CMB *B*-mode non-Gaussianity: Optimal bispectrum estimator and Fisher forecasts

Adriaan J. Duivenvoorden,[1,2,*] P. Daniel Meerburg,[3,4,5] and Katherine Freese[2,6,7]

[1]*Department of Physics: Joseph Henry Laboratories, Jadwin Hall, Princeton University, Princeton, New Jersey 08542, USA*
[2]*The Oskar Klein Centre for Cosmoparticle Physics, Department of Physics, Stockholm University, SE-106 91 Stockholm, Sweden*
[3]*Van Swinderen Institute for Particle Physics and Gravity, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands*
[4]*Kavli Institute for Cosmology, Madingley Road, Cambridge CB3 0HA, United Kingdom*
[5]*DAMTP, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WA, United Kingdom*
[6]*Department of Physics, University of Texas, Austin, Texas 78712, USA*
[7]*Department of Physics, University of Michigan, Ann Arbor, Michigan 48109, USA*

(Received 29 January 2020; accepted 23 June 2020; published 10 July 2020)

Upcoming cosmic microwave background (CMB) data can be used to explore harmonic 3-point functions that involve the *B*-mode component of the CMB polarization signal. We focus on bispectra describing the non-Gaussian correlation of the *B*-mode field and the CMB temperature anisotropies (T) and/or *E*-mode polarization, i.e., $\langle TTB\rangle$, $\langle EEB\rangle$, and $\langle TEB\rangle$. Such bispectra probe violations of the tensor consistency relation: the model-independent behavior of cosmological correlation functions that involve a large-wavelength tensor mode (gravitational wave). An observed violation of the tensor consistency relation would exclude a large number of inflation models. We describe a generalization of the Komatsu-Spergel-Wandelt (KSW) bispectrum estimator that allows statistical inference on this type of primordial non-Gaussianity with data of the CMB temperature and polarization anisotropies. The generalized estimator shares its statistical properties with the existing KSW estimator and retains the favorable numerical scaling with angular resolution. In this paper, we derive the estimator and present a set of Fisher forecasts. We show how the forecasts scale with various experimental parameters such as minimum and maximum multipole moments, relevant for, e.g., the upcoming ground-based Simons Observatory experiment and proposed *LiteBIRD* satellite experiment. We comment on possible contaminants due to secondary cosmological and astrophysical sources.

DOI: 10.1103/PhysRevD.102.023521

I. INTRODUCTION

Inflationary cosmology was proposed [1–3] to solve several cosmological puzzles: an early period of accelerated expansion explains the homogeneity, isotropy, and flatness of the Universe, as well as the lack of relic monopoles. One of the great successes of the inflationary paradigm is the production of small density inhomogeneities that grow to create the large-scale structure of the Universe today [4–8]. In addition, tensor modes produced during inflation lead to primordial gravitational waves that are potentially detectable in the polarization of the cosmic microwave background (CMB) [9–12]. Observations of the CMB provide tests of these predictions of inflation and can serve to distinguish between specific inflationary models.

As yet, CMB observations are consistent with a single slowly rolling scalar field as the inflaton, the field responsible for inflation [13]. For these single-field slow roll (SFSR) models, the fluctuations are described by primordial density fluctuations, which are nearly Gaussian, adiabatic, and nearly scale invariant [2,3]. Gaussianity implies that the 2-point correlation function of the density fluctuations uniquely determines all higher even n-point functions while all odd n-point functions vanish. In principle, inflation could be described by variants other than SFSR that introduce significant non-Gaussianity, such as multifield inflation; models with noncanonical kinetic terms or non–Bunch-Davies vacua [14]. As yet no evidence for primordial non-Gaussianity has been found in the *Planck* data [15]; hence many of these models have been ruled out. Conversely, evidence for non-Gaussian statistics

[*] adriaand@princeton.edu

in upcoming data would imply deviations from SFSR inflation and would provide an informative probe of the inflationary dynamics and the associated high-energy physics [16].

While the usual searches for primordial non-Gaussianity focus on the n-point statistics of scalar fluctuations, in this paper we concentrate on the relatively unexplored observational signatures of non-Gaussian correlations involving tensor fluctuations, as previously discussed by [17]. We propose to extend the search for primordial non-Gaussianity from one that only looks for the "scalar-scalar-scalar" correlation to one that also searches for the "scalar-scalar-tensor" correlation [17–23]: the non-Gaussian correlation between two modes of the primordial scalar perturbation and a mode of the tensor perturbation produced during inflation. To enable this goal, we generalize the statistical inference framework used for primordial non-Gaussianity.

The scalar-scalar-tensor correlation is parametrized in terms of the Fourier coefficients of the curvature (scalar) perturbation ζ [24,25] and the two helicity modes ${}^{\pm2}h_{\mathbf{k}}$ that describe the tensor perturbation [18]:

$$\langle \zeta_{\mathbf{k}_1} \zeta_{\mathbf{k}_2} {}^{\pm2}h_{\mathbf{k}_3} \rangle = (2\pi)^3 \delta^{(3)}(\mathbf{q}) {}^{\pm2}F(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3), \quad (1)$$

with $\mathbf{q} = \mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3$. The ${}^{\pm2}F$ functions depend on the inflationary dynamics and can differ between models. Both $\zeta_{\mathbf{k}}$ and ${}^{\pm2}h_{\mathbf{k}}$ are described in the early radiation-dominated Universe at a time when their comoving wavelength $2\pi/k$ (with $k \equiv |\mathbf{k}|$) is larger than the Universe's "comoving Hubble radius" $(aH)^{-1}$ in natural units. $H(t)$ and $a(t)$ are the Hubble parameter and the Robinson-Walker scale factor as a function of cosmic time. Both types of perturbations are assumed to be "adiabatic," implying that they do not evolve on these "superhorizon" scales [26].

Evidence for a nonzero $\zeta\zeta h$ 3-point function would not only point toward a deviation from SFSR inflation [18] but also would potentially rule out the majority of currently formulated models of inflation [21]. The reason for this is a robust consistency relation for the "squeezed limit": $|\mathbf{k}_3| \ll |\mathbf{k}_1| \approx |\mathbf{k}_2|$, of the $\zeta\zeta h$ correlation. In the squeezed limit, $\zeta\zeta h$ is completely determined by $P_\zeta(k)$ and $P_h(k)$, the power spectra of $\zeta_{\mathbf{k}}$ and ${}^{\pm2}h_{\mathbf{k}}$ [18,27]:

$$\frac{{}^{\pm2}F(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3)}{P_\zeta(k_1) P_h(k_3)} = \left(\frac{4 - n_s}{2}\right) (\hat{k}_1)^a (\hat{k}_2)^b e^{\pm2}_{ab}(\hat{\mathbf{k}}_3). \quad (2)$$

The relation is independent from the dynamics of scalar fields present during inflation and holds as long as modes of the tensor perturbation become adiabatic directly after reaching a superhorizon scale during inflation [21]. The "polarization tensors" $e^{\pm2}_{ab}$ with $a, b \in \{1, 2, 3\}$ are two traceless, transverse tensor fields that will be precisely defined later. $n_s - 1$ parametrizes how much $P_\zeta(k)$ deviates from the scale-invariant form; see Appendix C 1.

The tensor consistency relation in Eq. (2) is powerful because its predictions are falsified if a significant $\zeta\zeta h$ correlation is detected in the squeezed limit [28,29]. An observed violation of the tensor consistency relation would indicate that inflation is described by a nonstandard variant. For example, the relation is violated by inflationary models with light, nonzero spin fields that do not decay quickly after leaving the horizon [21]. As a consequence, falsification of the tensor consistency relation allows ruling out models that approximately respect the de Sitter isometries [30], except for isometry-respecting models with so-called partially massless spin fields [27,31]. Other inflationary models that cannot be ruled out by falsification are those that weakly break some of the de Sitter isometries and couple the extra spin fields to the resulting preferred spatial slicing [32–34]. Furthermore, in models where a subset of the de Sitter isometries is strongly broken, there is no reason for the consistency relation to hold [22,23]. This last class includes models in which the tensor perturbations are produced by additional fields [35,36]. These models generally also make predictions for large tensor non-Gaussianity in different forms than just the squeezed $\zeta\zeta h$ type [37].

It should be noted that tests for the consistency relation of the squeezed scalar-scalar-scalar ($\zeta\zeta\zeta$) correlation [38,39], which are similar to the tests for the tensor consistency relation, are already underway [15]. The consistency relation for the squeezed $\zeta\zeta\zeta$ correlation holds for single-field inflation models [28].[1] A detection of a significant $\zeta\zeta\zeta$ correlation in the squeezed limit would experimentally rule out the validity of the consistency relation and would provide evidence for the presence of more than one time-evolving scalar field during inflation. The tensor consistency relation in Eq. (2) is arguably more general than the $\zeta\zeta\zeta$ counterpart as it will, in principle, still hold for models with multiple scalar fields [21,45].

The CMB contains cosmological information both in its temperature anisotropies (T) and in its linear polarization. The polarization field can be divided into two components: the parity-even E-mode and parity-odd B-mode fields [7,8]. Primordial scalar perturbations source T and E-mode polarization, while primordial tensor perturbations source the T, E-, and B-mode fields. Observational searches using T and E constrain both scalar and tensor perturbations. However, the contributions to T and E from scalars are much larger than those of tensors, and so cosmic variance, due to the limited number of measurable modes, prohibits strong constraints on tensor perturbations with T and E data. The inclusion of B-mode data allows for much tighter

[1] The exception are single-field models that relax the standard assumption of a Bunch-Davies vacuum state [40,41]. Single-field nonattractor models [42,43] also do not conform to the consistency relation, but still do not produce an observable $\zeta\zeta\zeta$ correlation in the squeezed limit [44].

constraints on tensor perturbations [46]. Furthermore, unlike T, current B-mode observations are not cosmic-variance limited; hence sensitivity to primordial tensor perturbations can significantly increase with B-mode polarization data [47].

For these reasons, this paper focuses on bispectra, the harmonic equivalent of 3-point correlation functions, that describe how a single B-mode perturbation is correlated to perturbations in the E-mode field or the CMB temperature, i.e., the $\langle TTB\rangle$, $\langle EEB\rangle$, and $\langle TEB\rangle$ bispectra. These correlations are currently unconstrained, but will be within reach of observations by currently operating [48–51], upcoming [52–57], and proposed [47,58] experiments. Since B-modes are sourced by primordial tensor modes, these bispectra directly probe the $\zeta\zeta h$ correlation. The use of the $\langle TTB\rangle$, $\langle EEB\rangle$, and $\langle TEB\rangle$ bispectra avoids much of the scalar-induced cosmic variance that plagues current constraints on the $\zeta\zeta h$ correlation.[2] These constraints are expected to improve by an order of magnitude with the inclusion of current B-mode data [17,22,47]. CMB constraints on $\zeta\zeta\zeta$, already close to the cosmic-variance limit, will not see such improvements.[3] Future constraints on $\zeta\zeta h$ will benefit from the ongoing, unified experimental effort to collect B-mode data in order to constrain the ratio of the primordial tensor-to-scalar ratio r:

$$r_{k_0} \equiv \frac{P_h(k_0)}{P_\zeta(k_0)}. \tag{3}$$

Besides the fact that a detection of a roughly scale-invariant tensor power spectrum $P_h(k)$ would provide a strong argument against a range of alternatives to inflation [72–75], constraints on r are used to differentiate between models of inflation [13]. For slow-roll models, r also provides the energy scale of inflation $V^{1/4}$: $V^{1/4} \sim r^{1/4} \times 10^{16}$ GeV [76]. The upper limit on r is determined by the BICEP2/*Keck Array* and *Planck* CMB data to be $r_{0.002} < 0.064$ (at 95% confidence level) [13]. In the case of a nondetection, upcoming B-mode observations have the potential to improve over the current 95% upper limit by factors of approximately 10 [52,56] and 30 [47,58].

Statistical inference on primordial non-Gaussianity is generally done using statistical "estimators." Loosely speaking, an estimator is a rule to transform observed data into a statistical estimate of a parameter of interest. Here we concentrate on a CMB bispectrum estimator that transforms CMB data into an estimate of the amplitude of a given bispectrum and, simultaneously, the amplitude of the primordial 3-point function responsible for this bispectrum. There is a complication associated with the $\zeta\zeta h$ 3-point function that prohibits a straightforward implementation of the standard bispectrum estimator, see Eq. (40) [77–81]. Existing bispectrum estimators rely on a summary statistic of the CMB bispectrum: the so-called reduced bispectrum $b_{\ell_1\ell_2\ell_3}$, defined in Sec. III of this paper [82]. Data are usually compared to a version of the reduced bispectrum that is separable (factorizable) in ℓ_1, ℓ_2, and ℓ_3. For data with a large harmonic band-limit $\ell_{\max}$ this separable form reduces the computational scaling of the estimator from $\mathcal{O}(\ell_{\max}^5)$ to $\mathcal{O}(\ell_{\max}^3)$ [77]. The problem is that the $(\hat{k}_1)^a(\hat{k}_2)^b e_{ab}^{\pm2}(\hat{\mathbf{k}}_3)$ term that is present in the $\zeta\zeta h$ 3-point correlation function results in reduced bispectra that are not separable into ℓ_1, ℓ_2, and ℓ_3 [83]. Without a separable form of the reduced bispectrum, inference on $\zeta\zeta h$ likely becomes an enormous computational challenge.[4]

We demonstrate that a numerically efficient estimation of the amplitude of the $\zeta\zeta h$ 3-point correlation is still possible by making use of the *full* bispectrum instead of the reduced bispectrum, and we propose a generalization of the standard bispectrum estimator [Eq. (59)]. This generalization, which can be seen as the main result of this paper, allows for computationally efficient (and statistically optimal) estimation for all $\zeta\zeta h$ 3-point functions that include the $(\hat{k}_1)^a(\hat{k}_2)^b e_{ab}^{\pm2}(\hat{\mathbf{k}}_3)$ term in the following way:

$${}^{\pm2}F(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3) = f(k_1,k_2,k_3)(\hat{k}_1)^a(\hat{k}_2)^b e_{ab}^{\pm2}(\hat{\mathbf{k}}_3). \tag{4}$$

Here it is assumed that f can be expressed as (a sum of terms) separable in the three wave numbers k_1, k_2, and k_3. It is argued how numerical evaluation still scales as $\mathcal{O}(\ell_{\max}^3)$ and how the proposed estimator is exact: it does not rely on

[2] The only relevant dedicated searches have been for a parity-violating 3-point tensor-tensor-tensor correlation using the *Planck* data in [15] and a search in the *WMAP* data in [59] for a $\zeta\zeta h$ correlation that violates the tensor consistency relation.

[3] While inference on certain standardized types of $\zeta\zeta\zeta$ non-Gaussianity will only improve by a factor of approximately two with upcoming CMB data [47], it is possible that more complicated non-Gaussian features would still be hidden in the data. This is especially true for models with oscillating or nonsmooth inflationary potentials (see, e.g., [60–64]) or models that predict non-Gaussian n-point correlation functions with $n > 3$ [27,65,66]. In terms of improving constraints on (especially squeezed) $\zeta\zeta\zeta$ correlation functions, observables such as galaxy clustering [67,68], 21 cm tomography [69], the cross-correlation between CMB lensing and galaxy clustering [70] or the cross-correlation between the primary CMB anisotropies and small-scale spectral distortions of the CMB [71] have the potential to significantly improve constraints in the (far) future.

[4] Approximate methods that retain some computational efficiency without relying on a separable form do exist (the binned bispectrum estimator [84] and the modal estimator [85–87]) and have been successfully applied in the *Planck* analysis [15,88,89]. The first constraint on the amplitude of the $\zeta\zeta h$ 3-point function in [59] was made with a modified [90] version of the modal estimator. Despite the fact that the binned and modal estimators are broadly applicable, they are relatively involved, are not strictly statistically optimal, and have an unnecessary computational overhead in the case of reduced bispectra that are already in separable form. For inference on such bispectra the dedicated estimator developed in Refs. [77–81] provides a simpler and more efficient solution.

lossy data compression or on the flat-sky approximation [17,91].

In Appendix A, it is shown how the estimator can be adapted to other nonstandard 3-point correlation functions. We derive an estimator for scalar 3-point functions that are sensitive to the presence of higher-spin fields during inflation [65,92,93] and provide estimators for 3-point functions that involve two or three tensor components. The 3-point functions with multiple tensor components are relevant for inflation models with pseudoscalar–gauge field interactions [37,94–96], models with higher-derivative terms in the inflationary gravitational sector [97], and bimetric gravity models [98].

To illustrate the potential of the generalized estimator for testing the tensor consistency relation we provide a number of Fisher forecasts that represent idealized experimental outcomes. These forecasts demonstrate the ℓ_{min} and ℓ_{max} dependence of constraints on the amplitude of the squeezed $\zeta\zeta h$ correlation. The forecasts also show the influence of the lensing B-mode power spectrum, the effects of reionization, and the advantage of using both temperature and E-mode data in addition to the B-mode data. We comment on the expected contamination that is associated with B-mode data and the high-resolution data needed for squeezed 3-point functions. In future work the generalized estimator will be applied to simulated microwave sky data to evaluate the Fisher forecasts.

The current paper is organized as follows. We first review the CMB anisotropies, the bispectrum, and the primordial 3-point correlation function in Sec. II. We then introduce the generalized bispectrum estimator in Sec. III and present Fisher forecasts for the tensor-scalar-scalar bispectrum in Sec. IV. We discuss future work in Sec. V and conclude in Sec. VI.

II. PRELIMINARIES

A. CMB anisotropies

The data we consider are spherical harmonic modes of the CMB temperature and linear polarization anisotropies on the celestial sphere. After a brief review of the general properties of the harmonic modes, we will demonstrate the linear relation between the CMB anisotropies and the primordial scalar and tensor perturbations.

The temperature harmonic modes are related to the CMB temperature T measured at position $\hat{\mathbf{n}} \in S^2$ on the celestial sphere by

$$a_{T,\ell m} = \int_{S^2} \mathrm{d}\Omega(\hat{\mathbf{n}}) T(\hat{\mathbf{n}}) Y^*_{\ell m}(\hat{\mathbf{n}}), \tag{5}$$

where $\mathrm{d}\Omega(\hat{\mathbf{n}})$ and $Y^*_{\ell m}$ are the differential solid angle and a complex-conjugated spherical harmonic function, respectively. See Appendix B 1 for a summary of our notation.

The symmetric, traceless tensor field that describes the linearly polarized component of the microwave sky can be decomposed into two (real) fields: $Q(\hat{\mathbf{n}})$ and $U(\hat{\mathbf{n}})$. These fields are coordinate-dependent quantities that transform among themselves when the local coordinate basis (the tangent space) on the sphere at $\hat{\mathbf{n}}$ is rotated. For that reason, it is convenient to combine these fields into a complex "spin-2" field on the sphere, ${}^{(\pm 2)}P$, which is defined as follows:

$${}^{(\pm 2)}P(\hat{\mathbf{n}}) \equiv (Q \pm iU)(\hat{\mathbf{n}}). \tag{6}$$

Under a right-handed rotation of the local coordinate system around the point $\hat{\mathbf{n}}$ we then have

$${}^{(\pm 2)}P(\hat{\mathbf{n}}) \mapsto {}^{(\pm 2)}P(\hat{\mathbf{n}}) e^{\mp 2i\psi}, \tag{7}$$

where ψ is the angle of rotation. The sign of the exponent is a convention.

Instead of directly using ${}^{(\pm 2)}P$, we will describe polarization in terms of the harmonic modes of two fields that are scalars under coordinate rotations around $\hat{\mathbf{n}}$: the parity-even E field and the parity-odd B field. The harmonic modes of these two fields, the E- and B-modes, are related to the locally observable field as follows:

$$a_{E,\ell m} = -\frac{1}{2}\sum_{s\in\pm 2} \int_{S^2} \mathrm{d}\Omega(\hat{\mathbf{n}})\, {}^{(s)}P(\hat{\mathbf{n}})\, {}_sY^*_{\ell m}(\hat{\mathbf{n}}),$$
$$a_{B,\ell m} = -\frac{1}{2i}\sum_{s\in\pm 2} \mathrm{sgn}(s) \int_{S^2} \mathrm{d}\Omega(\hat{\mathbf{n}})\, {}^{(s)}P(\hat{\mathbf{n}})\, {}_sY^*_{\ell m}(\hat{\mathbf{n}}). \tag{8}$$

The spin-weighted spherical harmonics ${}_sY_{\ell m}$ form a complete and orthonormal basis for spin-s functions on the sphere, analogous to the regular spherical harmonics. See Appendix B 1 for a brief overview.

The parity-even E and parity-odd B harmonic modes transform differently under the parity transformation of the underlying spherical coordinates. Under parity, the odd moments of the temperature anisotropies and the E-mode field gain a minus sign. The opposite behavior holds for the B-mode field:

$$a_{T,\ell m} \mapsto (-1)^{\ell} a_{T,\ell m},$$
$$a_{E,\ell m} \mapsto (-1)^{\ell} a_{E,\ell m},$$
$$a_{B,\ell m} \mapsto (-1)^{\ell+1} a_{B,\ell m}. \tag{9}$$

To describe the primordial adiabatic scalar perturbations that source the CMB anisotropies, we use the gauge invariant curvature perturbation ζ [24,25].[5] As the initial

[5] The invariance under the choice of gauge (the choice of constant-time spacelike hypersurfaces and constant-position timelike worldlines) of ζ explains why it can simultaneously be interpreted as, e.g., the spatial curvature on hypersurfaces with constant energy density or as the energy density perturbation on spatially flat hypersurfaces [99].

adiabatic state is constant on superhorizon scales, we only need to consider the amplitude of ζ on some spacelike hypersurface in the early radiation-dominated era when all Fourier modes of interest were superhorizon. The Fourier coefficients of this amplitude at early time t_i are given by

$$\zeta_{\mathbf{k}} \equiv \int \mathrm{d}^3\mathbf{x}\,\zeta(\mathbf{x},t)|_{t=\tilde{t}(t_i,\mathbf{x})} e^{-i\mathbf{k}\cdot\mathbf{x}}, \tag{10}$$

where $\tilde{t} = t + \delta t(\mathbf{x},t)$ parametrizes weakly perturbed spacelike hypersurfaces relative to comoving coordinates $\{\mathbf{x},t\}$ of the flat Friedmann-Lemaître-Robertson-Walker (FLRW) background. Throughout this work, $\mathbf{k}$ denotes a three-dimensional (3D) comoving wave vector.

The primordial tensor perturbation h is the traceless and divergenceless linear perturbation to the flat FLRW metric:

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + a^2(t)[\delta_{ab} + h_{ab}(\mathbf{x},t)]\mathrm{d}x^a\mathrm{d}x^b, \tag{11}$$

with $h^a_a = \partial_a h^{ab} = 0$. Instead of using the coordinate basis to describe the tensor perturbation, we use a basis that sits perpendicular to the unit wave vector $\hat{\mathbf{k}}$, spanned by the $\hat{e}_{(\pm)}$ unit vectors.[6] On this new basis, the tensor perturbation conveniently reduces to two helicity states with Fourier coefficients given by

$${}_{(\pm2)}h_{\mathbf{k}} \equiv \frac{e^{ab}_{\pm2}(\hat{\mathbf{k}})}{2}\int \mathrm{d}^3\mathbf{x}\,h_{ab}(\mathbf{x},t)|_{t=\tilde{t}(\mathbf{x},t_i)} e^{-i\mathbf{k}\cdot\mathbf{x}}. \tag{12}$$

The polarization tensors $e_{\pm2}$ are two symmetric, traceless, and transverse tensor fields that transform h from the comoving coordinate basis to the $\hat{e}_{(\pm)}$ basis. The polarization tensors have the following properties:

$$(e^{ab}_{\pm2})^*(\hat{\mathbf{k}}) = e^{ab}_{\mp2}(\hat{\mathbf{k}}), \tag{13}$$

$$e^{\lambda}_{ab}(\hat{\mathbf{k}})e^{ab}_{\lambda'}(\hat{\mathbf{k}}) = 2\delta^{\lambda}_{-\lambda'} \quad (\lambda \in \pm 2). \tag{14}$$

The tensor perturbation h is gauge invariant (in the same sense as ζ is) [100]. The helicity components ${}_{(\pm2)}h$ are scalars under coordinate transformations up to a phase factor depending on the orientation of the basis spanned by $\hat{e}_{(\pm)}$.[7]

Let us categorize the stochastic primordial (superhorizon) amplitudes in terms of their helicity λ:

$${}^{(\lambda)}\xi_{\mathbf{k}} = \begin{cases} \zeta_{\mathbf{k}} & \text{for } \lambda = 0 \\ {}^{(\lambda)}h_{\mathbf{k}} & \text{for } \lambda = \pm 2 \end{cases}. \tag{15}$$

Following the notation set by [83], we then write down a compact expression for the observed CMB modes in terms of these helicity-dependent superhorizon amplitudes and a set of rotationally invariant transfer functions $\mathcal{T}_\ell(k)$:

$$a^{(Z)}_{X,\ell m} = 4\pi(-i)^\ell \sum_\lambda \mathrm{sgn}(\lambda)^{\lambda+x} \times \int \frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^3}\,{}^{(-\lambda)}\xi_{\mathbf{k}}\mathcal{T}^{(Z)}_{X,\ell}(k)\,{}_{-\lambda}Y^*_{\ell m}(\hat{\mathbf{k}}), \tag{16}$$

with $Z \in \{\zeta(\text{scalar}), h(\text{tensor})\}$, $\mathrm{sgn}(0) \equiv 0$, $0^0 \equiv 1$, $X \in \{T, E, B\}$, and helicity and parity determined by

$$\lambda = \begin{cases} 0 & \text{for } Z = \zeta \\ \pm 2 & \text{for } Z = h \end{cases}, \qquad x = \begin{cases} 0 & \text{for } X = T, E \\ 1 & \text{for } X = B \end{cases}.$$

Note that by defining ${}_{\mp2}Y^*_{\ell m}$ in Eq. (16) on the transverse basis spanned by $\hat{e}_{(\pm)}$, we ensure that the $a_{X,\ell m}$ for $Z = h$ are independent of the orientation of this basis. This approach is fully analogous to the decomposition of the spin-2 polarization field in Eq. (8).

The transfer functions $\mathcal{T}_\ell(k)$ transform the superhorizon amplitudes $\zeta_{\mathbf{k}}$ and ${}^{(\lambda)}h_{\mathbf{k}}$ to the CMB radiation and its polarization seen today [8,101]. In short, once the comoving Hubble radius (growing after inflation has ended) becomes larger than the comoving wavelengths of $\zeta_{\mathbf{k}}$ and ${}^{(\lambda)}h_{\mathbf{k}}$, they "enter the horizon" and start to evolve with time. The scalar perturbations sourced by ζ begin to oscillate under the effects of gravity and photon pressure, resulting in the acoustic oscillations seen in the CMB angular power spectra. The helicity components ${}^{\pm2}h$ start to propagate through space as the two polarization states of a gravitational wave, virtually decoupled from the other components of the Universe, and decay away with the expansion of space [76,102,103]. As a result, the most prominent difference between the scalar and tensor transfer functions is that the latter result in small values for CMB fluctuations on small ($\ell > 100$) angular scales. Small-scale tensor perturbations that entered the horizon before recombination decay significantly before leaving their imprint on the CMB. The transfer functions depend only on the unperturbed background cosmology and are readily available through numerical Einstein-Boltzmann solvers such as CAMB [104,105] or CLASS [106].[8] The projection onto the celestial sphere is also handled by the transfer functions.

[6] To relate the basis vectors of the comoving coordinates $\hat{e}_{(a)}$ to those of the noncoordinate basis, we introduce a set of "polarization" vectors: $\{e_+, e_-, e_0\}$, such that $\hat{e}_{(\lambda)} = e_\lambda{}^a\hat{e}_{(a)}$ with $\lambda \in \{+,-,0\}$. Geometrically, the $\hat{e}_{(\pm)}$ basis vectors span the plane perpendicular to the wave vector, while $\hat{e}_{(0)}$ points along the wave vector. The three vectors form a complete orthonormal basis. We let $\hat{e}_{(\pm)}$ describe states of circular polarization; i.e., the polarization vectors obey $(e_\pm{}^a)^* = e_\mp{}^a$.

[7] The polarization tensors are defined in terms of the $\pm$ polarization vectors as $e^{ab}_{\pm2} \equiv \sqrt{2}e_\pm{}^a e_\pm{}^b$. In the Cartesian basis, we may define the polarization vector as $e_\pm = \{1, \pm i, 0\}/\sqrt{2}$ for a wave vector aligned with the $\hat{\mathbf{z}}$ direction. The addition of a complex phase $\exp(-i\psi)$ to this definition amounts to an equally suitable basis that is simply rotated around the wave vector. The polarization tensors and helicity components are thus defined up to $\exp(-2i\psi)$.

[8] See https://camb.info and http://class-code.net.

See Appendix C 1 for more details on the transfer functions used in this work.

With Eq. (16), we have quantified the relation between the CMB anisotropies and the primordial scalar and tensor fields. The relation reiterates an important point: the primordial scalar fluctuations do not source the parity-odd B-mode field at linear order [8]. Higher-order cosmological effects, such as weak lensing by matter along the line of sight [107] or second order time evolution of the scalar perturbations [108–110], create a B-mode signal even in the absence of a primordial tensor contribution. Such effects are not included in the linear transfer functions so their influence has to be described separately. The same is true for a signal from astrophysical foregrounds. We will briefly discuss these contributions in Sec. V but will consider them in more detail in a future paper.

