High-resolution CMB bispectrum estimator

Wu Hyun Sohn

Department of Applied Mathematics and Theoretical Physics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Trinity College   May 2022

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

<div align="right">

Wu Hyun Sohn
May 2022

</div>

Acknowledgements

If this thesis was to be shared between everyone who made it happen, my share would be about a page. I would like to use this page to thank them and apologise to those who were not mentioned.

I would first like to thank my supervisors, James Fergusson and Paul Shellard. James has been the most patient and considerate supervisor one can ever hope for and was always there to give me valuable advice whenever I needed it. I am lucky to have been introduced to cosmology by him and benefit from his scientific intuition on many subjects. I am grateful to Paul for his insightful comments throughout my studies and for running our department. I would like to thank Kacper Kornet for his technical support; he is the reason why my code could run at all. I also express my gratitude to Trinity College for creating such an inspiring environment to study and for providing financial support during my postgraduate studies.

I am immensely grateful to Philip Clarke for helping me throughout the ups and downs of my PhD as a close collaborator, office mate, badminton buddy and an awesome friend. DAMTP is the place I will miss forever thanks to Áron Kovács, Fred Alford, Owain Salter Fitz-Gibbon, Miren Radia, Roland Bittleston, Daniel Zhang, João Melo, Charles Board, Roxana and Mycroft Rosca-Mead, and many more amazing people who I have had the privilege to meet.

I am forever indebted to the members of the Donut who accompanied me throughout the pandemic. Chuyi Yang, Kaimyn Chapman, Paul Tao, and Zen Ng—thank you for making me feel at home during the wildest of times.

Last but not least, I am incredibly privileged to have a caring family. Mom, dad, and Hyojie, your bottomless love and boxes full of Korean food sent from five thousand miles away gave me the energy to keep moving forward. And grandpa, sorry I am a little late, but I hope you get to celebrate your grandson's PhD degree all the same from where you are.

Abstract

The Cosmic Microwave Background (CMB) is one of the most valuable probes of the universe we have today. Anisotropies present in the ancient light contain rich statistical information about the perturbations in the early universe and their subsequent evolution until now. The CMB bispectrum, the Fourier equivalent of the three-point correlation function, allows us to study weak non-Gaussian signatures of the primordial fluctuations. Primordial non-Gaussianity is a key prediction of many physically well-motivated inflation models, and measuring its shape and amplitude allows us to constrain them.

This thesis comprises two sections. In the first section, we present forecasts on primordial non-Gaussianity constraints from upcoming CMB surveys. We focus our attention on models favoured by the Planck analysis, where a sharp feature in either the inflationary potential or sound speed causes oscillations in the bispectrum. Using preliminary specifications, we find that the Simons Observatory will have up to a factor of 1.6 improvements over Planck, increased to 1.7-2.2 for the CMB Stage-4 experiment. Motivated by bright prospects, we developed a novel CMB bispectrum estimator suited for the resolution and sensitivity of future surveys.

We discuss our high-resolution bispectrum estimator in the second section. Our code, named CMB-BEst, utilises a set of general basis functions to accurately constrain a wide variety of models. Implementing such a flexible and precise estimator was a computationally challenging task. We detail our algorithm design, code optimisation and parallelisation for high-performance computing clusters, which made this challenging computation tractable. Validation tests, both for internal consistency and comparisons against conventional estimators, are provided together with a proof-of-concept application. We highlight how CMB-BEst can be used for both general and targeted analyses of previously unconstrained models.

Table of contents

Chapter 1

Introduction

The galaxies we observe today are in fact their images from tens of thousands of years ago. This is due to the speed of light being finite, so the more distant an object is from us, the older the image we observe. This raises the question "how far back can we see?"

The radio astronomers Arno Penzias and Robert Wilson were calibrating their 50-foot-long horn antenna in 1964 when they found a mysterious background noise [1]. The measurements were independent of time and location in the sky and persisted after the removal of various potential contaminants. After the theoretical work of Robert Dicke, Jim Peebles, and David Wilkinson was brought forward, Penzias and Wilson identified the noise as Cosmic Microwave Background (CMB) radiation: ancient light from the early universe reaching us after billions of years [1, 2]. The discovery provided us with one of the most valuable probes of the physical universe.

Our modern mathematical formulation of cosmology is built on Einstein's work on general relativity in 1915 [3]. Using his framework, Friedmann, Lemaître, Robertson, and Walker contributed to writing down the unique metric for a spatially homogeneous and isotropic universe [4–8]. It was then realised that this FLRW metric has no stable static solution, and the universe likely started from a Big Bang and has been continuously expanding since. This was supported by Edwin Hubble's measurements of Cepheid variables in 1929 [9] and confirmed by the aforementioned discovery of the CMB, which was one of the key predictions of the Big Bang cosmology.

The Big Bang theory by itself, however, could not explain the apparent uniformity of the CMB in the early universe, as well as the flatness of the universe and the lack of magnetic monopoles today. The theory of cosmic inflation emerged to address these problems; the universe undergoes a phase of accelerated expansion shortly after the Big Bang [10]. Inflation successfully resolved the problem and further provided a mechanism for generating the initial density fluctuations which grow into the structure in the universe we observe today.

These primordial fluctuations originate from quantum fluctuations of the inflaton which has a nearly scale-invariant spectrum. This was confirmed when the Cosmic Background Explorer (COBE) discovered small anisotropies within the CMB in 1992 consistent with the inflationary predictions [11].

What is widely accepted to be the standard model of modern cosmology, the ΛCDM model, appeared only in the late 1990s. In 1998, measurements of distant type Ia supernovae showed that the universe's expansion is accelerating [12] which implied the presence of dark energy (Λ). Various astrophysical observations including galaxy rotation curves suggested the existence of dark matter [13], and that most of the matter in the universe is dark matter. The cold dark matter (CDM) scenario is currently the most favoured. The ΛCDM model considers the primary constituents of the universe to be dark energy, cold dark matter, baryons and radiation. The ΛCDM model has been extremely successful in explaining modern cosmological observations. CMB measurements from the *Planck* satellite, in particular, show exceptional agreement with the model [14]. Planck has mapped the CMB anisotropy to remarkable precision and resolution. The Planck CMB data by itself has placed one of the most stringent bounds on the theoretical parameters of ΛCDM to date [14, 15].

How does the CMB contain so much information about the universe? The answer is twofold. The first reason is that CMB anisotropy is linearly related to the primordial perturbations from inflation. Statistical properties of these initial fluctuations can thus be deduced from analysing correlation functions of the CMB, which lets us constrain the early universe physics. The second reason is that the CMB tracks the expansion history of the universe as it travels from the last scattering surface to us today. CMB photons scatter with baryons before free-streaming all the way to us, which then experience both the growth of the universe and the gravitational potential of matter perturbations. These signatures are engraved in the CMB anisotropy spectrum.

The CMB anisotropy is observed to be nearly Gaussian distributed. Statistical characteristics of a Gaussian random field can be summarised entirely using two-point correlation functions, or their Fourier counterpart: power spectra. The CMB power spectra have been thoroughly studied to constrain various cosmological parameters [16–19]. Meanwhile, higher-order statistics such as three-point correlation functions (bispectra in Fourier space) allow us to test the validity of the Gaussian assumption. They probe the non-Gaussian statistics of the CMB which can arise from either primordial dynamics or late-time effects [20].

Primordial non-Gaussianity is a key statistic for studying the physics of the early universe [21, 22]. The theory of inflation has been successful in describing the observed data, but its exact mechanism is yet undetermined. Currently, there are numerous viable inflationary models with well-founded physical motivations. Non-Gaussian signatures of primordial

fluctuations are robust predictions of various models, and measuring their shape and amplitude allows us to constrain particular inflationary scenarios. CMB bispectrum analysis from Planck yielded the most precise measurements of primordial non-Gaussianity to date [23]. So far, no statistically significant amount of non-Gaussianity has been detected.

In the near future, we expect several new major CMB experiments. The Simons Observatory (SO) [24] is a ground-based experiment currently under construction in the Atacama Desert of Chile. SO is expected to measure both CMB temperature and polarisation to unprecedented precision, largely improved compared to Planck. The first light from SO is planned to be observed in early 2022. Many more next-generation experiments such as CMB Stage-4 (CMB-S4) [25, 26] and LiteBIRD [27] are due course, brightening prospects for CMB science. In particular, the upcoming measurements will allow us to further constrain primordial non-Gaussianity, providing discovery potential [28].

This thesis is organised as follows. In Chapter **1**, we review the standard formulation of cosmology. We discuss the evolution equations of a homogeneous and isotropic universe as well as the basic theory of inflation. Chapter **2** details the cosmological perturbation theory with a focus on CMB anisotropy. We summarise how to compute the CMB power spectrum from primordial perturbations. Next in Chapter **3**, we outline how the CMB bispectrum can be used to probe primordial non-Gaussianity. We provide an example scenario to demonstrate how to calculate the bispectrum from the inflationary Lagrangian using the in-in formalism. The theory of CMB bispectrum estimation is reviewed together with the details of current implementations.

Chapters 4 and 5 contain my original work based on research conducted myself under the supervision of James Fergusson. Chapter **4** contains the forecast for future CMB-S4 surveys on the primordial non-Gaussianity parameter f_{NL} which was published in [29]. SO experiment specifications and expected CMB-S4 setup were used to predict their improved constraints via Fisher information analysis. We focussed on models with oscillatory features, where the significant enhancement in polarisation sensitivity significantly enhances the constraining power [30].

Motivated by the positive prospects from forecasts detailed in Chapter 4, we worked on developing a high-resolution bispectrum estimation pipeline suitable for future surveys. Chapter **5** contains the formulation and development details of the developed program, as well as many consistency checks performed during the extensive verification process. We outline the advantages of the new pipeline compared to conventional methods. Some working examples, studied in collaboration with Philip Clarke and Paul Shellard, are presented

[31]. Lastly, Chapter **6** concludes the thesis by summarising and laying out plans for future research.

Chapters 1 and 2 are purely introductory and contain textbook materials for cosmology. The formalism and interpretations presented are largely based on [32] and publicly available lecture notes [33–36]. Chapter 3 consists of literature reviews on the topic with references in the text. All written texts, codes and plots in Chapters 4 and 5 are results of my own work unless explicitly specified otherwise.

1.1   The homogeneous universe

In this section we review the standard cosmological formulation for the homogeneous universe, neglecting any perturbations. What we derive here will serve as a background solution for the full perturbative result discussed in the next chapter. We assume general relativity to be an accurate theory of gravity for the relevant scales we consider.

1.1.1   Geometry

In general relativity, spacetime is represented by a 4-dimensional Lorentzian manifold equipped with a metric. Distance measure in curved spacetime is given by the metric tensor g;

$$ds^2 = g_{\mu\nu}\, dx^\mu dx^\nu, \tag{1.1}$$

where the Greek letters $\mu, \nu = 0, 1, 2, 3$ represent time (0) and spatial $(1, 2, 3)$ indices of local coordinates. Flat spacetime has metric $g_{\mu\nu} = \eta_{\mu\nu} = \text{diag}\{-1, 1, 1, 1\}$, also known as the Minkowski metric. Throughout this thesis, we adopt the sign convention $(-, +, +, +)$ and work in units where $c = 1$. Unless specified otherwise, the Einstein summation convention is assumed.

In curved spacetime, free particles follow a trajectory given by the geodesic equation;

$$\frac{d^2 x^\mu}{ds^2} + \Gamma^\mu_{\nu\rho} \frac{dx^\nu}{ds} \frac{dx^\rho}{ds} = 0, \tag{1.2}$$

with s an affine parameter parametrising the trajectory, and $\Gamma^\mu_{\nu\rho}$ the Christoffel symbol representing the metric connection. Its value is given in terms of the metric tensor by

$$\Gamma^\mu_{\nu\rho} = \frac{1}{2} g^{\mu\sigma} \left(\partial_\rho g_{\nu\sigma} + \partial_\nu g_{\rho\sigma} - \partial_\sigma g_{\nu\rho} \right). \tag{1.3}$$

Here and throughout this thesis, ∂_μ denotes the partial derivative with respect to local coordinate x^μ. Note $g^{\nu\sigma}$ is the inverse metric satisfying $g^{\mu\nu}g_{\nu\rho} = \delta^\nu_\rho$.

Defining the tangent vector as $U^\mu = dx^\mu/ds$, the equation can be rewritten in a covariant form given by

$$(\nabla_U U)^\mu = U^\nu \nabla_\nu U^\mu = 0. \tag{1.4}$$

Note that in terms of local coordinates, the covariant derivative of a vector field is defined as $\nabla_\nu U^\mu = \partial_\nu U^\mu + \Gamma^\mu_{\nu\rho} U^\rho$.

The distance between two geodesics that are initially parallel may change in curved spacetime. Such geometric information is encapsulated within the Riemann curvature tensor R^a_{bcd}. [1] From R^a_{bcd} we can evaluate the Ricci curvature tensor R_{ab}, the Ricci scalar R, and finally the Einstein tensor G_{ab}. They are defined as follows.

$$R^\mu_{\nu\rho\sigma} := \partial_\rho \Gamma^\mu_{\nu\sigma} - \partial_\sigma \Gamma^\mu_{\nu\rho} + \Gamma^\tau_{\nu\sigma}\Gamma^\mu_{\tau\rho} - \Gamma^\tau_{\nu\rho}\Gamma^\mu_{\tau\sigma}, \tag{1.5}$$

$$R_{\mu\nu} := R^\rho_{\mu\rho\nu}, \tag{1.6}$$

$$R := g^{\mu\nu} R_{\mu\nu}, \tag{1.7}$$

$$G_{\mu\nu} := R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R. \tag{1.8}$$

The Einstein tensor is symmetric, i.e. $G_{\mu\nu} = G_{\nu\mu}$. It is also important to note that its divergence vanishes; $\nabla^\mu G_{\mu\nu} = 0$, which can be proven using the contracted Bianchi identity.

1.1.2 The FLRW universe

On very large scales, our universe is observed to be uniform in space (homogeneous) and does not have a favoured direction (isotropic). The spatial part of the homogeneous and isotropic metric has constant curvature and can be categorised into three classes depending on its sign: spherical ($\mathbb{S}^3$), Euclidean ($\mathbb{E}^3$), and hyperbolic ($\mathbb{H}^3$). They are induced from embedding $\mathbb{R}^3$ into submanifolds of $\mathbb{R}^4$ equipped with the Euclidean metric, defined as $K|\mathbf{x}|^2 + u^2 = 1$. Here $K = 1, 0, -1$ for $\mathbb{S}^3$, $\mathbb{E}^3$, and $\mathbb{H}^3$, respectively. Writing the embedding as $f : x^i = (x,y,z) \mapsto X^I = (x,y,z, \sqrt{1 - K(x^2 + y^2 + z^2)})$, the induced metric is

$$\gamma_{ij} := \frac{\partial X^I}{\partial x^i} \frac{\partial X^J}{\partial x^j} \delta_{IJ} = \delta_{ij} + \frac{x_i x_j}{1 - K x_k x^k}. \tag{1.9}$$

---

[1] Consider a 1-parameter family of geodesics $\gamma(s,t)$, where t is an affine parameter. The geodesic deviation equation states $T^\rho \nabla_\rho (T^\nu \nabla_\nu S^\mu) = R^\mu_{\nu\rho\sigma} T^\nu T^\rho S^\sigma$, where tangent vectors $T = \partial/\partial t$, $S = \partial/\partial s$.

The spatial line element is therefore given by

$$dl^2 = \gamma_{ij}dx^i dx^j = d\mathbf{x} \cdot d\mathbf{x} + \frac{K(\mathbf{x} \cdot d\mathbf{x})^2}{1 - K(\mathbf{x} \cdot \mathbf{x})} \tag{1.10}$$

$$= \frac{1}{1 - Kr^2}dr^2 + r^2 d\Omega^2, \tag{1.11}$$

where the angular line element $d\Omega^2 = d\theta^2 + \sin^2\theta d\phi^2$.

We may now write down the form of the metric describing our universe on large scales;

$$ds^2 = -dt^2 + a(t)^2 \left(\frac{1}{1 - Kr^2}dr^2 + r^2 d\Omega^2 \right). \tag{1.12}$$

This is known as the FLRW metric, named after independent researchers who worked on the topic. The function $a(t)$ is called the scale factor and it dictates the growth of the universe over time. Note that the metric is invariant under rescaling $a \to \lambda a$, $r \to r/\lambda$, and $K \to k := \lambda^2 K$. Hence we may set the scale factor to be $a(t_0) = 1$ at present time, at the cost of replacing $K \in \{-1, 0, 1\}$ by $k \in \mathbb{R}$.

The Levi-Civita connection corresponding to the FLRW metric can be computed using the definition (1.3). Its non-zero components are given in terms of γ (1.9) as follows.

$$\Gamma^0_{ij} = \frac{\dot{a}}{a}\gamma_{ij}, \tag{1.13}$$

$$\Gamma^i_{j0} = \Gamma^i_{0j} = \frac{\dot{a}}{a}\delta^i_j, \tag{1.14}$$

$$\Gamma^i_{jk} = \frac{1}{2a^2}\gamma^{il}\left(\partial_k \gamma_{jl} + \partial_j \gamma_{kl} - \partial_l \gamma_{jk}\right). \tag{1.15}$$

The overdot denotes time derivative $(\dot{}) := \partial/\partial t$ here, and in the remainder of this thesis. Indices for γ are raised and lowered using γ, not g.

Note that a path defined by $t(\tau) = \tau$ and $\mathbf{x}(\tau) = const$ is a timelike geodesic satisfying the geodesic equations (1.2). *Comoving* observers who follow these paths continue to perceive the expanding universe to be isotropic. Meanwhile, they find that they drift apart, as the physical distance $r_{phys} = a(t)r$ grows in time.

The Ricci curvature and Einstein tensor of the FLRW metric follows from definitions (1.5-1.8);

$$R_{00} = -\frac{3\ddot{a}}{a} \tag{1.16}$$

$$R_{ij} = \left[\frac{\ddot{a}}{a} + 2\left(\frac{\dot{a}}{a}\right)^2 + \frac{2k}{a^2}\right]a^2\gamma_{ij}, \tag{1.17}$$

$$R = 6\left[\frac{\ddot{a}}{a} + \left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2}\right], \tag{1.18}$$

$$G_{00} = 3\left[\left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2}\right], \tag{1.19}$$

$$G_{ij} = \left[-\frac{2\ddot{a}}{a} - \left(\frac{\dot{a}}{a}\right)^2 - \frac{k}{a^2}\right]a^2\gamma_{ij}. \tag{1.20}$$

While deriving (1.17) we used the fact that the Ricci tensor of three-dimensional spatial metric γ is equal to $2k\gamma_{ij}$. [2] Also note that components G_{0i} vanish and $G_{ij} \propto g_{ij}$, which is expected for a spatially homogeneous and isotropic spacetime.

1.1.3 Cosmic inventory

According to general relativity, spacetime is curved by its contents. Particles interact with gravity through the energy-momentum tensor $T_{\mu\nu}$, which encapsulates their energy, momentum flux, and stress. Thanks to spatial homogeneity and isotropy, components of the homogenous universe can be modelled as *perfect* fluids; they are completely characterised by their rest frame energy density and isotropic pressure. Defining the 4-velocity to be $U^\mu = dx^\mu/ds$ [3], the energy-momentum tensor of a perfect fluid is given by

$$T_{\mu\nu} = (\rho + P)U_\mu U_\nu + Pg_{\mu\nu}, \tag{1.21}$$

where the energy density ρ and pressure P only depends on time. For an observer comoving with the fluid, $T = \text{diag}\{\rho, P, P, P\}$.

---

[2]In general, the Ricci tensor of any *n*-dimensional constant-curvature space with metric g_{ij} is given by $R_{ij} = (n-1)\kappa g_{ij}$. Here κ denotes sectional curvature of the space, which is equal to k for γ defined in (1.9).

[3]Here s is an affine parametrisation of geodesic followed by the fluid. It is equal to proper time τ for massive particles geodesics, and $g_{\mu\nu}U^\mu U^\nu = -1$. Massless particles such as photons follow null trajectory, and s is chosen so that $g_{\mu\nu}U^\mu U^\nu = 0$.

The energy-momentum tensor is conserved by construction and $\nabla_\nu T^{\mu\nu} = 0$. We can check that this is indeed the case for a perfect fluid;

$$\nabla_\nu T^{\mu\nu} = (\rho + P)\nabla_\nu (U^\mu U^\nu) + P\nabla_\mu g^{\mu\nu} \tag{1.22}$$

$$= (\rho + P)(U^\nu \nabla_\nu U^\mu + U^\mu \nabla_\nu U^\nu) + \nabla_\mu g^{\mu\nu} \tag{1.23}$$

$$= 0. \tag{1.24}$$

The first term in (1.23) vanishes because the fluid's 4-velocity satisfies the geodesic equation (1.4). Incompressibility implies that $\nabla_\nu U^\nu = 0$, so the second term also vanishes. The last term is zero since the Levi-Civita connections are *metric*; i.e., $\nabla g = 0$.

Using the connections of the FLRW metric computed in the previous section, the $\mu = 0$ component of (1.24) yields the continuity equation;

$$\dot{\rho} + \frac{3\dot{a}}{a}(\rho + P) = 0. \tag{1.25}$$

Further imposing a constant equation of state $w = P/\rho$,

$$\frac{\dot{\rho}}{\rho} + 3(1+w)\frac{\dot{a}}{a} = 0, \tag{1.26}$$

$$\rho \propto a^{-3(1+w)}. \tag{1.27}$$

The universe contains a number of different components, but all known particles can be broadly categorised into three: radiation, matter, and dark energy.

Radiation consists of relativistic particles: photons and neutrinos. The energy-momentum tensor of radiation is traceless, fixing the equation of state to be $w = 1/3$. While the number density of photons decrease as $\propto a^{-3}$, their energy density scales as $\propto a^{-4}$ instead because each individual photon loses energy as the universe expands. We define the *redshift z* to quantify this effect;

$$1 + z := \frac{1}{a}. \tag{1.28}$$

Hence, a photon with original wavelength λ_0 gets redshifted by $\Delta\lambda = z\lambda_0$. Note that the redshift is directly related to the scale factor $a(t)$. It can be used to parametrise time at and distance to a light source in a model-independent way. Neutrinos show similar behaviours to photons since they remain ultra-relativistic.

Matter includes cold dark matter, electrons and protons. The latter two are often grouped as baryons, even though electrons are not technically baryonic. Pressure from non-relativistic

matter is negligible, and $w = 0$. Their number density scales $\propto a^{-3}$ as the universe expands, and so does their energy density. Cold dark matter constitutes about 84% of the matter and roughly a quarter of the total energy density today.

Dark energy is perhaps the most mysterious of the three, despite having the largest contribution to the total energy density at present time. The first evidence of its existence came from Type Ia supernovae measurements which implied that the universe's expansion is accelerating. Subsequent observations of CMB and baryonic acoustic oscillations provided further proof. The exact physical mechanism for dark energy is not yet known, but potential explanations include the cosmological constant and quintessence [37, 38]. For purposes of the ΛCDM model, dark energy has negative pressure ($w = -1$), hence its energy density is independent of the scale factor.

Table 1.1 summarises the species of the universe and their properties discussed above. The last column contains the fractional energy density of particles and shows how abundant each species are at present time. A precise definition of fractional density will appear in the next section.

Table 1.1 Cosmic inventory. The fractional density values are quoted from Planck CMB analysis [14].

	Examples	Equation of state	Density growth	Fractional density today
Radiation (r)	Photon (γ)	$w = 1/3$	$\rho \propto a^{-4}$	$\Omega_\gamma \approx 1 \times 10^{-4}$
	Neutrino (ν)			$\Omega_\nu < 2 \times 10^{-2}$
Matter (m)	Cold dark matter (c)	$w = 0$	$\rho \propto a^{-3}$	$\Omega_c \approx 0.27$
	Baryon (b)			$\Omega_b \approx 0.05$
Dark Energy (Λ)		$w = -1$	$\rho = \text{const}$	$\Omega_\Lambda \approx 0.68$

1.1.4 Evolution of the universe

We are now ready to calculate the time evolution of the homogeneous universe. The Einstein field equation of general relativity reads

$$G_{\mu\nu} = 8\pi G T_{\mu\nu}, \tag{1.29}$$

where G is the Newtonian constant of gravitation. [4] Substituting in the Einstein tensor for the FLRW metric from (1.19-1.20) and the energy-momentum tensor from (1.21), we obtain

$$3 \left[\left(\frac{\dot{a}}{a} \right) + \frac{k}{a^2} \right] = 8\pi G \rho, \tag{1.30}$$

$$-\frac{2\ddot{a}}{a} - \left(\frac{\dot{a}}{a} \right)^2 - \frac{k}{a^2} = 8\pi G P. \tag{1.31}$$

Rearranging above yields the Friedmann equations;

$$\left(\frac{\dot{a}}{a} \right)^2 = \frac{8\pi G}{3} \rho - \frac{k}{a^2}, \tag{1.32}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3} (\rho + 3P). \tag{1.33}$$

Note that the continuity equation (1.25) can be obtained from (1.33) and the time derivative of (1.32).

The Hubble parameter is defined as $H := \dot{a}/a$. From (1.32) we may compute the critical energy density for which the curvature k vanishes;

$$\rho_{\text{crit},0} := \frac{3H_0^2}{8\pi G}. \tag{1.34}$$

Subscripts 0 indicate that they are evaluated at present time $t = t_0$, where $a(t_0) = 1$.

In reality, the energy density and pressure appearing in the Friedmann equations are sums of contributions from different fluid components. The fractional density of a given fluid X is defined as

$$\Omega_X := \frac{\rho_X}{\rho_{\text{crit},0}}. \tag{1.35}$$

In the previous section we derived how each fluid's energy density depends on the scale factor. Quoting results summarised in Table 1.1, the Friedmann equations can be rewritten as follows;

$$\dot{a}^2 = H_0^2 \left(\frac{\Omega_{r,0}}{a^2} + \frac{\Omega_{m,0}}{a} + \Omega_{\Lambda,0} \, a^2 \right) - k, \tag{1.36}$$

$$\ddot{a} = H_0^2 \left(-\frac{\Omega_{r,0}}{a^3} - \frac{\Omega_{m,0}}{2a^2} + \Omega_{\Lambda,0} \, a \right). \tag{1.37}$$

---

[4]Constant G is not to be confused with the Einstein tensor $G_{\mu\nu}$ on the left-hand side.

These equations dictate the growth (or shrinking) of the universe given its curvature and energy density composition today. For $k \leq 0$, the right-hand side of (1.36) is always positive regardless of fractional density value. In this case, the fact that the universe is currently expanding suffices to show that scale factor a has been increasing monotonically. The universe began with the Big Bang at $a = 0$.

When $k > 0$, $\dot{a}$ vanishes at one or two values of a. There are multiple scenarios in this case, including Einstein's Static Universe (ESU) where $\dot{a} = \ddot{a} = 0$. The ESU is however unstable; perturbing around the static solution as $a(t) = a_{\mathrm{ESU}}(1 + \xi(t))$ in (1.37) gives $\ddot{\xi} > 0$ to leading order, which implies that there exists a growing solution for ξ. In fact, there are no stable static solutions to the Friedmann equations, rendering such models implausible. Another possibility is a closed universe without dark energy ($\Omega_\Lambda = 0$), where the scale factor grows until it hits the maximum and then decreases. This scenario requires $\ddot{a} < 0$ at all times. Measurements of Type Ia supernovae strongly suggest that the universe is currently in an accelerating phase, ruling out this option as well. Lastly, there is the bouncing universe model. The scale factor starts large, drops to a minimum value, and bounces back to accelerating growth. This model has been disregarded due to the need for the introduction of various new physics in the early universe but has recently regained popularity as an alternative to inflation [39].

Constraints from modern cosmological observations indicate that our universe is extremely flat, with $k \approx 0$. For the rest of this thesis, we set the curvature $k = 0$ and follow a standard Big Bang theory.

Note the different powers of the scale factor a are associated with each component of the universe in (1.36). According to the CMB measurements $\Omega_{r,0} \ll \Omega_{m,0} < \Omega_{\Lambda,0}$. The energy density is therefore dominated by a single component at a time, resulting in three different eras:

- Radiation domination (RD) where $0 < a < \Omega_{r,0}/\Omega_{m,0}$,

- Matter domination (MD) with $\Omega_{r,0}/\Omega_{m,0} < a < (\Omega_{m,0}/\Omega_{\Lambda,0})^{1/3}$, and

- Dark energy domination (ΛD) for $a > (\Omega_{m,0}/\Omega_{\Lambda,0})^{1/3}$.

When the universe consists mainly of a single fluid component X, we can simplify the Friedmann equations as follows;

$$\dot{a}^2 = H_0^2 \Omega_{X,0}\, a^{-1-3w_X}, \tag{1.38}$$

$$a \propto t^{\frac{2}{3(1+w_X)}}, \quad \text{if } w_X \neq -1. \tag{1.39}$$

Note that the scale factor grows exponentially for ΛD as $w_\Lambda = -1$; $a \propto \exp(Ht)$.

It is often convenient to consider *conformal* time defined as $\tau := \int_{t_i}^{t}(1/a(t'))dt'$, for some initial time reference t_i. As $dt = a d\tau$, the flat FLRW metric is given in terms of conformal time by

$$ds^2 = -dt^2 + a(t)^2 d\mathbf{x}^2 = a(t)^2 \left[-d\tau^2 + d\mathbf{x}^2 \right]. \tag{1.40}$$

Rewriting (1.38) with $da/d\tau = a\dot{a}$, we obtain

$$a \propto \tau^{\frac{2}{3w_X+1}}. \tag{1.41}$$

Calculations in this section are summarised in Table 1.2.

Table 1.2 Evolution of the universe. Ranges for the scale factor are computed from the cosmological parameters estimated in [14].

Era	Scale Factor	Growth (comoving)	Growth (conformal)
Radiation Domination	$a < 2.9 \times 10^{-4}$	$a \propto t^{1/2}$	$a \propto \tau$
Matter Domination	$2.9 \times 10^{-4} < a < 0.77$	$a \propto t^{2/3}$	$a \propto \tau^2$
Dark Energy Domination	$a > 0.77$	$a \propto e^{Ht}$	$a \propto -1/\tau$

1.2   Inflation

Soon after the quantitative formulation of Big Bang cosmology, several issues with the initial conditions were raised [10]. The CMB was observed to be nearly homogeneous, even though many parts of it should have been causally disconnected at the time. The curvature of the universe is extremely close to zero, while $k = 0$ is an unstable stationary point. The standard Big Bang cosmology provided little justification for such initial smoothness and fine-tuning.

The theory of cosmic inflation not only resolved most of these problems successfully but also provided a physical mechanism for the generation of inhomogeneities in the universe. Quantum fluctuations of the inflationary field seed the initial conditions, whose statistical properties are consistent with current observations. Inflation has therefore become the most widely accepted theory of the early universe to date.

In this section, we formulate the puzzles which led to the introduction of inflation (1.2.1) and outline the basic inflationary paradigm (1.2.2).

1.2.1   The horizon problem

According to relativity, information cannot travel faster than the speed of light. It is therefore possible to have two different points in spacetime that are causally disconnected; their lightcones do not intersect, so no events since the Big Bang could have affected both. Physical properties at such two points are independent of each other. The aim of this section is to compute the size of causally disjoint regions at the epoch of recombination when most of the CMB photons start free-streaming.

Consider a photon travelling in a straight line. Writing the radial part of $d\mathbf{x}^2$ as $d\chi^2$, our spacetime metric (1.40) becomes

$$ds^2 = a(\tau)^2(-d\tau^2 + d\chi^2). \tag{1.42}$$

Photons follow null geodesics, meaning $ds^2 = 0$ along their trajectories. Thus $\chi(\tau) = \tau + const$ or $\chi(\tau) = -\tau + const$. They appear to be straight and diagonal lines on the χ-τ plane, as shown in Figure 1.1. The distance light travels starting from some initial time τ_i to τ_f is then given by

$$\chi_{\text{PH}} := \tau_f - \tau_i = \int_{a_i}^{a_f} \frac{d\tau}{da} da = \int_{a_i}^{a_f} \frac{1}{a\dot{a}} da. \tag{1.43}$$

Here, χ_{PH} is called the particle horizon. No particles could have travelled further than this distance since the initial time τ_i.

Suppose that the universe is dominated by a single perfect fluid X in between a_i and a_f. The simplified Friedmann equation (1.38) then gives

$$\chi_{\text{PH}} = \int_{a_i}^{a_f} \frac{a^{(3w_X-1)/2}}{H_0\sqrt{\Omega_{X,0}}} da = \frac{2}{(3w_X+1)H_0\sqrt{\Omega_{X,0}}} \left(a_f^{(3w_X+1)/2} - a_i^{(3w_X+1)/2} \right). \tag{1.44}$$

Note that the particle horizon is bounded as $a_i \to 0$ if and only if $3w_X + 1 > 0$.

According to the conventional big bang cosmology, the universe begins at $a_i = 0$ with radiation contributing the most to energy density. The CMB last scattering surface lies around redshift $z \sim 1090$ (or $a_{\text{rec}} = 9.17 \times 10^{-4}$) in the matter domination era. Quoting cosmological parameters from tables 1.1 and 1.2, as well as treating radiation and matter domination separately in the integral, we get $\chi_{\text{CMB}} \approx 340$ Mpc.

The conformal distance to the last scattering surface χ_* can also be computed using the same formula (1.44), now integrating from a_{rec} to $a_0 = 1$. Approximating again by separating matter and dark energy domination era, $\chi_* \approx 15000$ Mpc. This is much larger than χ_{PH}!

As shown in Figure 1.1, CMB photons at two different opposite sides of the sky had no causal contact at all; their particle horizons have zero overlaps. Furthermore, $\chi_{\mathrm{CMB}}/\chi_* \approx 1.3$ degrees. Every disjoint 1.3 deg patch in the sky was causally unrelated at the time of recombination. There is no obvious reason for cosmological parameters in these patches to be similar. Despite this fact, the observed CMB is isotropic everywhere in temperature to order $O(10^{-5})$. This is the horizon problem; the universe at recombination is too homogeneous considering the small particle horizon then.
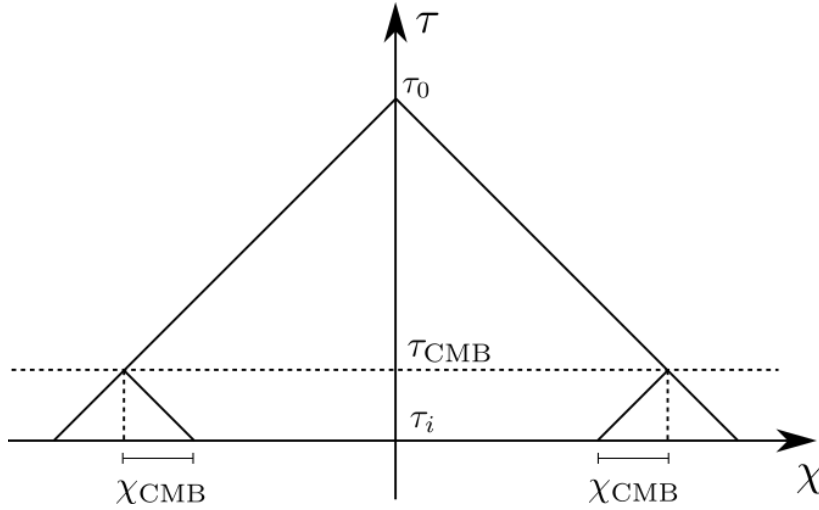

Fig. 1.1 The horizon problem. According to the conventional big bang cosmology, different regions of the CMB we observe today have had no overlap in their particle horizon. Yet, the CMB is measured close to uniform everywhere.

To resolve this issue, we need the particle horizon at recombination to be larger. The theory of cosmic inflation achieves this by having a period in the early universe where $3w + 1 < 0$. In this case, we see from (1.44) that χ_{PH} is unbounded as $a_i \to 0$. The initial conformal time $\tau_i \to -\infty$, allowing enough proper time for the particle horizon growth. Then even the two opposite regions of the CMB we see can have had causal contact in the past, as depicted in Figure 1.2.

Inflation can alternatively be characterised using the comoving Hubble radius defined as

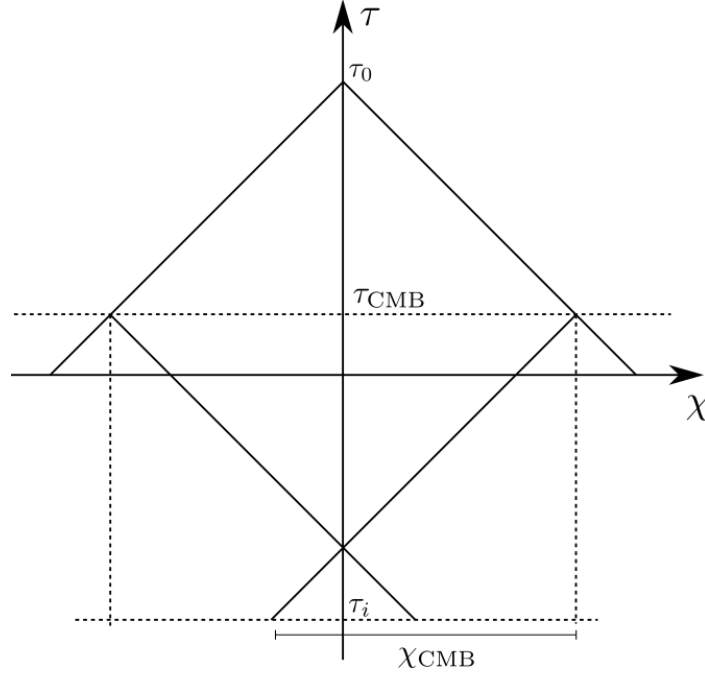$$\mathscr{H}^{-1} := \frac{1}{aH}. \tag{1.45}$$

Fig. 1.2 Solution to the horizon problem. Inflation allows more conformal time for different regions to have been in causal contact before recombination.

Note that the particle horizon can then be expressed in terms of the comoving hubble radius as

$$\chi_{\text{PH}} = \int_{a_i}^{a_f} \frac{1}{a\dot{a}} da = \int_{\ln a_i}^{\ln a_f} \mathcal{H}^{-1} d\ln a. \tag{1.46}$$

The particle horizon represents the distance where objects could have *ever* talked to each other. On the other hand, the comoving Hubble radius is a scale for how far information can reach *now*. [5] As can be seen from (1.44), $\mathcal{H}^{-1} \propto a^{(3w_X+1)/2}$, which is $\propto a$ for radiation domination and $\propto a^{1/2}$ during matter domination.

Inflation explains the homogeneity of the observed CMB by requiring $\mathcal{H}^{-1}$ to have shrunk rapidly in early times; $d\mathcal{H}^{-1}/d\ln a < 0$. We find that $\ddot{a} > 0$, so the universe undergoes accelerated expansion. The steep decrease in $\mathcal{H}^{-1}$ is parametrised using a 'slow-roll' parameter

$$\varepsilon := -\frac{d\ln H}{d\ln a} = -\frac{\dot{H}}{H^2} \ll 1. \tag{1.47}$$

---

[5] Two points $\mathcal{H}^{-1}$ apart drifts away with relative physical velocity $v_{\text{phys}} = \dot{a}\mathcal{H}^{-1} = 1$, which is equal to c in our units. It is difficult for such points to have causal interaction right now, especially within Hubble time H^{-1}.

Inflation also needs to last long enough for the particle horizon to grow sufficiently large. We define another parameter to denote this constraint;

$$\eta := \frac{d\ln\varepsilon}{d\ln a} = \frac{\dot{\varepsilon}}{\varepsilon H} \ll 1. \tag{1.48}$$

1.2.2 Slow-roll inflation

The simplest model of inflation consists of a single scalar field ϕ. The action for a real scalar field with canonical kinetic term and potential $V(\phi)$ is given by

$$S_\phi = \int dt d^3\mathbf{x} \sqrt{-g} \left[-\frac{1}{2} g^{\mu\nu} \partial_\mu \phi \partial_\nu \phi - V(\phi) \right], \tag{1.49}$$

where $g := \det g$. Denoting the integrand as the Lagrangian density $\mathscr{L}[t, \mathbf{x}, \phi, \partial_\mu \phi]$, the energy-momentum tensor can be expressed using functional derivatives as

$$T_{\mu\nu} = \frac{-2}{\sqrt{-g}} \frac{\delta \mathscr{L}}{\delta g^{\mu\nu}} = -2 \frac{\delta \mathscr{L}}{\delta g^{\mu\nu}} + g_{\mu\nu} \mathscr{L}. \tag{1.50}$$

Here, we used the identity $\delta\sqrt{-g} = -(1/2)\sqrt{-g}g_{\mu\nu}\delta g^{\mu\nu}$. Substituting (1.49),

$$T_{\mu\nu} = \left(\partial_\mu \phi\right)\left(\partial_\nu \phi\right) + g_{\mu\nu}\left(-\frac{1}{2}g^{\rho\sigma}\partial_\rho \phi \partial_\sigma \phi - V(\phi)\right). \tag{1.51}$$

Now suppose that ϕ drives inflation in the FLRW background. Due to the symmetries present in the homogeneous and isotropic metric, the inflation field can only depend on time; $\phi(\mathbf{x},t) = \bar{\phi}(t)$. We can read off the inflation field's energy density and pressure from the energy-momentum tensor.

$$\rho_{\bar{\phi}} = -T^0{}_0 = \frac{1}{2}\dot{\bar{\phi}}^2 + V(\bar{\phi}), \tag{1.52}$$

$$P_{\bar{\phi}}\delta^i{}_j = T^i{}_j = \delta^i{}_j \left(\frac{1}{2}\dot{\bar{\phi}}^2 - V(\bar{\phi})\right). \tag{1.53}$$

The two terms $\frac{1}{2}\dot{\bar{\phi}}^2$ and $V(\bar{\phi})$ can be interpreted as the kinetic and potential energy of the inflation field, respectively. The equation of state is also expressed in terms of the two;

$$w_{\bar{\phi}} = \frac{P_{\bar{\phi}}}{\rho_{\bar{\phi}}} = \frac{\frac{1}{2}\dot{\bar{\phi}}^2 + V(\bar{\phi})}{\frac{1}{2}\dot{\bar{\phi}}^2 - V(\bar{\phi})}. \tag{1.54}$$

It follows that $\bar{\phi}$ satisfies the condition $3w_{\bar{\phi}} + 1 < 0$ required for inflation, as long as the potential energy dominates over kinetic energy.

The classical equations of motion are obtained from the Euler-Lagrange equation. After some calculations we obtain

$$\ddot{\bar{\phi}} + 3H\dot{\bar{\phi}} = -V'(\bar{\phi}). \tag{1.55}$$

Even though $\bar{\phi}$ is a field, its dynamics given in (1.55) are identical to those of a particle rolling down a one-dimensional potential. Its movement should be slow for the kinetic energy to be much smaller than the potential, and hence the name 'slow-roll' inflation.

The Friedmann equations (1.32-1.33) can now be expressed in terms of the background inflationary field.

$$H^2 = \frac{8\pi G}{3}\left(\frac{1}{2}\dot{\bar{\phi}}^2 + V(\bar{\phi})\right) \approx \frac{8\pi G}{3}V(\bar{\phi}), \tag{1.56}$$

$$\frac{\ddot{a}}{a} = -\frac{8\pi G}{3}\left(\dot{\bar{\phi}}^2 - \frac{1}{2}V(\bar{\phi})\right) \approx \frac{8\pi G}{3}V(\bar{\phi}), \tag{1.57}$$

where the slow-roll approximations have been used in the last step. Taking the derivative of (1.56) gives us $3H\dot{\bar{\phi}} \approx -V'$. The parameters defined as

$$\varepsilon_V := \frac{1}{16\pi G}\left(\frac{V'}{V}\right)^2, \quad \eta_V := \frac{1}{8\pi G}\frac{V''}{V}, \tag{1.58}$$

are then shown to be small. Therefore the slow roll parameters from (1.47-1.48) satisfy

$$\varepsilon \approx \varepsilon_V \ll 1, \quad \eta \approx -2\eta_V + 4\varepsilon_V \ll 1. \tag{1.59}$$

We see that as long as the potential $V(\phi)$ is chosen such that $\varepsilon_V, \eta_V \ll 1$, the field ϕ can drive a period of accelerated expansion. Here, $\bar{\phi}(t)$ acts as a clock; it measures the progress of inflation until ε eventually grows comparable to 1 and inflation ends.

1.2.3 Quantum fluctuations

So far our considerations on the inflation field ϕ have been entirely classical. Moving on to quantum theory, field values are no longer fixed at each point in spacetime. The goal of this section is to quantify the statistical properties of these quantum fluctuations.

We assume that the metric remains unperturbed for simplicity. Gravity is coupled with perturbations of the inflation field in reality, but this approximation will still allow us to

derive most of the crucial results. We further neglect any terms suppressed by slow-roll parameters.

We write the inflation field as a sum of the classical solution and perturbation;

$$\phi(\mathbf{x}, t) = \bar{\phi}(t) + \frac{v(\mathbf{x}, t)}{a(t)}, \tag{1.60}$$

where the factor of $1/a$ has been introduced for later convenience. First rewriting the scalar field action (1.49) in terms of conformal time,

$$S_\phi = \int d\tau d^3\mathbf{x} a(\tau)^2 \left[\frac{1}{2}(\phi')^2 - \frac{1}{2}(\nabla\phi)^2 - a(\tau)^2 V(\phi) \right]. \tag{1.61}$$

When we include perturbations, terms linear in v vanish from the equations of motion of $\bar{\phi}$. Further removing terms with derivatives of $V(\phi)$ using the slow-roll condition,

$$\delta S_\phi = \int d\tau d^3\mathbf{x} \left[\frac{1}{2}\left(v' - \frac{a'}{a}v\right)^2 - \frac{1}{2}(\nabla v)^2 \right] \tag{1.62}$$

$$= \int d\tau d^3\mathbf{x} \left[\frac{1}{2}(v')^2 + \frac{1}{2}\frac{a''}{a}v^2 - \frac{1}{2}(\nabla v)^2 \right]. \tag{1.63}$$

Integration by parts has been used to obtain the last line. The equations of motion for v follows;

$$v'' - \frac{a''}{a}v - \nabla^2 v = 0. \tag{1.64}$$

Defining Fourier transforms of v as

$$v(\mathbf{x}, \tau) = \int \frac{d^3\mathbf{k}}{(2\pi)^3} e^{i\mathbf{k}\cdot\mathbf{x}} \tilde{v}(\mathbf{k}, \tau), \tag{1.65}$$

we obtain the Mukhanov-Sasaki equation;

$$v'' + (k^2 - \frac{a''}{a})v = 0. \tag{1.66}$$

Tildes above v have been omitted for brevity. Each $\mathbf{k}$ mode of the perturbative field $v(\mathbf{k}, \tau)$ evolves independently from each other. During slow-roll inflation $a \propto -1/\tau$, and $a''/a =$

$2/\tau^2$. The general form of the solution is given by the *mode functions* $v_k(\tau)$.

$$v_k(\tau) = c_+ \left(1 - \frac{i}{k\tau}\right) e^{-ik\tau} + c_- \left(1 + \frac{i}{k\tau}\right) e^{ik\tau}. \tag{1.67}$$

We would like to canonically quantise the field $v(\mathbf{k}, \tau)$. To achieve this goal, we first convert from the Lagrangian to Hamiltonian formalism.

$$\pi := \frac{\partial \mathscr{L}}{\partial v'} = v', \tag{1.68}$$

$$\mathscr{H} := v' \frac{\partial \mathscr{L}}{\partial v'} - \mathscr{L} = \frac{1}{2}\pi^2 + \frac{1}{2}(\nabla v)^2 - \frac{1}{2}\frac{a''}{a}v^2. \tag{1.69}$$

We now promote classical fields $v(\mathbf{k}, \tau)$, $\pi(\mathbf{k}, \tau)$ to operators $\hat{v}_{\mathbf{k}}(\tau)$, $\hat{\pi}_{\mathbf{k}}(\tau)$ satisfying equal-time commutation relations [6]

$$[\hat{v}_{\mathbf{k}_1}(\tau), \hat{\pi}_{\mathbf{k}_2}(\tau)] = (2\pi)^3 \delta^{(3)}(\mathbf{k}_1 + \mathbf{k}_2), \tag{1.70}$$

$$[\hat{v}_{\mathbf{k}_1}(\tau), \hat{v}_{\mathbf{k}_2}(\tau)] = [\hat{\pi}_{\mathbf{k}_1}(\tau), \hat{\pi}_{\mathbf{k}_2}(\tau)] = 0. \tag{1.71}$$

Defining operators $\hat{a}_{\mathbf{k}}$ and $\hat{a}_{\mathbf{k}}^\dagger$ appropriately, we may write

$$\hat{v}_{\mathbf{k}}(\tau) = v_k(\tau)\hat{a}_{\mathbf{k}} + v_k(\tau)^* \hat{a}_{-\mathbf{k}}^\dagger, \tag{1.72}$$

$$\hat{\pi}_{\mathbf{k}}(\tau) = v_k'(\tau)\hat{a}_{\mathbf{k}} + v_k'(\tau)^* \hat{a}_{-\mathbf{k}}^\dagger. \tag{1.73}$$

As long as we normalise the mode functions $v_k(\tau)$ so that its Wronskian $W := v_k v_k'^* - v_k^* v_k' = i$ (purely imaginary since v_k is real), we obtain

$$[\hat{v}_{\mathbf{k}_1}(\tau), \hat{\pi}_{\mathbf{k}_2}(\tau)] = i \left[\hat{a}_{\mathbf{k}_1}, \hat{a}_{\mathbf{k}_2}^\dagger\right], \tag{1.74}$$

$$\left[\hat{a}_{\mathbf{k}_1}, \hat{a}_{\mathbf{k}_2}^\dagger\right] = (2\pi)^3 \delta^{(3)}(\mathbf{k}_1 - \mathbf{k}_2). \tag{1.75}$$

---

[6]This follows from its Fourier equivalent: $[\hat{v}(\mathbf{x}_1, \tau), \hat{\pi}(\mathbf{x}_2, \tau)] = i\delta^{(3)}(\mathbf{x}_1 - \mathbf{x}_2)$.

The constructed $\hat{a}$ and $\hat{a}^\dagger$ are analogous to the creation and annihilation operators of a quantum harmonic oscillator. Our next step is to compute the Hamiltonian operator;

$$\hat{H} = \int d^3\mathbf{x} \left[\frac{1}{2}\hat{\pi}^2 + \frac{1}{2}(\nabla \hat{v})^2 - \frac{1}{2}\frac{a''}{a}\hat{v}^2 \right] \tag{1.76}$$

$$= \int d^3\mathbf{x} \frac{d^3\mathbf{k}_1}{(2\pi)^3}\frac{d^3\mathbf{k}_2}{(2\pi)^3}\frac{1}{2}e^{i(\mathbf{k}_1+\mathbf{k}_2)\cdot\mathbf{x}} \left[\hat{\pi}_{\mathbf{k}_1}\hat{\pi}_{\mathbf{k}_2} - (\mathbf{k_1}\cdot\mathbf{k_2})\hat{v}_{\mathbf{k}_1}\hat{v}_{\mathbf{k}_2} - \frac{a''}{a}\hat{v}_{\mathbf{k}_1}\hat{v}_{\mathbf{k}_2} \right] \tag{1.77}$$

$$= \int \frac{d^3\mathbf{k}}{(2\pi)^3}\frac{1}{2} \left[\hat{\pi}_{\mathbf{k}}\hat{\pi}_{-\mathbf{k}} + k^2\hat{v}_{\mathbf{k}}\hat{v}_{-\mathbf{k}} - \frac{a''}{a}\hat{v}_{\mathbf{k}}\hat{v}_{-\mathbf{k}} \right] \tag{1.78}$$

$$= \int \frac{d^3\mathbf{k}}{(2\pi)^3} \left[E_k \left(\hat{a}_{\mathbf{k}}\hat{a}_{\mathbf{k}}^\dagger + \hat{a}_{-\mathbf{k}}^\dagger\hat{a}_{-\mathbf{k}} \right) + F_k\hat{a}_{\mathbf{k}}\hat{a}_{-\mathbf{k}} + F_k^*\hat{a}_{\mathbf{k}}^\dagger\hat{a}_{-\mathbf{k}}^\dagger \right], \tag{1.79}$$

where

$$\omega_k^2 := k^2 - \frac{a''}{a}, \quad E_k := \frac{1}{2}(|v_k'|^2 + \omega_k^2|v_k|^2), \quad F_k := \frac{1}{2}(v_k'^2 + \omega_k^2 v_k^2). \tag{1.80}$$

Note that we are currently in the Heisenberg picture where the operators depend on time. Our mode functions have two degrees of freedom: c_+ and c_-. One of them has been fixed by the normalisation condition $W = 2ik(|c_+|^2 - |c_-|^2) = i$. The other degree of freedom remains, availing us a one-parameter family of possible initial mode functions. Note that $\hat{a}_{\mathbf{k}}^\dagger$ are defined in terms of $\hat{v}_{\mathbf{k}}$ and $\hat{\pi}_{\mathbf{k}}$. Fixing $\hat{a}_{\mathbf{k}}$ is thus equivalent to choosing $\hat{v}_{\mathbf{k}}$.

We define the vacuum $|0\rangle$ to be the state satisfying $\hat{a}_{\mathbf{k}}|0\rangle = 0$ for all $\mathbf{k} \in \mathbb{R}^3$. The expected energy of the vacuum state is given by

$$\langle 0|\hat{H}|0\rangle = \int \frac{d^3\mathbf{k}}{(2\pi)^3}E_k\langle 0|[\hat{a}_{\mathbf{k}}, \hat{a}_{\mathbf{k}}^\dagger]|0\rangle \tag{1.81}$$

$$= \int d^3\mathbf{k}E_k\delta^{(3)}(0), \tag{1.82}$$

where the divergence $\delta^{(3)}(0)$ arises only because we are integrating over the whole space. Removing this factor, we may interpret E_k as the vacuum energy density for mode k.

We now ask the vacuum state to be a ground state of the Hamiltonian. Minimising the energy density E_k while keeping normalisation condition $W = i$, we obtain the *Bunch-Davies* mode function;

$$v_{\mathbf{k}}(\tau) = \frac{1}{\sqrt{2k}}\left(1 - \frac{i}{k\tau}\right)e^{-ik\tau}. \tag{1.83}$$

For this choice of mode function and vacuum, $E_k \to \hbar \omega_k / 2$ as $\tau \to -\infty$. [7] This is analogous to the case of the quantum harmonic oscillator.

