High-resolution CMB bispectrum estimator with flexible modal bases

Wuhyun Sohn[*]

Korea Astronomy and Space Science Institute, Daejeon 34055, South Korea

James R. Fergusson[†] and E. P. S. Shellard[‡]

Centre for Theoretical Cosmology, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, United Kingdom

(Received 1 June 2023; accepted 9 August 2023; published 7 September 2023)

We present a new independent pipeline for the Cosmic Microwave Background bispectrum estimation of primordial non-Gaussianity and release a public code for constraining bispectrum shapes of interest based on the Planck 2018 temperature and polarization data. The estimator combines the strengths of the conventional Komatsu-Spergel-Wandelt and modal estimators at the cost of increased computational complexity, which has been made manageable through intensive algorithmic and implementation optimization. We also detail some methodological advances in numerical integration over a tetrapyd—domain where the bispectrum is defined on—via new quadrature rules. The pipeline has been validated both internally and against Planck. As a proof-of-concept example, we constrain some highly oscillatory models that were out of reach in conventional analyses using a targeted basis with a fixed oscillation frequency, and no significant evidence for primordial non-Gaussianity of these shapes is found. The methodology and code developed in this work will be directly applicable to future surveys where we expect a notable boost in sensitivity.

DOI: 10.1103/PhysRevD.108.063504

I. INTRODUCTION

To what degree the primordial perturbations are, if at all, non-Gaussian, is a key question of cosmology with many implications for early Universe physics [1–9]. Most theoretically well-motivated models predict various amplitudes and shapes of primordial non-Gaussianity (PNG, see e.g. the reviews of [1–4]), while the simplest single-field slow-roll inflation predicts an undetectably small PNG [8]. In a non-Gaussian universe, statistics beyond the two-point function are required to capture the full statistical information. The primordial bispectrum, or the harmonic counterpart of the three-point function, is the next leading-order measure of PNG. All cosmological datasets at present are consistent with vanishing PNG [10–15]. Among them, Planck's Cosmic Microwave Background (CMB) bispectrum analysis has placed the most stringent constraints on PNG to date [12].

The CMB bispectrum estimation of PNG, often parametrized by its amplitude $f_{\rm NL}$, is a computationally challenging task. A naïve computation is prohibitively expensive, so all existing implementations utilize some simplifying techniques. The Komatsu-Spergel-Wandelt (KSW) or KSW-like estimators [16–20] and skew-C_ℓ statistics [21] utilize separable (or factorizable) bispectrum templates which dramatically simplifies the integrals involved. The modal estimator [22–24] uses separable modal expansions of the primordial and the late-time bispectrum. The binned bispectrum estimator [25,26] compresses the data by binning the bispectrum in harmonic space with minimal loss of optimality.

A great variety of physically well-motivated models have been tested by the Planck Collaboration [10–12] using these estimators. Among them are models with oscillations in the bispectrum induced by, e.g., features in the inflationary potential [27–34], a transient reduction in the speed of sound [35–39], multifield dynamics [40–43], or resonances arising in axion monodromy models [44–50] (see e.g., [2,51–53] for reviews). Many of these models predict some enveloped oscillations in the primordial power spectrum and/or bispectrum that are linearly or logarithmically spaced. Previous works have placed constraints on these models using the CMB power spectrum [54–70], bispectrum [10–12,71–74], both in a joint analysis [75–77], and in combination with or solely from large-scale structure data [78–82]. However, despite its significant implications for early Universe physics, constraining highly oscillatory bispectrum shapes with general (nonseparable) envelopes has been out of reach using conventional methods due to computational challenges [12]. There remains a variety of such models that are yet unconstrained or only partially constrained.

Furthermore, to the authors' knowledge, there currently is no publicly available code that allows one to get Planck CMB

[*] wuhyun@kasi.re.kr
[†] jf334@cam.ac.uk
[‡] epss@damtp.cam.ac.uk

constraints on a general bispectrum shape of interest.[1] Having access to CMB bispectrum constraints would greatly benefit researchers in testing various models of interest.