B. Bispectrum and the primordial 3-point function

In Sec. II B 1, we summarize general properties of the observable of interest: the CMB bispectrum. As we are interested in bispectra that include a B-mode component, we explicitly discuss the inclusion of B-mode polarization. In Sec. II B 2, we then introduce the concept of a linearly propagated, or primary, bispectrum: a primordial 3-point correlation function that is evolved to the CMB bispectrum today by the linear transfer functions introduced in Sec. II A. In addition, we describe the primordial $\zeta\zeta h$ 3-point correlation function in more detail.

1. General properties of the bispectrum

The bispectrum is defined as the isotropic 3-point correlation function represented in terms of spherical harmonic coefficients. The bispectrum is proportional to the multivariate generalization of the skewness of a probability distribution and thus vanishes for purely Gaussian coefficients.

We can formulate a bispectrum for every combination of the temperature and polarization components $X_1, X_2, X_3 \in \{T, E, B\}$:

$$B^{\ell_1\ell_2\ell_3}_{m_1m_2m_3,X_1X_2X_3} \equiv \langle a_{X_1,\ell_1m_1} a_{X_2,\ell_2m_2} a_{X_3,\ell_3m_3}\rangle. \quad (17)$$

The $a_{X,\ell,m}$ are defined in Eqs. (5) and (8). Statistical isotropy constrains the azimuthal dependence such that the bispectrum may always be factored into a Wigner 3-j symbol and a factor independent of m_1, m_2, and m_3 [111,112]:

$$B^{\ell_1\ell_2\ell_3}_{m_1m_2m_3,X_1X_2X_3} = \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix} B^{X_1X_2X_3}_{\ell_1\ell_2\ell_3}. \quad (18)$$

We will refer to the left-hand side (lhs) as the bispectrum, while B on the right-hand side (rhs) is the angle-averaged bispectrum. See Appendix B for an overview of the Wigner 3-j symbols.

TABLE I. Factor gained after a parity transformation $\mathsf{P}\colon \hat{\mathbf{n}} \mapsto -\hat{\mathbf{n}}$, for bispectra $B^{\ell_1\ell_2\ell_3}_{m_1m_2m_3,X_1X_2X_3}$ grouped by X_1, X_2, X_3 polarization indices.

$\mathsf{P}\colon \hat{\mathbf{n}} \mapsto -\hat{\mathbf{n}}$	
TTT, TTE, TEE, TBB, EEE, EBB	$(-1)^{\ell_1+\ell_2+\ell_3}$
TTB, TEB, EEB, BBB	$(-1)^{\ell_1+\ell_2+\ell_3+1}$

It is possible to construct a parity-invariant bispectrum from three fields regardless of the parity behavior of the individual fields. This means that we can form a parity-invariant bispectrum for all combinations of T, E, and B. This is not the case for the angular power spectrum.[9] From Eq. (9), we see that invariance under parity alone imposes that $\ell_1+\ell_2+\ell_3 = \text{even}$ for bispectra with an even number of B-mode contributions and $\ell_1+\ell_2+\ell_3 = \text{odd}$ otherwise; see Table I [17,113].

Isotropy forces the $\ell_1+\ell_2+\ell_3 = \text{even}$ components of bispectra to be real while the $\ell_1+\ell_2+\ell_3 = \text{odd}$ parts are purely imaginary. This constraint can be deduced from the condition for isotropy in Eq. (18) and the reality condition of the harmonic coefficients:

$$a^*_{X,\ell m} = a_{X,\ell -m}(-1)^m, \quad (19)$$

which holds because the underlying $X = \{T, E, B\}$ fields are real valued. The combination of these two conditions together with the reality of the 3-j symbols then implies

$$(B^*)^{\ell_1\ell_2\ell_3}_{m_1m_2m_3} = \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ -m_1 & -m_2 & -m_3 \end{pmatrix} B_{\ell_1\ell_2\ell_3}(-1)^{\sum_{n=1}^{3} m_n},$$

which, through the property of the 3-j symbol in Eq. (B11), means that complex-conjugating the bispectrum results in the following behavior:

$$(B^*)^{\ell_1\ell_2\ell_3}_{m_1m_2m_3} = B^{\ell_1\ell_2\ell_3}_{m_1m_2m_3}(-1)^{\sum_{n=1}^{3}(m_n+\ell_n)}.$$

Note that the Wigner 3-j symbol vanishes for $m_1+m_2+m_3 \neq 0$. Clearly, the above relation also holds for the angle-averaged bispectrum $B_{\ell_1\ell_2\ell_3}$. We have

[9] Given the parity transformation rules in Eq. (9), we see that the 2-point cross correlation function between a B-mode coefficient and a T coefficient transforms under parity as

$$\langle a_{B,\ell_1m_1} a^*_{T,\ell_2m_2}\rangle \mapsto \langle a_{B,\ell_1m_1} a^*_{T,\ell_2m_2}\rangle(-1)^{\ell_1+\ell_2+1}.$$

Taken together with isotropy, which demands that the cross-correlation is proportional to $\delta_{\ell_1\ell_2}\delta_{m_1m_2}$, we see that there is no parity-invariant configuration. The BE power spectrum vanishes by extension.

TABLE II. Parity conservation forces the bispectrum to be purely real, to be purely imaginary, or to vanish, depending on its ℓ_1, ℓ_2, and ℓ_3 multipole indices and its X_1, X_2, and X_3 polarization indices.

	$\sum_{n=1}^{3}\ell_n = \text{odd}$	$\sum_{n=1}^{3}\ell_n = \text{even}$
TTT, TTE, TEE, TBB, EEE, EBB	Vanish	Real
TTB, TEB, EEB, BBB	Imaginary	Vanish

suppressed the X indices as the above holds for all combinations of polarization indices. See Table II for an overview of the geometric constraints on parity-invariant, isotropic bispectra. The fact that the bispectra of interest here—$\langle TTB\rangle$, $\langle TEB\rangle$, and $\langle EEB\rangle$—are purely imaginary is a consequence of the complex representation of the spherical harmonics that we use. Expressed in terms of the Stokes parameters Q and U, the corresponding 3-point correlations would be real valued and thus observable.

2. Linearly propagated bispectrum and primordial 3-point correlation function

We start by defining the linearly propagated, or primary, bispectrum in its most general form. As mentioned before, the linearly propagated bispectrum is formed by time evolving a primordial 3-point correlation function to the CMB bispectrum today using the linear transfer functions introduced in Sec. II A. We then introduce the standard scalar-only ($\zeta\zeta\zeta$) primordial 3-point correlation function as well as our main focus: the $\zeta\zeta h$ 3-point correlation function.

Let us parametrize the superhorizon 3-point correlation function, the object we are ultimately interested in, as a helicity-dependent quantity using the amplitudes introduced in Eq. (15):

$$ {}^{(\lambda_1\lambda_2\lambda_3)}B(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3) \equiv \langle {}^{(\lambda_1)}\xi_{\mathbf{k}_1}{}^{(\lambda_2)}\xi_{\mathbf{k}_2}{}^{(\lambda_3)}\xi_{\mathbf{k}_3}\rangle, \qquad (20)$$

where the helicity λ is 0 for scalar perturbations and ±2 for tensor perturbations. We can then, using Eq. (16), form the linearly propagated bispectrum [114]:

$$ B^{\ell_1\ell_2\ell_3(Z_1Z_2Z_3)}_{m_1m_2m_3,X_1X_2X_3} = \left(\prod_{n=1}^{3} 4\pi(-i)^{\ell_n}\sum_{\lambda_n}\mathrm{sgn}(\lambda_n)^{\lambda_n+x_n} \times \int \frac{\mathrm{d}^3\mathbf{k}_n}{(2\pi)^3}\, {}_{-\lambda_n}Y^{*}_{\ell_n m_n}(\hat{\mathbf{k}}_n)\mathcal{T}^{(Z_n)}_{X_n,\ell_n}(k_n)\right) \times {}^{(-\lambda_1-\lambda_2-\lambda_3)}B(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3). \qquad (21)$$

Note that the three Z indices of the bispectrum in Eq. (21) may each be either ζ or h.

We now consider the symmetries of the primordial 3-point function. The assumed translational invariance of the process generating the primordial fluctuations implies momentum conservation in Fourier space:

$$ {}^{(\lambda_1\lambda_2\lambda_3)}B(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3) = (2\pi)^3\delta^{(3)}(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3) \times {}^{(\lambda_1,\lambda_2,\lambda_3)}F(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3). \qquad (22)$$

What remains now is to consider certain expressions for the helicity-dependent ${}^{(\lambda_1,\lambda_2,\lambda_3)}F(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3)$ functions. In a regular analysis, these functions would be given by the model under consideration. Here we are more interested in classes of models, and so we use general parametrizations.

For the scalar-only ($\zeta\zeta\zeta$) 3-point function, isotropy demands that F depends only on scalar products of the three wave vectors: the individual amplitudes and $\mathbf{k}_1\cdot\mathbf{k}_2$, $\mathbf{k}_1\cdot\mathbf{k}_3$, and $\mathbf{k}_2\cdot\mathbf{k}_3$. F cannot depend on a pseudoscalar such as $\mathbf{k}_1\cdot(\mathbf{k}_2\times\mathbf{k}_3)$ in case of a parity-invariant 3-point correlation function. For simplicity, we use the following template:

$$ {}^{(000)}F(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3) = f^{(\zeta\zeta\zeta)}(k_1,k_2,k_3), \qquad (23)$$

where f is generally referred to as the shape of the bispectrum. We will make use of this standard $\zeta\zeta\zeta$ template to introduce the reader to existing estimation techniques later in this paper.

For the $\zeta\zeta h$ case, we use the following parametrization:

$$ {}^{(00\pm2)}F(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3) = f^{(\zeta\zeta h)}(k_1,k_2,k_3) \times (\hat{k}_1)^a(\hat{k}_2)^b e^{\pm2}_{ab}(\hat{\mathbf{k}}_3). \qquad (24)$$

Recall that roman indices denote three-dimensional spatial comoving coordinates; they are summed over when repeated. Note that ${}^{(00+2)}F$ and ${}^{(00-2)}F$ correspond to two independent 3-point functions; by denoting the shape function $f^{(\zeta\zeta h)}$ independent of helicity, we, however, implicitly assume parity invariance.

The class of $\zeta\zeta h$ 3-point functions described by Eq. (24) include those predicted by SFSR inflation [18]. The amplitude of the $\zeta\zeta h$ 3-point function will be too small to be observable with CMB data in the SFSR case. More importantly, the template in Eq. (24) also applies to the majority of mentioned models that violate the tensor consistency relation in Eq. (2) and thus potentially produce an observable signal [19–23]. We may therefore use Eq. (24) as the basis for inference on such models.

To gain intuition for the characteristics of the $\zeta\zeta h$ template, it is useful to realize that the delta function in Eq. (22) imposes that $\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3=\mathbf{0}$; i.e., the 3-point function is defined on triangular configurations of the three wave vectors. The $f^{(\zeta\zeta h)}(k_1,k_2,k_3)$ part of the $\zeta\zeta h$ template thus assigns a weight to each triangle based on the lengths of the three sides. While these weights completely determine the 3-point function in the $\zeta\zeta\zeta$ case, the $\zeta\zeta h$ case requires that two more aspects are taken into account. First, the $\zeta\zeta h$ 3-point function is always suppressed in triangular configurations wherein the wave vector of the $h_{\mathbf{k}}$ Fourier mode is roughly (anti)parallel to the wave vector(s) of one or both of the scalar modes. This suppression is not due to the $f^{(\zeta\zeta h)}$ weight function but is a consequence of the nature of the polarization tensors. Their transverse property demands that $(\hat{k})^a e^{\pm 2}_{ab}(\hat{\mathbf{k}}')$ vanishes as $\hat{\mathbf{k}}$ becomes equal to $\hat{\mathbf{k}}'$. We thus see a suppression when $\hat{\mathbf{k}}_3$ aligns with $\hat{\mathbf{k}}_2$ and/or $\hat{\mathbf{k}}_1$ in Eq. (24). Second, the transverse traceless behavior of the $\zeta\zeta h$ 3-point function is also reflected in its helicity dependence.

We have demonstrated how the CMB is affected by a nonzero primordial 3-point correlation function through the bispectrum. We have also introduced the $\zeta\zeta h$ 3-point correlation function in Eq. (24). The bulk of this work will focus on this 3-point function. Note that the estimation technique that will be presented in the following sections is, in principle, also applicable to other types of 3-point functions. For conciseness, the discussion of some other templates (including the SFSR scalar-tensor-tensor and tensor-tensor-tensor 3-point functions [18]) is placed in Appendix A.

III. ESTIMATOR

This section is organized as follows. We first introduce the general form of the bispectrum estimator in Sec. III A. In Sec. III B, we then summarize the existing numerically efficient implementation of the estimator, and in Sec. III C we present our new work: the generalization of the fast implementation to the $\zeta\zeta h$ case.

A. General bispectrum estimation

We summarize the properties of the now standard CMB bispectrum estimation method [78,82,115]: a parametric search for the amplitudes of theoretically motivated bispectrum templates using an estimator that consists of a cubic and a linear statistic. This method has been the basis for the *Planck* non-Gaussianity analysis [15]. A derivation of the estimator and discussion of its properties can be found in Appendix D.

The estimator yields an estimate of the overall (dimensionless) amplitude $f_{\mathrm{NL}} \in \mathbb{R}$ of a bispectrum. We thus parametrize the bispectrum of interest as

$$B(f_{\mathrm{NL}}) = f_{\mathrm{NL}} B_1, \tag{25}$$

where $B_1 \equiv B(f_{\mathrm{NL}} = 1)$ is a fixed theoretical template with suppressed ℓ and m indices.

In searches for primordial non-Gaussianity, the template B_1 is given by a normalized version of the linearly propagated bispectrum in Eq. (21). The linear nature implies that the f_{NL} parameter corresponds to the overall amplitude of the primordial 3-point correlation function ${}^{(-\lambda_1-\lambda_2-\lambda_3)}B$ in Eq. (21). In principle, the amplitudes of several templates can be jointly estimated (see Appendix D). Here we only need the single parameter variant.

The estimator for f_{NL} is given by

$$\begin{aligned}\hat{f}_{\mathrm{NL}} = \frac{1}{6\mathcal{I}_0} \sum_{\text{all}\,\ell,m} \sum_{\text{all}\,X} (B_1)^{\ell_1\ell_2\ell_3}_{m_1m_2m_3,X_1X_2X_3} \\ \times \{[(C^{-1}a)^{X_1}_{\ell_1m_1}(C^{-1}a)^{X_2}_{\ell_2m_2}(C^{-1}a)^{X_3}_{\ell_3m_3}] \\ - [(C^{-1})^{X_1X_2}_{\ell_1m_1\ell_2m_2}(C^{-1}a)^{X_3}_{\ell_3m_3} + \text{cyclic}]\},\end{aligned} \tag{26}$$

where $X \in \{T, E, B\}$. The data, $a_{X,\ell m}$, only enter in inverse-covariance-weighted form:

$$(C^{-1}a)^X_{\ell m} = \sum_{X'} \sum_{\ell',m'} (C^{-1})^{XX'}_{\ell m\ell' m'} a_{X',\ell' m'}. \tag{27}$$

Here C^{-1} is the inverse of the block matrix:

$$C_{\ell m\ell' m'} \equiv \begin{pmatrix} C_{TT} & C_{TE} & C_{TB} \\ C_{ET} & C_{EE} & C_{EB} \\ C_{BT} & C_{BE} & C_{BB} \end{pmatrix}_{\ell m\ell' m'}. \tag{28}$$

Each element is defined as

$$C_{XX',\ell m\ell' m'} = \langle a_{X,\ell m} a^*_{X',\ell' m'} \rangle, \tag{29}$$

with $X, X' \in \{T, E, B\}$. This covariance matrix includes both the signal and the noise covariance and is therefore generally not diagonal. The estimating procedure considers the covariances as fixed and known *a priori*.

Intuitively, the first and second lines, the "cubic term," in Eq. (26) serve as a matched filter that correlates the observed bispectrum with the theoretical template B_1. The terms linear in the data (first times third line) are usually jointly referred to as the "linear term" and effectively serve to counter the estimator variance induced by the anisotropic parts of the covariance matrix [78,81]. Only the cubic part of the estimator is needed in cases where the covariance matrix in Eq. (28) is rotationally invariant.[10] With weakly anisotropic covariance, the linear term can be neglected for nonsqueezed bispectrum templates and/or analyses without large-scale ($\ell \lesssim 100$) data [91].

The normalization of the estimator is given by the following (dimensionless) number:

$$\begin{aligned}\mathcal{I}_0 = \frac{1}{6} \sum_{\text{all}\,\ell,m} \sum_{\text{all}\,X} (B_1)^{\ell_1\ell_2\ell_3}_{m_1m_2m_3,X_1X_2X_3} \\ \times [(C^{-1})^{X_1X_4}_{\ell_1m_1\ell_4m_4}(C^{-1})^{X_2X_5}_{\ell_2m_2\ell_5m_5}(C^{-1})^{X_3X_6}_{\ell_3m_3\ell_6m_6}] \\ \times (B^*_1)^{\ell_4\ell_5\ell_6}_{m_4m_5m_6,X_4X_5X_6}.\end{aligned} \tag{30}$$

[10] One can check that the rotational invariance of the bispectrum forces the linear term to be proportional to the (unobservable) CMB monopole perturbation when the covariance matrix is rotationally invariant [115].

Note that $\mathcal{I}_0$ is completely independent from the observed data.

The estimator is often referred to as "optimal." The word optimal refers to the fact that, in the appropriate limit, the estimator yields an unbiased point estimate of $f_{\rm NL}$ with variance given by the inverse of the model's Fisher information on $f_{\rm NL}$. It should be noted that this behavior is strictly true only in the limit where all non-Gaussian signal vanishes, and this includes $f_{\rm NL} \to 0$. The expression in Eq. (30) becomes equal to the Fisher information on $f_{\rm NL}$ in this limit. In Appendix D, we specify the likelihood function of the data to make the above statements more precise.

The estimator in Eq. (26) is well-suited to estimate upper limits on $f_{\rm NL}$. When a weak non-Gaussian signal is present, the estimator is still usable, but one has to be wary of biases and nonoptimal variance [116,117]. This is especially relevant for *B*-mode data contaminated by Galactic signal or high-resolution data with relatively strong non-Gaussian contributions from, e.g., weak lensing. See the discussion in Sec. V for more details.

We end this summary with a practical note on the inverse covariance matrix C^{-1}. The matrix frequently appears as part of the matrix-vector product $C^{-1}a$ [see Eq. (27)]. The C matrix, given by the sum of the signal and noise covariance matrices, is generally too large and dense to allow for regular matrix operations, such as matrix inversion. However, by separating C into the signal covariance, which is diagonal in the harmonic basis, and the noise covariance, which is typically close to diagonal in the (pixel) coordinate basis, it is straightforward to apply the C matrix to a vector. This makes it possible to avoid explicitly calculating the inverse covariance matrix C^{-1} when computing the matrix-vector product $C^{-1}a$. Simply put, one recasts the problem into the linear equation $Cx = a$ and solves for x using an iterative method. The method converges to the correct answer as the equation is solved by $x = C^{-1}a$. Suitable iterative methods (e.g., the conjugate gradient method) start with an initial guess for the vector x and iteratively update this guess until $a - Cx$ falls below a predetermined threshold. Crucially, these methods only rely on the capability of applying C, or related matrices, to a vector. The inverse operation, where C^{-1} is applied to a vector, is not required. See Appendix A from [118] for a detailed description of such an iterative approach.

However, there are also cases in which the C^{-1} matrix does not appear as part of a matrix-vector product. The linear term and estimator normalization rely on sums over the elements of the C^{-1} matrix itself; see the third line of Eq. (26) and the second line of Eq. (30), respectively. The distinction between the matrix-vector product and the matrix itself is important. The iterative methods for the $C^{-1}a$ operation are not suitable for direct computation of the full C^{-1} matrix. Whenever an isolated C^{-1} matrix appears, it typically has to be replaced by a Monte Carlo estimate of C^{-1} that is generated using inverse-covariance-weighted Gaussian $a_{X,\ell m}$ with the same signal covariance, noise covariance, masking, etc., as the data, i.e., drawn from the distribution specified by Eq. (28):

$$(C^{-1})^{XX'}_{\ell m \ell' m'} \approx \langle (C^{-1}a)^{X}_{\ell m} (C^{-1}a^{\dagger})^{X'}_{\ell' m'} \rangle_{\rm MC}. \tag{31}$$

This Monte Carlo average converges to C^{-1} because

$$\langle (C^{-1}a)^{X}_{\ell m} (C^{-1}a^{\dagger})^{X'}_{\ell' m'} \rangle = (C^{-1})^{XX'}_{\ell m \ell' m'}, \tag{32}$$

where $(\langle \cdots \rangle)$ denotes the ensemble average over the multivariate $\mathcal{N}(0, C)$ distribution. The Monte Carlo approach allows one to indirectly compute C^{-1} using the iterative methods used for the matrix-vector product $C^{-1}a$. For this reason we will encounter Monte Carlo estimates, denoted by $\langle \cdots \rangle_{\rm MC}$, throughout this paper.

B. Fast bispectrum estimation

In this section we motivate the need for an efficient way to evaluate the estimator in Eq. (26) and review the standard method to do so: the Komatsu, Spergel, and Wandelt (KSW) estimator [77]. When used to estimate the amplitude of primordial 3-point functions, the KSW estimator applies to the $\zeta\zeta\zeta$ correlation but not to our main interest: the $\zeta\zeta h$ correlation. We will introduce the generalized version of the KSW estimator that can be used for $\zeta\zeta h$ in Sec. III C.

The number of numerical operations needed to evaluate the estimator in Eq. (26) quickly grows to enormous sizes as the resolution of the data, i.e., $\ell_{\rm max}$, increases. Even when the costs of computing $C^{-1}a$ are ignored, direct evaluation of the estimator in Eq. (26) will asymptotically scale as $\mathcal{O}(\ell^6_{\rm max})$. The isotropy of the bispectrum may be used to reduce this scaling to $\mathcal{O}(\ell^5_{\rm max})$ by, for instance, fixing $m_3 = -(m_1 + m_2)$, but this scaling is still unmanageable.

To avoid the $\mathcal{O}(\ell^5_{\rm max})$ scaling, bispectrum estimation generally focuses on separable bispectrum templates to reduce the scaling to $\mathcal{O}(\ell^3_{\rm max})$ (albeit possibly with a relatively large prefactor). The most straightforward implementation of this idea is formulated by Komatsu, Spergel, and Wandelt [77], in what we will refer to as the KSW estimator. See Ref. [119] for technical details and Refs. [80,81] for a generalization that uses *E*-mode data in addition to *T* data.

Simply put, the KSW estimator exploits the idea that for a hypothetical bispectrum template

$$B^{\ell_1 \ell_2 \ell_3}_{m_1 m_2 m_3} = F_{\ell_1, m_1} G_{\ell_2, m_2} H_{\ell_3, m_3}, \tag{33}$$

the sum in Eq. (26) can be factored into three independent parts, thereby reducing the scaling to $\mathcal{O}(\ell^2_{\rm max})$. Of course, this hypothetical bispectrum template is not suitable, as it is not rotationally invariant. The decomposition in Eq. (18) forbids isotropic templates that are explicitly factored like this. In reality, the KSW approach therefore uses a slightly modified version of the above decomposition. The numerical advantage is largely maintained with the modified version.

The modification comes in the form of the Gaunt integral expression. It allows the rotationally invariant part of the product of three (spin-weighted) spherical harmonics to be expressed in terms of Wigner 3-j symbols. The general expression can be found in Eq. (B10). Here we only need the following version:

$$\begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix} J^{000}_{\ell_1\ell_2\ell_3} = \int_{S^2} \mathrm{d}\Omega(\hat{\mathbf{n}}) \prod_{i=1}^{3} Y_{\ell_i m_i}(\hat{\mathbf{n}}), \tag{34}$$

with $J^{000}_{\ell_1\ell_2\ell_3}$ given by

$$J^{000}_{\ell_1\ell_2\ell_3} = \sqrt{\frac{(2\ell_1+1)(2\ell_2+1)(2\ell_3+1)}{4\pi}} \times \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ 0 & 0 & 0 \end{pmatrix}. \tag{35}$$

Now consider the reduced bispectrum [82]

$$b_{\ell_1\ell_2\ell_3} \equiv \frac{B_{\ell_1\ell_2\ell_3}}{J^{000}_{\ell_1\ell_2\ell_3}}, \tag{36}$$

where $B_{\ell_1\ell_2\ell_3}$ is the angle-averaged bispectrum. We have suppressed the polarization indices for simplicity. Note that the reduced bispectrum is only defined for $\ell_1+\ell_2+\ell_3 =$ even.[11] By expressing the bispectrum in Eq. (18) in terms of the reduced bispectrum, we may insert the Gaunt integral as follows:

$$B^{\ell_1\ell_2\ell_3}_{m_1m_2m_3} = \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix} J^{000}_{\ell_1\ell_2\ell_3} b_{\ell_1\ell_2\ell_3}, \\ = \int_{S^2} \mathrm{d}\Omega(\hat{\mathbf{n}}) \left(\prod_{i=1}^{3} Y_{\ell_i m_i}(\hat{\mathbf{n}}) \right) b_{\ell_1\ell_2\ell_3}. \tag{37}$$

The crucial insight is that isotropy does not constrain the reduced bispectrum in any way. Given a reduced bispectrum that is separable in N_{fact} sets of functions as

$$b_{\ell_1\ell_2\ell_3} = \frac{1}{6}\sum_{i=1}^{N_{\mathrm{fact}}} f^{(i)}_{\ell_1} g^{(i)}_{\ell_2} h^{(i)}_{\ell_3} + (5\text{ perm}), \tag{38}$$

we may thus express the bispectrum as

$$B^{\ell_1\ell_2\ell_3}_{m_1m_2m_3} = \int_{S^2} \mathrm{d}\Omega(\hat{\mathbf{n}}) \left(\frac{1}{6}\sum_{i=1}^{N_{\mathrm{fact}}} f^{(i)}_{\ell_1} Y_{\ell_1 m_1} \times g^{(i)}_{\ell_2} Y_{\ell_2 m_2} h^{(i)}_{\ell_3} Y_{\ell_3 m_3} + (5\text{ perm}) \right)(\hat{\mathbf{n}}). \tag{39}$$

We will refer to bispectra that can be written as the above expression as “locally separable.” This name refers to the fact that the integrand of the angular integral is separable in (ℓ_1, m_1), (ℓ_2, m_2), and (ℓ_3, m_3). We conclude that, while factored bispectra as in Eq. (33) are forbidden, isotropy allows a locally separable template such as Eq. (39).

It remains to be demonstrated how separable reduced bispectra actually lead to a reduction in computational cost. To see this, we insert Eq. (39) into Eq. (26) and write down the cubic part of the resulting expression:

$$\hat{f}_{\mathrm{NL,cubic}} = \frac{1}{6\mathcal{I}_0} \int_{S^2} \mathrm{d}\Omega(\hat{\mathbf{n}}) \times \left(\sum_{i=1}^{N_{\mathrm{fact}}} \mathcal{A}[f^{(i)}_{\ell}] \mathcal{A}[g^{(i)}_{\ell}] \mathcal{A}[h^{(i)}_{\ell}] \right)(\hat{\mathbf{n}}). \tag{40}$$

The $\mathcal{A}$ functionals yield spin-0 fields on the sphere given by the inverse covariance-weighted data, weighted by the factors of the reduced bispectrum [f_ℓ, g_ℓ, and h_ℓ, see Eq. (38)]. For example,

$$\mathcal{A}[f_{X,\ell}](\hat{\mathbf{n}}) = \sum_{\ell,m}\sum_{X} f_{X,\ell} (C^{-1}a)^X_{\ell m} Y_{\ell m}(\hat{\mathbf{n}}). \tag{41}$$

Note that we have reintroduced the polarization indices and assume they only run over $X \in \{T, E\}$ here. The Monte Carlo expression for the linear term in Eq. (26) becomes equal to

$$\hat{f}_{\mathrm{NL,lin}} = \frac{1}{6\mathcal{I}_0} \int_{S^2} \mathrm{d}\Omega(\hat{\mathbf{n}}) \left(\sum_{i=1}^{N_{\mathrm{fact}}} \mathcal{A}[f^{(i)}_{\ell}] \times \langle \mathcal{A}[g^{(i)}_{\ell}] \mathcal{A}[h^{(i)}_{\ell}] \rangle_{\mathrm{MC}} + \text{cyclic} \right)(\hat{\mathbf{n}}). \tag{42}$$

The two additional terms denoted by “cyclic” are obtained by cyclic permutations of $f^{(i)}_\ell$, $g^{(i)}_\ell$, and $h^{(i)}_\ell$.

Evaluating Eq. (40) does not quite scale as $\mathcal{O}(\ell^2_{\max})$ as one might expect but as $\mathcal{O}(N_{\mathrm{fact}}\ell^3_{\max})$. Simply put, the scaling is determined by the $\mathcal{O}(\ell^3_{\max})$ scaling of the recursive algorithms needed to compute the spherical harmonics that have to be recomputed N_{fact} times.[12]

[11] Restricting to $\ell_1+\ell_2+\ell_3 =$ even does not introduce a loss of generality for the parity-invariant $\langle TTT\rangle$, $\langle TTE\rangle$, $\langle TEE\rangle$, and $\langle EEE\rangle$ angle-averaged bispectra that are usually considered (see Table II), but, as was shown in Sec. III A, angle-averaged bispectra can in general be nonzero for $\ell_1+\ell_2+\ell_3 =$ odd.