Lastly, we compute the zero-point fluctuation of the inflation field with respect to the vacuum. The two-point function follows from the commutation relation (1.75);

$$\langle \hat{v}_{\mathbf{k}_1}(\tau)\hat{v}_{\mathbf{k}_2}(\tau)\rangle = \langle 0|[\hat{a}_{\mathbf{k}_1}, \hat{a}^\dagger_{-\mathbf{k}_2}]|0\rangle \, v_{k_1}(\tau)v^*_{k_2}(\tau) \tag{1.84}$$

$$= (2\pi)^3 \delta^{(3)}(\mathbf{k}_1 + \mathbf{k}_2) \, P_v(k_1, \tau), \tag{1.85}$$

where we defined the *power spectrum* of v as $P_v(k, \tau) := |v_k(\tau)|^2$. Fourier transforming back using

$$\hat{v}(\mathbf{x}, \tau) = \int \frac{d^3\mathbf{k}}{(2\pi)^3} \left[v_k \hat{a}_\mathbf{k} + v^*_k \hat{a}^\dagger_\mathbf{k} \right] e^{i\mathbf{k}\cdot\mathbf{x}}, \tag{1.86}$$

we compute the local variation in v;

$$\langle |\hat{v}(\mathbf{x}, \tau)|^2 \rangle = \int \frac{d^3\mathbf{k}_1}{(2\pi)^3} \frac{d^3\mathbf{k}_2}{(2\pi)^3} \langle \mathbf{0}|v_{k_1}\hat{a}_{\mathbf{k}_1} v^*_{k_2}\hat{a}^\dagger_{\mathbf{k}_2}|\mathbf{0}\rangle \tag{1.87}$$

$$= \int \frac{d^3\mathbf{k}_1}{(2\pi)^3} \frac{d^3\mathbf{k}_2}{(2\pi)^3} v_{k_1} v^*_{k_2} \langle \mathbf{0}| \left[\hat{a}_{\mathbf{k}_1}, \hat{a}^\dagger_{\mathbf{k}_2} \right] |\mathbf{0}\rangle \tag{1.88}$$

$$= \int \frac{d^3\mathbf{k}}{(2\pi)^3} |v_k|^2 \tag{1.89}$$

$$= \int d(\ln k) \mathscr{P}_v(k, \tau). \tag{1.90}$$

The dimensionless power spectrum is defined as

$$\mathscr{P}_v(k, \tau) := \frac{k^3}{2\pi^2} P_v(k, \tau). \tag{1.91}$$

Recall that perturbations in ϕ is given by $\delta\phi = v/a$ from (1.60). The *dimensionless power spectrum* for fluctuations in ϕ is therefore

$$\mathscr{P}_{\delta\phi}(k, \tau) = \frac{\mathscr{P}_v(k, \tau)}{a(\tau)^2} = \left(\frac{H}{2\pi} \right)^2 \left(1 + (k\tau)^2 \right), \tag{1.92}$$

where we used the fact that $a(\tau) = -1/H\tau$ during slow-roll inflation. For scales larger than the comoving Hubble radius we have $k\tau \ll 1$. In this limit, $\mathscr{P}_{\delta\phi} \to (H/2\pi)^2$ which is nearly

---

[7] $\hbar$ has been reinstated here for clarity.

constant. This is a key prediction from our simplistic model of inflation; we expect a nearly scale-invariant power spectrum of the primordial (end-of-inflation) perturbations.

Summary

In this introductory chapter, we covered the basic foundations of modern cosmology. Under a simplifying assumption that the universe is homogeneous and isotropic, we wrote down the most general metric tensor in which the scale factor $a(t)$ parametrises the expansion of the universe. We then considered radiation, matter and dark energy as the main constituents of the universe and derived Friedmann equations, which can be solved to determine the universe's expansion history given its energy composition. The universe likely began with a Big Bang and has been continuously expanding since. This led to another puzzle, the horizon problem; the early universe observed through the CMB seems exceptionally uniform despite supposedly having many causally disconnected regions. We discussed how inflation solves the problem by introducing a period of accelerated expansion shortly after the Big Bang, dramatically increasing the size of the particle horizon. Inflation also successfully provides a mechanism for generating the initial density fluctuations in the universe required to create the structure we observe today. We showed that quantum fluctuations in the inflation field follow a nearly scale-invariant spectrum which is consistent with observations to date.

The universe observed today is not quite homogeneous and isotropic at all scales. It is the distribution of inhomogeneities—perturbations about the homogeneous background—that contains valuable information about the universe and provides crucial observables for precision cosmology. We formulate the cosmological perturbation theory that describes the evolution of these primordial perturbations in the following chapter in the context of the CMB.

Chapter 2

Cosmic Microwave Background Anisotropy

The early universe consisted of hot plasma. Photons were tightly coupled to baryons through Compton scattering and remained in thermal equilibrium with them. Since the photon energy density dropped more rapidly with the universe's expansion compared to that of matter, the universe eventually transitioned from radiation-dominated to matter-dominated. Its temperature also continued to drop until it was low enough for the electrons and protons to combine and form neutral hydrogen. Following this rather abrupt event at redshift ≈ 1100, called *recombination*, the photons no longer had electrons to scatter off from and instead started streaming freely through space. These photons are now observed by us today as the Cosmic Microwave Background (CMB).

The CMB follows a blackbody spectrum, since the radiation was in thermal equilibrium before it decoupled at recombination. Its temperature has been redshifted since then from approximately 3000K to $T_{\mathrm{CMB}} = 2.725$K, which we observe today. Looking closer into the map of CMB temperature, there are fluctuations of order 10^{-5} that were seeded by the primordial perturbations and evolved subsequently in time. Such anisotropy in the CMB therefore contains valuable information about the universe. Satellite surveys such as COBE [11], WMAP [40] and Planck [41], as well as numerous ground-based experiments including ACT [17], SPT [42], and BICEP/Keck array [43] have produced precise measurements of the CMB which have allowed us to constrain various cosmological parameters. The CMB anisotropy is the gold standard dataset for precision cosmology.

This chapter provides mathematical background and some physical intuitions for studying the CMB anisotropy, based on some standard materials on the topic including [32–34]. We start by taking a step forward from the homogeneous and isotropic universe discussed in the previous chapter and formulate the cosmological perturbation theory in Section 2.1.

Focussing on scalar perturbations we derive the linearised equations governing the time evolution of the perturbative variables in Section 2.2. In Section 2.3 we introduce the Boltzmann equation for photons and outline how it is solved numerically. We connect the primordial power spectrum to the angular power spectrum of the CMB anisotropy measured today.

2.1 The inhomogeneous universe

2.1.1 Metric perturbations

Recall that the FLRW metric is given using conformal time by

$$ds^2 = a(t)^2(-d\tau^2 + d\mathbf{x}^2). \tag{2.1}$$

We now perturb the metric as follows.

$$ds^2 = a(t)^2 \left[-(1+2A)dt^2 + 2B_i \, d\tau dx^i + (\delta_{ij} + h_{ij})dx^i dx^j \right]. \tag{2.2}$$

The spatial indices $i, j, \cdots$ here are lowered and raised using δ_{ij}. Note that the scale factor has not been perturbed, since any variation of it can be absorbed into other perturbative variables. There are 1, 3 and 6 degrees of freedom coming from A, B_i and h_{ij}, respectively, adding up to 10 as expected from a metric tensor of the 3+1 dimensional spacetime.

We further extract the divergence part from B_i and h_{ij}, as well as the trace of h_{ij}. Using V and T to denote vector and tensor quantities,

$$B_i = \partial_i B + B_i^V \tag{2.3}$$

$$h_{ij} = 2C\delta_{ij} + 2\partial_{\langle i}\partial_{j\rangle}E + (\partial_i E_j^V + \partial_j E_i^V). + 2E_{ij}^T, \tag{2.4}$$

where $\partial_{\langle i}\partial_{j\rangle} := \partial_i\partial_j - \frac{1}{3}\delta_{ij}\nabla^2$. The variables are chosen such that B_i^V is divergence-free, while h_{ij}^T is traceless and transverse.

The Scalar-Vector-Tensor (SVT) theorem states that to linear order in the perturbations, these modes decouple and evolve in three independent groups: scalars (A, B, C, E), vectors (B_i^V, E_i^V), and tensors (E_{ij}^T). These groups each contain 4, 4, and 2 degrees of freedom, again adding up to 10 as required. In this section, we are only interested in the scalar perturbations

responsible for generating the temperature anisotropy and E-mode polarisation observed in the CMB. [1] We will therefore set the vector and tensor modes of the perturbations to zero.

Keeping only the scalar modes, the perturbed metric becomes

$$ds^2 = a(\tau)^2 \left[-(1+2A)dt^2 + 2\partial_i B \, d\tau dx^i + \left((1+2C)\delta_{ij} + 2\partial_{\langle i}\partial_{j\rangle} E \right) dx^i dx^j \right]. \quad (2.5)$$

Coordinate transformations are *gauge* symmetries in General Relativity; they correspond to redundancies in our mathematical representation of the system. Redefining the coordinate variables may change how the theory looks, but all physical results derived in the new set are equivalent to the original theory. These redundant degrees of freedom may still appear to evolve in a non-trivial manner and therefore need to be treated properly. Here, we outline how to fix a gauge in the perturbed metric and use gauge-invariant quantities to connect results from different gauges.

Consider the coordinate transformations $x^\mu \rightarrow \tilde{x}^\mu = x^\mu + \xi^\mu$ for some small ξ. As long as ξ is of the same order as other perturbative variables, the background metric remains to be FLRW. We dcompose ξ as above, keeping only the scalar perturbations. The transformations can then be described by two arbitrary functions T and L so that $\xi^0 = T(\tau, \mathbf{x})$ and $\xi^i = \partial_i L(\tau, \mathbf{x})$. The transformation matrices take the form

$$\frac{d\tilde{x}^\alpha}{dx^\mu} = \begin{pmatrix} 1+T' & \partial_i T \\ \partial_i L' & \delta_{ij} + \partial_i \partial_j L \end{pmatrix}, \qquad \frac{dx^\mu}{d\tilde{x}^\alpha} = \begin{pmatrix} 1-T' & -\partial_i T \\ -\partial_i L' & \delta_{ij} - \partial_i \partial_j L \end{pmatrix}. \quad (2.6)$$

The metric tensor in the new set of coordinates can be found using the tensor transformation rule

$$g_{\mu\nu} \rightarrow \tilde{g}_{\alpha\beta} = \frac{dx^\mu}{d\tilde{x}^\alpha} \frac{dx^\nu}{d\tilde{x}^\beta} g_{\mu\nu}. \quad (2.7)$$

Note also that the scale factor also changes under metric transforms due to variations in its argument; $a(\tau)^2 \rightarrow a(\tilde{\tau})^2 = a(t)^2(1+2\mathcal{H}T)$ at first order in perturbations. After some calculations, we see that each perturbation variable in the metric (2.5) transforms like

$$\tilde{A} = A - T' - \mathcal{H}T, \qquad\qquad \tilde{B} = B + T - L', \quad (2.8)$$

$$\tilde{C} = C - \mathcal{H}T - \frac{1}{3}\nabla^2 L, \qquad \tilde{E} = E - L. \quad (2.9)$$

---

[1] Tensor modes also generate the temperature anisotropy and E-mode polarisation as well as B-mode polarisation, but their contributions are negligible compared to those of the scalar modes.

Hence, the two free functions T and E can be chosen in a way that the metric perturbations satisfy some desirable properties containing two degrees of freedom. Doing so *fixes* the gauge and no further redundancies remain in our formulation. Popular choices include the *spatially flat* gauge, where $C = E = 0$ and the spatial perturbations vanish as $h_{ij} = 0$, and the *synchronous* gauge, for which $A = B = 0$ and the time is left unperturbed.

Results from different gauge choices may appear dissimilar. In order to link them together, it is convenient to work with quantities constructed from the perturbative variables such that they remain invariant under the transformations (2.8)-(2.9). The Bardeen potentials are two such examples of gauge-invariant variables;

$$\Psi_B := A + \mathscr{H}(B - E') + B' - E'', \tag{2.10}$$

$$\Phi_B := -C - \mathscr{H}(B - E') + \frac{1}{3}\nabla^2 E. \tag{2.11}$$

The gauge we will be using for the rest of this chapter is the *Newtonian* gauge where $B = E = 0$. This allows us to identify A and C with the Bardeen potentials, and the metric takes the form (dropping the subscript)

$$ds^2 = a(\tau)^2 \left[-(1 + 2\Psi)d\tau^2 + (1 - 2\Phi)d\mathbf{x}^2 \right]. \tag{2.12}$$

The perturbations $\Psi(\tau, \mathbf{x})$ and $\Phi(\tau, \mathbf{x})$ are directly related to the classical gravitational potential in the Newtonian limit.

2.1.2 Matter perturbations

In order to complete the Einstein field equation, we need the energy-momentum tensor as well as the metric written to first order in perturbations. Recall that a perfect fluid in a homogeneous universe has $\bar{T}_{\mu\nu} = (\bar{\rho} + \bar{P})\bar{U}_\mu \bar{U}_\nu + \bar{P}\bar{g}_{\mu\nu}$. The perturbations in $T_{\mu\nu}$ come from two sources: a) directly from perturbations in the energy density and pressure, and b) indirectly through variations in the metric and comoving observer's four-velocity. We consider the local inertial frame to separate these two effects. Let

$$(E_0)^\mu := a^{-1}(1 - \Psi)\delta_0^\mu, \quad (E_i)^\mu := a^{-1}(1 + \Phi)\delta_i^\mu. \tag{2.13}$$

The four vectors above form an orthonormal frame where $g_{\mu\nu}(E_\alpha)^\mu(E_\beta)^\nu = \eta_{\alpha\beta}$. In this locally-Minkowski frame, each component of the tensor $\tilde{T}^{\alpha\beta}$ represent a physical property

of the fluid. We write

$$\tilde{T}^{00} = \rho = \bar{\rho} + \delta\rho, \qquad \tilde{T}^{0i} = q^i = (\bar{\rho} + \bar{P})v^i, \tag{2.14}$$

$$\tilde{T}^{ij} = P\delta^{ij} - \Pi^{ij} = (\bar{P} + \delta P)\delta^{ij} - \Pi^{ij}. \tag{2.15}$$

The momentum density **q** and the mean velocity **v** are first-order quantities which vanish in the absence of perturbations. The symmetric and traceless matrix Π^{ij} represents anisotropic stress. This quantity vanishes for non-relativistic fluids such as dark matter, while it remains small but non-zero for relativistic particles like photons.

Transforming back to the original coordinates through $T^{\mu\nu} = (E_\alpha)^\mu (E_\beta)^\nu \tilde{T}^{\alpha\beta}$, we get

$$T^{00} = a^{-2}\bar{\rho} \quad + a^{-2}(\delta\rho - 2\Psi\bar{\rho}), \tag{2.16}$$

$$T^{0i} = 0 \quad + a^{-2}q^i, \tag{2.17}$$

$$T^{ij} = a^{-2}\bar{P}\delta^{ij} + a^{-2}\left[(\delta P + 2\Phi\bar{P})\delta^{ij} - \Pi^{ij}\right], \tag{2.18}$$

at first order in perturbations. Note that when only the scalar modes are considered, we may write $v^i = \partial_i v$ and $\Pi^{ij} = \partial_{\langle i}\partial_{j\rangle}\Pi$ for some v and Π. If multiple fluids contribute, then the total energy-momentum tensor can be obtained by summing over the individual values for each constituent I; $\delta\rho = \sum_I \delta\rho_I$, $\delta P = \sum_I \delta P_I$, and $\delta\mathbf{q} = \sum_I \mathbf{q}_I$.

The energy-momentum tensor is also subject to the gauge transformations (2.6). Each perturbation variable appearing in $\delta T^{\mu\nu}$ transforms as

$$\tilde{\delta\rho} = \delta\rho - T\bar{\rho}', \qquad\qquad \tilde{\delta P} = \delta P - T\bar{P}', \tag{2.19}$$

$$\tilde{q}^i = q^i + (\bar{\rho} + \bar{P})L', \qquad\qquad \tilde{\Pi}^{ij} = \Pi^{ij}. \tag{2.20}$$

We may therefore choose to work in a gauge where some of the above vanish. For example, $\delta\rho = B = 0$ in *uniform density* gauge, and spatial slices of the fluid have constant energy density.

2.1.3 Initial perturbations

With the metric and matter perturbations written down, we only need one more ingredient to solve the linearised Einstein's equations: initial conditions. In inflationary ΛCDM, deviations from the background solution originate from the quantum fluctuations of the inflation field as we discussed in Section 1.2.3. Statistical properties of the perturbations' distribution depend on the details of the inflationary scenario.

In models where a single-field drives inflation, such as slow-roll inflation, the field can be thought of as a local *clock*. Classical equations of motions dictate the trajectory followed by the field, which then regulates the expansion of the universe. The spatial fluctuations of the inflation field can therefore be considered as the differences in how far the field has moved down the trajectory at each point in space. Any such perturbations can be determined solely from the variations in the local 'clock' time $\delta\tau(\tau,\mathbf{x})$;

$$\delta f = \bar{f}(\tau + \delta\tau, \mathbf{x}) - \bar{f}(\tau, \mathbf{x}) = \frac{\partial\bar{f}}{\partial\tau}\delta\tau, \tag{2.21}$$

for any f. By substituting $f = \rho_I$ and P_I we find that $\delta\rho_I/\bar{\rho}'_I = \delta P_I/\bar{P}'_I = \delta\tau$, which is locally constant for each constituent I. Combined with the background continuity equation (1.25),

$$\frac{\delta_I}{1 + w_I} = \frac{\delta_J}{1 + w_J}, \tag{2.22}$$

where we have defined the density contrast as $\delta_I := \delta\rho_I/\bar{\rho}_I$.[2] In particular, $\delta_r = (4/3)\delta_m$ at each point $\mathbf{x}$. Furthermore, for a fluid with a constant equation of state $w_I = \bar{P}_I/\bar{\rho}_I$, the *sound speed* c_s is defined as

$$c_s^2 := \frac{\delta P_I}{\delta\rho_I} = \frac{\bar{P}'_I}{\bar{\rho}'_I} = \frac{d\bar{P}_I}{d\bar{\rho}_I} = w_I. \tag{2.23}$$

Hence, the perturbed energy density and pressure also satisfy the equation of state; $P_I = w_I\rho_I$.

Perturbations satisfying the strong constraint of (2.22) are called to be *adiabatic*. They affect every component of the universe equally so that the ratio between density contrasts is uniform in space. Orthogonal to adiabatic modes are *isocurvature* modes where the total density contrast vanishes everywhere and $\delta_r = -\delta_m$ at the end of inflation. Some multi-field inflation models are expected to seed isocurvature perturbations. However, all observations so far are consistent with purely adiabatic initial conditions [15, 44].

2.2 Linearised equations

In this section, we derive equations governing the time evolution of the perturbations defined in the previous section. The equations are of first order in the perturbative variables. Although all calculations here are performed in the Newtonian gauge, the methodology illustrated here

---

[2]To avoid confusion with the differential operator and Dirac delta function, we always write the density contrast δ_I with a subscript I indicating which fluid it represents. The *total* density contrast will be denoted as δ_{tot}.

is general and applies to any choice of gauge. In order to avoid conflicting notations later, we use η instead of τ for conformal time in this section. [3]

2.2.1   Kinematics

Non-interacting perfect fluids such as cold dark matter are *conserved*; they should satisfy the perturbed metric's equivalent of the continuity equation (1.25), as well as the Euler equations regarding time evolution of the perturbative velocity field $\mathbf{v}(\eta, \mathbf{x})$. In this section, we derive the conservation equations for the perturbation theory using the constraint $\nabla_\nu T^\nu{}_\mu = 0$.

The first step is to compute the Christoffel symbols (1.3). Working in Newtonian gauge, they are given by

$$\Gamma^0_{00} = \mathscr{H} \quad + \Psi', \qquad\qquad\qquad \Gamma^0_{i0} = 0 \qquad + \partial_i \Psi', \qquad (2.24)$$

$$\Gamma^0_{ij} = \mathscr{H}\delta_{ij} - \left[\Phi' + 2\mathscr{H}(\Psi + \Phi)\right]\delta_{ij}, \qquad \Gamma^i_{00} = 0 \qquad + \partial_i\Psi, \qquad (2.25)$$

$$\Gamma^i_{jk} = 0 \qquad + (\partial_k\Phi)\delta_{ij} - (\partial_i\Phi)\delta_{jk} - (\partial_j\Phi)\delta_{ik}, \qquad \Gamma^i_{j0} = \mathscr{H}\delta_{ij} - \Phi'\delta_{ij}. \qquad (2.26)$$

The background values appear as the first term in each equation. Note that these look different to (1.13)-(1.15) because the conformal time η is used here instead of the comoving time t. Spatial indices are lowered and raised using the delta function above.

There are four components in the equation $\nabla_\nu T^\nu{}_\mu = \partial_\nu T^\nu{}_\mu + \Gamma^\nu_{\nu\gamma}T^\gamma{}_\mu - \Gamma^\gamma_{\nu\mu}T^\nu{}_\gamma = 0$. We take $\mu = 0$ and substitute in the perturbed energy-momentum tensor (2.16)-(2.18). After removing the background part $\bar{\rho}' = -3\mathscr{H}(\bar{\rho} + \bar{P})$, we obtain

$$(\delta\rho)' + \nabla \cdot \mathbf{q} + 3\mathscr{H}(\delta\rho + \delta P) - 3\Phi'(\bar{\rho} + \bar{P}) = 0. \qquad (2.27)$$

This is the continuity equation. Only the first two terms would be present if the spacetime were flat; any changes in the energy density perturbation $\delta\rho$ are then due to the fluid flow parametrised by the momentum density $\mathbf{q} = (\bar{\rho} + \bar{P})\mathbf{v}$. The third term accounts for the dilution of the fluid density from the expansion of the universe. The last term corrects for the perturbations in the expansion, where Φ' comes from the time derivative of the effective spatial scale factor $a(\eta)(1 - \Phi(\eta, \mathbf{x}))$. All factors of 3 here originate from having three spatial dimensions.

(2.27) applies to every fluid component which does not interact with others. Rewriting in terms of the equation of state $w_I = \bar{P}_I/\bar{\rho}_I$, sound speed $c_s^2 = \delta P/\delta\rho$, and density contrast

---

[3]Unfortunately, we could not have avoided the clash by using η to denote conformal time to begin with, since η is also commonly used as one of the slow-roll parameters in inflation.

$\delta_I = \delta \rho_I / \bar{\rho}_I$, we have

$$\delta_I' + (1 + w_I)\nabla \cdot \mathbf{v}_I + 3\mathcal{H}(c_s^2 - w_I)\delta_I - 3(1 + w_I)\Phi' = 0. \tag{2.28}$$

This is a first-order differential equation for δ_I.

Meanwhile, the $\mu = i$ part of $\nabla_\nu T^\nu{}_\mu = 0$ yields the Euler equations;

$$\mathbf{v}_I' + \mathcal{H}\left(1 - \frac{3\bar{P}'}{\bar{\rho}'}\right)\mathbf{v}_I + \frac{\nabla(\delta P)}{\bar{\rho} + \bar{P}} + \nabla\Psi = 0. \tag{2.29}$$

We do not have the zeroth order terms in this case since $\mathbf{v} = \mathbf{0}$ in the homogeneous universe. Although (2.29) is a vector equation, it has only one independent scalar component after rewriting $\mathbf{v} = \nabla v$ for some v. This is because we are only considering the scalar perturbations.

The physical meaning of the individual terms in (2.29) is as follows. The second term with a factor of $\mathcal{H}$ represents the dilution, or redshift, of the velocity due to the universe's expansion. The third relates to the pressure; the fluid's velocity accelerates in the direction of the steepest pressure gradient. Lastly, the gravitational force pulling towards a potential well is encapsulated in the final term.

If we assume adiabatic initial conditions where $\bar{P}'/\bar{\rho}' = \delta P/\delta \rho = c_s^2$, then the Euler equation further simplifies to

$$\mathbf{v}_I' + \mathcal{H}(1 - 3c_s^2)\mathbf{v}_I + \frac{c_s^2}{1 + w_I}\nabla\delta_I + \nabla\Psi = 0, \tag{2.30}$$

for a given fluid component I.

2.2.2 Dynamics

Our next step is to solve the linearised Einstein's equations. Using the Christoffel symbols (2.24)-(2.26), we may calculate the Einstein tensor for the perturbed metric. After a rather long but straightforward algebra,

$$G_{00} = 3\mathcal{H}^2 \qquad\qquad\qquad + 2\nabla^2\Phi - 6\mathcal{H}\Phi', \tag{2.31}$$

$$G_{0i} = 0 \qquad\qquad\qquad\qquad + 2\partial_i\Phi' + 2\mathcal{H}\partial_i\Psi, \tag{2.32}$$

$$G_{ij} = -(2\mathcal{H}' + \mathcal{H}^2)\delta_{ij} \quad -\partial_{\langle i}\partial_{j\rangle}(\Psi - \Phi) + \left[\frac{2}{3}\nabla^2(\Psi - \Phi) + 2\Phi''\right.$$
$$\left. + (4\mathcal{H}' + 2\mathcal{H}^2)(\Psi + \Phi) + 2\mathcal{H}\Psi' + 4\mathcal{H}\Phi'\right]\delta_{ij}, \tag{2.33}$$

where we have separated the zeroth and first order terms as before.

Einstein's equations, $G_{\mu\nu} = 8\pi G T_{\mu\nu}$, consists of 10 parts. Among these only 4 of them relate to scalar modes, similarly to how there are 4 independent scalar perturbations in the metric tensor (2.5). We obtain one equation for each of G_{00}, G_{i0}, the trace of G_{ij}, and lastly the trace-free part of G_{ij}.

We start with the trace-free component of G_{ij}. Combining (2.33) and (2.18) gives

$$-\partial_{\langle i}\partial_{j\rangle}(\Psi - \Phi) = \partial_{\langle i}\partial_{j\rangle}\Pi. \tag{2.34}$$

The anisotropic stress Π is negligible in reality. Non-relativistic matter does not contribute at all to Π, and photons induce non-zero but small anisotropic stress. Hence, we set $\Pi \approx 0$ and let $\Psi = \Phi$ for the rest of our derivations.

The 00, $i0$, and ij trace parts of Einstein's equations then take the form

$$\nabla^2\Phi - 3\mathscr{H}\Phi' - 3\mathscr{H}^2\Phi = 4\pi G\, a^2\delta\rho, \tag{2.35}$$

$$\Phi' + \mathscr{H}\Phi = -4\pi G\, a^2 q, \tag{2.36}$$

$$\Phi'' + 3\mathscr{H}\Phi' + (2\mathscr{H}' + \mathscr{H}^2)\Phi = 4\pi G\, a^2\delta P. \tag{2.37}$$

We used the background equations to remove terms of zeroth order in perturbations. The $i0$ part generally gives a vector identity but it only has one scalar degree of freedom since $\mathbf{q} = (\bar\rho + \bar P)\mathbf{v} = \nabla q$ for some $q := (\bar\rho + \bar P)v$.

Time derivatives of Φ on the left-hand side can be removed by putting (2.35) and (2.36) together. The result is Poisson's equation;

$$\nabla^2\Phi = 4\pi G\, a^2(\delta\rho - 3\mathscr{H}q). \tag{2.38}$$

This resembles Poisson's equation for Newtonian gravity: $\nabla^2\phi_{\text{Newt.}} = 4\pi G\rho$. We confirm our previous claim that Φ corresponds to the perturbations of the Newtonian gravitational potential. Note that (2.38) is best solved in Fourier space where $\nabla^2\Phi(\eta, \mathbf{x})$ reduces to $-k^2\tilde\Phi(\eta, \mathbf{k})$.

2.2.3 Curvature perturbations

The Bardeen potential Φ_B is an extremely useful quantity for studying cosmological perturbation theory. It is not only gauge-invariant but also representative of a physically meaningful quantity—gravitational potential—in Newtonian gauge which can be solved using the Pois-

son's equation (2.38). Gauge-invariant quantities also let us connect results from different gauge choices in a consistent manner.

Another gauge-invariant variable which plays a crucial role in cosmology is the constant density curvature perturbation, or simply *curvature perturbation*, defined as

$$\zeta := -C + \frac{1}{3}\nabla^2 E + \mathcal{H}\frac{\delta\rho}{\bar{\rho}'}. \tag{2.39}$$

The name originates from the fact that it closely relates to the spatial curvature $R^{(3)}$ in the uniform density gauge $B = \delta\rho = 0$. In Newtonian gauge,

$$\zeta = \Phi + \mathcal{H}\frac{\delta\rho}{\bar{\rho}'} = \Phi - \frac{\delta\rho}{3(\bar{\rho} + \bar{P})}, \tag{2.40}$$

where we used the background continuity equation $\bar{\rho}' = -3\mathcal{H}(\bar{\rho} + \bar{P})$ for the second equality. The time derivative of ζ can then be evaluated using (2.27):

$$\zeta' = \Phi' + \left[\frac{1}{3}\nabla \cdot \mathbf{v} + \mathcal{H}\frac{\delta\rho + \delta P}{\bar{\rho} + \bar{P}} - \Phi'\right] + \frac{\delta\rho(\bar{\rho}' + \bar{P}')}{3(\bar{\rho} + \bar{P})^2} \tag{2.41}$$

$$= \frac{1}{3}\nabla^2 v + \frac{\mathcal{H}}{\bar{\rho} + \bar{P}}\left(\delta P - \frac{\bar{P}'}{\bar{\rho}'}\delta\rho\right). \tag{2.42}$$

The second term in (2.42) vanishes for the adiabatic modes sourced by single-field inflation, as long as we assume the constant equations of state shown in (2.23). We are then left with $\zeta' = (1/3)\nabla^2 v$.

Now consider scales much larger than the comoving Hubble radius $\mathcal{H}^{-1}$, namely the 'super-horizon' scales. The Fourier space equivalent of this condition is $k \ll \mathcal{H}$. The Fourier transform of $\nabla^2 v$ equals $-k^2\tilde{v}$, which is much smaller than the typical time scale $\mathcal{H}$ in this limit. It follows that $\zeta' \approx 0$; curvature perturbations are conserved in super-horizon scales.

For this reason, the curvature perturbations play a major role in connecting inflation to the primordial initial conditions. Recall that the comoving Hubble radius $\mathcal{H}^{-1}$ shrinks during inflation due to exponential growth of the universe. Most physical scales of interest that were originally sub-horizon, or $k^{-1} \gg \mathcal{H}^{-1}$, eventually exit the horizon as $\mathcal{H}^{-1}$ drops below k^{-1}. After the end of inflation, $\mathcal{H}^{-1}$ grows back so that the modes can re-enter the horizon. Here, since ζ remains constant at super-horizon scales, the value of ζ at horizon re-entry is equal to that evaluated at the time the mode left the horizon. ζ serves as a bridge which links the quantum fluctuations generated during inflation to the initial perturbations after inflation.

2.3   CMB anisotropy

The linearised Einstein field equations (2.35)-(2.37) dictate the time evolution of the gravitational perturbations given the total energy, momentum, and pressure of the universe's constituents. The cold dark matter's contributions to these quantities can be solved using the perturbed continuity equation (2.27) and Euler equations (2.29). This section covers how perturbations in the photon temperature evolve and translate into the observed CMB anisotropy, the main cosmological dataset of our interest.

Studying photon perturbations is similar in principle to cold dark matter but involves two major complications. First, photons interact directly with baryons through Compton scattering. Their perturbations hence stay tightly coupled until recombination, when the electrons and protons combine to form neutral hydrogen and the photons start free-streaming instead. We outline in Section 2.3.1 how the Boltzmann equations are used to find the time evolution of photon perturbations while accounting for the scattering effects. Second, unlike cold dark matter whose perturbations are characterised by its density contrast $\delta \rho_m$ and velocity field $\mathbf{v}_m$, we require a whole hierarchy of functions to accurately describe photons. This is because the photons do not necessarily travel parallel to the wavevector when perturbed; $\mathbf{p} \nparallel \mathbf{k}$. We discuss the details in Section 2.3.2.

2.3.1   Boltzmann equations

Photons

We start by calculating the path a photon takes while free-streaming within the perturbed metric (2.12). The 4-momentum of a photon in local coordinates satisfies $\eta_{\alpha\beta}\tilde{P}^\alpha\tilde{P}^\beta = 0$ since photons are massless. We may hence write its components as

$$\tilde{P}^0 = p, \qquad \tilde{P}^i = p\hat{p}^i, \tag{2.43}$$

where $\hat{\mathbf{p}}$ is a unit vector indicating the direction of propagation and p is the magnitude of the 3-momentum. We can revert back to the perturbed metric using the tetrad in (2.13); since $P^\mu = (E_\alpha)^\mu \tilde{P}^\alpha$,

$$P^0 = a^{-1}(1 - \Psi)p, \qquad P^i = a^{-1}(1 + \Phi)p\hat{p}^i. \tag{2.44}$$

Recall that the 4-momentum is defined as $P^\mu = dx^\mu/\lambda$ for some affine parameter λ. Using our definition above,

$$\frac{dx^i}{d\eta} = \frac{dx^i}{d\lambda}\frac{d\lambda}{d\eta} = \frac{P^i}{P^0} = (1 + \Psi + \Phi)\hat{p}^i \tag{2.45}$$

to first order in perturbations. A comoving observer finds photons to be travelling more slowly while passing through overdense regions with $\Psi, \Phi < 0$.

The geodesic equation (1.2) takes the form

$$\frac{dP^\mu}{d\lambda} + \Gamma^\mu_{\nu\rho}P^\nu P^\rho = P^0\left(\frac{dP^0}{d\eta}\right) + \Gamma^0_{\nu\rho}P^\nu P^\rho = 0. \tag{2.46}$$

Meanwhile, differentiating (2.44) gives

$$\frac{dP^0}{d\eta} = \frac{1 - \Psi}{a}\left(\frac{dp}{d\eta} - \mathcal{H}p\right) - \frac{p}{a}\left(\frac{\partial\Psi}{\partial\eta} + \frac{dx^i}{d\eta}\frac{\partial\Psi}{\partial x^i}\right). \tag{2.47}$$

Note that the total derivative $d/d\eta$ here is taken along the trajectory $x^\mu = x^\mu(\eta)$ of the photon. We may combine (2.46) and (2.44) to obtain an expression for the total derivative of p;

$$\frac{1}{p}\left(\frac{dp}{d\eta}\right) = -\mathcal{H} + \Phi' - \hat{p}^i\partial_i\Psi. \tag{2.48}$$

Only the first term on the right-hand side is present at the zeroth order in perturbations, where we get $p \propto a^{-1}$. We confirm that photons are redshifted as the universe expands, justifying the definition of the redshift z in (1.28). The other two terms quantify the effect of gravitational perturbations on photon energy. Photons at a place where the potential is increasing over time *gain* energy $\propto \Phi'$. On the other hand, photons moving out of a potential well *lose* energy $\propto (\mathbf{p} \cdot \nabla)\Psi$. Dividing by p gives the last term in (2.48).

Distribution function

In order to fully understand the physical properties of photon perturbations, we need to study their distribution function $f(\eta, \mathbf{x}, \mathbf{p})$ which measures the number of particles in a unit phase space volume. Photons in a thermal equilibrium within the homogenous universe follow the Bose-Einstein distribution where

$$\bar{f}(\eta, p) = \left[\exp\left\{\frac{p}{\bar{T}(\eta)}\right\} - 1\right]^{-1}, \tag{2.49}$$

up to a factor and the Boltzmann constant, both of which we set to 1 for convenience. This is equivalent to the blackbody radiation with temperature $\bar{T}(\eta)$.

We define the fractional temperature anisotropy Θ by perturbing the photon distribution function as follows;

$$f(\eta,\mathbf{x},p,\hat{\mathbf{p}}) = \left[\exp\left\{\frac{p}{\bar{T}(\eta)\left[1+\Theta(\eta,\mathbf{x},\hat{\mathbf{p}})\right]}\right\} - 1\right]^{-1}. \tag{2.50}$$

We have made an implicit assumption here that $\Theta(\eta,\mathbf{x},\hat{\mathbf{p}})$ does not depend on p, the photon energy. This claim will be justified later as we show that the Thomson scattering between photons and electrons leaves the energy of photons virtually unchanged.

The distribution function (2.50) can be expanded to linear order in perturbations as

$$f = \bar{f} - \frac{\partial \bar{f}}{\partial(\ln p)}\,\Theta, \tag{2.51}$$

by replacing p with $p(1-\Theta)$ in (2.49). The temperature anisotropy Θ therefore closely relates to perturbations in the distribution function.

Collisionless equations

Liouville's theorem states that the phase space distribution function remains constant along the system's classical trajectories. A generalisation of this to systems with collisions is the Boltzmann equation. For our photon distribution function, it reads

$$\frac{df}{d\eta} = \frac{\partial f}{\partial \eta} + \frac{\partial f}{\partial x^i}\frac{dx^i}{d\eta} + \frac{\partial f}{\partial(\ln p)}\frac{d(\ln p)}{d\eta} + \frac{\partial f}{\partial \hat{p}^i}\frac{d\hat{p}^i}{d\eta} \tag{2.52}$$

$$= \left.\frac{df}{d\eta}\right|_{\text{scattering}}. \tag{2.53}$$

In (2.52) we used the chain rule to expand out the total derivative again. Note that the last term only appears at second order in perturbations since both $(\partial f/\partial \hat{p}^i)$ and $(d\hat{p}^i/d\eta)$ appear at first order. This term corresponding to gravitational lensing opens up a whole subject of its own but will be neglected for the purposes of this discussion.

Without scattering, the term in (2.53) vanishes and the distribution function is conserved along each trajectory. The zeroth order part of the equation yields

$$\frac{d\bar{f}}{d\eta} = \frac{\partial \bar{f}}{\partial \eta} - \frac{\partial \bar{f}}{\partial(\ln p)}\frac{d(\ln \bar{p})}{d\eta} = 0. \tag{2.54}$$

Substituting the form of $\bar{f}$ we see that the background temperature scales as $\bar{T} \propto a^{-1}$, consistent with the redshift caused by the expansion of the universe.

We can write down the first order part of (2.52) in terms of the temperature anisotropy Θ using our previous results (2.45), (2.48), and (2.51);

$$\frac{df}{d\eta} = -\frac{\partial}{\partial \eta}\left(\frac{\partial \bar{f}}{\partial (\ln p)}\Theta\right) - \frac{\partial \bar{f}}{\partial (\ln p)}\hat{p}^i\frac{\partial \Theta}{\partial x^i} + \frac{\partial^2 \bar{f}}{\partial (\ln p)^2}\mathscr{H}\Theta + \frac{\partial \bar{f}}{\partial (\ln p)}\left(\frac{\partial \Phi}{\partial \eta} - \hat{p}^i\frac{\partial \Psi}{\partial x^i}\right) \tag{2.55}$$

$$= -\frac{\partial \bar{f}}{\partial (\ln p)}\left[\Theta' + \hat{p}^i\partial_i\Theta - \Phi' + \hat{p}^i\partial_i\Psi\right], \tag{2.56}$$

after some lengthy algebra. There are four terms in (2.56). The first two naturally arise from the free-streaming of the photon and correspond to the temporal and spatial changes in Θ. The other two in brackets account for the changes in photon energy caused by the metric perturbations Φ and Ψ. The sum of the two is in fact equal to $-p^{-1}(d\ln(ap)/d\eta)$. The *comoving* energy ap of the photon remains constant in the homogeneous universe.

Thomson scattering

The early universe contained hot plasma in which photons bounce off charged particles via Compton scattering. In particular, photons and electrons interact through Thomson scattering $e^- + \gamma \leftrightarrow e^- + \gamma$. This interaction provides the dominant contribution to the Boltzmann equation for photon temperature anisotropies until recombination, when electrons and protons combine to form neutral hydrogen, letting photons travel freely.

Thomson scattering involves a photon and a non-relativistic electron with energy $E_e(\mathbf{q}) = m_e + \frac{1}{2m_e}q^2$, where m_e is the electron rest mass. In the electron's rest frame, the scattering term (2.53) takes the form

$$\left.\frac{df}{d\eta_e}\right|_{\text{scattering}} = \bar{n}_e \int d^2\hat{\mathbf{p}}_{\text{in}} \frac{d\sigma}{d\Omega}(\hat{\mathbf{p}}, \hat{\mathbf{p}}_{\text{in}}) \left[f(p, \hat{\mathbf{p}}_{\text{in}}) - f(p, \hat{\mathbf{p}})\right], \tag{2.57}$$

where η_e and $\bar{n}_e$ denote the proper time and (background) mean number density of the electrons, respectively. We have suppressed the η and $\mathbf{x}$ dependencies from the photon distribution function $f(\eta, \mathbf{x}, p, \hat{\mathbf{p}})$ for now. The integrand of (2.57) comprises two parts. The former corresponds to the photons originally travelling in the direction $\hat{\mathbf{p}}_{\text{in}}$ but then scattered *into* $\hat{\mathbf{p}}$. The latter is the opposite; it accounts for the photons scattering *out of* $\hat{\mathbf{p}}$ to some other direction, say $\hat{\mathbf{p}}_{\text{in}}$. Both *in* and *out* scatters are equally effective since Thomson scattering is reversible. The rate of interaction depends on the number densities $\bar{n}_e$ and f, as well as the

angle-dependent function given by

$$\frac{d\sigma}{d\Omega}(\hat{\mathbf{p}}, \hat{\mathbf{p}}_{\text{in}}) = \frac{3\sigma_T}{16\pi} \left[1 + (\hat{\mathbf{p}} \cdot \hat{\mathbf{p}}_{\text{in}})^2 \right] \tag{2.58}$$

for Thomson scattering [32]. The Thomson differential cross section σ_T is a fixed constant equal to $8\pi\hbar^2\alpha^2/3c^2m_e^2$. Note from (2.57) and (2.58) that Thomson scattering leaves the photon energy p unchanged. This justifies our previous assumption to drop the p dependence from Θ; the photon energy remains unperturbed by scattering in our linear perturbation theory.

We now account for the bulk velocity $\mathbf{v}_b$ of electrons (baryons). In the non-relativistic limit where $\|\mathbf{v}_b\| \ll 1$, the Doppler effect alters the distribution function by an amount proportional to $\bar{n}_e(\hat{\mathbf{p}} \cdot \mathbf{v}_e)$. We also rewrite $f = \bar{f} - (\partial \bar{f}/\partial \ln p)\,\Theta$ and substitute (2.58) into (2.57) to obtain

$$\left.\frac{df}{d\eta}\right|_{\text{scattering}} = a\bar{n}_e\sigma_T \frac{d\bar{f}}{d(\ln p)} \left[\Theta(\hat{\mathbf{p}}) - \frac{3}{16\pi} \int d^2\hat{\mathbf{n}}\, \left[\Theta(\hat{\mathbf{n}})\left(1 + (\hat{\mathbf{p}} \cdot \hat{\mathbf{n}})^2\right) \right] - \hat{\mathbf{p}} \cdot \mathbf{v}_b \right], \tag{2.59}$$

where we relabelled $\hat{\mathbf{p}}_{\text{in}}$ as $\hat{\mathbf{n}}$ and performed the $\int d^2\hat{\mathbf{n}}$ integral for the *out* term cancelling out the factor $3/16\pi$. We now have a complete form of the scattering term appearing in the Boltzmann equations.

Line of sight solution

We obtain the full photon Boltzmann equation by combining (2.56) and (2.59).

$$\Theta' + (\hat{\mathbf{p}} \cdot \nabla)\Theta - \Phi' + (\hat{\mathbf{p}} \cdot \nabla)\Psi$$
$$= a\bar{n}_e\sigma_T \left[-\Theta + \frac{3}{16\pi} \int d^2\hat{\mathbf{n}}\, \left[\Theta(\hat{\mathbf{n}})\left(1 + (\hat{\mathbf{p}} \cdot \hat{\mathbf{n}})^2\right) \right] + \hat{\mathbf{p}} \cdot \mathbf{v}_b \right]. \tag{2.60}$$

Note that every term on the right-hand side has a factor of $a\bar{n}_e\sigma_T$ from scattering, which motivates us to define the *optical depth*;

$$\tau(\eta) := \int_{\eta}^{\eta_0} d\eta' \left[a(\eta')\bar{n}_e(\eta')\sigma_T \right], \tag{2.61}$$

where τ is not to be confused with conformal time, also denoted τ elsewhere in other chapters. Here, $-\delta\tau = a\bar{n}_e\sigma_T\delta\eta$ corresponds to the probability of scattering to occur between some small time interval η and $\eta + \delta\eta$. The optical depth hence measures how *opaque* the universe has been for photons to travel without running into something until today. The

probability of a scatter between some time η and today η_0 is in fact equal to $e^{-\tau(\eta)}$. This is 1 for $\tau = 0$ and tends to 0 as $\tau \to \infty$. Taking the time derivative yields the *visibility function* $g(\eta) := -\tau'(\eta)e^{-\tau(\eta)}$, or the probability of having last scattered at η.

Using τ and η we may rewrite (2.60) as follows.

$$\frac{d}{d\eta}\left[e^{-\tau}(\Theta + \Psi)\right] = e^{-\tau}(\Psi' + \Phi') + g\left[\frac{3}{16\pi}\int d^2\hat{\mathbf{n}}\,\left[\Theta(\hat{\mathbf{n}})(1 + (\hat{\mathbf{p}}\cdot\hat{\mathbf{n}})^2)\right] + \hat{\mathbf{p}}\cdot\mathbf{v}_b\right] \quad (2.62)$$

$$=: S(\eta, \mathbf{x}, \hat{\mathbf{p}}), \quad (2.63)$$

where we defined the source term $S(\eta, \mathbf{x}, \hat{\mathbf{p}})$ as the right-hand side of (2.62). Integrating along the line of sight yields a formal solution to the Boltzmann equation of the form

$$\Theta(\eta_0, \mathbf{x}_0, \hat{\mathbf{p}}) + \Psi(\eta_0, \mathbf{x}_0) = \int_0^{\eta_0} d\eta'\, S(\eta', \mathbf{x}_0 - (\eta_0 - \eta')\hat{\mathbf{p}}, \hat{\mathbf{p}}), \quad (2.64)$$

since $\tau(\eta_0) = 0$ and $\tau(0) = \infty$.

We can simplify the problem further by assuming that all photons last scattered at a fixed $\eta = \eta_*$ and have started free-streaming since. The visibility function is then equal to $g(\eta) \approx g(\eta_*)\delta(\eta - \eta_*)$, and $e^{-\tau}$ switches immediately from 0 to 1 at recombination $\eta = \eta_*$. The integral term in (2.62) reduces to the *monopole* $\Theta_0 := (1/4\pi)\int d^2\hat{\mathbf{n}}\,\Theta(\hat{\mathbf{n}})$ as long as we neglect the anisotropic stress, which we will detail in the next section.

With these approximations,

$$\Theta(\eta_0, \mathbf{x}_0, \hat{\mathbf{p}}) + \Psi(\eta_0, \mathbf{x}_0)$$
$$\approx \Theta_0(\eta_*, \mathbf{x}_*) + \Psi(\eta_*, \mathbf{x}_*) + \hat{\mathbf{p}}\cdot\mathbf{v}_b(\eta_*, \mathbf{x}_*) + \int_{\eta_*}^{\eta_0} d\eta'\,(\Psi' + \Phi')(\eta', \mathbf{x}_0 - (\eta_0 - \eta')\hat{\mathbf{p}}),$$
$$(2.65)$$

Here, $\mathbf{x}_* = \mathbf{x}_0 - (\eta_0 - \eta_*)\hat{\mathbf{p}}$ is the comoving coordinate of the point within the last scattering surface where the given photon comes from.

There are several different contributions to Θ in (2.65). Firstly, the difference in the potential Ψ between the last scattering $(\eta_*, \mathbf{x}_*)$ and today $(\eta_0, \mathbf{x}_0)$ induces gravitational redshift for any photons climbing out of the potential well. Next, the monopole Θ_0 at recombination is closely related to the photon energy density contrast, as we will see shortly. The sum of Ψ and Θ_0 is often called the Sachs-Wolfe (SW) contribution and is dominant at large scales. The third term $\hat{\mathbf{p}}\cdot\mathbf{v}_b$ measures the Doppler effect caused by the electron's peculiar velocity at the last scattering. Lastly, the last integral accounts for the Integrated Sachs-Wolfe (ISW) effect, where time variations in the gravitational potentials affect the

photon energy. Accelerated expansion in the late universe due to dark energy contributes the most to this integral.

2.3.2 Boltzmann hierarchy

The full Boltzmann equation (2.60) is not in an ideal structure for us to solve numerically. It is not only in an integro-differential form, but also a function of three quantities η, $\mathbf{x}$ and $\hat{\mathbf{p}}$, totalling 6 dimensions. In this section, we outline the formulation that the Boltzmann solvers today such as CAMB [45] and CLASS [46] use to find the time evolution of the photon temperature anisotropy.

We work in the Fourier space of $\mathbf{x}$ so that $\tilde{\Theta} = \tilde{\Theta}(\eta, \mathbf{k}, \hat{\mathbf{p}})$. The term $(\hat{\mathbf{p}} \cdot \nabla)\Theta$ then simplifies to $(\hat{\mathbf{p}} \cdot i\,\mathbf{k})\tilde{\Theta} = ik\mu\tilde{\Theta}$, where $\mu := \hat{\mathbf{p}} \cdot \hat{\mathbf{k}}$ measures the angle between wavevector $\mathbf{k}$ and the direction of propagation $\hat{\mathbf{p}}$.

The next major simplification comes from the fact that even though $\Theta(\eta, \mathbf{k}, \hat{\mathbf{p}})$ depends on $\hat{\mathbf{p}}$, it is axisymmetric with respect to $\mathbf{k}$ for the scalar perturbations we are studying. The scattering term in (2.59) also preserves this axisymmetry. Thus, $\tilde{\Theta}(\eta, \mathbf{k}, \hat{\mathbf{p}}) = \tilde{\Theta}(\eta, \mathbf{k}, \mu)$, where we managed to cut down one dimension. This motivates us to define the lth *multipole* as [4]

$$\Theta_l(\eta, \mathbf{k}) := i^l \int \frac{d^2\hat{\mathbf{p}}}{4\pi} \, P_l(\hat{\mathbf{k}} \cdot \hat{\mathbf{p}}) \, \Theta(\eta, \mathbf{k}, \hat{\mathbf{p}}) \tag{2.66}$$

$$= i^l \int_{-1}^{1} \frac{d\mu}{2} \, P_l(\mu) \, \Theta(\eta, \mathbf{k}, \mu), \tag{2.67}$$

where we dropped the tildes on Θ representing Fourier variables for brevity. $P_l(\mu)$ denotes the lth order Legendre polynomial, constructed to satisfy the orthogonality condition

$$\int_{-1}^{1} \frac{d\mu}{2} \, P_{l_1}(\mu) P_{l_2}(\mu) = \frac{1}{2l_1 + 1}\delta_{l_1 l_2}. \tag{2.68}$$

Legendre polynomials form a complete basis for smooth functions defined on interval $[-1, 1]$. Hence, we may expand $\Theta(\eta, \mathbf{k}, \mu)$ in μ using these polynomials for each $(\eta, \mathbf{k})$. After substituting into (2.67), (2.68) yields

$$\Theta(\eta, \mathbf{k}, \mu) = \sum_{l=0}^{\infty} (-i)^l (2l + 1) P_l(\mu)\Theta_l(\eta, \mathbf{k}) \tag{2.69}$$

---

[4]Note that some literature adopt a different convention where a factor of $2l + 1$ is included in the definition of Θ_l.

The lowest multipole at $l = 0$ corresponds to the monopole we discussed earlier; $\Theta_0 = (1/4\pi) \int d^2\hat{\mathbf{n}}\, \Theta(\hat{\mathbf{p}})$, which is a simple angle average. Next is the dipole at $l = 1$ given by $\Theta_1 = (i/4\pi) \int d^2\hat{\mathbf{n}}\, (\hat{\mathbf{k}} \cdot \hat{\mathbf{p}})\Theta(\hat{\mathbf{p}})$, then the quadrupole at $l = 2$, and so on. The integrated term in (2.60) now takes a simpler form

$$\frac{3}{16\pi} \int d^2\hat{\mathbf{n}}\, \left[\Theta(\eta, \mathbf{k}, \hat{\mathbf{n}})(1 + (\hat{\mathbf{p}} \cdot \hat{\mathbf{n}})^2)\right] = \Theta_0(\eta, \mathbf{k}) + \frac{1}{2}P_2(\hat{\mathbf{k}} \cdot \hat{\mathbf{p}})\Theta_2(\eta, \mathbf{k}), \qquad (2.70)$$

after some calculations and noting that $P_2(\mu) = (1/2)(3\mu^2 - 1)$.

Incorporating the results so far, the Boltzmann equation becomes

$$\Theta'(\mu) + ik\mu\Theta(\mu) - \Phi' + ik\mu\Psi = -\tau'\left[-\Theta(\mu) + \Theta_0 + \frac{1}{2}P_2(\mu)\Theta_2 + ik\mu v_b\right], \qquad (2.71)$$

where $\mathbf{v}_b = \nabla v_b$, and the η and $\mathbf{k}$ dependences are omitted for clarity. We combine the above with the Legendre basis expansion (2.69) to get

$$\sum_{l=0}^{\infty}(-i)^l(2l+1)P_l(\mu)\left[\Theta_l' + ik\mu\Theta_l\right]$$
$$= \sum_{l=0}^{\infty}(-i)^l(2l+1)P_l(\mu)\left[\tau'\Theta_l + \delta_{l0}(\Phi' - \tau'\Theta_0) + \delta_{l1}(\frac{1}{3}k\Psi + \frac{1}{3}k\tau'v_b) + \delta_{l2}(\frac{1}{10}\tau'\Theta_2)\right], \qquad (2.72)$$

since $P_0(\mu) = 1$ and $P_1(\mu) = \mu$. Note that the term $ik\mu\Theta_l$ related to free-streaming of photons has an extra factor of μ, which can be removed using the recursion relation for Legendre polynomials given by

$$(2l+1)\mu P_l(\mu) = (l+1)P_{l+1}(\mu) + lP_{l-1}(\mu). \qquad (2.73)$$

We can then equate the coefficients of $P_l(\mu)$ from both sides for each l, since the Legendre polynomials are orthogonal when $l \neq l'$. The result is a *hierarchy* of differential equations for the multipoles:

$$\Theta_l' - \frac{l}{2l+1}k\Theta_{l-1} + \frac{l+1}{2l+1}k\Theta_{l+1}$$
$$= \tau'\Theta_l + \delta_{l0}(\Phi' - \tau'\Theta_0) + \delta_{l1}(\frac{1}{3}k\Psi + \frac{1}{3}k\tau'v_b) + \delta_{l2}(\frac{1}{10}\tau'\Theta_2), \qquad (2.74)$$

for $l = 0, 1, 2, \cdots$ (Θ_{-1} set to 0). Finally, the first-order differential equations from each l can be solved in conjunction with other linearised equations including (2.27), (2.29), and (2.35)-(2.37) derived from the perturbation theory.

We note two key properties of the full set of differential equations which dictate how perturbations in the metric, matter, and radiation vary as the universe expands. First, the Fourier modes 'decouple'; they evolve independently from each other. The perturbative variables evaluated at different $\mathbf{k}$s do not affect each other to leading order and hence can be studied separately. Second, the time evolution of each Fourier mode only depends on $k = \|\mathbf{k}\|$ instead of $\mathbf{k}$ for scalar perturbations. This is because our background theory is isotropic and there is no preferred direction in our formulation. The Euler equation (2.29) may seem to depend on the full vector $\mathbf{k}$ at first, but it simplifies greatly after rewriting $\mathbf{v} = \nabla v$ for some v. The same is true for the $i0$ component of the linearised Einstein's equations (2.36).

2.3.3   Late-time anisotropy

We have derived the set of differential equations governing the time evolution of photon temperature perturbations. Adiabatic initial conditions mean that there is only one degree of freedom for scalar perturbations at the end of inflation. Hence, $\Phi(\eta_i, \mathbf{k})$ for example fully specifies the initial values needed for the evolution equations, which themselves depend only on the magnitude k of the wavevector as discussed earlier. It follows that the *radiation transfer function* defined as $\Delta_l(k) := \Theta(k, \eta_0)/\Phi(k, \eta_i)$ encapsulates all the relevant information about the photons' time evolution, from the early times, $\eta = \eta_i$, to present, $\eta = \eta_0$. In this section, we derive the relations between the photon temperature perturbation $\Theta(k, \eta_0)$ and the CMB anisotropy observed today.

