(\eta_1, \vec{p}_1; \eta, \vec{p}_2; \eta) \times \delta^{(3)}(\vec{p}_2 + \vec{p}_3 - \vec{p}_1) \equiv \Gamma^{f_1}{}_{f_2 f_3} Q_{f_1}{}^{ij}(\eta_1, \vec{p}_1). \tag{5.62}$$

We consider an insertion of the metric perturbation h_{ij}, rather than a helicity eigenvector f_1, because the decomposition into irreducible representations obscures the analyticity properties of the vertex. We also define the propagator of the fields with a momentum-conserving delta function stripped off, which in the case of coincident times defines the power spectrum,

$$W_{f_1 f_2}(\eta_1, \vec{p}_1; \eta_2, \vec{p}_2) \equiv \bar{W}_{f_1 f_2}(\eta_1, \vec{p}_1; \eta_2) \times \delta(\vec{p}_1 + \vec{p}_2). \tag{5.63}$$

We work in the reduced formulation of the theory, in a gauge in which $\varphi \equiv 0$ and the breaking term is

$$S_{\rm gf} = -\frac{M^2}{2} \int d\eta \, d^3p \, \left[H_T(\eta, \vec{p}) H_T(\eta, -\vec{p}) + H_+(\eta, \vec{p}) H_+(\eta, -\vec{p}) + H_-(\eta, \vec{p}) H_-(\eta, -\vec{p}) \right]. \tag{5.64}$$

We take the limit $M \to \infty$ at the end of the calculation, which decouples both the scalar H_T and the two vectors $H_\pm$. This implies that we assume the fields f_2 and f_3 to stand for the remaining light fields H_L or $H_{\pm\pm}$, but not for any of the massive fields. Then, from equation (5.55), invariance under spatial diffeomorphisms ξ^j implies that $\bar{\Gamma}^{ij}{}_{f_2 f_3}$ obeys the equation

$$\begin{aligned} 2a_1^2 p_i^1 \bar{\Gamma}^{ik}{}_{f_2 f_3} \delta_{kj} - \Bigg[\Bigg(2p_k^1 Q_{f_2}{}^{ik}(\vec{p}_2) Q_{ij}{}^{f_3}(\vec{p}_2 - \vec{p}_1) + (p_j^2 - p_j^1) Q_{f_2}{}^{ik}(\vec{p}_2) Q_{ik}{}^{f_3}(\vec{p}_2 - \vec{p}_1) \Bigg) \bar{W}_{f_3 f_3}(\eta, \vec{p}_2 - \vec{p}_1) \\ + \Bigg(2p_k^1 Q_{f_3}{}^{ik}(\vec{p}_1 - \vec{p}_2) Q_{ij}{}^{f_2}(-\vec{p}_2) - p_j^2 Q_{f_3}{}^{ik}(\vec{p}_1 - \vec{p}_2) Q_{ik}{}^{f_2}(-\vec{p}_2) \Bigg) \bar{W}_{f_2 f_2}(\eta, -\vec{p}_2) \Bigg] \frac{\delta(\eta - \eta_1)}{(2\pi)^{3/2}} \\ = -i(\bar{\Gamma}_{\Delta S, j})_{f_2 f_3}, \end{aligned} \tag{5.65}$$

where we have defined

$$(\bar{\Gamma}_{\Delta S,j})_{f_2 f_3}(\eta_1,\vec{p}_1;\eta,\vec{p}_2;\eta)\delta^{(3)}(\vec{p}_2+\vec{p}_3-\vec{p}_1) \equiv \left(\frac{\delta\Gamma_{\Delta S}}{\delta\xi^j(\eta_1,\vec{p}_1)}\right)^{\tilde{f}_2\tilde{f}_3} W_{\tilde{f}_2 f_2} W_{\tilde{f}_3 f_3}. \quad (5.66)$$

Note that because in the limit $M \to \infty$ the propagator is diagonal in field space, there is no need to sum over the repeated indices f_2 and f_3 in equation (5.65). On the other hand, because the breaking term is proportional to M^2, one needs to be careful with terms in the propagator that only decay like $1/M^2$ when dealing with $(\bar{\Gamma}_{\Delta S,j})_{f_2 f_3}$.

We can constrain the components of $\bar{\Gamma}^{ij}{}_{f_2 f_3}$ if we assume it to be analytic for momenta $\vec{p}_1$ in the vicinity of zero. As argued in [158], this is a non-trivial assumption even at tree-level because gravitons are massless particles and we are working in the reduced formulation of the theory, in which h_{00} and h_{0i} have been integrated out. Nevertheless, if the assumption holds, we can solve for $\bar{\Gamma}^{ij}{}_{f_2 f_3}$ as power series in the components of $\vec{p}_1$. Say, at zeroth order in $\vec{p}_1$, we find that the equation is satisfied provided that

$$(\bar{\Gamma}_{\Delta S,j})_{f_2 f_3}(\eta_1,\vec{p}_1=0;\eta,\vec{p}_2;\eta) = 0. \quad (5.67)$$

We shall verify this property below. At first order we obtain then the unique solution

$$2a_1^2\bar{\Gamma}^{ik}_{(0)f_2 f_3}\delta_{kj} = -i\frac{\partial(\bar{\Gamma}_{\Delta S,j})_{f_2 f_3}}{\partial p_i^1} + \frac{\delta(\eta-\eta_1)}{(2\pi)^{3/2}}\Bigg[-\delta^i{}_j\bar{W}_{f_2 f_3}(\eta,\vec{p}_2) - p_j^2\frac{\partial\bar{W}_{f_2 f_3}(\eta,\vec{p}_2)}{\partial p_i^2}$$
$$+ 2Q_{f_2}{}^{ik}(\vec{p}_2)Q_{kj}{}^{f_3}(\vec{p}_2)\bar{W}_{f_3 f_3}(\eta,\vec{p}_2) + 2Q_{f_3}{}^{ik}(-\vec{p}_2)Q_{kj}{}^{f_2}(-\vec{p}_2)\bar{W}_{f_2 f_2}(\eta,-\vec{p}_2)\Bigg], \quad (5.68)$$

Although this is not immediately apparent, it is straight-forward to check that for a rotationally-invariant state of the perturbations the right-hand side is always symmetric in ij. For instance, $p_j^2\,\partial\bar{W}_{f_2 f_3}/\partial p_i^2$ is symmetric if $\bar{W}_{f_2 f_3}$ only depends on the magnitude of the vector $\vec{p}_2$.

Along the same lines, one can derive the solution of equation (5.65) at higher

orders in the momentum $\vec{p}_1$. At first order the solution is again unique, but the proliferation of indices makes its manipulation rather cumbersome beyond the hard scalar case, in which $f_2 = H_L$, $f_3 = H_L$. At yet higher orders the solution is not unique because equation (5.65) only constrains the longitudinal component of the vertex.

In order to proceed, we need to determine $(\bar{\Gamma}_{\Delta S,j})_{f_2 f_3}$. With a breaking term of the form (5.64) the variation of the quadratic part of the action under spatial diffeomorphisms $\alpha = \xi^j(\eta_1, \vec{p}_1)$ becomes

$$i(\Gamma_{\Delta S,j})_{f_2 f_3} = \frac{M^2}{(2\pi)^{3/2}} \Big\{ \bar{W}_{f_2 H_T}(\eta, \vec{p}_2; \eta_1) \bar{W}_{f_3 \tilde{f}_3}(\eta, \vec{p}_3; \eta_1) \Big[2p_k^1 Q_{H_T}{}^{ik}(\vec{p}_2) Q_{ij}{}^{\tilde{f}_3}(-\vec{p}_3) - p_j^3 Q_{H_T}{}^{ik}(\vec{p}_2) Q_{ik}{}^{\tilde{f}_3}(-\vec{p}_3) \Big] + 2 \leftrightarrow 3 \Big\} \delta^{(3)}(\vec{p}_2 + \vec{p}_3 - \vec{p}_1). \quad (5.69)$$