[12] It should be noted that Ref. [119] describes an alternative, significantly more efficient $\mathcal{O}(N_{\mathrm{fact}}\ell^3_{\max})$ algorithm for Eq. (40) that only runs the expensive $Y_{\ell m}$ recursion once.

This is still a significant improvement over the general $\mathcal{O}(\ell_{\rm max}^5)$ scaling. Evaluation of the linear term scales as $\mathcal{O}(N_{\rm sim}N_{\rm fact}\ell_{\rm max}^3)$, where $100 \lesssim N_{\rm sim} \lesssim 1000$ iterations are typically needed for a sufficiently accurate estimate [119].

The estimator normalization $\mathcal{I}_0$ in Eq. (30) is evaluated by a Monte Carlo estimate. We omit the details of this aspect of the estimation procedure and mention only the two methods that are used in practical applications. The most straightforward estimate of $\mathcal{I}_0$ is given by the variance of the unnormalized estimator applied to an ensemble of simulated Gaussian data effectively drawn from the distribution specified by Eq. (28). A similar but slightly more involved Monte Carlo procedure is described in [119]. This second method is shown to converge for smaller ensembles than the first method.

Up to now, this general discussion has not specified the origin of the bispectrum; the fast estimation technique applies to all bispectra that can be described by Eq. (38). With regards to primordial $\zeta\zeta\zeta$ non-Gaussianity, the above construction is useful only when theoretical bispectrum templates can be reduced to the form of Eq. (38). Fortunately, this is the case for a large class of linearly propagated bispectra sourced by the $\zeta\zeta\zeta$ correlation. For such bispectra, the condition in Eq. (38) is met when the shape of the 3-point function in Eq. (23) is separable in k:

$$f^{(\zeta\zeta\zeta)}(k_1,k_2,k_3) = \frac{1}{6}\sum_{i=1}^{N_{\rm prim}} f^{(i)}(k_1)g^{(i)}(k_2)h^{(i)}(k_3) + (5\ \text{perm}). \tag{43}$$

The local shape in Eq. (C7) is an example of a separable shape template. The equilateral and orthogonal shape templates used in the *Planck* analysis [15] have been specifically derived to be separable [78,120].

The KSW estimator is expressed slightly differently for primordial $\zeta\zeta\zeta$ 3-point functions than Eq. (40), but the difference is notational. The cubic estimator corresponding to a 3-point function described by Eq. (43) is expressed as follows:

$$\hat{f}^{\zeta\zeta\zeta}_{\rm NL,cubic} = \frac{1}{6\mathcal{I}_0}\int_{S^2} {\rm d}\Omega(\hat{\mathbf{n}}) \sum_{i=1}^{N_{\rm prim}} \int_0^\infty r^2{\rm d}r \times (\mathcal{A}^{(\zeta)}_{(0,0)}[f^{(i)}]\mathcal{A}^{(\zeta)}_{(0,0)}[g^{(i)}]\mathcal{A}^{(\zeta)}_{(0,0)}[h^{(i)}])(r,\hat{\mathbf{n}}). \tag{44}$$

This expression and related ones will be described in more detail in Sec. III C and Appendix A. Here, we show that the above expression conforms to the general case in Eq. (40). The main difference between the two expressions is the appearance of the integral over comoving distance r in Eq. (44). Without going into details at this point, we simply note that replacing the integral by a finite number of quadrature points N_r allows the integral and sum over $i \in \{1,\ldots,N_{\rm prim}\}$ to be replaced by a single summation over $i \in \{1,\ldots,N_{\rm fact}\}$ with $N_{\rm fact} = N_r N_{\rm prim}$. This already brings Eq. (44) closer to Eq. (40). The second difference is that the $\mathcal{A}^{(\zeta)}$ functionals in Eq. (44) directly take the shape functions $f(k)$, etc., as their argument while the $\mathcal{A}$ in Eq. (40) take the reduced bispectrum factors f_ℓ as their argument. This difference can be understood as follows: $\mathcal{A}^{(\zeta)}$, just as $\mathcal{A}$, filters the inverse-covariance weighed data by f_ℓ [see Eq. (41)], but first transforms $f(k)$ to f_ℓ using the radiative transfer functions introduced in Eq. (16). In the end, applying $\mathcal{A}^{(\zeta)}$ to a function $f(k)$ yields a scalar field on the sphere, just as applying $\mathcal{A}$ to f_ℓ does. See Appendix A for the precise definition of $\mathcal{A}^{(\zeta)}$.

In summary, a primordial $\zeta\zeta\zeta$ 3-point correlation function described by a separable shape function will source a separable reduced bispectrum. We have established that the separability of the reduced bispectrum allows the use of the KSW estimator [see Eq. (40)]. Finally, the KSW estimator is a prescription that alleviates the scaling of the estimator in Eq. (26) from $\mathcal{O}(\ell_{\rm max}^5)$ to a more manageable $\mathcal{O}(N_{\rm fact}\ell_{\rm max}^3)$.

C. Fast scalar-scalar-tensor bispectrum estimation

1. Overview

We now turn to the situation for the $\zeta\zeta h$ 3-point correlation function. We explain why the standard KSW estimator, derived in Sec. III B, does not apply to this type of correlation. We then come to the main new result of this paper: we introduce an alternative approach that allows the construction of an efficient estimator for the $\zeta\zeta h$ correlation.

Recall that for the $\zeta\zeta\zeta$ correlation the necessary condition for a separable reduced bispectrum is given by Eq. (43): a separable shape function. Unlike the $\zeta\zeta\zeta$ 3-point correlation function, the $\zeta\zeta h$ correlation is not uniquely specified by a shape function. It turns out that when the reduced bispectrum for the $\zeta\zeta h$ template in Eq. (24) is computed, the result is nonseparable in ℓ_1, ℓ_2, and ℓ_3 [59]. This holds true even when the $f^{(\zeta\zeta h)}$ shape function in Eq. (24) is separable in k_1, k_2, and k_3, which means that the responsible piece is the angular term

$$\langle \zeta_{\mathbf{k}_1}\zeta_{\mathbf{k}_2}{}^{(\pm 2)}h_{\mathbf{k}_3}\rangle \propto (\hat{k}_1)^a(\hat{k}_2)^b e^{\pm 2}_{ab}(\hat{\mathbf{k}}_3). \tag{45}$$

Despite the angular dependence, this term is a scalar under spatial coordinate transformations. The term provides a weight and complex phase to each $\{\hat{\mathbf{k}}_1,\hat{\mathbf{k}}_2\}$ configuration relative to the wave vector of the tensor perturbation but has no preference for a global orientation of the three wave vectors. The associated CMB bispectrum is therefore isotropic and has a trivial dependence on its m_1, m_2, and m_3 azimuthal numbers, given by Eq. (18). With the azimuthal numbers constrained by isotropy, the geometrical coupling between the wave vectors in Eq. (45) can then only manifest itself in an explicit coupling between the ℓ_1, ℓ_2, and ℓ_3 multipole orders, which in turn prevents the reduced bispectrum to be separable.

Without a separable reduced bispectrum we cannot construct the KSW estimator for the $\zeta\zeta h$ template by simply inserting the factors of the reduced bispectrum into Eq. (40). To derive a generalized KSW estimator for this template, let us observe that each term in the sum over spatial indices in Eq. (45) is factored in the three wave vectors. Of course, unlike the summed expression, the individual terms are not 3-scalars; the decomposition is coordinate dependent. By itself, each term can be interpreted as a homogeneous but anisotropic 3-point function. Homogeneity is still preserved by the overall delta function in Eq. (22). The 3-point functions of this form result in anisotropic bispectra[13] that are locally separable in the sense of Eq. (39). The anisotropic expressions differ from the isotropic one in Eq. (39) by the f_ℓ, g_ℓ, and h_ℓ factors; they gain a dependence on m in addition to ℓ.

Roughly speaking, we thus exchange isotropy for separability. The estimates of the amplitudes of the anisotropic terms combine into an estimate of the amplitude of the original isotropic template. The trade-off is that several anisotropic templates have to be considered for one isotropic template. Constructing analogues of the cubic and linear estimator terms in Eqs. (40) and (42) for an anisotropic template will turn out to be rather straightforward. The generalizations of the $\mathcal{A}$ functionals in Eq. (41) will transform the data in an anisotropic manner, but note that this operation does not scale differently than the regular isotropic transformation. The overall scaling of the estimator with $\ell_{\max}$ will thus be unchanged. The number of anisotropic terms needed for 3-point functions of the type in Eq. (45) turns out to be only five. The amount of extra computations compared to the $\zeta\zeta\zeta$ estimator is thus rather insignificant.

Guided by the rough arguments provided in this section, we now turn to the actual derivation of the proposed estimator. We will first derive the expression for the linearly propagated bispectrum for the $\zeta\zeta h$ 3-point function and demonstrate how it is indeed given by a sum of anisotropic bispectra. We will then construct the actual estimator.

2. Full bispectrum for the scalar-scalar-tensor template

In this section, we derive the linearly propagated bispectrum for the $\zeta\zeta h$ 3-point correlation function. As mentioned in Sec. C 1, we require an expression for the full bispectrum instead of the angle-averaged or reduced bispectrum.

The general expression for the linearly propagated bispectrum in Eq. (21) is most easily evaluated by separating the integrals over the three wave vectors in angular and radial integrals. In order to do so we need to rewrite the delta function that imposes momentum conservation in Eq. (22). Additionally, we express all angular terms of the 3-point function as spin-weighed spherical harmonics in order to simplify the angular integrals.

We start with the delta function. We make use of the plane wave expansion in terms of spherical harmonics and spherical Bessel functions:

$$e^{i\mathbf{k}\cdot\mathbf{x}} = 4\pi\sum_{L,M} i^L j_L(kr) Y^*_{LM}(\hat{\mathbf{k}}) Y_{LM}(\hat{\mathbf{n}}), \tag{46}$$

with $\mathbf{k} = k\hat{\mathbf{k}}$ and $\mathbf{x} = r\hat{\mathbf{n}}$. The unit vector $\hat{\mathbf{n}}$ represents the direction of the line of sight from the origin of the comoving coordinate system (our location). Using this expansion we decompose the delta function into radial and angular parts:

$$\begin{aligned}\delta^{(3)}(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3) &= 8\sum_{L_1,M_1}\sum_{L_2,M_2}\sum_{L_3,M_3}\int_{S^2} \mathrm{d}\Omega(\hat{\mathbf{n}})\left(\prod_{i=1}^{3} Y_{L_i,M_i}(\hat{\mathbf{n}})\right)\\ &\quad\times\int_0^\infty r^2 \mathrm{d}r\left(\prod_{i=1}^{3} i^{L_i} j_{L_i}(k_i r) Y^*_{L_i M_i}(\hat{\mathbf{k}}_i)\right). \end{aligned}\tag{47}$$

See Appendix B 3 for details. Although the integral over $\hat{\mathbf{n}}$ is given by the Gaunt integral expression in Eq. (B10), it will turn out to be important to leave the expression factorizable in L_1, L_2, and L_3 so we do not solve the angular integral.

We then move on to the angular part of the $\zeta\zeta h$ template in Eq. (24). As discussed, this part is already expressed as a sum of factorized terms, so we leave it in its uncontracted form. However, we express the unit vectors and polarization tensor in terms of spherical harmonics. In a general coordinate system, not necessarily aligned with $\mathbf{k}_1$, $\mathbf{k}_2$, or $\mathbf{k}_3$, the two unit vectors in Eq. (24) are decomposed into dipole ($\ell = 1$) moments with a longitudinal ($m = 0$) and two solenoidal ($m = \pm 1$) modes, while the polarization tensor is decomposed into quadrupole ($\ell = 2$) moments with longitudinal ($m = 0$), solenoidal ($m = \pm 1$), and transverse ($m = \pm 2$) modes. To retain the correct transformation properties, the quadrupole moment is expressed in terms of spin-±2 spherical harmonics on the plane perpendicular to $\hat{\mathbf{k}}_3$. As the 45 resulting combinations have to sum to a 3-scalar, each combination has to be weighted by the appropriate Wigner 3-j symbol. The resulting expression is given by [83]

$$\begin{aligned}(\hat{k}_1)^a(\hat{k}_2)^b e^{\pm 2}_{ab}(\hat{\mathbf{k}}_3) &= \frac{(8\pi)^{3/2}}{6}\sum_{m_a,m_b,M}\begin{pmatrix}1 & 1 & 2\\ m_a & m_b & M\end{pmatrix}\\ &\quad\times Y^*_{1m_a}(\hat{\mathbf{k}}_1)Y^*_{1m_b}(\hat{\mathbf{k}}_2)\,{}_{\mp 2}Y^*_{2M}(\hat{\mathbf{k}}_3).\end{aligned}\tag{48}$$

[13] The bispectrum is isotropic by definition so an anisotropic bispectrum should be understood as a shorthand for a harmonic 3-point function that does not obey Eq. (18).

The selection rules of the 3-j symbol limit the azimuthal modes to only nine combinations: those that obey $m_a + m_b + M = 0$.

We may now use Eqs. (47) and (48) to decompose the $\zeta\zeta h$ 3-point function in radial and angular parts, resulting in the following expression:

$$
{}^{(00\pm2)}B(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3) = (2\pi)^3\frac{(8\pi)^{3/2}}{6}8\sum_{m_a,m_b,M}\begin{pmatrix}1 & 1 & 2\\ m_a & m_b & M\end{pmatrix}Y^*_{1m_a}(\hat{\mathbf{k}}_1)Y^*_{1m_b}(\hat{\mathbf{k}}_2)\,{}_{\mp2}Y^*_{2M}(\hat{\mathbf{k}}_3)
\times\int_0^\infty r^2\mathrm{d}r\int_{S^2}\mathrm{d}\Omega(\hat{\mathbf{n}})\left(\prod_{i=1}^{3}\sum_{L_i,M_i}(-1)^{L_i/2}j_{L_i}(k_ir)Y^*_{L_i,M_i}(\hat{\mathbf{k}}_i)Y_{L_iM_i}(\hat{\mathbf{n}})\right)f^{(\zeta\zeta h)}(k_1,k_2,k_3). \tag{49}
$$

As is required for the KSW estimator, we assume that the shape function $f^{(\zeta\zeta h)}$ is separable; i.e., it obeys

$$
f^{(\zeta\zeta h)}(k_1,k_2,k_3) = \frac{1}{6}\sum_{i=1}^{N_{\mathrm{prim}}} f^{(i)}(k_1)g^{(i)}(k_2)h^{(i)}(k_3) + (5\ \mathrm{perm}). \tag{50}
$$

The N_{prim} sets of f, g, and h functions depend on the model under investigation so we leave them unspecified.

We have now gathered all ingredients to form the linearly propagated CMB bispectrum for the $\zeta\zeta h$ 3-point correlation function. We do so by combining Eqs. (49) and (50) and inserting the result into Eq. (21). Because we have separated the 3-point function in radial and angular parts, the expression neatly factors into six independent integrals. We evaluate the angular integrals using the generalized Gaunt integral relation in Eq. (B10). The resulting contribution to the CMB bispectrum is then as follows:

$$
\begin{aligned}
B^{\ell_1\ell_2\ell_3(\zeta\zeta h)}_{m_1m_2m_3X_1X_2X_3} &= \frac{(8\pi)^{3/2}}{36}\sum_{m_a,m_b,M}\begin{pmatrix}1 & 1 & 2\\ m_a & m_b & M\end{pmatrix}\int_{S^2}\mathrm{d}\Omega(\hat{\mathbf{n}})\sum_{i=1}^{N_{\mathrm{prim}}}\int_0^\infty r^2\mathrm{d}r\\
&\times\sum_{L_1,M_1}\left[i^{\ell_1+L_1}J^{000}_{1L_1\ell_1}\begin{pmatrix}1 & L_1 & \ell_1\\ m_a & M_1 & m_1\end{pmatrix}(\mathcal{K}^{(\zeta)}_{(X_1)}[f^{(i)}])_{\ell_1,L_1}(r)\right]Y_{L_1M_1}(\hat{\mathbf{n}})\\
&\times\sum_{L_2,M_2}\left[i^{\ell_2+L_2}J^{000}_{1L_2\ell_2}\begin{pmatrix}1 & L_2 & \ell_2\\ m_b & M_2 & m_2\end{pmatrix}(\mathcal{K}^{(\zeta)}_{(X_2)}[g^{(i)}])_{\ell_2,L_2}(r)\right]Y_{L_2M_2}(\hat{\mathbf{n}})\\
&\times\sum_{L_3,M_3}\left[i^{\ell_3+L_3}J^{-202}_{2L_3\ell_3}[1+(-1)^{x_3+L_3+\ell_3}]\begin{pmatrix}2 & L_3 & \ell_3\\ M & M_3 & m_3\end{pmatrix}(\mathcal{K}^{(h)}_{(X_3)}[h^{(i)}])_{\ell_3,L_3}(r)\right]Y_{L_3M_3}(\hat{\mathbf{n}})\\
&+(5\ \mathrm{perm}).
\end{aligned} \tag{51}
$$

Here we have, as a shorthand, defined the following set of functionals for all $Z\in\{\zeta,h\}$, $X\in\{T,E,B\}$:

$$
(\mathcal{K}^{(Z)}_{(X)}[f])_{\ell,L} \equiv \frac{2}{\pi}\int_0^\infty k^2\mathrm{d}k f(k)\mathcal{T}^{(Z)}_{X,\ell}(k)j_L(kr). \tag{52}
$$

The $\mathcal{T}_\ell(k)$ transfer functions were introduced in Eq. (16). To evaluate the sum over the tensor helicities we have made use of the following relation:

$$
\sum_{\lambda_3\in\pm2}\mathrm{sgn}(\lambda_3)^{\lambda_3+x_3}J^{-\lambda_30\lambda_3}_{2L_3\ell_3} = J^{-202}_{2L_3\ell_3}[1+(-1)^{x_3+L_3+\ell_3}],
$$

which reflects that the $f^{(\zeta\zeta h)}$ shape function in Eq. (49) is helicity independent. Recall that $x_3\in\{0,1\}$ indicates whether the X_3 CMB field is parity even or parity odd. The J symbols are defined in Eq. (B9).

The expression for the bispectrum in Eq. (51) is a bit verbose, but this expanded form will make it easier to construct the estimator in Sec. III C 3. The expression shows how the bispectrum can be separated into factors that only depend on ℓ_1, ℓ_2, or ℓ_3. Of course, the expression, taken as a whole, ought to be isotropic. This may be checked by summing over all azimuthal dummy indices (m_a, m_b, M, M_1, M_2, M_3).[14] As expected, the

[14] First express the angular integral over $Y_{L_1M_1}$, $Y_{L_2M_2}$, and $Y_{L_3M_3}$ in terms of the Gaunt integral and then sum over the five 3-j symbols that depend on azimuthal numbers using Eq. (B17) [83].

resulting expression reduces to the isotropic form in Eq. (18) but yields a nonseparable angle-averaged/reduced bispectrum.

Each term in the sum over m_a, m_b, and M in Eq. (51) describes an anisotropic bispectrum. Each of these bispectra is "locally" separable in the sense of Eq. (39). The integral over the comoving radial coordinate r in Eq. (51) may be replaced with a weighted sum over N_{quad} integration points. Combined with the N_{prim} terms in the primordial shape function there will then be $N_{\mathrm{fact}} = N_{\mathrm{prim}} N_{\mathrm{quad}}$ locally separable terms.

The allowed combinations of L_1, L_2, and L_3 per (ℓ_1, ℓ_2, ℓ_3) triplet in Eq. (51) are quite limited; depending on the polarization indices of the bispectrum only 8 or 12 combinations are allowed [83]. Recall that the capital L's arise from the expansion of the delta function in Eq. (47). The specific values can be found by systematically going over the 3-j symbols, including the ones hidden in the J symbols [see Eq. (B9)]. First, note that $J^{000}_{1L_1\ell_1}$ and $J^{000}_{1L_2\ell_2}$ require $L_1 + \ell_1$ and $L_2 + \ell_2$ to be odd. The triangle conditions of the 3-j symbols in the second and third lines then enforce $L_1 = |\ell_1 \pm 1|$ and $L_2 = |\ell_2 \pm 1|$. The term in square brackets in the fourth line forces $L_3 + \ell_3$ to be even when $x_3 = 0$ or odd when $x_3 = 1$. The triangle condition of the 3-j symbol in the fourth line then requires $L_3 = \{\ell_3, |\ell_3 \pm 2|\}$ for $x_3 = 0$ and $L_3 = \{|\ell_3 \pm 1|\}$ for $x_3 = 1$. Finally, when the angular integral over $Y_{L_1 M_1}$, $Y_{L_2 M_2}$, and $Y_{L_3 M_3}$ is performed using Eq. (B10), it becomes clear how $L_1 + L_2 + L_3 =$ even is (again) imposed as well as $|L_1 - L_2| \leq L_3 \leq L_1 + L_2$.

We have derived the linearly propagated bispectrum for the $\zeta\zeta h$ 3-point correlation function: a crucial ingredient for the derivation of the estimator. The resulting bispectrum is given in Eq. (51). We have showed that the bispectrum can be viewed as a sum of anisotropic bispectra. As a sanity check of the derivation one may verify that the bispectrum holds up to the general constraints due to parity invariance that were formulated in Sec. II B 1. For polarization triplets X_1, X_2, X_3 with even parity, i.e., $X_3 \neq B$, the bispectrum is real and nonzero when $\ell_1 + \ell_2 + \ell_3 =$ even. On the other hand, when $X_3 = B$ (so $x_3 = 1$), the bispectrum becomes purely imaginary and nonzero only for $\ell_1 + \ell_2 + \ell_3 =$ odd.

3. $\mathcal{K}$ functionals

Before constructing the estimator it is instructive to take a more detailed look at the $\mathcal{K}_{\ell,L}$ functionals defined in Eq. (52). They will become an important part of the estimator. We thus have a brief digression in which we illustrate the role of the functionals in Eq. (51). Readers who are more interested in the actual estimator may skip this section.

The $\mathcal{K}$'s are a straightforward generalization of the $\alpha_\ell(r)$ and $\beta_\ell(r)$ functions introduced in the KSW description for the local model [77].[15] In the original KSW description the $\mathcal{K}$'s serve to transform the factors of the 3-point function to the factors of the reduced bispectrum, i.e., $f(k) \mapsto \mathcal{K}[f] = f_\ell$. For the $\zeta\zeta h$ estimator, the $\mathcal{K}$'s still serve to transform factors of the 3-point function into factors of the bispectrum. The difference is that, as can be seen in Eq. (51), the factors of the 3-point function now each require multiple transformations to account for their nonscalar nature.

Let us first focus on the $\mathcal{K}$ functionals that are relevant for regular $\zeta\zeta\zeta$ non-Gaussianity estimation: the $\mathcal{K}$'s with $L = \ell$ and transfer functions for $Z = \zeta$. It is convenient to consider a constant input function $f(k) = 1$, and the resulting functions are equal to the $\alpha^X_\ell(r)$ functions defined in Ref. [80],

$$\begin{aligned}(\mathcal{K}^{(\zeta)}_{(X)}[1])_{\ell,\ell}(r) &= \alpha^X_\ell(r) \\ &= \frac{2}{\pi}\int_0^\infty k^2 \mathrm{d}k \mathcal{T}^{(\zeta)}_{X,\ell}(k) j_\ell(kr), \end{aligned} \tag{53}$$

where $X \in \{T, E\}$ because of the ζ transfer function. The $\alpha^X_\ell(r)$ functions have a special interpretation: they serve as the transfer functions in coordinate space instead of Fourier space. Equation (53) is an inverse Fourier transform (i.e., inverse spherical Hankel transform) of the transfer function $\mathcal{T}_\ell(k)$, and it is true that the observable CMB harmonic modes sourced by ζ may be expressed as follows [79]:

$$a^{(\zeta)}_{X,\ell m} = \int_0^\infty r^2 \mathrm{d}r \zeta_{\ell m}(r) \alpha^X_\ell(r), \tag{54}$$

for $X \in \{T, E\}$. Here $\zeta_{\ell m}(r)$ are the spherical harmonic coefficients of the same initial amplitude of the curvature perturbation as in Eq. (10) but now decomposed on spherical shells around the origin of the comoving coordinate system:

$$\zeta_{\ell m}(r) = \int_{S^2} \mathrm{d}\Omega(\hat{\mathbf{n}}) \zeta(\mathbf{x}, t)|_{t=\tilde{t}(\mathbf{x}, t_i)} Y^*_{\ell m}(\hat{\mathbf{n}}). \tag{55}$$

Recall that $\tilde{t}(\mathbf{x}, t_i)$ denotes a spacelike hypersurface in the early radiation-dominated era.

The solid lines in Fig. 1 show $\alpha^X_\ell(r)$ for $X = T$ and $X = E$ as a function of the comoving radius on the initial spatial hypersurface. The lines show how $\zeta_{\ell m}(r)$ contributes to $a_{X,\ell m}$ for $\ell = 60$ over a range of comoving radii around 14000 Mpc. In terms of the conformal time along

[15] The functions $\alpha_\ell(r)$ and $\beta_\ell(r)$ from Ref. [77] are given by $(\mathcal{K}^{(\zeta)}_{(T)}[1])_{\ell,\ell}$ and $(\mathcal{K}^{(\zeta)}_{(T)}[P_\Phi])_{\ell,\ell}$ respectively. P_Φ is the power spectrum of the gauge-invariant Φ_H Bardeen potential [100] instead of the curvature perturbation ζ we use; the two gauge-invariant quantities are related as $\zeta = -3\Phi_H/2$ and $\zeta = -5\Phi_H/3$ for superhorizon adiabatic perturbations in the radiation and matter dominated eras respectively [99].

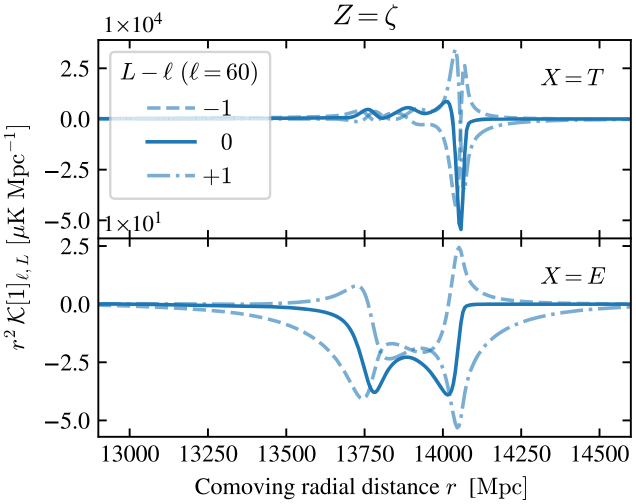

FIG. 1. Radial transfer functions (solid lines) demonstrating the response of the $\ell = 60$ temperature (top) and E-mode (bottom) CMB anisotropies to the curvature perturbation at comoving radial distance r. The response shown here corresponds to the epoch of recombination. The dashed and dot-dashed lines show the radial parts of the functions used to project a dipole moment constructed from the curvature perturbation to the ($\ell = 60$) CMB harmonic modes. For low multipole orders, such as the one depicted here, these functions are significantly less localized in r than the transfer function and thus require a wider range of integration points.

the path of a radially traveling photon ($\Delta\tau = r/c$), this range of r is roughly centered around the epoch of recombination. Another response at $r \approx 9000$ Mpc corresponds to the rescattering of CMB photons at reionization. Finally, at $r \lesssim 3000$ Mpc there is a slowly rising response as r approaches zero for $X = T$ and $\ell \lesssim 150$ that corresponds to the late-time integrated Sachs-Wolfe (ISW) effect.

The fact that $\mathcal{K}^{(\zeta)}[1]$ yields the radial transfer functions provides a physical reason why the $\mathcal{K}$ functionals result in functions that are highly localized in r. During bispectrum estimation the integral over r has to be evaluated as efficiently as possible; the localized nature of the radial functions is thus highly beneficial. We will now see how and why the radial functions used for the $\zeta\zeta h$ bispectrum differ from the ones used for regular scalar-sourced bispectra. These new functions will turn out to be slightly less localized in r, but the difference is minor.

Equation (54) must hold because the harmonic modes of the curvature perturbation on spherical shells $\zeta_{\ell m}(r)$ in Eq. (55) are related to the harmonic modes of the Fourier representation of ζ through the following simple relation:

$$\zeta_{\ell m}(k) = 4\pi(-i)^{\ell} \int_0^\infty r^2 \mathrm{d}r \zeta_{\ell m}(r) j_\ell(kr). \tag{56}$$

Here the $\zeta_{\ell m}(k)$ are the coefficients of the spherical harmonic decomposition of the angular part of $\zeta_{\mathbf{k}}$ in Eq. (10):

$$\zeta_{\mathbf{k}} = \sum_{\ell,m} \zeta_{\ell m}(k) Y_{\ell m}(\hat{\mathbf{k}}). \tag{57}$$

One can check that Eq. (54) holds by inserting Eqs. (57) and (56) into Eq. (16) and making use of the orthonormality of the spherical harmonics.

In turn, Eq. (56) is valid because ζ is a 3-scalar, and it has no intrinsic angular dependence. The projection from the Fourier basis to a basis of spherical shells at comoving radii r is thus completely determined by the "orbital" angular momentum of the field; i.e., the projection is determined by the plane wave decomposition of the 3D Fourier basis functions in Eq. (46). Simply put: projecting a Fourier mode of a 3-scalar to an angular mode with multipole order ℓ and azimuthal mode m sitting on a shell at radius r only requires transformations involving j_ℓ and $Y_{\ell m}$. Inserting Eq. (56) into Eq. (57) demonstrates this behavior.