We measure the CMB while sitting on a fixed point in spacetime: here ($\mathbf{x} = \mathbf{x}_0$) and now ($\eta = \eta_0$). Small variations in our time and location have effects that are completely negligible considering the Hubble scale today. Anisotropy in the CMB temperature we observe hence takes the simple form

$$\left(\frac{\Delta T}{T} \right)(\hat{\mathbf{p}}) = \Theta(\eta_0, \mathbf{x}_0, \hat{\mathbf{p}}, \eta_0). \tag{2.75}$$

The vector $\hat{\mathbf{p}}$ relates to which direction in the sky we point our telescopes. The observed data thus lie on a two-dimensional sphere containing $\hat{\mathbf{p}}$. The equivalent of Fourier transform for functions defined on a sphere is the spherical harmonic transform;

$$\Theta(\mathbf{x}, \hat{\mathbf{p}}, \eta) = \sum_{l,m} a_{lm}(\mathbf{x}, \eta) Y_{lm}(\hat{\mathbf{p}}). \tag{2.76}$$

In physical terms, the spherical harmonics $Y_{lm}(\hat{\mathbf{p}})$ are joint eigenstates of the angular momentum operators $\hat{L}^2$ and $\hat{L}_3$ with associated quantum numbers l and m, respectively. In

mathematical terms, they form a basis of harmonic (vanishing Laplacian) polynomials of degree l in three dimensions, their domain restricted to a sphere. For each l, the number m may take one of the $2l+1$ values in $\{-l, -l+1, \cdots, l\}$. Note that Y_{lm}s are orthogonal and normalised by construction:

$$\int d^2\hat{\mathbf{p}}\, Y^*_{lm}(\hat{\mathbf{p}}) Y_{l'm'}(\hat{\mathbf{p}}) = \delta_{ll'}\delta_{mm'}. \tag{2.77}$$

Another useful identity is the addition theorem for spherical harmonics given by

$$P_l(\hat{\mathbf{k}}\cdot\hat{\mathbf{p}}) = \frac{4\pi}{2l+1}\sum_{m=-l}^{l} Y^*_{lm}(\hat{\mathbf{k}})Y_{lm}(\hat{\mathbf{p}}), \tag{2.78}$$

where $P_l(\mu)$ are Legendre polynomials.

As long as sufficiently many multipoles l are included, the spherical harmonic coefficients contain full information of the original function, just like the Fourier transform. Thanks to their orthonormality (2.77), the spherical harmonic coefficients can be computed using

$$a_{lm}(\mathbf{x},\eta) = \int d^2\hat{\mathbf{p}}\, \Theta(\mathbf{x},\hat{\mathbf{p}},\eta)Y^*_{lm}(\hat{\mathbf{p}}). \tag{2.79}$$

$$= \int \frac{d^3k}{(2\pi)^3}e^{i\mathbf{k}\cdot\mathbf{x}}\int d^2\hat{\mathbf{p}}\, \Theta(\mathbf{k},\hat{\mathbf{p}},\eta)Y^*_{lm}(\hat{\mathbf{p}}). \tag{2.80}$$

Note that the evolution equation (2.71) for the perturbations $\Theta(\mathbf{k},\hat{\mathbf{p}},\eta)$ above only depends on $k = \|\mathbf{k}\|$, the angle $\mu = \hat{\mathbf{k}}\cdot\hat{\mathbf{p}}$, and time η. The rest are determined from initial conditions. The following ratio therefore only depends on k, μ and η as well;

$$\frac{\Theta(\mathbf{k},\hat{\mathbf{p}},\eta)}{\Phi(\mathbf{k},\eta_i)} = f(k,\mu,\eta) \tag{2.81}$$

$$= \sum_l (-i)^l(2l+1)P_l(\mu)\frac{\Theta_l(\mathbf{k},\eta)}{\Phi(\mathbf{k},\eta_i)} \tag{2.82}$$

$$= \sum_l (-i)^l(2l+1)P_l(\mu)\Delta_l(k,\eta), \tag{2.83}$$

where we expanded using Legendre polynomials in (2.82) and used the definition of transfer functions. Substituting this into (2.80),

$$a_{lm}(\mathbf{x}, \eta) = \int \frac{d^3\mathbf{k}}{(2\pi)^3} e^{i\mathbf{k}\cdot\mathbf{x}} \Phi(\mathbf{k}, \eta_i) \int d^2\hat{\mathbf{p}} \, \frac{\Theta(\mathbf{k}, \hat{\mathbf{p}}, \eta)}{\Phi(\mathbf{k}, \eta_i)} Y_{lm}^*(\hat{\mathbf{p}}) \tag{2.84}$$

$$= \int \frac{d^3\mathbf{k}}{(2\pi)^3} e^{i\mathbf{k}\cdot\mathbf{x}} \Phi(\mathbf{k}, \eta_i) \int d^2\hat{\mathbf{p}} \sum_{l'} \left[(-i)^{l'}(2l'+1) P_{l'}(\hat{\mathbf{k}}\cdot\hat{\mathbf{p}}) \Delta_{l'}(k, \eta) Y_{lm}^*(\hat{\mathbf{p}}) \right] \tag{2.85}$$

$$= \int \frac{d^3\mathbf{k}}{(2\pi)^3} \left[e^{i\mathbf{k}\cdot\mathbf{x}} \Phi(\mathbf{k}, \eta_i)(-i)^l(2l+1)\Delta_l(k, \eta) Y_{lm}^*(\hat{\mathbf{k}}) \right]. \tag{2.86}$$

For the last line, we expanded $P_{l'}$ using the addition theorem (2.78) and performed the $\int d\hat{\mathbf{p}}$ integral. Orthonormality (2.77) forces $l' = l$ and simplifies the summation over l'.

Setting $\mathbf{x} = \mathbf{x_0} = \mathbf{0}$ and choosing $\eta = \eta_0$, we obtain a formula for the observed temperature anisotropy:

$$a_{lm} = 4\pi(-i)^l \int \frac{d^3\mathbf{k}}{(2\pi)^3} \Phi(\mathbf{k})\Delta_l(k)Y_{lm}^*(\hat{\mathbf{k}}). \tag{2.87}$$

2.3.4   CMB power spectrum

Slow-roll inflation generates quantum fluctuations that are mostly Gaussian. In Section 1.2.3 we derived the dimensionless power spectrum for $\delta\phi$ (1.85), the fluctuations in the inflationary field, by evaluating the vacuum expectation value $\langle \delta\phi(\mathbf{k}_1, \tau)\delta\phi(\mathbf{k}_2, \tau) \rangle$ through canonical quantisation of $\delta\phi$. The equivalent expression for the potential Φ evaluated at the end of inflation is

$$\langle \Phi(\mathbf{k})\Phi(\mathbf{k}') \rangle = (2\pi)^3 \delta^{(3)}(\mathbf{k} + \mathbf{k}')P_\Phi(k), \tag{2.88}$$

where $P_\Phi(k)$ is the power spectrum of Φ. The Dirac delta function above enforces the momentum conservation, a manifestation of statistical homogeneity. This can be seen by considering a spatial translation $\mathbf{x} \to \mathbf{x} + \mathbf{c}$ under which the field transforms as $\Phi(\mathbf{k}) \to \exp(i\mathbf{k}\cdot\mathbf{c})\Phi(\mathbf{k})$. The correlation function on the left-hand side of (2.88) then gains a factor of $\exp(i(\mathbf{k}+\mathbf{k}')\cdot\mathbf{c})$. For this to be equal to 1 for arbitrary $\mathbf{c}$, we require $\mathbf{k} + \mathbf{k}' = \mathbf{0}$ as above.

We can now compute the power spectrum of late-time CMB temperature anisotropy from the given primordial power spectrum $P_\Phi(k)$. Using (2.87),

$$\langle a_{lm}^* a_{l'm'}\rangle = (4\pi)^2 i^{l-l'} \int \frac{d^3\mathbf{k}}{(2\pi)^3}\frac{d^3\mathbf{k}'}{(2\pi)^3}\langle\Phi(\mathbf{k})\Phi(\mathbf{k}')\rangle(\mathbf{k}')\Delta_l(k)\Delta_{l'}(k)Y_{lm}(\hat{\mathbf{k}})Y_{l'm'}^*(\hat{\mathbf{k}'}) \quad (2.89)$$

$$= (4\pi)^2 i^{l-l'} \int \frac{d^3\mathbf{k}}{(2\pi)^3}\Delta_l(k)\Delta_{l'}(k)P_\Phi(k)Y_{lm}(\hat{\mathbf{k}})Y_{l'm'}^*(\hat{\mathbf{k}}) \quad (2.90)$$

$$= \frac{2}{\pi}i^{l-l'} \int dk\, k^2 \Delta_l(k)\Delta_{l'}(k)P_\Phi(k)\delta_{ll'}\delta_{mm'} \quad (2.91)$$

$$= C_l \delta_{ll'}\delta_{mm'}, \quad (2.92)$$

where the *angular power spectrum* is defined as

$$C_l := \frac{2}{\pi}\int dk\, k^2 |\Delta_l(k)|^2 P_\Phi(k) = 4\pi\int d(\ln k)\,|\Delta_l(k)|^2 \mathscr{P}_\Phi(k). \quad (2.93)$$

The dimensionless power spectrum $\mathscr{P}_\Phi(k) := (1/2\pi^2)P_\Phi(k)$ as before.

The Gaussianity of primordial perturbations from slow-roll inflation means that the a_{lm}s, which evolved linearly from the initial fluctuations, are also Gaussian. CMB measurements show that the a_{lm}s are indeed consistent with a multivariate Gaussian distribution. The form of (2.87) shows that the multipoles a_{lm} are uncorrelated with each other and, since they are Gaussian distributed, vanishing correlation implies independence. In particular, the a_{lm}s with the same l but distinct ms are independent and identically distributed. We can therefore combine them to get a more precise estimate for C_l;

$$\hat{C}_l = \frac{1}{2l+1}\sum_{m=-l}^{l} a_{lm}^* a_{lm}. \quad (2.94)$$

In cosmology, we are often interested in the theoretical distribution from which an observable is sampled from. Unfortunately, there is only one realisation of the universe we may observe: the one we live in. The number of samples used to estimate the underlying distribution is therefore limited, and it imposes an inherent lower bound on the estimation error called *cosmic variance*.

This also applies to estimating the angular power spectrum C_l from a single set of a_{lm}s observed from CMB experiments. Even in a perfect experiment with zero measurement noise, we expect the variance of the estimator to be

$$\text{Var}\left[\hat{C}_l\right] = \frac{2}{2l+1}C_l^2. \quad (2.95)$$

Having access to $2l+1$ independent samples a_{lm} for each l is therefore crucial for doing precision cosmology with the CMB data; the cosmic variance for C_l is suppressed by a factor of $2l+1$.

Summary

In this chapter we formulated the cosmological perturbation theory with a focus on the CMB anisotropy. Starting from the perturbed metric of the universe, we derived the conservation and Einstein equations to first order in perturbations. We then discussed how to fix a gauge and introduced several important gauge-invariant quantities, including the curvature perturbation ζ which is constant at super-horizon scales. Working in Newtonian gauge, we derived the time evolution equations for both matter and radiation perturbations. The latter involved writing down the Boltzmann equation with scattering terms. The result is a system of differential equations for the multipoles Θ_l of the photon temperature perturbations, which can be solved to obtain the transfer functions encapsulating the full evolution history of the CMB.

CMB anisotropy is the most powerful probe for constraining cosmological parameters of the ΛCDM model. The CMB power spectrum obtained from recent observations provided extremely precise estimates of the inflationary ΛCDM parameters such as the energy composition and age of the universe. In addition to the temperature anisotropy mainly discussed in this chapter, weak polarisation present in the CMB caused by Thompson scattering is also a key observable for cosmology. There are two polarisation modes, *E* and *B*, mainly sourced by the scalar and tensor perturbations, respectively. The E-mode polarisation data have been combined with the temperature data which significantly improved the CMB's estimation power thanks to recent advances in measurement sensitivity [14].

Furthermore, the fact that the CMB anisotropy depends linearly on the primordial perturbations makes it an ideal dataset for studying higher-order correlation functions. In the next chapter, we will discuss how the CMB bispectrum can be used to test the Gaussian assumption and constrain various inflation models via primordial non-Gaussianity.

Chapter 3

Bispectrum and Primordial Non-Gaussianity

The primordial perturbations are consistent with being Gaussian distributed according to observations to date [23]. The statistical properties of a Gaussian random field are completely characterised by its mean and two-point functions. The latter is directly related to the CMB angular power spectrum since the CMB is a near-linear probe of the early universe. In fact, the power spectrum is a sufficient statistic if the perturbations are Gaussian, which means that it contains all the information about the distribution of the primordial perturbations we may ever extract from the CMB. The inverse is also true however; if the initial perturbations were non-Gaussian, it is essential to go beyond the power spectra and study higher-order statistics. Such non-Gaussian contributions are well captured in the three-point correlation functions, or their Fourier counterpart, the bispectrum.

In single-field slow-roll inflation with standard kinetic term and vacuum, the primordial bispectrum is suppressed by slow-roll parameters [47]. However, numerous other inflationary scenarios that are physically well-motivated violate these simple assumptions and hence predict non-Gaussian signatures. They are expected to leave imprints on the CMB bispectrum with a characteristic shape and amplitude, where the latter is parametrised by f_{NL}.

We discuss the bispectrum in relation to primordial non-Gaussianity in this section, from both theoretical and observational sides. Section 3.1 covers the basic formalism of bispectrum analysis. Section 3.2 then introduces the theoretical tools for calculating the primordial bispectrum from a given inflation model Lagrangian. On the CMB side, we formulate the optimal bispectrum estimator and its existing implementations in Section 3.3.

3.1 Bispectrum

Consider the three-point correlation function $\langle \zeta(\mathbf{x}_1)\zeta(\mathbf{x}_2)\zeta(\mathbf{x}_3)\rangle$ of the curvature perturbation ζ at the end of inflation. Its Fourier transform is given by

$$\langle \zeta(\mathbf{k}_1)\zeta(\mathbf{k}_2)\zeta(\mathbf{k}_3)\rangle = \int d^3\mathbf{x}_1 d^3\mathbf{x}_2 d^3\mathbf{x}_3 \, e^{-i(\mathbf{k}_1\cdot\mathbf{x}_1+\mathbf{k}_2\cdot\mathbf{x}_2+\mathbf{k}_3\cdot\mathbf{x}_3)} \langle \zeta(\mathbf{x}_1)\zeta(\mathbf{x}_2)\zeta(\mathbf{x}_3)\rangle. \quad (3.1)$$

Assuming that the universe is statistically homogeneous, these correlation functions are invariant under arbitrary spatial translation $\mathbf{x} \to \tilde{\mathbf{x}} = \mathbf{x} + \mathbf{c}$. In particular, we have $\langle \zeta(\mathbf{x}_1)\zeta(\mathbf{x}_2)\zeta(\mathbf{x}_3)\rangle = \langle \zeta(\mathbf{x}_1-\mathbf{x}_3)\zeta(\mathbf{x}_2-\mathbf{x}_3)\zeta(\mathbf{0})\rangle$. Relabelling the integration variables so that $\mathbf{x}_1' = \mathbf{x}_1 - \mathbf{x}_3$ and $\mathbf{x}_2' = \mathbf{x}_2 - \mathbf{x}_3$, (3.1) becomes

$$\int d^3\mathbf{x}_1' d^3\mathbf{x}_2' d^3\mathbf{x}_3 \, e^{-i[\mathbf{k}_1\cdot(\mathbf{x}_1'+\mathbf{x}_3)+\mathbf{k}_2\cdot(\mathbf{x}_2'+\mathbf{x}_3)+\mathbf{k}_3\cdot\mathbf{x}_3]} \langle \zeta(\mathbf{x}_1')\zeta(\mathbf{x}_2')\zeta(\mathbf{0})\rangle$$

$$= (2\pi)^3 \delta^{(3)}(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3) \int d^3\mathbf{x}_1' d^3\mathbf{x}_2' \, e^{-i(\mathbf{k}_1\cdot\mathbf{x}_1'+\mathbf{k}_2\cdot\mathbf{x}_2')} \langle \zeta(\mathbf{x}_1')\zeta(\mathbf{x}_2')\zeta(\mathbf{0})\rangle \quad (3.2)$$

$$=: (2\pi)^3 \delta^{(3)}(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3) \, B(\mathbf{k}_1,\mathbf{k}_2), \quad (3.3)$$

where we defined $B(\mathbf{k}_1,\mathbf{k}_2)$ to be the integral expression appearing in (3.2). The delta function enforces $\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3 = \mathbf{0}$ which corresponds to the conservation of momentum.

Assuming further that the universe is statistically isotropic, the correlators remain constant under rotations $\mathbf{x} \to \tilde{\mathbf{x}} = R\mathbf{x}$ for any orthogonal matrix R. It is straightforward to see that $B(\mathbf{k}_1,\mathbf{k}_2)$ remains invariant under rotations as well;

$$B(R\mathbf{k}_1, R\mathbf{k}_2) = \int d^3\mathbf{x}_1' d^3\mathbf{x}_2' \, e^{-i[(R\mathbf{k}_1)\cdot\mathbf{x}_1'+(R\mathbf{k}_2)\cdot\mathbf{x}_2']} \langle \zeta(\mathbf{x}_1')\zeta(\mathbf{x}_2')\zeta(\mathbf{0})\rangle \quad (3.4)$$

$$= \int d^3\mathbf{x}_1'' d^3\mathbf{x}_2'' \, e^{-i(\mathbf{k}_1\cdot\mathbf{x}_1''+\mathbf{k}_2\cdot\mathbf{x}_2'')} \langle \zeta(R\mathbf{x}_1'')\zeta(R\mathbf{x}_2'')\zeta(\mathbf{0})\rangle \; = B(\mathbf{k}_1,\mathbf{k}_2), \quad (3.5)$$

where we have used the fact that $R^T = R^{-1}$. We may therefore fix $\mathbf{k}_1$ to be aligned with the z-axis, for example, and rotate further to have $\mathbf{k}_2$ lie on the xz plane. B then depends only on three variables: two lengths k_1, k_2 and the angle between the two given by $\mathbf{k}_1 \cdot \mathbf{k}_2/(k_1 k_2)$. Since $\|\mathbf{k}_3\|^2 = \|\mathbf{k}_1+\mathbf{k}_2\|^2 = k_1^2 + k_2^2 + 2(\mathbf{k}_1\cdot\mathbf{k}_2)$, the angle can be replaced by k_3. Putting everything together, we obtain

$$\langle \zeta(\mathbf{k}_1)\zeta(\mathbf{k}_2)\zeta(\mathbf{k}_3)\rangle = (2\pi)^3 \delta^{(3)}(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3) \, B(k_1,k_2,k_3). \quad (3.6)$$

Note that the bispectrum $B(k_1,k_2,k_3)$ is a three-dimensional function, as opposed to the one-dimensional power spectrum $P(k)$. The domain of B is further restricted by the constraint given by the delta function; k_1, k_2, and k_3 must form three sides of a triangle.
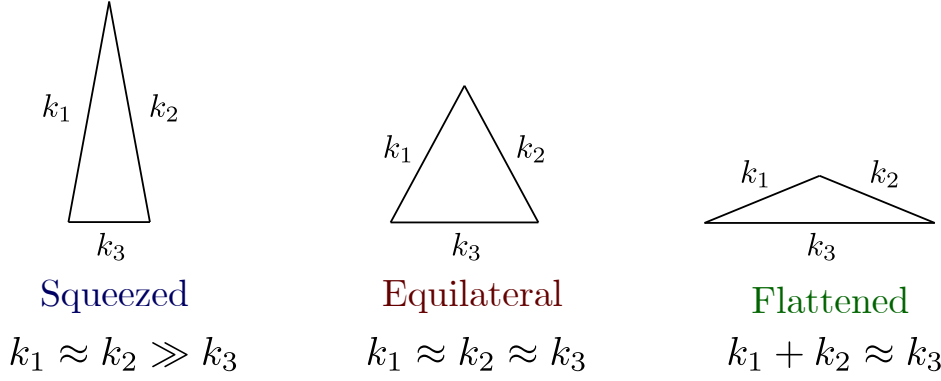
Fig. 3.1 Three notable triangle configurations for the bispectrum $B(k_1, k_2, k_3)$.

The functional form of the bispectrum comprises two parts: its dependence on the overall scaling $K = k_1 + k_2 + k_3$, called *running*, and the *shape* of the triangle formed by k_1, k_2 and k_3. Some notable shapes are depicted in Figure 3.1.

The squeezed, equilateral and flattened limits of bispectra all have distinct physical meanings. For example, the squeezed limit corresponds to a configuration with two small-scale (large k) modes and one large-scale (small k) mode. Heuristically, this relates to how the small scale covariances are affected by an encompassing large-scale mode.

One of the most studied bispectrum shapes comes from the *local* model. In this model, the perturbative field $\zeta(\mathbf{x})$ is expanded as a local function of some Gaussian field $\zeta_G(\mathbf{x})$ as

$$\zeta(\mathbf{x}) = \zeta_G(\mathbf{x}) + \frac{3}{5} f_{NL} \left(\zeta_G^2(\mathbf{x}) - \langle \zeta_G^2 \rangle \right) + \cdots . \tag{3.7}$$

The non-linearity parameter f_{NL} measures the amplitude of quadratic contributions to the field. A factor of $(3/5)$ here is conventional; the original definition of f_{NL} was written in terms of the potential Φ, which is equal to $(3/5)\zeta$ on super-horizon scales.

Substituting this form into (3.1) gives an expression involving correlation functions of four Gaussian fields, to leading order in f_{NL}. Isserlis' theorem, also known as Wick's theorem to physicists, allows us to write such four-point correlators in terms of the sum over all possible contractions:

$$\langle \zeta_G(\mathbf{x}_1)^2 \zeta_G(\mathbf{x}_2) \zeta_G(\mathbf{x}_3) \rangle = \langle \zeta_G(\mathbf{x}_1)^2 \rangle \langle \zeta_G(\mathbf{x}_2) \zeta_G(\mathbf{x}_3) \rangle + \langle \zeta_G(\mathbf{x}_1) \zeta_G(\mathbf{x}_2) \rangle \langle \zeta_G(\mathbf{x}_1) \zeta_G(\mathbf{x}_3) \rangle$$
$$+ \langle \zeta_G(\mathbf{x}_1) \zeta_G(\mathbf{x}_3) \rangle \langle \zeta_G(\mathbf{x}_1) \zeta_G(\mathbf{x}_2) \rangle . \tag{3.8}$$

Contributions from the first term in (3.8) cancel out with the ones coming from $\langle \zeta_G^2 \rangle$ in the definition (3.7). After expressing the two-point correlations in terms of the power spectrum in Fourier space as $\langle \zeta_G(\mathbf{k}_1) \zeta_G(\mathbf{k}_2) \rangle = (2\pi)^3 \delta^{(3)}(\mathbf{k}_1 + \mathbf{k}_2) P_\zeta(k_1, k_2)$, we obtain the *local*

bispectrum;

$$B(k_1, k_2, k_3) = \frac{3}{5} f_{\mathrm{NL}} \left[2P_\zeta(k_2)P_\zeta(k_3) + 2P_\zeta(k_3)P_\zeta(k_1) + 2P_\zeta(k_1)P_\zeta(k_2) \right]. \tag{3.9}$$

Due to this fact, it is common in literature (e.g. [48]) to define 'the reduced bispectrum' f_{NL} to be $f_{\mathrm{NL}}(k_1, k_2, k_3) := 5B(k_1, k_2, k_3)/6(P(k_1)P(k_2) + P(k_2)P(k_3) + P(k_3)P(k_1))$. f_{NL} is then a three-dimensional function in general. This is potentially confusing because of a conflicting convention in the observational community where f_{NL} represents a parameter measuring the amplitude of a specific bispectrum shape.

In this thesis, we follow the latter convention and associate one scalar f_{NL} with each of the bispectrum shapes. For example, f_{NL} in (3.9) is denoted $f_{\mathrm{NL}}^{\mathrm{local}}$, tied to the local shape. The connection between the theoretical predictions and observations is often made using approximate analytic templates for the bispectrum.

We parametrise the power spectrum as $P_\zeta(k) = A_\zeta k^{n_s - 4}$ using the power spectrum amplitude A_ζ and scalar spectral index n_s. The local bispectrum template is then defined as

$$B_\zeta^{\mathrm{local}}(k_1, k_2, k_3) := 2A^2 \left[\frac{1}{k_1^{4-n_s} k_2^{4-n_s}} + 2 \text{ cyc.} \right], \tag{3.10}$$

where $A^2 := (3/5)A_\zeta^2$. Note that we do not include $f_{\mathrm{NL}}^{\mathrm{local}}$ in our definition of the template either; it is set to one.

The primordial bispectrum from multi-field inflation models typically falls into this category (reviewed in e.g., [49]). The local shape peaks in the squeezed limit where $k_1 \approx k_2 \gg k_3$, diverging as $k_3 \to 0$. General single-field models, on the other hand, are better described by the equilateral and orthogonal templates [50, 51]:

$$B_\zeta^{\mathrm{equil}}(k_1, k_2, k_3) := 6A^2 \left[-\left(\frac{1}{k_1^{4-n_s} k_2^{4-n_s}} + 2 \text{ cyc.} \right) - \frac{2}{(k_1 k_2 k_3)^{2(4-n_s)/3}} \right.$$
$$\left. + \left(\frac{1}{k_1^{(4-n_s)/3} k_2^{2(4-n_s)/3} k_3^{(4-n_s)}} + 5 \text{ cyc.} \right) \right], \tag{3.11}$$

$$B_\zeta^{\mathrm{ortho}}(k_1, k_2, k_3) := 6A^2 \left[-3 \left(\frac{1}{k_1^{4-n_s} k_2^{4-n_s}} + 2 \text{ cyc.} \right) - \frac{8}{(k_1 k_2 k_3)^{2(4-n_s)/3}} \right.$$
$$\left. + 3 \left(\frac{1}{k_1^{(4-n_s)/3} k_2^{2(4-n_s)/3} k_3^{(4-n_s)}} + 5 \text{ cyc.} \right) \right]. \tag{3.12}$$

The equilateral and orthogonal shapes are peaked at the equilateral and flattened limits shown in Figure 3.1, respectively. The latter was constructed explicitly to probe bispectrum shapes perpendicular to the equilateral shape.

3.2 Primordial non-Gaussianity

Previously in Section 1.2.3, we derived the power spectrum of perturbations in the inflation field assuming a homogeneous background. This section outlines how to compute cosmological correlation functions from single-field inflation in general.

We introduce the necessary tools in Section 3.2.2 and 3.2.1. The bispectrum from a general type of single-field inflation is derived in Section 3.2.3. We focus on illustrating the framework without delving too much into the technical details. A simple example is used to demonstrate how explicit calculations are done, while we refer to, e.g., [21, 47, 48] for the full (and laborious) calculations.

3.2.1 ADM formalism

In order to study the perturbed metric in the presence of one or more inflation fields, it is convenient to use the Arnowitt-Deser-Misner (ADM) formalism [52]:

$$ds^2 = -Ndt^2 + h_{ij}(dx^i + N^i dt)(dx^j + N^j dt), \tag{3.13}$$

where the lapse and shift functions N and N^i, respectively, are non-dynamical variables of the action. They act as Lagrange multipliers and provide constraint equations. We focus on the scalar perturbations here and ignore vector and tensor modes since they evolve independently owing to the SVT theorem. The spatial metric h_{ij} contains two scalar degrees of freedom. When inflation is driven by a single scalar field, there is one extra scalar mode from perturbations of the field. Out of the three dynamical scalar degrees of freedom, two can be fixed by a gauge choice as we saw in 2.1.1. Here we choose the gauge so that the inflation field is uniform. The spatial metric takes the form

$$h_{ij} = a(t)^2 \, e^{2\zeta} \, \delta_{ij}, \tag{3.14}$$

where $\zeta(\mathbf{x},t)$ is equivalent to the curvature perturbation we defined earlier in (2.39). Note that, if we define $N(t) = \int_{t_0}^{t} H dt' = \ln a(t) - \ln a(t_0)$ (not to be confused with the lapse function N) to represent the number of e-folds of expansion between time t and t_0, then $\zeta = \delta N$. [1]

We now write down the action. Compared to (1.49), the Einstein-Hilbert term is added since the metric is no longer fixed. We consider a generalised form of the field Lagrangian with the action given by

$$S = \int dt d^3\mathbf{x} \sqrt{-g} \left[\frac{M_p}{2} R + P(X, \phi) \right],$$ (3.15)

where $M_p := (8\pi G)^{-1/2}$ is the reduced Planck mass and $X := -\frac{1}{2} g^{\mu\nu} \partial_\mu \phi \partial_\nu \phi$ is the canonical kinetic term. We set $M_p = 1$ for convenience. $P(X, \phi)$ is an arbitrary function which generalises $X - V(\phi)$ of the standard slow-roll inflation [53].

One notable consequence of the non-trivial kinetic term P is that the fluctuations in the inflation field no longer necessarily propagate with the speed of light. The sound speed, defined as the ratio $\delta P/\delta \rho$, in these models is given by

$$c_s^2 = \frac{P_{,X}}{P_{,X} + 2X P_{,XX}},$$ (3.16)

where the subscript '$,X$' represents taking a partial derivative with respect to X.

In the ADM formalism, (3.15) becomes

$$S = \frac{1}{2} \int dt d^3\mathbf{x} \sqrt{h} N \left[R^{(3)} + 2P(X, \phi) \right] + \frac{1}{2} \int dt d^3\mathbf{x} \sqrt{h} N^{-1} \left[E_{ij} E^{ij} - E^2 \right],$$ (3.17)

where the $R^{(3)}$ is the three-dimensional Ricci scalar of h_{ij}, and

$$E_{ij} := \frac{1}{2} \left(\dot{h}_{ij} - \nabla_i N_j - \nabla_j N_i \right), \qquad E := E_{ij} h^{ij}.$$ (3.18)

The spatial indices i, j are lowered and raised by h_{ij}.

In order to obtain an action for ζ, we substitute (3.14) into the expression and expand perturbatively in ζ. We keep terms of order up to three as they are all that is relevant for three-point correlation functions. It is sufficient to evaluate N and N_i to first order in ζ since they multiply with constraint equations which vanish up to first order.

---

[1] To be precise, δN is defined as in (2.21) with $f = N$. We have $\zeta(t, \mathbf{x}) = \delta N = N(t + \delta t) - N(t)$, where $\delta t(t, \mathbf{x})$ represents the perturbations in the inflationary 'clock time' at the given spatial position $\mathbf{x}$.

The quadratic part of the action is given by

$$S_2 = \int dt\, d^3\mathbf{x} \left[\frac{\varepsilon}{c_s^2} a^3 \dot{\zeta}^2 - \varepsilon a (\partial \zeta)^2 \right], \tag{3.19}$$

where the slow-roll parameter $\varepsilon = -\dot{H}/H^2$ has made a reappearance. Here, $(\partial \zeta)^2$ is short-hand for $\partial_i \zeta \partial^i \zeta$ which involves the spatial derivatives of ζ. Note that there are no terms proportional to ζ^2 in the quadratic action, meaning that ζ is massless even without the slow-roll assumptions.

The third order action contains many terms with varying amounts of contributions to the bispectrum. We write only one particular term here for an example calculation which will be shown later. The full result can be found in [53].

$$S_3 \supset \int d\tau\, d^3\mathbf{x}\, a^3 \left[-\frac{\varepsilon}{Hc_s^2} \left(1 - \frac{1}{c_s^2} \right) + \frac{2\lambda c_s^2}{H^2 \varepsilon} \right] \dot{\zeta}^3. \tag{3.20}$$

3.2.2   In-in formalism

We need to calculate the Hamiltonian from the action in order to quantise the field. When only the quadratic part (3.19) is considered, the corresponding Hamiltonian, say H_0, is also quadratic in ζ and its conjugate momentum $\pi := \partial \mathscr{L}/\partial(\dot{\zeta})$. H_0 describes the free theory in which the equations of motion are linear and we can find the mode functions that solve them, as we did in Section 1.2.3. Meanwhile, third-order and higher terms in the action such as (3.20) appear in the Hamiltonian as couplings that contribute to the correlation functions. Denoting these 'interacting' terms as H_{int}, the total Hamiltonian for the perturbations takes the form

$$H = H_0 + H_{\text{int}}. \tag{3.21}$$

There are multiple formalisms in quantum mechanics regarding the time evolution of operators and states. In the Schrödinger picture, the states evolve in time according to the Hamiltonian, while the operators remain fixed over time. On the other hand, the Heisenberg picture has the operators varying in time and the states constant. Yet another alternative is the interaction picture where the Hamiltonian (3.21) is split into two parts, responsible for the time evolution of the states and the operators, respectively. All three pictures are physically equivalent; all the correlation function values are identical between them.

We choose to work in the interaction picture where the operators and states evolve in time through H_0 and H_{int}, respectively. The main advantage of this formalism is that we can use the results from the free ($H = H_0$) theory to evolve operators in time. In particular, we

may canonically quantise the field $\hat{\zeta}(\mathbf{x},t)$ using its Fourier transform;

$$\hat{\zeta}_{\mathbf{k}}(t) = u_k(t)\,\hat{a}_{\mathbf{k}} + u_k^*(t)\,\hat{a}_{-\mathbf{k}}^{\dagger}, \tag{3.22}$$

where the mode functions $u_k(\tau)$ satisfy the equations of motion from the quadratic action (3.19) related to H_0, which we can solve analytically. Note that $u_k(\tau)$ depends only on $k = \|\mathbf{k}\|$ because the spatial derivatives enter through $(\partial\zeta)^2$ only. The annihilation and creation operators $\hat{a}_{\mathbf{k}}$ and $\hat{a}_{-\mathbf{k}}^{\dagger}$ satisfy the usual commutation relations: $[\hat{a}_{\mathbf{k}_1}, \hat{a}_{\mathbf{k}_2}^{\dagger}] = (2\pi)^3 \delta^{(3)}(\mathbf{k}_1 - \mathbf{k}_2)$.

Meanwhile, all the information about the interactions is captured in the time evolution operator $\hat{U}_I(t,t_0) = T\exp\left(-i\int_{t_0}^t H_I(t')dt'\right)$, where H_I denotes H_{int} in the interaction picture. The time-ordering operator T arranges the operators so that the ones evaluated at earlier times appear on the right. The states at time t are given by $|\psi_I(t)\rangle = U_I(t,t_0)|\psi_I(t_0)\rangle$.

The in-in formalism ([54–56]) allows us to calculate correlation functions at a given time t using two 'in' states: two copies of the vacuum at infinite past in our case. For some local operator $Q(t)$ written as a product of $\hat{\zeta}(\mathbf{x},t)$ and $\hat{\pi}(\mathbf{x},t)$, we define its expected value as follows;

$$\begin{aligned}
\langle Q(t)\rangle &:= \langle\text{in}|\, Q(t)\, |\text{in}\rangle \\
&= \langle\mathbf{0}|\left[\bar{T}\exp\left(i\int_{t_0}^t H_I(t')dt\right)\right]Q_I(t)\left[T\exp\left(-i\int_{t_0}^t H_I(t')dt\right)\right]|\mathbf{0}\rangle. \tag{3.23}
\end{aligned}$$

The interaction picture vacuum, $|\mathbf{0}\rangle$, lies in the infinite past $t = t_0$, where we assume that it can be identified as the vacuum of non-interacting theory and hence $a_{\mathbf{k}}|\mathbf{0}\rangle = 0$.

Note that the operators within the correlator in (3.23) are not time-ordered thanks to the anti-time-ordering operator $\bar{T}$. To complicate things further, we do not have an equivalent of the Feynmann propagator in Minkowski space that handles the time ordering for us. We instead treat the time ordering manually. Let us define the *contraction* between two terms $\zeta(\mathbf{k}_1,t_1)$ on the left and $\zeta(\mathbf{k}_2,t_2)$ on the right as the commutator given by

$$[\zeta^+(\mathbf{k}_1,t_1),\zeta^-(\mathbf{k}_2,t_2)] = u_{k_1}(t_1)u_{k_2}^*(t_2)[\hat{a}_{\mathbf{k}_1},\hat{a}_{\mathbf{k}_2}^{\dagger}] \tag{3.24}$$

$$= u_{k_1}(t_1)u_{k_2}^*(t_2)(2\pi)^3\delta^{(3)}(\mathbf{k}_1 + \mathbf{k}_2), \tag{3.25}$$

where the positive and negative frequency solutions ζ^+ and ζ^- refer to the terms with $\hat{a}$ and $\hat{a}^{\dagger}$ in (3.22), respectively. Note that swapping the order of $\zeta(\mathbf{k}_1,t_1)$ and $\zeta(\mathbf{k}_2,t_2)$ does alter the value of this commutator. We make equivalent definitions for the quantised conjugate momenta $\hat{\pi}_{\mathbf{k}}$.

We now have a result comparable to Wick's theorem;

$$\langle \mathbf{0}|\hat{O}(\mathbf{k}_1,t_2)\cdots\hat{O}(\mathbf{k}_n,t_n)|\mathbf{0}\rangle =\; :\; \hat{O}(\mathbf{k}_1,t_2)\cdots\hat{O}(\mathbf{k}_n,t_n)\; :\; +\; :\; \text{all possible contractions}\; :\; .$$

$$(3.26)$$

The normal ordering $::$ keeps $\hat{a}$s to the right and $\hat{a}^\dagger$ to the left so that the vacuum expectation value vanishes. The fields $\hat{O}$ are either $\hat{\zeta}$ or $\hat{\pi}$.

3.2.3 Bispectrum from single-field inflation

The mode function defined in (3.22) can be obtained analytically by solving the equations of motion from the quadratic action (3.19). If the slow-roll and other variation parameters remain small and change slowly in time, then

$$u_k(\tau) = \frac{iH}{\sqrt{4\varepsilon c_s k^3}}(1+ikc_s\tau)e^{-ikc_s\tau},$$

$$(3.27)$$

where τ is conformal time. The choice of $u_k(\tau)$ being the positive frequency solution as shown here is analogous to (1.83) from before; we are fixing the vacuum state to be the Bunch-Davies vacuum.

Before proceeding to compute correlation functions using the in-in formalism, we first need to obtain the Hamiltonian of the system. The conjugate momentum is defined as $\pi = \partial\mathcal{L}/\partial\dot{\zeta}$, where $\mathcal{L}$ is the Lagrangian density. The Hamiltonian density then takes the form $\mathcal{H} = \pi\dot{\zeta} - \mathcal{L} = \mathcal{H}_0 + \mathcal{H}_{\text{int}}$. At the order of perturbations we are studying, $\mathcal{H}_{\text{int}}$ consists only of cubic terms $\mathcal{H}_3$, which is simply equal to $-\mathcal{L}_3$ obtained from the cubic action S_3. It is also sufficient to consider the leading contribution to ζ, so that $\pi \propto \dot{\zeta}$. [2] We will rewrite π in terms of $\dot{\zeta}$ the rest of this section.

When working perturbatively, the leading contributions to the three-point correlation function comes from the terms with one factor of the interaction Hamiltonian. Noting that the two terms coming from the two H_{I}s in (3.23) are complex conjugates to each other, we have

$$\langle \zeta_{\mathbf{k}_1}(t)\zeta_{\mathbf{k}_2}(t)\zeta_{\mathbf{k}_3}(t)\rangle = 2\text{Re}\left[\langle \mathbf{0}| - i\zeta_{\mathbf{k}_1}(t)\zeta_{\mathbf{k}_2}(t)\zeta_{\mathbf{k}_3}(t)\int_{t_0}^t H_{\text{I}}(t')dt'|\mathbf{0}\rangle\right],$$

$$(3.28)$$

---

[2]There are some subtleties here. We first use $\pi = \partial\mathcal{L}/\partial\dot{\zeta}$, which does contain contributions from higher order terms coming from $\mathcal{L} = \mathcal{L}_0 + \mathcal{L}_{int}$. The Hamiltonian density $\mathcal{H} = \pi\dot{\zeta} - \mathcal{L}$ is obtained using this π. We then rewrite π in terms of ζ and its derivatives within $\mathcal{H}$. Next, the quadratic terms in $\mathcal{H}$ are combined as $\mathcal{H}_0$. From this free Hamiltonian we define the conjugate momentum in the interaction picture; $\pi_I := \partial\mathcal{H}_0/\partial\dot{\zeta}$. This π_I is what we refer to as π, simply proportional to $\dot{\zeta}$ in our case.

in tree-level. This is the key formula for computing the primordial bispectrum from a given single-field inflation Lagrangian.

We will now demonstrate how to compute (3.28) using a simple example. We take the term proportional to $\dot{\zeta}^3$ within the cubic action (3.20). To simplify further, we write it as $\mu a^3 \dot{\zeta}^3$, where $\mu = \mu(t)$ is given in terms of the variation parameters such as λ and ε. The corresponding contribution to the interaction Hamiltonian is given by

$$H_{\mathrm{I},\dot{\zeta}^3}(t) := -\mu(t)a(t)^3 \int d^3\mathbf{x} \left[\dot{\zeta}(\mathbf{x},t)^3 \right] \tag{3.29}$$

$$= -\mu a^3 \int d^3\mathbf{x} \frac{d^3\mathbf{p}_1}{(2\pi)^3} \frac{d^3\mathbf{p}_2}{(2\pi)^3} \frac{d^3\mathbf{p}_3}{(2\pi)^3} e^{i(\mathbf{p}_1+\mathbf{p}_2+\mathbf{p}_3)\cdot\mathbf{x}} \left[\dot{\zeta}_{\mathbf{p}_1} \dot{\zeta}_{\mathbf{p}_2} \dot{\zeta}_{\mathbf{p}_3} \right] \tag{3.30}$$

$$= -(2\pi)^3 \delta^{(3)}(\mathbf{p}_1 + \mathbf{p}_2 + \mathbf{p}_3) \cdot \mu a^3 \int \frac{d^3\mathbf{p}_1}{(2\pi)^3} \frac{d^3\mathbf{p}_2}{(2\pi)^3} \frac{d^3\mathbf{p}_3}{(2\pi)^3} \left[\dot{\zeta}_{\mathbf{p}_1} \dot{\zeta}_{\mathbf{p}_2} \dot{\zeta}_{\mathbf{p}_3} \right], \tag{3.31}$$

where we used the definition of Fourier transformation and omitted the time dependence for brevity. Note also that $\zeta_{\mathbf{p}}(t) = const$ after the mode crosses the horizon because the curvature perturbation freezes out at super-horizon scales—a result we have shown in Section 2.2.3.

We substitute (3.31) into (3.28) and swap the order of expectations and integrals. The integrand include a six-point correlation function which can be evaluated using (3.26);

$$\langle \mathbf{0} | \zeta_{\mathbf{k}_1}(t) \zeta_{\mathbf{k}_2}(t) \zeta_{\mathbf{k}_3}(t) \zeta_{\mathbf{p}_1}(t') \zeta_{\mathbf{p}_2}(t') \zeta_{\mathbf{p}_3}(t') | \mathbf{0} \rangle$$

$$= \langle \mathbf{0} | : \text{all contractions} : | \mathbf{0} \rangle \tag{3.32}$$

$$= [\zeta^+(\mathbf{k}_1,t), \zeta^-(\mathbf{p}_1,t')][\zeta^+(\mathbf{k}_2,t), \zeta^-(\mathbf{p}_2,t')][\zeta^+(\mathbf{k}_3,t), \zeta^-(\mathbf{p}_3,t')] + 5 \text{ perms.} \tag{3.33}$$

$$= \prod_{j=1}^{3} \left[u_{k_j} \dot{u}_{p_j}^*(t')(2\pi)^3 \delta^{(3)}(\mathbf{k}_j + \mathbf{p}_j) \right] + 5 \text{ perms..} \tag{3.34}$$

Note that we discarded some terms which involve contracting the fields at equal times. Such contractions necessarily cause one of the $\mathbf{p}_j$ to be $\mathbf{0}$ due to the delta function present in (3.31). $\zeta_{\mathbf{0}}$ corresponds to a constant scaling of the background $a(t)$, and after absorbing that factor into $a(t)$, $\zeta_{\mathbf{0}}$ can be set to zero.

Combining the results so far, we obtain the following expression for the bispectrum induced by the $\dot{\zeta}^3$ term

$$B_{\dot{\zeta}^3}(k_1,k_2,k_3,t) = \mathrm{Re}\left[2iu_{k_1}(t)u_{k_2}(t)u_{k_3}(t) \int_{t_0}^{t} dt' \mu(t')a(t')^3 \dot{u}_{k_1}^*(t')\dot{u}_{k_2}^*(t')\dot{u}_{k_3}^*(t') + 5 \text{ perms} \right]. \tag{3.35}$$

In order to evaluate the bispectrum at the end of inflation, we convert back from comoving to conformal time. The time integral is then taken from $\tau = -\infty$ to $\tau = 0$. The lower limit of the integral can be troublesome since $e^{ikc_s\tau}$ in the mode function displays oscillatory behaviour as $\tau \to -\infty$. We shift $-\infty$ to $-\infty(1+i\varepsilon)$ so that the integrand is suppressed at the lower limit. We further approximate the variation parameters including $\mu(\tau)$ to be constant and use the de Sitter background $a(\tau) \approx -1/(H\tau)$ within the integral;

$$B_{\dot\zeta^3}(k_1, k_2, k_3) \approx -12\mu \cdot \text{Imag}\left[u_{k_1}(0)u_{k_2}(0)u_{k_3}(0)\int_{-\infty(1+i\varepsilon)}^{0} d\tau \, a(\tau)u'^*_{k_1}(\tau)u'^*_{k_2}(\tau)u'^*_{k_3}(\tau)\right]$$

(3.36)

$$= -12\mu \cdot \text{Imag}\left[\frac{H^5 c_s^3}{64\varepsilon^3}\frac{1}{k_1k_2k_3}\int_{-\infty(1+i\varepsilon)}^{0} d\tau \, \tau^2 e^{ic_s\tau(k_1+k_2+k_3)}\right]$$

(3.37)

$$= \frac{3\mu H^5}{8\varepsilon^3}\frac{1}{k_1k_2k_3}\frac{1}{(k_1+k_2+k_3)^3}.$$

(3.38)

The corresponding shape function takes the form

$$S_{\dot\zeta^3}(k_1, k_2, k_3) \propto \frac{k_1k_2k_3}{(k_1+k_2+k_3)^3},$$

(3.39)

which is maximised in the equilateral limit and vanishes at the squeezed limit.

The full bispectrum can be computed using the general methodology shown in this section. We quote one of the most important results; for canonical slow-roll inflation with $P(X, \phi) = X - V(\phi)$, the bispectrum is suppressed by the slow-roll parameters ε and η [47]. However, relaxing any of the assumptions—canonical kinetic term, slow-roll, single-field, and Bunch-Davies vacuum—may yield significant and observable signatures in the bispectrum (see [21, 22] for reviews).

3.3 CMB bispectrum estimation

The presence of a non-vanishing bispectrum at the end of inflation due to primordial non-Gaussianity leaves imprints on the CMB. The bispectrum of the observed CMB anisotropy directly relates to the primordial counterpart thanks to the linear nature of the evolution and hence is a key statistic for constraining primordial non-Gaussianity.

The CMB appears consistent with being Gaussian distributed, so its bispectrum has relatively small signal-to-noise. In contrast to the power spectrum analysis where we treat each C_l (after binning) as an independent data point, the noise-dominated $b_{l_1l_2l_3}$s are not so significant individually. Therefore, we instead fit the whole bispectrum data to the

theoretical prediction from a given model in order to estimate a single parameter measuring the amplitude: f_{NL}.

In this section, we formulate the theory of CMB bispectrum estimation. Estimating f_{NL} is an extremely challenging task due to its computational complexity and the oscillatory integrals involved. We review several conventional approaches to handle these challenges and summarise their respective strengths and weaknesses. Lastly, we discuss some significant non-primordial sources of non-Gaussianity that needs to be carefully accounted for in the estimation process.

3.3.1 CMB bispectrum

Consider the three-point correlation function of the spherical harmonic coefficients a_{lm}^X from (2.87);

$$\left\langle a_{l_1 m_1}^{X_1} a_{l_2 m_2}^{X_2} a_{l_3 m_3}^{X_3} \right\rangle = (4\pi)^3 (-i)^{l_1 + l_2 + l_3} \left\langle \prod_{j=1}^{3} \left[\int \frac{d^3 \mathbf{k}_j}{(2\pi)^3} \zeta(\mathbf{k}_j) \Delta_{l_j}^{X_j}(k_j) Y_{l_j m_j}^*(\hat{\mathbf{k}}_j) \right] \right\rangle, \quad (3.40)$$

where we replaced Φ with ζ in the integrand. Here, X_js can be either T or E which correspond to the temperature and E-mode polarisation of the CMB anisotropy, respectively. The transfer functions $\Delta_l^X(k)$ for ζ depend only on $k = \|\mathbf{k}\|$. They incorporate all information about the evolution of the primordial perturbations ζ and then projection onto the observed sky today.

We may take everything but $\zeta(\mathbf{k_j})$s outside the brackets $\langle \cdot \rangle$. From the definition of bispectrum, we have

$$\langle \zeta(\mathbf{k}_1) \zeta(\mathbf{k}_2) \zeta(\mathbf{k}_3) \rangle = (2\pi)^3 \delta^{(3)}(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3) B(k_1, k_2, k_3) \quad (3.41)$$

$$= \int d^3 \mathbf{r} \, e^{-i\mathbf{r} \cdot (\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3)} B(k_1, k_2, k_3), \quad (3.42)$$

where an integral expression is substituted for the Dirac δ-function in the second line. At the cost of introducing an extra integral, we managed to express the δ-function in a separable form: $\exp(\mathbf{r} \cdot (\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3)) = \exp(\mathbf{r} \cdot \mathbf{k}_1) \exp(\mathbf{r} \cdot \mathbf{k}_2) \exp(\mathbf{r} \cdot \mathbf{k}_3)$. The remaining exponentials are rewritten using the plane wave expansion;

$$e^{-i\mathbf{k} \cdot \mathbf{r}} = \sum_{l=0}^{\infty} (2l+1)(-i)^l j_l(kr) P_l(\hat{\mathbf{k}} \cdot \hat{\mathbf{r}}) \quad (3.43)$$

$$= \sum_{l=0}^{\infty} \sum_{m=-l}^{l} 4\pi (-i)^l j_l(kr) Y_{lm}(\hat{\mathbf{k}}) Y_{lm}^*(\hat{\mathbf{r}}). \quad (3.44)$$

The Legendre polynomial $P_l(\hat{\mathbf{k}} \cdot \hat{\mathbf{r}})$ has been expanded using the spherical harmonic addition theorem in the last line. Note that $\mathbf{k}$ and $\mathbf{r}$ mix only through their amplitudes within the spherical bessel functions as $j_l(kr)$. Once substituted into (3.40), we can perform the angular integral $d^2\hat{\mathbf{k}}_j$ separately for each $j = 1, 2, 3$, since $d^3\mathbf{k}_j = dk_j k_j^2 d^2\hat{\mathbf{k}}_j$. Note also that the spherical harmonic orthogonality relation is given by

$$\int d^2\hat{\mathbf{n}} \, Y_{lm}(\hat{\mathbf{n}})Y_{l'm'}^*(\hat{\mathbf{n}}) = \delta_{ll'}\delta_{mm'}. \tag{3.45}$$

There are no implied summations over indices throughout this section. Incorporating (3.40), (3.42), and (3.44), we obtain

$$\left\langle a_{l_1 m_1}^{X_1} a_{l_2 m_2}^{X_2} a_{l_3 m_3}^{X_3} \right\rangle$$
$$= \frac{(4\pi)^3}{(2\pi)^9}(-1)^{l_1+l_2+l_3} \int d^3\mathbf{r}\, d^3\mathbf{k}_1 d^3\mathbf{k}_2 d^3\mathbf{k}_3 \, B(k_1,k_2,k_3)$$
$$\times \prod_{j=1}^{3}\left[\sum_{l'_j=0}^{\infty}\sum_{m'_j=-l'_j}^{l'_j} j_{l'_j}(k_j r)Y_{l'_j m'_j}(\hat{\mathbf{k}}_j)Y_{l'_j m'_j}^*(\hat{\mathbf{r}})\Delta_{l_j}^{X_j}(k_j)Y_{l_j m_j}^*(\hat{\mathbf{k}}_j)\right]$$
$$\tag{3.46}$$

$$= \left(\frac{2}{\pi}\right)^3 \int d^3\mathbf{r}\, dk_1 dk_2 dk_3 \, (k_1 k_2 k_3)^2 B(k_1,k_2,k_3)\prod_{j=1}^{3}\left[j_{l_j}(k_j r)Y_{l_j m_j}^*(\hat{\mathbf{r}})\Delta_{l_j}^{X_j}(k_j)\right] \tag{3.47}$$

$$= \left(\frac{2}{\pi}\right)^3 \mathscr{G}_{m_1 m_2 m_3}^{l_1 l_2 l_3*} \int dr\, dk_1 dk_2 dk_3 r^2 \, (k_1 k_2 k_3)^2 B(k_1,k_2,k_3)\prod_{j=1}^{3}\left[j_{l_j}(k_j r)\Delta_{l_j}^{X_j}(k_j)\right], \tag{3.48}$$

where the Gaunt integral is defined as

$$\mathscr{G}_{m_1 m_2 m_3}^{l_1 l_2 l_3} := \int d^2\hat{\mathbf{n}} \, Y_{l_1 m_1}(\hat{\mathbf{n}})Y_{l_2 m_2}(\hat{\mathbf{n}})Y_{l_3 m_3}(\hat{\mathbf{n}}). \tag{3.49}$$

This value is always real, so we may omit the complex conjugate in (3.48). Note also that we dropped a factor of $(-1)^{l_1+l_2+l_3}$ in (3.47). This is due to parity reasons. Spherical harmonics have definite parity; $Y_{lm}(-\hat{\mathbf{n}}) = (-1)^l Y_{lm}(\hat{\mathbf{n}})$. Applying parity transformation to the integral in (3.49) gives $\mathscr{G}_{m_1 m_2 m_3}^{l_1 l_2 l_3} = (-1)^{l_1+l_2+l_3}\mathscr{G}_{m_1 m_2 m_3}^{l_1 l_2 l_3}$. The Gaunt integral therefore evaluates to zero unless $l_1 + l_2 + l_3$ is even.

We define the *reduced* bispectrum as

$$b_{l_1 l_2 l_3}^{X_1 X_2 X_3} := \left(\frac{2}{\pi}\right)^3 \int dr dk_1 dk_2 dk_3 \, (rk_1 k_2 k_3)^2 B(k_1,k_2,k_3)\prod_{j=1}^{3}\left[j_{l_j}(k_j r)\Delta_{l_j}^{X_j}(k_j)\right]. \tag{3.50}$$

The late-time bispectrum can now be written in a concise form;

$$\left\langle a_{l_1 m_1}^{X_1} a_{l_2 m_2}^{X_2} a_{l_3 m_3}^{X_3} \right\rangle = \mathscr{G}_{m_1 m_2 m_3}^{l_1 l_2 l_3} b_{l_1 l_2 l_3}^{X_1 X_2 X_3}. \tag{3.51}$$

Recall that the three-point function in (3.41) is given by a product of delta function enforcing $\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3 = \mathbf{0}$ and the primordial bispectrum $B(k_1, k_2, k_3)$. Its spherical harmonic counterpart (3.51) takes an analogous form. The Gaunt integral $\mathscr{G}_{m_1 m_2 m_3}^{l_1 l_2 l_3}$ contains all geometrical information, enforcing the triangle condition on l_1, l_2, l_3 and angular momentum conservation $m_1 + m_2 + m_3 = 0$. Meanwhile, the reduced bispectrum encodes statistical information about the underlying three-point functions, just like $B(k_1, k_2, k_3)$.

The value of the Gaunt integral is best represented using Wigner 3-j symbols;

$$\mathscr{G}_{m_1 m_2 m_3}^{l_1 l_2 l_3} = \sqrt{\frac{(2l_1+1)(2l_2+1)(2l_3+1)}{4\pi}} \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} \begin{pmatrix} l_1 & l_2 & l_3 \\ 0 & 0 & 0 \end{pmatrix}. \tag{3.52}$$

The Wigner 3-j symbols, written here as a 2-by-3 matrix, are closely related to the addition of angular momenta. They are real coefficients appearing in the expansion of the zero-total-angular-momentum state $|0\,0\rangle$;

$$|0\,0\rangle = \sum_{l_1,m_1} \sum_{l_2,m_2} \sum_{l_3,m_3} \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} |l_1 m_1\rangle |l_2 m_2\rangle |l_3 m_3\rangle. \tag{3.53}$$

For further details on Wigner 3-j symbols see, e.g., [57]. We quote the following two identities for our purposes.

$$\sum_{m_1,m_2,m_3} \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix}^2 = 1, \tag{3.54}$$

$$\begin{pmatrix} l_1 & l_2 & l_3 \\ 0 & 0 & 0 \end{pmatrix}^2 = \frac{1}{2} \int_{-1}^{1} d\mu \, P_{l_1}(\mu) P_{l_2}(\mu) P_{l_3}(\mu). \tag{3.55}$$

The normalisation condition (3.54) can be easily derived by computing the norm of the state $|0\,0\rangle$ in the definition. The second identity (3.55) allows us to rewrite a square of any given 3-j symbol satisfying $m_1 = m_2 = m_3 = 0$ in terms of a separable integral.