It is easy to check that $(\bar{\Gamma}_{\Delta S,\xi^k})_{f_2 f_3}$ vanishes at $\vec{p}_1 = 0$, as required by condition (5.67). In fact, as we also stated above, for two hard tensors f_2 and f_3 the breaking term vanishes at all momenta and can be therefore ignored. On the other hand, this simplification does not generically occur when the two fields f_2 and f_3 are scalars. If, for instance, $f_2 = H_L^{(2)}(\eta, \vec{p}_2)$ and $f_3 = H_L^{(3)}(\eta, \vec{p}_3)$, the first derivative of $(\bar{\Gamma}_{\Delta S,j})_{f_2 f_3}$ equals

$$\frac{\partial (\bar{\Gamma}_{\Delta S,j})_{H_L^{(2)} H_L^{(3)}}}{\partial p_i^1} \propto -M^2 Q_{H_T}{}^{ik}(\vec{p}_2) Q_{kj}{}^{H_L} \big[\bar{W}_{H_L H_T}(\eta, \vec{p}_2; \eta_1) \bar{W}_{H_L H_L}(\eta, -\vec{p}_2; \eta_1) + \bar{W}_{H_L H_T}(\eta, -\vec{p}_2; \eta_1) \bar{W}_{H_L H_L}(\eta, \vec{p}_2; \eta_1) \big], \quad (5.70)$$

where we have used that $Q_{H_T}{}^{ij}(-\vec{p}_2) Q_{ij}{}^{H_L}(\vec{p}_3) \equiv 0$. Note that $\bar{W}_{H_L H_T}$ is proportional to $1/M^2$, so the right hand side remains finite in the limit $M \to \infty$. Since the finite limit of $M^2 \bar{W}_{H_L H_T}$ depends on details of the theory, we cannot hence determine the contribution of this term from symmetry arguments alone. There is however an exception. If we were interested in correlation functions with a soft scalar, we would

f_1	f_2	f_3
H_L	H_L	H_L
H_L	H_L	$H_{\pm\pm}$
H_L	$H_{\pm\pm}$	$H_{\pm\pm}$
$H_{\pm\pm}$	$H_{\pm\pm}$	$H_{\pm\pm}$

Table 5.1: The different combination of fields for which the consistency relation (5.71a) in unitary gauge holds. The field f_1 carries the soft momentum, whereas the two hard fields f_2 and f_3 can be interchanged.

need to contract the vertex with ${Q_{ij}}^{H_L}$, since $\bar{\Gamma}^{H_L}{}_{H_L H_L} \equiv \bar{\Gamma}^{ij}{}_{H_L H_L} {Q_{ij}}^{H_L}$. Because the right-hand side of equation (5.70) is traceless, the contribution of the breaking term would then vanish. Similarly, one can also check that the breaking term does not contribute to $\Gamma^{H_L}{}_{H_L H_{\pm\pm}}$. We summarize the combination of fields for which the symmetry-breaking term can be discarded at zeroth order in $\vec{p}_1$ in table 5.1.

The offshoot of the previous analysis is that at zeroth order in $\vec{p}_1$, and for the combination of fields listed in table 5.1, the mixed vertex $\bar{\Gamma}^{f_1}{}_{f_2 f_3}$ is determined by symmetry alone. We are then just a step away from the consistency relation for spatial diffeomorphisms. Convolving $\bar{\Gamma}^{f_1}{}_{f_2 f_3}$ with the propagator of f_1 we finally arrive at

$$\frac{\bar{W}_{f_1 f_2 f_3}(\eta, \vec{p}_1; \eta, \vec{p}_2; \eta)}{\bar{W}_{f_1 f_1}(\eta, \vec{p}_1; \eta)} = \frac{{Q_{ik}}^{f_1}(-\vec{p}_1)\delta^{kj}}{2(2\pi)^{3/2}} \Bigg[- \delta^i{}_j \bar{W}_{f_2 f_3}(\vec{p}_2) - p_j^2 \frac{\partial \bar{W}_{f_2 f_3}}{\partial p_i^2} + 2{Q_{f_2}}^{ik}(\vec{p}_2) {Q_{jk}}^{f_3}(\vec{p}_2) \bar{W}_{f_3 f_3}(\vec{p}_2) + 2{Q_{f_3}}^{ik}(-\vec{p}_2) {Q_{jk}}^{f_2}(-\vec{p}_2) \bar{W}_{f_2 f_2}(\vec{p}_2) + \mathcal{O}(\vec{p}_1) \Bigg], \tag{5.71a}$$

which holds for all the combinations of fields listed in table 5.1. This is the counterpart of the consistency relation from spatial diffeomorphisms derived in [158]. It is a consequence of invariance under spatial diffeomorphisms, analyticity and translation and rotational invariance. The consistency relation simplifies significantly in the soft

scalar case $f_1 = H_L$, in which it takes the form

$$\frac{\bar{W}_{H_L f_2 f_3}(\eta, \vec{p}_1; \eta, \vec{p}_2; \eta)}{\bar{W}_{H_L H_L}(\eta, \vec{p}_1; \eta)} = \frac{\delta_{f_2 f_3}}{(2\pi)^{3/2}} \left(1 - \vec{p}_2 \cdot \frac{\partial}{\partial \vec{p}_2}\right) \bar{W}_{f_2 f_2}(\eta, \vec{p}_2; \eta) + \mathcal{O}(\vec{p}_1). \tag{5.71b}$$

Note that the right hand side is proportional to $\delta_{f_2 f_3}$ because the three-point function in the soft-momentum limit $H_L(\vec{p}_1 \to 0)$ is expected to be roughly the two-point function of the two hard fields f_2, f_3, which is diagonal in field space. For two hard scalars, it is also relatively easy to obtain the $\mathcal{O}(\vec{p}_1)$ correction by solving equation (5.65) to next-to-leading order. Again, one can show that the breaking term can be discarded, thereby yielding

$$\begin{aligned}\frac{\bar{W}_{H_L H_L H_L}(\eta, \vec{p}_1; \eta, \vec{p}_2; \eta)}{\bar{W}_{H_L H_L}(\eta, \vec{p}_1; \eta)} = \frac{1}{(2\pi)^{3/2}} \Bigg[1 - \vec{p}_2 \cdot \frac{\partial}{\partial \vec{p}_2} \\ + \vec{p}_1 \cdot \frac{\partial}{\partial \vec{p}_2} + (\vec{p}_2 \cdot \frac{\partial}{\partial \vec{p}_2})(\vec{p}_1 \cdot \frac{\partial}{\partial \vec{p}_2}) - \frac{1}{2} \vec{p}_1 \cdot \vec{p}_2 \left(\frac{\partial}{\partial \vec{p}_2}\right)^2 \Bigg] \bar{W}_{H_L H_L}(\eta, \vec{p}_2; \eta) + \mathcal{O}({\vec{p}_1}^2).\end{aligned} \tag{5.71c}$$

Spatially Flat Gauge Most, if not all, of the consistency relations that have been derived so far apply only in unitary gauge. Yet consistency relations can also be formulated in other gauges. Consider for instance spatially flat gauge, which we recover with a breaking term of the form

$$S_{\text{gf}} = -\frac{M^2}{2} \int d\eta \, d^3p \left(H_L^2 + H_T^2 + H_+^2 + H_-^2\right) \tag{5.72}$$

when we take M to infinity at the end of the calculation. In this limit, the fields H_L, H_T and $H_\pm$ decouple from the rest of the perturbations, which amounts to working in a gauge where $H_L \equiv H_T \equiv H_\pm \equiv 0$. In this gauge, the fluctuations of the scalar field φ contain all the information about the scalar sector of the perturbations.

In order to arrive at a consistency relation in spatially flat gauge that follows from spatial diffeomorphisms, we choose $\alpha = \xi^k(\eta_1, \vec{p}_1)$, $a = f_2(\eta_2, \vec{p}_2)$ and $b = f_3(\eta_3, \vec{p}_3)$ in

f_1	f_2	f_3
$H_{\pm\pm}$	φ	φ
$H_{\pm\pm}$	$H_{\pm\pm}$	$H_{\pm\pm}$

Table 5.2: The different combination of fields for which the consistency relation (5.71a) in spatially flat gauge holds. The field f_1 carries the soft momentum, whereas the two hard fields f_2 and f_3 can be interchanged.

equation (5.55). Since we have the limit of spatially flat gauge in mind, the fields f_2 and f_3 therefore stand for either the scalar φ or the two tensors $H_{\pm\pm}$. Proceeding as above we are led to

$$\begin{aligned}
2a_1^2 p_i^1 \bar{\Gamma}^{ik}{}_{f_2 f_3}\delta_{kj} - \Bigg[& \Big(2p_k^1 Q_{f_2}{}^{ik}(\vec{p}_2) Q_{ij}^{f_3}(\vec{p}_2 - \vec{p}_1) + (p_j^2 - p_j^1) Q_{f_2}{}^{ik}(\vec{p}_2) Q_{ik}{}^{f_3}(\vec{p}_2 - \vec{p}_1)\Big) \bar{W}_{f_3 f_3}(\eta, \vec{p}_2 - \vec{p}_1) \\
& + \Big(2p_k^1 Q_{f_3}{}^{ik}(\vec{p}_1 - \vec{p}_2) Q_{ij}^{f_3}(-\vec{p}_2) - p_j^2 Q_{f_3}{}^{ik}(\vec{p}_1 - \vec{p}_2) Q_{ik}{}^{f_3}(-\vec{p}_2)\Big) \bar{W}_{f_2 f_2}(\eta, -\vec{p}_2) \\
& + (p_j^2 - p_j^1)\delta_{f_2}{}^{\varphi}\delta_{f_3}{}^{\varphi} \bar{W}_{\varphi\varphi}(\eta, \vec{p}_2 - \vec{p}_1) - p_j^2 \delta_{f_3}{}^{\varphi}\delta_{f_2}{}^{\varphi} \bar{W}_{\varphi\varphi}(\eta, -\vec{p}_2) \Bigg] \frac{\delta(\eta - \eta_1)}{(2\pi)^{3/2}} = -i(\Gamma_{\Delta S, j})_{f_2 f_3},
\end{aligned} \tag{5.73}$$