For fields that are not 3-scalars, a relation such as Eq. (56) will not hold. In these cases, the coupling between the intrinsic angular dependence of the field and that of the plane wave contributes to the projection. The exact expressions for these "total angular momentum" projection operators may be found in Refs. [121–123]. We will use the general properties of these operators to gain a better understanding of the role of the second multipole index of the $\mathcal{K}_{\ell,L}$ functionals.

In the above we argued that the projection of a single Fourier mode, i.e., a plane wave, to an angular mode with multipole order ℓ and azimuthal mode m sitting on a shell at radius r will only involve j_ℓ and $Y_{\ell m}$. The same projection for an intrinsically dipolelike ($\ell' = 1$) field that is modulated by a plane wave will involve operators constructed out of $j_{\ell\pm1}$ and $Y_{\ell\pm1m}$. Two distinct projections exist in this case: one for the longitudinal ($m' = 0$) component of the dipolelike field and one for the solenoidal ($m' = \pm1$) components [121]. Similarly, the projection of an intrinsically quadrupolelike ($\ell' = 2$) field modulated by a plane wave will involve j_ℓ, $Y_{\ell m}$; $j_{\ell\pm1}$, $Y_{\ell\pm1m}$; and $j_{\ell\pm2}$, $Y_{\ell\pm2m}$. Again, there are distinct projections for the longitudinal ($m' = 0$), solenoidal ($m' = \pm1$), and transverse ($m' = \pm2$) components of the field. This time, a projection using ℓ and $\ell \pm 2$ only contributes to the parity-even component of the resulting field; the $\ell \pm 1$ projections only contribute to the parity-odd component [121].

Having gained this intuition, it is now understood why only the terms with $L_1 = |\ell_1 \pm 1|$ and $L_2 = |\ell_2 \pm 1|$ contribute in the second and third lines of Eq. (51), respectively. Each of the two lines describes how a dipole moment constructed out of one of the two unit wave vectors in the 3-point function template in Eq. (45) is projected to a set of angular modes on spherical shells at radius r. The prefactor given to $\mathcal{K}_{\ell,L} Y_{LM}$ in the second and third lines of Eq. (51) will change depending on whether the longitudinal mode (e.g., $m_a = 0$) or the solenoidal modes (e.g., $m_a = \pm1$) are projected.

The $\mathcal{K}$ functionals with $L = \ell \pm 1$ differ substantially from the $L = \ell$ variants used in the $\zeta\zeta\zeta$ KSW estimator. This is especially true for low ($\ell \lesssim 500$) multipole orders. We plot the $\mathcal{K}[1]_{\ell,\ell\pm1}$ functions next to the regular radial transfer functions in Fig. 1 to illustrate this. Note that for $\ell \gtrsim 500$ the functions with $L = \ell \pm 1$ converge to the shape of those with $L = \ell$ although there remains a small phase shift in r regardless of ℓ.

In a similar way, the fourth line of Eq. (51) describes the projection of the quadrupole moment constructed out of the polarization tensor in Eq. (45). As we established before, the $L = \ell \pm 1$ components are needed for the B-mode field while the $L = \ell, \ell \pm 2$ components are used for the parity-even T and E fields. The prefactor of $\mathcal{K}_{\ell_3,L_3} Y_{L_3 M_3}$ is now dependent on M, which denotes whether the longitudinal ($M = 0$), solenoidal ($M = \pm1$), or transverse ($M = \pm2$) components of the quadrupole are taken into account. To illustrate how the $\mathcal{K}_{\ell,L}$ functionals change when the $Z = h$ transfer functions are used instead of the $Z = \zeta$ transfer functions we considered before, we plot $\mathcal{K}_{\ell,L}[1]$ for $Z = h$ and $L = \ell, \ell \pm 2$ in Fig. 2. The plotted range again roughly corresponds to the recombination era. Not shown is another small response that corresponds to the reionization era. There is no equivalent for the late-time ISW effect. In Fig. 3 we plot the same functions but for $L = \{\ell \pm 1\}$. These functions are used to project the quadrupole moment of the 3-point function to the CMB B-mode field.

The small aliasing effects seen in Figs. 2 and 3 are purely numerical; both j_L and $\mathcal{T}_\ell$ in Eq. (52) oscillate rapidly with k. The integral thus requires a large number of k integration points to completely converge for each value of r. It should be noted that the integral is a candidate for the FFTLog algorithm described in Appendix B of Ref. [124]. This algorithm, which uses the fast Fourier transform to speed up Hankel transforms such as Eq (52), would significantly lower the evaluation cost of the integral and would likely increase the accuracy of the result. The increased speed would be particularly useful for analyses that recompute the transfer functions often in order to marginalize over uncertainties in the ΛCDM parameters. We have not used the FFTLog algorithm in this work, but we have verified that the bispectrum and the results in Sec. IV are not sensitive to the numerical artifacts seen in the figures.

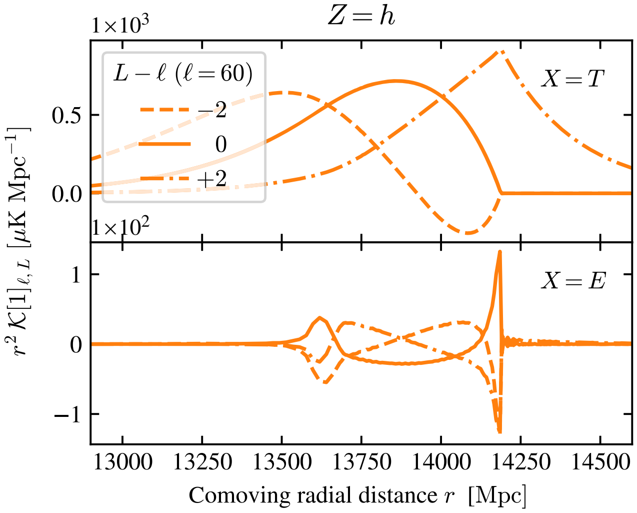

FIG. 2. The three radial functions needed to compute the response of the $\ell = 60$ temperature (top) and E-mode (bottom) CMB anisotropies to a quadrupole moment constructed from the 3-tensor metric perturbation at comoving radial distance r.

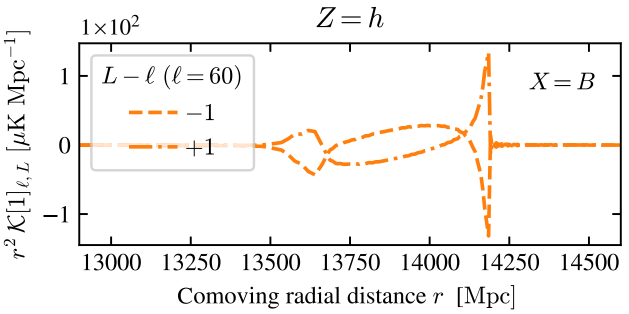

FIG. 3. Similar to Fig. 2 but instead showing the two radial functions needed to compute the response of the $\ell = 60$ B-mode CMB anisotropies to a quadrupole moment constructed from the 3-tensor metric perturbation at comoving radial distance r.

The point of this section was to explain the role of the $\mathcal{K}_{\ell,L}$ functionals present in the $\zeta\zeta h$ bispectrum in Eq. (51). As illustrated in the figures, the functionals with $\ell \neq L$, i.e., the ones needed for the $\zeta\zeta h$ bispectrum, differ substantially from the $\ell = L$ functionals that are used for the standard $\zeta\zeta\zeta$ bispectrum.

4. Estimator

Using Eq. (51), the expression for the $\zeta\zeta h$ bispectrum, we now write down the estimator for the amplitude of this bispectrum template. For simplicity we start by neglecting the linear term in the estimator in Eq. (26) and focus on the cubic term.

The expression for the bispectrum in Eq. (51) is sourced by the $\zeta\zeta h$ template. The order matters, the observed CMB bispectrum is also sourced by the 3-point functions with permuted ζ and h indices. However, it will be convenient to keep ignoring the $\zeta h\zeta$ and $h\zeta\zeta$ contributions for now and start by constructing the estimator for the $\zeta\zeta h$ template only. We thus divide the (cubic part of) the estimator in three parts,

$$\hat{f}^{\mathrm{tot}}_{\mathrm{NL,cubic}} = \hat{f}^{\zeta\zeta h}_{\mathrm{NL,cubic}} + \hat{f}^{\zeta h\zeta}_{\mathrm{NL,cubic}} + \hat{f}^{h\zeta\zeta}_{\mathrm{NL,cubic}}, \tag{58}$$

and start with the first term on the rhs.

We reap the benefits of our work in the previous sections; the cubic estimator is simply constructed by inserting the expression for the bispectrum, Eq. (51), into the general expression for the estimator in Eq. (26) and keeping the terms cubic in the data. Let us again stress that this result is only achieved through the use of the full bispectrum as opposed to the angle-averaged or reduced bispectrum. The resulting expression for the cubic part of the estimator, and the main result of this paper, is given by

$$\hat{f}^{\zeta\zeta h}_{\mathrm{NL,cubic}} = \frac{\sqrt{2}}{54\mathcal{I}_0}\sum_{\substack{m_a,m_b,\\ M}}\begin{pmatrix} 1 & 1 & 2 \\ m_a & m_b & M \end{pmatrix}\int_{S^2}\mathrm{d}\Omega(\hat{\mathbf{n}}) \times \sum_{i=1}^{N_{\mathrm{prim}}}\int_0^\infty r^2\mathrm{d}r(\mathcal{A}^{(\zeta)}_{(1,m_a)}[f^{(i)}]\mathcal{A}^{(\zeta)}_{(1,m_b)}[g^{(i)}] \times \mathcal{A}^{(h)}_{(2,M)}[h^{(i)}])(r,\hat{\mathbf{n}}) + (2\ \mathrm{cyclic}). \tag{59}$$

We have again made use of the shape template in Eq. (50). The two extra terms are cyclic permutations of $f^{(i)}$, $g^{(i)}$, $h^{(i)}$. The six permutations of the shape function in Eq. (51) thus reduce to three. This is possible due to the invariance under a simultaneous interchange of $f^{(i)}$, $g^{(i)}$ and m_a, m_b in Eq. (59) or, more physically, the indistinguishability of the two scalar components of the 3-point function.

The similarity of Eq. (59) to the standard $\zeta\zeta\zeta$ KSW estimator in Eq. (44) is evident. The most important difference between the expressions is the anisotropy in the dependence on m_a, m_b, and M in Eq. (59); as a reminder, in order to construct the equivalent of a KSW estimator for the $\zeta\zeta h$ template, we needed to construct pieces of the bispectrum separable in ℓ_1, ℓ_2, and ℓ_3, and this could only be done at the expense of introducing several anisotropic templates. As discussed previously, the estimates of the amplitudes of the anisotropic terms combine into an estimate of the amplitude of the original isotropic template. The anisotropy appears in the Wigner 3-j symbol and the m_a, m_b, and M indices of the generalized $\mathcal{A}$ functionals in Eq. (59). We will discuss the meaning of the m_a, m_b, and M indices and the 3-j symbol in more detail in the remainder of this section, but in short each (m_a, m_b, M) triplet corresponds to a combination of the longitudinal, solenoidal, and/or transverse angular modes of the contracted angular term [see Eq. (45)] that is present in the $\zeta\zeta h$ 3-point function. For a given (m_a, m_b, M) triplet, Eq. (59) estimates the contribution from the corresponding combination of angular modes to the data; the 3-j symbols then provide a relative weight to each contribution when all are summed into the final estimate $\hat{f}^{\zeta\zeta h}_{\mathrm{NL,cubic}}$.

Before coming to the computational scaling of the estimator, let us focus our attention to the generalized $\mathcal{A}$ functionals in Eq. (59). For a given input function $f(k)$, $\mathcal{A}^{(Z)}_{(S,n)}[f]$ returns a scalar field on a spherical shell at comoving radial coordinate r. The S index denotes whether the associated factor of the 3-point function is a monopole ($S = 0$), dipole ($S = 1$), or quadrupole ($S = 2$) source. The n index tells us whether we are considering the longitudinal ($n = 0$), solenoidal ($n = \pm 1$), or transverse ($n = \pm 2$) part of the source. From Eq. (59) we see that for the $\zeta\zeta h$ bispectrum we only need the $S = 1$ functionals for the $Z = \zeta$ part and the $S = 2$ functionals for the $Z = h$ part.

At each radial coordinate r we may decompose the $\mathcal{A}$ functionals in terms of spherical harmonics:

$$\mathcal{A}^{(Z)}_{(S,n)}[f](r,\hat{\mathbf{n}}) = \sum_{L,M}(\mathcal{A}^{(Z)}_{(S,n)}[f])_{LM}(r)Y_{LM}(\hat{\mathbf{n}}). \tag{60}$$

The resulting harmonic modes are given by linear transformations of the inverse-covariance-weighted data. Based on the primordial index Z, we identify two cases:

$$(\mathcal{A}^{(Z)}_{(S,n)}[f])_{LM}(r) \equiv \begin{cases} (4\pi)^{1/2}\sum_{\ell,m} i^{\ell+L}J^{000}_{SL\ell}\begin{pmatrix} S & L & \ell \\ n & M & m \end{pmatrix}\sum_X (\mathcal{K}^{(\zeta)}[f])^X_{\ell,L}(r)(C^{-1}a)^X_{\ell m} & Z = \zeta, \\ (4\pi)^{1/2}\sum_{\ell,m} i^{\ell+L}J^{-202}_{SL\ell}\begin{pmatrix} S & L & \ell \\ n & M & m \end{pmatrix}\sum_X [1 + (-1)^{x+L+\ell}](\mathcal{K}^{(h)}[f])^X_{\ell,L}(r)(C^{-1}a)^X_{\ell m} & Z = h. \end{cases} \tag{61}$$

Note that for the $Z = \zeta$ case, the sum over X only runs over $\{T, E\}$, while for $Z = h$ it runs over $\{T, E, B\}$. The parity behavior associated with a given polarization index X is denoted by x. The $\mathcal{K}$ functionals are defined in Eq. (52). The data are filtered by the different $\mathcal{K}$ functionals in an anisotropic manner depending on the value of n. For example, the $(\mathcal{A}_{(S,2)})_{LM}$ modes are sourced by the $m = -(M + 2)$ modes of the data.

The inverse spherical harmonic transformation needed to evaluate Eq. (60) scales as $\mathcal{O}(\ell^3_{\mathrm{max}})$ and will in reality determine the overall scaling of the estimator evaluation. One might worry that the sums over ℓ and m needed to construct the harmonic coefficients in Eq. (61) will contribute significantly to the computational scaling. This is not the case, as the selection rules of the Wigner 3-j symbols forbid most values of ℓ and m. Only $\ell \in L \pm 1$ and $m = -(M + n)$ are needed to compute $\mathcal{A}^{(\zeta)}_{LM}$ while for $\mathcal{A}^{(h)}_{LM}$ only $\ell \in \{L, L \pm 1, L \pm 2\}$ and $m = -(M + n)$ are required.

To compute the angular integral in Eq. (59), the pixelization scheme used for the $\mathcal{A}[f]$ fields (or "maps") must support harmonic band limits given by the sum of the band limits of the three individual maps [see Eq. (B8)]. In reality, the $\mathcal{A}^{(\zeta)}$ maps will likely be bandlimited by the instrumental beam or noise covariance. On the other hand, the $\mathcal{A}^{(h)}$ maps only contain information on large ($\ell \lesssim 200$) scales; the tensor transfer functions suppress all information in the data on smaller scales. Small-scale tensor perturbations produced by an approximately scale-invariant process are inaccessible through the primary anisotropies. Unlike scalar perturbations, small-scale tensor perturbations decay away with cosmic expansion before recombination.

It is instructive to take a closer look at how the symmetries of the spherical harmonics and the 3-j symbols relate the harmonic coefficients of the $\mathcal{A}$ functionals with (S, n) to those with $(S, -n)$. This relation can be used to approximately half the number of inverse harmonic transformations needed to evaluate Eq. (59). Assuming that the input function $f(k)$ is real valued, the coefficients transform as follows under complex conjugation:

$$(\mathcal{A}^{(Z)}_{(S,n)}[f])^*_{LM} = (\mathcal{A}^{(Z)}_{(S,-n)}[f])_{L-M}(-1)^{n+M+S}. \tag{62}$$

It follows that the functionals in Eq. (60) map input functions to complex fields on the sphere that obey

$$(\mathcal{A}^{(Z)}_{(S,n)}[f])^*(\hat{\mathbf{n}}) = (\mathcal{A}^{(Z)}_{(S,-n)}[f])(\hat{\mathbf{n}})(-1)^{n+S}. \tag{63}$$

Going back to the estimator in Eq. (59), we see that only five out of the nine allowed combinations of m_a, m_b, and M need to be considered: the remaining terms may be found with the use of Eq. (63). We may, for example, use the following five combinations:

$$(m_a, m_b, M) \in \{(1,1,-2), (1,0,-1), (0,1,-1), (1,-1,0), (0,0,0)\}. \tag{64}$$

The $m_a = m_b = M = 0$ case is unique, the other four combinations in Eq. (64) are related to the remaining four combinations by a factor of (-1). The 3-j symbol in Eq. (59) does not change if this minus sign is added to its lower indices. For the $\mathcal{A}$ maps, Eq. (63) tells us that the addition of a minus sign to the n index is equivalent to complex conjugation. For the products of $\mathcal{A}$ maps the following thus holds:

$$\mathcal{A}^{(\zeta)}_{(1,m_a)}\mathcal{A}^{(\zeta)}_{(1,m_b)}\mathcal{A}^{(h)}_{(2,M)} + \mathcal{A}^{(\zeta)}_{(1,-m_a)}\mathcal{A}^{(\zeta)}_{(1,-m_b)}\mathcal{A}^{(h)}_{(2,-M)} = 2\mathrm{Re}(\mathcal{A}^{(\zeta)}_{(1,m_a)}\mathcal{A}^{(\zeta)}_{(1,m_b)}\mathcal{A}^{(h)}_{(2,M)}). \tag{65}$$

Note that we have suppressed the $f^{(i)}$, $g^{(i)}$, and $h^{(i)}$ input functions to the $\mathcal{A}$'s. Equation (65) implies that, instead of computing nine products, one can only calculate five products of complex $\mathcal{A}$ maps and discard the imaginary parts to evaluate Eq. (59). The fact that only five out of nine terms are needed can be understood from the original expression for the angular term. Starting with the nine terms in the sum over a and b in Eq. (45), the symmetry under the simultaneous exchange of a, b and $\zeta_{\mathbf{k}_1}$, $\zeta_{\mathbf{k}_2}$ removes 3 degrees of freedom. The vanishing trace of the polarization tensor removes the fourth.

It is easy to see that the two additional estimator terms with permuted indices in Eq. (58) are constructed by permuting the columns of the 3-j symbol together with the (m_a, ζ), (m_b, ζ), and (M, h) index pairs of the three $\mathcal{A}$ functionals in Eq. (59). The 3-j symbol is invariant under such permutations. The product of $\mathcal{A}$ maps is also invariant under such permutations because of the symmetrized form of the shape function in Eq. (50). The total cubic term of the estimator is therefore simply given by

$$\hat{f}^{\mathrm{tot}}_{\mathrm{NL,cubic}} = 3\hat{f}^{\zeta\zeta h}_{\mathrm{NL,cubic}}. \tag{66}$$

After deriving the cubic part of the estimator, the linear term is obtained in an analogous way. It can be found by inserting the bispectrum in Eq. (51) into Eq. (26) and keeping the terms linear in the data:

$$\hat{f}^{\zeta\zeta h}_{\mathrm{NL,lin}} = -\frac{\sqrt{2}}{54\mathcal{I}_0}\sum_{M,m_a,m_b}\begin{pmatrix}1 & 1 & 2\\ m_a & m_b & M\end{pmatrix}\int_{S^2}\mathrm{d}\Omega(\hat{\mathbf{n}})\sum_{i=1}^{N_{\mathrm{prim}}}\int_0^\infty r^2\mathrm{d}r \times(\langle\mathcal{A}^{(\zeta)}_{(1,m_a)}[f^{(i)}]\mathcal{A}^{(\zeta)}_{(1,m_b)}[g^{(i)}]\rangle_{\mathrm{MC}}\mathcal{A}^{(h)}_{(2,M)}[h^{(i)}] + 8\,\mathrm{perm})(r,\hat{\mathbf{n}}). \tag{67}$$

We again assume an input shape function parametrized by Eq. (50). The eight additional permutations in Eq. (67) are those constructed by cyclic permutations of $f^{(i)}$, $g^{(i)}$, $h^{(i)}$ and by varying which pair of $\mathcal{A}$'s sits in the $\langle\rangle_{\mathrm{MC}}$ brackets.

Similar to the total cubic term, it may be checked that including the two cyclic permutations of $\zeta\zeta h$ simply amounts to

$$\hat{f}^{\mathrm{tot}}_{\mathrm{NL,lin}} = 3\hat{f}^{\zeta\zeta h}_{\mathrm{NL,lin}}. \tag{68}$$

Finally, the normalization of the estimator $\mathcal{I}_0$ may be estimated by simply applying the unnormalized estimator to an ensemble of simulated data. Given the expressions for the cubic and linear terms presented here, the efficient algorithm from Ref. [119] for the estimation of the normalization can also be used for this type of bispectrum. We omit the details of this implementation.

This concludes the derivation of the estimator for the $\zeta\zeta h$ 3-point function. The resulting expression is given in Eq. (59). In Appendix A, we show how one would repeat this effort for several more involved 3-point functions.

IV. FISHER FORECASTS

We forecast the expected uncertainty on an upper limit on the amplitude of a squeezed $\zeta\zeta h$ 3-point correlation function. We illustrate the constraining power of current and upcoming CMB experiments, and demonstrate how the upper limit depends on certain instrumental effects. We expand on previous forecasts in Refs. [17,22] by taking into account the dependence on the lower harmonic band limit of the data, the addition of E-mode data and the extra variance induced by weak lensing. In a future paper, we apply the derived estimator to a set of map-based simulations to better judge the effects of foreground contamination, nontrivial noise covariances, and secondary non-Gaussian contamination. In this light, the forecasts presented here should be considered as a baseline for more realistic forecasts.

A. Procedure

Before presenting the results from the Fisher forecasts, this section specifies the exact parametrization of the $\zeta\zeta h$ 3-point function. We also explain the assumed experimental setup and the numerical implementation of forecast calculation.

We parametrize the k-dependent part of the $\zeta\zeta h$ template in Eq. (24) as follows:

$$f^{(\zeta\zeta h)}(k_1,k_2,k_3) = 16\pi^4 A_s^2 f_{\mathrm{NL}}^{\mathrm{tot}} f(k_1,k_2,k_3). \qquad (69)$$

A_s represents the amplitude of the curvature perturbation (see Appendix C 1). We imagine an analysis that looks for a deviation from the tensor consistency relation by placing an upper limit on the amplitude of the squeezed 3-point function; we thus use the standard local shape of the $f(k_1,k_2,k_3)$ template as a generic squeezed shape template. See Eq. (C7) for the precise expression. The local shape differs slightly from the SFSR shape template [18] used in Refs. [17,22,59]. However, the two templates give almost equal weight to squeezed configurations with a large-wavelength tensor perturbation. Given that the tensor perturbation only sources CMB anisotropies on large angular scales, we may, for all practical purposes, consider the shapes as equal here. This is reflected in the results we obtain: our forecasts agree with those in Refs. [17,22] when parameters overlap.[16]

For simplicity, we only consider the $\langle TTB\rangle$, $\langle EEB\rangle$, and $\langle TEB\rangle$ bispectra in the forecasts. We thus do not take into account the information contained in the $\langle TTT\rangle$, $\langle TTE\rangle$, $\langle TEE\rangle$, and $\langle EEE\rangle$ bispectra. The main justification for this choice is the associated extra cosmic variance due to the lack of a B-mode component. Additionally, it should be noted that the squeezed $\langle TTT\rangle$ bispectrum is expected to be relatively strongly contaminated by a secondary non-Gaussian signal [125]. It is expected to be of limited use for our purpose; see the discussion in Sec. V.

We use the inverse Fisher information $\mathcal{I}_0$ as an estimate for the estimator variance. The 1σ upper limits that we will quote are simply given by $1/\sqrt{\mathcal{I}_0}$. We calculate the Fisher information in the limit of no non-Gaussian signal contribution; i.e., we use Eq. (30). We further simplify the situation by assuming isotropic signal and noise covariances. The resulting diagonal covariance matrices, together with the orthonormality relation of the Wigner 3-j symbols in Eq. (B14) allow the Fisher information to be expressed in terms of angle-averaged bispectra. The effects from incomplete sky coverage are treated in a simplified manner by taking into account an increase in estimator variance proportional to the observed fraction of the sky (f_{sky}). Given this trivial scaling, we assume $f_{\mathrm{sky}} = 1$ in all of the following. Finally, we use the lensed version of the CMB power spectra, but neglect the non-Gaussian aspects of CMB lensing. See the discussion in Sec. V D.

The resulting simplified expression for the Fisher information $\mathcal{I}_0$ is given by

$$\mathcal{I}_0 = f_{\mathrm{sky}} \sum_{\ell_1\le\ell_2\le\ell_3} \sum_{\mathrm{all}\,X} \frac{1}{\Delta_{\ell_1\ell_2\ell_3}} (B_1)^{X_1X_2X_3}_{\ell_1\ell_2\ell_3} \times [(C^{-1})^{X_1X_4}_{\ell_1}(C^{-1})^{X_2X_5}_{\ell_2}(C^{-1})^{X_3X_6}_{\ell_3}](B_1^*)^{X_4X_5X_6}_{\ell_1\ell_2\ell_3}, \qquad (70)$$

with $(B_1^*)^{X_1X_2X_3}_{\ell_1\ell_2\ell_3} = (B_1)^{X_1X_2X_3}_{\ell_1\ell_2\ell_3}(-1)^{\ell_1+\ell_2+\ell_3}$ and with total angle-averaged bispectrum given by

$$(B_1)^{X_1X_2X_3}_{\ell_1\ell_2\ell_3} = (B_1)^{X_1X_2X_3(\zeta\zeta h)}_{\ell_1\ell_2\ell_3} + (B_1)^{X_1X_2X_3(\zeta h\zeta)}_{\ell_1\ell_2\ell_3} + (B_1)^{X_1X_2X_3(h\zeta\zeta)}_{\ell_1\ell_2\ell_3}. \qquad (71)$$

The factor of $\Delta_{\ell_1\ell_2\ell_3}$ in Eq. (70) simply results from using $(1/6)\sum_{\ell_1,\ell_2,\ell_3} = \sum_{\ell_1\le\ell_2\le\ell_3} 1/\Delta_{\ell_1\ell_2\ell_3}$ where $\Delta_{\ell_1\ell_2\ell_3}$ is defined to equal 6 for identical ℓ indices, 1 for unequal indices, and 2 otherwise. This simplification is possible because the bispectrum is invariant under all six permutations of its (ℓ, m) index pairs.[17] Written as such, permutations of $\{X_1,X_2,X_3\}$, $\{X_4,X_5,X_6\}$, and $\{\zeta,h\}$ become distinct and have to be explicitly summed over.

As explained in Sec. III C 2, we may obtain the angle-averaged version of the $\zeta\zeta h$ bispectrum by summing over the m_a, m_b, M, M_1, M_2, and M_3 indices in Eq. (51) and inserting the resulting bispectrum into Eq. (18). This will yield the expression first derived in Ref. [83]. The first term in Eq. (71) for the primordial shape in Eq. (50) is given by

[16] The definition in Eq. (69) differs from the one used in Refs. [17,31] by a factor of $\sqrt{r}$: $f_{\mathrm{NL}}^{\mathrm{here}} = \sqrt{r} f_{\mathrm{NL}}^{\mathrm{there}}$, where r is the tensor-to-scalar ratio. To compare our results to those in Ref. [22], use $f_{\mathrm{NL}}^{\mathrm{here}} = (\lambda_{sst}\epsilon)^{\mathrm{there}}$.

[17] Note that the angle-averaged bispectrum used in Eq. (70) is only invariant under cyclic permutations of ℓ_1, ℓ_2, and ℓ_3. For odd permutations, it picks up a factor of $(-1)^{\ell_1+\ell_2+\ell_3}$. Although we consider the $\ell_1+\ell_2+\ell_3 = \mathrm{odd}$ case here, the factors of (-1) cancel in the expression for the Fisher information, so we may still use the $1/\Delta_{\ell_1\ell_2\ell_3}$ simplification.

$$(B_1)^{X_1X_2X_3(\zeta\zeta h)}_{\ell_1\ell_2\ell_3} = \frac{(8\pi)^{3/2}}{3}\sum_{L_1,L_2,L_3}\left(\prod_{i=1}^{3}(-i)^{\ell_i-L_i}\right)J^{000}_{L_1L_2L_3}J^{000}_{\ell_1L_11}J^{000}_{\ell_2L_21}J^{20-2}_{\ell_3L_32}\begin{Bmatrix}\ell_1 & \ell_2 & \ell_3\\ L_1 & L_2 & L_3\\ 1 & 1 & 2\end{Bmatrix}$$
$$\times\frac{1}{6}\sum_{i=1}^{N_{\mathrm{prim}}}\int_0^\infty r^2\mathrm{d}r[(\mathcal{K}^{(\zeta)}_{(X_1)}[f^{(i)}])_{\ell_1,L_1}(\mathcal{K}^{(\zeta)}_{(X_2)}[g^{(i)}])_{\ell_2,L_2}(\mathcal{K}^{(h)}_{(X_3)}[h^{(i)}])_{\ell_3,L_3}](r) + (5\text{ perm}). \tag{72}$$

The other two terms in Eq. (71) are obtained by permuting the ζ and h indices. The five permuted terms in Eq. (72) refer to permutations of the $f^{(i)}$, $g^{(i)}$, and $h^{(i)}$ functions. The $\mathcal{K}$ functionals were introduced in Eq. (52).