Motivated by these reasons, we developed an independent bispectrum estimation pipeline CMB-BEST (a shorthand for CMB Bispectrum ESTimator). The method generalizes the KSW estimator by allowing a flexible choice of basis functions, which is used for a modal-like expansion of general bispectrum templates. It combines the accuracy of KSW with the broad applicability of modal at the cost of increased computational cost. We have extensively optimized the algorithm and implementation of CMB-BEST to make computation manageable. We have also thoroughly tested the code for self-consistency and against Planck's modal estimator. Some new constraints on a class of highly oscillatory templates, as well as reproductions of the Planck 2018 analyses, are presented in this paper as proof-of-concept examples.

We publicly release the frontend of our code as a Python package named CMBBEST [85], where users can provide general bispectrum shapes of interest and obtain Planck 2018 constraints on the corresponding $f_{\rm NL}$'s in a matter of seconds, or minutes for a higher-resolution basis, on a laptop. This was made possible by precomputing the computationally expensive parts of the pipeline on supercomputing clusters and providing the results as a data file. We plan on providing more basis function choices and future survey data in time.

There have been some methodological advances during the development of CMB-BEST. This paper includes our novel quadrature rule for efficient numerical integration of functions over a "tetrapyd" domain where the bispectrum is defined. The method is much more efficient and accurate than the simple 3D trapezoidal rule and is expected to benefit some numerical analyses of the large-scale structure as well. The method is implemented and shared as a Python package TETRAQUAD [86].

Upcoming CMB experiments such as the Simons Observatory [87] and CMB-S4 [88,89] are expected to dramatically improve the sensitivity in polarization measurements. The future datasets, in combination with the existing ones, will provide constraints on PNG that are almost as stringent as they can be from the CMB alone [89]. The methodology and code developed here will be directly applicable to upcoming surveys.

This paper is organized as follows. In Sec. II, we review the CMB bispectrum estimation procedure. Section III details the main formalism of CMB-BEST and various basis function choices made in this work. Section IV introduces our novel numerical method for evaluating integrals over a tetrapyd domain. Section V details the public release of our code and provides some proof-of-concept examples of CMB-BEST. Appendixes contain some computational details and various consistency checks. The conclusion is given in Sec. VI.

II. CMB BISPECTRUM ESTIMATION

In this section, we review how the CMB bispectrum can be used to study PNG. We provide a derivation of the CMB bispectrum estimator from the CMB bispectrum likelihood written analogously to [4,77,90]. We note that the core ideas behind this formalism are heavily based on the original works on the topic [16–18,20,90–93], and the formalism is mathematically equivalent to the ones described in Planck PNG papers [10–12]. This way of presentation was chosen to clearly state the assumptions we make to write down the CMB bispectrum estimator and to draw parallels to the CMB power spectrum likelihoods. We consider only the CMB temperature here for simplicity, but the formalism can be extended to include polarization, as described in Appendix C.

A. CMB bispectrum likelihood

CMB observations provide us with the spherical harmonic coefficients $a_{\ell m}$ of the temperature or polarization anisotropy maps. We are interested in their statistical properties, most of which are captured by their angular power spectrum estimated via

$$\hat{C}_\ell \equiv \sum_m \frac{1}{2\ell+1} a_{\ell m} a^*_{\ell m}. \tag{1}$$

This is a sufficient statistic only if the $a_{\ell m}$'s are Gaussian distributed. Going 1 order beyond the power spectrum, we can construct an estimate for the bispectrum as

$$\hat{B}_{\ell_1\ell_2\ell_3} \equiv \sum_{m_j} \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix} \left[a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3} - [\langle a_{\ell_1 m_1} a_{\ell_2 m_2} \rangle_{\rm MC} a_{\ell_3 m_3} + (2\,{\rm cyc})] \right]. \tag{2}$$

The weights are the Wigner-3j symbols related to angular momentum conservation; $\hat{B}_{\ell_1\ell_2\ell_3}$ vanishes unless ℓ_1, ℓ_2, and ℓ_3 form a triangle. The two terms in square brackets, cubic and linear in $a_{\ell m}$'s, together estimate the full bispectrum $\langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3} \rangle$. The linear term is necessary to reduce the variance of the estimate and keep it unbiased in the presence of an anisotropic sky mask, which introduces extra off diagonal correlations in the power spectrum which correlate with $a_{\ell m}$'s. Gaussian simulations are used for Monte Carlo approximations of the full

[1] Existing codes such as [83] and [84] do not apply to general shapes.

covariance appearing in the linear term. $\hat{B}_{\ell_1\ell_2\ell_3}$ can be understood as a summary statistic for the CMB 3-point functions under the assumption of statistical isotropy.