which has essentially the same structure as equation (5.65), since both capture invariance under spatial diffeomorphisms in the reduced formulation of the theory. Because the propagator $W_{g\varphi}$ must fall like $1/M^2$, and because $\mathcal{T}$ does not mix metric perturbations and φ, inspection of equation (5.56) reveals that $(\Gamma_{\Delta S, j})_{f_2 f_3}$ vanishes at all orders in $\vec{p}_1$ for the combination of fields listed on table 5.2. Again, using analyticity we can solve equation (5.73) by taking partial derivatives wrt p_i^1 on both sides of the equation. The unique solution at zeroth order is again given by equation (5.68) with $\partial(\bar{\Gamma}_{\Delta S, j})_{f_2 f_3}/\partial p_1^i$ set to zero. Therefore, the ensuing consistency relations then take the form of equation (5.71a), where this time f_1, f_2 and f_3 are drawn from the values listed in table 5.2. In this gauge we can for instance reliably determine the cubic vertex $\bar{\Gamma}^{H_{\pm\pm}}{}_{\varphi\varphi}$ for a soft tensor and two hard scalars.

Time Diffeomorphisms

Proceeding along similar lines it is possible to derive a consistency relation that follows from invariance under time diffeomorphisms, among a few other assumptions. To do so, we need to work in spatially flat gauge, as the gauge-fixing condition $\varphi = 0$ we used in unitary gauge breaks time diffeomorphisms in an uncontrolled way. One may be tempted to introduce a symmetry breaking mass term for the scalar φ instead, but this choice is not useful, because the scalar field φ would appear in the cubic vertices that are constrained by the Slavnov-Taylor identities (see equation (5.90a)). Because of that, we rather choose to work in the analogue of spatially flat gauge, with the gauge-fixing terms in equation (5.72).

We begin the analysis with equation (5.55) for time diffeomorphisms, $\alpha = \xi^0(\eta_1, \vec{p}_1)$, by setting as usual $a = f_2(\eta_2, \vec{p}_2)$ and $b = f_3(\eta_3, \vec{p}_3)$. Factoring out the momentum-conserving delta function we obtain

$$
\begin{aligned}
&2a_1^2\mathcal{H}_1\delta_{ij}\bar{\Gamma}^{ij}{}_{f_2f_3}(\eta_1,\vec{p}_1;\eta,\vec{p}_2;\eta) + \frac{\partial\bar{\phi}}{\partial\eta_1}\bar{\Gamma}^{\varphi}{}_{f_2f_3}(\eta_1,\vec{p}_1;\eta,\vec{p}_2;\eta) \\
&- \frac{\delta(\eta-\eta_1)}{(2\pi)^{3/2}}\Bigg\{\Big[Q_{f_2}{}^{ij}(\vec{p}_2)Q_{ij}{}^{f_3}(\vec{p}_2-\vec{p}_1)\Big(2\mathcal{H}+\frac{\partial}{\partial\eta}\Big)+\delta_{f_2}{}^{\varphi}\delta^{f_3\varphi}\frac{\partial}{\partial\eta}\Big]\bar{W}_{f_3f_3}(\eta,\vec{p}_2-\vec{p}_1;\eta) \\
&+\Big[Q_{f_3}{}^{ij}(\vec{p}_1-\vec{p}_2)Q_{ij}{}^{f_2}(-\vec{p}_2)\Big(2\mathcal{H}+\frac{\partial}{\partial\eta}\Big)+\delta^{f_2\varphi}\delta_{f_3}{}^{\varphi}\frac{\partial}{\partial\eta}\Big]\bar{W}_{f_2f_2}(\eta,-\vec{p}_2;\eta)\Bigg\} = (\bar{\Gamma}_{\Delta S,0})_{f_2f_3},
\end{aligned}
\tag{5.74}
$$

where the time derivatives only act on the first time argument of the power spectrum. Two different sources can potentially contribute to the symmetry breaking term on the right hand side of the equation: The first is due to the gauge-fixing term (5.72), but, as in the case of spatial diffeomorphisms in spatially flat gauge, combining equation (5.56) with equation (5.91a) we immediately find that in the limit $M \to \infty$ this contribution vanishes for the combination of fields listed on table 5.2. The second contribution arises because the reduced formulation of the theory is only invariant

under space-independent diffeomorphisms ($\vec{p}_1 = 0$), but not for time diffeomorphisms with arbitrary spatial dependence. This implies that $(\bar{\Gamma}_{\Delta S,0})_{f_2 f_3}$ must vanish at zeroth order in $\vec{p}_1$, but not at higher orders. If the two hard fields are of the same type however, $f_2 = f_3$, the invariance of $(\bar{\Gamma}_{\Delta S,0})_{f_2 f_2}(\eta_1, \vec{p}_1; \eta_2, \vec{p}_2; \eta_3)$ under $\vec{p}_2 \to \vec{p}_1 - \vec{p}_2$ implies that the breaking term cannot contain a linear piece in $\vec{p}_1$ either. In what follows we shall restrict our consideration to the two cases in which the breaking term certainly vanishes.

The problem with equation (5.74) is that it contains the vertex $\delta_{ij}\bar{\Gamma}^{ij}{}_{f_2 f_3}$, which does not play any role in the limit in which H_L becomes infinitely heavy. But fortunately, we have already calculated this vertex in the analysis of Section 5.5.2 that led to the solution (5.68), with $\partial(\bar{\Gamma}_{\Delta S,j})_{f_2 f_3}/\partial p_i^1 = 0$ for the cases listed on table 5.2. Substituting that solution into (5.74) we then obtain a Slavnov-Taylor identity for time diffeomorphisms that contains the relevant vertex $\bar{\Gamma}^{\varphi}{}_{f_2 f_3}$ alone. Convolving the resulting equation with the φ propagator, integrating and evaluating at equal times we finally arrive at the consistency relation

$$\frac{\bar{W}_{\varphi f_2 f_3}(\eta, \vec{p}_1; \eta, \vec{p}_2; \eta)}{\bar{W}_{\varphi\varphi}(\eta, \vec{p}_1; \eta)} = \frac{1}{(2\pi)^{3/2}\bar{\phi}'}\Bigg\{2\Big[\delta_{f_2 f_3}\Big(2\mathcal{H} + \frac{\partial}{\partial\eta}\Big) - 2\mathcal{H}\delta_{f_2\varphi}\delta_{f_3\varphi}\Big]\bar{W}_{f_2 f_2}(\eta, \vec{p}_2)$$
$$- \mathcal{H}\Big[-p_2^i\frac{\partial\bar{W}_{f_2 f_2}(\eta, \vec{p}_2)}{\partial p_2^i}\delta_{f_2 f_3} + \delta_{f_2 f_3}\bar{W}_{f_2 f_2}(\eta, \vec{p}_2) - 4\delta_{f_2\varphi}\delta_{f_3\varphi}\bar{W}_{\varphi\varphi}(\eta, -\vec{p}_2)\Big]\Bigg\} + \mathcal{O}(\vec{p}_1). \tag{5.75a}$$

This consistency relation relies on invariance under time diffeomorphisms, although it also depends on the invariance under spatial diffeomorphisms, analyticity and rotational and translational invariance of the quantum state of the perturbations. Whereas most if not all of the consistency relations that have been discussed in the literature so far involve spatial derivatives of the power spectrum, this relation also contains its time derivatives.

Perhaps because equation (5.75a) is quite general, it is a rather formidable ex-

pression. We can simplify its form by setting for instance $f_2 = \varphi$ and $f_3 = \varphi$. This was one of the cases in which the gauge fixing terms do not contribute at zeroth or linear order in momentum, which allows to extend the previous consistency relation to the next order in the soft momentum. The calculation progresses in the same way, the only difference being that the equations for the vertices need to be solved to a higher order.

$$\frac{\bar{W}_{\varphi\varphi\varphi}(\eta,\vec{p}_1;\eta,\vec{p}_2;\eta)}{\bar{W}_{\varphi\varphi}(\eta,\vec{p}_1;\eta)} = \frac{1}{(2\pi)^{3/2}\bar{\phi}'}\Bigg\{2\frac{\partial\bar{W}_{\varphi\varphi}(\eta,\vec{p}_2;\eta)}{\partial\eta} - \vec{p}_1\cdot\frac{\partial^2\bar{W}_{\varphi\varphi}(\eta,\vec{p}_2;\eta)}{\partial\vec{p}_2\,\partial\eta}$$
$$-\mathcal{H}\Bigg[-3+(3\vec{p}_1-\vec{p}_2)\cdot\frac{\partial}{\partial\vec{p}_2}+\Big(\vec{p}_2\cdot\frac{\partial}{\partial\vec{p}_2}\Big)\Big(\vec{p}_1\cdot\frac{\partial}{\partial\vec{p}_2}\Big)-\frac{\vec{p}_1\cdot\vec{p}_2}{2}\left(\frac{\partial}{\partial\vec{p}_2}\right)^2\Bigg]\bar{W}_{\varphi\varphi}(\eta,\vec{p}_2;\eta)\Bigg\}+\mathcal{O}({\vec{p}_1}^2). \tag{5.75b}$$

5.6 Summary and Conclusions

We have explored the constraints that diffeomorphism invariance imposes on the correlation functions of cosmological perturbations. Because these basically follow from symmetry, we have relied on the Lagrangian formulation of the theory, and the corresponding functional integral approach for its perturbative quantization. In this approach, expectation values can be calculated by introducing a closed time contour. Other than that, the formalism is formally identical to the one used to calculate in-out matrix elements.