The evaluation of Eq. (70) has an overall $\mathcal{O}(\ell^3_{\mathrm{max}})$ scaling. The computation is feasible because the $\mathcal{K}$ functionals in Eq. (72) can be precomputed. However, for high band limits (e.g., $\ell_{\mathrm{max}} = 5000$ used below) the procedure is unwieldy. This is especially true when multiple choices for the inverse signal + noise covariance matrix C^{-1} are to be explored. The computation of multiple Wigner 9-j symbols at every valid (ℓ_1, ℓ_2, ℓ_3) triplet exacerbates the situation compared to the Fisher information for a $\zeta\zeta\zeta$ bispectrum.

To get around the computational complexity of Eq. (70), we split the problem into two parts: We first store a sparsely sampled representation of Eq. (72). We then interpolate this representation over all multipole orders when the sums over ℓ_1, ℓ_2, and ℓ_3 are performed. This approach results in an insignificant reduction in accuracy but reduces evaluation time significantly. Computing $\mathcal{I}_0$ with $\ell_{\mathrm{max}} = 5000$ takes roughly 30 CPU minutes. The method is effective because the smoothness of the primordial templates and transfer functions (in k and ℓ, respectively) translate into an angle-averaged bispectrum that is rather smooth with ℓ_1, ℓ_2, and ℓ_3.[18]

The sparse sampling is determined by the following binning scheme: $\Delta\ell = 1$ for $\ell \leq 50$, $\Delta\ell = 4$ for $50 < \ell \leq 200$, $\Delta\ell = 12$ for $200 < \ell \leq 500$, $\Delta\ell = 24$ for $500 < \ell \leq 2000$, and finally $\Delta\ell = 40$ for $\ell > 2000$. This binning scheme is used for the ℓ_1, ℓ_2, and ℓ_3 dimensions. In each resulting three-dimensional bin, a single valid sample (depending on the parity and triangle constraints) is selected. The angle-averaged bispectrum for each (X_1, X_2, X_3) polarization tuple is then calculated over all selected samples. The integral over r in Eq. (72) is evaluated using the trapezoidal rule with 500 integration points that span $0 \leq r \leq 18000$. Most points are placed around regions corresponding to the reionization and recombination eras. With some effort, we expect that the number of r samples can be reduced by a factor of 10. The resulting sparse, angle-averaged bispectra are compact enough to be saved to disk. Finally, to evaluate Eq. (70) the sparse representations are interpolated over all valid multipole combinations using a three-dimensional linear interpolation scheme. The result is weighed by the (unbinned) inverse covariance matrices in Eq. (70).

The above algorithm is implemented in a publicly available Python code library.[19] The code makes heavy use of the scientific SciPy and NumPy libraries.[20] Performance-critical steps are compiled to optimized machine code at runtime by Numba: a just-in-time Python compiler [126]. The Wigner symbols are evaluated using the WIGXJPF library [127]. The radiation transfer functions and CMB power spectra are computed using CAMB. Finally, every step of the code has been written with the message passing interface (MPI) standard in mind; computing in parallel on distributed memory systems is therefore possible. The code should be relatively easily adaptable to other (smooth) bispectrum templates. The repository also contains the necessary scripts to reproduce the results in the following section.

In summary, we use the Fisher information to forecast the expected upper limits on the amplitude of the squeezed $\zeta\zeta h$ 3-point function. The exact form of the $\zeta\zeta h$ correlation is specified in Eqs. (24) and (69) with the standard local shape template for $f(k_1, k_2, k_3)$.

B. Results

The results presented in this section fall into three categories. We first study how the expected upper limits on the $\zeta\zeta h$ amplitude vary as functions of minimum and maximum multipole moments. Second, we explore how advantageous it is to use both T- and E-mode data together with the B-mode data. Finally, we investigate the deterioration of the upper limits due to gravitational lensing. We emphasize that by using the Fisher information, Eq. (70), to determine the best-achievable upper limits we are effectively investigating how well a finite number of purely Gaussian distributed data points can constrain the $\zeta\zeta h$ amplitude to be zero.

We start by exploring how the minimum multipole moment of the B-mode data affects the constraining power. The flat-sky forecasts in Ref. [17] did not probe this regime. The lowest achievable lower band limit ℓ^B_{min} is one of the

[18] This is only true when the factor of $(-i)^{\ell_1+\ell_2+\ell_3}$ in Eq. (72) is ignored. If required (for the cross-correlation of two different templates), this phase can be included after the interpolation step.

[19] https://github.com/adrijd/cmb_sst_ksw.

[20] https://www.scipy.org.

main distinctions between ground-based and satellite CMB experiments. The atmosphere prohibits measurements over large angular scales. Current B-mode data from ground-based observatories reach $\ell^B_{\rm min} \approx 50$. Polarization modulation techniques, such as spinning half-wave plates, might allow future efforts to reach an effective $\ell^B_{\rm min} \approx 30$ [52]. Without atmospheric contamination satellite missions can in principle reach $\ell^B_{\rm min} = 2$. In reality, it remains to be seen if uncertainty on systematic instrumental effects and Galactic foregrounds will allow such a challenging measurement to be made. A more conservative estimate for a satellite (or balloon-borne) experiment would be $\ell^B_{\rm min} \approx 20$.

In Fig. 4 we show the achievable 1σ upper limits on $f^{\rm tot}_{\rm NL}$ as a function of overall band-limit $\ell_{\rm max}$ and lower band-limit $\ell^B_{\rm min}$. There is no contribution from instrumental noise, and the only source of uncertainty is the cosmic variance induced by the Gaussian components of ζ and h.

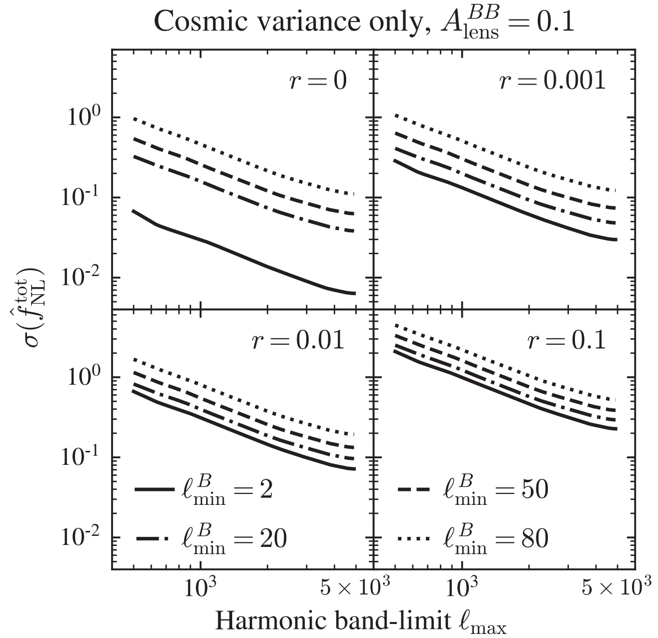

FIG. 4. Achievable 1σ upper limits on $f^{\rm tot}_{\rm NL}$, i.e., the unavoidable errors solely caused by cosmic variance, as a function of maximum harmonic band limit $\ell_{\rm max}$. Here, the $f^{\rm tot}_{\rm NL}$ parameter is the amplitude of the $\zeta\zeta h$ 3-point function with a local shape function. The lines in each panel correspond to lower band limits $\ell^B_{\rm min}$ of the B-mode data. The vast improvement due to low-multipole B-mode data seen in the upper-left panel is caused by the contribution from reionization to the bispectrum. When the tensor power is increased (the other three panels) the scaling with $\ell^B_{\rm min}$ becomes more regular: the contribution from reionization gets suppressed by the B-mode power spectrum. Still, the low multipole orders contain a significant amount of information on the 3-point function. These results take into account the Fisher information in the $\langle TTB\rangle$, $\langle TEB\rangle$, and $\langle EEB\rangle$ CMB bispectra. The lensing contribution to the B-mode power spectrum is assumed to be "delensed" to only 10% of the ΛCDM amplitude ($A^{BB}_{\rm lens} = 0.1$).

The lensing contribution to the B power spectrum is assumed to be "delensed" to only 10% of the ΛCDM amplitude ($A^{BB}_{\rm lens} = 0.1$). It is clear that as long as the Gaussian contribution to h is neglected, i.e., $r = 0$, the upper limits strongly benefit from a low $\ell^B_{\rm min}$. Scattering at reionization significantly contributes to the $\ell \lesssim 20$ B-mode components of the bispectrum for $r < 0.001$. The lensing contribution to B is essentially negligible at such large angular scales, so the low-ℓB-mode data become a highly sensitive probe of the squeezed bispectrum. When $r \neq 0$, the additional cosmic variance induced by h quickly closes this window, even though there still remains a significant dependence on $\ell^B_{\rm min}$ for $r \neq 0$. We find that for $r \geq 10^{-2}$, the 1σ upper limits conform rather well to the $\ell_{\rm max}(\log(\ell^B_{\rm max}/\ell^B_{\rm min}))^{1/2}$ scaling conjectured in Ref. [21]. Here $\ell_{\rm max}$ refers to the band limit of the T- and E-mode data, while $\ell^B_{\rm max}$ refers to the band limit of the B-mode data. The scaling fits well when $\ell^B_{\rm max} \approx 150$: roughly the maximum multipole order that contains usable information on the primordial tensor perturbation for a 90% delensed B-mode power spectrum. The curves in the two panels in Fig. 4 that have $r < 10^{-2}$ do not fit the scaling: the relatively strong contributions from reionization and lensing are not captured by the analytic relation. Finally, the observation that a lower r will tighten the upper limit on the $\zeta\zeta h$ amplitude should not be mistaken with the idea that a lower r will increase the potential of detecting the $\zeta\zeta h$ correlation. Letting $r \to 0$ increases the potential of ruling out a nonzero $f^{\rm tot}_{\rm NL}$ because of lower cosmic variance, but for a detection of $f^{\rm tot}_{\rm NL}$, r has to be nonzero. The precise relation between $f^{\rm tot}_{\rm NL}$ and r is model dependent. Our choice to parametrize the $\zeta\zeta h$ amplitude in a model-independent way hides this subtlety; see Footnote 16.

The relative importance of the low-ℓB-mode data also grows when the lensing contribution to the B power spectrum is increased. This behavior is depicted in Fig. 5. As we move from no lensing BB contribution to the full ΛCDM amplitude, the low-ℓB-mode data become more relevant. This is a simple consequence of the shape of the lensing contribution relative to the bispectrum. The dominant lensing contribution to the estimator variance, i.e., the B lensing power spectrum, is roughly constant with ℓ on large scales while the $\langle TTB\rangle$, $\langle EEB\rangle$, and $\langle TEB\rangle$ bispectra peak at configurations with large-scale B-mode components.

Note that the lower band limit used for the T and E data is set at $\ell = 2$ for all results presented in this section. The rationale behind this choice is that the *WMAP* and *Planck* data already provide cosmic-variance limited data for T and E on large angular scales. Note that this is not strictly true for the E-mode data. Current $\ell \lesssim 30$ E-mode data are systematic limited [128,129]. We have checked that by conservatively removing the $\ell \leq 30$ E-mode data the curves do not visibly change.

We now focus on the individual and combined contribution of the $\langle TTB\rangle$, $\langle TEB\rangle$, and $\langle EEB\rangle$ bispectra.

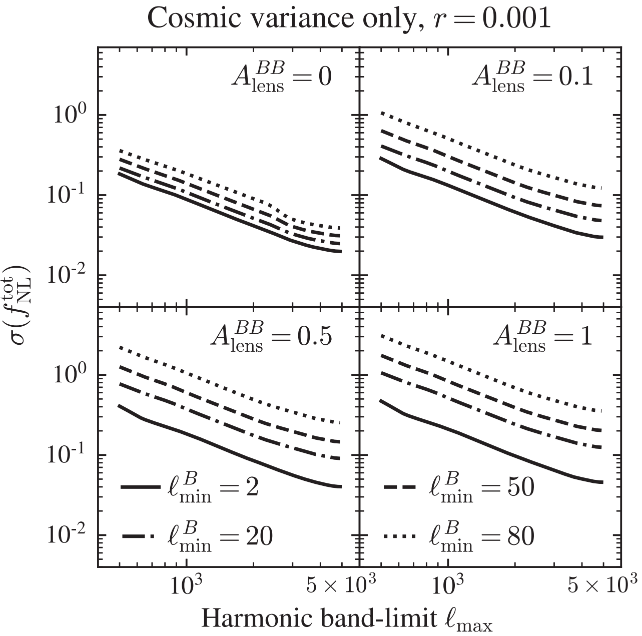

FIG. 5. Cosmic variance limited 1σ upper limits on $f^{\rm tot}_{\rm NL}$ as a function of maximum harmonic band limit $\ell_{\rm max}$. Here, the $f^{\rm tot}_{\rm NL}$ parameter is the amplitude of the $\zeta\zeta h$ 3-point function with a local shape function. The lines in each panel correspond to lower band limits $\ell^B_{\rm min}$ of the B-mode data. As the lensing contribution to the B-mode power spectrum $A^{BB}_{\rm lens}$ is increased from the upper-left panel to the lower-right panel, upper limits worsen and become more dependent on the low-multipole B-mode data. These limits take into account the Fisher information in the $\langle TTB\rangle$, $\langle TEB\rangle$, and $\langle EEB\rangle$ CMB bispectra. The tensor contribution to the CMB power spectra is sourced by an $r = 0.001$ primordial tensor power spectrum.

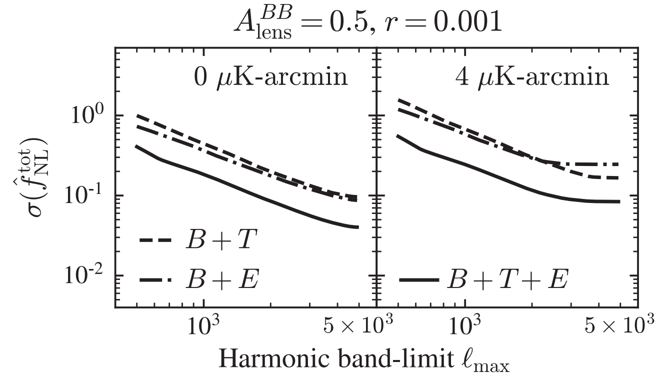

FIG. 6. How the expected upper limits (1σ) on the amplitude $f^{\rm tot}_{\rm NL}$ of the $\zeta\zeta h$ 3-point function change when E-mode data are excluded (dashed lines) or when T data are excluded (dot-dashed lines). Combined constraints (i.e., from the Fisher information in the $\langle TTB\rangle$, $\langle TEB\rangle$, and $\langle EEB\rangle$ bispectra) (solid lines) are significantly stronger than those obtained from a naive addition of the $B + T$ and $B + E$ Fisher information. This effect holds when (white) noise is added to the data; the left panel shows the noiseless case, while in the right panel 4 ($4\sqrt{2}$) μK-arcmin noise is added to the T (E, B) harmonic modes. For these noise levels, the T (E) data are cosmic-variance limited up to $\ell \approx 4000$ (2500). For data with higher band limits ($\ell_{\rm max}$) the constraints saturate due to the noise. The addition of white noise to the B-mode data is responsible for the overall upward shift of the curves in the right panel. Note that the lower harmonic band limit of the B-mode data is set to $\ell = 2$ for this figure.

In Ref. [17] only the $\langle TTB\rangle$ bispectrum was taken into account. Reference [22] additionally calculated the Fisher information associated with the $\langle EEB\rangle$ bispectrum. In Fig. 6 we demonstrate how combining the information in T and E (in addition to B) yields much better results than the Fisher information of the individual cases would suggest. This effect is also seen in the $\zeta\zeta\zeta$ non-Gaussianity estimation and can be traced back to the fact that the T and E transfer functions for ζ are out of phase [79]. The same is true for the radial transfer functions we use; see Fig. 1. This effect holds up under slightly more realistic circumstances: by adding 4 μK-arcmin white noise to the T harmonic modes and $4\sqrt{2}$ μK-arcmin to the E and B harmonic modes, we see the same behavior.

Finally, we investigate the relation between the lensing amplitude and the instrumental noise level. As mentioned before, the lensing signal serves as a cosmic variance contribution to the estimator variance. The lensing contribution to the T- and E-mode power spectra provides a relatively minor contribution, while the contribution from the lensed B power spectrum is significant. Fortunately, the lensing contribution to the B-mode field is not entirely irreducible: with knowledge of the lensing potential, the lensing contribution can be reduced, or delensed [130]. In Fig. 7 we show upper limits as a function of instrumental B-mode noise for the case of only 10% lensing contribution to the B-mode power spectrum ($A^{BB}_{\rm lens} = 0.1$) and for the full lensing contribution. The instrumental B-mode noise ranges from 50 to 0.3 μK-arcmin. To put this in context: the upper value roughly corresponds to the noise level in the *Planck* data. The Simons Observatory [52] and *LiteBIRD* [58] experiments aim to achieve a B-mode noise level of approximately 3 μK-arcmin, while the CMB-S4 proposal [47] aims for approximately 1 μK-arcmin. From Fig. 7 it becomes clear that the lensing BB contribution starts to dominate over the instrumental noise for noise amplitudes below 5 μK-arcmin. This is unsurprising given that the large-scale B-mode lensing contribution is well approximated by 5 μK-arcmin white noise [131]. We can thus infer that for the Simons Observatory or *LiteBIRD* experiments the gain from B-mode delensing would be noticeable but relatively minor, while an experiment such as CMB-S4 would need at least a factor of 10 of delensing in the B-mode power to make use of the potential of the instrumental sensitivity.

In summary, the forecasts demonstrate that the statistical improvement with minimum and maximum multipole moments roughly follows the expected behavior for a squeezed 3-point function, with the exception that a low $\ell_{\rm min}$ for the B-mode data is more advantageous than one

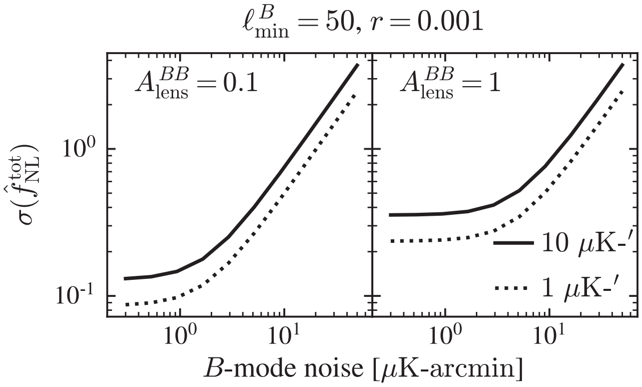

FIG. 7. Expected upper limits (1σ) on the amplitude $f^{\rm tot}_{\rm NL}$ of the $\zeta\zeta h$ 3-point function as a function of the (white) noise amplitude of the B harmonic modes. It can be seen how decreasing the B-mode noise is useful only up to a certain limit given by the amplitude of the lensing BB contribution. For the B-mode data that are delensed to only 10% of the ΛCDM amplitude ($A^{BB}_{\rm lens} = 0.1$) (left panel), the constraining power saturates roughly below 1 μK-arcmin. Without delensing, the constraints already start to saturate below 5 μK-arcmin. This behavior is essentially independent of the noise level of the T and E data: the same curve is seen regardless of whether 1 ($\sqrt{2}$) (solid lines) or 10 ($10\sqrt{2}$) (dotted lines) μK-arcmin noise is added to the T (E) data. Note that the harmonic band limit of the data is set to $\ell_{\rm max} = 5000$.

would naively expect. Constraints benefit significantly from the simultaneous use of T- and E-mode data. Last, future experiments will need to delens their B-mode data significantly to keep improving upper limits. It should be noted that these conclusions will likely differ for shapes that are not squeezed.

V. DISCUSSION

Generally, we expect two effects that will influence our ability to measure primordial non-Gaussianity. The first effect is a bias in the estimated amplitude of the primordial signal, i.e., a mismatch between the true amplitude and the expectation value of the estimate, due to other non-Gaussian signals that mimic the primordial signal. The nonprimordial, non-Gaussian signal is, for example, caused by secondary extragalactic sources and Galactic foregrounds. In some cases, these biases may be subtracted from the estimate or captured by a joint estimate; see, e.g., the lensing-ISW bias in the *Planck* analysis [15]. A second, more irreducible effect comes from the fact that the non-Gaussian signal, primordial or secondary, will contribute to the estimator variance. When this contribution exceeds the contribution from (cosmic) variance from the Gaussian CMB component and detector noise, simulations of the responsible non-Gaussian signal are needed to accurately characterize the estimator variance. While we will leave a detailed discussion of both effects to a future publication, here we provide a brief discussion. We focus on contaminants for squeezed bispectra with one large-scale B-mode component, as such bispectra will provide the largest constraining power for the primordial $\zeta\zeta h$ 3-point function.

A. Polarized Galactic foregrounds

The large-scale polarization B- and (to a lesser extent) E-mode fields are dominated by Galactic emission: at low frequencies by synchrotron radiation and at higher frequencies by polarized dust emission [132]. Because the primordial B-mode signature is expected at large angular scales ($\ell \lesssim 100$), inference on the tensor-to-scalar ratio r relies heavily on multifrequency data to break the degeneracy between foreground and CMB power. Similarly, inference on the bispectra we are interested in would require uncontaminated large-scale B-mode data.

One would naively expect that component-separated B-mode data suitable for constraints on r are also suitable for constraints on bispectra with B-mode components. However, there is an extra complication for bispectrum inference: residual anisotropic or non-Gaussian correlations between foreground B and foreground T or E signals. Residual correlations of this type might not be important for a power spectrum analysis but will bias a bispectrum analysis. Unfortunately, it is quite natural to expect a Galactic signal to source a squeezed bispectrum: small-wavelength foreground power in a given direction is likely not independent from the foreground signal on larger wavelengths in the same direction. The question is thus whether multifrequency cleaning of the data will suppress such correlations enough.

Characterization of the non-Gaussian aspects of the polarized Galactic signal is relatively unexplored at this point. Early results obtained from the *Planck* data in Ref. [133] suggest that there are indeed significant squeezed $\langle TTB\rangle$, $\langle TEB\rangle$, and $\langle EEB\rangle$ bispectra on large angular scales in the thermal dust component of the Galactic signal. No significant bispectrum is found in the synchrotron emission. Reference [133] does not find a significant non-Gaussian correlation when foreground-cleaned *Planck* B-mode data are correlated with the T and/or E components of the Galactic dust. Although this analysis omits the very large angular scales ($\ell \leq 40$), it does suggests that the standard component separation methods sufficiently suppress Galactic foregrounds given the *Planck* noise level. It should also be noted that in a related study no evidence was found for a dust bispectrum template in the foreground-cleaned *Planck* temperature data [134]. More investigation is clearly still needed; just as it seems to be the case for inference on r, one would expect foreground uncertainty to be the limiting factor for inference on the $\zeta\zeta h$ 3-point correlation function.

B. Secondaries sourced by ζ

We now consider non-Galactic secondary non-Gaussian signals that are sourced by the curvature perturbation ζ (as opposed to h). We again focus on squeezed bispectra with a large wavelength B-mode, as such bispectra may bias the inference on the primordial signal.

The most well-studied secondary signal is sourced by the correlation between the late-time ISW effect and the lensing potential [135]. A similar correlation exists between the quadrupole perturbation that sources the polarized reionization signal and the lensing potential [136]. The ISW effect and the polarization generated at reionization only affect the CMB over large angular scales. On the other hand, the lensing potential modulates small-scale power. The associated bispectra are thus of the squeezed type. The ISW effect only affects the temperature anisotropies, and the polarized reionization signal is purely E. This means that although $\langle TTB\rangle$, $\langle TEB\rangle$, $\langle EEB\rangle$ bispectra are produced [136,137], the only significant configurations will have large-scale T- or E-mode components instead of a B-mode component.

In general, the requirement of a squeezed bispectrum with a large-scale B-mode contribution is highly constraining. There are no obvious (non-Galactic) candidates that preferentially source a B-mode signal on large angular scales. Nonlinear effects other than lensing that produce the B-mode signal, such as patchy reionization [138] and the polarized Sunyaev Zel'dovich (PSZ) effect [139,140], do so only at relatively small angular scales. Unclustered, extragalactic point sources may be weakly polarized and have a reduced bispectrum that is approximately constant with multipole order [141]. They thus contaminate all bispectra, regardless of shape. However, especially for squeezed models, the point-source bias is found to be negligible: the two types of bispectra can be estimated independently [15,141].

C. Secondaries sourced by *h*

The $\zeta\zeta h$ 3-point correlation function is contingent upon the existence of the primordial tensor perturbation h. For completeness, we thus briefly discuss a possible secondary non-Gaussian signal sourced by a purely Gaussian tensor perturbation h.

In this case, the most obvious single-B-mode bispectrum candidate will be due to the interplay between two effects: (1) the standard correlation between the lensing potential ϕ and the ISW and polarized reionization signal, together with (2) the fact that lensing will now convert some of the (Gaussian) primordial B-mode signal to E-mode polarization [142]. In the resulting $\langle BEX\rangle$ bispectrum, B is the standard primordial B-mode signal, E is the primordial B-mode signal lensed to an E-mode signal, and X is the standard scalar-induced T or E signal. To first order in the lensing potential, the bispectrum should be given by the triangular configurations of $C_\ell^{BB}C_{\ell'}^{\phi X}$. The suppression by r, due to the presence of the primordial B-mode power spectrum, makes this bispectrum lower in amplitude than the standard lensing-ISW bispectrum discussed in Sec. V B. More importantly however, the fact that the lensing-ISW and lensing-reionization correlation $C_\ell^{\phi X}$ is only nonzero for $\ell \lesssim 100$ [136] means that there will be no significant bispectrum configurations with a large-scale B-mode component and two small-scale ($\ell > 100$) T and/or E components: the relevant configuration for a bias.

Analogous to the E-mode-lensing correlation in Sec. V B, the B-mode signal from reionization, present when $r \neq 0$, is also correlated to small-scale power through a correlation with the lensed signal. The difference is that isotropy and parity invariance forbid a correlation between B and the regular gradient-type lensing potential. Instead the B-mode signal is correlated to the curl-type lensing potential sourced by the h perturbation [143,144]. Unlike the $\zeta\zeta\zeta$ case, there will now exist $\langle BXX'\rangle$ bispectra, where B is the unlensed B-mode field and X the curl-lensed T- or E-mode field (and $X' \in \{T, E\}$). To leading order we expect such bispectra to be proportional to the triangular configurations of $C_\ell^{B\omega}C_{\ell'}^{XX'}$, where $C_\ell^{B\omega}$ is the cross-correlation between the curl component of the lensing deflection angle and the reionization B signal. The power spectrum of the tensor-induced ω, i.e., $C_\ell^{\omega\omega}$, is strongly suppressed compared to scalar-induced lensing and decays rapidly for $\ell > 2$ [143,144]. One would expect similar behavior for the amplitude of $C_\ell^{B\omega}$ and thus expect that $C_\ell^{B\omega}C_{\ell'}^{XX'}$ is negligible. Still, the associated bispectra are maximized in the squeezed limit with a large-scale B-mode, so they should be considered as a potential bias to a primordial signal.

The tensor-induced temperature quadrupole on the last-scattering surface seen by galaxy clusters will source the PSZ effect [145,146]. The resulting small-scale power will be correlated with the primary B-mode field from reionization and will thus source a squeezed $\langle BEE\rangle$ bispectrum (among others). The B-mode component is on large angular scales, which means that the bispectrum has the right shape to be a potentially relevant contaminant of the primordial bispectrum.

Finally, second-order perturbation theory predicts that a correlation between short-wavelength scalar modes and primordial tensor modes emerges as the latter reenter the horizon during matter domination. This occurs regardless of any primordial $\zeta\zeta h$ correlation [147]. The correlation is usually studied as a quadrupole asymmetry in the matter/galaxy power spectrum, but can be understood as a squeezed $\zeta\zeta h$ 3-point correlation. Consequently, this second order effect is, in principle, imprinted in the CMB bispectra that we are interested in, but the imprint is likely too small to be observable.

D. Contributions to the covariance

In the previous three sections, we focused on possible biases to the estimator. All discussed effects will also contribute to the covariance of the estimate. Fortunately, in most cases these effects are subdominant to the Gaussian contribution to the covariance, given by the inverse of

Eq. (30). However, as we illustrate in Appendix D, the covariance of the estimator receives additional contributions from any connected 4- and 6-point correlation function present in the data. For example, for the $\zeta\zeta\zeta$, temperature-only bispectrum, the connected moments due to lensing will introduce significant additional covariance on small angular scales. The variance due to the connected 4-point function alone is expected to dominate the cosmic-variance induced estimator variance for local-type non-Gaussianity for $\ell_{\max} \gtrsim 3500$ [148] (and hence will be a concern for experiments such as Simons Observatory and CMB-S4). The total effect on the estimator covariance will depend on the shape of the primordial bispectra that are estimated: local, or squeezed, shapes will likely be affected the most.