We define the theoretical bispectrum as the expected value of the estimate

$$B^{\mathrm{th}}_{\ell_1\ell_2\ell_3} \equiv \langle \hat{B}_{\ell_1\ell_2\ell_3} \rangle = h_{\ell_1\ell_2\ell_3} b^{\mathrm{th}}_{\ell_1\ell_2\ell_3}, \tag{3}$$

where $h_{\ell_1\ell_2\ell_3}$ is a geometric factor [(A3) in Appendix A] and $b_{\ell_1\ell_2\ell_3}$ is the reduced bispectrum.

In order to write down the likelihood, we make the following approximations:

(1) $\{\hat{B}_{\ell_1\ell_2\ell_3}\}_{\ell_1\leq\ell_2\leq\ell_3}$ are multivariate normal distributed.

(2) $\{\hat{B}_{\ell_1\ell_2\ell_3}\}_{\ell_1\leq\ell_2\leq\ell_3}$ have a diagonal covariance matrix with diagonal entries $(6/\Delta_{\ell_1\ell_2\ell_3})C_{\ell_1}C_{\ell_2}C_{\ell_3}$.

Here, the symmetry factor $\Delta_{\ell_1\ell_2\ell_3} = 6, 3, 1$ when there are 3, 2, 1 distinct values of l_j's, respectively.

The first approximation is justified by the central limit theorem, analogously to the power spectrum analysis. There are $O(\ell_1\ell_2)$ independent terms appearing in (2) for each (ℓ_1, ℓ_2, ℓ_3), so the weighted sum follows a near-Gaussian distribution. We note that this approximation can be inaccurate for a handful of terms for which ℓ_1 and ℓ_2 are small (and so are ℓ_3 by the triangle condition).

Omitting the nondiagonal terms in the covariance matrix is a choice to reduce computational complexity and was studied in [17,20]. The values of diagonal entries then follow by Wick's theorem in the limit where $a_{\ell m}$'s are only weakly non-Gaussian. The equivalent case for polarization is given in Appendix C.

We note that, for a simple estimation of f_{NL} using a single bispectrum template, the second approximation only affects the optimality of the estimator. The optimality of the estimator has been tested thoroughly in Planck analysis [10–12] using multiple pipelines and is shown to be near optimal. This approximation therefore would have little effect on our analysis. However, for future CMB surveys which will grant access to higher ℓ's, this may no longer be the case. Contributions from the connected 4-point functions due to CMB lensing studied in [94], for example, are expected to be more significant, and the approximation above would underestimate the true covariance. Delensing has been proposed as a solution in [94].

Under the assumptions above, our statistical model is

$$\hat{\mathbf{B}} = \mathbf{B}^{\mathrm{th}} + \boldsymbol{\epsilon}, \tag{4}$$

where the errors $\epsilon_{\ell_1\ell_2\ell_3}$ are multivariate normal with diagonal covariance. The CMB bispectrum likelihood is thus given by[2]

$$\mathcal{L}(\mathbf{B}^{\mathrm{th}}|\hat{\mathbf{B}}) = A\ \exp\Bigg[-\frac{1}{2}\sum_{\ell_1\leq\ell_2\leq\ell_3}\frac{\Delta_{\ell_1\ell_2\ell_3}}{6C_{\ell_1}C_{\ell_2}C_{\ell_3}} \times \left(\hat{B}_{\ell_1\ell_2\ell_3} - B^{\mathrm{th}}_{\ell_1\ell_2\ell_3}\right)^2\Bigg]. \tag{5}$$

The normalization constant $A = (2\pi)^{-n_B/2}\prod[6C_{\ell_1}C_{\ell_2}C_{\ell_3})/\Delta_{\ell_1\ell_2\ell_3}]^{-1/2}$, where n_B denotes the number of ordered triplets (ℓ_1, ℓ_2, ℓ_3) in consideration.