Our most general constraints take the form of master identities for the generator of connected correlators iW and the generator of one-particle-irreducible diagrams $i\Gamma$ in an arbitrary gauge. The former are closely related to Schwinger-Dyson equations, which merely state that the classical equations of motion hold in the quantum theory, whereas the latter assume the form of Slavnov-Taylor identities that mirror the (broken) symmetry of the underlying theory. We showed that both sets of identities are equivalent.

Because diffeomorphism invariance has to be broken in order to quantize the the-

ory, the change of the action under the broken diffeomorphisms plays a crucial role in the Schwinger-Dyson and Slavnov-Taylor identities. The broken symmetry enters through an additional generator, containing an insertion of a single vertex determined by the change of the action under a diffeomorphism transformation. Therefore, these identities are also a direct reflection of how the symmetry is broken, and not just of the invariance of the theory. In order to keep such breaking under control, we cannot gauge-fix some of the cosmological perturbations to zero, but have to give some of these perturbations a mass term. This is analogous to the use of R_ξ gauges in gauge theories. As a result, all the fields survive in the gauge-fixed theory, even though some of them are just gauge artifacts in the original invariant theory. A compromise emerges if one lets the symmetry-breaking masses approach infinity, which effectively decouples the corresponding fields from the theory, while keeping the symmetry breaking under control. In theories in which the metric components h_{00} and h_{0i} are auxiliary fields, it is also possible to integrate the latter out; the thus "reduced theory" remains invariant under spatial diffeomorphisms, although it loses invariance under spatially dependent time diffeomorphisms, at least in their original form.

We have formulated our identities in DeWitt notation, which allowed us to focus on the conceptual aspects of the identities, rather than on the specific details of diffeomorphism transformations. Consequently, our identities in fact hold in any theory invariant under a set of symmetries that acts linearly (though possibly inhomogeneously) on the fields. For all those theories, for instance, the Slavnov-Taylor identities state how a three-point function with an insertion of the change of one of the fields under the inhomogeneous component of the transformation is related to the change of the two-point function solely under the linear component of the transformation. These identities provide useful checks of the self-consistency of the theory, and could be used to diagnose inconsistencies in any calculation of expectation values of cosmological perturbations.

Yet perhaps the most important application of these identities is the formulation

of consistency relations that relate expectation values of products of different numbers of cosmological perturbation fields. To do so, we had to appeal to the reduced formulation of the theory in order to limit the total number of scalars to a single field. In this case, diffeomorphism invariance alone does not suffice to extract definite predictions from the theory, mostly because it is not possible to reduce the scalar field sector to a single field without losing control of the symmetry. On the other hand, the additional assumption of analyticity allowed us to derive consistency relations that constrain the single dynamically relevant field in the gauge-fixed theory, in the limit in which one of the field momenta approaches zero. We were thus able to reproduce in a different field parameterization the consistency relations presented in [158]. These are captured in our equations (5.71), which embody consistency relations that follow from spatial diffeomorphisms in unitary gauge, and hold for the set of fields listed on table 5.1. We also extended these results to novel relations in spatially flat gauge, still in the reduced formulation of the theory. In this gauge, the consistency relations also take the form of equation (5.71a), with the fields for which it applies being listed in Table 5.2. In the same gauge, we finally derived new consistency relations that follow from invariance under time diffeomorphisms. These relations are listed in equations (5.75), and, as opposed to those involving invariance under spatial diffeomorphisms, contain time derivatives of power spectra. Although it is natural to think of these consistency relations as constraints on the properties of the primordial perturbations created during inflation, they apply no matter what the evolution of the background is, provided that the analyticity assumption applies.

All the consistency relations that we have discussed here are close analogues of the relation between the gravitational vertex and matter self-energy that ultimately enforces the equivalence principle in general relativity. In that sense, unfortunately, our conclusions do not appear to resolve the tension between Kretschmann's objection to Einstein's principle of general covariance, and the apparent physical implications of local symmetries, such as the equivalence principle or the consistency relations we

just derived. If it turned out that the primordial perturbations did not obey the consistency relations we presented here, we would probably argue that they were not generated during a period of single-field inflation, rather than concluding that diffeomorphism invariance is somehow broken.

Appendix

5.A Irreducible Tensors

As we discuss in the main text, in cosmological perturbation theory it is convenient to work with perturbations that transform irreducibly under the isometries of the cosmological background: spatial rotations and translations. We thus introduce a set of eleven irreducible tensors $Q_{\mu\nu}{}^{f}(\vec{x};\vec{p})$ and $Q^{\varphi}(\vec{x},\vec{p})$ that we use as basis elements in an expansion of arbitrary cosmological perturbations,

$$h_{\mu\nu}(\eta,\vec{x}) = \sum_{f}\int d^3p\, Q_{\mu\nu}{}^{f}(\vec{x};\vec{p})\, f(\eta,\vec{p}), \quad \varphi(\eta,\vec{x}) = \int d^3p\, Q^{\varphi}(\vec{x};\vec{p})\varphi(\eta,\vec{p}). \quad (5.76)$$

These tensors are plane waves, and although they depend on time through the scale factor, we suppress the time argument for simplicity,

$$Q_{\mu\nu}{}^{f}(\vec{x};\vec{p}) \equiv a^2\frac{e^{i\vec{p}\cdot\vec{x}}}{(2\pi)^{3/2}}Q_{\mu\nu}{}^{f}(\vec{p}), \quad Q^{\varphi}(\vec{x};\vec{p}) \equiv \frac{e^{i\vec{p}\cdot\vec{x}}}{(2\pi)^{3/2}}Q^{\varphi}(\vec{p}), \quad (5.77)$$

with non-vanishing momentum-dependent components

$$\textbf{scalars} \quad Q^{\varphi} = 1, \tag{5.78a}$$

$$Q_{00}{}^{A} = -2, \tag{5.78b}$$

$$Q_{0i}{}^{B} = \frac{ip_i}{p}, \tag{5.78c}$$

$$Q_{ij}{}^{H_L} = 2\delta_{ij}, \tag{5.78d}$$

$$Q_{ij}{}^{H_T} = 2\left(\frac{1}{3}\delta_{ij} - \frac{p_i p_j}{p^2}\right), \tag{5.78e}$$

$$\textbf{vectors} \quad Q_{0i}{}^{B_\pm} = -\hat{\epsilon}_i^{\pm} \tag{5.79a}$$

$$Q_{ij}{}^{H_\pm} = -i\left(\frac{p_i}{p}\hat{\epsilon}_j^{\pm} + \frac{p_j}{p}\hat{\epsilon}_i^{\pm}\right), \tag{5.79b}$$

$$\textbf{tensors} \quad Q_{ij}{}^{H_{\pm\pm}} = 2\hat{\epsilon}_i^{\pm}\hat{\epsilon}_j^{\pm}. \tag{5.80}$$

Here, $\hat{\epsilon}^{\pm}(\vec{p})$ are two orthonormal transverse vectors with[2]

$$\vec{p}\cdot\hat{\epsilon}^{\pm} = 0, \tag{5.81a}$$

$$\vec{p}\times\hat{\epsilon}^{\pm} = \mp i\, p\, \hat{\epsilon}^{\pm}. \tag{5.81b}$$

Note that the polarization vectors are complex, and that $(\hat{\epsilon}^{\pm})^* = \hat{\epsilon}^{\mp}$. Hence, it follows that $(\hat{\epsilon}^{\pm})^*\cdot\hat{\epsilon}^{\pm} = \hat{\epsilon}^{\mp}\cdot\hat{\epsilon}^{\pm} = 1$, but $\hat{\epsilon}^{\pm}\cdot\hat{\epsilon}^{\pm} = (\hat{\epsilon}^{\mp})^*\cdot\hat{\epsilon}^{\pm} = 0$.