We make one last definition which will prove to be useful in the next section;

$$h^2_{l_1 l_2 l_3} := \sum_{m_1, m_2, m_3} \left(\mathscr{G}^{l_1 l_2 l_3}_{m_1 m_2 m_3} \right)^2 \tag{3.56}$$

$$= \frac{(2l_1 + 1)(2l_2 + 1)(2l_3 + 1)}{4\pi} \begin{pmatrix} l_1 & l_2 & l_3 \\ 0 & 0 & 0 \end{pmatrix}^2 \tag{3.57}$$

$$= \frac{(2l_1 + 1)(2l_2 + 1)(2l_3 + 1)}{8\pi} \int_{-1}^{1} d\mu \, P_{l_1}(\mu) P_{l_2}(\mu) P_{l_3}(\mu). \tag{3.58}$$

By squaring the Gaunt integral and summing over ms, we get a simpler quantity $h^2_{l_1 l_2 l_3}$ which preserves all of the important geometrical information. Here l_1, l_2 and l_3 must still satisfy triangle inequalities and add up to an even number. Otherwise, the integral over Legendre polynomials in (3.58) vanishes.

3.3.2 Optimal estimator

The bispectrum of the CMB anisotropy is a powerful statistic for studying primordial non-Gaussianity. The Planck collaboration's CMB bispectrum analyses provided the most stringent bounds on the amplitude of primordial bispectrum with respect to various shapes and constrained a wide class of inflationary models. Due to the largely linear evolution of the CMB, there are few contributions to the bispectrum from non-primordial origins compared to other probes such as the large scale structure. Here, we derive the optimal CMB bispectrum estimator using mathematical foundations laid out previously.

Consider a number of inflation models which predict non-zero primordial bispectra. We often use template bispectra which capture some common characteristics of a class of models, like the local, equilateral, and orthogonal shapes (3.10-3.12). We would like to find out if the true underlying bispectrum, if any, can be expanded in terms of the functions $B^{(i)}$ chosen;

$$B(k_1, k_2, k_3) = \sum_i f^{(i)}_{NL} B^{(i)}(k_1, k_2, k_3), \tag{3.59}$$

where the primordial non-Gaussianity (or non-linearity) parameter $f^{(i)}_{NL}$ measures the magnitude of the ith bispectrum shape found in reality. Detection of a non-zero f_{NL} would serve as strong evidence for the corresponding inflation models. Non-detection of f_{NL}, on the other hand, still allows us to place bounds on it and hence constrain models which predict larger bispectra.

Expressing (3.59) in terms of the late-time CMB anisotropies,

$$\left\langle a_{l_1m_1}^{X_1} a_{l_2m_2}^{X_2} a_{l_3m_3}^{X_3} \right\rangle = \sum_i f_{\text{NL}}^{(i)} \, \mathcal{G}_{m_1m_2m_3}^{l_1l_2l_3} \, b_{l_1l_2l_3}^{X_1X_2X_3,(i)}. \tag{3.60}$$

The goal of the CMB bispectrum estimation is to compute $f_{\text{NL}}^{(i)}$s that best describes the observed data. In reality, we can only observe one realisation of the universe and therefore a single set of a_{lm}s. The expectation values $\langle \cdot \rangle$ on the left-hand side of (3.60) are replaced by sample estimates, which introduces some errors;

$$a_{l_1m_1}^{X_1} a_{l_2m_2}^{X_2} a_{l_3m_3}^{X_3} = \sum_i f_{\text{NL}}^{(i)} \, \mathcal{G}_{m_1m_2m_3}^{l_1l_2l_3} \, b_{l_1l_2l_3}^{X_1X_2X_3,(i)} + \varepsilon_{l_1l_2l_3,m_1m_2m_3}^{X_1X_2X_3}. \tag{3.61}$$

For simplicity, we drop the X_j's from now on. The derivation of the bispectrum estimator here can easily be generalised to include both temperature and E-mode polarisation. We also define some shorthand notations for the harmonic multipole indices to improve readability;

$$\mathbf{l}_j := (l_j, m_j), \ \ L := (l_1, l_2, l_3), \ \text{and} \ \mathbf{L} := (\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3). \tag{3.62}$$

The estimation problem is summarised as follows.

$$B_{\mathbf{L}}^{\text{obs}} = \sum_i B_{\mathbf{L}}^{(i)} f_{\text{NL}}^{(i)} + \varepsilon_{\mathbf{L}}, \tag{3.63}$$

$$\text{where} \ B_{\mathbf{L}}^{\text{obs}} := a_{\mathbf{l}_1} a_{\mathbf{l}_2} a_{\mathbf{l}_3} \ \text{and} \ B_{\mathbf{L}}^{(i)} := \mathcal{G}_{\mathbf{L}} b_L^{(i)}. \tag{3.64}$$

The form of (3.63) makes it clear that the bispectrum estimation is linear regression in essence. Borrowing words from statistics, we specify the components of our analysis below.

- $\mathbf{B}^{\text{obs}}$ is the *regressand*, in our case the noisy observed bispectrum samples $B_{\mathbf{L}}^{\text{obs}}$ obtained for each *observation* $\mathbf{L}$.

- $\mathbf{B}^{(i)}$s are the *regressors*, theoretical bispectra $B_{\mathbf{L}}^{(i)}$ motivated from inflationary models.

- $f_{\text{NL}}^{(i)}$s are the *regression coefficients* which parametrise how much each regressor contributes to the regressand. Our main objective is to estimate them.

- ε is the *error term* which represents the noise in the bispectrum sample compared to the true underlying value. $\varepsilon_{\mathbf{L}}$ can be sourced by the sampling error (from having limited number of samples) and/or inaccuracies in the measurements.

We adopt the least squares method to estimate f_{NL}s from given $\mathbf{B}^{\text{obs}}$ and $B_{\mathbf{L}}^{(i)}$. Before doing so, each element of the regressand and regressors needs to be normalised so that the

expected variances in the error are constant across different observations $\mathbf{L}$. We compute the theoretical variance in $B_{\mathbf{L}}^{\text{obs}}$ under two assumptions. First, we work in a weak non-Gaussian limit where the dominant contributions to correlation functions come from the Gaussian part of $a_{\mathbf{l}}$s. Wick's theorem then reduces the problem down to summing over all possible contractions. Second, we neglect inaccuracies in the a_{lm} measurements as we are cosmic variance limited; the sampling error dominates the total error.

Under these assumptions, the expected covariance of the observed bispectra is given by

$$\left\langle B_{\mathbf{L}}^{\text{obs}} B_{\mathbf{L}'}^{\text{obs}} \right\rangle = \left\langle a_{\mathbf{l}_1} a_{\mathbf{l}_2} a_{\mathbf{l}_3} a_{\mathbf{l}'_1} a_{\mathbf{l}'_2} a_{\mathbf{l}'_3} \right\rangle \tag{3.65}$$

$$= \left[\left\langle a_{\mathbf{l}_1} a_{\mathbf{l}'_1} \right\rangle \left\langle a_{\mathbf{l}_2} a_{\mathbf{l}'_2} \right\rangle \left\langle a_{\mathbf{l}_3} a_{\mathbf{l}'_3} \right\rangle + \left\langle a_{\mathbf{l}_1} a_{\mathbf{l}'_1} \right\rangle \left\langle a_{\mathbf{l}_2} a_{\mathbf{l}'_3} \right\rangle \left\langle a_{\mathbf{l}_3} a_{\mathbf{l}'_2} \right\rangle + \cdots \right]^{\ddagger}$$

$$+ \left[\left\langle a_{\mathbf{l}_1} a_{\mathbf{l}_2} \right\rangle \left\langle a_{\mathbf{l}'_1} a_{\mathbf{l}'_2} \right\rangle \left\langle a_{\mathbf{l}_3} a_{\mathbf{l}'_3} \right\rangle + \left\langle a_{\mathbf{l}_1} a_{\mathbf{l}_2} \right\rangle \left\langle a_{\mathbf{l}'_2} a_{\mathbf{l}'_3} \right\rangle \left\langle a_{\mathbf{l}_3} a_{\mathbf{l}'_1} \right\rangle + \cdots \right]^{\ddagger\ddagger}. \tag{3.66}$$

Here, (3.66) contains all possible contractions of the 6 $a_{\mathbf{l}}$s, 3 from each of $B_{\mathbf{L}}^{\text{obs}}$ and $B_{\mathbf{L}'}^{\text{obs}}$. There are 15 such contractions: 6 consisting only of terms between the two bispectra (in $\ddagger$) and 9 which also have internal contractions ($\ddagger\ddagger$). Our aim is to modify the regression variables so that the resulting covariance matrix reduces to the identity. The error in each $\mathbf{L}$ is required to be independent and of unit variance.

The terms in ($\ddagger$) originate from symmetries present in the bispectrum; $B_{\mathbf{L}}^{\text{obs}} = a_{\mathbf{l}_1} a_{\mathbf{l}_2} a_{\mathbf{l}_3}$ is invariant under permutations of $\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3$. In order to remove duplicate elements, we restrict our $\mathbf{L}$s to ones satisfying $\mathbf{l}_1 \leq \mathbf{l}_2 \leq \mathbf{l}_3$. [3]

On the other hand, terms in ($\ddagger\ddagger$) reveal a more complex issue about our formulation. The observed anisotropies $a_{\mathbf{l}}$ are not necessarily independent since various factors such as partial sky coverage and correlated noise can induce correlations between them. Even if they are independent, our construction of $B_{\mathbf{L}}^{\text{obs}}$ can yield non-zero terms in ($\ddagger\ddagger$) when at least two out of $\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3$ are identical. This is similar, in essence, to incorrectly estimating the variance of a random variable X from samples X_i by computing $\sum_i X_i^2$ instead of the correct $\sum_i (X_i - \langle X \rangle)^2$.

We redefine the observed bispectra in terms of the $a_{\mathbf{l}}$s by subtracting off these extra contributions.

$$B_{\mathbf{L}}'^{\text{obs}} := a_{\mathbf{l}_1} a_{\mathbf{l}_2} a_{\mathbf{l}_3} - \left\langle a_{\mathbf{l}_1} a_{\mathbf{l}_2} \right\rangle a_{\mathbf{l}_3} - \left\langle a_{\mathbf{l}_2} a_{\mathbf{l}_3} \right\rangle a_{\mathbf{l}_1} - \left\langle a_{\mathbf{l}_3} a_{\mathbf{l}_1} \right\rangle a_{\mathbf{l}_2}. \tag{3.67}$$

The newly introduced terms linear in $a_{\mathbf{l}}$s have zero mean since $\langle a_{\mathbf{l}} \rangle = 0$ except for the monopole $l = 0$ which is excluded from the CMB bispectrum analysis. Thus $B_{\mathbf{L}}'^{\text{obs}}$ is still an

---

[3] $(l_1, m_1) < (l_2, m_2)$ if and only if $(l_1 < l_2)$ **or** $(l_1 = l_2$ and $m_1 < m_2)$.

unbiased estimate of the underlying bispectra. Substituting into (3.65), all terms that were previously in (‡‡) now vanish.

In theory, we can compute (3.65) using the full covariance matrix $C_{l_1 l_2} = \langle a_{l_1} a_{l_2} \rangle$ and invert it to get the least squares estimate for f_{NL}s. In practice, however, this is a costly and numerically demanding operation. [4] We instead approximate the non-diagonal covariances in (3.67) using the Monte Carlo method: $\langle a_{l_1}^G a_{l_2}^G \rangle_{MC}$ from an ensemble of realistic Gaussian simulations. In other places, we assume $C_{l_1 l_2} \approx C_{l_1} \delta_{l_1 l_2}$ to simplify our calculations.

Each element of the regressand is now independent;

$$\left\langle B'^{\text{obs}}_{\mathbf{L}} B'^{\text{obs}}_{\mathbf{L}'} \right\rangle = C_{l_1} C_{l_2} C_{l_3} \, \Delta_{\mathbf{l_1 l_2 l_3}} \, \delta_{\mathbf{l_1 l'_1}} \delta_{\mathbf{l_2 l'_2}} \delta_{\mathbf{l_3 l'_3}}, \tag{3.68}$$

where $\Delta_{\mathbf{l_1 l_2 l_3}}$ is a symmetry factor equal to 6 if $\mathbf{l}_1 = \mathbf{l}_2 = \mathbf{l}_3$, 2 if exactly two of them are identical, and 1 if all three are distinct. The regressand and regressors are rescaled so that each element of the error term has unit variance;

$$\tilde{B}^{\text{obs}}_{\mathbf{L}} := \frac{1}{\sqrt{\Delta_{\mathbf{l_1 l_2 l_3}} C_{l_1} C_{l_2} C_{l_3}}} \left[a_{\mathbf{l}_1} a_{\mathbf{l}_2} a_{\mathbf{l}_3} - \langle a_{\mathbf{l}_1} a_{\mathbf{l}_2} \rangle a_{\mathbf{l}_3} - \langle a_{\mathbf{l}_2} a_{\mathbf{l}_3} \rangle a_{\mathbf{l}_1} - \langle a_{\mathbf{l}_3} a_{\mathbf{l}_1} \rangle a_{\mathbf{l}_2} \right], \tag{3.69}$$

$$\tilde{B}^{(i)}_{\mathbf{L}} := \frac{1}{\sqrt{\Delta_{\mathbf{l_1 l_2 l_3}} C_{l_1} C_{l_2} C_{l_3}}} \mathscr{G}_{\mathbf{L}} b^{(i)}_L. \tag{3.70}$$

The ordinary least squares estimate for a linear model $\mathbf{y} = X\beta + \varepsilon$ is given by $\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$. We define the equivalent objects to $(X^T X)_{ij}$ and $(X^T \mathbf{y})_i$ in our linear model (3.63) as F_{ij} and S_i, respectively. The matrix F is given by

$$F_{ij} := \tilde{\mathbf{B}}^{(i)} \cdot \tilde{\mathbf{B}}^{(j)} \tag{3.71}$$

$$= \sum_{\mathbf{l_1 \leq l_2 \leq l_3}} \frac{\mathscr{G}_{\mathbf{L}}^2 \, b^{(i)}_L b^{(j)}_L}{\Delta_{\mathbf{l_1 l_2 l_3}} \, C_{l_1} C_{l_2} C_{l_3}} = \sum_{\mathbf{l_1, l_2, l_3}} \frac{\mathscr{G}_{\mathbf{L}}^2 \, b^{(i)}_L b^{(j)}_L}{6 \, C_{l_1} C_{l_2} C_{l_3}} = \sum_{l_1, l_2, l_3} \frac{h_L^2 \, b^{(i)}_L b^{(j)}_L}{6 \, C_{l_1} C_{l_2} C_{l_3}}. \tag{3.72}$$

The symmetry factor in (3.72) was removed by relieving the restriction $\mathbf{l}_1 \leq \mathbf{l}_2 \leq \mathbf{l}_3$ in the summation. For the last equality we used the function $h^2_{l_1 l_2 l_3}$ defined in (3.56).

Similarly, the vector S can be written as

$$S_i := \tilde{\mathbf{B}}^{(i)} \cdot \tilde{\mathbf{B}}^{\text{obs}} \tag{3.73}$$

$$= \sum_{\mathbf{l_1, l_2, l_3}} \frac{\mathscr{G}_{\mathbf{L}} b^{(i)}_L}{6 \, C_{l_1} C_{l_2} C_{l_3}} \left[a_{\mathbf{l}_1} a_{\mathbf{l}_2} a_{\mathbf{l}_3} - \langle a_{\mathbf{l}_1} a_{\mathbf{l}_2} \rangle a_{\mathbf{l}_3} - \langle a_{\mathbf{l}_2} a_{\mathbf{l}_3} \rangle a_{\mathbf{l}_1} - \langle a_{\mathbf{l}_3} a_{\mathbf{l}_1} \rangle a_{\mathbf{l}_2} \right]. \tag{3.74}$$

---

[4] For Planck-like experiments, the matrix is about $2500^2 \times 2500^2$ in size and hence costly and numerically unstable to invert.

We finally have the least squares estimate of f_{NL} in terms of F and S;

$$\hat{f}_{NL}^{(i)} = \sum_j (F^{-1})_{ij} S_j. \tag{3.75}$$

One of the main strengths of the ordinary least squares estimator lies in its *optimality*; it has the smallest possible variance among all unbiased estimators that are linear in data. That is to say, it saturates the Cramer-Rao bound. Since $\langle S_i S_j \rangle = F_{ij}$, we have $\mathrm{Var}(\hat{f}_{NL}^{(i)}) = (F^{-1})_{ii}$ (no sum over i). This value is indeed equal to the Cramer-Rao bound computed from the Fisher information matrix F here.

The Fisher matrix of the estimator naturally motivates the following definition of an inner product on the space of reduced bispectra.

$$\left\langle \mathbf{b}^{(i)}, \mathbf{b}^{(j)} \right\rangle := \sum_L \frac{h_L^2 b_L^{(i)} b_L^{(j)}}{6\, C_{l_1} C_{l_2} C_{l_3}} = F_{ij}. \tag{3.76}$$

In particular, the *correlation* between two bispectrum shapes can be defined as

$$\mathrm{Corr}\left(\mathbf{b}^{(i)}, \mathbf{b}^{(j)}\right) := \frac{\left\langle \mathbf{b}^{(i)}, \mathbf{b}^{(j)} \right\rangle}{\sqrt{\left\langle \mathbf{b}^{(i)}, \mathbf{b}^{(i)} \right\rangle \left\langle \mathbf{b}^{(j)} \mathbf{b}^{(j)} \right\rangle}}. \tag{3.77}$$

If all bispectrum shapes under consideration are uncorrelated with respect to this metric so that $|\mathrm{Corr}(\mathbf{b}^{(i)}, \mathbf{b}^{(j)})| \ll 1$ whenever $i \neq j$, then the Fisher matrix F is approximately diagonal and $(F^{-1})_{ii} \approx (F_{ii})^{-1}$ (no sum over i implied). In this case, the estimated $\hat{f}_{NL}^{(i)}$s in (3.75) are identical to the values obtained using a single regressor $\mathbf{B}^{(i)}$. In other words, we may analyse individual bispectrum shapes independently.

We write down the estimator for single shape analysis and restore the shortened indices;

$$\hat{f}_{NL} = \frac{1}{N} \sum_{l_j, m_j} \frac{\mathscr{G}_{m_1 m_2 m_3}^{l_1 l_2 l_3} b_{l_1 l_2 l_3}}{C_{l_1} C_{l_2} C_{l_3}} \left[a_{l_1 m_1} a_{l_2 m_2} a_{l_3 m_3} - 3 \left\langle a_{l_1 m_1} a_{l_2 m_2} \right\rangle a_{l_3 m_3} \right], \tag{3.78}$$

$$N := 6F = \sum_{l_j, m_j} \frac{h_{l_1 l_2 l_3}^2 b_{l_1 l_2 l_3}^2}{C_{l_1} C_{l_2} C_{l_3}}, \tag{3.79}$$

where the symmetry in indices was used to combine the linear terms into one. We also defined the normalisation factor N to follow conventions in the literature. The theoretical error on $\hat{f}_{NL}$ is then $\sigma(\hat{f}_{NL}) = \sqrt{F^{-1}} = \sqrt{6/N}$.

3.3.3 CMB bispectrum estimators

The estimator (3.78) can be intimidating. The multipole moments l can go up to 2500 for the Planck survey [41], meaning that the sum over all possible l_j, m_j contains $\approx 2.7 \times 10^{15}$ terms.[5] Direct evaluation of the Gaunt integral $\mathscr{G}^{l_1 l_2 l_3}_{m_1 m_2 m_3}$ involves calculating Wigner 3-j symbols, which are expensive to compute and store. Not to mention that the number of terms scales as $\propto l_{\max}^5$.

All known CMB bispectrum estimation methods therefore deploy some ingenious techniques to reduce the computational cost. In this section, we review the three most common approaches: KSW [50, 58], Modal [59, 60], and Binned [61, 62].

KSW estimator

The Komatsu-Spergel-Wandelt (KSW) estimator was first introduced in [58] together with the construction of the bispectrum estimator and has been studied extensively since then [50, 63, 64] (see e.g, [22] for reviews). The core idea is to exploit the *separability* of the bispectrum. Consider the local shape for example:

$$S^{\text{local}}(k_1, k_2, k_3) = 2A_\Phi^2 \left(\frac{k_1^2}{k_2 k_3} + \frac{k_2^2}{k_3 k_1} + \frac{k_3^2}{k_1 k_2} \right), \tag{3.80}$$

where A_Φ is the amplitude of the power spectrum of Φ when the scalar spectral index n_s is set to 1. Note that each term can be expressed as a product of three separate functions that only depend on one of the variables: $k_1^2/(k_2 k_3) = k_1^2 \cdot k_2^{-1} \cdot k_3^{-1}$, for example. The reduced bispectrum then simplifies to an integral of separable terms;

$$b^{\text{local}}_{l_1 l_2 l_3} = 2A^2 \int dr\, r^2 \left[\alpha_{l_1}(r)\beta_{l_2}(r)\beta_{l_3}(r) + \beta_{l_1}(r)\alpha_{l_2}(r)\beta_{l_3}(r) + \beta_{l_1}(r)\beta_{l_2}(r)\alpha_{l_3}(r) \right], \tag{3.81}$$

where

$$\alpha_l(r) := \frac{2}{\pi} \int dk\, k^2 \Delta_l(k) j_l(kr), \tag{3.82}$$

$$\beta_l(r) := \frac{2}{\pi} \int dk\, k^{-1} \Delta_l(k) j_l(kr). \tag{3.83}$$

---

[5]This number was calculated considering the symmetry ($l_1 \geq l_2 \geq l_3$), triangle inequality ($l_2 + l_3 \geq l_1$), and restrictions on m_js from angular momentum conservation ($m_1 + m_2 + m_3 = 0$).

Substituting (3.81) into the bispectrum estimator (3.78) gives

$$\hat{f}_{\text{NL}}^{\text{local}} = \frac{6A^2}{N} \int dr \, r^2 \int d^2\hat{\mathbf{n}} \left[A(r,\hat{\mathbf{n}})B(r,\hat{\mathbf{n}})^2 - 2\left\langle A(r,\hat{\mathbf{n}})B(r,\hat{\mathbf{n}})\right\rangle A(r,\hat{\mathbf{n}}) - \left\langle B(r,\hat{\mathbf{n}})^2\right\rangle A(r,\hat{\mathbf{n}}) \right],$$
(3.84)

where we have defined the *filtered maps*

$$A(r,\hat{\mathbf{n}}) := \sum_{l,m} \frac{\alpha_l(r)}{C_l} a_{lm} Y_{lm}(\hat{\mathbf{n}}),$$
(3.85)

$$B(r,\hat{\mathbf{n}}) := \sum_{l,m} \frac{\beta_l(r)}{C_l} a_{lm} Y_{lm}(\hat{\mathbf{n}}).$$
(3.86)

The estimator (3.84) has significantly lower computational complexity compared to the original form. The integral over $\hat{\mathbf{n}}$ becomes a summation over map pixels, roughly 50 million in Planck-like settings. The filtered maps (3.85-3.86) are efficiently obtained using Spherical Harmonic Transforms (SHTs).

We are now only left with the normalisation factor. An integral representation of $h_{l_1 l_2 l_3}^2$ in (3.58) allows a fast computation of N, as introduced in [65];

$$N = (2A^2)^2 \int d\mu \int dr \, r^2 \int dr' \, r'^2 \left[3R_{\alpha\alpha}^2 R_{\beta\beta} + 6R_{\alpha\beta}R_{\beta\alpha}R_{\alpha\alpha} \right],$$
(3.87)

where

$$R_{XY}(r,r',\mu) := \sum_l \frac{2l+1}{(8\pi)^{2/3}} X_l(r)Y_l(r')P_l(\mu)$$
(3.88)

for $X, Y = \alpha, \beta$.

The KSW formalism can be used to constrain various other separable shapes by replacing α_l and β_l with appropriate functions, which we refer to as *modes*. The equilateral and orthogonal shapes, for example, call for four such functions with $k^{-1}, 1, k$, and k^2 in place of k^2 in the integral (3.82). Constant feature models with shape function $S(k_1, k_2, k_3) = \sin(\omega(k_1 + k_2 + k_3) + \phi)$ can be written in terms of the two modes $\sin(\omega k)$ and $\cos(\omega k)$ via trigonometric identities [66].

In some cases, a non-separable function can be expanded to separable ones using an analytic formula. For instance, the Schwinger parametrisation

$$\frac{1}{(k_1 + k_2 + k_3)^n} = \frac{1}{(n-1)!} \int_0^\infty du \, u^{n-1} e^{-u(k_1 + k_2 + k_3)}$$
(3.89)

allows us to approximate the left-hand side with a sum over separable terms parametrised by u [65]. These numerical tricks, however, are restricted to cases where the integrand is relatively well-behaved. Otherwise, we require a large number of terms for an accurate expansion, and the trick causes a net increase in computation time instead of a net decrease.

Modal estimator

The KSW estimator is fast and numerically stable but restrictive in the type of bispectra it can tackle. The Modal estimator, first developed in [59] and expanded much further through [60, 67–69], builds on a simple but effective idea to address this issue; when the bispectrum shape cannot be factorised, we may instead *expand* it using a basis constructed from separable functions.

The essence of Modal estimator is captured in the following 'modal' expansion;

$$\frac{v_{l_1} v_{l_2} v_{l_3}}{\sqrt{C_{l_1} C_{l_2} C_{l_3}}} \, b_{l_1 l_2 l_3} = \sum_{n \leftrightarrow (p_1, p_2, p_3)} \alpha_n^Q Q_{n l_1 l_2 l_3}. \tag{3.90}$$

The left-hand side is the reduced bispectrum, rescaled using C_ls and $v_l := (2l+1)^{1/6}$ for later convenience. On the right-hand side is the mode expansion, where each triplet (p_1, p_2, p_3) is associated with a number n. α_n^Q is the expansion coefficient with respect to the basis function Q, which is defined as

$$Q_{n \, l_1 l_2 l_3} := \frac{1}{6} \left[q_{p_1}(l_1) q_{p_2}(l_2) q_{p_3}(l_3) + q_{p_1}(l_1) q_{p_2}(l_3) q_{p_3}(l_2) + \cdots \right], \tag{3.91}$$

with appropriate mode functions $q_p(l)$. Polynomials and Fourier modes are common choices for these mode functions, but the formalism itself is completely general. All we require is that the modal expansion (3.90) accurately describes the given bispectrum.

As seen from the KSW estimator, separability greatly simplifies (3.78). We define the filtered maps analogous to the KSW ones up to some factors;

$$M_p(\hat{\mathbf{n}}) = \sum_{l,m} \frac{q_p(l)}{v_l \sqrt{C_l}} a_{lm} Y_{lm}(\hat{\mathbf{n}}). \tag{3.92}$$

Note that we do not have the line-of-sight integral r appearing in the KSW formalism since the mode expansion was performed in the late-time l space, instead of the primordial k space.

The estimator (3.78) now becomes

$$\hat{f}_{\rm NL} = \frac{1}{N} \sum_{n \leftrightarrow (p_1, p_2, p_3)} \alpha_n^Q \int d^2\mathbf{n} \left[M_{p_1}(\hat{\mathbf{n}}) M_{p_2}(\hat{\mathbf{n}}) M_{p_3}(\hat{\mathbf{n}}) - 3 \left\langle M_{p_1}(\hat{\mathbf{n}}) M_{p_2}(\hat{\mathbf{n}}) \right\rangle M_{p_3}(\hat{\mathbf{n}}) \right].$$

$$(3.93)$$

One of the Modal pipeline's key objectives is to compute

$$\beta_n^Q := \int d^2\mathbf{n} \left[M_{p_1}(\hat{\mathbf{n}}) M_{p_2}(\hat{\mathbf{n}}) M_{p_3}(\hat{\mathbf{n}}) - 3 \left\langle M_{p_1}(\hat{\mathbf{n}}) M_{p_2}(\hat{\mathbf{n}}) \right\rangle M_{p_3}(\hat{\mathbf{n}}) \right], \qquad (3.94)$$

so that the estimator simply becomes a dot product of α and β: $\hat{f}_{\rm NL} = \frac{1}{N} \sum_n \alpha_n^Q \beta_n^Q$. Note that β_n^Q depends on the observed data and choice of basis functions, but is independent of theoretical model in consideration. This is an important property of the Modal estimator; the time-consuming integral of (3.94) only needs to be performed once per dataset. We can then constrain a wide range of models simultaneously by computing α_n^Q for each bispectrum shape. The modal decomposition costs much less than the $f_{\rm NL}$ estimation in general.

Before moving on to calculating the normalisation factor, we first modify the inner product defined in (3.76) as

$$\left\langle \mathbf{b}^{(i)}, \mathbf{b}^{(j)} \right\rangle := \sum_{l_j} \frac{h_{l_1 l_2 l_3}^2}{v_{l_1}^2 v_{l_2}^2 v_{l_3}^2} \, b_{l_1 l_2 l_3}^{(i)} b_{l_1 l_2 l_3}^{(j)}. \qquad (3.95)$$

The weights appearing in the inner product above, $(h_{l_1 l_2 l_3}/v_{l_1} v_{l_2} v_{l_3})^2 \approx const.$, are now nearly uniform across the allowed l configurations [59]. This serves as a natural inner product for basis functions $\mathbf{Q}_n$, especially for polynomial modes.

Expanding the bispectrum in (3.79) gives

$$N = \sum_{n_1} \sum_{n_2} \alpha_{n_1}^Q \alpha_{n_2}^Q \left(\sum_{l_j} \frac{h_{l_1 l_2 l_3}^2}{v_{l_1}^2 v_{l_2}^2 v_{l_3}^2} \, Q_{n_1 \, l_1 l_2 l_3} Q_{n_2 \, l_1 l_2 l_3} \right) \qquad (3.96)$$

$$= \sum_{n_1} \sum_{n_2} \alpha_{n_1}^Q \alpha_{n_2}^Q \left\langle \mathbf{Q}_{n_1}, \mathbf{Q}_{n_2} \right\rangle. \qquad (3.97)$$

Another key quantity to be computed in the Modal formalism is the matrix

$$\gamma_{n_1 n_2} := \left\langle \mathbf{Q}_{n_1}, \mathbf{Q}_{n_2} \right\rangle. \qquad (3.98)$$

Even if the mode functions $q_p(l)$ are chosen to be orthogonal, the three-dimensional basis functions $\mathbf{Q}_n$ are not necessarily orthogonal with respect to the inner product. Once γ is

computed, however, we may transform our basis functions to become orthonormal. By definition γ is symmetric and positive semi-definite. As long as we choose the basis $\mathbf{Q}_n$s to be linearly independent, γ is always non-degenerate and hence invertible. Since γ^{-1} is also symmetric and positive-definite, we may perform a Cholesky decomposition on it;

$$\gamma^{-1} = \lambda \lambda^T, \tag{3.99}$$

for some lower triangular matrix λ. We now define a new set of basis

$$\mathbf{R}_t := \sum_n \lambda_{nt} \, \mathbf{Q}_n. \tag{3.100}$$

The new basis functions are now orthonormal with respect to the inner product $\langle \cdot, \cdot \rangle$;

$$\langle \mathbf{R}_{t_1}, \mathbf{R}_{t_2} \rangle = \sum_{n_1, n_2} \lambda_{n_1 t_1} \langle \mathbf{Q}_{n_1}, \mathbf{Q}_{n_2} \rangle \lambda_{n_2 t_2} = (\lambda^T \gamma \lambda)_{t_1 t_2} = \delta_{t_1 t_2}. \tag{3.101}$$

Orthonormality of the basis is especially useful for the modal decomposition. Taking the inner product with $\mathbf{R}_t$ in the modal expansion (3.90) for R,

$$\alpha_t^R = \left\langle \frac{v_{l_1} v_{l_2} v_{l_3}}{\sqrt{C_{l_1} C_{l_2} C_{l_3}}} \, b_{l_1 l_2 l_3} \, , \, \mathbf{R}_t \right\rangle. \tag{3.102}$$

After we obtain α_t^R, the expansion coefficients for $\mathbf{Q}_n$ can be found by

$$\alpha_n^Q = \sum_t \lambda_{nt} \alpha_t^R, \tag{3.103}$$

from which it is straightforward to check $\sum_n \alpha_n^Q \mathbf{Q}_n = \sum_t \alpha_t^R \mathbf{R}_t$. There is no new information gained from converting our basis from $\mathbf{Q}$ to $\mathbf{R}$; it is simply a change of basis.

The normalisation factor in (3.97) further simplifies when we use R;

$$N = \sum_{t_1} \sum_{t_2} \alpha_{t_1}^R \alpha_{t_2}^R \langle \mathbf{R}_{t_1}, \mathbf{R}_{t_2} \rangle = \sum_t \left| \alpha_t^R \right|^2 \tag{3.104}$$

So far, we have been working on the late-time harmonic space directly related to the CMB observations. The theoretical bispectrum $B(k_1, k_2, k_3)$ predicted by inflation models, however, lies within the primordial Fourier space. The remaining work is about how to bridge this gap. In particular, we would like to obtain α^R from a given shape function $S(k_1, k_2, k_3) = (k_1 k_2 k_3)^2 B(k_1, k_2, k_3)$ so that we can get an f_{NL} estimate for $S(k_1, k_2, k_3)$.

The primordial and late-time spaces are formulated similarly in the Modal approach. We expand the shape function using an independent set of *primordial* basis functions;

$$S(k_1, k_2, k_3) = \sum_n \bar{\alpha}_n^{\bar{Q}} \bar{Q}_n(k_1, k_2, k_3), \qquad \text{where} \tag{3.105}$$

$$\bar{Q}_n(k_1, k_2, k_3) := \frac{1}{6} \left[\bar{q}_{p_1}(k_1) \bar{q}_{p_2}(k_2) \bar{q}_{p_3}(k_3) + \bar{q}_{p_1}(k_1) \bar{q}_{p_2}(k_3) \bar{q}_{p_3}(k_2) + \cdots \right]. \tag{3.106}$$

A bar is placed above each variable to denote that they are primordial quantities. The inner product is defined following [59];

$$\left\langle S^{(i)}, S^{(j)} \right\rangle := \int_{V_{\mathbf{k}}} dk_1 dk_2 dk_3 \, \frac{1}{k_1 + k_2 + k_3} \, B^{(i)}(k_1, k_2, k_3) B^{(j)}(k_1, k_2, k_3). \tag{3.107}$$

The mathematical formulation from here on is identical to that of the late-time case. We therefore simply write down the primordial counterparts;

$$\bar{\gamma}_{n_1 n_2} := \left\langle \bar{\mathbf{Q}}_{n_1}, \bar{\mathbf{Q}}_{n_2} \right\rangle, \tag{3.108}$$

$$\bar{\gamma}^{-1} = \bar{\lambda} \bar{\lambda}^T, \tag{3.109}$$

$$\bar{\mathbf{R}}_t := \sum_n \bar{\lambda}_{nt} \, \bar{\mathbf{Q}}_n, \tag{3.110}$$

$$\bar{\alpha}_t^{\bar{R}} = \left\langle S(k_1, k_2, k_3) \, , \, \bar{\mathbf{R}}_t \right\rangle, \tag{3.111}$$

$$\bar{\alpha}_n^{\bar{Q}} = \sum_t \bar{\lambda}_{nt} \bar{\alpha}_t^{\bar{R}}. \tag{3.112}$$

A key step in connecting primordial results to late-time is the projection via transfer functions (3.50). Projecting $\bar{Q}_n(k_1, k_2, k_3)$ yields

$$\tilde{Q}_{n \, l_1 l_2 l_3} := \left(\frac{2}{\pi} \right)^3 \int dr dk_1 dk_2 dk_3 \, (r k_1 k_2 k_3)^2 \, \bar{Q}_n(k_1, k_2, k_3) \prod_{j=1}^3 \left[j_{l_j}(k_j r) \Delta_{l_j}(k_j) \right], \tag{3.113}$$

where a tilde indicates that it is a projected quantity, as opposed to a bar for primordial quantities and none for inherently late-time variables.

The following inner product serves as the bridge between objects defined in the early and late universe:

$$\Gamma_{tn} = \left\langle \mathbf{R}_t, \tilde{\mathbf{Q}}_n \right\rangle. \tag{3.114}$$

By the virtue of the projected bispectrum's dual representation,

$$\sum_t \alpha_t^R R_{n\,l_1 l_2 l_3} = \frac{v_{l_1} v_{l_2} v_{l_3}}{\sqrt{C_{l_1} C_{l_2} C_{l_3}}} \, b_{l_1 l_2 l_3} = \sum_n \bar{\alpha}_n^{\bar{Q}} \tilde{Q}_{n\,l_1 l_2 l_3}. \tag{3.115}$$

Taking an inner product with $\mathbf{R}_t$ on both sides gives

$$\alpha_t^R = \sum_n \Gamma_{tn} \bar{\alpha}_n^{\bar{Q}}. \tag{3.116}$$

The matrix Γ therefore lets us convert the primordial modal expansion coefficients $\bar{\alpha}_n^{\bar{Q}}$ to the late time ones α_t^R.

We have discussed the inner workings of the Modal estimator in great detail. Despite the long list of formulae here, its core idea is captured in the modal decomposition (3.90). The main computational challenge in the Modal approach is precomputing β^Q, which contains complete information about the fit between the basis and the observed bispectrum. Afterwards, the bispectrum estimation problem reduces down to a matter of finding the modal coefficients α for models under consideration. This decomposition is performed fast and efficiently using the orthonormal basis functions $\mathbf{R}$ and $\bar{\mathbf{R}}$.

Binned estimator

The binned bispectrum estimator, introduced in [61] and used for Planck analyses [23, 30, 62, 70], uses carefully chosen bins to reduce the computational complexity. In this section, we highlight the main concepts of its formalism while relabelling some notations in the original literature [61] in order to be consistent with our previous discussions.

The estimator (3.74) can be rewritten as

$$S_i = \sum_{l_1 l_2 l_3} \frac{b_{l_1 l_2 l_3}^{(i)}}{6\,C_{l_1} C_{l_2} C_{l_3}} \sum_{m_1 m_2 m_3} \mathcal{G}_{m_1 m_2 m_3}^{l_1 l_2 l_3} \left[a_{l_1 m_1} a_{l_2 m_2} a_{l_3 m_3} - \left(\langle a_{l_1 m_1} a_{l_2 m_2} \rangle a_{l_3 m_3} + 2\,\text{cyc.} \right) \right] \tag{3.117}$$

$$= \sum_{l_1 l_2 l_3} \frac{b_{l_1 l_2 l_3}^{(i)}}{6\,C_{l_1} C_{l_2} C_{l_3}} \int d^2 \hat{\mathbf{n}} \left[M_{l_1}(\hat{\mathbf{n}}) M_{l_2}(\hat{\mathbf{n}}) M_{l_3}(\hat{\mathbf{n}}) - \left(\langle M_{l_1}(\hat{\mathbf{n}}) M_{l_2}(\hat{\mathbf{n}}) \rangle M_{l_3}(\hat{\mathbf{n}}) + 2\,\text{cyc.} \right) \right], \tag{3.118}$$

where M_ls analogous to the filtered maps of the KSW and Modal estimators are defined as

$$M_l(\hat{\mathbf{n}}) := \sum_{m=-l}^{l} a_{lm} Y_{lm}(\hat{\mathbf{n}}). \tag{3.119}$$

They correspond to the l-multipole component of the original map. We define the observed bispectrum $B_{l_1 l_2 l_3}^{\text{obs}}$ to be the integrand of (3.118).[6]

The core idea of this formulation is based on the fact that a_{lm}s with similar ls are distributed in a comparable fashion and can be binned without big loss of information. The multipole moments l in the range $[l_{\min}, l_{\max}]$ are divided into N_{bin} intervals $\Delta_i := [l_i, l_{i+1} - 1]$, where $i = 0, 1, \cdots, N_{\text{bin}} - 1$. The filtered map corresponding to the ith bin is given by

$$M_i(\hat{\mathbf{n}}) := \sum_{l \in \Delta_i} \sum_{m=-l}^{l} a_{lm} Y_{lm}(\hat{\mathbf{n}}). \tag{3.120}$$

The binned bispectrum take the form

$$B_{i_1 i_2 i_3}^{\text{obs}} = \frac{1}{\Xi_{i_1 i_2 i_3}} \int d^2\hat{\mathbf{n}} \left[M_{i_1}(\hat{\mathbf{n}}) M_{i_2}(\hat{\mathbf{n}}) M_{i_3}(\hat{\mathbf{n}}) - \left(\langle M_{i_1}(\hat{\mathbf{n}}) M_{i_2}(\hat{\mathbf{n}}) \rangle M_{i_3}(\hat{\mathbf{n}}) + 2 \text{ cyc.} \right) \right], \tag{3.121}$$

where $\Xi_{i_1 i_2 i_3}$ counts the number of allowed (l_1, l_2, l_3) triplets in the bin $(\Delta_{i_1}, \Delta_{i_2}, \Delta_{i_3})$ satisfying the triangle inequality and parity condition $l_1 + l_2 + l_3 \in 2\mathbb{Z}$. We further impose that i_1, i_2, and i_3 are ordered in a way that $i_1 \leq i_2 \leq i_3$.

We now compute the expected variance of the binned bispectrum. Assuming that the covariance matrix of the a_{lm}s is diagonal, calculations similar to (3.65) yield

$$\langle B_{i_1 i_2 i_3}^{\text{obs}} B_{i_1', i_2', i_3'}^{\text{obs}} \rangle = \delta_{i_1 i_1'} \delta_{i_2 i_2'} \delta_{i_3 i_3'} \frac{g_{i_1 i_2 i_3}}{(\Xi_{i_1 i_2 i_3})^2} \sum_{l_1 \in \Delta_{i_1}} \sum_{l_2 \in \Delta_{i_2}} \sum_{l_3 \in \Delta_{i_3}} h_{l_1 l_2 l_3}^2 C_{l_1} C_{l_2} C_{l_3} \tag{3.122}$$

$$=: \delta_{i_1 i_1'} \delta_{i_2 i_2'} \delta_{i_3 i_3'} V_{i_1 i_2 i_3}, \tag{3.123}$$

for some V_{i_1, i_2, i_3}. The symmetry factor $g_{i_1 i_2 i_3} = 6, 2$, and 1 for 3, 2, and no identical numbers within i_1, i_2, i_3, respectively.

We write the theoretical binned bispectrum as

$$B_{i_1 i_2 i_3}^{\text{th}} = \frac{1}{\Xi_{i_1 i_2 i_3}} \sum_{l_1 \in \Delta_{i_1}} \sum_{l_2 \in \Delta_{i_2}} \sum_{l_3 \in \Delta_{i_3}} h_{l_1 l_2 l_3}^2 h_{l_1 l_2 l_3}^2 b_{l_1 l_2 l_3}^{\text{th}}, \tag{3.124}$$

---

[6]This is the convention from the original literature [61], which is different to some conventions elsewhere by a factor of $h_{l_1 l_2 l_3}$. In terms of the theoretical reduced bispectrum $b_{l_1 l_2 l_3}$, we have $\langle B_{l_1 l_2 l_3}^{\text{obs}} \rangle = h_{l_1 l_2 l_3}^2 b_{l_1 l_2 l_3}^{\text{th}}$.

where b^{th} is the usual reduced bispectrum. Then the binned estimator takes the form

$$\hat{f}_{\text{NL}} = \frac{\langle \mathbf{B}^{\text{obs}}, \mathbf{B}^{\text{th}} \rangle}{\langle \mathbf{B}^{\text{th}}, \mathbf{B}^{\text{th}} \rangle}, \quad \text{where} \tag{3.125}$$

$$\langle \mathbf{B}^{(1)}, \mathbf{B}^{(2)} \rangle = \sum_{i_1 \le i_2 \le i_3} \frac{B^{(1)}_{i_1 i_2 i_3} B^{(2)}_{i_1 i_2 i_3}}{V_{i_1 i_2 i_3}}. \tag{3.126}$$

The ordering $i_1 \le i_2 \le i_3$ can be removed similarly as before by replacing $g_{i_1 i_2 i_3}$ in the denominator of (3.126) with 6.

Binning different multipoles together inevitably causes some information loss and hence increases the variance of the final estimate. This effect is quantified by the ratio

$$R := \frac{\text{Var}(\hat{f}_{\text{NL}}^{\text{ideal}})}{\text{Var}(\hat{f}_{\text{NL}}^{\text{binned}})} = \frac{\langle \mathbf{B}^{\text{th}}, \mathbf{B}^{\text{th}} \rangle^{\text{binned}}}{\langle \mathbf{B}^{\text{th}}, \mathbf{B}^{\text{th}} \rangle^{\text{no binning}}}, \tag{3.127}$$

which always lies between 0 and 1. As long as R is close to 1, the binned estimator performs just as well as the full bispectrum estimator.

The main advantage of the binned estimator is the significant reduction in computational complexity. Using optimised binning strategies outlined in [62], this can be done with minimal loss of information. The Binning formalism works particularly well for models with smooth bispectra or features in l-space. The full binned bispectrum can also be used for non-parametric studies of non-Gaussianity from observations after certain smoothing operations. It is also worth noting that the most computationally expensive part of the estimation process—obtaining $B^{\text{obs}}_{i_1 i_2 i_3}$—only needs to be done once per dataset. The result can then be used to constrain various models instantaneously as long as their theoretical bispectra are provided.

Meanwhile, the binned estimator is not as effective for constraining models with general features or oscillations. This is not only because the binning smooths out high-frequency oscillations in the bispectrum, but also since evaluating the theoretical bispectra in l space through (3.124) becomes numerically challenging. The reduced bispectrum needs to be computed for every (l_1, l_2, l_3) triplet and then summed over within each bin. Such computation of the bispectrum is often practically intractable unless the given shape is separable. For the same reason, the binned estimator struggles to constrain non-separable shapes, even though the formalism itself applies to any shape.

3.3.4 Other sources of non-Gaussianity.

The CMB is one of the cleanest probes of the primordial bispectrum. CMB anisotropy remains linear in initial perturbations to a very good approximation so that its bispectrum contains a little contribution from non-linear evolution. Nevertheless, there are non-primordial sources of non-Gaussianity that require close attention. In this section, we discuss both physical and experimental complications to the CMB bispectrum estimation. Instrumental effects including the measurement noise and the width of the probing beam will be covered in the next chapter (Section 4.2.4).

Lensing bispectrum

After its departure from the last scattering surface, the CMB photons travel through the perturbed universe until they reach us. Overdense regions on their path act as gravitational lenses and alter the trajectories, which leave observable imprints on the CMB map. Overdensities magnify the CMB around them and thus shift the power spectrum towards larger scales. The deflection α away from observation angle $\hat{\mathbf{n}}$ can be computed from the line-of-sight integral given by

$$\alpha = -2 \int_0^{\chi_*} d\chi \, \frac{\chi_* - \chi}{\chi_* \chi} \nabla_{\hat{\mathbf{n}}} \Psi_{\mathrm{W}}(\chi \hat{\mathbf{n}}; \eta_0 - \chi), \tag{3.128}$$

where the Weyl potential $\Psi_{\mathrm{W}} = (\Psi + \Phi)/2$ is defined as the average of two Newtonian potentials. The integral is performed along the unperturbed path, an assumption known as the Born approximation. This is valid as long as the deflection angle α remains small. The angular derivative $\nabla_{\hat{\mathbf{n}}}$ is taken orthogonal to $\hat{\mathbf{n}}$ and within the tangent plane on the sphere. The integration weight quantifies how earlier deflections cause greater displacements in the sky today. However, most contributions to the integral in fact come from the late-time universe with redshift $z \lesssim 2$. We define the *lensing potential* as

$$\psi := -2 \int_0^{\chi_*} d\chi \, \frac{\chi_* - \chi}{\chi_* \chi} \Psi_{\mathrm{W}}(\chi \hat{\mathbf{n}}; \eta_0 - \chi), \tag{3.129}$$

so that $\alpha(\hat{\mathbf{n}}) = \nabla_{\hat{\mathbf{n}}} \psi$ in our regime. ψ is a secondary observable of the CMB which opens up a whole field by itself. We refer to [71–73] for detailed reviews on the topic of CMB lensing, including derivations of the equations above.

Another significant late-time contribution to CMB anisotropy comes from the Integrated Sachs-Wolfe (ISW) effect. This is the last term in (2.65) with a simple relabelling of variables:

$$\Theta_{\text{ISW}}(\eta_0, \mathbf{x}_0, \hat{\mathbf{n}}) = 2 \int_0^{\chi_*} d\chi \, \Psi'_W(\chi \hat{\mathbf{n}}; \eta_0 - \chi).$$ (3.130)

The integrand is small except during the accelerated expansion phase at low redshift, again $z \lesssim 2$. The two integrals (3.129) and (3.130) are highly correlated through Ψ_W and thus generate a non-trivial bispectrum, mainly in the squeezed limit. This lensing-ISW bispectrum is well approximated by the template given by [20];

$$b_{l_1 l_2 l_3}^{\text{lens}} = \frac{1}{2} \left[l_1(l_1 + 1) - l_2(l_2 + 1) + l_3(l_3 + 1) \right] \tilde{C}_{l_1}^{TT} C_{l_3}^{T\psi} + 5 \text{ perms.,}$$ (3.131)

where $\tilde{C}_l^{TT}$ is the *lensed* angular power spectrum and $C_l^{T\psi}$ is the cross power spectrum between temperature (T) and ψ. Assuming the standard ΛCDM cosmology we expect to find the bispectrum above in the CMB without any additional factors. We can measure its amplitude in the CMB data through the usual bispectrum estimator (3.78)-(3.79) by setting $\mathbf{b} = \mathbf{b}^{\text{lens}}$ so that $\hat{f}_{\text{NL}}^{\text{lens}} = \langle \mathbf{b}^{\text{lens}}, \mathbf{b}^{\text{obs}} \rangle / \langle \mathbf{b}^{\text{lens}}, \mathbf{b}^{\text{lens}} \rangle$, where we used the inner product $\langle \cdot, \cdot \rangle$ from (3.76). This value has been estimated in the Planck 2013 analysis to a 2.5σ level [74]. The most recent Planck analysis estimates $f_{\text{NL}}^{\text{lens}} = 0.73 \pm 0.27$ (68% CL) using the SMICA foreground-cleaned CMB temperature maps [23].

The lensing-ISW bispectrum is an observable originating from the non-linear evolution of the CMB. When studying primordial contributions to the observed bispectrum, however, such late-time effects need to be subtracted. The lensing-ISW *bias* is given by

$$\Delta f_{\text{NL}}^{(t)} = \frac{\langle \mathbf{b}^{(t)}, \mathbf{b}^{\text{lens}} \rangle}{\langle \mathbf{b}^{(t)}, \mathbf{b}^{(t)} \rangle}$$ (3.132)

for the bispectrum template $\mathbf{b}^{(t)}$ of interest. This bias can be significant for templates with a large squeezed limit, such as the local template, since they are highly correlated with the lensing-ISW bispectrum (3.131).

Note that CMB lensing also affects the bispectrum estimation by increasing the estimator's covariance [75]. The covariance of the bispectrum estimator, which is a six-point function of the CMB anisotropies, gets an extra contribution through the connected four-point component from gravitational lensing. This effect has been small but will be important for the future CMB surveys with increased resolution power.

Sky mask and inpainting

CMB measurements are affected by various foreground contaminants. A number of component separation methods such as SMICA [76], NILC [77], SEVEM [78], and Commander [79] have been developed to clean unwanted signals from data [80], but some regions with an overwhelming amount of contaminations need to be *masked* away. The corresponding parts of the sky are simply removed from any analysis. Even in a full-sky survey like Planck, only $\approx 78\%$ of the sky can be used for science [80].

The partial sky coverage due to masks induces multiple complications. For the CMB bispectrum analysis, there are two main issues to manage. First, the statistical isotropy of the data is broken due to the irregular mask shape. The covariance matrix $C_{l_1 m_1, l_2, m_2} = \langle a^*_{l_1 m_1} a_{l_2 m_2} \rangle$ is no longer diagonal. Using the full covariance matrix can be costly for bispectrum analysis, but the linear part of the bispectrum estimator introduced in (3.67) can handle most of the errors caused by off-diagonal covariances.

Second, the sky mask generates a non-vanishing bispectrum by itself which introduces additional noise in the estimation. Sharp cuts around the edges of the mask induce some high-frequency signals. A low-frequency mode from the overall mask shape then correlates with them to produce bispectrum in squeezed configurations. We follow [81] and utilise a simple inpainting method to reduce the effect. In this approach, a linear diffusive method is used to interpolate and hence smoothen the mask edges; at each iteration, every pixel within previously masked regions takes the average value of its neighbouring pixels. The resulting inpainted map after a sufficient number of iterations is shown to contain little excess contamination from the mask.

Lastly, we note that not all contaminants can be masked away, especially if they are localised and scattered. As detailed in [70], extra-Galactic point sources such as synchrotron emission and thermal emission from dust appear in the measured CMB data. Their reduced bispectrum is expected to be constant. Similarly to the case of the lensing bispectrum, these bispectrum contributions from point sources bias the f_{NL} estimates and therefore need to be subtracted off. We use the bias given in (3.132) with $\mathbf{b}^{\mathrm{PS}} = 1$ instead of $\mathbf{b}^{\mathrm{lens}}$ to account for this effect.

Summary

In this chapter, we introduced the theory of the bispectrum and explained why it is a key quantity for testing inflation against the CMB data. We first outlined how to compute the primordial bispectrum from the inflationary Lagrangian using the in-in formalism. The bispectrum is expected to be unobservably small for the simplest single-field, slow-roll inflation with a canonical kinetic term and Bunch-Davies vacuum. However, we showed using an example of a non-canonical kinetic term that breaking some of these assumptions can induce detectable non-Gaussianity with distinct shapes. Moving on to the late-time universe, we reviewed the theory of CMB bispectrum estimation while keeping an analogy to linear regression in mind. It was noted that the high-dimensional integrals appearing in the bispectrum estimator are computationally expensive and numerically challenging. Exploiting separability is the key to reducing computational complexity. We detailed the three main estimators used in the latest Planck analysis: KSW, Modal and binned.

No observations so far have found statistically significant non-Gaussianity of primordial origin. In the most recent Planck analysis, however, there were some interesting 'hints' from some oscillatory shapes that appear in the feature and resonance type models. The signal has not yet been decisive after accounting for the look-elsewhere effect [23, 82]. The constraining power of these oscillatory models has benefited considerably from the inclusion of the E-mode polarisation data in the Planck analysis and is expected to improve further in future CMB experiments with enhanced polarisation sensitivity. We present a forecast for oscillatory models in Chapter 4.

Chapter 4

CMB Stage-4 Forecast

The Cosmic Microwave Background (CMB) radiation is one of our most valuable probes of the primordial universe. The temperature and polarisation of this ancient light contain rich statistical information both about the primordial perturbations created during inflation and also their subsequent evolution until now. This allows us to test our inflationary theories and also the history of our universe. The recent Planck CMB experiments have provided stringent tests on various models of inflation through the estimation of cosmological parameters and via primordial non-Gaussianity [23, 30].

The simplest model of inflation involves a single scalar field slowly rolling down a smooth potential. In this case, the CMB temperature fluctuations are expected to be Gaussian distributed with only tiny deviations (e.g. [47]). However, many other physically well-motivated models generate larger non-Gaussian signatures at the end of inflation (see reviews of [21]). Such primordial non-Gaussianities are well constrained by three-point correlation functions of the CMB anisotropies or their Fourier transform, the CMB bispectrum. Different inflationary models predict bispectra with different momentum dependence, or 'shapes'. We constrain these models by using an optimal estimator for their amplitude parameter, f_{NL}, for each specific bispectrum shape (see, e.g, [22, 83] for reviews).