B. CMB bispectrum estimator of f_{NL}

In power spectrum analyses, ΛCDM or other models provide theoretical predictions of $C_\ell(\boldsymbol{\theta})$'s through Boltzmann solvers such as CAMB [95] and CLASS [96], where $\boldsymbol{\theta}$ denotes the model parameters. The likelihood, combined with some priors on the parameters, then provides a posterior distribution which places bounds on the parameters. A similar analysis can be done for the bispectrum likelihood with any n-parameter model $\mathbf{B}^{\mathrm{th}}(\boldsymbol{\theta})$. However, since the signal-to-noise ratio of the CMB bispectrum is much smaller, the constraining power is often limited. Instead, we study some bispectrum templates that are well-motivated by models of the early Universe and perform a simple linear fit, which constrains the amplitude of PNG. In our work, the cosmological parameters are fixed to their best-fit values at the power spectrum level, since their variation does not significantly affect the bispectrum amplitude [11].

The bispectrum of the curvature perturbations at the end of inflation is defined through

$$\langle\zeta(\mathbf{k}_1)\zeta(\mathbf{k}_2)\zeta(\mathbf{k}_3)\rangle = (2\pi)^3\delta^{(3)}(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3) \times B_\zeta(k_1, k_2, k_3). \tag{6}$$

Given one or more bispectrum templates $B^{(i)}_\zeta(k_1, k_2, k_3)$, we introduce some free parameters $f^{(i)}_{\mathrm{NL}}$ to represent the amplitude of PNG,

$$B^{\mathrm{th}}_{\ell_1\ell_2\ell_3}(\mathbf{f}_{\mathrm{NL}}) = \sum_i f^{(i)}_{\mathrm{NL}} B^{(i)}_{\ell_1\ell_2\ell_3} = \sum_i f^{(i)}_{\mathrm{NL}} h_{\ell_1\ell_2\ell_3} b^{(i)}_{\ell_1\ell_2\ell_3}, \tag{7}$$

where the reduced CMB bispectra are related to their primordial counterparts through

$$b^{(i)}_{\ell_1\ell_2\ell_3} = \left(\frac{2}{\pi}\right)^3 \int dr dk_1 dk_2 dk_3 (rk_1k_2k_3)^2 B^{(i)}_\zeta(k_1, k_2, k_3) \times \prod_{j=1}^{3}[j_{\ell_j}(k_j r) T_{\ell_j}(k_j)]. \tag{8}$$

Here, $j_\ell(k)$ denotes the spherical Bessel function. The CMB transfer functions $T_\ell(k)$ are obtained from the background cosmology which is fixed as ΛCDM best-fit

[2] We denote the likelihood as $\mathcal{L}(\boldsymbol{\theta}|\mathrm{data}) \equiv P(\mathrm{data}|\boldsymbol{\theta})$, which is the probability of observing the data given the model with parameters $\boldsymbol{\theta}$.

parameters to the CMB power spectrum likelihood. For Planck data, the bispectrum analyses are insensitive to this choice of background parameters [10–12].

In most cases, a model of inflation predicts a bispectrum whose shape is approximated by a single template, so an independent analysis with a single f_{NL} parameter suffices. However, if a class of models predicts bispectra that are expressed as linear combinations of two or more templates, then it is appropriate to have a joint analysis with multiple f_{NL} parameters at once. The Planck team provides joint constraints to the equilateral and orthogonal shapes [12], for example, since general single-field inflation models often yield a combination of the two.

The CMB bispectrum likelihood is then a function of $f_{\mathrm{NL}}^{(i)}$'s:

$$-2\ln\mathcal{L}(\mathbf{f}_{\mathrm{NL}}|\hat{\mathbf{B}}) = (\mathrm{const}) - 2\sum_i S_i f_{\mathrm{NL}}^{(i)} + \sum_{i,j} F_{ij} f_{\mathrm{NL}}^{(i)} f_{\mathrm{NL}}^{(j)}, \tag{9}$$

where we have defined

$$S_i \equiv \sum_{\ell_1\le\ell_2\le\ell_3} \frac{\Delta_{\ell_1\ell_2\ell_3}}{6C_{\ell_1}C_{\ell_2}C_{\ell_3}} \hat{B}_{\ell_1\ell_2\ell_3} B^{(i)}_{\ell_1\ell_2\ell_3} \tag{10}$$