Given arbitrary metric and scalar perturbations $h_{\mu\nu}(x)$ and $\varphi(x)$ we would like to find their components in the basis of tensors above. We thus introduce a correspond-

[2]These vectors can be taken to be $\hat{\epsilon}^{\pm} = R(\hat{p})\frac{1}{\sqrt{2}}(\hat{e}_x \pm i\hat{e}_y)$, where $R(\hat{p})$ is a standard rotation mapping the z axis to the $\hat{p}$ direction.

ing set of projection operators to project onto those components,

$$f(\eta,\vec{p}) = \int d^3x\, {Q_f}^{\mu\nu}(\vec{p};\vec{x}) h_{\mu\nu}(\eta,\vec{x}), \quad \varphi(\eta,\vec{p}) = \int d^3x\, Q_\varphi(\vec{p};\vec{x})\varphi(\eta,\vec{x}). \tag{5.82}$$

These tensors are

$${Q_f}^{\mu\nu}(\vec{p};\vec{x}) \equiv \frac{1}{a^2}\frac{e^{-i\vec{p}\cdot\vec{x}}}{(2\pi)^{3/2}} {Q_f}^{\mu\nu}(\vec{p}), \quad Q_\varphi(\vec{p};\vec{x}) \equiv \frac{e^{-i\vec{p}\cdot\vec{x}}}{(2\pi)^{3/2}} Q_\varphi(\vec{p}), \tag{5.83}$$

where the non-vanishing momentum-dependent components read

$$Q_\varphi = 1, \tag{5.84a}$$

$${Q_A}^{00} = -\frac{1}{2}, \tag{5.84b}$$

$${Q_B}^{0i} = -\frac{i}{2}\frac{p^i}{p}, \tag{5.84c}$$

$${Q_{H_L}}^{ij} = \frac{1}{6}\delta^{ij}, \tag{5.84d}$$

$${Q_{H_T}}^{ij} = \frac{3}{4}\left(\frac{1}{3}\delta^{ij} - \frac{p^i p^j}{p^2}\right), \tag{5.84e}$$

$${Q_{B_\pm}}^{0i} = -\frac{1}{2}\hat{\epsilon}^i_\mp, \tag{5.85a}$$

$${Q_{H_\pm}}^{ij} = \frac{i}{2}\left(\frac{p^i}{p}\hat{\epsilon}^j_\mp + \frac{p^j}{p}\hat{\epsilon}^i_\mp\right), \tag{5.85b}$$

$${Q_{H_{\pm\pm}}}^{ij} = \frac{1}{2}\hat{\epsilon}^i_\mp\hat{\epsilon}^j_\mp, \tag{5.86}$$

and vector and tensor indices are raised with the Euclidean metric δ^{ij} (note that this convention does not apply to the projectors Q themselves.) These projection

operators satisfy the completeness relation

$$\int d^3x\, Q_{f_1}{}^{\mu\nu}(\vec{p}_1;\vec{x})Q_{\mu\nu}{}^{f_2}(\vec{x};\vec{p}_2) = \delta_{f_1}{}^{f_2}\,\delta^{(3)}(\vec{p}_1-\vec{p}_2). \tag{5.87}$$

It is also convenient to work with the irreducible components of the four-vectors ξ^μ that parameterize the different infinitesimal diffeomorphisms. We hence write

$$\xi^\alpha(\eta,\vec{x}) = \int d^3p\, Q^\alpha{}_{\bar{\alpha}}(\vec{x};\vec{p})\,\xi^{\bar{\alpha}}(\eta,\vec{p}) \quad \text{and} \quad \xi^{\bar{\alpha}}(\eta,\vec{p}) = \int d^3x\, Q^{\bar{\alpha}}{}_\alpha(\vec{p};\vec{x})\,\xi^\alpha(\eta,\vec{x}), \tag{5.88}$$

where the non-vanishing components of these tensors are

$$Q^0{}_0(\vec{x};\vec{p}) = \frac{e^{i\vec{p}\cdot\vec{x}}}{(2\pi)^{3/2}}, \quad Q^0{}_0(\vec{p};\vec{x}) = \frac{e^{-i\vec{p}\cdot\vec{x}}}{(2\pi)^{3/2}}, \tag{5.89a}$$

$$Q^i{}_L(\vec{x};\vec{p}) = -\frac{ip_i}{p}\frac{e^{i\vec{p}\cdot\vec{x}}}{(2\pi)^{3/2}}, \quad Q^L{}_i(\vec{p};\vec{x}) = \frac{ip_i}{p}\frac{e^{-i\vec{p}\cdot\vec{x}}}{(2\pi)^{3/2}}, \tag{5.89b}$$

$$Q^i{}_\pm(\vec{x};\vec{p}) = \epsilon^i_\pm(\vec{p})\frac{e^{i\vec{p}\cdot\vec{x}}}{(2\pi)^{3/2}}, \quad Q^\pm{}_i(\vec{p};\vec{x}) = \epsilon^\mp_i(\vec{p})\frac{e^{-i\vec{p}\cdot\vec{x}}}{(2\pi)^{3/2}}. \tag{5.89c}$$

As we discuss in the main text, relations that involve all these projection tensors simplify considerably in DeWitt notation.

5.B Transformation under Diffeomorphisms

In order to calculate how the irreducible perturbations introduced above transform under diffeomorphisms, we need to combine equations (5.12) and (5.13) with (5.17). Using the results of Appendix 5.A we find for time diffeomorphisms ($\alpha = \xi^0$), longitudinal diffeomorphisms ($\alpha = L$) and transverse diffeomorphisms of either helicity ($\alpha = \pm$)

$$\mathcal{S}_{f(\eta_1,\vec{p}_1)\xi^0(\eta_2,\vec{p}_2)} = \delta(\vec{p}_1-\vec{p}_2)\left[\delta_f{}^A\left(\mathcal{H}_1+\frac{\partial}{\partial\eta_1}\right) - p_1\delta_f{}^B + \mathcal{H}_1\delta_f{}^{H_L} + \frac{\partial\bar{\phi}}{\partial\eta_1}\delta_f{}^\varphi\right]\delta(\eta_1-\eta_2), \tag{5.90a}$$

$$\mathcal{S}_{f(\eta_1,\vec{p}_1)\xi^L(\eta_2,\vec{p}_2)} = \delta(\vec{p}_1-\vec{p}_2)\left[-\delta_f{}^B\frac{d}{d\eta_1}+\frac{p_1}{3}\delta_f{}^{H_L}-p_1\delta_f{}^{H_T}\right]\delta(\eta_1-\eta_2), \tag{5.90b}$$

$$\mathcal{S}_{f(\eta_1,\vec{p}_1)\xi^\pm(\eta_2,\vec{p}_2)} = \delta(\vec{p}_1-\vec{p}_2)\left[-\delta_f{}^{B_\pm}\frac{d}{d\eta_1}-p_1\delta_f{}^{H_\pm}\right]\delta(\eta_1-\eta_2). \tag{5.90c}$$

In these equations a prime denotes a derivative with respect to conformal time η, $\mathcal{H}\equiv a'/a$ and $p\equiv|\vec{p}|$. The components of the transformations linear in the fields are

$$\begin{aligned}\mathcal{T}_{f_1(\eta_1,\vec{p}_1)}{}^{f_2(\eta_2,\vec{p}_2)}{}_{\xi^0(\eta_3,\vec{p}_3)} = \frac{\delta(\vec{p}_2+\vec{p}_3-\vec{p}_1)}{(2\pi)^{3/2}}\Bigg[&\delta^A_{f_1}\delta^{f_2}_A\left(2\mathcal{H}_1-\frac{\partial}{\partial\eta_2}-2\frac{\partial}{\partial\eta_3}\right)\\ -4ip_i^3Q_{f_1}{}^{0i}(\vec{p}_1)\delta^{f_2}_A+&Q_{f_1}{}^{0i}(\vec{p}_1)Q_{0i}{}^{f_2}(\vec{p}_2)\left(4\mathcal{H}_1-2\frac{\partial}{\partial\eta_2}-2\frac{\partial}{\partial\eta_3}\right)+2ip_j^3Q_{f_1}{}^{ij}(\vec{p}_1)Q_{i0}{}^{f_2}(\vec{p}_2)+\\ &+Q_{f_1}{}^{ij}(\vec{p}_1)Q_{ij}{}^{f_2}(\vec{p}_2)\left(2\mathcal{H}_1-\frac{\partial}{\partial\eta_2}\right)-\delta^\varphi_{f_1}\delta^{f_2}_\varphi\frac{\partial}{\partial\eta_2}\Bigg]\delta(\eta_1-\eta_2)\delta(\eta_1-\eta_3),\end{aligned} \tag{5.91a}$$