Although all observations to date are consistent with vanishing non-Gaussianity, the models most favoured by the latest Planck CMB analysis were the ones with oscillations in the primordial power spectrum [30]. Among them are feature models, where the oscillations are caused by a sharp feature in either the inflationary potential [84–89], sound speed [90, 91], or multi-field potentials [92] (see [21, 93] for reviews). The primordial power spectrum then becomes scale-dependent, displaying sinusoidal oscillations that are linearly spaced in momentum space. The resulting bispectrum also oscillates and is highly uncorrelated with other popular bispectrum templates [94], therefore allowing us to constrain them independently.

Planck constrained f_{NL} for feature models from CMB bispectra, but no signals above 3σ significance were found after accounting for the 'look elsewhere effect' using the method introduced in [95]. The multi-peak statistic analysis, however, revealed some non-standard signals up to 4σ level that deserves closer attention [30]. There have been many other searches on signatures of oscillations. Constraints also come from the CMB power spectrum [82, 96–100], the large scale structure [101, 102], and a combination of the two [103, 104]. We expect stronger constraints on feature models from future LSS experiments [105]. This paper covers the prospects of upcoming CMB experiments in constraining f_{NL} for feature models.

Currently, there are two implementations of the optimal estimator for constraining f_{NL} for feature type models. The Planck analysis adopted the Modal estimator for which the given bispectrum is expanded using a separable basis [60, 67]. This method is efficient and can flexibly account for various oscillatory shapes and can easily constrain all frequencies simultaneously. However, when the oscillation frequency is large the modal basis fails to converge within a reasonable number of basis elements, making the method impractical. The other approach using a KSW-type estimator is viable for various shapes including the feature model [58, 66]. Although this method mainly applies to models with separable bispectra, even highly oscillatory templates can be computed. This method however tends to be more expensive as each frequency must be dealt with separately. We present further optimisations to the fast KSW estimator introduced in [106] and apply it to the feature models in our forecasts.

CMB Stage-4 (CMB-S4) is the next-generation CMB experiment located at the South Pole and in the Atacama Desert in Chile [25]. One of the main goals of CMB-S4 is to measure the polarisation signal in the CMB to the cosmic variance limit. In the Planck analysis, constraints on oscillatory models benefited much more from the inclusion of E-mode polarisation data compared to others, and therefore are anticipated to improve significantly with the future CMB experiments. Preliminary specifications have been released for these experiments [24, 25] and have been used to produce some forecasts for the standard f_{NL} templates, but not yet for feature type models. Here we address this by presenting the Fisher forecasts on f_{NL} for feature models based on these specifications and observe that the feature type models do indeed receive larger improvements from the extra polarisation information than the standard templates.

This chapter is organised as follows. First, we briefly review the theory of the CMB bispectrum in Section 4.1, building on our previous formulation to include polarisation. A bispectrum template for the feature model is defined and computed here. In Section 4.2 we formulate the bispectrum estimator and introduce methods to further optimise its computation.

The technique is applied to the case of feature models to obtain equations for the Fisher forecast of $f_{\rm NL}$. The implementation details are briefly discussed. In Section 4.3 we present the results of our forecast and how they depend on the model and experimental parameters. In particular, forecasts for the Simons Observatory are compared with the Planck results. We provide a summary at the end. The materials presented in this chapter are from my original research supervised by James Fergusson and published in [29].

4.1 Feature model bispectrum

4.1.1 CMB bispectrum

Recall that the three-point correlation function of the primordial perturbations is defined as

$$\langle \Phi(\mathbf{k}_1)\Phi(\mathbf{k}_2)\Phi(\mathbf{k}_3)\rangle = (2\pi)^3 \delta^{(3)}(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3) B_\Phi(k_1, k_2, k_3), \tag{4.1}$$

under assumptions of the statistical homogeneity and isotropy. The primordial bispectrum B_Φ vanishes for Gaussian perturbations, but more general inflation models predict non-zero bispectra with various shapes. In order to constrain these models we re-parameterise the bispectrum into an amplitude parameter $f_{\rm NL}$ and a normalised shape part. Constraining $f_{\rm NL}$ from the CMB measurements allows to determine how well a particular shape under consideration aligns with the data, which we translate into constraints on the model itself.

In order to compare the theory with measurements, we first need to relate the primordial perturbations to spherical multipole modes of the late-time CMB anisotropies. The form of (2.87) derived previously generalises to include the CMB polarisation as well;

$$a_{lm}^X = 4\pi(-i)^l \int \frac{d^3\mathbf{k}}{(2\pi)^3} \Phi(\mathbf{k})\Delta_l^X(k) Y_{lm}(\hat{\mathbf{k}}). \tag{4.2}$$

Here the index X is either T or E, representing the CMB temperature and E-mode polarisation, respectively. The linear CMB radiation transfer function $\Delta_l^X(k)$ can be computed from the Boltzmann solvers like CAMB [45] and CLASS [46].

The three-point correlation function of a_{lm}^Xs yields the reduced bispectrum $b_{l_1 l_2 l_3}$ times a geometrical factor $\mathcal{G}_{m_1 m_2 m_3}^{l_1 l_2 l_3}$, or the Gaunt integral. After some algebraic manipulations we obtain the formula (3.50) for the reduced bispectrum ;

$$b_{l_1 l_2 l_3}^{X_1 X_2 X_3} = \left(\frac{2}{\pi}\right)^3 \int_0^\infty r^2 dr \int_{\mathcal{V}_k} d^3\mathbf{k}\, (k_1 k_2 k_3)^2 B_\Phi(k_1, k_2, k_3) \prod_{i=1}^{3} \left[j_{l_i}(k_i r) \Delta_{l_i}^{X_i}(k_i) \right], \tag{4.3}$$

where j_l is the spherical Bessel function arising from the Rayleigh expansion formula. Using this equation, we can compute the projected bispectrum from any given primordial bispectrum. Direct computation of this four-dimensional integral for every l combination, however, is practically impossible. Not only is the integral in 4D but also the oscillatory integrand requires a large number of sample points in each of k_i, making the full calculation for every l_i triple prohibitively expensive. All bispectrum estimators get around this problem by expanding B_Φ as a sum of *separable* terms. We will use the feature model template to explain this in more detail later.

4.1.2   Feature model

We follow the works of [30, 66, 82, 95] and assume the following template for the bispectrum of feature models;

$$B_\Phi^{\text{feat}}(k_1, k_2, k_3) = \frac{6A_\Phi^2}{(k_1 k_2 k_3)^2} \sin(\omega K + \phi), \tag{4.4}$$

where $K = k_1 + k_2 + k_3$, A represents the primordial power spectrum amplitude, and ϕ denotes a phase. The oscillation 'frequency' ω is associated with the location and scale of the feature in the inflationary potential. It is often written in terms of the oscillation scale k_c as $\omega = 2\pi/3k_c$. ω is measured in Mpc but we omit the unit for notational convenience.

The feature model template has two free parameters that need to be fixed before we can constrain the model: the frequency ω and phase ϕ. The phase can be easily dealt with by observing that

$$B_\Phi^{\text{feat}}(k_1, k_2, k_3) = \cos\phi\, B_\Phi^{\text{sin}}(k_1, k_2, k_3) + \sin\phi\, B_\Phi^{\text{cos}}(k_1, k_2, k_3). \tag{4.5}$$

Here B_Φ^{sin} and B_Φ^{cos} correspond to feature models with $\phi = 0, \pi/2$ respectively. A template with some non-zero phase simply corresponds to a linear combination of the sine and cosine templates. As we will see later these two shapes are in fact highly uncorrelated and hence can be constrained independently from each other.

On the other hand, one still has a complete freedom of choice on the oscillation frequency. Such freedom dramatically expands the size of the parameter space. In practice, we constrain f_{NL} for each fixed value of the oscillation frequency, which yields hundreds of estimates. Since there are so many numbers we are looking at, there is a good chance that we find notable signals by sheer luck. This is called the *look elsewhere effect* and has been accounted for in the Planck analysis [23, 30] using the methods introduced in [82, 95]. The look-elsewhere-adjusted statistics used in the literature can be employed for the future CMB-S4

data analysis. This work, however, focuses on forecasting the raw estimates and comparing them with those of Planck.

4.1.3 Separability

The bispectrum template of feature models (4.4) is an example of a separable shape. It can be expressed as a sum of terms in the form $f(k_1)g(k_2)h(k_3)$ for some functions f, g and h, dramatically simplifying the computation of reduced bispectrum $b_{l_1l_2l_3}$ as we have seen in Section 3.3. The three-dimensional integral over the k space in (4.3) splits into three individual one-dimensional integrals for separable shapes. Feature models for example have

$$
\begin{aligned}
b_{l_1l_2l_3}^{X_1X_2X_3,\text{feat}} &= 6A^2 \left(\frac{2}{\pi}\right)^3 \int_0^\infty r^2 dr \int_{\mathscr{V}_k} d^3\mathbf{k}\, e^{i\omega(k_1+k_2+k_3)} \prod_{i=1}^3 \left[j_{l_i}(k_i r)\Delta_{l_i}^{X_i}(k_i) \right] \\
&= 6A^2 \left(\frac{2}{\pi}\right)^3 \int_0^\infty r^2 dr \prod_{i=1}^3 \left[\int_0^\infty dk_i\, e^{i\omega k_i} j_{l_i}(k_i r)\Delta_{l_i}^{X_i}(k_i) \right].
\end{aligned}
\tag{4.6}
$$

Here the real and imaginary parts of b^{feat} correspond to the bispectra of cosine and sine feature models, respectively. Now define

$$
s_l^X(r) := \frac{2A^{2/3}}{\pi} \int_0^\infty dk \sin(\omega k) j_l(kr)\Delta_l^X(k),
\tag{4.7}
$$

$$
c_l^X(r) := \frac{2A^{2/3}}{\pi} \int_0^\infty dk \cos(\omega k) j_l(kr)\Delta_l^X(k).
\tag{4.8}
$$

These are analogous to $\alpha_l^X(r)$ and $\beta_l^X(r)$ we defined in (3.82)-(3.83) for the vanilla KSW estimator. Then (4.6) reduces to

$$
\begin{aligned}
b_{l_1l_2l_3}^{X_1X_2X_3,\text{feat}} &= 6 \int_0^\infty r^2 dr \left(c_{l_1}^{X_1} c_{l_2}^{X_2} c_{l_3}^{X_3} - c_{l_1}^{X_1} s_{l_2}^{X_2} s_{l_3}^{X_3} - s_{l_1}^{X_1} c_{l_2}^{X_2} s_{l_3}^{X_3} - s_{l_1}^{X_1} s_{l_2}^{X_2} c_{l_3}^{X_3} \right) \\
&\quad + 6i \int_0^\infty r^2 dr \left(s_{l_1}^{X_1} c_{l_2}^{X_2} c_{l_3}^{X_3} + c_{l_1}^{X_1} s_{l_2}^{X_2} c_{l_3}^{X_3} + c_{l_1}^{X_1} c_{l_2}^{X_2} s_{l_3}^{X_3} - s_{l_1}^{X_1} s_{l_2}^{X_2} s_{l_3}^{X_3} \right) \\
&= b_{l_1l_2l_3}^{X_1X_2X_3,\cos} + i\, b_{l_1l_2l_3}^{X_1X_2X_3,\sin}.
\end{aligned}
\tag{4.9}
$$

4.2   Efficient computation of the estimator with polarisation

4.2.1   Estimator

The full optimal estimator for a given bispectrum in the weak non-Gaussian limit including both temperature and E-mode polarisation is given by [22, 58];

$$
S_i = \frac{1}{6} \sum_{l_j, m_j} \sum_{X_j} \mathcal{G}^{l_1 l_2 l_3}_{m_1 m_2 m_3} b^{X_1 X_2 X_3, (i)}_{l_1 l_2 l_3} (C^{-1}_{l_1 m_1, l_4 m_4})^{X_1 X_4} (C^{-1}_{l_2 m_2, l_5 m_5})^{X_2 X_5} (C^{-1}_{l_3 m_3, l_6 m_6})^{X_3 X_6}
$$
$$
\left[a^{X_4}_{l_4 m_4} a^{X_5}_{l_5 m_5} a^{X_6}_{l_6 m_6} - \left(C_{l_4 m_4, l_5 m_5} a^{X_6}_{l_6 m_6} + 2 \text{ cyclic} \right) \right]. \quad (4.10)
$$

Computing this form involves an inversion of the full covariance matrix which is computationally expensive. As a result we will follow the diagonal covariance approximation in [106] for the inverse covariances so that $C^{-1}_{l_1 l_4 m_1 m_4} \approx 1/C_{l_1} \delta^D_{l_1 l_4} \delta^D_{m_1 - m_4}$, and approximate the covariance in the linear term with an ensemble average of realistic simulations $C^{X_1 X_2}_{l_4 l_5 m_4 m_5} \approx \left\langle a^{X_1}_{l_1 m_1} a^{X_2}_{l_2 m_2} \right\rangle$. With these the estimator takes the form;

$$
\hat{f}_i = \sum_j (F^{-1})_{ij} S_j, \quad (4.11)
$$

where

$$
S_i = \frac{1}{6} \sum_{l_j, m_j} \sum_{X_j, X'_j} \mathcal{G}^{l_1 l_2 l_3}_{m_1 m_2 m_3} b^{X_1 X_2 X_3, (i)}_{l_1 l_2 l_3} (C^{-1}_{l_1})^{X_1 X'_1} (C^{-1}_{l_2})^{X_2 X'_2} (C^{-1}_{l_3})^{X_3 X'_3}
$$
$$
\left[a^{X'_1}_{l_1 m_1} a^{X'_2}_{l_2 m_2} a^{X'_3}_{l_3 m_3} - \left(\left\langle a^{X'_1}_{l_1 m_1} a^{X'_2}_{l_2 m_2} \right\rangle a^{X'_3}_{l_3 m_3} + 2 \text{ cyclic} \right) \right]. \quad (4.12)
$$

The summations here are taken over l_j, m_j, X_j and X'_j for each $j = 1, 2,$ and 3. The spherical multipole moments a^X_{lm}'s are computed from either the observed or simulated CMB maps. The covariance matrix C^{XY}_l is a 2×2 matrix consisting of values C^{TT}_l, C^{TE}_l, C^{ET}_l and C^{EE}_l. [1] The linear terms (the second term in square brackets) are required to account for anisotropies induced by masking and anisotropic noise. The bracket $\langle \cdot \rangle$ denotes averaging over Monte Carlo simulations of Gaussian realisations.

The Fisher matrix of the estimator is given by

$$
F_{ij} = \frac{f_{\text{sky}}}{6} \sum_{\text{all } X, X'} \sum_{\text{all } l} h^2_{l_1 l_2 l_3} b^{X_1 X_2 X_3, (i)}_{l_1 l_2 l_3} (C^{-1}_{l_1})^{X_1 X'_1} (C^{-1}_{l_2})^{X_2 X'_2} (C^{-1}_{l_3})^{X_3 X'_3} b^{X'_1 X'_2 X'_3, (j)}_{l_1 l_2 l_3}. \quad (4.13)
$$

---

[1]Note that this is equivalent to the other convention in literature [67], which instead uses a $2l \times 2l$ matrix containing four diagonal $l \times l$ block matrices C^{TT}, C^{TE}, C^{ET} and C^{EE}.

Here, f_{sky} denotes the fraction of the sky covered by the experiment, and $h^2_{l_1 l_2 l_3} := \sum_{m_j} \left(\mathscr{G}^{l_1 l_2 l_3}_{m_1 m_2 m_3} \right)^2$. Since the estimator $\hat{f}_i$ in (4.11) is nearly optimal, its 68% confidence (1σ) interval can be computed from the Fisher matrix as $\sigma_i := \Delta f^{(i)}_{\text{NL}} = (F^{-1})_{ii}$.

Note that the CMB-S4 experiment is ground-based and inevitably probe smaller fractions of the sky compared to satellite experiments like Planck. Having a smaller fraction of the sky available leads to increased uncertainties for the estimator. The current estimate is that many of the new experiments will cover approximately 40% of the sky, which is much less than the 74% of Planck. The error bars will thus increase by a factor of 1.38 from the decrease in f_{sky} alone. This loss in constraining power can be ameliorated by adding the Planck data from the pixels unobserved by these experiments.

4.2.2 Orthonormalising the covariance matrix

In [67] it was noted that orthogonalising the multipoles of temperature and polarisation maps dramatically reduces the number of terms required for computing the Modal estimator. This technique can also be applied to KSW estimators, or indeed any optimal bispectrum estimator, which is yet to be done to the authors' knowledge.

In both (4.12) and (4.13) there are summations over indices X and X' to account for correlations between the CMB temperature and E-mode polarisation. This can be simplified by essentially making a change of basis in X space for each l so that every C_l becomes orthonormal. Perform a Cholesky decomposition on C_l and invert the matrix. Then $C^{-1}_l = L^T_l L_l$, where L_l is a lower triangular matrix given by

$$L_l = \begin{pmatrix} \frac{1}{\sqrt{C^{TT}_l}} & 0 \\ \frac{-C^{TE}_l}{\sqrt{C^{TT}_l}\sqrt{C^{TT}_l C^{EE}_l - C^{TE^2}_l}} & \frac{C^{TT}_l}{\sqrt{C^{TT}_l}\sqrt{C^{TT}_l C^{EE}_l - C^{TE^2}_l}} \end{pmatrix} \tag{4.14}$$

Now let

$$\tilde{\Delta}^X_l(k) = \sum_{X'} L^{XX'}_l \Delta^{X'}_l(k), \quad \text{and} \quad \tilde{a}^X_{lm} = \sum_{X'} L^{XX'}_l a^{X'}_{lm}. \tag{4.15}$$

When a_{lm}'s are generated from simulations, the second transformation is not required as long as the new transfer function $\tilde{\Delta}^X_l$ is used in the process.

Defining $\tilde{b}_{l_1 l_2 l_3}$ to be the corresponding reduced bispectrum, (4.12) and (4.13) simplify to

$$S_i = \frac{1}{6} \sum_{l_j, m_j} \sum_{X_j} \mathscr{G}^{l_1 l_2 l_3}_{m_1 m_2 m_3} \tilde{b}^{X_1 X_2 X_3, (i)}_{l_1 l_2 l_3} \left[\tilde{a}^{X_1}_{l_1 m_1} \tilde{a}^{X_2}_{l_2 m_2} \tilde{a}^{X_3}_{l_3 m_3} - \left(\left\langle \tilde{a}^{X_1}_{l_1 m_1} \tilde{a}^{X_2}_{l_2 m_2} \right\rangle \tilde{a}^{X_3}_{l_3 m_3} + 2 \text{ cyclic} \right) \right] \tag{4.16}$$

and

$$F_{ij} = \frac{f_{\text{sky}}}{6} \sum_{\text{all } X} \sum_{\text{all } l} h_{l_1 l_2 l_3}^2 \, \tilde{b}_{l_1 l_2 l_3}^{X_1 X_2 X_3, (i)} \, \tilde{b}_{l_1 l_2 l_3}^{X_1 X_2 X_3, (j)}. \tag{4.17}$$

Using this method not only is mathematically concise but also halves the number of terms involved in the summation. The linear transformations in (4.15) only need to be done once at the beginning of the program and cost little compared to the main computation. We also found it easier to optimise the code using instruction-level parallelism after this simplification.

The only downside of this formalism is that we can no longer get breakdowns of signal from each of the TTT, TTE, TEE and EEE bispectrum since our new modes are linear combinations of the T and E modes. However, we are interested in either T-only or $T + E$ results in most cases, and this method works perfectly well for these cases.

4.2.3   Estimator for feature models

We now consider the estimator for feature models. The method is similar to the one seen in [66] except that the polarisation is included in a more concise way so that the covariance matrices are trivial, thanks to the orthonormalisation process outlined above.

Consider the bispectrum shape of

$$B_\Phi(k_1, k_2, k_3) = f_{\text{NL}}^{\text{sin}} B^{\text{sin}}(k_1, k_2, k_3) + f_{\text{NL}}^{\text{cos}} B^{\text{cos}}(k_1, k_2, k_3), \tag{4.18}$$

for a fixed value of oscillation frequency ω. Here B^{sin} and B^{cos} correspond to the reduced bispectra b^{sin} and b^{cos} defined in (4.9). The Fisher matrix F is 2×2 but its off-diagonal entries are 2-3 orders of magnitude smaller than the diagonal ones in most cases, as we will see in the next section. Thus, the two shapes are assumed to be uncorrelated and constrained individually. Here we present detailed computations for $f_{\text{NL}}^{\text{sin}}$ only but the cosine one can be computed similarly.

From (4.9) and the definition of Gaunt integral $\mathscr{G}_{m_1 m_2 m_3}^{l_1 l_2 l_3} = \int d\hat{\mathbf{n}} \, Y_{l_1 m_1}(\hat{\mathbf{n}}) Y_{l_2 m_2}(\hat{\mathbf{n}}) Y_{l_3 m_3}(\hat{\mathbf{n}})$ it follows that

$$S^{\text{cub}} = \int_0^\infty r^2 dr \int d^2 \hat{\mathbf{n}} \left[-M_s^3 + 3 M_s M_c^2 \right], \quad \text{and} \tag{4.19}$$

$$S^{\text{lin}} = -3 \int_0^\infty r^2 dr \int d^2 \hat{\mathbf{n}} \left[-M_s \langle M_s^2 \rangle + M_s \langle M_c^2 \rangle + 2 M_c \langle M_s M_c \rangle \right], \tag{4.20}$$

where

$$M_s(r, \hat{\mathbf{n}}) = \sum_X \sum_{lm} \tilde{s}_l^X(r) \tilde{a}_{lm}^X Y_{lm}(\hat{\mathbf{n}}), \tag{4.21}$$

$$M_c(r, \hat{\mathbf{n}}) = \sum_X \sum_{lm} \tilde{c}_l^X(r) \tilde{a}_{lm}^X Y_{lm}(\hat{\mathbf{n}}). \tag{4.22}$$

The sum of S^{cub} and S^{lin} gives the final value of S for the sine feature model.

For efficient Fisher matrix calculation, we follow [65] and deploy the identity

$$h_{l_1 l_2 l_3}^2 = \frac{(2l_1 + 1)(2l_2 + 1)(2l_3 + 1)}{8\pi} \int_{-1}^{1} d\mu\, P_{l_1}(\mu) P_{l_2}(\mu) P_{l_3}(\mu), \tag{4.23}$$

where $P_l(\mu)$ represents the Legendre polynomial. Then,

$$F = \frac{3}{4\pi} \int r^2 dr \int r'^2 dr' \int d\mu \left[P_{ss}^3 + 3 P_{ss} P_{cc}^2 - 3 P_{cs}^2 P_{ss} - 3 P_{sc}^2 P_{ss} + 6 P_{cs} P_{sc} P_{cc} \right]. \tag{4.24}$$

where we have defined

$$P_{ss}(r, r', \mu) := \sum_X \sum_l (2l + 1) \tilde{s}_l^X(r) \tilde{s}_l^X(r') P_l(\mu),$$

$$P_{sc}(r, r', \mu) := \sum_X \sum_l (2l + 1) \tilde{s}_l^X(r) \tilde{c}_l^X(r') P_l(\mu). \tag{4.25}$$

and similarly P_{cs} and P_{cc}.

Calculation of (4.20) and (4.25) are two of the most computationally expensive steps. If we did not orthonormalise the covariance matrix, there would be extra summations over X' and some 2×2 matrix computations involving $(C_l^{-1})^{XX'}$ in these steps.

4.2.4 Probing beam and instrumental noise

In an ideal experiment where measurements are made on each point of the sky perfectly, the covariance matrices $C_l^{XX'}$ in (4.12) and (4.13) would consist purely of the signal. In reality, however, the probing beam has finite width and the sensors are noisy. These effects can be incorporated by modifying the covariance matrices and bispectra as follows;

$$C_l^{X_1 X_2} \to W_l^{X_1} W_l^{X_2} C_l^{X_1 X_2} + N_l^{X_1 X_2}, \qquad b_{l_1 l_2 l_3}^{X_1 X_2 X_3} \to W_{l_1}^{X_1} W_{l_2}^{X_2} W_{l_3}^{X_3} b_{l_1 l_2 l_3}^{X_1 X_2 X_3}, \tag{4.26}$$

where W_l^X and $N_l^{X_1 X_2}$ represent the beam window function and the noise covariance matrix, respectively. When substituted into the KSW estimator, these changes are equivalent to

modifying

$$C_l^{X_1X_2} \rightarrow C_l^{X_1X_2} + \left(W_l^{X_1} W_l^{X_2} \right)^{-1} N_l^{X_1X_2} = (C_l^{\text{sig}})^{X_1X_2} + (C_l^{\text{noise}})^{X_1X_2}, \qquad (4.27)$$

while keeping the bispectrum same as before. We have defined the effective (beam-corrected) noise covariance matrix C_l^{noise}. Modes for which C_l^{noise} is much larger than C_l^{sig} contribute little to the f_{NL} estimator.

For forecasting purposes, we assume Gaussian beam and white uncorrelated noise until more detailed experiment specifications become available. Under these assumptions, the effective noise covariances reduce to [107]

$$C_l^{\text{noise},TT} = \exp\left(l(l+1)\sigma_{\text{beam}}^2 \right) N_{\text{white}}, \quad C_l^{\text{noise},EE} = 2\, C_l^{\text{noise},TT}, \quad C_l^{\text{noise},TE} = 0. \quad (4.28)$$

The factor of two for the *EE* mode comes from measuring the two Stokes parameters Q and U. The Gaussian beam profile is usually specified by its FWHM (full width at half maximum) in arcmin, which is then converted to standard deviations in radians for σ_{beam}. The noise level is in the units of $\mu K \cdot$ arcmin. This is then divided by $T_{\text{CMB}} = 2.725K$, converted to radians and squared to get N_{white}.

For the Planck component-separated maps, the beam FWHM equals 5 arcmin and the noise level of 47 $\mu K \cdot$arcmin gives a good approximation to the noise covariances. For CMB-S4 the details are not confirmed, but the beam FWHM is expected to lie between 1-5 arcmin, while the noise level will range from 1 to 9 $\mu K \cdot$arcmin. [25]

In reality, there exists extra contamination in large angular scales due to the $1/f$ noise. Though most of our analysis assumes a simpler form of the noise covariances elaborated above, for the Simons Observatory forecasts we follow [24] and model the $1/f$ noise as $N_l = N_{\text{red}}(l/l_{\text{knee}})^{\alpha_{\text{knee}}} + N_{\text{white}}$. The noise curves from each channel were then put together using the inverse variance method. This is a good approximation for the E-mode polarisation but not for temperature, since extra degradations occur during the component separation process. Still, this would be a reasonable approximation for our forecast since the dominant contributions to the feature model signal come from the polarisation data. For Planck, the full post-component-separation noise curves are available and were used for our analysis.

4.2.5 Implementation and validation

We implemented the pipeline outlined above using the C programming language and parallelised using hybrid MPI + openMP to fully benefit from modern computing architecture. The code was then run in the COSMOS super-computing system.

The transfer functions are generated from the CAMB code [45]. The spherical Bessel function values were pre-computed using recursion relations and stored, while the Legendre function values were computed on the fly using the GNU scientific library. We computed the angular power spectrum data from the ΛCDM parameters estimated in the Planck 2015 results.

Numerical integrations for the variables k, r and r' were performed using a simple trapezoidal method, which were found to be accurate and easier to vectorise for optimisation purposes. On the other hand, the integration of μ required more care because the Legendre polynomials are highly oscillatory. We adopted the Gauss-Legendre quadrature rule with $1.5\,l_{\max} + 1$ points which can integrate polynomials up to order $3\,l_{\max}$ exactly. The weights and nodes for the quadrature were computed using the QUADPTS code [108].

We performed various checks to ensure that the code runs correctly. First we used the code to reproduce the Planck results on the error bars for feature models, which agreed within 3% level. The error mainly comes from the difference in Modal and KSW approaches for representing primordial bispectra. We discuss this further in Chapter 5. The code was then used to compute the bispectrum of the constant model which corresponds to the feature model template with $\omega = \phi = 0$. There exists an approximate analytic form in this case [60] which we were able to reproduce accurately. We also performed convergence tests on r and r' integration by doubling the number of points for each of them. The grid was chosen to be very dense around recombination and quite dense near reionisation. We confirmed that changes in the integral are kept less than 0.5% for each value of ω.

4.3   CMB-S4 forecast results

4.3.1   Phase dependence

We now present the CMB-S4 forecast on the f_{NL} estimation power for feature models. For notational convenience, we denote the error bars for f_{NL} of sine and cosine feature models by $\sigma_{\sin}$ and $\sigma_{\cos}$. The superscripts T and $T + E$ are also put to distinguish temperature-only analyses from the full polarisation ones.

First of all, we check that the sine and cosine bispectrum templates defined in (4.6) are indeed uncorrelated and can be constrained separately. We see if the Fisher matrix of feature models is robust to changes in the phase for different ω values of interest. The feature model bispectrum with a specific phase ϕ can be represented as a sum of sine and cosine ones as in

(4.5). Hence, its Fisher matrix is given by

$$F(\omega, \phi) = \cos^2 \phi \, F_{ss}(\omega) + \sin^2 \phi \, F_{cc}(\omega) + 2 \cos \phi \sin \phi \, F_{sc}(\omega), \qquad (4.29)$$

where F_{ss} is the element F_{ij} of the Fisher matrix in (4.13) with the reduced bispectra $b^{(i)} = b^{(j)} = b^{\sin}$, and so on. Correlation between the sine and cosine templates can be expressed as $F_{sc}/(F_{ss}F_{cc})^{1/2}$, and this value can be learned from analysing the ϕ dependence of $F(\omega, \phi)$.

Figure 4.1 shows the forecast error bars for the full phase range $[0, \pi]$ in the most sensitive experiment specification of 1' beam and $1 \mu K \cdot arcmin$ noise. The forecast σ varies within a 1% level for every $\omega \geq 20$. In terms of the Fisher matrix, the cross term F_{sc} was 2-3 orders of magnitude smaller than F_{ss} and F_{cc} for all cases. The correlation between the sine and cosine templates was less than 1%. This justifies our previous choice of constraining $f_{\mathrm{NL}}^{\sin}$ and $f_{\mathrm{NL}}^{\cos}$ separately. We now focus our attention on $\sigma_{\sin}$ for our discussions.

For smaller values of ω, the phase affects the error bar primarily through modulating the amplitude of the acoustic oscillations in the CMB itself. The radiation transfer functions are non-zero for k values in $0 - 0.8 \, \mathrm{Mpc}^{-1}$. The argument ωk covers less than two full periods in this k range if $\omega \leq 10 \, \mathrm{Mpc}$, and phase has a direct influence on the amplitude of the acoustic peaks. In the extreme case of $\omega = 0$, the bispectrum vanishes completely for the sine feature model. Variations in the overall bispectrum amplitude therefore result in varying Fisher information for low frequencies.
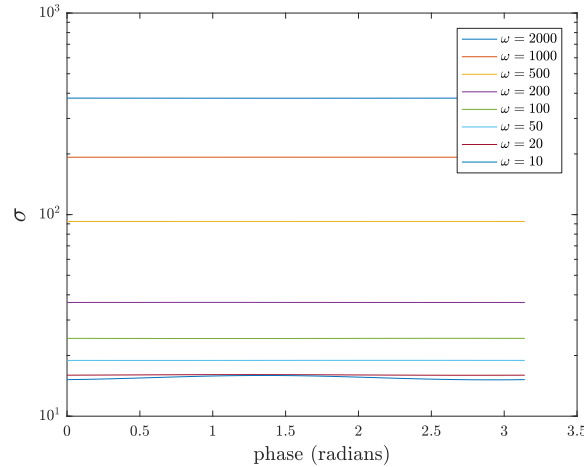
Fig. 4.1 Forecast error bars σ^{T+E} versus the phase ϕ. Apart from the smallest frequency $\omega = 10$, the error bar remains almost constant. This implies that the sine ($\phi = 0$) and cosine ($\phi = \pi/2$) feature models can be constrained independently.

4.3.2 $l_{\mathbf{max}}$ dependence

Figure 4.2 shows the graph of the forecast error bar $\sigma_{\sin}^{T+E}$ as we increase $l_{\max}$. The forecasts were done within angular scale range $2 \leq l \leq l_{\max}$, the oscillation frequency ω set to 100, and assuming 1' beam and $1\mu K \cdot$arcmin noise. The Planck noise curves were approximated by ones for 5' beam and 47 $\mu K \cdot$arcmin noise for this plot only since we extend $l_{\max}$ to 4000 here.

The Planck error bar essentially stalls out when $l_{\max}$ reaches 2000. The forecast error bar, on the other hand, keeps decreasing until $l_{\max} = 4000$ thanks to the improved sensitivity in measuring small scales (and large l's). Despite the information loss due to smaller sky coverage f_{sky}, the forecast error bar reduces to about 42% of Planck by $l_{\max} = 4000$. This corresponds to a factor of 2.4 times improvement to measurement precision on f_{NL}.
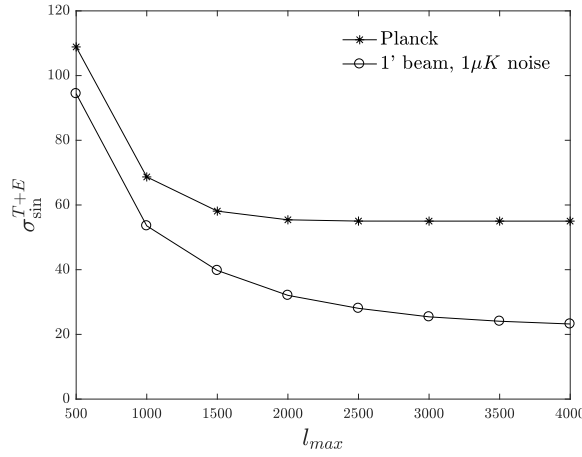
Fig. 4.2 Forecast error bars $\sigma_{\sin}^{T+E}$ when multipoles $2 \leq l \leq l_{\max}$ are included, in comparison with Planck. The oscillation frequency ω is set to 100 Mpc in all cases. Planck did not have access to the information from modes $l \geq 2000$, but CMB-S4 is expected to be able to explore modes up to $l = 4000$.

4.3.3 Beam and noise dependence

Now we explore the effects of different beam widths and noise levels on the forecast error bars. Figure 4.3 shows forecast $\sigma_{\sin}^{T+E}$ for ranges of beam and noise levels. Their oscillation frequencies are also varied, but only two representatives $\omega = 20$ and 2000 are chosen here. Forecasts for the other values of ω also display similar dependences on beam widths and noise levels.

First of all, note that all estimated error bars in the plot are smaller than Planck, for which $\sigma_{\sin}^{T+E} = 34$ when $\omega = 20$ and $\sigma_{\sin}^{T+E} = 610$ when $\omega = 2000$. In fact, even the least sensitive

CMB-S4 specification of 5' beam and $9\mu K\cdot$arcmin noise is expected to put better bounds on feature models. Note that $l_{\max} = 4000$ for both T and E here.

Wider beams and noisier detectors provide less signal and thus larger error bars, as expected. In this range of beam width and noise levels, noise has a bigger effect on the forecast; experiments with 1' beam and $5\mu K\cdot$arcmin noise yields larger error bars than the ones with 5' beam with $1\mu K\cdot$arcmin noise. Between the most sensitive specification of 1' beam and $1\mu K\cdot$arcmin and the least sensitive one with 5' beam and $9\mu K\cdot$arcmin, $\sigma_{\sin}$ differs by a factor of 1.6.

Fig. 4.3 Beam (left) and noise (right) dependences of the forecast error $\sigma_{\sin}^{T+E}$ for $\omega = 2000$ (top) and $\omega = 20$ (bottom). The noise level was set as $1\mu K\cdot$arcmin for the first plot, while the second plot had a fixed beam FWHM of 1'. We obtain less information from using wider beam and noisier sensors, as expected.

4.3.4 Oscillation frequency dependence

We present the main results of the forecast. Figure 4.4 summarises the $\sigma_{\sin}$ forecasts for several different CMB-S4 preliminary specifications, including the Simons Observatory (SO) baseline and goal. Note that the $1/f$ noise effect is incorporated in SO forecasts but not in other ones. We also provide 1σ errors for joint estimators, for which Planck signals from the fraction of the sky not covered by CMB-S4 are combined via a simple inverse variance method; $\sigma_{\text{joint}}^{-2} = \sigma_{\text{CMB-S4}}^{-2} + \sigma_{\text{Planck}}^{-2}$. This is not statistically optimal but serves as a conservative estimate of the joint estimation power.

The most sensitive setup with 1' beam and $1\mu K\cdot$arcmin noise would yield error bars that are 47-62% of Planck. These correspond to a factor of 1.6-2.1 improvements. Here relatively smaller improvements are made for high oscillation frequencies. They correspond to smaller momentum scales $k_* = 2\pi/3\omega$, or larger angular scales, which benefit less from

the increased sensitivity of CMB-S4. When the results are combined with Planck, the error bar reduces to 45-57% of Planck, or a factor of 1.7-2.2 improvement.

Forecast error bars from the SO baseline specification and the more ambitious one ('goal') do not differ very much. Quoting in terms of the baseline values, $\sigma_{\sin}$ lies about 68-86% of that of Planck or equivalently, 1.2-1.5 times smaller than Planck. Numbers change to 62-74% when combined with Planck, so the overall improvement ratio is about 1.3-1.6.
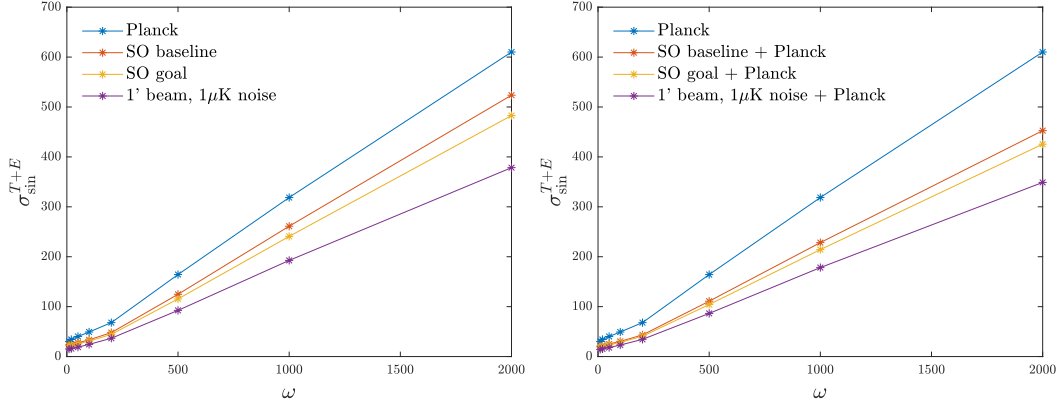
Fig. 4.4 Frequency dependence of the forecast error in comparison to Planck (left). All CMB-S4 specifications would improve constraints on feature models. The most sensitive setup with 1' beam and $1\mu K\cdot$arcmin noise is expected to yield error bars that are 1.6-2.1 times smaller than Planck. When the Planck results are combined with CMB-S4, we get even stronger constraints (right).

Figure 4.5 shows the results when only the CMB temperature data are used in the forecast. CMB-S4 would in fact be worse than Planck in terms of constraining $f_{\rm NL}^{\rm feat}$ for this case. The loss in information due to less sky coverage overwhelms the increased sensitivity. We see again that the real strength of CMB-S4 lies in measuring CMB polarisation.

Then how much information do we actually gain from adding E-mode polarisation? Figure 4.6 shows the ratio of $\sigma_{\sin}$s between the temperature-only and polarisation-included analyses. The forecast error bars reduce up to 4.6 times smaller when polarisation information is added, which is much larger than the corresponding Planck value of 2.2. The ratio decreases overall when the joint statistics with Planck are considered. An intriguing feature of this plot is that the ratio is maximised around $\omega = 200$ before it starts dropping again.

In order to gain insight into this behaviour, we performed some simplified computations using the power spectrum. We imposed oscillations on the primordial power spectrum as $P_{\rm feat}(k) = P(k)(1 + \sin(2\omega k + \phi))$, which is just like our feature model bispectrum template but with $\omega(k_1 + k_2 + k_3)$ replaced by $\omega(k + k)$. $P_{\rm feat}(k)$ is then projected to the late-time
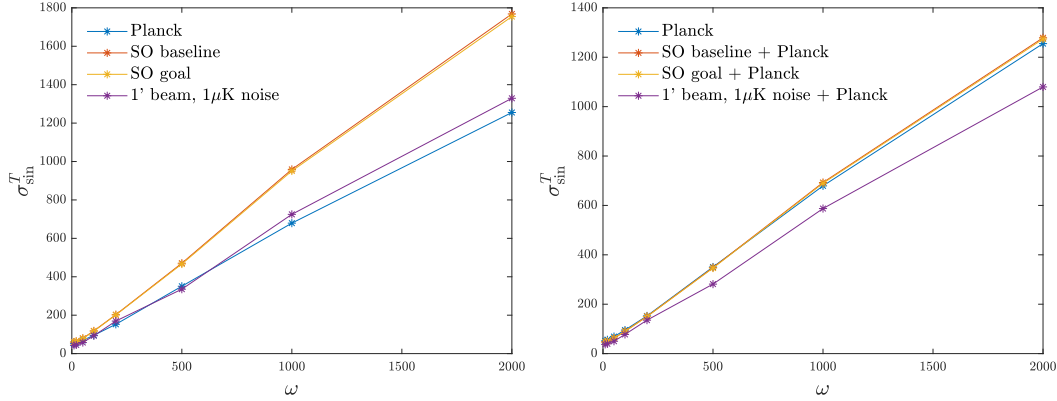
Fig. 4.5 Frequency dependence of the forecast error from temperature data only, in comparison to Planck (left). CMB-S4 would perform worse than Planck when only the temperature map is concerned. After the addition of Planck data the error bars improve only marginally (right). This shows that polarisation data is crucial for constraining feature models.

harmonic space using the transfer functions;

$$C_{\text{feat},l}^{X_1 X_2} = \frac{2}{\pi} \int k^2 dk P_{\text{feat}}(k) \Delta_l^{X_1}(k) \Delta_l^{X_2}(k). \tag{4.30}$$

We observed that the fractional variation $(C_{\text{feat},l} - C_l)/C_l$ displays some oscillations in l, and the largest contribution comes from a term $\propto \sin(2\omega l/\Delta\tau)$ where $\Delta\tau$ represents the conformal distance to the last scattering surface. This fact can be explained by approximating the transfer function as $\Delta_l(k) \approx (1/3) j_l(k\Delta\tau)$ and noting that the spherical Bessel function has a sharp peak at l for large l's. The integral in (4.30) therefore picks up a term proportional to $\sin(2\omega l/\Delta\tau)$.

   The amplitude of these 'maximal' oscillations in $(C_l' - C_l)/C_l$ were then computed using discrete Fourier transform for different values of oscillation scale ω and two different phases $\phi = 0, \pi/2$ (i.e. sine and cosine). The results are shown in Figure 4.7. Some extra wiggles to the graph come from the phase of oscillations imposed; we indeed see that graphs of sine and cosine oscillate between each other. Some peak features near $\omega \approx 70$ and 140 arise from resonances with Baryonic Acoustic Oscillations.

   We can think of the computed amplitude as a measure of information C_l's contain about the primordial oscillations. First of all, note that the amplitude in all four plots generally decreases as ω grows. Previously in Figure 4.4 we saw that the amount of information obtained from the CMB is smaller for larger ω's, consistent with what can be said from the amplitude analysis. Moreover, the amplitudes for the EE mode are generally larger than the TT mode ones, and their difference is the largest in the ω range of 70 to 300. This could

Fig. 4.6 Improvements on the forecast error when including E-mode polarisation data. Constraints from CMB-S4 would improve significantly from addition of the polarisation data. The improvement is maximised around $\omega \approx 200$ Mpc.

serve as a heuristic explanation for the improvement in forecast error bars from including polarisation data being maximised around $\omega = 200$, as depicted in Figure 4.6.

4.3.5   Comparison to scale-invariant models

Our pipeline for forecasting $f_{\mathrm{NL}}^{\mathrm{feat}}$ also yields forecasts for f_{NL} of the constant model, which is scale-invariant and have a trivial shape; $B(k_1, k_2, k_3) \propto (k_1 k_2 k_3)^{-2}$. Forecasts on $f_{\mathrm{NL}}^{\mathrm{const}}$ follow from our pipeline by simply setting the oscillation frequency $\omega = 0$ and phase $\phi = \pi/2$. Table 4.1 summarises the forecast results for several different CMB-S4 specifications mentioned before, using both T and E data and in combination with Planck data from the regions of the sky not covered by CMB-S4. For the 1' beam and $1\mu K{\cdot}$arcmin noise setup, the error bar is expected to be reduced by a factor of 2.3 compared to Planck.

Table 4.1 Forecasts on the estimation errors of f_{NL} for the constant model

	Planck	SO baseline + Planck	SO goal + Planck	1' beam, $1\mu K$ noise + Planck
$\sigma(f_{\mathrm{NL}}^{\mathrm{const}})$	23.4	14.9	14.0	10.4

The latest Planck constraints on f_{NL} of some popular bispectrum templates are given by $f_{\mathrm{NL}}^{\mathrm{local}} = 2.5 \pm 5.7$, $f_{\mathrm{NL}}^{\mathrm{equil}} = -16 \pm 70$, and $f_{\mathrm{NL}}^{\mathrm{ortho}} = -34 \pm 33$ [30]. CMB-S4 is expected to yield better estimates on these as well. Table 4.2 summarises the forecast improvement ratio given in [25] together with the constant and feature model ratios computed in this work.

Surprisingly, the estimation error for feature models does not improve as much as other templates. Feature models benefit much more from the inclusion of polarisation data than any
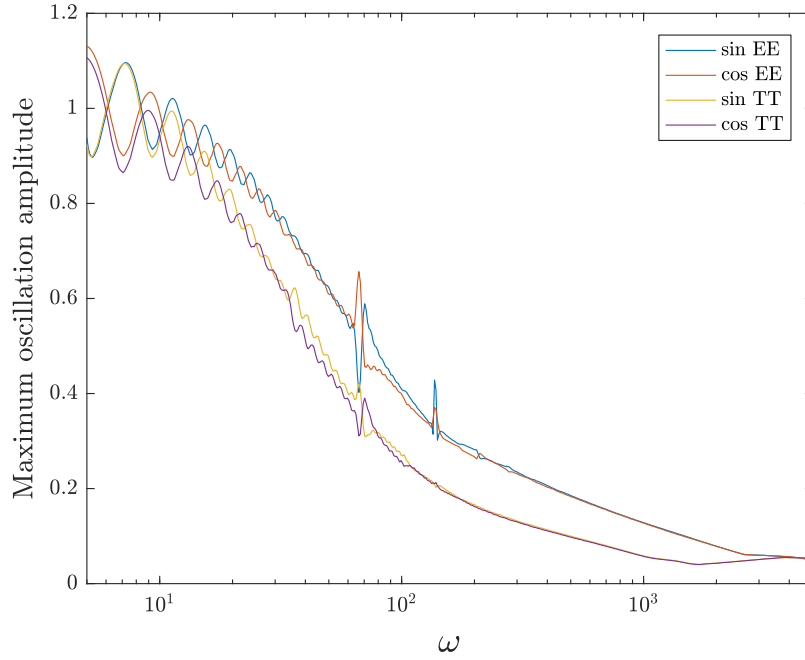
Fig. 4.7 The maximum amplitude of oscillations detected in fractional variations of the projected power spectrum C_l^{TT} and C_l^{EE}, when extra oscillations $\sin(2\omega k)$ and $\cos(2\omega k)$ were imposed on the primordial power spectrum. Heuristically this shows that E-mode polarisation is more sensitive to the primordial oscillations, especially in the ω range of 70 to 300. Some peaks near $\omega = 70$ and 140 arise from resonances with Baryonic Acoustic Oscillations.

other scale-independent shapes; for example, $\sigma^T/\sigma^{T+E} = 4.6$ for the feature models with $\omega = 200$ in CMB-S4 while the value equals 2.8 for the constant models. Because CMB-S4 would have a significantly enhanced polarisation measurement sensitivity, we originally expected the feature models to be constrained significantly better than Planck.

In order to investigate this lack of improvement, we performed a breakdown analysis on the improvements gained from CMB-S4 temperature and polarisation; we computed $\sigma(f_{NL})$ for the constant and feature models using each of the four combinations of Planck / CMB-S4 noise curves for temperature / polarisation (e.g. Planck T + CMB-S4 E). The results are summarised in Table 4.3.

We see that the constraints on feature models improve by a factor of 1.7 when swapping Planck polarisation noises with the CMB-S4 ones. This factor is indeed larger than 1.6 of the constant model. The difference is however not significant. It seems that the amount of feature signals in polarisation data left unexplored by Planck is not tremendously large compared to the constant model. The feature model improves less than the constant model when the temperature measurements are enhanced. In fact, for feature models, the signal

Table 4.2 Expected improvement ratios of the f_{NL} estimation errors for the CMB-S4 1'beam, $1\mu K$arcmin setup, for various bispectrum templates. The local, equilateral and orthogonal results are quoted from [25].

	Local	Equilateral	Orthogonal	Constant	Feature ($\omega = 200$)
$\sigma^{\text{Planck}}/\sigma^{\text{CMB-S4}}$	2.5	2.1	2.4	2.3	2.0

Table 4.3 Expected improvements on the estimation errors of f_{NL} for each combination of Planck / CMB-S4 temperature (T) and polarisation (E) data. Here the CMB-S4 assumes 1' beam and $1\mu K$arcmin noise. The feature model has oscillation frequency $\omega = 200$ and phase $\phi = 0$. The sky fraction $f_{\text{sky}} = 0.4$ for all cases except for Planck T + Planck E, for which $f_{\text{sky}} = 0.76$.

$\sigma(f_{NL}^{\text{const}})$		E		$\sigma(f_{NL}^{\text{feat}})$		E	
improvements		Planck	CMB-S4	improvements		Planck	CMB-S4
T	Planck	1.0	1.6	T	Planck	1.0	1.7
	CMB-S4	1.1	2.2		CMB-S4	0.9	2.0

loss from smaller sky fraction f_{sky} eclipses the signal gain from more sensitive temperature measurements. This lack of improvements from temperature causes the full CMB-S4 constraints on the feature model not to improve as much as the constant model overall.

Summary

Upcoming CMB Stage-4 experiments will provide an opportunity to measure the CMB temperature anisotropies and polarisation with ever greater precision. The CMB bispectrum estimators for constraining primordial non-Gaussianity would greatly benefit from the improvement in measurement sensitivity. In this research, we made forecasts on the f_{NL} for feature models, which have not been done so far despite the growing interest in inflation models with oscillations in the primordial spectrum. For efficient forecasts we simplified the bispectrum estimator for f_{NL} by orthonormalising the covariance matrix, further optimising the computation. When the most sensitive CMB Stage-4 experiment specification of 1' beam and $1\mu K$arcmin noise is concerned, we expect a factor of 1.7-2.2 times more stringent constraints compared to Planck. Under the realistic Simons Observatory conditions, the improvement would be about 1.3-1.6 to Planck.

Although this is not a massive boost in the estimation power, we can hope to verify the current 4σ-level signals found in the Planck 2015 analysis. It is also worth noting that CMB-S4 would allow us to explore the modes with $l > 2000$ and localised oscillations on these scales are currently unconstrained. Moreover, though we have only considered linearly-spaced oscillations in this work, we expect even better improvements on the models inducing logarithmically-spaced oscillations where small scale modes with higher l would greatly enhance the constraining power. Lastly, we can expect to cross-validate using these new statistically independent modes to test current constraints further.

We also extensively studied how the forecasts depend on various parameters. Frequency dependences of the ratio between the T and T+E forecasts were particularly illuminating; the improvement from adding polarisation information is maximised around $\omega = 200$. Some simplified calculations were presented to heuristically address this fact. Even though the estimation power on feature models massively benefits from the polarisation data, the expected improvements compared to Planck are quite underwhelming overall. A breakdown analysis of the temperature and polarisation contribution revealed that the feature models would indeed improve more than other scale-independent models if only the polarisation (and not temperature) measurement sensitivity were enhanced to the CMB-S4 standards. However, boosts in the temperature measurements affect scale-independent models more so that they gain more information overall.

The KSW estimator utilises separable bispectrum templates to constrain f_{NL} exactly even for highly oscillatory models but is inherently restrictive in the range of shapes it can handle. On the other hand, the Modal estimator uses separable basis functions to expand the given bispectrum shape and thus is able to cover a broad range of non-separable shapes. Due to limitations such as the basis size, however, its accuracy drops for oscillatory models with

high frequency. Hence, currently, there is a blind spot in the bispectrum analysis for general (non-separable) *and* highly oscillatory models barred by numerical challenges. We address this in our original research on the high-resolution CMB bispectrum estimator in Chapter 5.

Chapter 5

High-Resolution CMB Bispectrum Estimator

Despite their major role in constraining a wide range of inflation models, there are only a few implementations of the CMB bispectrum estimator due to computational challenges. The Planck collaboration mainly utilised three independent approaches for their bispectrum analysis [23, 30, 70]: KSW [58], Modal [60], and Binned [61]. Standard bispectrum templates for Local, Equilateral and Orthogonal shapes have been constrained using all three methods independently, whilst other more specific ones were covered by only a subset of them.

Some of the most challenging models to handle are those with oscillatory behaviour. Numerous theoretically well-motivated models ([88, 94, 109–111] and many more, see e.g., [21, 105] for reviews) fall into this category. Feature models, as discussed in the previous chapter, often predict linearly spaced oscillations in the bispectrum. Resonance models such as axion monodromy inflation [112, 113] produce log-spaced oscillations modulated with various envelopes. Constraining such models is difficult since oscillations in the bispectra degrade the numerical stability of the integrals involved.

Several simple models with oscillations have been studied in the latest Planck analysis using Modal and a few specific KSW-type estimators introduced in [66]. Modal estimator covers a wide range of oscillatory models, both with and without envelopes, thanks to the versatility of its mode expansion. It is however difficult to constrain models with high-frequency oscillations using generic mode functions. Modal code allows for choosing a more tailored set of modes for this purpose, but the fact that there are *two* independent sets of mode functions - primordial and late-time - complicates targeted analyses. For each specific selection of mode functions in the primordial 'k' domain, a suitable choice of basis has to

be made in the late-time 'l' space. Rapidly oscillating signals may also get lost during the projection from one to the other and cause some numerical issues.

On the other hand, adapted KSW-type estimators [66, 114, 115] may probe models with much faster oscillations in their bispectra. They exploit the separability of specific oscillatory templates and/or Fourier-type bases to obtain accurate constraints on a model-by-model basis. However, these methods are not easily applicable for models with more complicated bispectra, especially the ones that do not depend solely on $K = k_1 + k_2 + k_3$. The formalism is also restricted to certain types of bispectra and requires a separate implementation for each case. This was one of the main motivations behind the research project; can we have a bispectrum estimator which has the efficiency and versatility of Modal, whilst also benefiting from the precision of the KSW formalism for a more specialised analysis?

We developed a novel CMB bispectrum estimation pipeline CMB-BEST. [1] CMB-BEST has two main strengths over former methods. First of all, it is implemented for completely general basis sets, which allows broad analyses of inflationary models. In fact, the conventional KSW estimator is equivalent to a specific choice of basis in CMB-BEST. CMB-BEST can cover any model KSW estimator can, and furthermore, it can simultaneously constrain general bispectrum shapes through primordial mode expansion, just like Modal.

Secondly, CMB-BEST is designed to be able to handle complex and highly oscillatory signals. It does not require a separate set of basis functions for late-time l space, contrary to Modal. More diverse and specialised choices of basis can therefore be made. The code will be used on numerous inflation models with complex oscillations which are yet to be investigated due to lack of resolution from previous methods. Potential choices of the targeted basis for this purpose will be discussed in Section 5.1.2.

Every good thing comes at a price. For CMB-BEST, the price is computational cost. Combining the best of Modal and KSW estimators, CMB-BEST's formalism is more numerically demanding than both of them. Naïvely speaking, running CMB-BEST for one set of basis functions is equivalent to computing thousands of KSW estimators put together. It is prohibitively expensive unless properly and thoroughly optimised.