We focus primarily on bispectra with a single *B*-mode component; in the previous sections, we argued that such bispectra are less susceptible to secondary biases. However, this argument does not hold for the variance of the estimator: when lensing is introduced, it is expected that the estimator covariance is affected in a way that is rather similar to the temperature-only case mentioned above. For example, consider the $\langle TTB\rangle$ bispectrum; the variance of its estimate will be approximately proportional to the $\langle TTTTBB\rangle$ 6-point function. In the noiseless Gaussian case, this 6-point function reduces to terms proportional to $C_\ell^{TT} C_{\ell'}^{TT} C_{\ell''}^{BB}$. When lensing is introduced, the power spectra are replaced by their lensed versions (which has a large effect on C_ℓ^{BB}). However, there should also be a contribution proportional to the connected $\langle TTTT\rangle$ 4-point function from lensing. One would expect this contribution to saturate the constraining power for $\ell_{\max} \gtrsim 3500$, just as it does for the temperature-only case mentioned above. For the variance on estimates using the $\langle EEB\rangle$ or $\langle TEB\rangle$ bispectra a similar argument applies [131]. In other words, we expect that an estimate of the $\zeta\zeta h$ 3-point function using high-resolution data will have large non-Gaussian contributions to its (co)variance, at least for squeezed bispectrum shapes with a *B*-mode contribution on large angular scales.[21] Note that this non-Gaussian contribution to the variance is not included in the Fisher forecasts presented in Sec. IV.

In a future study we hope to identify all these contributions to the covariance and estimate their effects on our ability to extract the primordial signal. We note that, in principle, secondary biases and non-Gaussian contributions to the covariance from lensing can likely be reduced significantly by delensing [149]. As some of the contributions to the covariance might be hard to compute analytically, applying the developed estimator on a suite of realistically lensed simulations would be an important aspect of such a study.

VI. CONCLUSIONS

The CMB bispectrum sourced by primordial scalar-tensor interactions is a well-defined observable that can be probed effectively with upcoming CMB polarization data. Inference on these types of primordial interactions probes nonstandard early-Universe models that are essentially unconstrained by current studies. In addition, inference on the squeezed $\zeta\zeta h$ 3-point function provides a powerful consistency test of the standard inflationary paradigm.

In this work, we derived a numerically efficient and optimal estimator for the amplitude of CMB bispectra sourced by primordial $\zeta\zeta h$ 3-point correlation functions. We demonstrated that despite the intrinsic geometrical complexity of the bispectrum, an efficient estimator can be formulated; see Eq. (59). There is a limited computational overhead compared to standard $\zeta\zeta\zeta$ bispectrum estimation [see Eq. (44)], but the same asymptotic scaling with data resolution is reached. The derived estimator provides complementarity to the more general modal and binned bispectrum estimators [84–87,90] and should, due to its numerical advantage, be the preferred method for high-resolution data.

We studied the bispectrum sourced by a squeezed $\zeta\zeta h$ 3-point function in more detail. We presented a set of Fisher forecasts that form a baseline to which more realistic forecasts will be compared in future work. The presented forecasts demonstrate a relatively strong dependence on the size of the largest angular scale accessible in the data. We also demonstrated how constraints from the combination of temperature, *E*- and *B*-mode data are significantly better than those only from temperature and *B*-mode data or only from *E*- and *B*-mode data. Finally, we found that the lensing contribution to the *B*-mode data starts to significantly impact the constraints from experiments such as the Simons Observatory and *LiteBIRD*. For a more futuristic experiment like CMB-S4, delensing of the large-scale *B*-mode data will be crucial.

Although the Fisher forecasts provide us with a good indication of the ultimate constraining power of future CMB experiments, future forecasts will need to include more realism. This requires applying the estimator directly to simulated sky maps. Besides allowing the characterization of standard complications such as nontrivial noise properties and sky cuts, this approach is the appropriate way to study effects that are more specific to, e.g., the $\zeta\zeta h$ bispectrum. Examples of such effects include the incomplete removal of Galactic *B*-mode signal or non-Gaussian polarized secondary sources. Lensed sky simulations will also allow one to quantify the expected extra estimator variance due to non-Gaussian 4- and 6-point correlation

[21] Because the effect should only become dominant for $\ell \gtrsim 3500$, there should be a negligible effect on primordial bispectra with more than one *B*-mode component and/or shapes that are more equilateral. In these cases, the signal drops sharply for $\ell_B \gtrsim 200$.

functions in the lensed CMB fields, as well as the effects of delensing these fields. Although current data are inconclusive, it seems likely that the eventual limit on future constraints will be from foreground uncertainty on large angular scales and the non-Gaussian lensing contribution on small scales. Before this point is reached, however, the data will contain a large amount of unexplored cosmological information. With an efficient estimator in hand, we should now turn towards map-based simulations to predict the exact amount of information.

In the next decade, we will significantly improve our measurements of the CMB polarization field. With this in mind, we should consider interesting science targets beyond the tensor-to-scalar ratio that can provide insight into the early Universe. One of these targets is probing the primordial interactions between scalars and tensors as well as tensor self-interactions. Currently, the most sensitive probe of these interactions comes from including the B-mode field into CMB bispectrum inference. The work presented here is a contribution toward the development of a complete framework to constrain these interactions with upcoming CMB data.

ACKNOWLEDGMENTS

The authors thank Joel Meyers, Alex van Engelen, William Coulton, Eiichiro Komatsu, and Sebastian Baum for useful discussions and thank an anonymous referee for helpful comments and suggestions. P. D. M. acknowledges support from Senior Kavli Institute Fellowships at the University of Cambridge and the Netherlands organization for scientific research (NWO) VIDI Grant (dossier 639.042.730). A. J. D. and K. F. acknowledge support by the Vetenskapsrådet (Swedish Research Council) through Contract No. 638-2013-8993 and the Oskar Klein Centre for Cosmoparticle Physics. K. F. acknowledges support from the Jeff and Gail Kodosky Endowed Chair in Physics, DOE Award No. DE-SC007859, and the LCTP at the University of Michigan. Some computations have been performed at the Owl Cluster funded by the University of Oslo and the Research Council of Norway through Grant No. 250672.

APPENDIX A: ESTIMATOR FOR OTHER ANGULAR TERMS

In this appendix, we show how the $\hat{f}_{\mathrm{NL}}$ estimator for 3-point functions with other angular terms. Besides providing a few useful examples, it can be seen how each estimator still asymptotically scales as $\mathcal{O}(\ell_{\max}^3)$. For each template we show the expression for the bispectrum and the cubic part of the estimator. As demonstrated in Sec. III C 4, it is straightforward to derive the linear term of the estimator given the cubic term.

1. Scalar-scalar-scalar

a. Standard scalar-only template

For comparison and completeness, we first treat the standard $\zeta\zeta\zeta$ template, i.e., a template with no contracted angular term. Assuming a shape template such as Eq. (43), the expression for the bispectrum in Eq. (21) simplifies to

$$B^{\ell_1\ell_2\ell_3(\zeta\zeta\zeta)}_{m_1m_2m_3X_1X_2X_3} = \frac{1}{6}\int_{S^2} \mathrm{d}\Omega(\hat{\mathbf{n}}) \sum_{i=1}^{N_{\mathrm{prim}}} \int_0^\infty r^2 \mathrm{d}r \sum_{\ell_1,m_1} [(\mathcal{K}^{(\zeta)}_{(X_1)}[f^{(i)}])_{\ell_1,\ell_1}(r)] Y_{\ell_1 m_1}(\hat{\mathbf{n}}) \times \sum_{\ell_2,m_2} [(\mathcal{K}^{(\zeta)}_{(X_2)}[g^{(i)}])_{\ell_2,\ell_2}(r)] Y_{\ell_2 m_2}(\hat{\mathbf{n}}) \sum_{\ell_3,m_3} [(\mathcal{K}^{(\zeta)}_{(X_3)}[h^{(i)}])_{\ell_3,\ell_3}(r)] Y_{\ell_3 m_3}(\hat{\mathbf{n}}) + (5\ \mathrm{perm}). \tag{A1}$$

Note that the five extra terms are permutations of the input functions f, g, and h. With this bispectrum, the cubic term of the estimator becomes

$$\hat{f}^{\zeta\zeta\zeta}_{\mathrm{NL,cubic}} = \frac{1}{6\mathcal{I}_0} \int_{S^2} \mathrm{d}\Omega(\hat{\mathbf{n}}) \sum_{i=1}^{N_{\mathrm{prim}}} \int_0^\infty r^2 \mathrm{d}r (\mathcal{A}^{(\zeta)}_{(0,0)}[f^{(i)}] \mathcal{A}^{(\zeta)}_{(0,0)}[g^{(i)}] \mathcal{A}^{(\zeta)}_{(0,0)}[h^{(i)}])(r,\hat{\mathbf{n}}), \tag{A2}$$

which is the standard result [80], but rephrased in our notation. See Eqs. (60) and (61) for the definition of the $\mathcal{A}_{(S,n)}$ functionals. In the $(S,n) = (0,0)$ case used here, the functionals are much less complicated: the 3-j symbols reduce to a delta function which simplifies the expression to

$$\mathcal{A}^{(\zeta)}_{(0,0)}[f](r,\hat{\mathbf{n}}) = \sum_{\ell,m} (-1)^\ell \sum_{X\in\{T,E\}} (\mathcal{K}^{(\zeta)}[f])^X_{\ell,\ell}(r) (C^{-1}a)^X_{\ell m} Y_{\ell m}(\hat{\mathbf{n}}). \tag{A3}$$

Note that the $(-1)^\ell$ factors are not present in the original expression [80]. They do not change the estimator, as only configurations with $\ell_1 + \ell_2 + \ell_3 =$ even contribute. The $\mathcal{K}$ functionals are defined in Eq. (52).

b. Scalar-only template with angular dependence of massive spinning particles

The second $\zeta\zeta\zeta$ template is inspired by the three-point function template derived in Ref. [31]. The template captures the imprint of a massive spin-s field during inflation. Although the template only involves the curvature perturbation, it does include a contracted angular term

$$ {}^{(000)}F(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3) = \frac{1}{6} f^{(\zeta\zeta\zeta)}(k_1,k_2,k_3) P_s(\hat{\mathbf{k}}_2\cdot\hat{\mathbf{k}}_3) + (5\text{ perm}), \tag{A4} $$

$$ = \frac{1}{6}\sum_{i=1}^{N_{\text{prim}}} f^{(i)}(k_1) g^{(i)}(k_2) h^{(i)}(k_3) P_s(\hat{\mathbf{k}}_2\cdot\hat{\mathbf{k}}_3) + (5\text{ perm}). \tag{A5} $$

P_s is a Legendre polynomial of degree s. The five additional permutations are permutations of the three wave vectors.

In order to write down the corresponding bispectrum, we expand the Legendre polynomial in terms of spherical harmonics:

$$ P_s(\hat{\mathbf{k}}\cdot\hat{\mathbf{k}}') = \frac{4\pi}{2s+1}\sum_{m'=-s}^{s} Y_{sm'}(\hat{\mathbf{k}}) Y^*_{sm'}(\hat{\mathbf{k}}'). \tag{A6} $$

The bispectrum for a spin-s template then becomes

$$ \begin{aligned} B^{\ell_1\ell_2\ell_3(\zeta\zeta\zeta)}_{m_1m_2m_3X_1X_2X_3} = {} & \frac{4\pi}{6(2s+1)}\sum_{m'=-s}^{s}(-1)^{m'}\int_{S^2}\mathrm{d}\Omega(\hat{\mathbf{n}})\sum_{i=1}^{N_{\text{prim}}}\int_0^\infty r^2\mathrm{d}r[(\mathcal{K}^{(\zeta)}_{(X_1)}[f^{(i)}])_{\ell_1,\ell_1}(r)]Y_{\ell_1m_1}(\hat{\mathbf{n}}) \\ & \times\sum_{L_2,M_2}\left[i^{\ell_2+L_2}J^{000}_{sL_2\ell_2}\begin{pmatrix} s & L_2 & \ell_2 \\ -m' & M_2 & m_2\end{pmatrix}(\mathcal{K}^{(\zeta)}_{(X_2)}[g^{(i)}])_{\ell_2,L_2}(r)\right]Y_{L_2M_2}(\hat{\mathbf{n}}) \\ & \times\sum_{L_3,M_3}\left[i^{\ell_3+L_3}J^{000}_{sL_3\ell_3}\begin{pmatrix} s & L_3 & \ell_3 \\ m' & M_3 & m_3\end{pmatrix}(\mathcal{K}^{(\zeta)}_{(X_3)}[h^{(i)}])_{\ell_3,L_3}(r)\right]Y_{L_3M_3}(\hat{\mathbf{n}}) \\ & + (5\text{ perm}). \end{aligned} \tag{A7} $$

The five additional terms are obtained by simultaneously permuting $f^{(i)}$, $g^{(i)}$, and $h^{(i)}$ with the 1, 2, and 3 indices. The cubic term of the estimator for this bispectrum is given by

$$ \begin{aligned} \hat{f}^{\zeta\zeta\zeta}_{\text{NL,cubic}} = {} & \frac{1}{18\mathcal{I}_0}\frac{4\pi}{2s+1}\sum_{m'=-s}^{s}(-1)^{m'}\int_{S^2}\mathrm{d}\Omega(\hat{\mathbf{n}})\sum_{i=1}^{N_{\text{prim}}}\int_0^\infty r^2\mathrm{d}r(\mathcal{A}^{(\zeta)}_{(0,0)}[f^{(i)}]\mathcal{A}^{(\zeta)}_{(s,-m')}[g^{(i)}]\mathcal{A}^{(\zeta)}_{(s,m')}[h^{(i)}])(r,\hat{\mathbf{n}}) \\ & + (2\text{ cyclic}). \end{aligned} \tag{A8} $$

The two extra terms are cyclic permutations of $f^{(i)}$, $g^{(i)}$, and $h^{(i)}$.

2. Scalar-tensor-tensor

To illustrate the situation for a scalar-tensor-tensor 3-point function, we use a template inspired by the SFSR result [18]:

$$ {}^{(0\lambda_2\lambda_3)}F(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3) = f^{(\zeta hh)}(k_1,k_2,k_3)e^{\lambda_2}_{ab}(\hat{\mathbf{k}}_2)e^{ab}_{\lambda_3}(\hat{\mathbf{k}}_3) \tag{A9} $$

$$ = \sum_{i=1}^{N_{\text{prim}}} f^{(i)}(k_1)g^{(i)}(k_2)h^{(i)}(k_3)e^{\lambda_2}_{ab}(\hat{\mathbf{k}}_2)e^{ab}_{\lambda_3}(\hat{\mathbf{k}}_3). \tag{A10} $$

The polarization tensors $e^{\pm 2}$ are defined in Eqs. (13) and (14), the a and b indices run over the three spatial dimensions. We use the Einstein summation convention. Using the notation from Ref. [83], we may expand the polarization tensors as:

$$e^{\pm 2}_{ab} = \frac{3}{\sqrt{2\pi}} \sum_{M,m_a,m_b} {}_{\mp 2}Y^*_{2M}\, \alpha_a^{m_a}\alpha_b^{m_b} \begin{pmatrix} 2 & 1 & 1 \\ M & m_a & m_b \end{pmatrix}. \tag{A11}$$

The α coefficients obey the following orthogonality relation:

$$\alpha^m_a \alpha^b_{m'} = \frac{4\pi}{3}(-1)^m \delta^{-m'}_m. \tag{A12}$$

Using this relation together with the orthogonality relation of the Wigner 3-j symbols in Eq. (B13), the contraction of two polarization tensors can be expressed as follows:

$$e^\lambda_{ab}(\hat{\mathbf{k}})\, e^{ab}_{\lambda'}(\hat{\mathbf{k}}') = \frac{8\pi}{5}\sum_{m'=-2}^{2}(-1)^{m'}\, {}_{-\lambda}Y^*_{2-m'}(\hat{\mathbf{k}})\, {}_{-\lambda'}Y^*_{2m'}(\hat{\mathbf{k}}'). \tag{A13}$$

The bispectrum corresponding to the template in Eq. (A10) thus becomes

$$\begin{aligned} B^{\ell_1\ell_2\ell_3(\zeta hh)}_{m_1m_2m_3X_1X_2X_3} = {} & \frac{8\pi}{5}\sum_{m'=-2}^{2}(-1)^{m'}\int_{S^2}\mathrm{d}\Omega(\hat{\mathbf{n}})\sum_{i=1}^{N_{\mathrm{prim}}}\int_0^\infty r^2\mathrm{d}r[(\mathcal{K}^{(\zeta)}_{(X_1)}[f^{(i)}])_{\ell_1,\ell_1}(r)]Y_{\ell_1 m_1}(\hat{\mathbf{n}}) \\ & \times \sum_{L_2,M_2}\left[i^{\ell_2+L_2}J^{-202}_{2L_2\ell_2}[1+(-1)^{x_2+L_2+\ell_2}]\begin{pmatrix} 2 & L_2 & \ell_2 \\ -m' & M_2 & m_2 \end{pmatrix}(\mathcal{K}^{(h)}_{(X_2)}[g^{(i)}])_{\ell_2,L_2}(r)\right]Y_{L_2M_2}(\hat{\mathbf{n}}) \\ & \times \sum_{L_3,M_3}\left[i^{\ell_3+L_3}J^{-202}_{2L_3\ell_3}[1+(-1)^{x_3+L_3+\ell_3}]\begin{pmatrix} 2 & L_3 & \ell_3 \\ m' & M_3 & m_3 \end{pmatrix}(\mathcal{K}^{(h)}_{(X_3)}[h^{(i)}])_{\ell_3,L_3}(r)\right]Y_{L_3M_3}(\hat{\mathbf{n}}). \end{aligned} \tag{A14}$$

The cubic part of the estimator is given by

$$\hat{f}^{\zeta hh}_{\mathrm{NL,cubic}} = \frac{4\pi}{15\mathcal{I}_0}\sum_{m'=-2}^{2}(-1)^{m'}\int_{S^2}\mathrm{d}\Omega(\hat{\mathbf{n}})\sum_{i=1}^{N_{\mathrm{prim}}}\int_0^\infty r^2\mathrm{d}r(\mathcal{A}^{(\zeta)}_{(0,0)}[f^{(i)}]\mathcal{A}^{(h)}_{(2,-m')}[g^{(i)}]\mathcal{A}^{(h)}_{(2,m')}[h^{(i)}])(r,\hat{\mathbf{n}}). \tag{A15}$$

The expressions for the $h\zeta h$ and $hh\zeta$ parts are derived in an analogous way.

3. Tensor-tensor-tensor

Finally, we derive the estimator for a tensor-tensor-tensor 3-point function. We again take the SFSR prediction [18] as inspiration for our template:

$$\begin{aligned} {}^{(\lambda_1\lambda_2\lambda_3)}F(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3) = {} & f^{(hhh)}(k_1,k_2,k_3) \\ & \times[\hat{k}^a_2\hat{k}^b_2 e^{\lambda_1}_{ab}(\hat{\mathbf{k}}_1)e^{cd}_{\lambda_2}(\hat{\mathbf{k}}_2)e^{\lambda_3}_{cd}(\hat{\mathbf{k}}_3) - 2e^{\lambda_1}_{ab}(\hat{\mathbf{k}}_1)e^{\lambda_2}_{cd}(\hat{\mathbf{k}}_2)e^{bc}_{\lambda_3}(\hat{\mathbf{k}}_3)\hat{k}^a_2\hat{k}^d_3] + (2\text{ cyclic}) \end{aligned} \tag{A16}$$

$$\begin{aligned} = {} & \sum_{i=1}^{N_{\mathrm{prim}}} f^{(i)}(k_1)g^{(i)}(k_2)h^{(i)}(k_3) \\ & \times[\hat{k}^a_2\hat{k}^b_2 e^{\lambda_1}_{ab}(\hat{\mathbf{k}}_1)e^{cd}_{\lambda_2}(\hat{\mathbf{k}}_2)e^{\lambda_3}_{cd}(\hat{\mathbf{k}}_3) - 2e^{\lambda_1}_{ab}(\hat{\mathbf{k}}_1)e^{\lambda_2}_{cd}(\hat{\mathbf{k}}_2)e^{bc}_{\lambda_3}(\hat{\mathbf{k}}_3)\hat{k}^a_2\hat{k}^d_3] + (2\text{ cyclic}). \end{aligned} \tag{A17}$$

The two extra terms are cyclic permutations of the three wave vectors.

To derive the bispectrum, we need to expand the unit wave vectors in spherical harmonics [83]:

$$\hat{k}^a = \sum_m \alpha^a_m Y_{1m}(\hat{\mathbf{k}}). \tag{A18}$$

The α coefficients obey the relation in Eq. (A12). Together with Eqs. (A11), (B8), and (B13) we then expand the first angular term in Eq. (A17) as follows:

$$\hat{k}_2^a\hat{k}_2^b e_{ab}^{\lambda_1}(\hat{\mathbf{k}}_1)e_{\lambda_2}^{cd}(\hat{\mathbf{k}}_2)e_{cd}^{\lambda_3}(\hat{\mathbf{k}}_3) = \frac{64\pi^2}{75}\sum_{L,M,M',M''} J_{22L}^{0-\lambda_2\lambda_2}\begin{pmatrix} 2 & 2 & L \\ M & M'' & M' \end{pmatrix} {}_{-\lambda_1}Y_{2M}^*(\hat{\mathbf{k}}_1)\,{}_{-\lambda_2}Y_{LM'}^*(\hat{\mathbf{k}}_2)\,{}_{-\lambda_3}Y_{2M''}^*(\hat{\mathbf{k}}_3). \tag{A19}$$

The J symbols are defined in Eq. (B9). The second angular term in Eq. (A17) is expressed in terms of Wigner 6-j symbols by making use of the relation in Eq. (B16), see also Ref. [150]. The resulting expression is:

$$\begin{aligned} e_{ab}^{\lambda_1}(\hat{\mathbf{k}}_1)e_{cd}^{\lambda_2}(\hat{\mathbf{k}}_2)e_{\lambda_3}^{bc}(\hat{\mathbf{k}}_3)\hat{k}_2^a\hat{k}_3^d = {} & \frac{(8\pi)^{5/2}}{6}\sum_{\substack{L,J\\ M,M',M''}}(-1)^{L+1}J_{21L}^{\lambda_2 0-\lambda_2}J_{21J}^{\lambda_3 0-\lambda_3}\begin{pmatrix} J & L & 2 \\ M'' & M' & M \end{pmatrix}\begin{Bmatrix} J & L & 2 \\ 1 & 1 & 2 \end{Bmatrix}\begin{Bmatrix} 1 & 2 & J \\ 1 & 2 & 1 \end{Bmatrix} \\ & \times {}_{-\lambda_1}Y_{2M}^*(\hat{\mathbf{k}}_1)\,{}_{-\lambda_2}Y_{LM'}^*(\hat{\mathbf{k}}_2)\,{}_{-\lambda_3}Y_{JM''}^*(\hat{\mathbf{k}}_3). \end{aligned} \tag{A20}$$

It is convenient to separate the corresponding bispectrum into a part sourced by the first angular term and a part sourced by the second term. The first part is given by

$$\begin{aligned} B_{m_1m_2m_3X_1X_2X_3}^{\ell_1\ell_2\ell_3(hhh,1)} = {} & \frac{64\pi^2}{75}\sum_{L,M,M',M''}\begin{pmatrix} 2 & 2 & L \\ M & M'' & M' \end{pmatrix}\int_{S^2}\mathrm{d}\Omega(\hat{\mathbf{n}})\sum_{i=1}^{N_{\mathrm{prim}}}\int_0^\infty r^2\mathrm{d}r \\ & \times\sum_{L_1,M_1}\left[i^{\ell_1+L_1}J_{2L_1\ell_1}^{-202}[1+(-1)^{x_1+L_1+\ell_1}]\begin{pmatrix} 2 & L_1 & \ell_1 \\ M & M_1 & m_1 \end{pmatrix}(\mathcal{K}_{(X_1)}^{(h)}[f^{(i)}])_{\ell_1,L_1}(r)\right]Y_{L_1M_1}(\hat{\mathbf{n}}) \\ & \times\sum_{L_2,M_2}\left[i^{\ell_2+L_2}J_{2L_2\ell_2}^{-202}J_{22L}^{02-2}[1+(-1)^{x_2+L_2+\ell_2+L}]\begin{pmatrix} L & L_2 & \ell_2 \\ M' & M_2 & m_2 \end{pmatrix}(\mathcal{K}_{(X_2)}^{(h)}[g^{(i)}])_{\ell_2,L_2}(r)\right]Y_{L_2M_2}(\hat{\mathbf{n}}) \\ & \times\sum_{L_3,M_3}\left[i^{\ell_3+L_3}J_{2L_3\ell_3}^{-202}[1+(-1)^{x_3+L_3+\ell_3}]\begin{pmatrix} 2 & L_3 & \ell_3 \\ M'' & M_3 & m_3 \end{pmatrix}(\mathcal{K}_{(X_3)}^{(h)}[h^{(i)}])_{\ell_3,L_3}(r)\right]Y_{L_3M_3}(\hat{\mathbf{n}}) \\ & + (2\text{ cyclic}). \end{aligned} \tag{A21}$$

The two extra terms are given by cyclic permutations of the $f^{(i)}$, $g^{(i)}$, and $h^{(i)}$ input functions together with the 1, 2, and 3 indices. The second part is given by:

$$\begin{aligned} B_{m_1m_2m_3X_1X_2X_3}^{\ell_1\ell_2\ell_3(hhh,2)} = {} & \frac{(8\pi)^{5/2}}{3}\sum_{\substack{L,J\\ M,M',M''}}(-1)^{L+1}\begin{pmatrix} J & L & 2 \\ M'' & M' & M \end{pmatrix}\begin{Bmatrix} J & L & 2 \\ 1 & 1 & 2 \end{Bmatrix}\begin{Bmatrix} 1 & 2 & J \\ 1 & 2 & 1 \end{Bmatrix}\int_{S^2}\mathrm{d}\Omega(\hat{\mathbf{n}})\sum_{i=1}^{N_{\mathrm{prim}}}\int_0^\infty r^2\mathrm{d}r \\ & \times\sum_{L_1,M_1}\left[i^{\ell_1+L_1}J_{2L_1\ell_1}^{-202}[1+(-1)^{x_1+L_1+\ell_1}]\begin{pmatrix} 2 & L_1 & \ell_1 \\ M & M_1 & m_1 \end{pmatrix}(\mathcal{K}_{(X_1)}^{(h)}[f^{(i)}])_{\ell_1,L_1}(r)\right]Y_{L_1M_1}(\hat{\mathbf{n}}) \\ & \times\sum_{L_2,M_2}\left[i^{\ell_2+L_2}J_{2L_2\ell_2}^{-202}J_{21L}^{-202}[1+(-1)^{x_2+L_2+\ell_2+L+1}]\begin{pmatrix} L & L_2 & \ell_2 \\ M' & M_2 & m_2 \end{pmatrix}(\mathcal{K}_{(X_2)}^{(h)}[g^{(i)}])_{\ell_2,L_2}(r)\right]Y_{L_2M_2}(\hat{\mathbf{n}}) \\ & \times\sum_{L_3,M_3}\left[i^{\ell_3+L_3}J_{2L_3\ell_3}^{-202}J_{21J}^{-202}[1+(-1)^{x_3+L_3+\ell_3+J+1}]\begin{pmatrix} J & L_3 & \ell_3 \\ M'' & M_3 & m_3 \end{pmatrix}(\mathcal{K}_{(X_3)}^{(h)}[h^{(i)}])_{\ell_3,L_3}(r)\right]Y_{L_3M_3}(\hat{\mathbf{n}}) \\ & + (2\text{ cyclic}). \end{aligned} \tag{A22}$$

The cubic estimator is also most easily expressed in two parts. The part corresponding to the first bispectrum, Eq. (A21), is given by

$$\begin{aligned} \hat{f}_{\mathrm{NL,cubic}}^{hhh,1} = {} & \frac{32\pi^2}{225\mathcal{I}_0}\sum_{L,M,M',M''}\begin{pmatrix} 2 & 2 & L \\ M & M'' & M' \end{pmatrix}J_{22L}^{02-2}\int_{S^2}\mathrm{d}\Omega(\hat{\mathbf{n}})\sum_{i=1}^{N_{\mathrm{prim}}}\int_0^\infty r^2\mathrm{d}r \\ & \times(\mathcal{A}_{(2,M)}^{(h)}[f^{(i)}]\mathcal{B}_{(L,M')}^{(h)}[g^{(i)}]\mathcal{A}_{(2,M'')}^{(h)}[h^{(i)}])(r,\hat{\mathbf{n}}) + (2\text{ cyclic}). \end{aligned} \tag{A23}$$

The two extra terms are cyclic permutations of $f^{(i)}$, $g^{(i)}$, and $h^{(i)}$. The second part, corresponding to Eq. (A21), is given by

$$\hat{f}^{hhh,2}_{\mathrm{NL,cubic}} = \frac{(8\pi)^{5/2}}{18\mathcal{I}_0} \sum_{\substack{L,J \\ M,M',M''}} (-1)^{L+1} \begin{pmatrix} J & L & 2 \\ M'' & M' & M \end{pmatrix} J^{-202}_{21L} J^{-202}_{21J} \begin{Bmatrix} J & L & 2 \\ 1 & 1 & 2 \end{Bmatrix} \begin{Bmatrix} 1 & 2 & J \\ 1 & 2 & 1 \end{Bmatrix}$$
$$\times \int_{S^2} \mathrm{d}\Omega(\hat{\mathbf{n}}) \sum_{i=1}^{N_{\mathrm{prim}}} \int_0^\infty r^2 \mathrm{d}r (\mathcal{A}^{(h)}_{(2,M)}[f^{(i)}] \mathcal{C}^{(h)}_{(L,M')}[g^{(i)}] \mathcal{C}^{(h)}_{(J,M'')}[h^{(i)}])(r, \hat{\mathbf{n}}) + (2\ \text{cyclic}). \tag{A24}$$

We have introduced the $\mathcal{B}$ and $\mathcal{C}$ functionals. They are completely analogous to the $\mathcal{A}$ functionals, defined in Eqs. (60) and (61), but slightly differ in their spherical harmonic coefficients:

$$(\mathcal{B}^{(h)}_{(S,n)}[f])_{LM}(r) \equiv (4\pi)^{1/2} \sum_{\ell,m} i^{\ell+L} J^{-202}_{SL\ell} \begin{pmatrix} S & L & \ell \\ n & M & m \end{pmatrix} \sum_X [1 + (-1)^{x+L+\ell+S}] (\mathcal{K}^{(h)}[f])^X_{\ell,L}(r) (C^{-1}a)^X_{\ell m}, \tag{A25}$$

$$(\mathcal{C}^{(h)}_{(S,n)}[f])_{LM}(r) \equiv (4\pi)^{1/2} \sum_{\ell,m} i^{\ell+L} J^{-202}_{SL\ell} \begin{pmatrix} S & L & \ell \\ n & M & m \end{pmatrix} \sum_X [1 + (-1)^{x+L+\ell+S+1}] (\mathcal{K}^{(h)}[f])^X_{\ell,L}(r) (C^{-1}a)^X_{\ell m}. \tag{A26}$$

Computing the combination of Eqs. (A23) and (A24) will still asymptotically scale as $\mathcal{O}(\ell^3_{\mathrm{max}})$. Although more terms have to be computed compared to the previous templates, this computational overhead is easily outweighed by the fact that the h transfer functions impose $\ell_{\mathrm{max}} \approx 200$.