$$\begin{aligned}\mathcal{T}_{f_1(\eta_1,\vec{p}_1)}{}^{f_2(\eta_2,\vec{p}_2)}{}_{\xi^k(\eta_3,\vec{p}_3)} = \frac{\delta(\vec{p}_2+\vec{p}_3-\vec{p}_1)}{(2\pi)^{3/2}}\Bigg[&\delta^A_{f_1}Q_{0k}{}^{f_2}(\vec{p}_2)\frac{\partial}{\partial\eta_3}+i\delta^A_{f_1}\delta^{f_2}_Ap_k^2+\\ &+2iQ_{f_1}{}^{0i}(\vec{p}_1)Q_{0k}{}^{f_2}(\vec{p}_2)p_i^3-2Q_{f_1}{}^{0i}(\vec{p}_1)Q_{ik}{}^{f_2}(\vec{p}_2)\frac{\partial}{\partial\eta_3}+2iQ_{f_1}{}^{i0}(\vec{p}_1)Q_{i0}{}^{f_2}(\vec{p}_2)p_k^2+\\ &+2iQ_{f_1}{}^{ij}(\vec{p}_1)Q_{ik}{}^{f_2}(\vec{p}_2)p_j^3+iQ_{f_1}{}^{ij}(\vec{p}_1)Q_{ij}{}^{f_2}(\vec{p}_2)p_k^2+i\delta^\varphi_{f_1}\delta^{f_2}_\varphi p_k^2\Bigg]\delta(\eta_1-\eta_2)\delta(\eta_1-\eta_3),\end{aligned} \tag{5.91b}$$

where the non-vanishing components of the tensors $Q_f{}^{\mu\nu}(\vec{p})$ and $Q_{\mu\nu}{}^f(\vec{p})$ are given in Appendix 5.A. Contracting equation (5.91b) with $Q^k{}_L$ and $Q^k{}_\pm$ one readily recovers the transformations under longitudinal and transverse diffeomorphisms. Note that for some choices of the fields, these expressions can be further simplified. For instance, for $f_1=H_L$, $Q_{f_1}{}^{ij}(\vec{p}_1)Q_{ij}{}^{f_2}(\vec{p}_2)=\delta^{f_2}_{H_L}$.

Bibliography

[1] C. Armendariz-Picon and J. T. Neelakanta, JCAP **1403**, 049 (2014), 1309.6971.

[2] C. Armendariz-Picon and J. T. Neelakanta, JCAP **1212**, 009 (2012), 1210.3017.

[3] J. T. Neelakanta (2015), 1501.03513.

[4] C. Armendariz-Picon, J. T. Neelakanta, and R. Penco, JCAP **1501**, 035 (2015), 1411.0036.

[5] V. M. Slipher, Proc.Am.Phil.Soc. **56**, 403 (1917).

[6] E. Hubble, Proc.Nat.Acad.Sci. **15**, 168 (1929).

[7] G. Lemaitre, Annales Soc.Sci.Brux.Ser.I Sci.Math.Astron.Phys. **A47**, 49 (1927).

[8] A. Friedman, Zeitschrift fr Physik **10**, 377 (1922).

[9] A. Friedmann, Zeitschrift fr Physik **21**, 326 (1924).

[10] Robertson, H. P., Astrophys. J. **82**, 284 (1935).

[11] Walker, Proc. Lond. Math. Soc. **42**, 90 (1937).

[12] A. A. Penzias and R. W. Wilson, Astrophys.J. **142**, 419 (1965).

[13] G. F. Smoot, C. Bennett, A. Kogut, E. Wright, J. Aymon, et al., Astrophys.J. **396**, L1 (1992).

[14] P. Ade et al. (Planck), Astron.Astrophys. **571**, A15 (2014), 1303.5075.

[15] F. Zwicky, Helv.Phys.Acta **6**, 110 (1933).

[16] L. Evans and P. Bryant, JINST **3**, S08001 (2008).

[17] M. Aartsen et al. (IceCube), Phys.Rev.Lett. **111**, 021103 (2013), 1304.5356.

[18] S. Perlmutter et al. (Supernova Cosmology Project), Astrophys.J. **517**, 565 (1999), astro-ph/9812133.

[19] A. G. Riess et al. (Supernova Search Team), Astron.J. **116**, 1009 (1998), astro-ph/9805201.

[20] R. A. Alpher, H. Bethe, and G. Gamow, Phys. Rev. **73**, 803 (1948).

[21] S. Weinberg, Phys.Rev. **D70**, 043541 (2004), astro-ph/0401313.

[22] V. F. Mukhanov and G. V. Chibisov, JETP Lett. **33**, 532 (1981).

[23] J. M. Maldacena, JHEP **0305**, 013 (2003), astro-ph/0210603.

[24] J. L. Feng, Ann.Rev.Astron.Astrophys. **48**, 495 (2010), 1003.0904.

[25] S. Colombi, S. Dodelson, and L. M. Widrow, Astrophys.J. **458**, 1 (1996), astro-ph/9505029.

[26] J. Sommer-Larsen and A. Dolgov, Astrophys.J. **551**, 608 (2001), astro-ph/9912166.

[27] P. Colin, V. Avila-Reese, and O. Valenzuela, Astrophys.J. **542**, 622 (2000), astro-ph/0004115.

[28] P. Bode, J. P. Ostriker, and N. Turok, Astrophys.J. **556**, 93 (2001), astro-ph/0010389.

[29] Z. Haiman, R. Barkana, and J. P. Ostriker, AIP Conf.Proc. **586**, 136 (2001), astro-ph/0103050.

[30] H. de Vega and N. Sanchez (2011), 1109.3187.

[31] M. Viel, G. D. Becker, J. S. Bolton, and M. G. Haehnelt, Phys.Rev. **D88**, 043502 (2013), 1306.2314.

[32] L. G. van den Aarssen, T. Bringmann, and C. Pfrommer, Phys.Rev.Lett. **109**, 231301 (2012), 1205.5809.

[33] A. M. Green, S. Hofmann, and D. J. Schwarz, JCAP **0508**, 003 (2005), astro-ph/0503387.

[34] B. Moore, J. Diemand, J. Stadel, and T. R. Quinn (2005), astro-ph/0502213.

[35] S. Profumo, K. Sigurdson, and M. Kamionkowski, Phys.Rev.Lett. **97**, 031301 (2006), astro-ph/0603373.

[36] A. Loeb and M. Zaldarriaga, Phys.Rev. **D71**, 103520 (2005), astro-ph/0504112.

[37] E. Bertschinger, Phys.Rev. **D74**, 063509 (2006), astro-ph/0607319.

[38] F.-Y. Cyr-Racine, R. de Putter, A. Raccanelli, and K. Sigurdson, Phys.Rev. **D89**, 063517 (2014), 1310.3278.

[39] W. H. Press and P. Schechter, Astrophys.J. **187**, 425 (1974).

[40] J. Bond, S. Cole, G. Efstathiou, and N. Kaiser, Astrophys.J. **379**, 440 (1991).

[41] K. Abazajian, Phys.Rev. **D73**, 063513 (2006), astro-ph/0512631.

[42] M. Viel, J. Lesgourgues, M. G. Haehnelt, S. Matarrese, and A. Riotto, Phys.Rev. **D71**, 063534 (2005), astro-ph/0501562.

[43] A. Boyarsky, J. Lesgourgues, O. Ruchayskiy, and M. Viel, JCAP **0905**, 012 (2009), 0812.0010.

[44] R. S. de Souza, A. Mesinger, A. Ferrara, Z. Haiman, R. Perna, et al., Mon.Not.Roy.Astron.Soc. **432**, 3218 (2013), 1303.5060.

[45] T. Bringmann and S. Hofmann, JCAP **0407**, 016 (2007), hep-ph/0612238.

[46] C. M. Muller, Phys.Rev. **D71**, 047302 (2005), astro-ph/0410621.

[47] E. Calabrese, M. Migliaccio, L. Pagano, G. De Troia, A. Melchiorri, et al., Phys.Rev. **D80**, 063539 (2009).

[48] A. L. Serra and M. J. d. L. D. Romero, Mon.Not.Roy.Astron.Soc. **415**, 74 (2011), 1103.5465.

[49] I. H. Gilbert, Astrophys.J. **144**, 233 (1966).

[50] J. Bond and A. Szalay, Astrophys.J. **274**, 443 (1983).

[51] R. H. Brandenberger, N. Kaiser, and N. Turok, Phys.Rev. **D36**, 2242 (1987).

[52] D. Boyanovsky, H. de Vega, and N. Sanchez, Phys.Rev. **D78**, 063546 (2008), 0807.0622.

[53] P. McDonald et al. (SDSS Collaboration), Astrophys.J.Suppl. **163**, 80 (2006), astro-ph/0405013.

[54] S. Das, T. Louis, M. R. Nolta, G. E. Addison, E. S. Battistelli, et al., JCAP **1404**, 014 (2014), 1301.1037.

[55] K. Story, C. Reichardt, Z. Hou, R. Keisler, K. Aird, et al., Astrophys.J. **779**, 86 (2013), 1210.7231.

[56] S. Weinberg, *Cosmology* (2008).

[57] A. Lewis, A. Challinor, and A. Lasenby, Astrophys.J. **538**, 473 (2000), astro-ph/9911177.

[58] A. Lewis and S. Bridle, Phys.Rev. **D66**, 103511 (2002), astro-ph/0205436.