We invested a considerable amount of time and effort in optimising the code. Separable mode expansions and subsequent algorithm design were studied in detail for a maximal reduction in computational complexity. We also made full use of parallelisation techniques to exploit modern computing architecture. Improving data locality for efficient memory access yielded an order of magnitude speed-up. We illustrate our optimisation procedure in Section 5.2.

---

[1] Short for <u>CMB</u> <u>B</u>ispectrum <u>Est</u>imator. The goal of CMB-BEST is to be the best bispectrum estimator with wide applications and great precision.

The code was then tested thoroughly both internally and against Planck analysis. We used CMB maps and simulations from the Planck satellite experiment and checked that CMB-BEST agrees with previous routines map-by-map for various bispectrum templates. Different choices of basis functions within CMB-BEST also yielded consistent results. We dedicate Section 5.3 to presenting the outcome of these consistency checks.

Lastly, we present some applications of the code which serve as a proof of concept. In particular, we connect CMB-BEST to PRIMODAL, an efficient numerical code for computing bispectra of primordial perturbations from a given single field model [116]. PRIMODAL uses separable modes for its computation, which enables template-free, direct model-to-constraint analysis when combined with CMB-BEST appropriately. Section 5.3.3 highlights some results from the combined pipeline. We also identify areas where improvements could be made and discuss directions for future research.

Upcoming surveys will provide major improvements to constraints on primordial non-Gaussianity. Inflationary models with oscillations would also greatly benefit from the enhanced sensitivity, as we discussed in the previous chapter. It is therefore crucial to have a robust and flexible bispectrum estimation routine ready for future surveys, especially one which can handle high-frequency oscillations. Having another code independent from the existing ones would also be greatly beneficial for cross-validation. We expect CMB-BEST to fill this role in the near future.

5.1   Formalism

5.1.1   CMB-BEST formalism

Recall that the optimal CMB bispectrum estimator for a given template can be written as

$$\hat{f}_{\mathrm{NL}} = \frac{1}{N} \sum_{l_j, m_j} \frac{\mathscr{G}^{l_1 l_2 l_3}_{m_1 m_2 m_3} b_{l_1 l_2 l_3}}{C_{l_1} C_{l_2} C_{l_3}} \left[a_{l_1 m_1} a_{l_2 m_2} a_{l_3 m_3} - \left(\left\langle a^G_{l_1 m_1} a^G_{l_2 m_2} \right\rangle a_{l_3 m_3} + 2 \text{ cyc.} \right) \right]. \quad (5.1)$$

Here we omit superscripts X for temperature and polarisation for notational convenience. Even though the formalism in this section will be presented for CMB temperature data only, the method is general and can easily be extended to include polarisation. For estimation of the full covariance matrix $C_{lm,l'm'}$ needed for the linear term, we use an ensemble average from Gaussian simulations, as denoted by superscripts G and the bracket $\langle \cdot \rangle$.

The normalisation factor is given by

$$N = \sum_{l_j} \frac{h_{l_1 l_2 l_3}^2 b_{l_1 l_2 l_3}^2}{C_{l_1} C_{l_2} C_{l_3}}. \tag{5.2}$$

The core part of our estimation routine is the separable mode expansion of the shape function;

$$S(k_1, k_2, k_3) := (k_1 k_2 k_3)^2 B(k_1, k_2, k_3) = \sum_{p_j} \alpha_{p_1 p_2 p_3} q_{p_1}(k_1) q_{p_2}(k_2) q_{p_3}(k_3). \tag{5.3}$$

Choices for the basis functions $q_p(k)$ are detailed in the next section. Due to the separability, the reduced bispectrum reduces to a compact form of

$$b_{l_1 l_2 l_3} = \sum_{p_j} \alpha_{p_1 p_2 p_3} \int dr \tilde{q}_{p_1}(l_1, r) \tilde{q}_{p_2}(l_2, r) \tilde{q}_{p_3}(l_3, r), \tag{5.4}$$

where the *projected* mode functions are defined as

$$\tilde{q}_p(l, r) := \frac{2r^{\frac{2}{3}}}{\pi} \int dk q_p(k) \Delta_l(k) j_l(kr). \tag{5.5}$$

Radiative transfer functions $\Delta_l(k)$ and spherical Bessel functions $j_l(kr)$ are denoted the same way as in the previous chapter.

Every term appearing in (5.1) except the Gaunt integral is now separable. Using the definition $\mathcal{G}_{m_1 m_2 m_3}^{l_1 l_2 l_3} = \int d^2 \mathbf{n} Y_{l_1 m_1}(\mathbf{n}) Y_{l_2 m_2}(\mathbf{n}) Y_{l_3 m_3}(\mathbf{n})$, we can render it separable at the cost of introducing an extra integral.

We define the filtered maps as

$$M_p^{(i)}(\mathbf{n}, r) := \sum_{l,m} \frac{\tilde{q}_p(l, r)}{C_l} a_{lm}^{(i)} Y_{lm}(\mathbf{n}), \tag{5.6}$$

where $a_{lm}^{(i)}$'s are represent the spherical harmonic transform of the ith CMB map. For later convenience, we use a convention where $i = 0$ corresponds to the observed CMB map. Maps number 1-N_{sims} are Gaussian simulations. Note that without the factors involving $\tilde{q}$ and C_l's, M is simply equal to the original map in real space. Each mode extracts different anisotropy scales present on the map.

The bispectrum estimator (5.1) reduces to

$$\hat{f}_{\text{NL}}^{(i)} = \frac{1}{N} \sum_{p_j} \alpha_{p_1 p_2 p_3} (\beta_{p_1 p_2 p_3}^{\text{cub},(i)} - 3\beta_{p_1 p_2 p_3}^{\text{lin},(i)}), \tag{5.7}$$

where most of the computation required is now contained in the 'β's, given by

$$\beta_{p_1 p_2 p_3}^{\text{cub},(i)} := \int dr \int d^2 \mathbf{n} \, M_{p_1}^{(i)}(\mathbf{n}, r) M_{p_2}^{(i)}(\mathbf{n}, r) M_{p_3}^{(i)}(\mathbf{n}, r), \tag{5.8}$$

$$\beta_{p_1 p_2 p_3}^{\text{lin},(i)} := \frac{1}{N_{\text{sims}}} \sum_{j \neq i} \int dr \int d^2 \mathbf{n} \, M_{p_1}^{(j)}(\mathbf{n}, r) M_{p_2}^{(j)}(\mathbf{n}, r) M_{p_3}^{(i)}(\mathbf{n}, r). \tag{5.9}$$

Here we evaluate f_{NL} estimates for each of the Gaussian simulations $i = 1, \cdots, N_{\text{sims}}$. Naturally, they are normally distributed with a mean of zero. Under the null hypothesis that the initial fluctuations are purely Gaussian, the value of f_{NL} estimated from the observed CMB map is also drawn from the same normal distribution. Any statistically significant deviations from zero would therefore allow us to reject the null hypothesis.

It is important to note that the β matrices depend only on the choice of mode functions and input map data, and are independent of the theoretical bispectrum considered. Once β^{cub} and β^{lin} are computed and stored, we may constrain any model of interest by decomposing the template to get α. The f_{NL} estimate is then a simple dot product: $\alpha \cdot \beta / N$.

The normalisation can also be obtained in a similar fashion;

$$N = \sum_{p_j, p_j'} \alpha_{p_1 p_2 p_3} \Gamma_{p_1 p_2 p_3, p_1' p_2' p_3'} \alpha_{p_1' p_2' p_3'}, \tag{5.10}$$

or equivalently, $N = \alpha^T \Gamma \alpha$. We exploit separability once again to compute the Γ matrix;

$$\Gamma_{p_1 p_2 p_3, p_1' p_2' p_3'} := \int dr \int d\mu \, \mathscr{P}_{p_1 p_1'}(\mu, r, r') \mathscr{P}_{p_3 p_3'}(\mu, r, r') \mathscr{P}_{p_3 p_3'}(\mu, r, r'), \tag{5.11}$$

$$\mathscr{P}_{pp'}(\mu, r, r') := \sum_l \frac{2l+1}{(8\pi)^{1/3} C_l} \tilde{q}_{p'}(l, r) \tilde{q}_p(l, r') P_l(\mu), \tag{5.12}$$

where $P_l(\mu)$'s are the Legendre polynomials.

In summary, CMB-BEST computes the main quantities: β^{cub}, β^{lin}, and Γ. The most computationally expensive part is the linear term β^{lin} by a couple of orders of magnitude in most cases. Considerable effort has been made to optimise the corresponding part of the code, which will be detailed in the following sections.

5.1.2   Basis functions

One of the greatest strengths of CMB-BEST lies in its flexibility with the choice of mode functions. We decompose a given shape as a linear combination mode function in three-dimensional space. Hence, we shall refer to them as 'basis' functions. Adopting a specialised basis set provides optimised results to specific models of interest, whilst a more general construction of basis allows us to simultaneously constrain a wide range of models.

The main difference between CMB-BEST and the conventional Modal estimator [59] is that CMB-BEST has a single set of mode functions while Modal requires two; one in each of the primordial and late-time space (see section 3.3.3). While the Modal approach can accurately express a wide range of bispectra [60, 67, 70], the two-step modal expansion becomes less accurate for highly oscillatory shapes since the oscillations often get lost during the projection to late-time space [23, 95]. CMB-BEST avoids these issues in the second mode expansion by evaluating an exact analytical form analogous to the KSW estimator [58]. The results obtained from CMB-BEST is therefore as accurate as the primordial basis expansion.

We observe that the KSW estimator [58] is derived from a simple monomial basis in our notation;

$$q_p(k) = k^{p-1}, \quad p = 0, 1, 2, 3. \tag{5.13}$$

All three standard templates - local, equilateral, and orthogonal - can be expressed as a sum of separable terms in the form $q_{p_1}(k_1)q_{p_2}(k_2)q_{p_3}(k_3)$. The shape function of the local template, for example, is given by

$$S^{\text{local}}(k_1 k_2 k_3) := 2A^2 \left[\frac{k_1^2}{k_2 k_3} + \frac{k_2^2}{k_3 k_1} + \frac{k_3^2}{k_1 k_2} \right] \tag{5.14}$$

$$= 2A^2 \left[q_3(k_1)q_0(k_2)q_0(k_3) + q_0(k_1)q_3(k_2)q_0(k_3) + q_0(k_1)q_0(k_2)q_3(k_3) \right], \tag{5.15}$$

where A is the primordial power spectrum amplitude. Decomposition coefficients $\alpha_{p_1 p_2 p_3}$ have three non-zero components: $\alpha_{300} = \alpha_{030} = \alpha_{003} = 2A^2$. Coefficients for the equilateral and orthogonal templates can similarly be found.

We set the scalar spectral index as $n_s = 1$ above for simplicity. For a non-unit n_s, we modify the basis as follows;

$$q_p(k) = k^2 \left[k_* \left(\frac{k}{k_*} \right)^{(4-n_s)/3} \right]^{p-3}, \quad p = 0, 1, 2, 3. \tag{5.16}$$

The pivot scale k_* is defined such that the power spectrum evaluates to A at $k = k_*$. Including k_* here ensures that the $q_p(k)$'s have the right units. Note that the prefactor k^2 comes from the definition of the shape function and is therefore unaffected by n_s. We will refer to this choice of mode functions to be the 'KSW' basis.

For studying models with linearly spaced, high-frequency oscillation, a Fourier-like basis

$$q_0(k) = \sin(\omega k), \ q_1(k) = \cos(\omega k), \tag{5.17}$$

for a fixed ω is an appropriate choice. This is in fact equivalent to the method we used to study feature models in Chapter 4. The small size of the basis lets us efficiently constrain theoretical models with given characteristic scale ω. By scanning over a range of ω, or including modes with different values of ω in the basis, we may also perform a more comprehensive analysis of oscillatory features.

Lastly, for investigating a wide variety of models simultaneously, we construct large basis sets suitable for decomposing general shape functions. One such example is the following basis which consists of Legendre polynomials;

$$q_p(k) = P_p(\bar{k}), \ \text{where} \ \bar{k} = \frac{2k - k_{\min} - k_{\max}}{k_{\max} - k_{\min}}. \tag{5.18}$$

Here we linearly map the range from $k \in [k_{\min}, k_{\max}]$ to $\bar{k} \in [-1, 1]$, which is the interval where Legendre polynomials are defined. The number of modes are not bounded; the larger the basis, the more complete coverage of theory k space we get.

The Legendre basis functions are inherently orthogonal, so that $\int_{-1}^{1} d\bar{k} P_l(\bar{k}) P_{l'}(\bar{k}) = 0$ whenever $l \neq l'$. Orthogonality is essential for this class of basis sets since it allows us to greatly simplify the way of decomposing theoretical templates. We elaborate on the decomposition method in Section 5.1.3.

The Legendre basis also enables direct connection to PRIMODAL, a fast numerical code for computing primordial bispectra from general single-field inflation models [116]. PRIMODAL utilises the inherent separability present in the *in-in* formalism to efficiently compute the bispectrum using a separable basis. Using the Legendre polynomials for the basis is optimal for this step, but the methodology is not restricted to any particular choice of basis. The code then outputs the expansion coefficients of the primordial bispectrum with respect to the Legendre polynomials. Hence, the result can be plugged in directly to CMB-BEST, creating one fluid pipeline from the model Lagrangian to f_{NL} estimation.

As noted in [116], we may augment the base set of Legendre polynomials with one or more extra functions for a better description of some bispectrum templates. One solid choice

is to add a mode $q(k) = k^{n_{\mathrm{s}}-2}$, orthogonalised with respect to the rest of the basis. This significantly boosts the performance of decomposing local-type bispectra.

Ideally, we would like the k range used in the definition (5.18) to be as wide as possible so that more information from different scales is incorporated in the estimation process. A wider k range, however, also results in lower resolving power because the interval can fit more oscillations of a given frequency. Polynomials of higher degrees are necessary to handle the same bispectra. We found that $k_{\max}/k_{\min} = 1000$ is an overall sweet spot for analysing Planck data.

Throughout the rest of this thesis, we refer to a basis with the $k^{n_{\mathrm{s}}-2}$ mode and the first 29 Legendre polynomials, totalling 30 modes, defined in the k range with $(k_{\min}, k_{\max}) = (2.09 \times 10^{-4}, 2.09 \times 10^{-1})$, as the 'Legendre' basis. Any deviation from this set of parameters will be stated explicitly. Note that this is for our convenience during testing only; the code may take any choice of basis.

5.1.3 Primordial basis expansion

Our formalism assumes that the bispectrum shape of interest can be accurately represented as a linear combination of a chosen basis set;

$$S(k_1, k_2, k_3) = \sum_{p_j} \alpha_{p_1, p_2, p_3} \, Q_{p_1 p_2 p_3}(k_1, k_2, k_3), \qquad (5.19)$$

where the three-dimensional mode functions are defined as

$$Q_{p_1 p_2 p_3}(k_1, k_2, k_3) := q_{p_1}(k_1) \, q_{p_2}(k_2) \, q_{p_3}(k_3). \qquad (5.20)$$

For some models, the coefficients $\alpha_{p_1 p_2 p_3}$ are obtained analytically. The local shape function with respect to the KSW basis is a simple example; $\alpha_{300} = \alpha_{030} = \alpha_{003} = 2A^2$ and zero otherwise. For other models, the shape function needs to be expanded with respect to the chosen basis.

Shape functions are defined on the same domain as bispectra: $(k_1, k_2, k_3) \in \mathbb{R}^3$ where k_1, k_2, and k_3 form a triangle. We cannot observe scales smaller than a certain size in practice, which places an upper bound on k: $k < k_{\max}$. The resulting domain in three dimensions is shown in Figure 5.1.

We follow [59] and denote this domain consisting of two tetrahedrons glued together as the 'tetrapyd'. Note that all shape functions of interest are symmetric in the k's, so we may restrict our interest to a region where $k_1 \geq k_2 \geq k_3$ (right of Figure 5.1). We dub the resulting tetrahedron with one-sixth the original volume the 'sliced' tetrapyd, somewhat
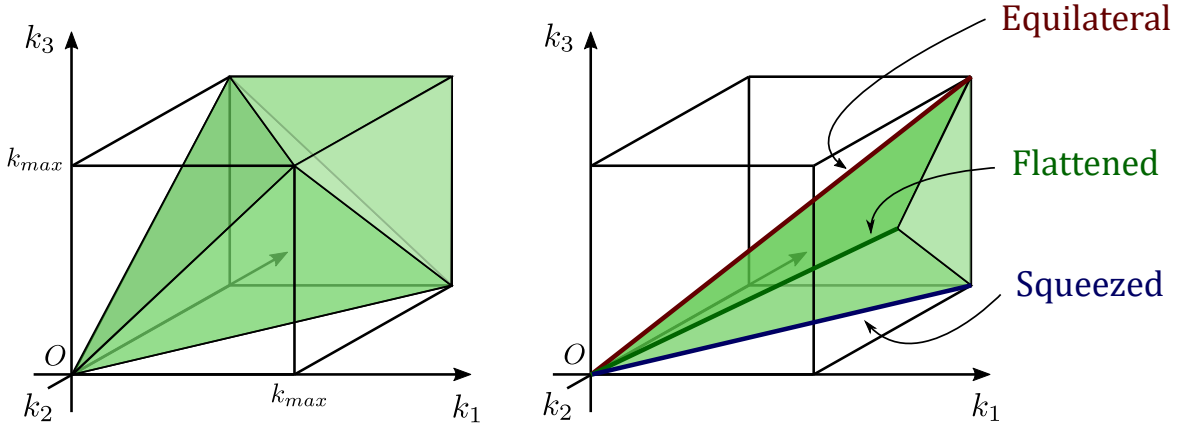
Fig. 5.1 Tetrapyd domain in which the shape functions are defined. Left: the full region specified by triangle inequalities and an upper bound k_{max}. Right: one-sixth slice of the tetrapyd from which shape functions are uniquely determined by symmetry in the k's. Edges representing the equilateral, flattened and squeezed configurations are annotated.

unoriginally. The three main types of triangle configurations correspond to three edges of the sliced tetrapyd: equilateral ($k_1 = k_2 = k_3$), flattened ($k_1 = k_2 + k_3$), and squeezed ($k_1 = k_2 \gg k_3$). [2] Figure 5.2 shows some of the Legendre basis functions plotted in three dimensions.

Let $V_{\mathbf{k}}$ be the subset of $\mathbb{R}^3$ representing the sliced tetrapyd. By choosing a natural inner product

$$\langle S_1, S_2 \rangle = \int_{V_{\mathbf{k}}} d^3\mathbf{k} \ S_1(\mathbf{k}) S_2(\mathbf{k}), \tag{5.21}$$

we restrict our attention to an L^2 space defined on $V_{\mathbf{k}}$: a vector space consisting of square-integrable functions. [3] A set of three-dimensional basis functions $Q_{\mathbf{p}}$ with size N spans a subspace $U_Q \subset L^2(V_{\mathbf{k}})$ of dimension at most N.

Using this notation, expanding a given shape function S with respect to a basis set Q is equivalent to finding its orthogonal projection into U_Q;

$$S = S^{\parallel} + S^{\perp}, \quad \text{where} \tag{5.22}$$

$$S^{\parallel} \in U_Q \quad \text{and} \quad \left\langle S^{\perp}, Q' \right\rangle = 0, \ \forall Q' \in U_Q. \tag{5.23}$$

---

[2] To be precise, the edges for flattened represent $k_1/2 = k_2 = k_3$, and the squeezed one is defined as $k_1 = k_2$, $k_3 = 0$. The edges by themselves are unphysical, but points near these lines correspond to the limits described.

[3] Note that the local shape function $S(k_1, k_2, k_3) = (k_1^2/k_2 k_3 + 2 \text{ cyc.})$ is not in fact square-integrable due to its divergence as $k_i \to 0$. We work around this problem by prescribing a lower bound $k \geq k_{min}$, which is also the case in numerical calculations.
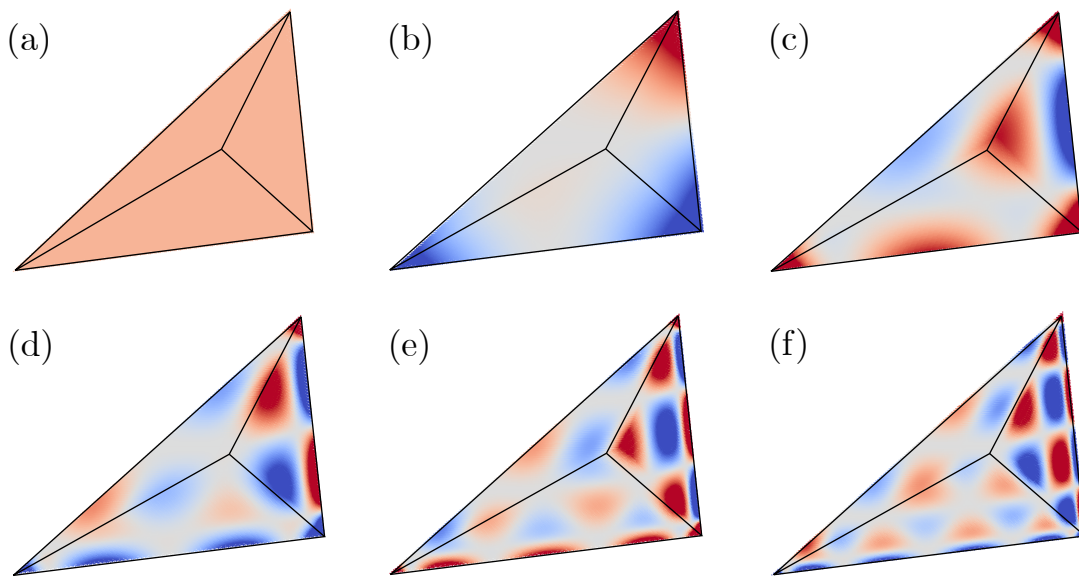
Fig. 5.2 Examples of our Legendre basis functions, evaluated on the sliced tetrapyd domain shown in Figure 5.1. Functions are defined as $Q_{p_1 p_2 p_3}(k_1, k_2, k_3) := P_{p_1}(k_1) P_{p_2}(k_2) P_{p_3}(k_3)$, where $P_l(k)$ are Legendre polynomials. Here we plot $p_1 = p_2 = p_3 = p$, where p equals (a) 0, (b) 1, (c) 2, (d) 3, (e) 4, and (f) 5. A single colour map is used across the plots: red and blue correspond to $+1$ and -1, respectively.

As long as $\left\|S^{\perp}\right\| \ll \|S\|$, we may approximate

$$S \approx S^{\|} = \sum_{\mathbf{p}} \alpha_{\mathbf{p}} Q_{\mathbf{p}}. \tag{5.24}$$

The decomposition coefficients α are obtained by taking inner product with $Q_{\mathbf{p}}$ on both sides.

$$\langle S, Q_{\mathbf{p}} \rangle = \langle Q_{\mathbf{p}}, Q_{\mathbf{p}'} \rangle \alpha_{\mathbf{p}'}. \tag{5.25}$$

From above, we can compute $\alpha_{\mathbf{p}}$ by inverting the matrix $\Gamma_{\mathbf{p}\mathbf{p}'} = \langle Q_{\mathbf{p}}, Q_{\mathbf{p}'} \rangle$. Alternatively, orthonormalising the basis with respect to the inner product $\langle, \rangle$ turns Γ to an identity matrix, trivialising the inversion process.

We rewrote the formalism using linear analysis language here to emphasize two aspects of our primordial basis expansion. First, even though we chose a simple inner product (5.21) for now, the method is completely general and we may freely choose a different inner product. For example, non-unit weights $w(\mathbf{k})$ can be included in the integrand and/or the domain of integration may be changed. Doing so alters how we decompose shape functions in terms of our basis.

Second, we highlight the fact that basis expansion is in essence a minimisation problem with respect to the chosen inner product, which is often oblivious of late-time physics. The orthogonal projection of S to subspace U_Q is a point in U_Q which has the minimal distance to S: $\left\|S - S^{\|}\right\|$ is minimised. The obtained $S^{\|}$, however, is not necessarily the best description of S when we consider their late-time counterparts in l space. Some errors with small norm $\|\Delta S\|$ might get enhanced when convolved with transfer functions, yielding large reduced bispectrum $\Delta b_{l_1 l_2 l_3}$. Meanwhile, sometimes large differences in S are completely unobservable late-time and provide mostly identical constraints on f_{NL}.

With these points in mind, we define the following metrics to compare shape functions in primordial space;

$$\mathrm{Corr}(S_1, S_2) := \frac{\langle S_1, S_2 \rangle}{\sqrt{\langle S_1, S_1 \rangle \langle S_2, S_2 \rangle}}, \tag{5.26}$$

$$\varepsilon^2(S_1, S_2) := \frac{\langle S_1 - S_2, S_1 - S_2 \rangle}{\sqrt{\langle S_1, S_1 \rangle \langle S_2, S_2 \rangle}}. \tag{5.27}$$

$\mathrm{Corr}(S_1, S_2)$ measures correlation between the two shapes and is independent of the normalisation. $\varepsilon(S_1, S_2)$ is the distance between the two shapes, which reduces to $\sqrt{2 - 2\,\mathrm{Corr}(S_1, S_2)}$ when S_1 and S_2 have equal amplitude. We use both of these values to test if our primordial basis expansion has converged to the target shape function.

Inverting the matrix $\Gamma_{\mathbf{pp'}} = \langle Q_{\mathbf{p}}, Q_{\mathbf{p'}} \rangle$ in (5.25) can be tricky in practice. The $p_{\max}^3 \times p_{\max}^3$ matrix is not only large in size for high $p_{\max}$, but also near singular. Some of the three-dimensional mode functions $Q_{\mathbf{p}}$ become linearly dependent within tetrapyd even if they are constructed from independent modes in one dimension. The presence of such degenerate modes introduces severe numerical instability during inversion.

For our Legendre basis, we use a small trick to get around this problem. Motivated by the fact that Legendre polynomials form an orthogonal basis in one dimension, we extend the domain of integral in the inner product (5.21) to a cube containing the tetrapyd instead. Most template shape functions have an analytic form and can easily be extended in this region. As long as $\int dk \, q_p(k) q_{p'}(k) = \delta_{pp'}$, the decomposed coefficients for a given S can be found by

$$\alpha_{p_1 p_2 p_3}^{(t)} = \int dk_1 \int dk_2 \int dk_3 \, S^{(t)}(k_1, k_2, k_3) q_{p_1}(k_1) q_{p_2}(k_2) q_{p_3}(k_3). \qquad (5.28)$$

Note that the three-dimensional integral can now be split into three separate one-dimensional integrals in range $[k_{\min}, k_{\max}]$. By performing the three integrals one by one, we obtain a numerically stable algorithm that is fast and memory efficient. Decomposing a shape evaluated on 1000^3 k grid points with respect to 30^3 Legendre basis functions only takes a few seconds in our C code.

After we obtain coefficients from (5.28), we evaluate the accuracy of the expansion using the inner product defined in tetrapyd because it is the only physical region where our shape function matters. Since the cube includes tetrapyd, good convergence in the cube guarantees a small error within the tetrapyd in most cases. The only exception is when the target shape function blows up outside the tetrapyd. We will see such examples in later sections.

5.2 Implementation and optimisation

CMB bispectrum estimation is a numerically challenging task. All existing approaches exploit the separability of multi-dimensional integrals to reduce computational complexity because it is practically impossible otherwise. Planck analysis provided constraints to primordial non-Gaussianity from various bispectrum templates, but having limited computing resources was what stopped us from exploring further. Numerous shapes such as complex oscillatory models have been outside our reach, despite being theoretically well-motivated.

The CMB-BEST formalism significantly reduces the amount of computation needed for CMB bispectrum estimation. Obtaining the linear term β^{lin} in (5.9), however, is still prohibitively expensive unless thoroughly optimised. Performance is especially crucial here

since it directly affects the breadth of models we can cover; the faster the code, the higher number of modes we can have, and so the more shapes we can expand in our basis accurately.

In this section, we provide details for various aspects of our optimisation process: algorithm design, parallel computing, and data locality improvements. The final specifications and data files used are outlined at the end.

Throughout this section, we will treat the functions of interest as discrete arrays. Our notation for indices and their limits are summarised in Table 5.1. We adopt a simple trapezoidal rule for most numerical integrals. For the μ integral in (5.11), however, we use the Gauss-Legendre quadrature computed from the public code QUADPTS [108] due to the highly oscillatory and unstable nature of the integral. Multi-dimensional arrays are stored in the row-major order following the C convention for efficiency. We use the HEALPIX library [117] for pixelisation of the sky, which includes the LIBSHARP [118] library for Spherical Harmonic Transforms (SHTs).

Table 5.1 Our index conventions for discretised arrays and their sizes.

Index	Range	Description
r	$[0, N_r)$	Line-of-sight integral r grid index.
p, p_j	$[0, p_{max})$	Mode number. p_j is a shorthand for (p_1, p_2, p_3).
i, j	$[0, N_{sims}]$	Map number. Index $i = 0$ corresponds to the observed CMB map, while $i > 0$ are for simulated Gaussian maps.
n	$[0, N_{pix})$	Map pixel number.
l, m	$[0, l_{max})$	Spherical harmonic multipole moments. Note $-l \leq m \leq l$.
μ	$[0, N_\mu)$	Gauss-Legendre quadrature μ grid index.

5.2.1 Algorithm

Our goal is to compute three key quantities: Γ (5.11), β^{cub} (5.8), and β^{lin} (5.9). The matrix Γ allows us to find the normalisation factor N for a given theoretical template, while the two β's provide the amplitude of f_{NL} for each of the CMB maps and simulations used up to normalisation.

In most cases of interest, the bottleneck point of our pipeline is computing the linear term β^{lin}. Even though the Γ matrix computation through (5.11) grows more rapidly with the number of basis functions ($\propto p_{max}^6$) than the β's ($\propto p_{max}^3$), it does not involve operations with high-definition maps and remains subdominant in terms of the total cost.

We dedicate this section to explaining our algorithm design for β computation in detail. The discretised versions of (5.8-5.9) are given by

$$\beta^{\text{cub}}(i, p_1, p_2, p_3) = \sum_r \sum_n M(r, i, p_1, n) \cdot M(r, i, p_2, n) \cdot M(r, i, p_3, n), \qquad (5.29)$$

$$\beta^{\text{lin}}(i, p_1, p_2, p_3) = \sum_r \sum_{j \neq i} \sum_n M(r, j, p_1, n) \cdot M(r, j, p_2, n) \cdot M(r, i, p_3, n). \qquad (5.30)$$

The order of indices is chosen such that later calculations have optimal memory layouts. Some integral weights and factors are absorbed into arrays for brevity.

Note that the data arrays for different values of r are completely independent of each other. This provides us with a natural way to distribute tasks. We compute and save contributions to β's for each r separately. The summation over r is performed at the end. Therefore, throughout the rest of this chapter, we assume that r is fixed and drop the r dependence in the descriptions of our algorithms.

The filtered map arrays $M(i, p, n)$ are obtained as follows. A given map i is first transformed into spherical harmonic coefficients $a^{(i)}(l, m)$s via SHT. We then compute $\tilde{q}_{(}p, l) * a(i, l, m)/C(l)$ from (5.6), which is fed into reverse SHT to synthesise the filtered maps.

As a rough guide to the size of each summation, we typically have $N_{\text{sim}} \approx 150$ simulations, $p_{\text{max}} = 30$ modes, and $N_{\text{pix}} = 50,331,648$ pixels. [4] Considering the fact that one double-precision array of size ~ 50 million pixels takes about 400MB of memory space, this is indeed a task for supercomputers.

Our first and most straightforward method of computing βs is outlined in Algorithm 1. The computational complexity of each innermost loop is annotated on the right-hand side.

Here, we first obtain the filtered maps $M(i, p, n)$ via SHT for each map and mode. The full results are stored in memory. Next, we iterate through each map and set of modes (p_1, p_2, p_3) and sum over each of the map pixels to obtain β^{cub} and β^{lin}. Symmetries in the indices are respected; we only loop over (p_1, p_2, p_3) satisfying $p_1 \geq p_2$ for the linear terms, and $p_1 \geq p_2 \geq p_3$ for the cubic ones.

SHT typically scales as $\propto l_{\text{max}}^3$, where l_{max} is the maximum degree of spherical harmonic functions used. The total number of pixels N_{pix} grows $\propto l_{\text{max}}^2$ when chosen appropriately for resolution, hence the SHT costs $O(N_{\text{sims}} \cdot p_{\text{max}} \cdot N_{\text{pix}}^{3/2})$. The most expensive part of this algorithm is still the loop where we calculate the linear term, which scales as $O(N_{\text{sims}}^2 \cdot p_{\text{max}}^3 \cdot N_{\text{pix}})$.

---

[4]This value corresponds to $N_{side} = 2048$ in Healpix. $N_{\text{pix}} = 12N_{side}^2$

Algorithm 1 Computing βs: the naïve method

1: Allocate $M(i,p,n)$ $\triangleright$ Memory $\sim N_{\text{sims}} \cdot p_{\text{max}} \cdot N_{\text{pix}}$
2:
3: **for** each map i **do**
4:     **for** each mode p **do** $\triangleright$ $O(N_{\text{sims}} \cdot p_{\text{max}} \cdot N_{\text{pix}}^{3/2})$
5:         **compute** $M(i,p,n)$ by SHT
6:     **end for**
7: **end for** $\triangleright$ $M(i,p,n)$ ready
8:
9: **for** each map i **do**
10:     **for** each set of modes (p_1,p_2,p_3) **do**
11:         **for** each pixel n **do** $\triangleright$ $O(N_{\text{sims}} \cdot p_{\text{max}}^3 \cdot N_{\text{pix}})$
12:             $\beta^{\text{cub}}(i,p_1,p_2,p_3) \mathrel{+}= M(i,p_1,n) \cdot M(i,p_2,n) \cdot M(i,p_3,n)$
13:         **end for**
14:     **end for**
15: **end for**
16:
17: **for** each map i **do**
18:     **for** each map $j \neq i$ **do**
19:         **for** each set of modes (p_1,p_2,p_3) **do**
20:             **for** each pixel n **do** $\triangleright$ $O(N_{\text{sims}}^2 \cdot p_{\text{max}}^3 \cdot N_{\text{pix}})$
21:                 $\beta^{\text{lin}}(i,p_1,p_2,p_3) \mathrel{+}= M(j,p_1,n) \cdot M(j,p_2,n) \cdot M(i,p_3,n)$
22:             **end for**
23:         **end for**
24:     **end for**
25: **end for**

Our first major optimisation comes from the observation that we may swap the order of summations to reduce the computation. By precomputing $C(p_1, p_2, n) = \sum_j M(j, p_1, n) \cdot M(j, p_2, n)$, we require one fewer loop over maps for β^{lin}, leading to a factor of $N_{\text{sims}} \sim 150$ improvement. Algorithm 2 shows a pseudocode for this method.

Algorithm 2 Computing βs: optimised for computation

1: Allocate $M(i, p, n)$ ▷ Memory $\sim N_{\text{sims}} \cdot p_{\text{max}} \cdot N_{\text{pix}}$
2: Allocate $C(p_1, p_2, n)$ ▷ Memory $\sim p_{\text{max}}^2 \cdot N_{\text{pix}}$
3:
4: **for** each map i **do**
5:      **for** each mode p **do**   ▷ $O(N_{\text{sims}} \cdot p_{\text{max}} \cdot N_{\text{pix}}^{3/2})$
6:          **compute** $M(i, p, n)$ by SHT
7:      **end for**
8: **end for**   ▷ $M(i, p, n)$ ready
9:
10: **for** each map j **do**
11:      **for** each pair of modes (p_1, p_2) **do**
12:          **for** each pixel n **do**   ▷ $O(N_{\text{sims}} \cdot p_{\text{max}}^2 \cdot N_{\text{pix}})$
13:              $C(p_1, p_2, n) \mathrel{+}= M(j, p_1, n) \cdot M(j, p_2, n)$
14:          **end for**
15:      **end for**
16: **end for**   ▷ $C(p_1, p_2, n)$ ready
17:
18: **for** each map i **do**
19:      **for** each set of modes (p_1, p_2, p_3) **do**
20:          **for** each pixel n **do**   ▷ $O(N_{\text{sims}} \cdot p_{\text{max}}^3 \cdot N_{\text{pix}})$
21:              $\beta^{\text{cub}}(i, p_1, p_2, p_3) \mathrel{+}= M(i, p_1, n) \cdot M(i, p_2, n) \cdot M(i, p_3, n)$
22:              $\beta^{\text{lin}}(i, p_1, p_2, p_3) \mathrel{+}= C(p_1, p_2, n) \cdot M(i, p_3, n)$
23:          **end for**
24:      **end for**
25: **end for**

Note that the resulting sum of β^{lin} includes an unwanted contribution from the case where $j = i$, since $C(p_1, p_2, n)$ is obtained by summing over all maps. Thankfully, this extra contribution is exactly equal to the cubic term. We just have to subtract β^{cub} from the total sum to get the correct value of β^{lin}.

We shaved off a whole loop at the cost of extra memory usage. For our purposes $N_{\text{sims}} \sim 150 > p_{\text{max}} = 30$, so the additional space required for $C(p_1, p_2, n)$ is relatively small compared to $M(i, p, n)$. The symmetry in p_1 and p_2 also means that we only need to store $p_{\text{max}}(p_{\text{max}} + 1)/2$ maps instead of p_{max}^2 for C.

Our next challenge is the amount of memory required to save $M(i,p,n)$. For the parameters mentioned above, this is ≈ 1.8TB, which is quite significant. Although it is not impossible to find supercomputing systems which can accommodate such large arrays, reducing the memory usage would be highly beneficial. We seek to achieve this whilst keeping the overall computational complexity to be $O(N_{\mathrm{sims}} \cdot p_{\mathrm{max}}^3 \cdot N_{\mathrm{pix}})$ as before.

We observe that most summations are done for a single map i. The array $C(p_1,p_2,n)$ is required to be computed before the main loop for β^{lin}, but otherwise there are no 'mixing' between the maps. We exploited this fact to develop Algorithm 3.

Algorithm 3 Computing βs: fast and memory efficient

1: Allocate $m(p,n)$ $\triangleright$ Memory $\sim p_{\mathrm{max}} \cdot N_{\mathrm{pix}}$
2: Allocate $C(p_1,p_2,n)$ $\triangleright$ Memory $\sim p_{\mathrm{max}}^2 \cdot N_{\mathrm{pix}}$
3:
4: **for** each map i **do**
5:    **for** each mode p **do** $\triangleright$ $O(N_{\mathrm{sims}} \cdot p_{\mathrm{max}} \cdot N_{\mathrm{pix}}^{3/2})$
6:      **compute** $M(i,p,n)$ by SHT and store in $m(p,n)$
7:    **end for**
8:
9:    **for** each pair of modes (p_1,p_2) **do**
10:      **for** each pixel n **do** $\triangleright$ $O(N_{\mathrm{sims}} \cdot p_{\mathrm{max}}^2 \cdot N_{\mathrm{pix}})$
11:        $C(p_1,p_2,n) \mathrel{+}= m(p_1,n) \cdot m(p_2,n)$
12:      **end for**
13:    **end for**
14: **end for** $\triangleright$ $C(p_1,p_2,n)$ ready
15:
16: **for** each map i **do**
17:    **for** each of mode p **do** $\triangleright$ $O(N_{\mathrm{sims}} \cdot p_{\mathrm{max}} \cdot N_{\mathrm{pix}}^{3/2})$
18:      **compute** $M(i,p,n)$ by SHT and store in $m(p,n)$
19:    **end for**
20:
21:    **for** each set of modes (p_1,p_2,p_3) **do**
22:      **for** each pixel n **do** $\triangleright$ $O(N_{\mathrm{sims}} \cdot p_{\mathrm{max}}^3 \cdot N_{\mathrm{pix}})$
23:        $\beta^{\mathrm{cub}}(i,p_1,p_2,p_3) \mathrel{+}= m(p_1,n) \cdot m(p_2,n) \cdot m(p_3,n)$
24:        $\beta^{\mathrm{lin}}(i,p_1,p_2,p_3) \mathrel{+}= C(p_1,p_2,n) \cdot m(p_3,n)$
25:      **end for**
26:    **end for**
27: **end for**

Algorithm 3 dramatically reduces the amount of memory required, at the cost of doubling the SHTs for computing $M(i,p,n)$s. The first time through, SHT results from each map are

used to find $C(p_1, p_2, n)$. After a full loop over maps we have $C(p_1, p_2, n)$ ready, another set of SHTs for each map allows us to obtain the β's.

SHTs have a subdominant contribution to the total computation time even after becoming doubled in number. One of the main strengths of Algorithm 3 is that both the memory and computation time scale linearly with the number of simulations used, N_{sims}. In the future when a larger number of Gaussian simulations are required to acquire a more accurate estimate of the linear term, it is straightforward to adapt our method accordingly.

We choose Algorithm 3 to be our main method of computing β's for CMB-BEST.

5.2.2 Parallel computing

In order to fully benefit from modern computer architecture, parallel computing is essential. In this section, we outline how CMB-BEST was parallelised for optimal performance on supercomputers.

Following [119], we discuss parallelisation at three different levels: domain, thread, and data. They mostly correspond to nodes, cores, and registers in modern computer clusters. Each level has distinct characteristics which make them ideal for different parallelisation techniques. We make full use of each level in our methodology.

Domain parallelism refers to dividing the task into many domains where each domain entails heavy computation while having limited data communication between them. Since the domains are largely independent, Message Passing Interface (MPI) [120] is a suitable tool for parallelisation. MPI is a communication protocol where many instances ('ranks') of the same code may transfer data while running separately from each other À perfect example of domain parallelism in CMB-BEST would be our line-of-sight integration over r. Almost no data is shared between different rs, despite the heavy computational cost from map operations within each domain. CMB-BEST therefore scales well with the number of MPI ranks and requires minimal effort to parallelise using MPI.

Thread parallelism opportunities arise when there are many independent computational tasks on a single set of data. Most modern supercomputers use multi-core processors. Each core, or processing unit, can run one or more threads, executing instructions independently from each other while sharing memory space. Note that MPI would not be as effective here due to a large amount of data sharing needed; ranks have to either continuously communicate with each other, or store individual copies of the data. This type of parallelism is extremely common in many applications, including SHTs and map integrations in CMB-BEST. Open Multi-Processing (OpenMP) [121] is one of the most popular implementations of multi-threading publicly available. We utilise the C OpenMP interface in our code.

Data parallelism applies when the same arithmetic operations are applied to multiple data items, ideally adjacent in memory space. This type of parallel processing is called SIMD: Single Instruction, Multiple Data. Many procedures involving large arrays fall into this category. We use Advanced Vector Extensions (AVX) supported by Intel processors to exploit data parallelism in CMB-BEST, especially for the operations involving large (filtered) map arrays. In particular, Intel's Xeon Phi series implement AVX-512, where 512-bit registers hold up to 8 double-precision floating numbers [119]. A single instruction can be applied to all of them at once in a single clock cycle, providing a major boost to the computation speed. We explicitly align large arrays in memory for optimal vectorisation performance.

Table 5.2 summarises the discussions in this section.

Table 5.2 Three levels of parallelism utilised in CMB-BEST. The levels and characteristics are chosen based on [119].

	Characteristics	Methods	Utilisation
Domain Parallelism	• Limited data communication between domains • Heavy computation within the domain	MPI	• Line of sight r integration split into independent tasks
Thread Parallelism	• Data sharing between threads • High proportion of independent tasks	OpenMP	• SHT • Map patches divided across threads
Data Parallelism	• Same operation applied to multiple data items	SIMD	• Map pixel integration vectorised using AVX-512 • Data vectors explicitly aligned in memory for optimal performance

5.2.3 Data locality

We implemented Algorithm 3 and profiled it using the Intel VTune Amplifier. Our program was found to be memory-bound, which means that its speed is limited mainly by the amount of memory available and the speed of memory access. The CPU speed, rate of the file I/O, and MPI communication all have subdominant contributions in comparison. This is somewhat expected since our method deals with large map arrays. The number of operations on each data element is small compared to the size of data, causing the CPUs to be 'starved' for data to work on most of the time. Our final set of optimisation focusses on improving memory access patterns.

CPUs of most modern computers contain a small amount of memory attached to them called *cache*. Recently used data and instructions are stored in cache memory so that reusing them is more efficient; accessing them is much faster than loading from the main, larger memory often shared with other CPUs. A cache is often divided into multiple levels. The smallest and fastest is the L1 cache, which is the first level a CPU checks for data. When the required data is not stored in the L1 cache, a *cache miss* occurs. The system then has to look further down the cache levels to fetch the wanted data, incurring a large time loss. It is therefore crucial to reuse data as many times as possible before it is lost from the cache.

Furthermore, the cache 'caches' memory locations in units of cache lines, or chunks of memory containing multiple data elements. Accessing memory locally therefore significantly increases the chance of cache hits, boosting overall performance. Array operations, for example, are optimised when they are stored in consecutive memory locations due to this fact.

CMB-BEST has been modified in two ways to improve data locality and memory performance. The first one was simple yet effective; we made sure to initialise large arrays within the same OpenMP construct as the main computation loop. This guarantees the physical memory of array elements to be allocated near the cores where they are going to be used. Memory access during the main computation loop is therefore much faster than it would be otherwise. Systems with non-uniform memory access (NUMA) especially benefit from this method. For CMB-BEST, we gained a two times speed-up compared to when a single master thread initialised the entire array.

The second optimisation centres around *cache blocking*, a technique used to maximise data reuse. Let us focus on the main computation loop in Algorithm 3:

for each set of modes (p_1, p_2, p_3) **do**
    for each pixel n **do**
        $\beta^{\text{cub}}(i, p_1, p_2, p_3) \mathrel{+}= m(p_1, n) \cdot m(p_2, n) \cdot m(p_3, n)$
        $\beta^{\text{lin}}(i, p_1, p_2, p_3) \mathrel{+}= C(p_1, p_2, n) \cdot m(p_3, n)$

       end for
   end for

In every loop over the set of modes (p_1, p_2, p_3), four large arrays are read from memory: $m(p_1, \cdot)$, $m(p_2, \cdot)$, $m(p_3, \cdot)$, and $C(p_1, p_2, \cdot)$. Each of them takes up around 400MB of memory. Since their size is greater than the cache storage capacity, data in front of them are gone from the cache by the time a loop over n completes. All four arrays will then have to be loaded from the main memory again when the next iteration starts.

Suppose that the map arrays had smaller sizes and all of them fit in one of the cache levels instead. After one loop iteration over n for (p_1, p_2, p_3), the arrays now remain in the cache. Then the next loop on $(p_1, p_2, p_3 + 1)$ begins, and the arrays $m(p_1, \cdot)$, $m(p_2, \cdot)$, and $C(p_1, p_2, \cdot)$ are required again for computation. They can now be accessed from the cache, which saves a significant amount of time. We see that each data element is reused multiple times before being discarded from the cache in this case.

Motivated by these facts, we divide the map array into equally sized blocks so that each patch can now fit inside the cache memory. We restructure the loop as follows.

  for each block b **do**
     for each set of modes (p_1, p_2, p_3) **do**
        for each pixel n' in block **do**
           $\beta^{\mathrm{cub}}(i, p_1, p_2, p_3) \mathrel{+}= m(p_1, n') \cdot m(p_2, n') \cdot m(p_3, n')$
           $\beta^{\mathrm{lin}}(i, p_1, p_2, p_3) \mathrel{+}= C(p_1, p_2, n') \cdot m(p_3, n')$
        end for
     end for
  end for

New pixel numbers are calculated as $n' = B \cdot b + n$, where B is the size of each block and $0 \le n < B$. We have not changed the total number of arithmetic operations required, so the computational complexity remains constant. Meanwhile, data locality within each block is greatly improved, as each of the blocked arrays is now small enough to fit in the cache. Each data element is accessed in closer succession temporarily as well.

One caveat here is that having too many blocks may degrade the overall performance. There exists a non-negligible overhead coming from an extra *for* loop and the OpenMP construct used over n'. The block size divided by the number of cores should not be smaller than the size of cache lines either. The optimal size of cache blocks depends on the memory architecture of the processor used. This often needs to be found empirically.

Figure (5.3) shows how the runtime of a small test code varies with the number of cache blocks used. The code was run on the Intel Gold 6154 processor, which is an 18-core high-performance chip with total L1, L2, and L3 cache memory sizes of 1.125MiB, 18MiB,

and 24.74MiB, respectively [122]. Even though some parameters such as N_{sims} are tuned down for faster testing, this is representative of the full computation.
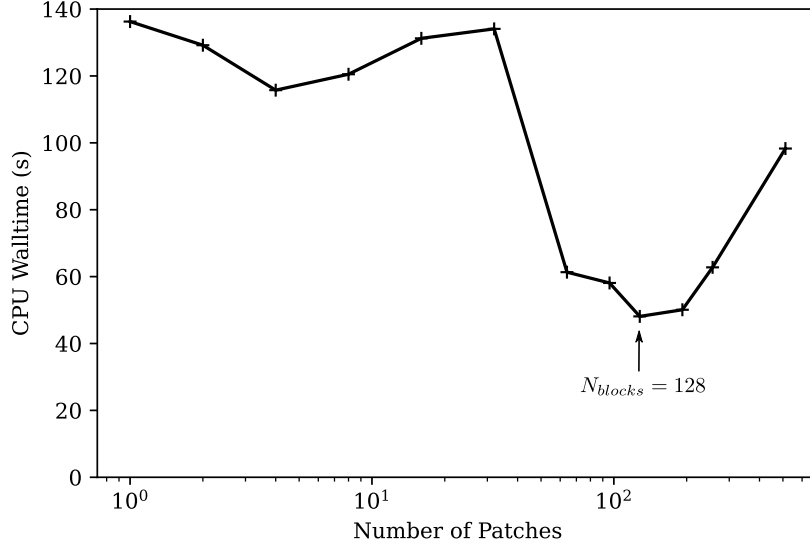
Fig. 5.3 Cache block optimisation. The total wall time of a test code for computing βs was measured while varying the number of cache blocks used. Having 128 blocks is optimal for the 18-core Intel Gold 6154 processor.

We find a significant improvement in performance moving from 32 to 48 blocks when the block size drops from 12MiB to 8MiB. This is precisely when three blocks are allowed to fit in the L3 cache (24.7MiB) at the same time. Improved data locality dramatically reduces time spent on memory access as we expected. The optimal number of blocks was found to be 128; each segment contains about 400,000 elements and takes up 3MiB of memory, which is 1/6 and 1/8 of L2 and L3 cache size, respectively. We gain roughly three times speed-up compared to the original code without cache blocking.

Algorithm 4 summarises our final implementation of β computation code, now with cache blocking and OpenMP construct indicators.

5.3 Validation

CMB bispectrum estimation is not only computationally challenging but also prone to numerical instabilities unless implemented carefully. We invested a considerable amount of time after the development of CMB-BEST validating various aspects of the code. We

---

Algorithm 4 Computing βs: our final implementation

1: Allocate $m(p,n)$
2: Allocate $C(p_1,p_2,n)$
3:   ▷ Both initialised within OpenMP construct over n
4: **for** each map i **do**
5:     **for** each mode p **do**
6:         **compute** $M(i,p,n)$ by SHT and store in $m(p,n)$ ▷ OpenMP within SHT
7:     **end for**
8:
9:     **for** each block b **do**
10:         **for** each pair of modes (p_1,p_2) **do**
11:             **for** each pixel n' in block **do**   ▷ OpenMP *for* construct
12:                 $C(p_1,p_2,n') \mathrel{+}= m(p_1,n') \cdot m(p_2,n')$
13:             **end for**
14:         **end for**
15:     **end for**
16: **end for**   ▷ $C(p_1,p_2,n)$ ready
17:
18: **for** each map i **do**
19:     **for** each of mode p **do**
20:         **compute** $M(i,p,n)$ by SHT and store in $m(p,n)$ ▷ OpenMP within SHT
21:     **end for**
22:
23:     **for** each block b **do**
24:         **for** each set of modes (p_1,p_2,p_3) **do**
25:             **for** each pixel n' in block **do**   ▷ OpenMP *for* construct
26:                 $\beta^{\mathrm{cub}}(i,p_1,p_2,p_3) \mathrel{+}= m(p_1,n') \cdot m(p_2,n') \cdot m(p_3,n')$
27:                 $\beta^{\mathrm{lin}}(i,p_1,p_2,p_3) \mathrel{+}= C(p_1,p_2,n') \cdot m(p_3,n')$
28:             **end for**
29:         **end for**
30:     **end for**
31: **end for**

---

highlight some of our validation efforts in this section, checking consistency within the program itself (section 5.3.1) and against existing code such as Modal [60] (section 5.3.2).

5.3.1   Internal consistency checks

CMB-BEST is a general code where one can freely choose a set of basis functions. Three of our main options are the 'KSW' basis (5.16), 'Legendre' basis (5.18, augmented with $q(k) = k^{n_s - 2}$) as discussed previously in Section 5.1.2, and Fourier basis (5.17). Both KSW and Legendre basis sets can cover the standard templates: local, equilateral, and orthogonal (see e.g., [70] for definitions). The KSW basis provides an exact form for the three templates by choosing appropriate powers of k as its basis elements. On the other hand, the templates are expanded in terms of separable Legendre polynomials up to some fixed degree p_{max} for the Legendre basis. As long as p_{max} is sufficiently large, most smooth bispectrum shapes can be represented accurately.

Our first consistency check is shown in Figure 5.4, where we compare the f_{NL} estimates from the Planck 2018 CMB map and 140 full focal plane (FFP10) realistic Gaussian simulations [123]. We use CMB maps obtained through the SMICA component separation method [76, 124]. On the left-hand side are scatter plots of f_{NL} values obtained using each of the two basis sets. In an ideal case where the two estimates are identical for every test map, all the points would lie on a straight line given by $y = x$. Drawn in dashed red line is the best linear fit to the data. Its slope, intercept, and the R-squared value are annotated below. On the right is a more detailed plot of the computed f_{NL} for each map.

We see that results from the two different sets of basis are in good agreement. The R-squared value of the linear fit is greater than 0.99 for Local, and 0.999 for Equilateral and Orthogonal shapes. The intercept and sample mean are also near zero. We do not find any systematic discrepancies across the shapes from individual map estimates either.

The fact that the results from the KSW and Legendre basis are consistent validates multiple aspects of our pipeline. First of all, we can deduce that the Legendre basis expansion (5.28) accurately represents the bispectrum shapes of interest, especially since the KSW basis is designed to be exact for the three standard shapes. Having $p_{max} = 30$ modes is more than sufficient to achieve sub-percent accuracy. Figure 5.5 explicitly compares results obtained from Legendre bases with $p_{max} = 29$ and 30 while keeping everything else the same. The two results show close agreement; the largest of the errors is less than $O(10^{-5})$. This further confirms that our expansion using Legendre polynomials has completely converged.