APPENDIX B: USEFUL MATHEMATICAL IDENTITIES

1. Spin-weighted spherical harmonics

The spin-weighted spherical harmonics (SWSHs) ${}_sY_{\ell m}$ are generalizations of the standard spherical harmonics $Y_{\ell m}$. Both types of spherical harmonics are functions on the sphere S^2. Indeed, one may relate

$${}_0Y_{\ell m} = Y_{\ell m}. \tag{B1}$$

The relation between the two sets of functions for nonzero s can be found in the literature [151,152].

The SWSHs are conveniently defined on the standard spherical coordinate system by taking the Wigner D-matrices (irreps of the three-dimensional rotation group) parametrized in terms of the Euler angles and fixing the polar axis as follows:

$${}_sY_{\ell m}(\theta, \phi) = (-1)^m \sqrt{\frac{2\ell+1}{4\pi}} D^{\ell}_{-ms}(\phi, \theta, \psi)\bigg|_{\psi=0}. \tag{B2}$$

With a slight abuse of notation, we use $\hat{\mathbf{n}}$ in the arguments of the spherical harmonics to refer to the θ and ϕ angles that describe the spherical decomposition of the 3D unit vector, i.e., $\hat{\mathbf{n}} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$. Similarly, we denote the differential solid angle with $\mathrm{d}\Omega(\hat{\mathbf{n}})$, i.e., $\int_{S^2} \mathrm{d}\Omega(\hat{\mathbf{n}}) \equiv \int_0^{2\pi} \mathrm{d}\phi \int_0^{\pi} \mathrm{d}\theta \sin\theta$.

The functions form an orthonormal and complete system for each integer[22] spin weight s:

$$\int_{S^2} \mathrm{d}\Omega(\hat{\mathbf{n}})\, {}_sY_{\ell m}(\hat{\mathbf{n}})\, {}_sY^*_{\ell' m'}(\hat{\mathbf{n}}) = \delta_{\ell\ell'}\delta_{mm'}, \tag{B3}$$

$$\sum_{\ell,m} {}_sY_{\ell m}(\hat{\mathbf{n}})\, {}_sY^*_{\ell m}(\hat{\mathbf{n}}') = \delta(\cos\theta - \cos\theta')\delta(\phi - \phi'). \tag{B4}$$

This leads to the following forward and inverse transformations for (square-integrable) spin-weighted functions on the sphere:

$${}_sf_{\ell m} = \int_{S^2} \mathrm{d}\Omega(\hat{\mathbf{n}})\, {}^{(s)}f(\hat{\mathbf{n}})\, {}_sY^*_{\ell m}(\hat{\mathbf{n}}) \quad \forall\ \ell \in \{|s|, \ldots, \ell_{\mathrm{max}}\},$$
$$\forall\ m \in \{-\ell, \ell\},$$
$${}^{(s)}f(\hat{\mathbf{n}}) = \sum_{\ell=|s|}^{\ell_{\mathrm{max}}} \sum_{m=-\ell}^{\ell} {}_sf_{\ell m}\, {}_sY_{\ell m}(\hat{\mathbf{n}}) \quad \forall\ \hat{\mathbf{n}} \in S^2. \tag{B5}$$

We include the Condon-Shortley phase convention in our definition of the SWSHs. Under complex conjugation and parity $[(\theta, \phi) \mapsto (\pi - \theta, \phi + \pi)]$ the functions therefore obey

$${}_sY^*_{\ell m}(\hat{\mathbf{n}}) = (-1)^{s+m}\, {}_{-s}Y_{\ell -m}(\hat{\mathbf{n}}), \tag{B6}$$

[22] Throughout this work we only describe (representation of) three-dimensional (3D) rotations so we limit ourselves to (non-negative) integer multipole order (ℓ) and integer magnetic or "azimuthal" numbers (m and s).

$$ {}_sY_{\ell m}(-\hat{\mathbf{n}}) = (-1)^{\ell}\,{}_{-s}Y_{\ell m}(\hat{\mathbf{n}}). \tag{B7}$$

In particular, this implies that ${}_sf^*_{\ell m} = {}_{-s}f_{\ell -m}(-1)^{m+s}$ holds for two spin-weighted functions ${}^{(\pm s)}f$ that obey $({}^{(s)}f)^* = {}^{(-s)}f$. For $s=0$ this simply means that f is a real-valued function.

A tensor product of SWSHs may be decomposed into a direct sum by making use of the Wigner 3-j symbols (see next section):

$$ {}_{s_1}Y_{\ell_1 m_1}(\hat{\mathbf{n}})\,{}_{s_2}Y_{\ell_2 m_2}(\hat{\mathbf{n}}) = \sum_{\ell_3=|\ell_1-\ell_2|}^{\ell_1+\ell_2}\sum_{m_3=-\ell_3}^{\ell_3}\sum_{s_3=-\ell_3}^{\ell_3} J^{-s_1-s_2-s_3}_{\ell_1\ell_2\ell_3} \times \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix} {}_{s_3}Y^*_{\ell_3 m_3}(\hat{\mathbf{n}}), \tag{B8}$$

with

$$ J^{s_1s_2s_3}_{\ell_1\ell_2\ell_3} \equiv \sqrt{\frac{(2\ell_1+1)(2\ell_2+1)(2\ell_3+1)}{4\pi}} \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ s_1 & s_2 & s_3 \end{pmatrix}. \tag{B9}$$

Note that the upper limit on the first sum in Eq. (B8) implies that the harmonic band limit of a product of functions on the sphere is given by the sum of their individual band limits. Equation (B8) also shows that the integral over a product of three SWSHs is given by

$$ \int_{S^2} \mathrm{d}\Omega(\hat{\mathbf{n}})\,{}_{s_1}Y_{\ell_1 m_1}(\hat{\mathbf{n}})\,{}_{s_2}Y_{\ell_2 m_2}(\hat{\mathbf{n}})\,{}_{s_3}Y_{\ell_3 m_3}(\hat{\mathbf{n}}) = J^{-s_1-s_2-s_3}_{\ell_1\ell_2\ell_3} \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix}. \tag{B10}$$

Note though that this only holds for the $s_1+s_2+s_3=0$ case. For the $s_1=s_2=s_3=0$ case this integral is referred to as the Gaunt integral.

2. Wigner 3-*j*, 6-*j*, and 9-*j* symbols

The Wigner 3-j symbols are real valued and serve to describe the decomposition of tensor products of SWSHs into direct sums of SWSHs [see Eq. (B8)] (this also holds, in more generality, for irreps of the rotation group such as the Wigner-D matrices). The 3-j symbols are closely related to the Clebsch-Gordan coefficients but are normalized such that they are the exact coefficients needed to form a rotationally invariant product of three SWSH coefficients [recall the definition of the angle-averaged bispectrum in Eq. (18)]. In the following, we list a limited number of symbol properties; see Ref. [153] for an exhaustive description.

The 3-j symbols pick up a (real) phase factor when the sign of the three “magnetic” indices is simultaneously changed,

$$ \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix} = (-1)^{\ell_1+\ell_2+\ell_3} \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ -m_1 & -m_2 & -m_3 \end{pmatrix}. \tag{B11}$$

The symbols are invariant under cyclic permutations of m_1, m_2, and m_3 but pick up a factor of $(-1)^{\ell_1+\ell_2+\ell_3}$ for anticyclic permutations. The symbols are only nonzero for $m_1+m_2+m_3=0$, $|\ell_1-\ell_2|\le\ell_3\le\ell_1+\ell_2$, and $|m_i|\le\ell_i\,\forall\, i\in\{1,2,3\}$. There are two orthogonality relations:

$$ \sum_{L,M}(2L+1)\begin{pmatrix} \ell_1 & \ell_2 & L \\ m_1 & m_2 & M \end{pmatrix}\begin{pmatrix} \ell_1 & \ell_2 & L \\ m_1' & m_2' & M \end{pmatrix} = \delta_{m_1m_1'}\delta_{m_2m_2'}, \tag{B12}$$

$$ \sum_{m_1,m_2}\begin{pmatrix} \ell_1 & \ell_2 & L \\ m_1 & m_2 & M \end{pmatrix}\begin{pmatrix} \ell_1 & \ell_2 & L' \\ m_1 & m_2 & M' \end{pmatrix} = \frac{\delta_{LL'}\delta_{MM'}}{2L+1}. \tag{B13}$$

In particular, in the case of equal symbols one has

$$ \sum_{m_1=-\ell_1}^{\ell_1}\sum_{m_2=-\ell_2}^{\ell_2}\sum_{m_3=-\ell_3}^{\ell_3}\begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix}^2 = 1. \tag{B14}$$

As mentioned, the Wigner 3-j symbols are used to couple two SWSHs or, equivalently, find the third angular state that combines two SWSHs into a rotationally invariant quantity. In general, there is no unique way to couple three SWSHs; there are two distinct sequences of applying Eq. (B8) to the product. The Wigner 6-j symbol is used to transform between these two possible final angular states [153]:

$$ \sum_L (2L+1)(-1)^{\ell_1+\ell_3+m_1+m_4}\begin{Bmatrix} \ell_1 & \ell_2 & \ell_4 \\ \ell_3 & \ell_5 & L \end{Bmatrix}\begin{pmatrix} \ell_1 & L & \ell_5 \\ m_1 & M & m_5 \end{pmatrix}\begin{pmatrix} \ell_3 & L & \ell_2 \\ -m_3 & M & m_2 \end{pmatrix} = \begin{pmatrix} \ell_1 & \ell_4 & \ell_2 \\ m_1 & m_4 & -m_2 \end{pmatrix}\begin{pmatrix} \ell_3 & \ell_4 & \ell_5 \\ m_3 & -m_4 & -m_5 \end{pmatrix}. \tag{B15}$$

By using one of the orthogonality relations of the 3-j symbols, the 6-j symbol may equivalently be expressed as

$$\begin{Bmatrix} \ell_1 & \ell_2 & \ell_3 \\ \ell_4 & \ell_5 & \ell_6 \end{Bmatrix}\begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix} = \sum_{m_4,m_5,m_6} (-1)^{\sum_{i=4}^{6}\ell_i+m_i}\begin{pmatrix} \ell_1 & \ell_5 & \ell_6 \\ m_1 & m_5 & -m_6 \end{pmatrix}\begin{pmatrix} \ell_4 & \ell_2 & \ell_6 \\ -m_4 & m_2 & m_6 \end{pmatrix} \times \begin{pmatrix} \ell_4 & \ell_5 & \ell_3 \\ m_4 & -m_5 & m_3 \end{pmatrix}. \tag{B16}$$

The 6-j symbols are invariant under all permutations of their columns and under the simultaneous permutation of upper and lower arguments in two columns. The symbols also obey several triangle conditions that can be deduced from the top rows of each of the 3-j symbols in the above expression. There also exist an orthogonality relation for the 6-j symbols [153].

Finally, the Wigner 9-j symbols are defined to describe the transformation between different couplings of four SWSHs. The symbols may be expressed in terms of either 6-j or 3-j symbols [153]. The latter expression is given by

$$\begin{Bmatrix} \ell_1 & \ell_2 & \ell_3 \\ \ell_4 & \ell_5 & \ell_6 \\ \ell_7 & \ell_8 & \ell_9 \end{Bmatrix}\begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix} = \sum_{m_4,\ldots,m_9}\begin{pmatrix} \ell_4 & \ell_5 & \ell_6 \\ m_4 & m_5 & m_6 \end{pmatrix}\begin{pmatrix} \ell_7 & \ell_8 & \ell_9 \\ m_7 & m_8 & m_9 \end{pmatrix}\begin{pmatrix} \ell_4 & \ell_7 & \ell_1 \\ m_4 & m_7 & m_1 \end{pmatrix} \times \begin{pmatrix} \ell_5 & \ell_8 & \ell_2 \\ m_5 & m_8 & m_2 \end{pmatrix}\begin{pmatrix} \ell_6 & \ell_9 & \ell_3 \\ m_6 & m_9 & m_3 \end{pmatrix}. \tag{B17}$$

The 9-j symbols are invariant under reflections of their arguments along either diagonal and even permutations of rows or columns; odd permutations result in a factor of $(-1)^{\sum_{i=1}^{9}\ell_i}$. Elements of each row and column are constrained by the triangle conditions of the 3-j symbols in the above expression. There also exists an orthogonality relation for the 9-j symbols; details can be found in Ref. [153].

3. Delta function

The delta function is expanded as

$$\delta^{(3)}(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3) = \frac{1}{(2\pi)^3}\int \mathrm{d}^3\mathbf{x}\, e^{i(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3)\cdot\mathbf{x}}. \tag{B18}$$

By making use of the Rayleigh equation

$$e^{i\mathbf{k}\cdot\mathbf{x}} = \sum_{\ell} i^{\ell}(2\ell+1)j_{\ell}(kr)P_{\ell}(\hat{\mathbf{k}}\cdot\hat{\mathbf{n}}) \tag{B19}$$

$$= 4\pi\sum_{\ell,m} i^{\ell} j_{\ell}(kr)Y^{*}_{\ell m}(\hat{\mathbf{k}})Y_{\ell m}(\hat{\mathbf{n}}), \tag{B20}$$

we produce two equivalent expressions for the delta function,

$$\delta^{(3)}(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3) = 8\int_0^{\infty} r^2\mathrm{d}r \sum_{\ell_1,m_1}\sum_{\ell_2,m_2}\sum_{\ell_3,m_3}\left(\prod_{i=1}^{3} j_{\ell_i}(k_i r)Y^{*}_{\ell_i m_i}(\hat{\mathbf{k}}_i)\right)\int_{S^2}\mathrm{d}\Omega(\hat{\mathbf{n}})\left(\prod_{i=1}^{3} Y_{\ell_i,m_i}(\hat{\mathbf{n}})\right) \tag{B21}$$

$$= 8\int_0^{\infty} r^2\mathrm{d}r \sum_{\ell_1,m_1}\sum_{\ell_2,m_2}\sum_{\ell_3,m_3}\left(\prod_{i=1}^{3} i^{\ell_1} j_{\ell_i}(k_i r)Y^{*}_{\ell_i m_i}(\hat{\mathbf{k}}_i)\right)J^{000}_{\ell_1\ell_2\ell_3}\begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix}. \tag{B22}$$

Note that $\mathbf{x} = r\hat{\mathbf{n}}$ and $\mathbf{k} = k\hat{\mathbf{k}}$. We have used Eq. (B10) to arrive at the second expression. The J symbol is defined in Eq. (B9). See Ref. [154] for these and alternative expressions.

APPENDIX C: COSMOLOGY CONVENTIONS

1. Power spectra

Because of the assumed statistical homogeneity of the superhorizon 2-point correlation functions of the amplitudes of both the spatial components of ζ and ${}^{(\lambda)}h$, these correlations are represented as diagonal in the 3D Fourier basis. Statistical isotropy limits the diagonal to only depend on the wave number k. We use the following conventions for the correlation functions:

$$\langle \zeta_{\mathbf{k}} \zeta^*_{\mathbf{k}'} \rangle = (2\pi)^3 \delta^{(3)}(\mathbf{k}-\mathbf{k}') P_\zeta(k), \tag{C1}$$

$$\langle {}^{(\lambda)}h_{\mathbf{k}} {}^{(\lambda')}h^*_{\mathbf{k}'} \rangle = (2\pi)^3 \delta^{(3)}(\mathbf{k}-\mathbf{k}') \delta_{\lambda,\lambda'} \frac{P_h(k)}{2}. \tag{C2}$$

The power spectra are parametrized as follows:

$$P_\zeta(k) = 2\pi^2 \frac{A_s(k_0)}{k^3} \left(\frac{k}{k_0}\right)^{n_s(k_0)-1}, \tag{C3}$$

$$P_h(k) = 2\pi^2 \frac{r_{k_0} A_s(k_0)}{k^3} \left(\frac{k}{k_0}\right)^{n_t(k_0)}, \tag{C4}$$

with tensor-to-scalar ratio r_{k_0} (i.e., the ratio at the pivot scale), scalar amplitude A_s, pivot scale k_0, and scalar (tensor) spectral tilt n_s (n_t). We have used fixed values for some of these parameters: $\{A_s(k_0) = 2.1056 \times 10^{-9}, k_0 = 0.05\ \mathrm{Mpc}^{-1}, n_s(k_0) = 0.9665, n_t(k_0) = 0\}$. The remaining cosmological parameters that govern the radiation transfer functions are set to $\{T_{\mathrm{CMB}} = 2.7255\ \mathrm{K}, H_0 = 67.66\ \mathrm{km\,s^{-1}\,Mpc^{-1}}, \Omega_{\mathrm{b}}h^2 = 0.02242, \Omega_{\mathrm{c}}h^2 = 0.11933, \tau = 0.0561\}$, and the CAMB defaults of December 2018.

We extract the radiation transfer functions from CAMB. We normalize the default output from CAMB such that the CMB power/cross spectra are related to the primordial power spectra defined above as

$$\langle a^{(Z)}_{X,\ell m} a^{(Z)*}_{Y,\ell' m'} \rangle = \delta_{\ell\ell'} \delta_{mm'} C^{(Z)}_{XY,\ell} \tag{C5}$$

$$= \delta_{\ell\ell'} \delta_{mm'} \frac{2}{\pi} \int_0^\infty k^2 \mathrm{d}k P_Z(k) \mathcal{T}^{(Z)}_{X,\ell}(k) \mathcal{T}^{(Z)}_{Y,\ell}(k), \tag{C6}$$

with $Z \in \{\zeta, h\}$ and $XY \in \{TT, EE, TE, ET, BB\}$.

2. Local 3-point correlation function

The local shape template used in Sec. IV is given by [89]

$$f^{\mathrm{local}}(k_1, k_2, k_3) = 2\left[\left(\frac{1}{(k_1 k_2)^3}\right) + 2\ \mathrm{perm}\right]. \tag{C7}$$

The template is symmetric under permutations of the three wave numbers and is perfectly scale invariant (i.e., proportional to k^{-6} for $k_1 = k_2 = k_3$). If desired, including the scalar or tensor spectral tilt simply amounts to the replacement $k \mapsto k(k_0/k)^{(n_s-1)/3}$ or $k \mapsto k(k_0/k)^{n_t/3}$, where k_0 is some fiducial pivot scale.

APPENDIX D: ESTIMATOR DERIVATION

We review the derivation of the estimator in Eq. (26) and its behavior in the presence of the non-Gaussian signal.

1. Estimation theory

The statistical estimate of a parameter produced by an unbiased estimator has an expectation value that is equal to the true value of the parameter. If such an unbiased estimator saturates the Cramér-Rao bound, it achieves the lowest possible variance (or covariance for multiple parameters) on the estimate, independent from the true value(s) of the parameter(s). We will briefly introduce the Cramér-Rao bound.

Consider a dataset $x = \{x_1, x_2, \ldots, x_n\}$ drawn from the likelihood $\Pr(x|\theta)$: a probability density function (PDF) with unknown fixed parameters $\theta = \{\theta_1, \theta_2, \ldots, \theta_d\}$. Under the assumption that the PDF satisfies the following regularity condition:

$$\int \mathrm{d}^n x \frac{\partial \log \Pr(x|\theta)}{\partial \theta_i} \Pr(x|\theta) = 0, \quad \forall\ i \in \{1, \ldots, d\}, \tag{D1}$$

it can be shown that the covariance matrix C_θ of an unbiased estimate of the parameters θ is bounded by the inverse of the Fisher information matrix:

$$C_{\hat{\theta}} \geq \mathcal{I}^{-1}(\theta). \tag{D2}$$

This bound is the Cramér-Rao bound. In the matrix notation used here, the inequality refers to the positive definiteness of the $C_{\hat{\theta}} - \mathcal{I}^{-1}$ matrix. The elements of the information matrix are directly obtained from the PDF:

$$\mathcal{I}_{ij}(\theta) = \int \mathrm{d}^n x \left(\frac{\partial \log \Pr(x|\theta)}{\partial \theta_i}\right) \left(\frac{\partial \log \Pr(x|\theta)}{\partial \theta_j}\right) \Pr(x|\theta). \tag{D3}$$

The Fisher information does not depend on the observed data; it only depends on the parameter vector.

It can be shown that an unbiased estimator $\hat{\theta} = \{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_d\}$ that saturates the bound for all values of the parameters θ must satisfy

$$\frac{\partial \log \Pr(x|\theta)}{\partial \theta_i} = \sum_j \mathcal{I}_{ij}(\theta)(\hat{\theta}_j - \theta_j). \tag{D4}$$

Although it is generally nontrivial to construct an estimator that fulfills this relation for all possible values of θ, the relation suggests a simple recipe for the construction of an estimator $\hat{\theta}$ that fulfills the Cramér-Rao bound in the case where θ is known,

$$\hat{\theta}_i = \sum_j \mathcal{I}_{ij}^{-1}(\theta)\frac{\partial \log \Pr(x|\theta)}{\partial \theta_j} + \theta_i, \tag{D5}$$

where $\mathcal{I}^{-1}$ is the inverse of the Fisher matrix. In reality, θ is unknown. However, an estimator constructed in this way may still be useful for estimates of θ that are close to the assumed value. This is the approach we will take.

2. CMB bispectrum estimation

We now construct the bispectrum estimator and provide a brief discussion of its statistical properties. We will see that the estimator is unbiased and becomes statistically optimal (saturates the Cramér-Rao bound) in the limit of vanishing non-Gaussianity.

a. Probability density function

It is clear from the previous section that a closed-form expression for the likelihood of the data is required to construct the estimator. However, there exists no such expression when the condition of Gaussian initial perturbations is relaxed. Without a closed-form expression, we thus construct an approximation to the full non-Gaussian likelihood by perturbing around the Gaussian form. The specifics of this perturbation are determined by the connected moments, or cumulants, predicted by the model.

Given a characteristic function and its associated probability distribution, one can distinguish between the moments about the origin of the distribution (the n-point correlation functions) and the connected moments about the origin (the cumulants). The connected moments are proportional to the MacLaurin coefficients of the natural logarithm of the characteristic function. The connected moments about the origin are proportional to the MacLaurin coefficients of the characteristic function itself (i.e., without the logarithm). In more practical terms the moments about the origin, the n-point correlation functions, may be expanded in terms of the connected moments with the help of Wick's theorem [155]. For the mean-zero distributions we are interested in, the first moments of a random field, expressed as a set of spherical harmonic modes $\{a_{\ell m}\}$, are expanded as follows:

$$\langle a_{\ell_1 m_1} a_{\ell_2 m_2} \rangle = \langle a_{\ell_1 m_1} a_{\ell_2 m_2} \rangle_c, \tag{D6}$$

$$\langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3} \rangle = \langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3} \rangle_c, \tag{D7}$$

$$\begin{aligned}\langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3} a_{\ell_4 m_4} \rangle &= \langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3} a_{\ell_4 m_4} \rangle_c + \langle a_{\ell_1 m_1} a_{\ell_2 m_2} \rangle_c \langle a_{\ell_3 m_3} a_{\ell_4 m_4} \rangle_c \\ &\quad + \langle a_{\ell_1 m_1} a_{\ell_3 m_3} \rangle_c \langle a_{\ell_2 m_2} a_{\ell_4 m_4} \rangle_c + \langle a_{\ell_1 m_1} a_{\ell_4 m_4} \rangle_c \langle a_{\ell_2 m_2} a_{\ell_3 m_3} \rangle_c,\end{aligned} \tag{D8}$$

$$\langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3} a_{\ell_4 m_4} a_{\ell_5 m_5} \rangle = \langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3} a_{\ell_4 m_4} a_{\ell_5 m_5} \rangle_c, \tag{D9}$$

$$\begin{aligned}\langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3} a_{\ell_4 m_4} a_{\ell_5 m_5} a_{\ell_6 m_6} \rangle &= \langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3} a_{\ell_4 m_4} a_{\ell_5 m_5} a_{\ell_6 m_6} \rangle_c \\ &\quad + \langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3} a_{\ell_4 m_4} \rangle_c \langle a_{\ell_5 m_5} a_{\ell_6 m_6} \rangle_c + 14\ \text{perm} \\ &\quad + \langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3} \rangle_c \langle a_{\ell_4 m_4} a_{\ell_5 m_5} a_{\ell_6 m_6} \rangle_c + 9\ \text{perm} \\ &\quad + \langle a_{\ell_1 m_1} a_{\ell_2 m_2} \rangle_c \langle a_{\ell_3 m_3} a_{\ell_4 m_4} \rangle_c \langle a_{\ell_5 m_5} a_{\ell_6 m_6} \rangle_c + 14\ \text{perm.}\end{aligned} \tag{D10}$$

The quantities on the lhs represent the moments and the quantities on the rhs are the connected moments (denoted by $\langle \cdots \rangle_c$). For a distribution with a vanishing mean, there is no distinction between the moments and connected moments for $n = 2$ and $n = 3$. For $n = 4$ and higher, we see a distinction. A Gaussian distribution is a distribution for which all connected moments with $n > 2$ vanish.

The approximation to the likelihood of the data we will use is known as the Edgeworth series. More specifically, it is an Edgeworth expansion around a mean zero multivariate Gaussian distribution. We truncate the series such that the only relevant cumulants are the 2- and 3-point functions. A detailed derivation of this procedure can be found in Ref. [156]. In short, one Taylor expands the non-Gaussian part of a general characteristic function to first order and discards all terms except the third-order moments. Fourier transforming this truncated series together with the unmodified Gaussian part yields the PDF. Although the Edgeworth expansion is an asymptotic series, truncating it to third order does not guarantee a well-defined (i.e., positive and normalized) PDF [157]. However, as long as we are only interested in the weakly non-Gaussian regime, where the third-order moment is subdominant to the second, we assume that these subtleties can be safely ignored.

Representing the likelihood for a measured set of n spherical harmonic modes $a = \{a_{X,\ell m}\}$ as the truncated Edgeworth series yields [78,115]

$$\mathrm{Pr}(a|C,B) = \Bigg(1 + \frac{1}{6}\sum_{\substack{\ell_1,\ell_2,\ell_3\\ m_1,m_2,m_3}} B^{\ell_1\ell_2\ell_3}_{m_1m_2m_3,X_1X_2X_3}\{[(C^{-1}a)^{X_1}_{\ell_1m_1}(C^{-1}a)^{X_2}_{\ell_2m_2}(C^{-1}a)^{X_3}_{\ell_3m_3}] - [(C^{-1})^{X_1X_2}_{\ell_1m_1\ell_2m_2}(C^{-1}a)^{X_3}_{\ell_3m_3} + \text{cyclic}]\}\Bigg)\frac{e^{-\frac{1}{2}a^\dagger C^{-1}a}}{\sqrt{(2\pi)^n\det C}}, \tag{D11}$$

where C and B denote the 2- and 3-point correlation functions of a. The notational shorthand $C^{-1}a$ is defined in Eq. (27). The above expression is that of a nested model: when B vanishes, we recover the mean zero Gaussian model. The extra terms denoted by “cyclic” are given by the two cyclic permutations of the three (ℓ, m, X) triplets.