[59] A. Gelman and D. B. Rubin, Statist.Sci. **7**, 457 (1992).

[60] C. Bennett et al. (WMAP), Astrophys.J.Suppl. **208**, 20 (2013), 1212.5225.

[61] B. A. Reid, W. J. Percival, D. J. Eisenstein, L. Verde, D. N. Spergel, et al., Mon.Not.Roy.Astron.Soc. **404**, 60 (2010), 0907.1659.

[62] J. Dunkley, E. Calabrese, J. Sievers, G. Addison, N. Battaglia, et al., JCAP **1307**, 025 (2013), 1301.0776.

[63] R. Barkana and A. Loeb, Phys.Rept. **349**, 125 (2001), astro-ph/0010468.

[64] K. Sigurdson, M. Doran, A. Kurylov, R. R. Caldwell, and M. Kamionkowski, Phys.Rev. **D70**, 083501 (2004), astro-ph/0406355.

[65] P. J. Fox, R. Harnik, J. Kopp, and Y. Tsai, Phys.Rev. **D84**, 014028 (2011), 1103.0240.

[66] P. J. Fox, R. Harnik, J. Kopp, and Y. Tsai, Phys.Rev. **D85**, 056011 (2012), 1109.4398.

[67] J. M. Cornell, S. Profumo, and W. Shepherd, Phys.Rev. **D88**, 015027 (2013), 1305.4676.

[68] M. Aartsen et al. (IceCube Collaboration), Phys.Rev. **D88**, 122001 (2013), 1307.3473.

[69] S. Desai et al. (Super-Kamiokande Collaboration), Phys.Rev. **D70**, 083523 (2004), hep-ex/0404025.

[70] T. Tanaka et al. (Super-Kamiokande Collaboration), Astrophys.J. **742**, 78 (2011), 1108.3384.

[71] C. Boehm and R. Schaeffer, Astron.Astrophys. **438**, 419 (2005), astro-ph/0410591.

[72] N. F. Bell, E. Pierpaoli, and K. Sigurdson, Phys.Rev. **D73**, 063523 (2006), astro-ph/0511410.

[73] G. Mangano, A. Melchiorri, P. Serra, A. Cooray, and M. Kamionkowski, Phys.Rev. **D74**, 043517 (2006), astro-ph/0606190.

[74] I. M. Shoemaker, Phys.Dark Univ. **2**, 157 (2013), 1305.1936.

[75] J. D. Wells (1994), hep-ph/9404219.

[76] E. W. Kolb and M. S. Turner, Front.Phys. **69**, 1 (1990).

[77] N. Arkani-Hamed, D. P. Finkbeiner, T. R. Slatyer, and N. Weiner, Phys.Rev. **D79**, 015014 (2009), 0810.0713.

[78] P. D. Serpico and G. G. Raffelt, Phys.Rev. **D70**, 043526 (2004), astro-ph/0403417.

[79] M. R. Buckley and P. J. Fox, Phys.Rev. **D81**, 083522 (2010), 0911.3898.

[80] J. L. Feng, M. Kaplinghat, and H.-B. Yu, Phys.Rev. **D82**, 083525 (2010), 1005.4678.

[81] L. G. van den Aarssen, T. Bringmann, and Y. C. Goedecke, Phys.Rev. **D85**, 123512 (2012), 1202.5456.

[82] R. Agnese et al. (CDMS Collaboration), Phys.Rev.Lett. **111**, 251301 (2013), 1304.4279.

[83] A. Lewis and A. Challinor, Phys.Rev. **D66**, 023531 (2002), astro-ph/0203507.

[84] O. Adriani et al. (PAMELA Collaboration), Nature **458**, 607 (2009), 0810.4995.

[85] J. Chang, J. Adams, H. Ahn, G. Bashindzhagyan, M. Christl, et al., Nature **456**, 362 (2008).

[86] A. A. Abdo et al. (Fermi LAT Collaboration), Phys.Rev.Lett. **102**, 181101 (2009), 0905.0025.

[87] M. Cirelli, M. Kadastik, M. Raidal, and A. Strumia, Nucl.Phys. **B813**, 1 (2009), 0809.2409.

[88] J. D. March-Russell and S. M. West, Phys.Lett. **B676**, 133 (2009), 0812.0559.

[89] I. Cholis, L. Goodenough, D. Hooper, M. Simet, and N. Weiner, Phys.Rev. **D80**, 123511 (2009), 0809.1683.

[90] A. Sommerfeld, Annalen der Physik **403**, 257 (1931), ISSN 1521-3889, URL http://dx.doi.org/10.1002/andp.19314030302.

[91] J. Hisano, S. Matsumoto, M. M. Nojiri, and O. Saito, Phys.Rev. **D71**, 063528 (2005), hep-ph/0412403.

[92] T. R. Slatyer, N. Padmanabhan, and D. P. Finkbeiner, Phys.Rev. **D80**, 043526 (2009), 0906.1197.

[93] S. Galli, F. Iocco, G. Bertone, and A. Melchiorri, Phys.Rev. **D84**, 027302 (2011), 1106.1528.

[94] P. Ade et al. (Planck Collaboration) (2015), 1502.01589.

[95] L. Ackerman, M. R. Buckley, S. M. Carroll, and M. Kamionkowski, Phys.Rev. **D79**, 023519 (2009), 0810.5126.

[96] M. Blennow, E. Fernandez-Martinez, O. Mena, J. Redondo, and P. Serra, JCAP **1207**, 022 (2012), 1203.5803.

[97] J. L. Feng, M. Kaplinghat, and H.-B. Yu, Phys.Rev.Lett. **104**, 151301 (2010), 0911.0422.

[98] B. Moore, Nature **370**, 629 (1994).

[99] A. A. Klypin, A. V. Kravtsov, O. Valenzuela, and F. Prada, Astrophys.J. **522**, 82 (1999), astro-ph/9901240.

[100] S. Mashchenko, H. Couchman, and J. Wadsley, Nature **442**, 539 (2006), astro-ph/0605672.

[101] J. S. Bullock, A. V. Kravtsov, and D. H. Weinberg, Astrophys.J. **539**, 517 (2000), astro-ph/0002214.

[102] N. Yoshida, V. Springel, S. D. White, and G. Tormen, Astrophys.J. **544**, L87 (2000), astro-ph/0006134.

[103] M. Kaplinghat, L. Knox, and M. S. Turner, Phys.Rev.Lett. **85**, 3335 (2000), astro-ph/0005210.

[104] E. Komatsu et al. (WMAP Collaboration), Astrophys.J.Suppl. **192**, 18 (2011), 1001.4538.

[105] J. Hamann, S. Hannestad, G. G. Raffelt, I. Tamborra, and Y. Y. Wong, Phys.Rev.Lett. **105**, 181301 (2010), 1006.5276.

[106] J. Dunkley, R. Hlozek, J. Sievers, V. Acquaviva, P. Ade, et al., Astrophys.J. **739**, 52 (2011), 1009.0866.

[107] M. Archidiacono, E. Calabrese, and A. Melchiorri, Phys.Rev. **D84**, 123008 (2011), 1109.2767.

[108] J. B. Dent, S. Dutta, and R. J. Scherrer, Phys.Lett. **B687**, 275 (2010), 0909.4128.

[109] C.-P. Ma and E. Bertschinger, Astrophys.J. **455**, 7 (1995), astro-ph/9506072.

[110] S. Weinberg, Phys.Rev. **D67**, 123504 (2003), astro-ph/0302326.

[111] M. Bucher, K. Moodley, and N. Turok, Phys.Rev. **D62**, 083508 (2000), astro-ph/9904231.

[112] K. A. Malik and D. Wands, JCAP **0502**, 007 (2005), astro-ph/0411703.

[113] C. Reichardt, P. Ade, J. Bock, J. R. Bond, J. Brevik, et al., Astrophys.J. **694**, 1200 (2009), 0801.1491.

[114] A. O'Hagan and J. J. Forster, *Kendall's advanced theory of statistics, volume 2B: Bayesian inference*, vol. 2 (Arnold, 2004).

[115] F. Debbasch and W. van Leeuwen, Physica A: Statistical Mechanics and its Applications **388**, 1818 (2009).

[116] S. Weinberg, *The Quantum theory of fields. Vol. 1: Foundations* (1995).

[117] J.-H. Yoon and C.-Y. Wong, Phys.Rev. **C61**, 044905 (2000), nucl-th/9908079.

[118] C. Eckart, Phys.Rev. **58**, 919 (1940).

[119] S. Weinberg, Phys.Rev. **D74**, 063517 (2006), astro-ph/0607076.

[120] G. Hinshaw et al. (WMAP), Astrophys.J.Suppl. **208**, 19 (2013), 1212.5226.

[121] P. Ade et al. (Planck Collaboration), Astron.Astrophys. **571**, A16 (2014), 1303.5076.

[122] A. R. Liddle and D. Lyth (2000).

[123] P. Ade et al. (Planck Collaboration) (2013), 1303.5084.