$$= \sum_{\ell_j,m_j} \frac{\mathcal{G}^{\ell_1\ell_2\ell_3}_{m_1m_2m_3} b^{(i)}_{\ell_1\ell_2\ell_3}}{6C_{\ell_1}C_{\ell_2}C_{\ell_3}} \times \left[a_{\ell_1m_1}a_{\ell_2m_2}a_{\ell_3m_3} - [\langle a_{\ell_1m_1}a_{\ell_2m_2}\rangle a_{\ell_3m_3} + (2\,\mathrm{cyc})]\right], \tag{11}$$

$$F_{ij} \equiv \sum_{\ell_1\le\ell_2\le\ell_3} \frac{\Delta_{\ell_1\ell_2\ell_3}}{6C_{\ell_1}C_{\ell_2}C_{\ell_3}} B^{(i)}_{\ell_1\ell_2\ell_3} B^{(j)}_{\ell_1\ell_2\ell_3} \tag{12}$$

$$= \sum_{\ell_1,\ell_2,\ell_3} \frac{h^2_{\ell_1\ell_2\ell_3} b^{(i)}_{\ell_1\ell_2\ell_3} b^{(j)}_{\ell_1\ell_2\ell_3}}{6C_{\ell_1}C_{\ell_2}C_{\ell_3}}. \tag{13}$$

We have replaced the sums over $\ell_1 \le \ell_2 \le \ell_3$ with the ones over ℓ_1, ℓ_2, ℓ_3 using the symmetry factor $\Delta_{\ell_1\ell_2\ell_3}$ above. The Gaunt integral comes from the geometric factors given in (A3).

The CMB bispectrum estimator is the maximum likelihood estimator (MLE) of (9):

$$\widehat{f_{\mathrm{NL}}}^{(i)} = \sum_j (F^{-1})_{ij} S_j. \tag{14}$$

Assuming that the assumptions made above are valid, the MLE is unbiased so that $\langle \widehat{f_{\mathrm{NL}}}^{(i)} \rangle = f_{\mathrm{NL}}^{(i)}$. Furthermore, the estimator is optimal by the Gauss-Markov theorem; it has the smallest variance amongst all unbiased estimators constructed from $\hat{B}_{\ell_1\ell_2\ell_3}$. Its variance is then given by the Cramér-Rao bound which is expressed in terms of the Fisher information matrix F_{ij} as

$$\mathrm{Cov}(\widehat{\mathbf{f}_{\mathrm{NL}}}) = F^{-1}. \tag{15}$$

Therefore, the marginal error on the parameter $f_{\mathrm{NL}}^{(i)}$ is equal to $\sigma(f_{\mathrm{NL}}^{(i)}) = (F^{-1})_{ii}$ (no summation implied). Note that this is in general different from an independent analysis with one bispectrum template, for which $\sigma(f_{\mathrm{NL}}^{(i)}) = (F_{ii})^{-1}$ (no sum).

In practice, we use simulations (FFP10 end-to-end CMB maps [97,98]) with Gaussian initial conditions to obtain the sample variance of $f_{\mathrm{NL}}^{(i)}$, instead of directly using the Fisher error bar above. In a weakly non-Gaussian regime with a sufficient number of simulations, this variance accurately represents the variance of the estimator $\widehat{f_{\mathrm{NL}}}$. The simulations are also used to approximate the full nondiagonal covariance $\langle a_{\ell_1m_1}a_{\ell_2m_2}\rangle_{\mathrm{MC}}$ appearing in (11).

C. Beam, noise, sky mask, and lensing

Our observations of the CMB are subject to having a finite beam width, discrete pixelization, instrumental noise, partial sky coverage, and CMB weak lensing. We outline here how these effects are taken into account.

The window function W_ℓ is a product of the beam and pixel window functions. We use the Planck CMB maps produced using the SMICA component-separation method [99]. The maps have an effective beam full width at half maximum (FWHM) of 5 arcmin for temperature and 10 arcmin for polarization. Having a finite beam width limits the resolution and suppresses powers in high-ℓ multipoles. The beam window function is given by $\exp(-\ell(\ell+1)/2\sigma^2_{\mathrm{beam}})$, where $\sigma_{\mathrm{beam}} = (\mathrm{FWHM})/\sqrt{8\ln 2}$, for temperature and polarization. Discrete pixelizations also have a similar effect of suppressing small-scale powers. Using the HEALPix pixelization [100] with $N_{\mathrm{side}} = 2048$, the pixel window function is computed using the Python library HEALPY [101].