It is straightforward to incorporate harmonic modes sourced by a combination of primordial scalar and tensor perturbations in the above description. Consider the following decomposition:

$$a_{X,\ell m} = a^{(\zeta)}_{X,\ell m} + a^{(h)}_{X,\ell m} + n_{X,\ell m}. \tag{D12}$$

Since the noise $n_{X,\ell m}$ is independent from the primordial fields and since all components have zero mean, the most general bispectrum then is expressed as

$$B = B^{(\zeta\zeta\zeta)} + 3B^{(\zeta\zeta h)} + 3B^{(\zeta hh)} + B^{(hhh)}. \tag{D13}$$

Inserting Eq. (D13) into Eq. (D11) produces a likelihood for a that takes into account the non-Gaussian correlation between the primordial scalar and tensor fields.

b. Estimator

The condition in Eq. (D4) implies that an unbiased estimator of a vector of parameters $\hat{f}_{\mathrm{NL}} = \{\hat{f}^1_{\mathrm{NL}}, \hat{f}^2_{\mathrm{NL}}, \ldots, \hat{f}^d_{\mathrm{NL}}\}$ constructed as follows saturates the Cramér-Rao bound in the limit where the parameter vector goes to the null vector, i.e., $f_{\mathrm{NL}} \to 0$:

$$\hat{f}^I_{\mathrm{NL}} = \sum_J \mathcal{I}^{-1}_{IJ}(f_{\mathrm{NL}})\frac{\partial \log \mathrm{Pr}(a|f_{\mathrm{NL}})}{\partial f^J_{\mathrm{NL}}}. \tag{D14}$$

The I and J indices run over the dimensions of the parameter vector space. Note that $\hat{f}_{\mathrm{NL}}$ and f_{NL} can be understood either as scalars or as vectors; in the latter case, the $\mathcal{I}^{-1}$ is the inverse of the $d \times d$ Fisher matrix instead of the scalar Fisher information. To identify $\mathrm{Pr}(a|C,B)$ in Eq. (D11) with $\mathrm{Pr}(a|f_{\mathrm{NL}})$, we treat the bispectrum as fixed up to a scaling $f_{\mathrm{NL}} \in \mathbb{R}^d$ and consider the shape of the bispectrum and the covariance as fixed. More specifically, we assume

$$B(f_{\mathrm{NL}}) = f_{\mathrm{NL}} \cdot B_1, \tag{D15}$$

where the inner product is defined in the parameter vector space. This expression is a generalization of Eq. (25) that allows the bispectrum to consist of a sum of bispectra each with its own f_{NL} parameter. To construct the estimator we now simply insert Eq. (D15) into the expression for the PDF in Eq. (D11) and insert the result into Eq. (D14). We may expand the logarithm in a power series and neglect all terms but the one that is $\mathcal{O}(B)$. This is a valid approach because the second term in the brackets in Eq. (D11) must be $\ll 1$ in the weak non-Gaussian regime. This yields the estimator constructed by Ref. [78] (which is a refinement to the cubic expression originally introduced in Ref. [82]):

$$\hat{f}^I_{\mathrm{NL}} = \frac{1}{6}\sum_J \mathcal{I}^{-1}_{0,IJ} \sum_{\text{all }\ell,m} \sum_{\text{all }X} (B^J_1)^{\ell_1\ell_2\ell_3}_{m_1m_2m_3,X_1X_2X_3} \times \{[(C^{-1}a)^{X_1}_{\ell_1m_1}(C^{-1}a)^{X_2}_{\ell_2m_2}(C^{-1}a)^{X_3}_{\ell_3m_3}] - [(C^{-1})^{X_1X_2}_{\ell_1m_1\ell_2m_2}(C^{-1}a)^{X_3}_{\ell_3m_3} + \text{cyclic}]\}. \tag{D16}$$

Note the use of $\mathcal{I}^{-1}_0 \equiv \mathcal{I}^{-1}(0)$ instead of $\mathcal{I}^{-1}(f_{\mathrm{NL}})$: strictly speaking, the inverse of the Fisher matrix will depend on the parameter vector. This reflects the fact that a true optimal estimator should vary between datasets based on the value of f_{NL}. Of course, such optimality is not possible with the point estimator we use here: f_{NL} is unknown. A true optimal weighting would be achieved with a Bayesian approach in which the likelihood of the data is calculated for each value of f_{NL}. In reality, this reweighting of the estimator is not important for values of f_{NL} that are of interest [158]. For $f_{\mathrm{NL}} = 0$ the estimator is optimal by construction and the Fisher matrix has a simple analytic solution:

$$\mathcal{I}_{0,IJ} = \frac{1}{6}\sum_{\text{all}\ell,m}\sum_{\text{all}X}(B^I_1)^{\ell_1\ell_2\ell_3}_{m_1m_2m_3,X_1X_2X_3} \times [(C^{-1})^{X_1X_4}_{\ell_1m_1\ell_4m_4}(C^{-1})^{X_2X_5}_{\ell_2m_2\ell_5m_5}(C^{-1})^{X_3X_6}_{\ell_3m_3\ell_6m_6}] \times (B^{J*}_1)^{\ell_4\ell_5\ell_6}_{m_4m_5m_6,X_4X_5X_6}. \tag{D17}$$

c. Statistical properties estimator

The estimator $\hat{f}_{\mathrm{NL}}$ is a function, or "statistic," of the data a, so the statistical properties of the estimator may be derived from the likelihood of the data. Here we present a heuristic overview of the statistical properties, given different models for the data. It should be understood that an analytic approach such as the one presented here is mainly useful to gain intuition; characterization of the estimator applied to a real dataset requires the use of simulations.

To derive the bias, covariance, and higher-order moments of the estimate, we first define what we mean by the pth moment of the estimate,

$$\langle \hat{f}^p_{\mathrm{NL}} \rangle \equiv \mathrm{E}(\hat{f}^p_{\mathrm{NL}} | f_{\mathrm{NL}}) \tag{D18}$$

$$= \int \mathcal{D}a \hat{f}^p_{\mathrm{NL}} \, \mathrm{Pr}(a | f_{\mathrm{NL}}), \tag{D19}$$

where

$$\int \mathcal{D}a \equiv \prod_{\ell,m} \int \mathrm{d}a_{\ell m}, \tag{D20}$$

and where $\hat{f}^p_{\mathrm{NL}}$ denotes the pth power of the estimate. This notation is understood to generalize to the multivariate case as, e.g., $\langle \hat{f}^2_{\mathrm{NL}} \rangle \to \langle \hat{f}^I_{\mathrm{NL}} \hat{f}^J_{\mathrm{NL}} \rangle$, $\langle \hat{f}^3_{\mathrm{NL}} \rangle \to \langle \hat{f}^I_{\mathrm{NL}} \hat{f}^J_{\mathrm{NL}} \hat{f}^K_{\mathrm{NL}} \rangle$. It is then convenient to note that the expression for $\mathrm{Pr}(a|f_{\mathrm{NL}})$ in Eq. (D11) consists of two parts: a regular Gaussian PDF and a second part that consists of a Gaussian PDF times terms cubic and linear in a. This means that we can divide the integral in Eq. (D19) into a purely Gaussian integral ($\langle \cdots \rangle_G$) and another Gaussian integral ($\langle \cdots \rangle_{G'}$) with an integrand that is multiplied with these cubic and linear terms. Since the estimator in Eq. (D16) is an odd function of a, the $\langle \cdots \rangle_G$ integral will always vanish for $p =$ odd. On the other hand, the $\langle \cdots \rangle_{G'}$ part will always vanish for a moment with $p =$ even.

With this knowledge and the likelihood of the data in Eq. (D11), deriving the bias of the estimator comes down to evaluating Eq. (D19) for $p = 1$. This is an odd moment, so only the $\langle \cdots \rangle_{G'}$ integral has to be evaluated. The result is that $\langle \hat{f}_{\mathrm{NL}} \rangle = f_{\mathrm{NL}}$; i.e., the estimate is unbiased regardless of the value of f_{NL}. For the (co)variance of the estimate, i.e., $\mathrm{Var}(\hat{f}_{\mathrm{NL}}) \equiv \langle \hat{f}^2_{\mathrm{NL}} \rangle - \langle \hat{f}_{\mathrm{NL}} \rangle^2$, we need to additionally evaluate Eq. (D19) for $p = 2$. Doing so, we find $\mathrm{Var}(\hat{f}_{\mathrm{NL}}) = \mathcal{I}_0^{-1} - f^2_{\mathrm{NL}}$, which is equal to the optimal value $\mathcal{I}^{-1}(f_{\mathrm{NL}})$ only when $f_{\mathrm{NL}} = 0$. So we establish that in the limit of $f_{\mathrm{NL}} \to 0$, the estimator is unbiased and optimal. In cases where $f_{\mathrm{NL}} \neq 0$, the estimator is still unbiased[23] but suffers from nonoptimal (co)variance [116,117]. This is expected, as the situation does not conform to Eq. (D4) anymore. Finally, note that for $f_{\mathrm{NL}} \neq 0$, the estimate itself becomes (weakly) non-Gaussian. For instance, there will be a nonzero $\langle \hat{f}^3_{\mathrm{NL}} \rangle$ moment with an $\mathcal{O}(f_{\mathrm{NL}} B_1^4 C^{-6} \mathcal{I}_0^{-3})$ amplitude.

In the above we assumed that the likelihood for the data is described by Eq. (D11). When an additional 3-point function, not parametrized by an f_{NL} parameter, is introduced in the likelihood, the estimator becomes biased. The exact bias depends on the shape of the added 3-point function; see the discussion in Sec. V.

An interesting situation arises when the data are drawn from a distribution with nonzero higher-order connected moments. This situation is not only hypothetical: lensing introduces a significant connected 4-point function, as well as smaller connected 6-, 8-, etc., point functions [131]. To describe the statistical properties of the estimator in the presence of lensing, we thus need to update the likelihood of the data in Eq. (D11) with these nonzero higher-order connected moments. Let us focus on the connected 4-point function, denoted by T. The Edgeworth expansion will now include $\mathcal{O}(Ta^4/C^4)$, $\mathcal{O}(Ta^2/C^3)$, and $\mathcal{O}(T/C^2)$ terms in addition to the $\mathcal{O}(1)$, $\mathcal{O}(B_1 a^3/C^3)$, and $\mathcal{O}(B_1 a/C^2)$ terms already present in Eq. (D11). With these additions, the bias of the estimator does not change, but the variance of the estimator receives an $\mathcal{O}(B_1^2 T C^{-5} \mathcal{I}_0^{-2})$ contribution. By extension, the addition of connected 6-, 8-, or higher-point functions to the likelihood will also contribute to the estimator variance. The estimate itself will also become non-Gaussian with these additions. For instance, there is an $\mathcal{O}(B_1^4 T C^{-8} \mathcal{I}_0^{-4})$ connected 4-point function of $\hat{f}_{\mathrm{NL}}$ when a connected 4-point function T is added to the likelihood. Computing a semianalytic estimate of the additional estimator variance is highly challenging due to the number of elements that make up the higher-order connected moments. See Refs. [91,159] for details on a semianalytic approach in the flat-sky approximation. We briefly discuss the expected additional lensing-induced estimator variance in Sec. V D.

[23] Of course, any statements about unbiasedness rely on the assumed validity of the truncated Edgeworth expansion, which, as mentioned, should be reconsidered in cases of large deviations from Gaussianity.

[1] A. H. Guth, Phys. Rev. D **23**, 347 (1981).
[2] A. D. Linde, Phys. Lett. **108B**, 389 (1982).
[3] A. Albrecht and P. J. Steinhardt, Phys. Rev. Lett. **48**, 1220 (1982).
[4] V. F. Mukhanov and G. V. Chibisov, JETP Lett. **33**, 532 (1981).
[5] S. W. Hawking, Phys. Lett. **115B**, 295 (1982).
[6] A. A. Starobinsky, Phys. Lett. **117B**, 175 (1982).
[7] M. Kamionkowski, A. Kosowsky, and A. Stebbins, Phys. Rev. Lett. **78**, 2058 (1997).
[8] M. Zaldarriaga and U. Seljak, Phys. Rev. D **55**, 1830 (1997).
[9] A. A. Starobinsky, JETP Lett. **30**, 682 (1979).
[10] L. P. Grishchuk, Sov. Phys. JETP **40**, 409 (1975).
[11] V. A. Rubakov, M. V. Sazhin, and A. V. Veryaskin, Phys. Lett. **115B**, 189 (1982).
[12] V. F. Mukhanov, H. A. Feldman, and R. H. Brandenberger, Phys. Rep. **215**, 203 (1992).
[13] Planck Collaboration, arXiv:1807.06211
[14] X. Chen, Adv. Astron. **2010**, 1 (2010).
[15] Planck Collaboration, arXiv:1905.05697.
[16] P. D. Meerburg, D. Green, R. Flauger, B. Wallisch, M. C. D. Marsh, E. Pajer, G. Goon, C. Dvorkin, A. M. Dizgah, D. Baumann, G. L. Pimentel, S. Foreman, E. Silverstein, E. Chisari, B. Wandelt, M. Loverde, and A. Slosar, Bull. Am. Astron. Soc. **51**, 107 (2019).
[17] P. D. Meerburg, J. Meyers, A. van Engelen, and Y. Ali-Haïmoud, Phys. Rev. D **93**, 123511 (2016).
[18] J. M. Maldacena, J. High Energy Phys. 05 (2003) 013.
[19] M. Akhshik, J. Cosmol. Astropart. Phys. 05 (2015) 043.
[20] E. Dimastrogiovanni, M. Fasiello, D. Jeong, and M. Kamionkowski, J. Cosmol. Astropart. Phys. 12 (2014) 050.
[21] L. Bordin, P. Creminelli, M. Mirbabayi, and J. Noreńa, J. Cosmol. Astropart. Phys. 09 (**2016**) 041.
[22] G. Domènech, T. Hiramatsu, C. Lin, M. Sasaki, M. Shiraishi, and Y. Wang, J. Cosmol. Astropart. Phys. 05 (2017) 034.
[23] S. Endlich, B. Horn, A. Nicolis, and J. Wang, Phys. Rev. D **90**, 063506 (2014).
[24] J. M. Bardeen, P. J. Steinhardt, and M. S. Turner, Phys. Rev. D **28**, 679 (1983).
[25] D. Wands, K. A. Malik, D. H. Lyth, and A. R. Liddle, Phys. Rev. D **62**, 043527 (2000).
[26] S. Weinberg, Phys. Rev. D **67**, 123504 (2003).
[27] D. Baumann, G. Goon, H. Lee, and G. L. Pimentel, J. High Energy Phys. 04 (2018) 140.
[28] P. Creminelli and M. Zaldarriaga, J. Cosmol. Astropart. Phys. 10 (2004) 006.
[29] E. Pajer, F. Schmidt, and M. Zaldarriaga, Phys. Rev. D **88**, 083502 (2013).
[30] S. Deser and A. Waldron, Phys. Lett. B **513**, 137 (2001).
[31] H. Lee, D. Baumann, and G. L. Pimentel, J. High Energy Phys. 12 (2016) 040.
[32] C. Cheung, P. Creminelli, A. L. Fitzpatrick, J. Kaplan, and L. Senatore, J. High Energy Phys. 03 (2008) 014.
[33] L. Senatore and M. Zaldarriaga, J. High Energy Phys. 04 (2012) 024.
[34] L. Bordin, P. Creminelli, A. Khmelnitsky, and L. Senatore, J. Cosmol. Astropart. Phys. 10 (2018) 013.
[35] L. Senatore, E. Silverstein, and M. Zaldarriaga, J. Cosmol. Astropart. Phys. 08 (2014) 016.
[36] P. Adshead, E. Martinec, and M. Wyman, Phys. Rev. D **88**, 021302(R) (2013).
[37] A. Agrawal, T. Fujita, and E. Komatsu, Phys. Rev. D **97**, 103526 (2018).
[38] K. Hinterbichler, L. Hui, and J. Khoury, J. Cosmol. Astropart. Phys. 01 (2014) 039.
[39] M. Mirbabayi and M. Zaldarriaga, J. Cosmol. Astropart. Phys. 03 (2015) 025.
[40] N. Agarwal, R. Holman, A. J. Tolley, and J. Lin, J. High Energy Phys. 05 (2013) 085.
[41] L. Berezhiani and J. Khoury, J. Cosmol. Astropart. Phys. 09 (2014) 018.
[42] M. H. Namjoo, H. Firouzjahi, and M. Sasaki, Europhys. Lett. **101**, 39001 (2013).
[43] J. Martin, H. Motohashi, and T. Suyama, Phys. Rev. D **87**, 023514 (2013).
[44] R. Bravo, S. Mooij, G. A. Palma, and B. Pradenas, J. Cosmol. Astropart. Phys. 05 (**2018**) 025.
[45] P. Creminelli, J. Gleyzes, J. Noreña, and F. Vernizzi, Phys. Rev. Lett. **113**, 231301 (2014).
[46] Planck Collaboration, arXiv:1807.06209.
[47] CMB-S4 Collaboration, arXiv:1610.02743.
[48] A. A. Fraisse *et al.*, J. Cosmol. Astropart. Phys. 04 (2013) 047.
[49] BICEP2 and Keck Array Collaborations, Phys. Rev. Lett. **121**, 221301 (2018).
[50] M. D. Niemack *et al.*, Proc. SPIE Int. Soc. Opt. Eng. **7741**, 77411S (2010).
[51] K. Arnold *et al.*, Proc. SPIE Int. Soc. Opt. Eng. **7741**, 77411E (2010).
[52] Simons Observatory Collaboration, J. Cosmol. Astropart. Phys. 02 (2018) 056.
[53] S. W. Henderson *et al.*, J. Low Temp. Phys. **184**, 772 (2016).
[54] B. A. Benson *et al.*, Proc. SPIE Int. Soc. Opt. Eng. **9153**, 91531P (2014).
[55] T. Essinger-Hileman *et al.*, Proc. SPIE Int. Soc. Opt. Eng. **9153**, 91531I (2014).
[56] H. Hui *et al.*, Proc. SPIE Int. Soc. Opt. Eng. **10708**, 1070807 (2018).
[57] Z. Ahmed *et al.*, Proc. SPIE Int. Soc. Opt. Eng. **9153**, 91531N (2014).
[58] Y. Sekimoto *et al.*,Proc. SPIE Int. Soc. Opt. Eng. **10698**, 106981Y (2018).
[59] M. Shiraishi, M. Liguori, and J. R. Fergusson, J. Cosmol. Astropart. Phys. 01 (2018) 016.
[60] X. Chen, R. Easther, and E. A. Lim, J. Cosmol. Astropart. Phys. 06 (2007) 023.
[61] R. Holman and A. J. Tolley, J. Cosmol. Astropart. Phys. 05 (2008) 001.
[62] P. D. Meerburg, J. P. van der Schaar, and P. S. Corasaniti, J. Cosmol. Astropart. Phys. 05 (**2009**) 018.
[63] R. Flauger and E. Pajer, J. Cosmol. Astropart. Phys. 01 (2011) 017.
[64] A. Achucarro, J.-O. Gong, S. Hardeman, G. A. Palma, and S. P. Patil, J. High Energy Phys. 05 (2012) 066.
[65] M. Shiraishi, E. Komatsu, M. Peloso, and N. Barnaby, J. Cosmol. Astropart. Phys. 05 (**2013**) 002.

[66] R. Flauger, M. Mirbabayi, L. Senatore, and E. Silverstein, J. Cosmol. Astropart. Phys. 10 (2017) 058.
[67] N. Dalal, O. Dore, D. Huterer, and A. Shirokov, Phys. Rev. D **77**, 123514 (2008).
[68] T. Baldauf, U. Seljak, and L. Senatore, J. Cosmol. Astropart. Phys. 04 (2011) 006.
[69] A. Cooray, Phys. Rev. Lett. **97**, 261301 (2006).
[70] M. Schmittfull and U. Seljak, Phys. Rev. D **97**, 123540 (2018).
[71] E. Pajer and M. Zaldarriaga, Phys. Rev. Lett. **109**, 021302 (2012).
[72] J. Khoury, B. A. Ovrut, P. J. Steinhardt, and N. Turok, Phys. Rev. D **64**, 123522 (2001).
[73] M. Gasperini and G. Veneziano, Phys. Rep. **373**, 1 (2003).
[74] R. H. Brandenberger, A. Nayeri, S. P. Patil, and C. Vafa, Phys. Rev. Lett. **98**, 231302 (2007).
[75] A. Ijjas and P. J. Steinhardt, Classical Quantum Gravity **35**, 135004 (2018).
[76] M. Kamionkowski and E. D. Kovetz, Annu. Rev. Astron. Astrophys. **54**, 227 (2016).
[77] E. Komatsu, D. N. Spergel, and B. D. Wandelt, Astrophys. J. **634**, 14 (2005).
[78] P. Creminelli, A. Nicolis, L. Senatore, M. Tegmark, and M. Zaldarriaga, J. Cosmol. Astropart. Phys. 05 (2006) 004.
[79] A. P. S. Yadav and B. D. Wandelt, Phys. Rev. D **71**, 123004 (2005).
[80] A. P. S. Yadav, E. Komatsu, and B. D. Wandelt, Astrophys. J. **664**, 680 (2007).
[81] A. P. S. Yadav, E. Komatsu, B. D. Wandelt, M. Liguori, F. K. Hansen, and S. Matarrese, Astrophys. J. **678**, 578 (2008).
[82] E. Komatsu and D. N. Spergel, Phys. Rev. D **63**, 063002 (2001).
[83] M. Shiraishi, D. Nitta, S. Yokoyama, K. Ichiki, and K. Takahashi, Prog. Theor. Phys. **125**, 795 (2011).
[84] M. Bucher, B. Racine, and B. van Tent, J. Cosmol. Astropart. Phys. 05 (2016) 055.
[85] J. R. Fergusson and E. P. S. Shellard, Phys. Rev. D **80**, 043510 (2009).
[86] J. R. Fergusson and E. P. S. Shellard, Phys. Rev. D **76**, 083523 (2007).
[87] J. R. Fergusson, M. Liguori, and E. P. S. Shellard, Phys. Rev. D **82**, 023502 (2010).
[88] Planck Collaboration, Astron. Astrophys. **571**, A24 (2014).
[89] Planck Collaboration, Astron. Astrophys. **594**, A17 (2016).
[90] M. Shiraishi, M. Liguori, and J. R. Fergusson, J. Cosmol. Astropart. Phys. 05 (2014) 008.
[91] W. R. Coulton *et al.*, J. Cosmol. Astropart. Phys. 09 (2018) 022.
[92] G. Franciolini, A. Kehagias, A. Riotto, and M. Shiraishi, Phys. Rev. D **98**, 043533 (2018).
[93] M. Shiraishi, E. Komatsu, and M. Peloso, J. Cosmol. Astropart. Phys. 04 (2014) 027.
[94] M. Shiraishi, A. Ricciardone, and S. Saga, J. Cosmol. Astropart. Phys. 11 (2013) 051.
[95] J. L. Cook and L. Sorbo, J. Cosmol. Astropart. Phys. 11 (2013) 047.
[96] E. Dimastrogiovanni, M. Fasiello, R. J. Hardwick, H. Assadullahi, K. Koyama, and D. Wands, J. Cosmol. Astropart. Phys. 11 (2018) 029.
[97] N. Bartolo, G. Orlando, and M. Shiraishi, J. Cosmol. Astropart. Phys. 01 (2019) 050.
[98] E. Dimastrogiovanni, M. Fasiello, and G. Tasinato, J. Cosmol. Astropart. Phys. 08 (2018) 016.
[99] D. H. Lyth and A. R. Liddle, *The Primordial Density Perturbation: Cosmology, Inflation and the Origin of Structure* (Cambridge University Press, Cambridge, England, 2009).
[100] J. M. Bardeen, Phys. Rev. D **22**, 1882 (1980).
[101] U. Seljak and M. Zaldarriaga, Astrophys. J. **469**, 437 (1996).
[102] A. G. Polnarev, Sov. Astron. **29**, 607 (1985).
[103] S. Weinberg, Phys. Rev. D **69**, 023503 (2004).
[104] A. Lewis, A. Challinor, and A. Lasenby, Astrophys. J. **538**, 473 (2000).
[105] C. Howlett, A. Lewis, A. Hall, and A. Challinor, J. Cosmol. Astropart. Phys. 04 (2012) 027.
[106] D. Blas, J. Lesgourgues, and T. Tram, J. Cosmol. Astropart. Phys. 07 (2011) 034.
[107] M. Zaldarriaga and U. Seljak, Phys. Rev. D **58**, 023003 (1998).
[108] M. Beneke, C. Fidler, and K. Klingmüller, J. Cosmol. Astropart. Phys. 04 (2011) 008.
[109] S. Mollerach, D. Harari, and S. Matarrese, Phys. Rev. D **69**, 063002 (2004).
[110] C. Fidler, G. W. Pettinari, M. Beneke, R. Crittenden, K. Koyama, and D. Wands, J. Cosmol. Astropart. Phys. 07 (2014) 011.
[111] D. N. Spergel and D. M. Goldberg, Phys. Rev. D **59**, 103001 (1999).
[112] W. Hu, Phys. Rev. D **64**, 083005 (2001).
[113] M. Kamionkowski and T. Souradeep, Phys. Rev. D **83**, 027301 (2011).
[114] L. Wang and M. Kamionkowski, Phys. Rev. D **61**, 063504 (2000).
[115] D. Babich, Phys. Rev. D **72**, 043003 (2005).
[116] P. Creminelli, L. Senatore, and M. Zaldarriaga, J. Cosmol. Astropart. Phys. 03 (2007) 019.
[117] M. Liguori, A. Yadav, F. K. Hansen, E. Komatsu, S. Matarrese, and B. Wandelt, Phys. Rev. D **76**, 105016 (2007).
[118] K. M. Smith, O. Zahn, and O. Doré, Phys. Rev. D **76**, 043510 (2007).
[119] K. M. Smith and M. Zaldarriaga, Mon. Not. R. Astron. Soc. **417**, 2 (2011).
[120] L. Senatore, K. M. Smith, and M. Zaldarriaga, J. Cosmol. Astropart. Phys. 01 (2010) 028.
[121] W. Hu and M. J. White, Phys. Rev. D **56**, 596 (1997).
[122] L. Dai, M. Kamionkowski, and D. Jeong, Phys. Rev. D **86**, 125013 (2012).
[123] L. Dai, D. Jeong, and M. Kamionkowski, Phys. Rev. D **87**, 043504 (2013).
[124] A. J. S. Hamilton, Mon. Not. R. Astron. Soc. **312**, 257 (2000).
[125] J. C. Hill, Phys. Rev. D **98**, 083542 (2018).

[126] S. K. Lam, A. Pitrou, and S. Seibert, in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, LLVM '15 (ACM, New York, 2015), pp. 7:1–7:6.

[127] H. T. Johansson and C. Forssén, SIAM J. Sci. Stat. Comput. **38**, A376 (2016).

[128] Planck Collaboration, arXiv:1807.06206.

[129] Planck Collaboration, arXiv:1807.06207.

[130] U. Seljak and C. M. Hirata, Phys. Rev. D **69**, 043005 (2004).

[131] A. Lewis and A. Challinor, Phys. Rep. **429**, 1 (2006).

[132] Planck Collaboration, arXiv:1801.04945.

[133] W. R. Coulton and D. N. Spergel, J. Cosmol. Astropart. Phys. 10 (2019) 056.

[134] G. Jung, B. Racine, and B. van Tent, J. Cosmol. Astropart. Phys. 11 (2018) 047.

[135] D. M. Goldberg and D. N. Spergel, Phys. Rev. D **59**, 103002 (1999).

[136] A. Lewis, A. Challinor, and D. Hanson, J. Cosmol. Astropart. Phys. 03 (2011) 018.

[137] W. Hu, Phys. Rev. D **62**, 043007 (2000).

[138] C. Dvorkin, W. Hu, and K. M. Smith, Phys. Rev. D **79**, 107302 (2009).

[139] M. Kamionkowski and A. Loeb, Phys. Rev. D **56**, 4511 (1997).

[140] A.-S. Deutsch, M. C. Johnson, M. Münchmeyer, and A. Terrana, J. Cosmol. Astropart. Phys. 04 (2018) 034.

[141] A. Curto, M. Tucci, J. Gonzalez-Nuevo, L. Toffolatti, E. Martinez-Gonzalez, F. Argueso, A. Lapi, and M. Lopez-Caniego, Mon. Not. R. Astron. Soc. **432**, 728 (2013).

[142] T. Okamoto and W. Hu, Phys. Rev. D **67**, 083002 (2003).

[143] S. Dodelson, E. Rozo, and A. Stebbins, Phys. Rev. Lett. **91**, 021301 (2003).

[144] A. Cooray, M. Kamionkowski, and R. R. Caldwell, Phys. Rev. D **71**, 123527 (2005).

[145] E. Alizadeh and C. M. Hirata, Phys. Rev. D **85**, 123540 (2012).

[146] A.-S. Deutsch, E. Dimastrogiovanni, M. Fasiello, M. C. Johnson, and M. Münchmeyer, Phys. Rev. D **100**, 083538 (2019).

[147] L. Dai, D. Jeong, and M. Kamionkowski, Phys. Rev. D **88**, 043507 (2013).

[148] D. Babich and M. Zaldarriaga, Phys. Rev. D **70**, 083005 (2004).

[149] D. Green, J. Meyers, and A. van Engelen, J. Cosmol. Astropart. Phys. 12 (2017) 005.

[150] M. Shiraishi, D. Nitta, and S. Yokoyama, Prog. Theor. Phys. **126**, 937 (2011).

[151] E. T. Newman and R. Penrose, J. Math. Phys. (N.Y.) **7**, 863 (1966).

[152] J. N. Goldberg, A. J. Macfarlane, E. T. Newman, F. Rohrlich, and E. C. G. Sudarshan, J. Math. Phys. (N.Y.) **8**, 2155 (1967).

[153] A. R. Edmonds, *Angular Momentum in Quantum Mechanics* (Princeton University Press, Princeton, NJ, 1957).

[154] R. Mehrem, Appl. Math. Comput. **217**, 5360 (2011).

[155] *Data Analysis in Cosmology*, edited by V. J. Martínez, E. Saar, E. Martínez-González, and M.-J. Pons-Bordería, Lecture Notes in Physics Vol. 665 (Springer, New York, 2009).

[156] A. Taylor and P. Watts, Mon. Not. R. Astron. Soc. **328**, 1027 (2001).

[157] E. Sellentin, A. H. Jaffe, and A. F. Heavens, arXiv:1709.03452.

[158] F. Elsner and B. D. Wandelt, Astrophys. J. **724**, 1262 (2010).

[159] I. Kayo, M. Takada, and B. Jain, Mon. Not. R. Astron. Soc. **429**, 344 (2013).