Secondly, we validated the accuracy of the linear term and lensing-ISW bias calculations through the sample mean of f_{NL} estimates from lensed simulations. When we exclude either of the two, we find the sample mean to be far from zero. The error is especially large for the
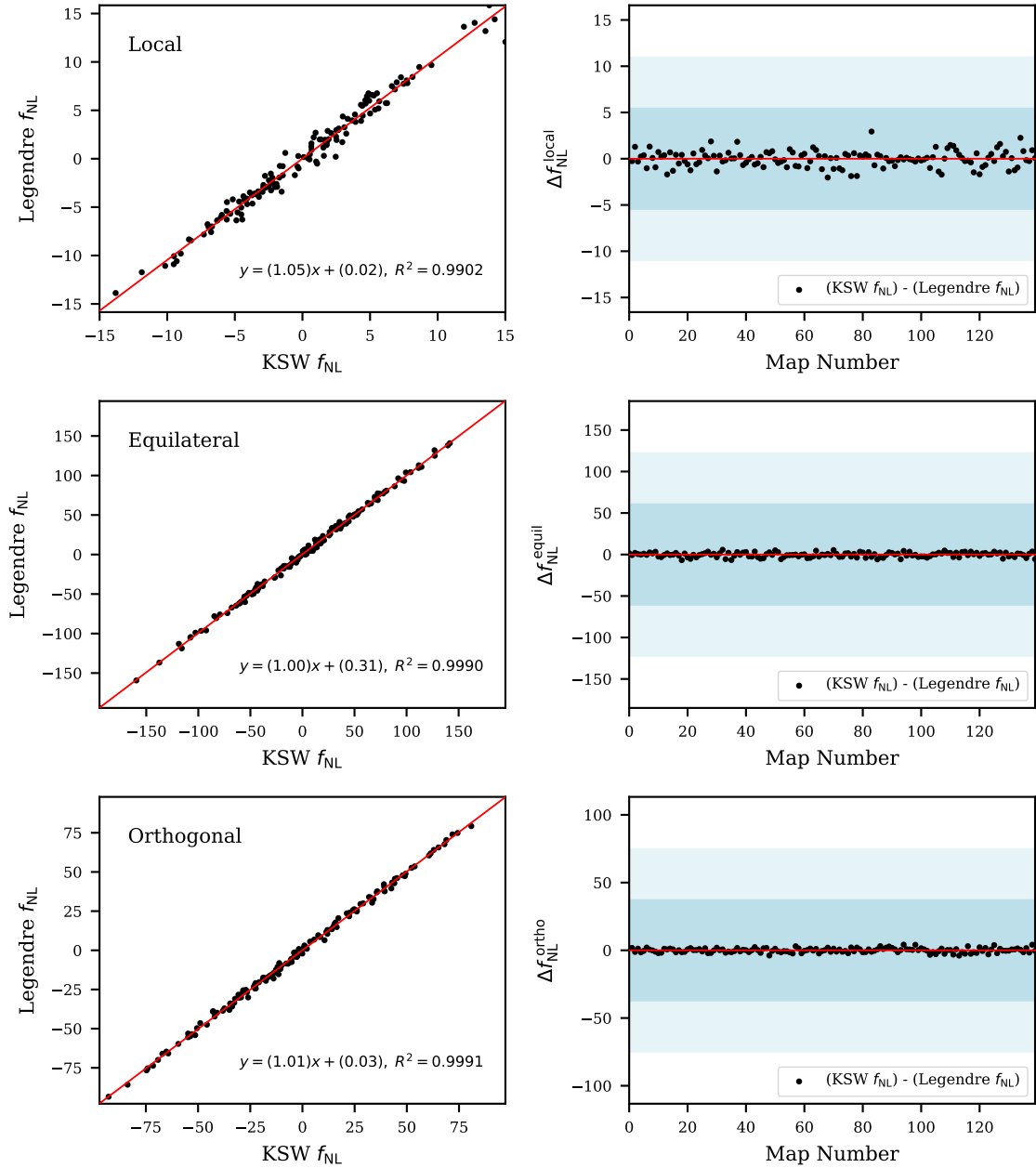
Fig. 5.4 A map-by-map comparison of the f_{NL} estimates evaluated using the KSW and Legendre basis sets for three standard templates. The Planck 2018 CMB map and 140 FFP10 simulations have been used, each representing a single point on the scatter plot (left). Details of the linear best-fit to data (red) are annotated below. Shown on the right-hand side are plots of the differences in the f_{NL} values for each map, together with the 1σ and 2σ intervals shaded in blue. In the ideal case where the two basis sets yield identical results, we should see all the points lie on the line $y = x$ for the left plot and $y = 0$ for the right plot. For more information on each of the three theoretical templates used, see e.g., [70].
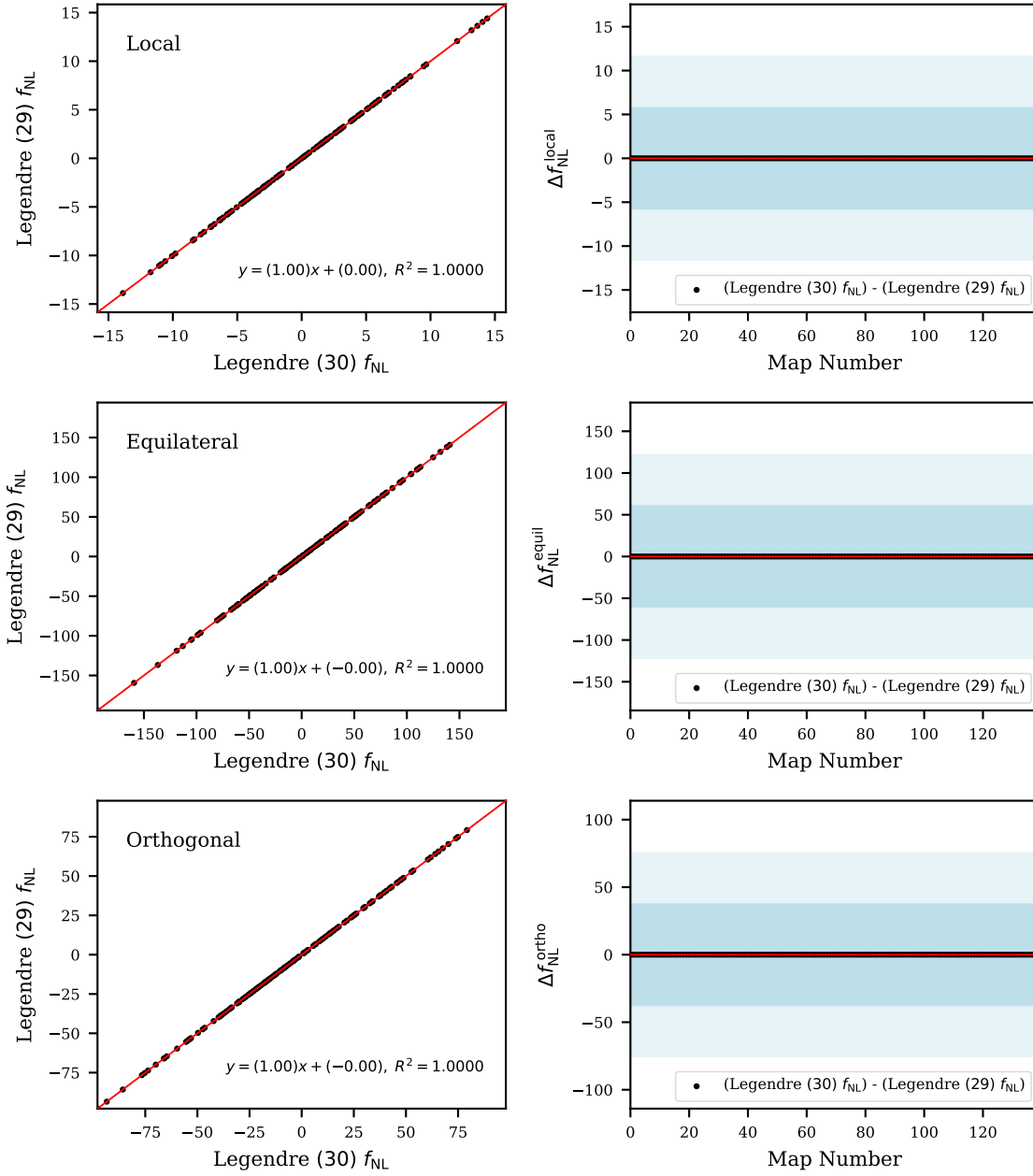
Fig. 5.5 A map-by-map comparison of the f_{NL} estimates for three standard templates evaluated using the Legendre basis with different numbers of modes: $p_{\max} = 29$ and 30. The two results agree with errors less than $O(10^{-5})$, as can be seen from the scatter plots (left) and the map-by-map residual plots (right). Shaded in blue are the 1σ and 2σ levels of the f_{NL} estimator. This confirms that our expansion using Legendre polynomials has completely converged for these bispectrum shapes.

local template. The linear term accounts for the anisotropy introduced by masking parts of the observed sky. The squeezed limit contribution to the bispectrum comes from couplings between a pair of small-scale modes and a large-scale one, making it more susceptible to bias from partial coverage. The presence of sky masks therefore offsets the f_{NL} estimates from the local shape which has a large squeezed limit. Lensing-ISW bias is also the largest in squeezed configurations and affects Local f_{NL}. The fact that our lensed Gaussian simulations have f_{NL}s fluctuating around 0 validates our bias subtractions.

Lastly, we check that CMB-BEST accurately preserves the optimality of the CMB bispectrum estimator. In the weak non-Gaussian limit, the estimator (5.1) saturates the Cramer-Rao bound. Its expected variance is the lowest amongst all possible unbiased bispectrum estimators of f_{NL}. Heuristically speaking, the estimator extracts as much information about non-Gaussianity as possible from the CMB bispectrum. As we discussed in Chapter 4, this bound only depends on the Fisher information determined from normalisation; $Var[\hat{f}_{\mathrm{NL}}] = F^{-1} = 6/N$. We refer to this value as the *theoretical* variance (the best possible from theory). Meanwhile, individual f_{NL} estimates from simulated maps and independent and normally distributed. Sample variance obtained from N_{sims} simulations should therefore approach the theoretical variance as $N_{\mathrm{sims}} \to \infty$. Table 5.3 summarises calculated values of the two types of variances discussed.

Table 5.3 Comparison of the sample and theoretical variances obtained from f_{NL} estimates of standard shapes, computed using each of the KSW and Legendre basis sets. Sample variances are within 1σ of theoretical values assuming that the 140 individual f_{NL}'s from simulated maps are normally distributed. This is statistically consistent with the optimality of our bispectrum estimator.

Template	Basis	Sample Variance	Theoretical Variance	(Sample)/(Theory)
Local	KSW	5.5	5.3	1.04
	Legendre	5.9	5.7	1.04
Equilateral	KSW	61.2	67.7	0.90
	Legendre	61.7	67.7	0.91
Orthogonal	KSW	37.7	33.7	1.12
	Legendre	38.0	33.9	1.12

Note that the numbers do not match up exactly between sample and theoretical variances, and the results for equilateral template appear to be more optimal than the theory allows. This is because the sample variance $\hat{S}^2 = (\sum_i (f_{\mathrm{NL}}^{(i)})^2)/N_{\mathrm{sims}}$ follows a chi-squared distribution with

$N_{\text{sims}} - 1$ degrees of freedom under our assumptions. The standard deviation corresponding to the (normalised) distribution equals $\sqrt{2/(N_{\text{sims}} - 1)}$, which is ≈ 0.12 for $N_{\text{sims}} = 140$. It is therefore not surprising to see our sample variances differ by up to 12% from the theoretical ones.

We performed further checks to ensure that this discrepancy is due to statistical fluctuations rather than systematic errors. The ratio between the sample and theoretical variances remained nearly constant across the KSW and Legendre basis from CMB-BEST, as well as Planck's Modal pipeline on the same set of simulated maps. Meanwhile, sample variances evaluated from independent sets of simulations do fluctuate around the theoretical value. A closer check has been done for each of the mode sets (p_1, p_2, p_3) in the Legendre basis. The decomposition coefficient α is set to be 1 at (p_1, p_2, p_3) and its permutations, and then vanishing everywhere else. Substituting into (5.7) and (5.10), we get

$$f_{\text{NL}}^{(i)} = \frac{1}{N} \left[\left(\beta_{p_1 p_2 p_3}^{\text{cub},(i)} - 3\beta_{p_1 p_2 p_3}^{\text{lin},(i)} \right) + 5 \text{ cyc.} \right], \tag{5.31}$$

$$N = 6 \left(\Gamma_{p_1 p_2 p_3, p_1 p_2 p_3} + \Gamma_{p_1 p_2 p_3, p_1 p_3 p_2} + \cdots + \Gamma_{p_1 p_2 p_3, p_3 p_2 p_1} \right). \tag{5.32}$$

Here we assumed that p_1, p_2, p_3 are distinct for convenience. Corresponding sample and theoretical variances have been compared for each of the modes. Overall, they are found to be statistically consistent as before.

While Figure 5.4 shows excellent agreement between results from the KSW and Legendre basis sets overall, there are small but noticeable scatters in the plot for the local shape. The slope of linear best fit is also slightly above 1, meaning that f_{NL} estimates from the Legendre routine tend to vary 5% more than the KSW ones. A larger variance means less information extracted. Our first hypothesis was that the Legendre polynomials are losing a small fraction of information from the CMB due to their fixed k range in their definition (5.18). We test this by increasing the range by changing $k_{\text{max}}/k_{\text{min}} = 1000$ to 2000. The results are shown in Figure 5.6.

Our main focus of this figure is on the local template results. Estimates from the KSW and Legendre basis sets match almost perfectly now. When altering the ratio $k_{\text{max}}/k_{\text{min}}$, we fixed the k_{max} and lowered k_{min}. Including more small-k, or large-scale modes provides extra information in the bispectrum. The local shape is especially affected by this since squeezed configurations including one of these extra large scale modes have a more significant contribution to the total estimate. Also, note that $k_{\text{max}}/k_{\text{min}} = 2000$ is comfortably larger than the equivalent ratio in harmonic space $l_{\text{max}}/l_{\text{min}} = 2500/2$.

Despite the improvements in the squeezed limit and local template, we may not simply set $k_{\text{max}}/k_{\text{min}} = 2000$ as default because it hurts convergence in other shapes of interest. As
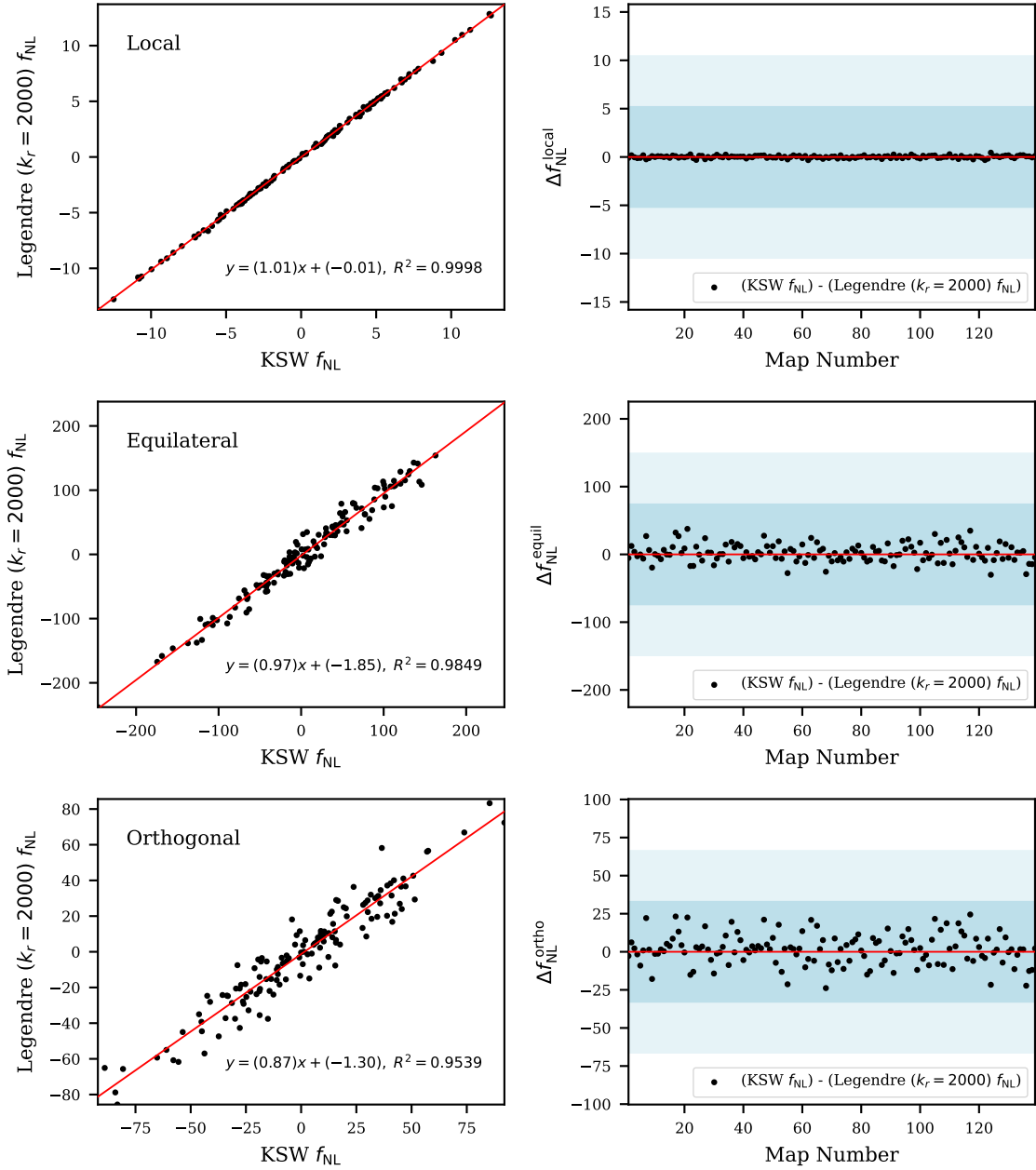
Fig. 5.6 A map-by-map comparison of the f_{NL} estimates for three standard templates from the KSW and Legendre basis sets, similar to Figure 5.4. Here, the Legendre basis has a wider k domain: $k_{max}/k_{min} = 2000$ instead of the usual 1000. The number of modes (p_{max}) has been reduced to 10 instead of 30 due to limited computational resources. We see that the additional information gained from large scales (low k) fixes the small scatters present in f_{NL}^{local} of the previous plot, however at the cost of increased spread in f_{NL}^{equil} and f_{NL}^{ortho}.

can be seen from the Equilateral and Orthogonal plots in Figure 5.6, newer estimates of f_{NL} are less accurate for shapes other than Local. The fact that $p_{max} = 10$ here rather than 30 is one of the main causes of the drop in precision, but having smaller k values within Legendre polynomials' domain also has a negative impact. Shapes with a $1/k$ scaling in their expressions vary more dramatically when k is small and tend to be harder to expand in terms of Legendre polynomials.

For the final check of internal consistency, we inspect how each point in the line of sight integral (r) contributes to f_{NL} estimates. CMB-BEST's formalism makes it straightforward to plot the r integrand since the integral is done at the very end. Figure 5.7 shows plots of f_{NL} contributions from different sources and for the three standard bispectrum templates. Shaded in light blue are the 1σ regions obtained from 140 simulations for each point in r.

As illustrated in plots on the right-hand side of Figure 5.7, the vast majority of signal comes from recombination around $r = 14,000$Mpc. In fact, its contribution is dominant enough that neglecting signals from everywhere else would still be a good approximation. CMB-BEST uses an adaptive r grid which is denser around recombination, following the works of [65]. Other small but notable contributions to the total estimate come from reionisation ($r \sim 10,000$Mpc) and the ISW effect ($r < 5,000$Mpc).

Zooming in on an interval around recombination, significant contributions from the cubic term to f_{NL} of the Local template is rather prominent (upper left of Figure 5.7). The signal is sufficiently larger than the expected random fluctuations evaluated from Gaussian simulations, which could be hinting at non-zero primordial non-Gaussianity. However, contributions from the linear term (shown orange) completely counterbalance it, bringing the total down to values consistent with zero. This again validates the accuracy of our methodology; bias to f_{NL} generated from anisotropic sky masks are precisely subtracted off using the linear terms.

We do not find any statistically significant signal across the whole r range otherwise, especially when taking the look-elsewhere effect into account. It is not meaningful to find a couple of 3-4σ values when the other 500 points are simply disregarded. If we detect primordial non-Gaussianity in the future, however, then these plots of f_{NL} contributions for each r will provide valuable insights into where the signal comes from.

5.3.2   Consistency with Planck

In the previous section, we demonstrated how the integrity of CMB-BEST was validated. The next set of validations involves comparing it against other existing codes for CMB bispectrum estimation.

We test primordial non-Gaussianity constraints on Local, Equilateral and Orthogonal templates against the Planck 2018 analysis [23]. Two sets of basis functions, KSW and
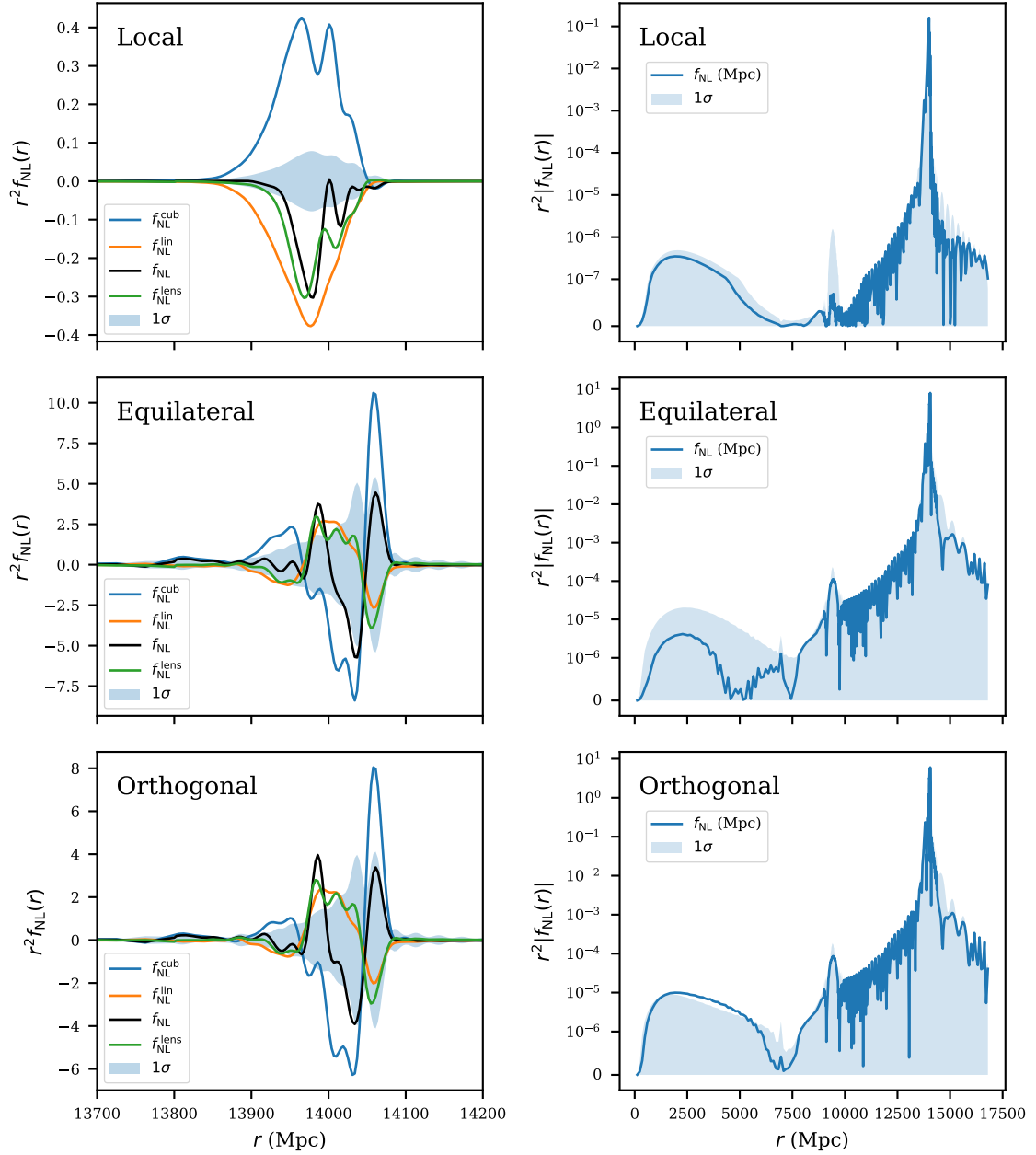
Fig. 5.7 Contributions to the total f_{NL} from each point in the line-of-sight integral over r for standard templates. On the left-hand side, we plot contributions from the cubic, linear, and lensing-ISW bias, as well as the total f_{NL}. Shown in blue is the 1σ interval obtained from the corresponding terms in 140 FFP10 Gaussian simulations. We focus on the r interval around recombination where most of the signal comes from. Plots on the right-hand side show the f_{NL} contributions over the whole r range in log scale. The ISW effect, reionisation, and recombination are responsible for the three most noticeable peaks in all three plots.

Legendre, have been used to compute f_{NL} from the foreground-cleaned CMB map included in the final data release. We choose SMICA as the main component separation method since it was shown to be the most reliable and robust for Planck bispectrum analysis [23, 30, 70, 124]. Table 5.4 summarises the constraints obtained, together with the quoted results from the Planck team's own KSW estimator and Modal 2 pipeline.

Table 5.4 Constraints on f_{NL} for the standard shapes from the KSW and Legendre basis of CMB-BEST, in comparison with the Planck 2018 analysis [23]. Only the temperature data from the SMICA foreground-cleaned map and FFP10 simulations were used for the analysis. Values shown are after the lensing bias subtraction, with uncertainties at 68% CL.

Shape	CMB-BEST		Planck 2018	
	KSW	Legendre	KSW	Modal
Local	-2.2 ± 5.5	-2.0 ± 5.9	-0.5 ± 5.6	-0.6 ± 6.4
Equilateral	17 ± 61	15 ± 62	7 ± 66	34 ± 67
Orthogonal	-7 ± 38	-9 ± 38	-15 ± 36	-26 ± 43

Note that while the constraints from CMB-BEST are largely consistent with Planck 2018 results, there are discrepancies of up to 0.3σ. However, there are variations around this level within the estimates from different pipelines of Planck as well. Equilateral constraints from Planck's own KSW and Modal estimators shown here, for example, differ by $\approx 0.4\sigma$. A similar amount of fluctuation can be found in the full result shown in [23]. In an ideal world, they should match exactly across different approaches as long as the same dataset is used. However, the statistical significance of the individual f_{NL}'s is not largely affected by such variations.

We invested a significant proportion of our time investigating this error. Here we discuss three potential areas which might account for the discrepancies. First of all, human error during the implementation and estimation process should not be neglected. Bispectrum analysis is complex and computationally expensive. Implementing it often involves writing a long and heavily optimised code. During the development and testing stages, we found and fixed many mistakes in our $10,000+$ lines of C code. Various unit tests were performed in the process to test individual sections of the code: basis expansion, projection to the l space, SHTs, parallelisation, and more. Internal consistency checks were then used to verify the integrity of the combined pipeline. The Planck team also went to great lengths to validate and cross-check different methods [70]. It therefore seems unlikely that trivial mistakes are causing the gap in results.

The next source of error we studied is the parameter set. The cosmological parameters we used for constraints in Table 5.4 are identical to those of the Planck Modal pipeline. Small changes in cosmological parameters are also shown to have little effect on f_{NL} [23, 30]. One major difference, however, is the number of Gaussian simulations used. We included 140 simulated maps while Modal has about 300. This number is mainly restricted by computational resources required for the large Legendre basis in CMB-BEST. While we found 140 to be sufficient for most cases, sample variance may cause some fluctuation in the linear term and error bar. A rough estimate for sampling error is $\sim \sqrt{2/N_{sims}} \approx 0.12$, as calculated in the previous section. Other than N_{sims}, more internal parameters such as the grid density of discretised arrays have also been checked to yield consistent results when varied.

We are left with systematic errors as potential causes for discrepancies. The most significant bias to f_{NL} constraints is the lensing-ISW bias discussed in Section 3.3.4. Table 5.5 shows biases from the lensing bispectrum [20] for both CMB-BEST and Planck 2018 analysis. The numbers vary but are consistent overall. Map-by-map comparisons of constraints from the Legendre basis and Modal are shown in Figure 5.8.

Table 5.5 Bias to f_{NL} of standard shapes originating from the lensing bispectrum. We compare CMB-BEST's two different basis sets and the Planck 2018 analysis [23], using SMICA map and FFP10 simulations.

Shape	CMB-BEST		Planck 2018	
	KSW	Legendre	KSW	Modal
Local	7.5	8.2	7.3	6.9
Equilateral	−0.7	−0.6	−0.7	4.0
Orthogonal	−22	−22	−23	−25

Out of the three templates, the local shape shows the largest scatter between constraints. Comparing f_{NL}'s of simulated maps from CMB-BEST and Modal, we find a correlation of 0.916. The intercept value of 0.78 in the linear fit is mainly due to the sample mean present in Modal. Legendre's sample mean is 0.069 for Local, while Modal's is −0.698. Otherwise, the difference between the two pipelines is distributed such that its sample standard deviation equals 2.5, skewness is −0.14, and kurtosis is 0.19. Having a low level of small skewness and kurtosis of less than 1 implies that the distribution is close to a univariate normal, consistent with random fluctuation. Similarly, we do not find any significant systematic error from the equilateral and orthogonal shapes either.
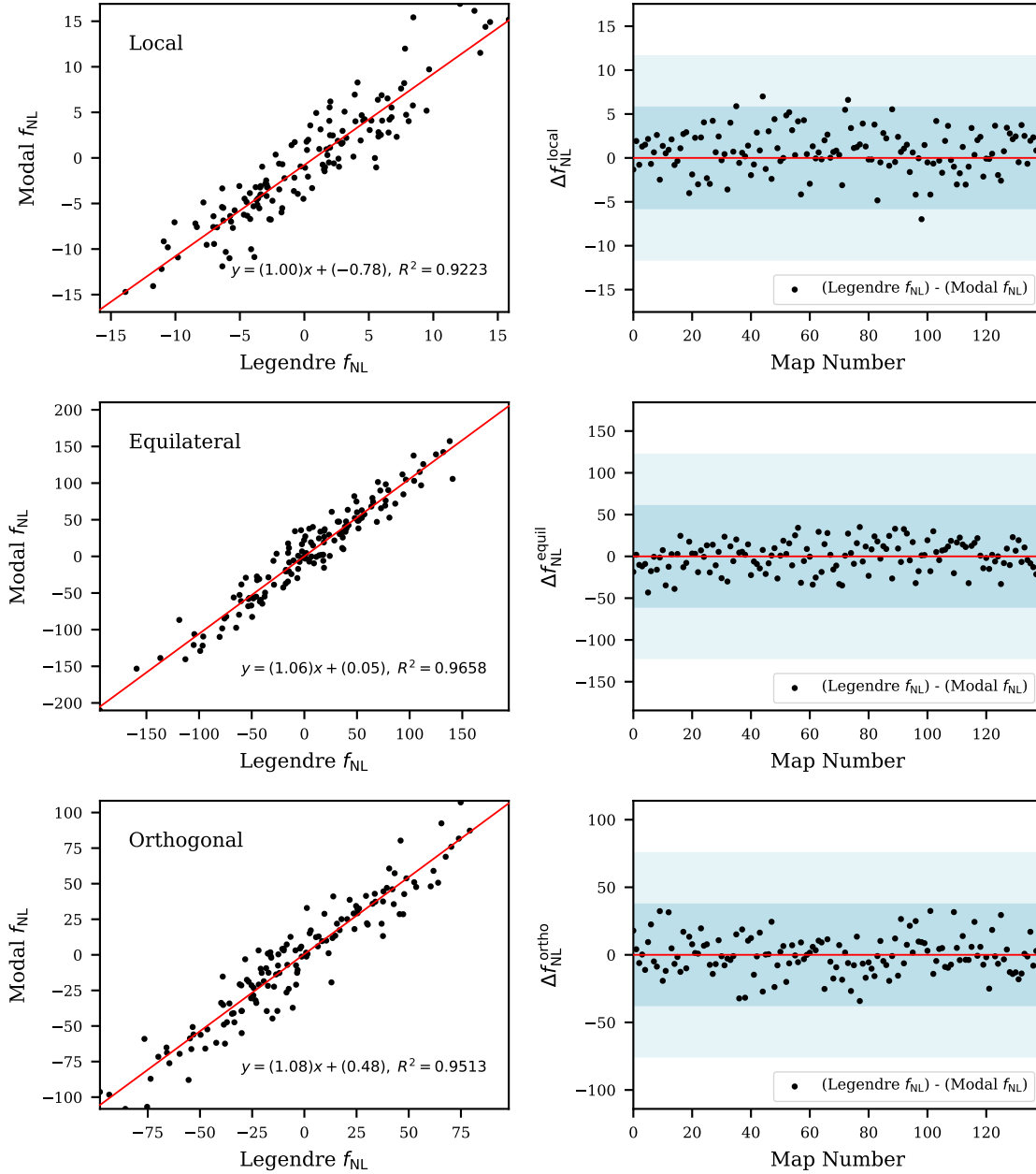
Fig. 5.8 A map-by-map comparison of the f_{NL} estimates obtained from the CMB-BEST's Legendre basis set against the Modal estimator results of the Planck 2018 analysis [23]. The first 140 FFP10 simulations are used here. On the left-hand side are scatter plots where each simulation is represented by a point according to f_{NL} estimates of standard templates. Their linear best fit lines are shown in red. Differences in the estimates from the two routines are shown map-by-map on the right-hand side, together with the 1σ and 2σ levels shaded in blue. Overall, CMB-BEST and Modal are in good agreement without any significant systematic errors.

Having not found a clear source of error, we conclude that the small gaps between the f_{NL} estimates in Table 5.4 mainly come from differences in methodology and are consistent with random fluctuations.

We now move on to constraining models with oscillations. The simplest template for oscillatory models is the feature model studied in Chapter 4. We use a template shape function of the form $S^{feat}(k_1, k_2, k_3) = \sin(\omega(k_1 + k_2 + k_3) + \phi)$. Figure 5.9 shows f_{NL} constraints obtained from the Legendre basis and compares them with Modal results from the Planck 2018 analysis. The 'phase' ϕ is set to zero, while the oscillation 'frequency' ω was varied from 10 Mpc to 350 Mpc. We follow [95] and increase ω in steps of 10 so that correlations between shapes with different ω are kept low.
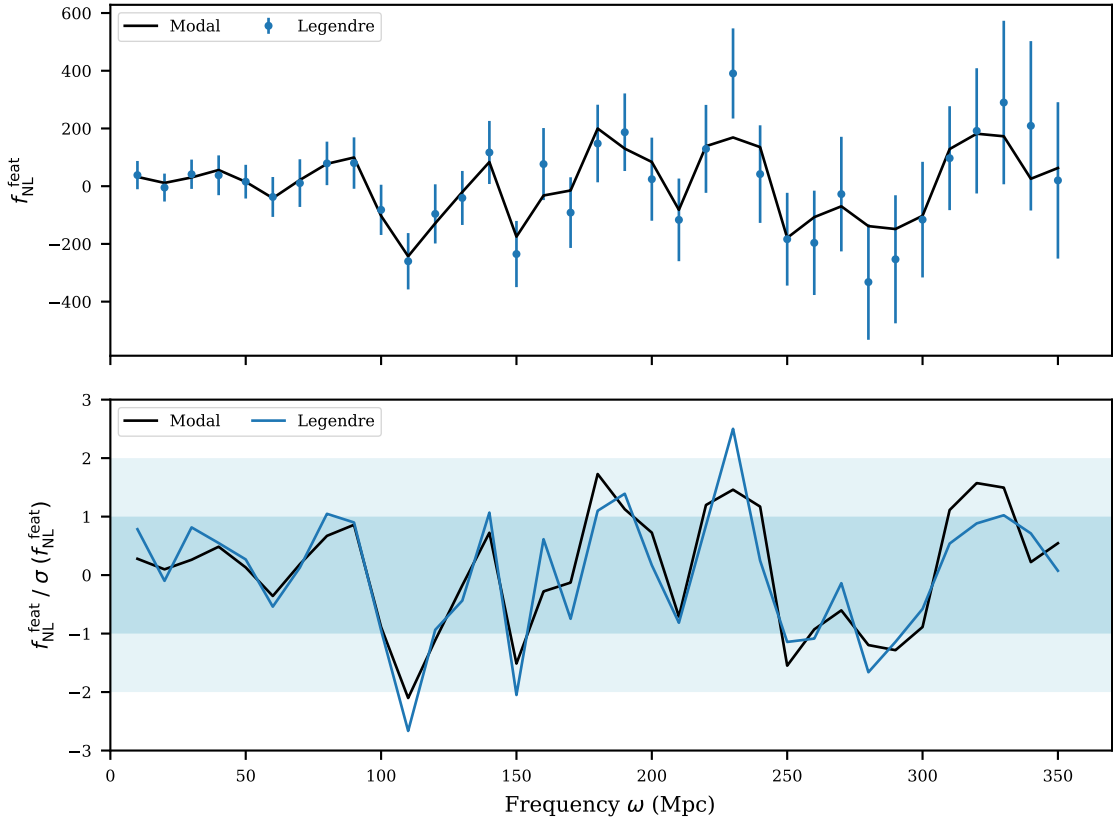
Fig. 5.9 The f_{NL} estimates for the feature models with $S(k_1, k_2, k_3) = \sin(\omega(k_1 + k_2 + k_3))$, obtained using the Legendre basis (blue) and Planck's Modal (black). Top: a direct comparison of f_{NL} for different values of ω. Error bars indicate the expected standard deviations in the estimator calculated using 140 Gaussian simulations. Bottom: signal-to-noise $f_{NL}/\sigma(f_{NL})$, again for a range of ω values. Shaded in blue are the 1σ and 2σ levels. We see that the two approaches yield coherent estimates overall.

The Legendre basis accurately expands the feature model template using Legendre polynomials via basis expansion outlined in (5.28). Estimates from the two methods, CMB-BEST's Legendre and Planck's Modal, are mostly compatible. The most notable difference is at $\omega = 230$Mpc where f_{NL} from Legendre is more than 1σ larger compared to Modal. Even though it is interesting that the new estimate now passes the 2σ threshold, having one such point out of 35 shown here has less statistical significance.

Both the Legendre and Modal approaches involve expanding the shape function with respect to a polynomial basis. Polynomials are versatile but have limited resolution for oscillatory signals; it cannot resolve shapes with a number of oscillations greater than the maximum degree of polynomials (p_{max} here). For the Legendre basis with k range $[2.09 \times 10^{-4}, 2.09 \times 10^{-1}]$Mpc^{-1} and $p_{\mathrm{max}} = 30$, frequencies greater than $\pi p_{\mathrm{max}}/(k_{\mathrm{max}} - k_{\mathrm{min}}) \approx 436$Mpc are unresolvable. In reality, numerics start to break before this value. We do not have a reference analytic value for true f_{NL}s, but evaluating correlations between shapes is effective for checking our numerics.

Figure 5.10 shows correlations between f_{NL}s from feature models with different frequency ωs. The plots show the correlation between f_{NL} estimates from 140 FFP10 Gaussian simulations ('sample'), together with the one evaluated using our 'late-time' inner product ('theoretical') given by

$$\left\langle b^{(i)}, b^{(j)} \right\rangle = \sum_{l_j} \frac{h_{l_1 l_2 l_3}^2 b_{l_1 l_2 l_3}^{(i)} b_{l_1 l_2 l_3}^{(j)}}{C_{l_1} C_{l_2} C_{l_3}} \tag{5.33}$$

$$= \sum_{p_j p_j'} \alpha_{p_1 p_2 p_3}^{(i)} \alpha_{p_1' p_2' p_3'}^{(j)} \sum_{l_j} \frac{h_{l_1 l_2 l_3}^2 (b_{p_1 p_2 p_3}^{(i)})_{l_1 l_2 l_3} (b_{p_1' p_2' p_3'}^{(j)})_{l_1 l_2 l_3}}{C_{l_1} C_{l_2} C_{l_3}} \tag{5.34}$$

$$= \sum_{p_j p_j'} \alpha_{p_1 p_2 p_3}^{(i)} \alpha_{p_1' p_2' p_3'}^{(j)} \Gamma_{p_1 p_2 p_3, p_1' p_2' p_3'} \tag{5.35}$$

$$= \alpha^{(\mathbf{i})} \, \Gamma \, \alpha^{(\mathbf{j})}, \tag{5.36}$$

where Γ is defined in (5.11). In the limit $N_{\mathrm{sims}} \to \infty$, the sample correlation can be shown to approach the theoretical value in the weakly non-Gaussian limit. We see from Figure 5.10 that they indeed display the same qualitative behaviour for $N_{\mathrm{sims}} = 140$.

The templates with $\omega < 300$ are highly uncorrelated with each other, as can be seen from small non-diagonal elements. Noticeable correlations on lines $\omega_2 = \omega_1 \pm c$ for $c = 70, 140$ arise from resonance between oscillations and transfer functions at the Baryonic Acoustic Oscillation (BAO) scale, as we observed in Figure 4.7. Faint lines can also be found near $\omega_2 = -\omega_1 + c$, $c = 70, 140, 210, \cdots$ for similar reasons.
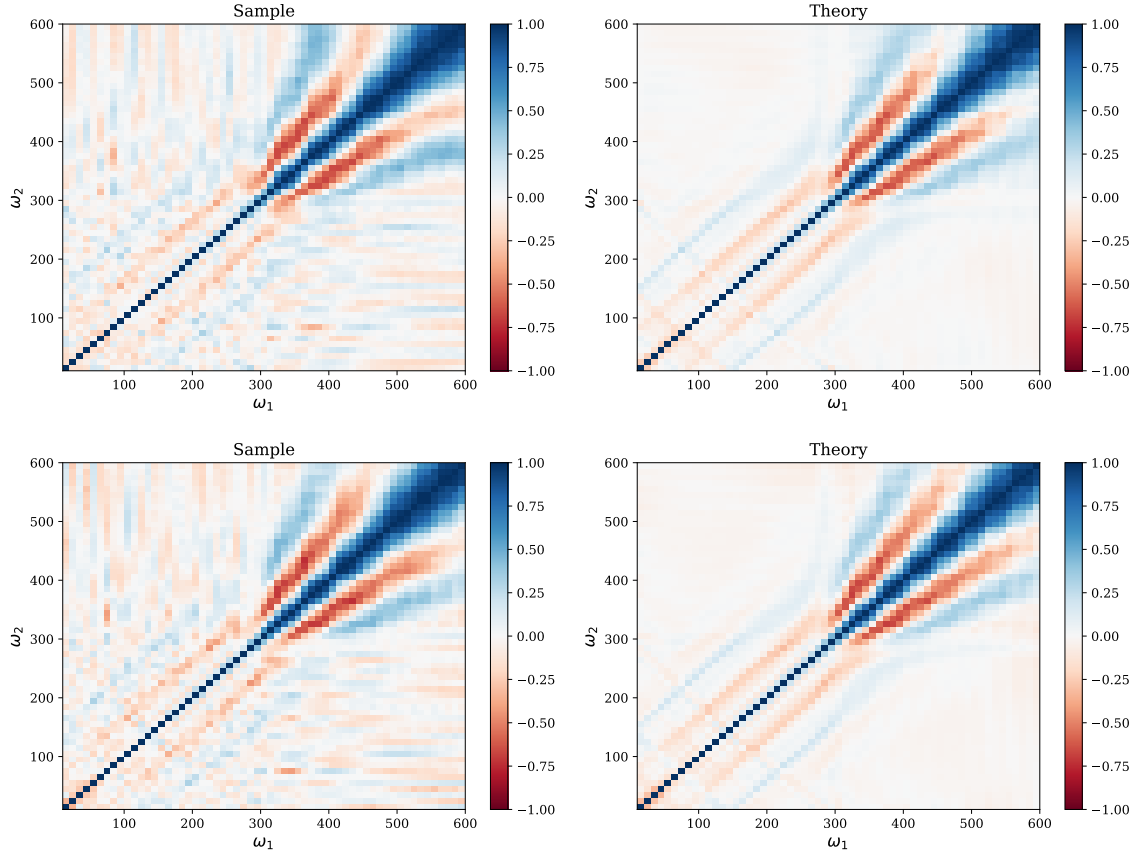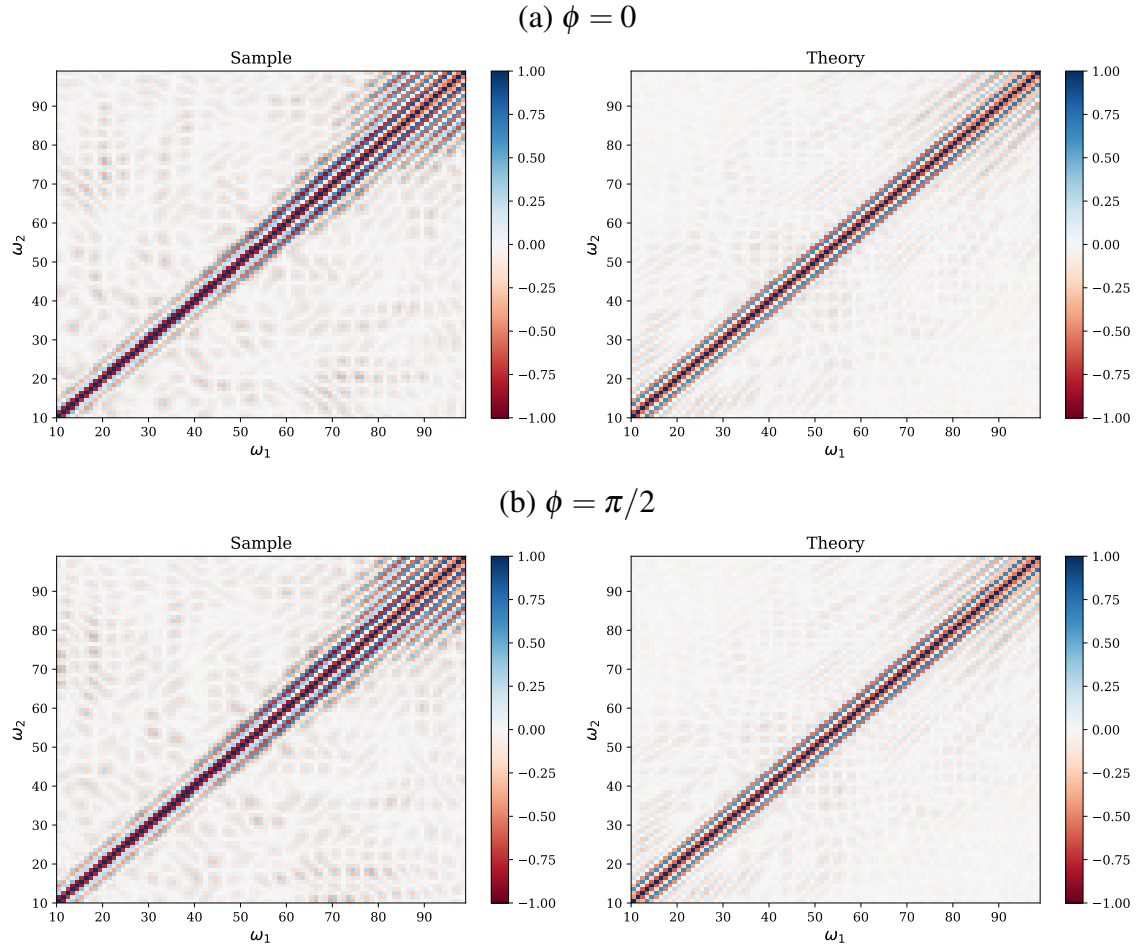
Fig. 5.10 Correlations between feature model templates $S(k_1, k_2, k_3) = \sin(\omega(k_1 + k_2 + k_3) + \phi)$ with different ω values, for $\phi = 0$ (top two) and $\phi = \pi/2$ (bottom two). Results are from the Legendre basis. 'Sample' correlations are obtained from f_{NL} estimates from 140 Gaussian simulations, while 'theory' correlations come from the inner product induced by $\Gamma_{p_1 p_2 p_3, p_1 p_2 p_3}$ matrix in (5.11). Large non-diagonal correlations appear around $\omega \approx 300$, after which oscillations in the shape are no longer resolved by polynomials of degree up to $p_{\max}$.

When $\omega > 300$, however, large non-diagonal correlations appear. This is about when the oscillation frequency surpasses the resolution set by the highest degree of polynomials. Our basis expansion becomes inaccurate after this point. There are interesting linear structures present, with slopes roughly equal to $2, 1/2$ and subsequently $4, 1/4$. These lines come from aliasing caused by subsampling rapid oscillations; our basis picks up $\omega' = \omega/2$ signal instead. $\omega_2/2 = -\omega_1 + 70$ and other pairs of (ω_1, ω_2) resonate to give large non-diagonal correlations.

We verify our claim that numerical inaccuracies at high ωs are caused by the lack of resolution due to a limited number of polynomial modes. Figure 5.11 illustrates how having a smaller k range allows us to constrain faster oscillations more accurately. Note that we also change l_{max} correspondingly since smaller scales are neglected. The plot on the right-hand side uses a Legendre basis with the (k_{min}, k_{max}) range rescaled by a factor of $3/4$. Fewer oscillations appear within the k interval and are therefore easier to expand using fewer polynomials. As we expected, the scale at which non-diagonal correlations blow up is now shifted to a higher frequency, $\omega \approx (4/3) \cdot 300$.
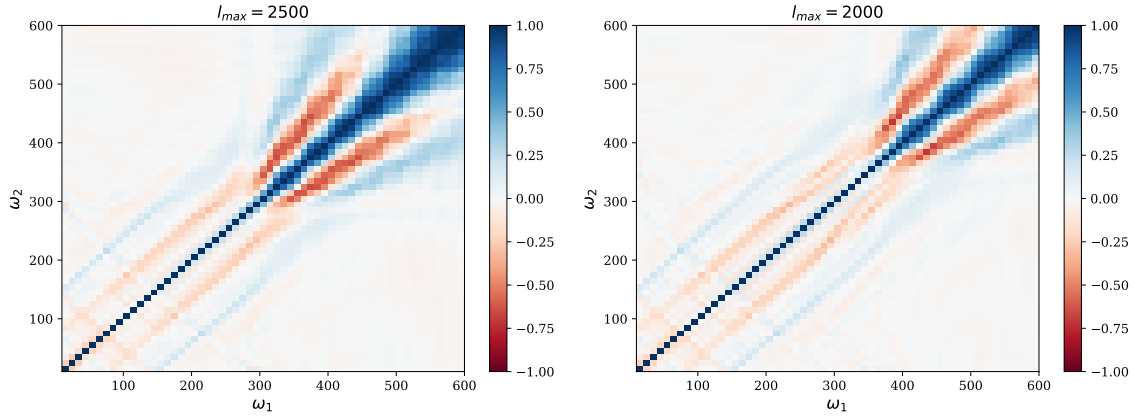
Fig. 5.11 Theoretical correlations between feature model templates with different frequency ωs, as described in Figure 5.10. The phase ϕ is set to zero in both cases, while the right plot is obtained from the Legendre basis with a different k range. The maximum l value is reduced from 2500 to 2000, shifting (k_{min}, k_{max}) by a factor of $2000/2500 = 0.75$. Having a smaller k range means fewer oscillations within the k interval for same ω which provides better effective resolution. The right plot does indeed show smaller non-diagonal correlations at high frequencies.

There is a subtlety around the f_{NL} estimates obtained from inaccurate primordial basis expansions. CMB-BEST computes βs and Γs with respect to a fixed Legendre basis, which has been shown to be accurate for each and every polynomial mode. Therefore, even when the primordial basis is unable to resolve rapid oscillations, the f_{NL} estimates we obtain are

nevertheless meaningful; they are simply probing a different model. The constraints are not for the given shape function, but rather its projection to the vector subspace spanned by the basis functions as shown in (5.24). Detecting a non-zero f_{NL} here can still have significant implications.

Another popular template for models with oscillations is the 'resonance' shape parametrised as $S(k_1, k_2, k_3) = \sin(\omega \log(k_1 + k_2 + k_3) + \phi)$. Log-spaced oscillations are numerically harder to deal with since the oscillation frequency diverges as $k \to 0$. Note also that any scaling factors to the k's can be absorbed into the phase via $\log(c(k_1 + k_2 + k_3)) = \log(k_1 + k_2 + k_3) + \log(c)$. For our Legendre basis with k_{max}/k_{min} fixed to 1000, the full k range includes $\approx 1.1\omega$ oscillations. Any frequency larger than ≈ 27 therefore cannot be expanded using $p_{max} = 30$ polynomials. We still explore shapes with higher ωs, however, since the basis can pick up slower oscillations in higher k values. Corresponding constraints should be taken with a grain of salt; they probe bispectrum shapes *similar* to the resonance template. Low-k oscillations are especially likely to be wiped out from these shapes.

Figure 5.12 compares f_{NL} constraints for the resonance template over a range of ωs. As can be seen from the top two plots, the signal-to-noise values obtained from the Legendre basis and Modal are relatively consistent up until $\omega \approx 35$, after which the two results start diverging significantly. The threshold is about the same for 'sinlog' ($\phi = 0$) and 'coslog' ($\phi = \pi/2$) shapes.

Recall that the sample σ from f_{NL} estimates of Gaussian simulations should converge to the theoretical value calculated from the Fisher information as $N_{sims} \to \infty$, since the CMB bispectrum estimator is optimal. We have checked that the sample and theoretical variances are consistent for standard templates in Section 5.3.1, using both KSW and Legendre basis sets. The bottom two plots in Figure 5.9 show the equivalent results for the resonance template with varying ω.

Similarly to the feature models studied in Chapter 4, uncertainty in the estimated f_{NL}s increases as we raise ω, exploring more rapid oscillations in the bispectrum. Both the Modal and Legendre methods lose their ability to resolve shapes with ωs larger than 35. The constraints are then no longer for the precise resonance shape but rather an approximation to it. A notable difference between Modal and CMB-BEST's Legendre basis is their behaviour at high ω. Modal's numerical accuracy is completely lost after $\omega \approx 50$, leading to disproportionately large bispectra and small σ. On the other hand, Legendre retains stability in its basis expansion, yielding constraints for approximate bispectrum shapes closest to the true, high-frequency ones.