We obtain the power spectra of instrumental noise, N_ℓ, from 300 end-to-end simulated noise maps [97,98]. This yields consistent results when compared with the SMICA postcomponent-separation noise. We further assume that the temperature and E-mode polarization noises are uncorrelated and that their three-point functions vanish. The noise terms therefore only appear in C_ℓ^{TT} and C_ℓ^{EE}.

The sky masks for temperature and polarization maps with $f_{\mathrm{sky}}^{\mathrm{T}} = 0.779$ and $f_{\mathrm{sky}}^{\mathrm{E}} = 0.781$ [99] have been used for our analysis. Assuming statistical isotropy of the Universe, the expectation value $\langle\cdot\rangle$ gains a factor of f_{sky}. Furthermore, sharp discontinuities in the CMB map due to sky masks have been shown to cause numerical issues for the bispectrum analysis [102]. In particular, the small-scale powers boosted by sharp cut sky correlate with the large-scale multipoles from the mask shape, inducing a significant bias of local-type bispectrum. We therefore adopt the method of [102] and inpaint the CMB maps around the edges of sky masks via linear isotropic diffusion. Inpainting smooths the mask around the edges and hence transfers the mask-induced small-scale CMB powers to larger scales.

The CMB gets weakly lensed by matter as it travels from the last scattering surface to us. This weak lensing affects the CMB power spectrum by smoothing out the acoustic peaks slightly [103]. For the CMB bispectrum, small-scale powers induced by lensing correlate with the integrated Sachs-Wolfe (ISW) contributions to the large-scale modes, which creates a bias in the squeezed shape [104,105]. The lensing-ISW bias is given by [106]

$$b^{\text{lensing-ISW}}_{\ell_1\ell_2\ell_3} = \frac{1}{2}[\ell_1(\ell_1+1) - \ell_2(\ell_2+1) + \ell_3(\ell_3+1)] \times \tilde{C}^{\mathrm{TT}}_{\ell_1} C^{\mathrm{T}\phi}_{\ell_3} + (5\,\text{perms}), \tag{16}$$

where $C^{\mathrm{T}\phi}_{\ell}$ is the temperature and lensing potential cross power spectra. A tilde above C_ℓ signifies that it is a lensed quantity. We explicitly include the bias above in our analysis.

In summary, the statistical model (4) is modified as the following:

$$\hat{B}_{\ell_1\ell_2\ell_3} = f_{\text{sky}} W_{\ell_1} W_{\ell_2} W_{\ell_3} (B^{\text{th}}_{\ell_1\ell_2\ell_3} + B^{\text{lensing-ISW}}_{\ell_1\ell_2\ell_3}) + \epsilon_{\ell_1\ell_2\ell_3}, \tag{17}$$

where the error has a vanishing mean and a diagonal covariance equal to

$$\mathrm{Cov}(\epsilon_{\ell_1\ell_2\ell_3}) = f_{\text{sky}} \mathrm{diag}\left(\frac{6}{\Delta_{\ell_1\ell_2\ell_3}} (W^2_{\ell_1} C_{\ell_1} + N_{\ell_1})(W^2_{\ell_2} C_{\ell_2} + N_{\ell_2})(W^2_{\ell_3} C_{\ell_3} + N_{\ell_3})\right), \tag{18}$$

and $B^{\text{lensing-ISW}}_{\ell_1\ell_2\ell_3} = h_{\ell_1\ell_2\ell_3} b^{\text{lensing-ISW}}_{\ell_1\ell_2\ell_3}$.