[124] G. Hinshaw et al. (WMAP Collaboration), Astrophys.J.Suppl. **148**, 135 (2003), astro-ph/0302217.

[125] M. Tegmark, A. de Oliveira-Costa, and A. Hamilton, Phys.Rev. **D68**, 123523 (2003), astro-ph/0302496.

[126] A. de Oliveira-Costa, M. Tegmark, M. Zaldarriaga, and A. Hamilton, Phys.Rev. **D69**, 063516 (2004), astro-ph/0307282.

[127] A. Berera, L.-Z. Fang, and G. Hinshaw, Phys.Rev. **D57**, 2207 (1998), astro-ph/9703020.

[128] C. J. Copi, D. Huterer, and G. D. Starkman, Phys.Rev. **D70**, 043515 (2004), astro-ph/0310511.

[129] D. J. Schwarz, G. D. Starkman, D. Huterer, and C. J. Copi, Phys.Rev.Lett. **93**, 221301 (2004), astro-ph/0403353.

[130] P. Bielewicz, H. K. Eriksen, A. Banday, K. Gorski, and P. Lilje, Astrophys.J. **635**, 750 (2005), astro-ph/0507186.

[131] C. Monteserin, R. B. Barreiro, P. Vielva, E. Martinez-Gonzalez, M. Hobson, et al., Mon.Not.Roy.Astron.Soc. **387**, 209 (2008), 0706.4289.

[132] F. K. Hansen, A. Banday, and K. Gorski, Mon.Not.Roy.Astron.Soc. **354**, 641 (2004), astro-ph/0404206.

[133] C.-G. Park, Mon.Not.Roy.Astron.Soc. **349**, 313 (2004), astro-ph/0307469.

[134] H. Eriksen, F. Hansen, A. Banday, K. Gorski, and P. Lilje, Astrophys.J. **605**, 14 (2004), astro-ph/0307507.

[135] Y. Akrami, Y. Fantaye, A. Shafieloo, H. Eriksen, F. Hansen, et al., Astrophys.J. **784**, L42 (2014), 1402.0870.

[136] P. Vielva, E. Martinez-Gonzalez, R. Barreiro, J. Sanz, and L. Cayon, Astrophys.J. **609**, 22 (2004), astro-ph/0310273.

[137] K. Land and J. Magueijo, Phys.Rev. **D72**, 101302 (2005), astro-ph/0507289.

[138] F. Hansen, A. Banday, K. Gorski, H. Eriksen, and P. Lilje, Astrophys.J. **704**, 1448 (2009), 0812.3795.

[139] P. Ade et al. (Planck Collaboration) (2013), 1303.5083.

[140] C. Bennett, R. Hill, G. Hinshaw, D. Larson, K. Smith, et al., Astrophys.J.Suppl. **192**, 17 (2011), 1001.4758.

[141] M. Quartin and A. Notari (2014), 1408.5792.

[142] J. Bobin, F. Sureau, J. L. Starck, A. Rassat, and P. Paykari, Astron.Astrophys. **563**, A105 (2014), 1401.6016.

[143] C. Armendariz-Picon, JCAP **1103**, 048 (2011), 1012.2849.

[144] M. A. Blanco, M. Flrez, and M. Bermejo, Journal of Molecular Structure: THEOCHEM **419**, 19 (1997).

[145] E. Wigner, Vieweg Verlag, Braunchweig (1931).

[146] C. J. Copi, D. Huterer, D. J. Schwarz, and G. D. Starkman, Adv.Astron. **2010**, 847541 (2010), 1004.5602.

[147] T. W. Anderson and D. A. Darling, Ann. Math. Statist. **23**, 193 (1952).

[148] M. Stephens, Journal of the American Statistical Association **69, 347**, 730 (1973).

[149] K. Pearson, G. B. Jeffery, and E. M. Elderton, Biometrika **21, 1/4**, 164 (1929).

[150] L. R. Abramo and T. S. Pereira, Adv.Astron. **2010**, 378203 (2010), 1002.3173.

[151] A. Berera, R. V. Buniy, and T. W. Kephart, JCAP **0410**, 016 (2004), hep-ph/0311233.

[152] K. Gorski, E. Hivon, A. Banday, B. Wandelt, F. Hansen, et al., Astrophys.J. **622**, 759 (2005), astro-ph/0409513.

[153] T. D. Lee and C. N. Yang, Phys. Rev. **104**, 254 (1956), URL http://link.aps.org/doi/10.1103/PhysRev.104.254.

[154] C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes, and R. P. Hudson, Phys. Rev. **105**, 1413 (1957), URL http://link.aps.org/doi/10.1103/PhysRev.105.1413.

[155] E. Kretschmann, Annalen der Physik **358**, 575 (1918).

[156] R. Brout and F. Englert, Phys.Rev. **141**, 12311232 (1966).

[157] B. S. DeWitt, Phys.Rev. **162**, 1239 (1967).

[158] L. Berezhiani and J. Khoury, JCAP **1402**, 003 (2014), 1309.4461.

[159] W. D. Goldberger, L. Hui, and A. Nicolis, Phys.Rev. **D87**, 103520 (2013), 1303.1193.

[160] K. Hinterbichler, L. Hui, and J. Khoury, JCAP **1401**, 039 (2014), 1304.5527.

[161] G. L. Pimentel, JHEP **1402**, 124 (2014), 1309.1793.

[162] L. Berezhiani, J. Khoury, and J. Wang, JCAP **1406**, 056 (2014), 1401.7991.

[163] L. Berezhiani and J. Khoury, JCAP **1409**, 018 (2014), 1406.2689.

[164] P. Creminelli, J. Norea, M. Simonović, and F. Vernizzi, JCAP **1312**, 025 (2013), 1309.3557.

[165] B. Horn, L. Hui, and X. Xiao, JCAP **1409**, 044 (2014), 1406.0842.

[166] P. Creminelli and M. Zaldarriaga, JCAP **0410**, 006 (2004), astro-ph/0407059.

[167] C. Cheung, A. L. Fitzpatrick, J. Kaplan, and L. Senatore, JCAP **0802**, 021 (2008), 0709.0295.

[168] L. Senatore and M. Zaldarriaga, JCAP **1208**, 001 (2012), 1203.6884.

[169] P. Creminelli, J. Norena, and M. Simonović, JCAP **1207**, 052 (2012), 1203.4595.

[170] K. Hinterbichler, L. Hui, and J. Khoury, JCAP **1208**, 017 (2012), 1203.6351.

[171] V. Assassi, D. Baumann, and D. Green, JCAP **1211**, 047 (2012), 1204.4207.

[172] H. Collins, R. Holman, and T. Vardanyan, JCAP **1412**, 007 (2014), 1405.0017.

[173] J. M. Bardeen, Phys.Rev. **D22**, 1882 (1980).

[174] B. S. DeWitt, Phys.Rev. **162**, 1195 (1967).

[175] L. Faddeev and V. Popov, Sov.Phys.Usp. **16**, 777 (1974).

[176] R. K. Unz, Nuovo Cim. **A92**, 397 (1986).

[177] S. Weinberg, Phys.Rev. **D72**, 043514 (2005), hep-th/0506236.

[178] E. A. Calzetta and B.-L. Hu, *Nonequilibrium quantum field theory*, vol. 10 (Cambridge University Press Cambridge, 2008).

[179] C. Armendariz-Picon and R. Penco, Phys.Rev. **D85**, 044052 (2012), 1108.6028.

Curriculum Vitae

Personal Data

Date of Birth	24 October 1988
Address	Department of Physics, Syracuse University
	Syracuse, NY 13244, USA
Phone	+1 315 744 7715
email	jtneelak@syr.edu

Education

2009 –	**PhD in Physics, Syracuse University, Syracuse, NY, USA**
2006 – 2009	BSc (Hons) in Physics, Chennai Mathematical Institute, Chennai, India.

Programming, Scripting Languages Known

Fortran, Mathematica, C++, Python, Bash

Awards, etc.

1. Outstanding Teaching Assistant Award 2013-14, Syracuse University
2. Levenstein Fellowship (Syracuse University), 2014

Talks, Schools and Conferences

1. Northeast Cosmology Workshop, *McGill University*, 30 September 30 - 2 October, 2011

2. East Coast Gravity Workshop, *Syracuse University*, 21 - 22 April, 2012

3. Cosmology and Fundamental Physics with Planck, *CERN, Geneva*, 17 - 28 June, 2013

4. Post-Planck Cosmology, *Les Houches Summer School, France*, 8 July - 2 August, 2013

5. Neighborhood Workshop on Astrophysics and Cosmology, *Penn State University*, 3 - 4 April, 2014.

6. Challenges in Modern Cosmology: Dark Matter and Dark Energy, *IIP, Natal, Brazil*, 5-9 May, 2014

7. COSMO-2014, KICP, Chicago, 25-29 Aug, 2014

8. ISCAP, Columbia University, 6 Nov, 2014