A detailed comparison between the sample and theoretical variances is depicted in Figure 5.13. This serves as a useful consistency check within the CMB-BEST pipeline. Sample
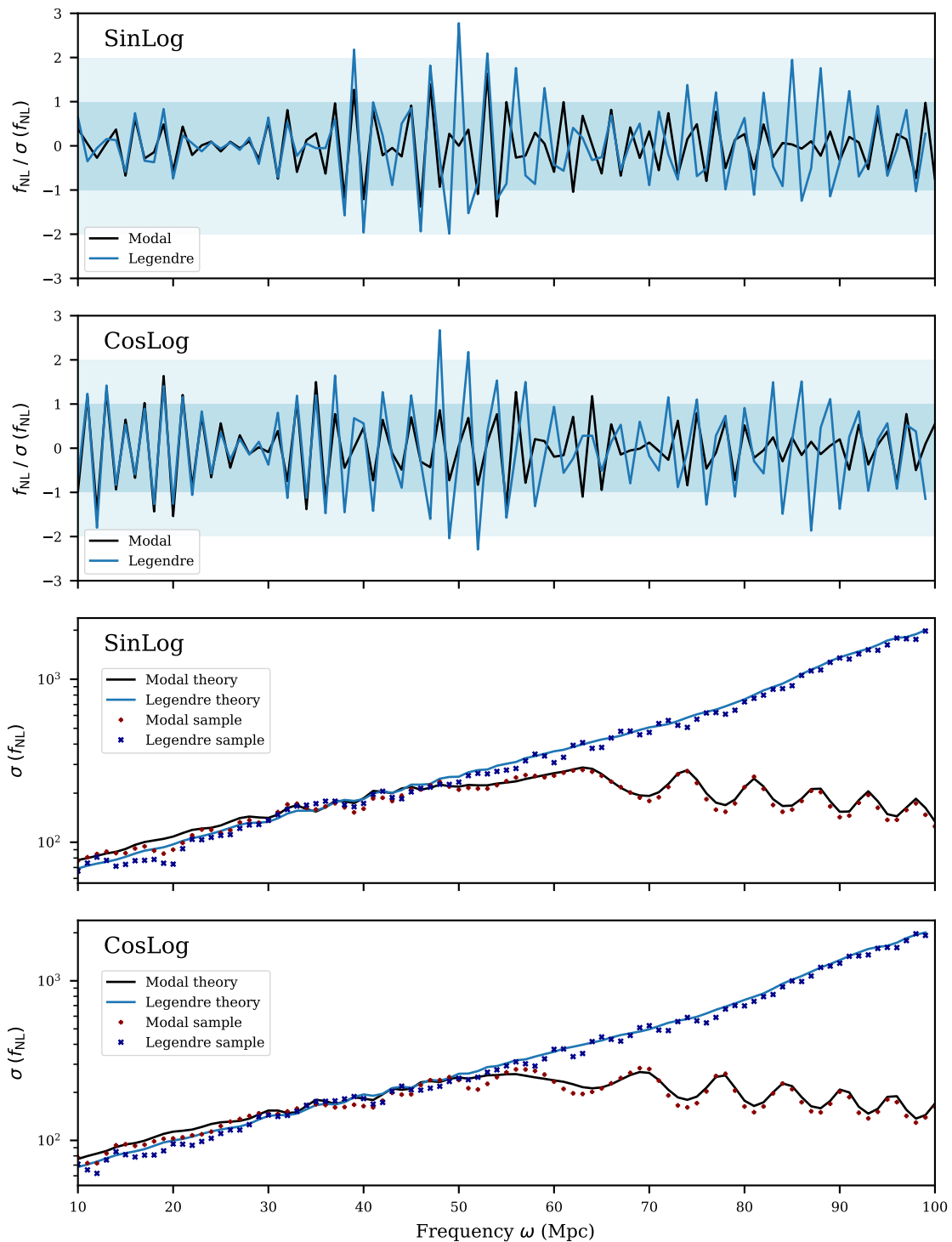
Fig. 5.12 Constraints for the resonance shape $S(k_1, k_2, k_3) = \sin(\omega \log(k_1 + k_2 + k_3) + \phi)$ with varying frequency ωs, while the phase ϕ is set to 0 ('SinLog') and $\pi/2$ ('CosLog'). Results are obtained using CMB-BEST's Legendre basis (blue) and Planck's Modal estimator (black). Top two: signal-to-noise significance of the estimated f_{NL}s with their 1σ and 2σ levels shaded in blue. Bottom two: standard deviations of the f_{NL} estimates against ω calculated from f_{NL}s of 140 Gaussian simulations (sample) and the Γ matrix (theory).
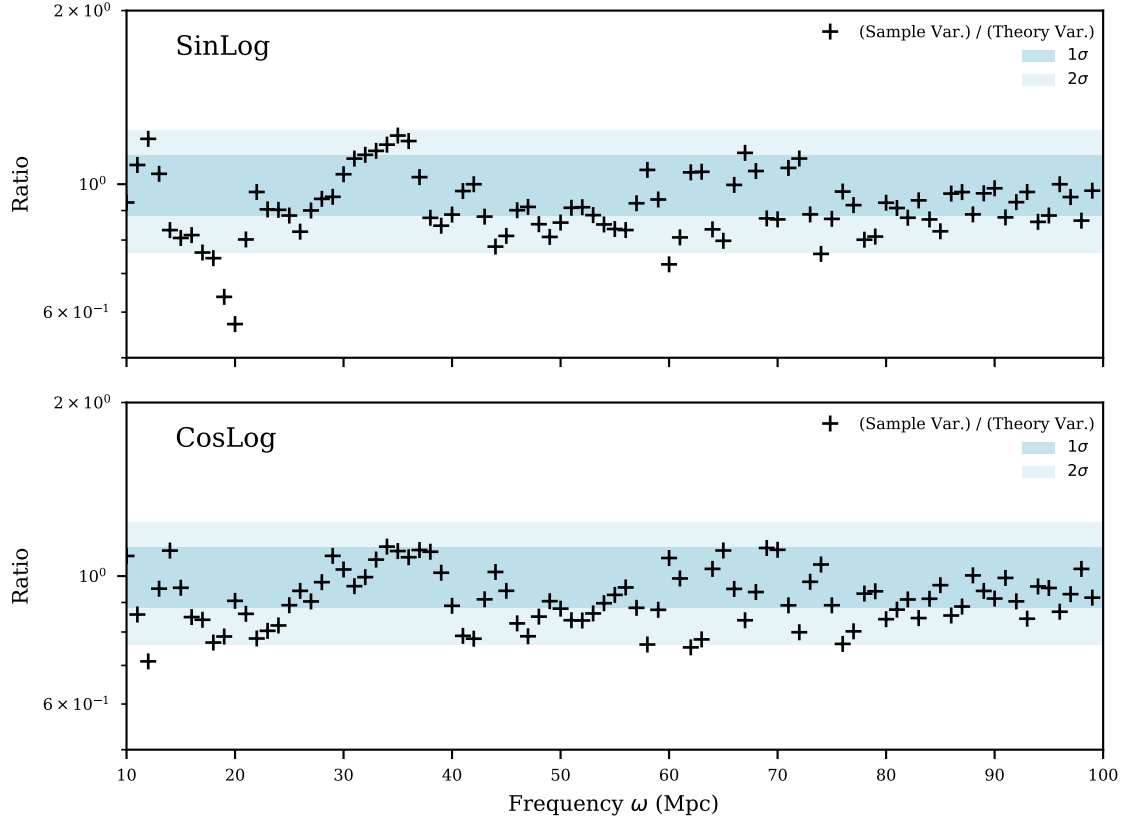
Fig. 5.13 Ratios between the sample and theoretical variances obtained from the Legendre basis for resonance shapes. The sample variance is calculated from the f_{NL} estimates of 140 Gaussian simulations. The 1σ and 2σ intervals (shaded in blue) are chosen based on a χ^2-distribution, which accurately describes the statistics of the sample variance as long as the f_{NL}s are Gaussian distributed. The results are statistically consistent with random fluctuations except potentially at $\omega = 20$, as discussed in the main text.

variances are calculated from a finite number of f_{NL} samples from simulations. We test if these estimates are compatible with the underlying distribution: χ^2 with $N_{sims} - 1$ degrees of freedom in this case. We achieve the desired consistency for both $\phi = 0$ and $\phi = \pi/2$. One potentially meaningful outlier is at $\omega = 20$ and $\phi = 0$, where the sample estimate is below 60% of the expected level. This anomaly is not likely to be a numerical error specific to CMB-BEST since a similar dip can be found from Modal results in Figure 5.12. No significant irregularity is found for different phases at the same frequency. We classify this point as a random fluctuation for now, but a further investigation using an independent set of Gaussian simulations is required to be certain.

As with the feature models before, we plot the correlation matrix between the templates with different ωs in Figure 5.14. Shapes with similar ωs are naturally correlated, but most off-diagonal terms of the matrix vanish. Even for $\omega > 35$ where the basis expansion becomes less accurate, cross-correlations tend to remain small until $\omega \approx 75$. We verify the stability of CMB-BEST's Legendre basis expansion; best approximations to the highly oscillatory template are found, allowing us to continue exploring independent bispectrum shapes with the characteristic log-spaced oscillations.

5.3.3   Proof of concept

PRIMODAL is a fast and efficient numerical code for computing bispectra of primordial perturbations given a single field inflation Lagrangian using in-in formalism [116]. The computed bispectrum is expressed as coefficients of a separable basis expansion in PRIMODAL, as opposed to a grid of discrete points like other in-in codes. Choosing an equivalent basis set for CMB-BEST therefore creates a direct link between the two codes. The combined pipeline is capable of constraining specific inflation models directly without the use of approximate templates. Such template-free bispectrum analysis also enables a fast and extensive scan of theory parameters.

Collaborating closely with the main author of PRIMODAL, Philip Clarke, we verified the integrity of the combined PRIMODAL + CMB-BEST routine. We thoroughly tested the consistency of basis functions, their orthogonalisation, and convergence. Both PRIMODAL and CMB-BEST are now in the exploitation phase. We present a working example of the combined pipeline as a proof of concept.

DBI inflation [125–128] is a well-studied single-field inflation model inspired by string theory. Its action follows the general form (3.15), with a non-canonical kinetic term given by

$$P(X, \phi) = -\frac{1}{f(\phi)} \left[\sqrt{1 - 2f(\phi)X} - 1 \right] - V(\phi), \tag{5.37}$$

Fig. 5.14 Correlations between the resonance model templates of shape given by $S(k_1, k_2, k_3) = \sin(\omega \log(k_1 + k_2 + k_3) + \phi)$ with different ω values, for (a) $\phi = 0$ and (b) $\phi = \pi/2$. 'Sample' correlations are obtained from the Legendre basis f_{NL} estimates from 140 Gaussian simulations. 'Theory' correlations come from the inner product induced by Γ as shown in (5.36).

where $X = -\frac{1}{2}g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi$ as before. $f(\phi)$ is an arbitrary function called the warp factor. We choose

$$f(\phi) = \frac{\lambda_{\text{DBI}}}{\phi^4}, \qquad V(\phi) = V_0 - \frac{1}{2}m^2\phi^2, \qquad m = \sqrt{\beta_{\text{IR}}}\,H, \tag{5.38}$$

for some theory parameters λ_{DBI}, V_0, and β_{IR}. The sound speed can then be obtained from (3.16), which evaluates to $c_{\text{s}}^{\text{DBI}} \approx 3/\beta_{\text{IR}}N_e$ under some slow-roll approximations [127]. Here, N_e denotes the number of e-folds until the end of inflation.

DBI inflation generates non-Gaussianity with its shape similar to the equilateral template (3.11). The latest constraints on the model come from the Planck CMB bispectrum analysis using an approximate template

$$B_\Phi^{\text{DBI}}(k_1,k_2,k_3) = \frac{6A_\Phi^2}{(k_1k_2k_3)^3}\frac{-3/7}{(k_1+k_2+k_3)^2}\left[\sum_i k_i^5 + \sum_{i\neq j}(2k_i^4 k_j - 3k_i^3 k_j^2) \right.$$
$$\left. + \sum_{i\neq j\neq l}(k_i^3 k_j k_l - 4k_i^2 k_j^2 k_l)\right]. \tag{5.39}$$

We switched to a convention where the potential Φ is used instead of ζ. On superhorizon scales at the end of inflation, they differ by a constant factor: $\Phi = (3/5)\zeta$. With respect to this template, the theoretical bispectrum from DBI inflation has an amplitude equal to

$$f_{\text{NL}}^{\text{DBI}} \approx -\frac{35}{108}\left[\frac{1}{(c_{\text{s}}^{\text{DBI}})^2} - 1\right], \tag{5.40}$$

an approximation which is accurate as long as the DBI sound speed $c_{\text{s}}^{\text{DBI}} \ll 1$. Constraints on $f_{\text{NL}}^{\text{DBI}}$ therefore place limits on c_{s}. Planck 2018 analysis [23] found $f_{\text{NL}}^{\text{DBI}} = 46 \pm 58$ with a 68% confidence level from CMB temperature data. The corresponding bound on the speed of sound is $c_{\text{s}}^{\text{DBI}} \geq 0.079$.

We reproduce the DBI sound speed constraint from Planck 2018 analysis using the PRIMODAL+CMB-BEST routine. Unlike the Planck analysis, our method does not require templates to connect inflationary predictions to CMB data. Instead of using a single estimate from an approximate template, we scan over DBI theory parameters and constrain the individual primordial bispectrum directly. A given parameter set can be ruled out if the corresponding constraint on f_{NL} excludes $f_{\text{NL}} = 1$ with high confidence.

One notable limitation in the current implementation of PRIMODAL is that the details of how inflation ends are not prescribed. This means that the scales during inflation and what we observe cannot be mapped directly, which restricts our ability to translate the results into

a constraint on more fundamental parameters such as β_{IR}. Our work here still serves as a useful validation of the combined pipeline. For the DBI sound speed scan, the approximate relation (5.40) is used to relate the output c_s^*—sound speed at some pivot scale during horizon crossing—from PRIMODAL with the physical DBI sound speed c_s^{DBI}. Numerous other validation tests we performed such as convergence tests with respect to the basis size p_{max} are unaffected by this caveat and still serve as a proof of concept for our combined pipeline.

The parameter β_{IR} is varied within the interval $[0.1885, 0.58]$ while fixing other parameters as $\lambda_{DBI} = 2.00475 \times 10^1 5$, $V_0 = 5.2 \times 10^{-12} M_{pl}^4$, and $\phi_0 = 0.46042 M_{pl}$, where M_{pl} is the Planck mass. Further details of the scan will be included in [31]. The results are depicted in Figure 5.15.

First of all, we check that our Legendre basis with $p_{max} = 30$ has converged and accurately represents the primordial bispectrum from DBI inflation. Values of the f_{NL} estimates obtained from $p_{max} = 30$ were tested against the equivalent result with $p_{max} = 29$ and shown to differ by less than 0.01%, as can be seen in Figure 5.15. In order to thoroughly investigate the degree of convergence, we studied a late-time equivalent of ε in (5.27) used to compare primordial bispectrum shapes;

$$\varepsilon^2(\mathbf{f}_1, \mathbf{f}_2) = \frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} \left(\frac{f_1^{(i)}}{\sigma_1} - \frac{f_2^{(i)}}{\sigma_2} \right)^2, \tag{5.41}$$

where $f_1^{(i)}$ and $f_2^{(i)}$ denote the f_{NL} estimates for the ith Gaussian simulated map from the two routines. ε^2 measures the mean squared error in the two sets of f_{NL} signal-to-noises for N_{sim} simulations. Comparing results from the Legendre basis with $p_{max} = 30$ and 25 across the scan, we confirmed that $\varepsilon < 0.01$ in the full range and $\varepsilon \approx 0.002$ for the majority. A plot with the full result will be provided in [31].

Secondly, we constrain the DBI sound speed as $c_s \gtrsim 0.056$ with 95% confidence using CMB temperature data from Planck. Each of the points in Figure 5.15 is normalised to some $\tilde{f}_{NL}$ such that $\tilde{f}_{NL} = 1$ corresponds to the bispectrum predicted by the model. Estimates from the Planck 2018 map take negative values with error bars small enough to exclude $f_{NL} = 1$ with 2σ confidence level when $c_s < 0.056$. Our constraint is similar but not identical to the result from Planck 2018 analysis which reads $c_s \geq 0.079$. The discrepancy stems from differences in the constraints for the DBI template (5.39): $f_{NL}^{DBI} = 27 \pm 58$ for PRIMODAL+CMB-BEST whereas $f_{NL}^{DBI} = 46 \pm 58$ for Planck. This $\approx 0.32\sigma$ gap in signal to noise is mainly because the Modal expansion is incomplete and the correlation between
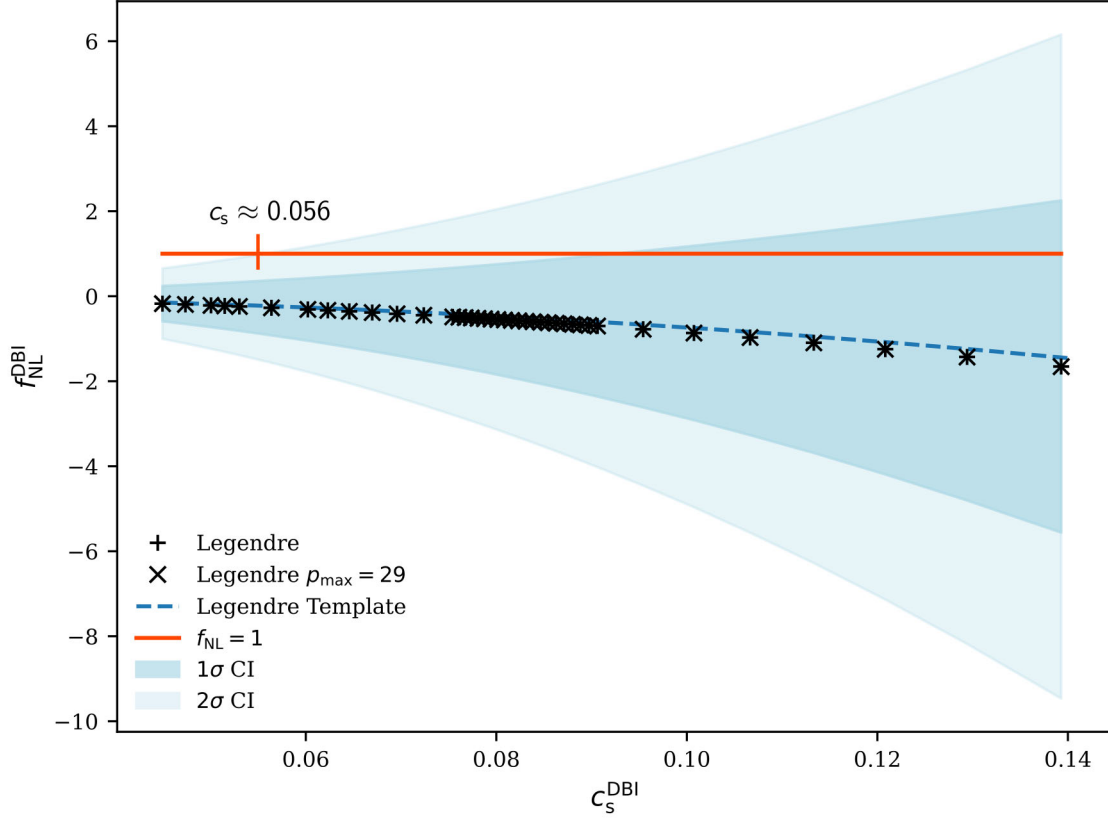
Fig. 5.15 Constraints on the DBI model with varying sound speed c_s. Primordial bispectrum computed from PRIMODAL is fed directly into CMB-BEST in the form of coefficients $\alpha_{p_1 p_2 p_3}$ appearing in Legendre basis expansion. Constraints from the Legendre basis with $p_{max} = 30$ (black +) are shown in comparison with $p_{max} = 29$ (black ×). 1σ and 2σ confidence intervals around 0 are shaded in blue and represent the uncertainty in the estimated f_{NL}s. Bispectrum shapes are normalised so that $f_{NL} = 1$ corresponds to the DBI model under consideration. Our constraint on DBI sound speed is $c_s^{DBI} > 0.056$ with 95% confidence level. Equivalent results from an approximate DBI bispectrum template (5.39) (blue dashed) give the c_s^{DBI} bound which is identical up to two significant figures.

the expanded and full DBI template equals $r \approx 0.95$. As discussed in Appendix B of [70], this level of correlation causes an expected scatter in f_{NL}s approximately equal to 0.3σ.

Lastly, we verify that the DBI template (5.39) accurately approximates the full numerical result obtained from PRIMODAL. Drawn in blue-dashed on Figure 5.15, the difference between the $\tilde{f}_{\mathrm{NL}}$ constraints obtained using PRIMODAL and the DBI template differ by $O(0.01\sigma)$.

We extend the PRIMODAL+CMB-BEST pipeline to study DBI models with resonance. The primordial bispectrum appears similar to the DBI template but has extra log-spaced oscillations superimposed. The Planck analysis covered the DBI shape and the resonance templates independently [23] but has not constrained the combination of the two. Our method studies the bispectrum obtained directly from inflation without using approximate templates. We vary the effective frequency ω of the oscillations while keeping other DBI theory parameters fixed. Figure 5.16 summarises the results of the scan.
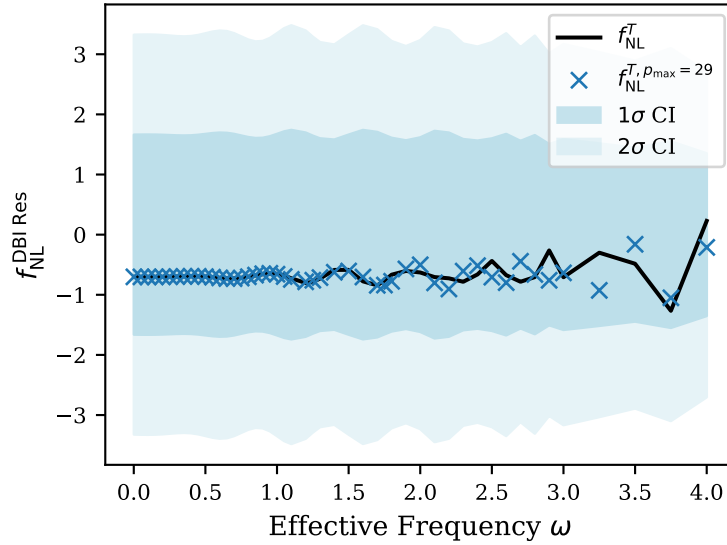
Fig. 5.16 Constraints on the DBI model with oscillations imposed on the inflationary potential, causing resonance-type bispectra approximately parametrised as $\sin(\omega \log(k_1 + k_2 + k_3) + \phi)$. The primordial bispectrum computed from PRIMODAL for various oscillation frequency ω has been fed directly to CMB-BEST. The results (black line) are shown together with the ones obtained from a smaller basis where $p_{\max} = 29$ (blue $\times$). The blue regions in the plot represent 1σ and 2σ confidence intervals.

We find no significant detection of non-Gaussianity in the parameter range we study. Note also that our f_{NL} estimates become unreliable after $\omega \gtrsim 2.0$ as the Legendre basis fails to represent rapid oscillations accurately. This can be seen by comparing the estimates from

the bases of size $p_{\max} = 30$ and 29 in the figure. Augmenting the Legendre basis with specific functions or choosing an alternative set of basis [116] can improve the convergence up to $\omega \approx 5$ (to be included in [31]), which is a significant improvement but still rather restrictive compared to the viable parameter range of $\omega \lesssim 35$ for the 'vanilla' resonance models shown in Figure 5.12.

Investigating the lack of convergence has revealed a more fundamental issue with the primordial basis expansion. Earlier in this chapter, we discussed how the tetrapyd domain of the primordial bispectrum can be extended to the cube given by $[k_{\min}, k_{\max}]^3$ such that the decomposition coefficients can be computed accurately and efficiently via (5.28). By doing so, we are expanding the given bispectrum in a cube which contains the tetrapyd while analytically continuing the function in between the two. Having good convergence in the cube then implies good convergence in the tetrapyd inside it. However, in cases where the target function behaves badly outside the tetrapyd, trying to fit the function in the whole cube instead of the tetrapyd may be suboptimal.

We demonstrate this issue by computing the f_{NL} constraints using a bispectrum template $S(k_1, k_2, k_3) = S^{\mathrm{equil}}(k_1, k_2, k_3) \sin(\omega \log(k_1 + k_2 + k_3))$: the equilateral shape with sine-log oscillations imposed. Figure 5.17 shows two plots analogous to Figure 5.14. The correlations between f_{NL}s estimated from different values of ω are computed using two different approaches for the primordial basis expansion. Shown on the left-hand side is the result for when we evaluate the given shape function $S(k_1, k_2, k_3)$ within the whole *cube*, as we have so far. We see that the basis fails to express the given bispectrum template accurately after $\omega \gtrsim 4$. Large correlations shown in the plot are simply artefacts of inaccuracy in the expansion.

On the right-hand side is the result for which the shape function values outside the tetrapyd were manually set to zero during the primordial basis expansion. The modified shape function is still continuous within the cubic domain, since the equilateral shape vanishes on boundaries of the tetrapyd despite its divergence outside of it. We see that this simple trick has dramatically improved the basis' ability to expand highly oscillatory bispectrum templates. The cross-correlations between templates remain small up until $\omega \approx 75$, quite similar to what we saw for the sine-log templates. We confirm that it is indeed the bad behaviour of the templates outside the tetrapyd domain during the primordial basis expansion that causes the lack of convergence. We plan to explore this topic further in the future.
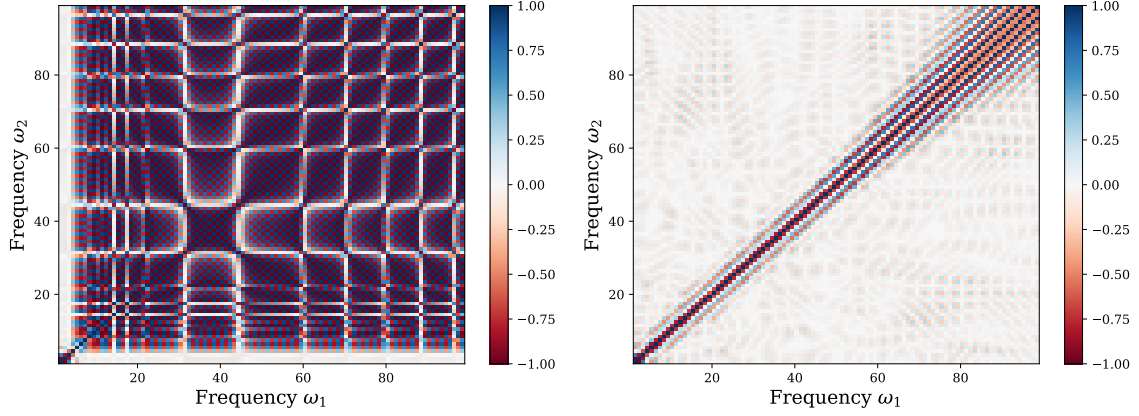
Fig. 5.17 Correlations between f_{NL} constraints for the enveloped resonance shape $S(k_1,k_2,k_3) = S^{\mathrm{equil}}(k_1,k_2,k_3)\sin(\omega\log(k_1+k_2+k_3))$ with different ωs. Since the given shape behaves badly in the region outside the tetrapyd but within the cubic domain of (5.28), we observe large off-diagonal correlations in the original Legendre basis expansion (Left). Manually setting the shape function to zero at these unphysical configurations therefore dramatically improves numerical performance (Right).

Summary

In this chapter, we presented a thorough review of our work on the high-resolution bispectrum estimator CMB-BEST. Starting with the mathematical framework of our pipeline, we showed that our formalism combines the strengths of conventional approaches to bispectrum estimation; we can handle highly oscillatory functions as accurately as the KSW estimator whilst covering a broad range of models like Modal. In order to tackle the computational complexity, we thoroughly optimised the code at both algorithmic and implementation levels. The code was then parallelised for maximal utilisation of high-performance computing clusters.

With the completed code we performed various tests to validate the integrity of CMB-BEST. We showed that the two different choices of basis functions—monomials and Legendre polynomials—give consistent constraints for the standard bispectrum shapes. Furthermore, we compared our results to those from the Modal estimator using the Planck 2018 data and found that they agree on a map-by-map basis. Oscillatory shapes such as the feature and resonance templates were also constrained; by analysing the correlation between the templates with different oscillation frequencies, we showed that our basis can handle oscillatory shapes as accurately as the Planck analysis using the Legendre basis of size $p_{\max} = 30$.

Having validated the pipeline, we provided a proof-of-concept example where CMB-BEST is combined with PRIMODAL to constrain inflation models with non-canonical kinetic terms. The combined pipeline is capable of constraining inflation models directly without using approximate templates. We studied the DBI inflation models while varying the speed of sound parameter to obtain the constraint $c_{\mathrm{s}}^{\mathrm{DBI}} \geq 0.056$ with 95% confidence. We are currently preparing to publish the work presented here as [31].

Chapter 6

Conclusion

The standard model of cosmology has been remarkably successful; the CMB power spectrum from Planck, for example, shows an exquisite fit to the ΛCDM model with only six free parameters. Many questions remain unanswered, however, especially regarding early universe physics. The most widely accepted theory for the early universe is inflation, whose prediction of the scale-invariant primordial spectra has been tested to be consistent with the observed data. Inflation also resolves the horizon and flatness problems present in the vanilla Big Bang cosmology. The simplest scenario of the single-field slow-roll inflation with a canonical kinetic term can explain the current observations, but numerous other physically well-motivated scenarios are yet to be ruled out.

The key to distinguishing different models of inflation is primordial non-Gaussianity. Deviations from the most canonical inflation model leave weak non-Gaussian signatures in the primordial perturbations. These imprints are best captured in the bispectrum of the primordial perturbations. Studying the shape and amplitude of the primordial bispectrum, therefore, allows us to constrain various inflation models directly. The CMB is the ideal probe for this job since its anisotropy depends linearly on the initial density perturbations. The CMB bispectrum hence directly relates to the primordial bispectrum and is used to construct the optimal estimator for the primordial non-Gaussianity parameter $f_{\rm NL}$.

The most recent Planck analysis has constrained $f_{\rm NL}$ to great precision using the CMB bispectrum estimator. No primordial source of non-Gaussianity has been found, and $f_{\rm NL}$s for the standard shapes are currently consistent with 0. More interesting 'hints' of non-Gaussianity have been found in models with oscillations in the bispectrum. Signals of 3-4σ significance have been found in these models, but scanning over a large parameter space meant that these are not yet conclusive detections. A further investigation using independent and more precise measurements would be extremely beneficial.

The next generation of CMB experiments is anticipated to measure the polarisation of the CMB with greatly enhanced sensitivity. Given that the Planck constraints on oscillatory models benefited much more from the inclusion of polarisation data than other shapes, we predict that the constraining power would benefit immensely from the future CMB data. Our work presented in Chapter 4 forecasts that the most sensitive CMB Stage-4 experiment specification is expected to yield a factor of 1.7-2.2 times more stringent constraints compared to Planck.

Despite the bright prospects and growing interest in constraining oscillatory models, a large part of the model and parameter space are currently unconstrained due to numerical and computational difficulties. The CMB bispectrum estimation is a challenging task where the naïve computation is practically impossible. There are two main approaches to this: KSW and Modal.

The KSW estimator exploits the inherent separability of the primordial bispectrum template and can accurately constrain even highly oscillatory shapes but is restrictive in the type of bispectrum shapes it can handle. On the other hand, the Modal estimator uses separable mode functions to expand the primordial and late-time bispectra and hence is able to constrain a wide range of non-separable shapes. However, dealing with high-frequency oscillations is challenging for Modal because there are two separate basis sets—primordial and late-time—to consider; high-frequency primordial basis functions lose a lot of their features during the projection, making it hard to expand them using the late-time basis functions accurately.

Our novel approach to the CMB bispectrum estimation, CMB-BEST, is designed to combine the strengths of the KSW and Modal estimators, and hence is suitable for constraining general oscillatory models using future CMB data. We use the primordial basis expansion to decompose the given bispectrum shape into separable terms like Modal but then apply a KSW-like method to obtain constraints on each of them. CMB-BEST works for general basis sets and contains the KSW estimator as a sub-case.

The main strength of CMB-BEST lies in its flexibility and accuracy; the code can take general basis functions which get projected using an exact analytic form. This would be especially useful for studying general bispectrum shapes with high-frequency oscillations that are currently unconstrained due to practical difficulties in numerical calculations.

The CMB-BEST pipeline is, however, computationally more costly than both the KSW and Modal ones. We therefore invested a significant amount of time optimising the code at both algorithmic and implementation levels. The code was then parallelised on multiple levels to fully benefit from the modern computing architecture. The total runtime has improved many orders of magnitude throughout this process. The completed code has then thoroughly

tested against the Modal pipeline used in the Planck 2018 analysis. We found that the results are consistent with Planck map-by-map using 140 simulated maps, for the standard shapes as well as the feature and resonance templates.

CMB-BEST has been validated and is now in the exploitation phase. Working in collaboration with the authors of PRIMODAL, we developed a fluid pipeline where a given inflationary Lagrangian can be directly constrained without the use of approximate templates. This allows a fast and accurate constraint to the specific model under consideration. As a proof-of-concept example, we presented work on constraining the DBI sound speed in Chapter 5, where we obtained $c_{\mathrm{s}}^{\mathrm{DBI}} \geq 0.056$ with 95% confidence. The combined PRIMODAL+CMB-BEST pipeline can perform similar scans on other theoretical parameters, which we plan to do in the near future.

We further demonstrated that CMB-BEST can indeed handle high-frequency oscillations of the resonance-type bispectra more stably compared to Modal. Its resolving power is limited only by the accuracy in its primordial basis expansion, unlike Modal which also depends on the late-time counterpart. Our next goal is to develop and utilise specialised basis sets tailored to highly oscillatory bispectrum shapes of interest.

There are currently two main routes for improvement for CMB-BEST. First, the code can be generalised to incorporate the E-mode polarisation data. Implementing this should be straightforward using an orthonormalisation defined in Chapter 4. The additional computational complexity involved in the extra set of spherical harmonic transforms for polarisation is expected to be subdominant and have a small impact overall. Second, we plan on improving the current method used for the decomposition of a given primordial shape function. We will investigate further the effect of the integration domain, as we saw from the example of the DBI resonance models in Chapter 5. We note that this concerns only the last step of the pipeline and the main parts of the bispectrum estimation remain unaffected.

It is truly an exciting time to be researching cosmology. Numerous future experiments will soon provide us with ever more immense and accurate measurements of the observable universe. We believe that the work presented here will add to the community's continued efforts to better understand the universe we live in.

List of figures

List of tables

References

[1] A.A. Penzias and R.W. Wilson, *A measurement of excess antenna temperature at 4080 Mc/s*, *The Astrophysical Journal* **142** (1965) 419.

[2] R.H. Dicke, P.J.E. Peebles, P.G. Roll and D.T. Wilkinson, *Cosmic black-body radiation.*, *The Astrophysical Journal* **142** (1965) 414.

[3] A. Einstein, *Die feldgleichungen der gravitation*, *Sitzung der physikalische-mathematischen Klasse* **25** (1915) 844.

[4] A. Friedman, *Über die krümmung des raumes*, *Zeitschrift für Physik* **10** (1922) 377.

[5] A. Friedmann, *Über die möglichkeit einer welt mit konstanter negativer krümmung des raumes*, *Zeitschrift für Physik* **21** (1924) 326.

[6] G. Lemaître, *Un univers homogène de masse constante et de rayon croissant rendant compte de la vitesse radiale des nébuleuses extra-galactiques*, in *Annales de la Société scientifique de Bruxelles*, vol. 47, pp. 49–59, 1927.

[7] H.P. Robertson, *Kinematics and world-structure*, *The Astrophysical Journal* **82** (1935) 284.

[8] A.G. Walker, *On milne's theory of world-structure*, *Proceedings of the London Mathematical Society* **2** (1937) 90.

[9] E. Hubble, *A relation between distance and radial velocity among extra-galactic nebulae*, *Proceedings of the national academy of sciences* **15** (1929) 168.

[10] A.H. Guth, *Inflationary universe: A possible solution to the horizon and flatness problems*, *Physical Review D* **23** (1981) 347.

[11] G.F. Smoot, C.L. Bennett, A. Kogut, E. Wright, J. Aymon, N. Boggess et al., *Structure in the COBE differential microwave radiometer first-year maps*, *The Astrophysical Journal* **396** (1992) L1.

[12] A.G. Riess, A.V. Filippenko, P. Challis, A. Clocchiatti, A. Diercks, P.M. Garnavich et al., *Observational evidence from supernovae for an accelerating universe and a cosmological constant*, *The Astronomical Journal* **116** (1998) 1009.

[13] V.C. Rubin, W.K. Ford Jr and N. Thonnard, *Rotational properties of 21 sc galaxies with a large range of luminosities and radii, from ngc 4605/r= 4kpc/to ugc 2885/r= 122 kpc*, *The Astrophysical Journal* **238** (1980) 471.

[14] Planck Collaboration, *Planck 2018 results. VI. Cosmological parameters*, *Astronomy & Astrophysics* **641** (2020) A6.

[15] Planck Collaboration, *Planck 2018 results. X. Constraints on inflation*, *Astronomy & Astrophysics* **641** (2020) A10.

[16] G. Hinshaw, D. Spergel, L. Verde, R. Hill, S. Meyer, C. Barnes et al., *First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: The Angular Power Spectrum*, *The Astrophysical Journal Supplement Series* **148** (2003) 135.

[17] J. Dunkley, R. Hlozek, J. Sievers, V. Acquaviva, P.A. Ade, P. Aguirre et al., *The Atacama Cosmology Telescope: cosmological parameters from the 2008 power spectrum*, *The Astrophysical Journal* **739** (2011) 52.

[18] Planck Collaboration, *Planck 2018 results. V. CMB power spectra and likelihoods*, *Astronomy & Astrophysics* **641** (2020) A5.

[19] BICEP/Keck Collaboration, *Improved constraints on primordial gravitational waves using planck, wmap, and bicep/keck observations through the 2018 observing season*, *Physical Review Letters* **127** (2021) 151301.

[20] A. Lewis, A. Challinor and D. Hanson, *The shape of the CMB lensing bispectrum*, *Journal of Cosmology and Astroparticle Physics* **2011** (2011) 018.

[21] X. Chen, *Primordial non-gaussianities from inflation models*, *Advances in Astronomy* **2010** (2010) .

[22] E. Komatsu, *Hunting for Primordial Non-Gaussianity in the Cosmic Microwave Background*, *Classical and Quantum Gravity* **27** (2010) [`1003.6097`].

[23] Planck Collaboration, *Planck 2018 results. IX. Constraints on primordial non-Gaussianity*, *Astronomy & Astrophysics* **641** (2020) A9.

[24] The Simons Observatory Collaboration, *The Simons Observatory: science goals and forecasts*, *Journal of Cosmology and Astroparticle Physics* **2019** (2019) 056.

[25] K.N. Abazajian et al., *CMB-S4 Science Book*, *arXiv preprint arXiv:1610.02743* (2016) [`1610.02743`].

[26] K. Abazajian, G. Addison, P. Adshead, Z. Ahmed, S.W. Allen, D. Alonso et al., *CMB-S4 science case, reference design, and project plan*, *arXiv preprint arXiv:1907.04473* (2019) .

[27] M. Hazumi, P.A. Ade, Y. Akiba, D. Alonso, K. Arnold, J. Aumont et al., *Litebird: a satellite for the studies of b-mode polarization and inflation from cosmic background radiation detection*, *Journal of Low Temperature Physics* **194** (2019) 443.

[28] P.D.M. et al, *Primordial non-gaussianity*, *Astro2020 Science White Paper* (2019) [`1903.04409`].

[29] W. Sohn and J.R. Fergusson, *CMB-S4 forecast on the primordial non-gaussianity parameter of feature models*, *Phys. Rev. D* **100** (2019) 063536.

[30] Planck Collaboration, *Planck 2015 results. XVII. Constraints on primordial non-Gaussianity*, Astronomy & Astrophysics **594** (2016) A17 [1502.01592].

[31] W. Sohn, P. Clarke, J.R. Fergusson and E.P.S. Shellard, *High-resolution CMB bispectrum estimator*, *in preparation* (2022) .

[32] S. Dodelson, *Modern cosmology*, Elsevier (2003).

[33] J.R. Fergusson, *Part III cosmology*, *Cambridge DAMTP Cosmology Lectures* (2020) .

[34] A. Challinor and H. Peiris, *Lecture notes on the physics of cosmic microwave background anisotropies*, in *AIP Conference Proceedings*, vol. 1132, pp. 86–140, American Institute of Physics, 2009.

[35] D. Baumann, *Primordial cosmology*, in *Theoretical Advanced Study Institute Summer School 2017" Physics at the Fundamental Frontier"*, vol. 305, p. 009, SISSA Medialab, 2018.

[36] E.A. Lim, *Advanced cosmology: Primordial non-gaussianities*, *Cambridge DAMTP Advanced Cosmology Lectures* (2012) .

[37] B. Ratra and P.J. Peebles, *Cosmological consequences of a rolling homogeneous scalar field*, *Physical Review D* **37** (1988) 3406.

[38] R.R. Caldwell, R. Dave and P.J. Steinhardt, *Cosmological imprint of an energy component with general equation of state*, *Physical Review Letters* **80** (1998) 1582.

[39] R. Brandenberger and P. Peter, *Bouncing cosmologies: progress and problems*, *Foundations of Physics* **47** (2017) 797.

[40] D.N. Spergel, L. Verde, H.V. Peiris, E. Komatsu, M. Nolta, C.L. Bennett et al., *First-year Wilkinson Microwave Anisotropy Probe (WMAP)* observations: determination of cosmological parameters*, *The Astrophysical Journal Supplement Series* **148** (2003) 175.

[41] Planck Collaboration, *Planck 2013 results. xvi. cosmological parameters*, Astronomy & Astrophysics **571** (2014) A16.

[42] J. Carlstrom, P.A. Ade, K. Aird, B. Benson, L. Bleem, S. Busetti et al., *The 10 meter South Pole Telescope*, *Publications of the Astronomical Society of the Pacific* **123** (2011) 568.

[43] P. Ade, Z. Ahmed, R. Aikin, K. Alexander, D. Barkats, S. Benton et al., *Improved constraints on cosmology and foregrounds from bicep2 and keck array cosmic microwave background data with inclusion of 95 ghz band*, *Phys. Rev. Lett.* (2016) .

[44] A. Kogut, D. Spergel, C. Barnes, C. Bennett, M. Halpern, G. Hinshaw et al., *First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Temperature-Polarization Correlation*, *The Astrophysical Journal Supplement Series* **148** (2003) 161.

[45] A. Lewis, A. Challinor and A. Lasenby, *Efficient Computation of Cosmic Microwave Background Anisotropies in Closed Friedmann-Robertson-Walker Models*, *The Astrophysical Journal* **538** (2000) 473 [9911177].

[46] D. Blas, J. Lesgourgues and T. Tram, *The cosmic linear anisotropy solving system (CLASS). Part II: approximation schemes*, *Journal of Cosmology and Astroparticle Physics* **2011** (2011) 034.

[47] J. Maldacena, *Non-Gaussian features of primordial fluctuations in single field inflationary models*, *Journal of High Energy Physics* **2003** (2002) 13 [0210603].

[48] C. Burrage, R.H. Ribeiro and D. Seery, *Large slow-roll corrections to the bispectrum of noncanonical inflation*, *Journal of Cosmology and Astroparticle Physics* **2011** (2011) 032.

[49] C.T. Byrnes and K.-Y. Choi, *Review of local non-gaussianity from multifield inflation*, *Advances in Astronomy* **2010** (2010) .

[50] P. Creminelli, A. Nicolis, L. Senatore, M. Tegmark and M. Zaldarriaga, *Limits on non-Gaussianities from WMAP data*, *Journal of Cosmology and Astroparticle Physics* **2006** (2006) 004.

[51] L. Senatore, K.M. Smith and M. Zaldarriaga, *Non-gaussianities in single field inflation and their optimal limits from the wmap 5-year data*, *Journal of Cosmology and Astroparticle Physics* **2010** (2010) 028.

[52] R. Arnowitt, S. Deser and C.W. Misner, *Republication of: The dynamics of general relativity*, *General Relativity and Gravitation* **40** (2008) 1997.

[53] X. Chen, M.-x. Huang, S. Kachru and G. Shiu, *Observational signatures and non-gaussianities of general single-field inflation*, *Journal of Cosmology and Astroparticle Physics* **2007** (2007) 002.

[54] J. Schwinger, *Brownian motion of a quantum oscillator*, *Journal of Mathematical Physics* **2** (1961) 407.

[55] R.D. Jordan, *Effective field equations for expectation values*, *Physical Review D* **33** (1986) 444.

[56] E. Calzetta and B.-L. Hu, *Closed-time-path functional formalism in curved spacetime: Application to cosmological back-reaction problems*, *Physical Review D* **35** (1987) 495.

[57] F. Olver, D. Lozier, R. Boisvert and C. Clark, *The NIST Handbook of Mathematical Functions*, Cambridge University Press (2010).

[58] E. Komatsu, D.N. Spergel and B.D. Wandelt, *Measuring Primordial Non-Gaussianity in the Cosmic Microwave Background*, *The Astrophysical Journal* **634** (2005) 14 [0305189].

[59] J.R. Fergusson, M. Liguori and E.P.S. Shellard, *General cmb and primordial bispectrum estimation: Mode expansion, map making, and measures of f_{NL}*, *Physical Review D* **82** (2010) 023502.

[60] J.R. Fergusson, M. Liguori and E.P.S. Shellard, *The CMB bispectrum*, *Journal of Cosmology and Astroparticle Physics* **2012** (2012) [1006.1642].

[61] M. Bucher, B. van Tent and C.S. Carvalho, *Detecting bispectral acoustic oscillations from inflation using a new flexible estimator*, *Monthly Notices of the Royal Astronomical Society* **407** (2010) 2193 [0911.1642].

[62] M. Bucher, B. Racine and B. Van Tent, *The binned bispectrum estimator: Template-based and non-parametric CMB non-Gaussianity searches*, *Journal of Cosmology and Astroparticle Physics* **2016** (2016) [1509.08107].

[63] D. Babich, *Optimal estimation of non-Gaussianity*, *Physical Review D* **72** (2005) 043003.

[64] P. Creminelli, L. Senatore and M. Zaldarriaga, *Estimators for local non-Gaussianities*, *Journal of Cosmology and Astroparticle Physics* **2007** (2007) 019.

[65] K.M. Smith and M. Zaldarriaga, *Algorithms for bispectra: Forecasting, optimal analysis and simulation*, *Monthly Notices of the Royal Astronomical Society* **417** (2011) 2 [0612571].

[66] M. Münchmeyer, F. Bouchet, M.G. Jackson and B. Wandelt, *The Komatsu Spergel Wandelt estimator for oscillations in the cosmic microwave background bispectrum*, *Astronomy & Astrophysics* **570** (2014) A94.

[67] J.R. Fergusson, *Efficient optimal non-Gaussian CMB estimators with polarization*, *Physical Review D* **90** (2014) [1403.7949].

[68] M. Shiraishi, M. Liguori and J.R. Fergusson, *General parity-odd cmb bispectrum estimation*, *Journal of Cosmology and Astroparticle Physics* **2014** (2014) 008.

[69] M. Shiraishi, M. Liguori, J.R. Fergusson and E.P.S. Shellard, *General modal estimation for cross-bispectra*, *Journal of Cosmology and Astroparticle Physics* **2019** (2019) 046.

[70] Planck Collaboration, *Planck 2013 results. XXIV. Constraints on primordial non-Gaussianity*, *Astronomy & Astrophysics* **571** (2014) A24.

[71] M. Bartelmann and P. Schneider, *Weak gravitational lensing*, *Physics Reports* **340** (2001) 291.

[72] A. Lewis and A. Challinor, *Weak gravitational lensing of the CMB*, *Physics Reports* **429** (2006) 1.

[73] D. Hanson, A. Challinor and A. Lewis, *Weak lensing of the CMB*, *General Relativity and Gravitation* **42** (2010) 2197.

[74] Planck Collaboration, *Planck 2013 results. XIX. The integrated Sachs-Wolfe effect*, *Astronomy & Astrophysics* **571** (2014) A19.

[75] W.R. Coulton, P.D. Meerburg, D.G. Baker, S. Hotinli, A.J. Duivenvoorden and A. Van Engelen, *Minimizing gravitational lensing contributions to the primordial bispectrum covariance*, *Physical Review D* **101** (2020) 123504.

[76] J.-F. Cardoso, M. Le Jeune, J. Delabrouille, M. Betoule and G. Patanchon, *Component separation with flexible models-application to multichannel astrophysical observations*, *IEEE Journal of Selected Topics in Signal Processing* **2** (2008) 735.

[77] S. Basak and J. Delabrouille, *A needlet internal linear combination analysis of WMAP 7-year data: estimation of CMB temperature map and power spectrum*, *Monthly Notices of the Royal Astronomical Society* **419** (2012) 1163.

[78] E. Martínez-González, J. Diego, P. Vielva and J. Silk, *Cosmic microwave background power spectrum estimation and map reconstruction with the expectation-maximization algorithm*, *Monthly Notices of the Royal Astronomical Society* **345** (2003) 1101.

[79] H. Eriksen, J. Jewell, C. Dickinson, A. Banday, K. Górski and C. Lawrence, *Joint Bayesian component separation and CMB power spectrum estimation*, *The Astrophysical Journal* **676** (2008) 10.

[80] Planck Collaboration, *Planck 2018 results. IV. Diffuse component separation*, *Astronomy & Astrophysics* **641** (2020) A4.

[81] H.F. Gruetjen, J.R. Fergusson, M. Liguori and E.P.S. Shellard, *Using inpainting to construct accurate cut-sky CMB estimators*, *Physical Review D* **95** (2017) 043532.

[82] J.R. Fergusson, H.F. Gruetjen, E.P.S. Shellard and B. Wallisch, *Polyspectra searches for sharp oscillatory features in cosmic microwave sky data*, *Physical Review D* **91** (2015) [1412.6152].

[83] M. Liguori, E. Sefusatti, J.R. Fergusson and E.P. Shellard, *Primordial non-gaussianity and bispectrum measurements in the cosmic microwave background and large-scale structure*, *Advances in Astronomy* **2010** (2010) [1001.4707].

[84] A.A. Starobinskij, *Spectrum of adiabatic perturbations in the universe when there are singularities in the inflationary potential*, *Soviet Journal of Experimental and Theoretical Physics Letters* **55** (1992) 489.

[85] J. Adams, R. Easther and B. Cresswell, *Inflationary perturbations from a potential with a step*, *Physical Review D* **64** (2001) 6 [0102236].

[86] X. Chen, R. Easther and E.A. Lim, *Large non-Gaussianities in single-field inflation*, *Journal of Cosmology and Astroparticle Physics* **06** (2007) [0611645v3].

[87] P. Adshead, C. Dvorkin, W. Hu and E.A. Lim, *Non-Gaussianity from step features in the inflationary potential*, *Physical Review D* **85** (2012) 1 [1110.3050].

[88] D.K. Hazra, A. Shafieloo, G.F. Smoot and A.A. Starobinsky, *Wiggly whipped inflation*, *Journal of Cosmology and Astroparticle Physics* **2014** (2014) [1404.0360].

[89] C. Dvorkin and W. Hu, *Generalized slow roll approximation for large power spectrum features*, *Physical Review D* **81** (2010) 1 [0910.2237].

[90] V. Miranda, W. Hu and P. Adshead, *Warp features in DBI inflation*, *Physical Review D* **86** (2012) 1 [1207.2186].

[91] N. Bartolo, D. Cannone and S. Matarrese, *The effective field theory of inflation models with sharp features*, *Journal of Cosmology and Astroparticle Physics* **2013** (2013) [1307.3483].

[92] A. Achúcarro, J.O. Gong, S. Hardeman, G.A. Palma and S.P. Patil, *Features of heavy physics in the CMB power spectrum*, *Journal of Cosmology and Astroparticle Physics* **2011** (2011) [1010.3693].

[93] J. Chluba, J. Hamann and S.P. Patil, *Features and new physical scales in primordial observables: theory and observation*, *International Journal of Modern Physics D* **24** (2015) 1530023.

[94] P.D. Meerburg, J.P. Van Der Schaar and P.S. Corasaniti, *Signatures of initial state modifications on bispectrum statistics*, *Journal of Cosmology and Astroparticle Physics* **2009** (2009) 018.

[95] J.R. Fergusson, H.F. Gruetjen, E.P.S. Shellard and M. Liguori, *Combining power spectrum and bispectrum measurements to detect oscillatory features*, 2015.

[96] J. Martin and C. Ringeval, *Superimposed oscillations in the WMAP data?*, *Physical Review D* **69** (2004) 9 [0310382].

[97] M. Benetti, M. Lattanzi, E. Calabrese and A. Melchiorri, *Features in the primordial spectrum: New constraints from WMAP7 and ACT data and prospects for the Planck mission*, *Physical Review D* **84** (2011) 1 [1107.4992].

[98] P.D. Meerburg, R.A. Wijers and J.P. van der Schaar, *WMAP7 constraints on oscillations in the primordial power spectrum*, *Monthly Notices of the Royal Astronomical Society* **421** (2012) 369 [1109.5264].

[99] P.D. Meerburg, D.N. Spergel and B.D. Wandelt, *Searching for oscillations in the primordial power spectrum. I. Perturbative approach*, *Physical Review D* **89** (2014) 19 [1308.3704].

[100] P.D. Meerburg, D.N. Spergel and B.D. Wandelt, *Searching for oscillations in the primordial power spectrum. II. Constraints from Planck data*, *Physical Review D* **89** (2014) 1 [1308.3705].

[101] T. Chantavat, C. Gordon and J. Silk, *Large scale structure forecast constraints on particle production during inflation*, *Physical Review D* **83** (2011) 1 [1009.5858].

[102] M. Ballardini, F. Finelli, C. Fedeli and L. Moscardini, *Probing primordial features with future galaxy surveys*, *Journal of Cosmology and Astroparticle Physics* **2016** (2016) [1606.03747].

[103] B. Hu and J. Torrado, *Searching for primordial localized features with CMB and LSS spectra*, *Physical Review D* **91** (2015) 1 [1410.4804].

[104] M. Benetti and J.S. Alcaniz, *Bayesian analysis of inflationary features in Planck and SDSS data*, *Physical Review D* **94** (2016) 1 [1604.08156].

[105] X. Chen, C. Dvorkin, Z. Huang, M.H. Namjoo and L. Verde, *The future of primordial features with large-scale structure surveys*, *Journal of Cosmology and Astroparticle Physics* **2016** (2016) [1605.09365].

[106] A.P.S. Yadav, E. Komatsu and B.D. Wandelt, *Fast estimator of primordial non-gaussianity from temperature and polarization anisotropies in the cosmic microwave background*, *The Astrophysical Journal* **664** (2007) 680.

[107] K.W. Ng and G.C. Liu, *Correlation Functions of Cmb Anisotropy and Polarization*, *International Journal of Modern Physics D* **08** (1999) 61 [9710012].

[108] N. Hale and A. Townsend, *Fast and Accurate Computation of Gauss–Legendre and Gauss–Jacobi Quadrature Nodes and Weights*, *SIAM journal on scientific computing* **35** (2013) A652.

[109] X. Chen, *Folded resonant non-Gaussianity in general single field inflation*, *Journal of Cosmology and Astroparticle Physics* **2010** (2010) 003.

[110] P.D. Meerburg, J.P. van der Schaar and M.G. Jackson, *Bispectrum signatures of a modified vacuum in single field inflation with a small speed of sound*, *Journal of Cosmology and Astroparticle Physics* **2010** (2010) 001.

[111] P.D. Meerburg and J.P. van der Schaar, *Minimal cutoff vacuum state constraints from cmb bispectrum statistics*, *Physical Review D* **83** (2011) 043520.

[112] E. Silverstein and A. Westphal, *Monodromy in the CMB: gravity waves and string inflation*, *Physical Review D* **78** (2008) 106003.

[113] R. Flauger, L. McAllister, E. Pajer, A. Westphal and G. Xu, *Oscillations in the CMB from axion monodromy inflation*, *Journal of Cosmology and Astroparticle Physics* **2010** (2010) 009.

[114] M. Münchmeyer, P.D. Meerburg and B.D. Wandelt, *Optimal estimator for resonance bispectra in the cmb*, *Physical Review D* **91** (2015) 043534.

[115] P.D. Meerburg, M. Münchmeyer and B. Wandelt, *Joint resonant cmb power spectrum and bispectrum estimation*, *Physical Review D* **93** (2016) 043536.

[116] P. Clarke and E.P.S. Shellard, *Probing inflation with precision bispectra*, *Journal of Cosmology and Astroparticle Physics* **2021** (2021) 002.

[117] K.M. Gorski, E. Hivon, A.J. Banday, B.D. Wandelt, F.K. Hansen, M. Reinecke et al., *Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere*, *The Astrophysical Journal* **622** (2005) 759.

[118] M. Reinecke and D.S. Seljebotn, *Libsharp–spherical harmonic transforms revisited*, *Astronomy & Astrophysics* **554** (2013) A112.

[119] J. Jeffers, J. Reinders and A. Sodani, *Intel Xeon Phi Processor High Performance Programming: Knights Landing Edition*, Elsevier Science (2016).

[120] A.S. William Gropp, Ewing Lusk, *Using MPI: portable parallel programming with the message-passing interface*, vol. 1, MIT press (1999).

[121] L. Dagum and R. Menon, *OpenMP: an industry standard API for shared-memory programming*, IEEE computational science and engineering **5** (1998) 46.

[122] Intel, *Second generation Intel Xeon Scalable Processors. Datasheet, Volume One, Electrical*, 4, 2019.

[123] Planck Collaboration, *Planck 2015 results. XII. Full Focal Plane Simulations*, Astronomy & Astrophysics **594** (2016) A12.

[124] Planck Collaboration, *Planck 2013 results. XII. Diffuse component separation*, Astronomy & Astrophysics **571** (2014) A12.

[125] E. Silverstein and D. Tong, *Scalar speed limits and cosmology: Acceleration from D-cceleration*, Physical Review D **70** (2004) 103505.

[126] M. Alishahiha, E. Silverstein and D. Tong, *DBI in the sky: non-Gaussianity from inflation with a speed limit*, Physical Review D **70** (2004) 123505.

[127] X. Chen, *Running non-Gaussianities in Dirac-Born-Infeld inflation*, Physical Review D **72** (2005) 123518.

[128] R. Bean, X. Chen, H. Peiris and J. Xu, *Comparing infrared Dirac-Born-Infeld brane inflation to observations*, Physical Review D **77** (2008) 023527.