III. HIGH-RESOLUTION CMB BISPECTRUM ESTIMATOR

Computation of the CMB bispectrum estimator (14) can be prohibitively expensive; the most naive method would require summing over $O(\ell^5_{\max}) \sim O(10^{16})$ terms. All existing techniques rely on one or more tricks and/or assumptions to simplify the computation process. The KSW estimator [16] utilizes separable bispectrum templates for which the three-dimensional integral $dk_1 dk_2 dk_3$ splits into a product of three separate one-dimensional integrals. The formalism is fast and efficient but is restricted to a limited range of separable templates. The modal estimator [22–24] expands the primordial and late-time bispectra in terms of separable basis functions. The bispectrum information is compressed with respect to the basis and stored. This allows fast and thorough analyses of general bispectrum templates. Lastly, the binned bispectrum estimator [25,26] bins the bispectrum into different ℓ bins, which makes the total size more computationally tractable with minimal loss of optimality. We refer to [11] and references therein for detailed reviews on the CMB bispectrum estimation.

In this section, we introduce our novel, independent CMB bispectrum estimator CMB-BEST, which combines ideas from the KSW estimator [16] and the modal estimator [22,23]. While being computationally more expensive than the two, CMB-BEST combines the best of both worlds to be general and efficient, and has the flexibility in the choice of basis for high-resolution analyses.

A. CMB-BEST formalism

In CMB-BEST, the primordial bispectra are expanded using separable basis functions similar to modal [22], followed by the compression of the CMB bispectrum information with respect to this basis in a way similar to the KSW estimator [16]. We detail the formalism here.

Given a choice of one-dimensional mode functions $q_p(k)$, a given primordial bispectrum template is expanded as

$$(k_1k_2k_3)^2 B_\zeta(k_1,k_2,k_3) = \sum_{p_1,p_2,p_3} \alpha_{p_1p_2p_3} q_{p_1}(k_1) q_{p_2}(k_2) q_{p_3}(k_3). \tag{19}$$

The expansion above holds for bispectrum templates that are accurately represented using the basis functions $Q_{p_1p_2p_3}(k_1,k_2,k_3) \equiv q_{p_1}(k_1)q_{p_2}(k_2)q_{p_3}(k_3)$.[3] Note that, by symmetry, the above is equivalent to having the sum over $p_1 \geq p_2 \geq p_3$ with an additional symmetry factor $\Delta_{p_1p_2p_3}$. This convention will be used in the next section, where different choices of mode functions and the expansion procedure are discussed. Here, there are no restrictions to each p_j which runs from 1 to $p_{\max}$, the number of mode functions.

The basis functions are products of terms that depend on only one of the k's. The reduced bispectrum simplifies thanks to this separability:

$$b_{\ell_1\ell_2\ell_3} = \sum_{p_j} \alpha_{p_1p_2p_3} \int dr \tilde{q}_{p_1}(\ell_1, r) \tilde{q}_{p_2}(\ell_2, r) \tilde{q}_{p_3}(\ell_3, r), \tag{20}$$

[3] The truncation error is small as long as B is within or close to the function space spanned by the basis functions $\{q_{p_1}(k_1)q_{p_2}(k_2)q_{p_3}(k_3)\}$.

where the projected mode functions are defined as

$$\tilde{q}_p(\ell, r) \equiv \frac{2r^{\frac{2}{3}}}{\pi} \int dk q_p(k) T_\ell(k) j_\ell(kr). \tag{21}$$

In the modal estimator [22], another set of mode functions is introduced in the ℓ space to form a late-time basis so that $b_{\ell_1\ell_2\ell_3} = \sum_{p_j} \tilde{\alpha}_{p_1p_2p_3} \tilde{Q}_{p_1p_2p_3}(\ell_1\ell_2\ell_3)$. This effectively removes the line-of-sight integral $\int dr$ appearing in (20) and reduces the computational complexity by a couple of orders of magnitude. The expansion accurately approximates most bispectrum templates; the correlation levels between the template and the basis expansion vary between ∼0.95 and ∼0.99 for modal analysis in Planck [10]. However, for bispectrum templates motivated by feature or resonance models, the modal estimator had limited coverage compared with the KSW-type estimators [12]. In CMB-BEST, we do not perform this second late-time basis expansion step and instead follow a KSW-like formalism to compute the estimator exactly, albeit with increased computational complexity.

We define the filtered maps from $a_{\ell m}$'s as

$$M_p^{(i)}(\hat{\mathbf{n}}, r) \equiv \sum_{\ell, m} \frac{\tilde{q}_p(l, r)}{C_l} a_{\ell m} Y_{\ell m}(\hat{\mathbf{n}}). \tag{22}$$

The observed map corresponds to $i = 0$ in our notation above, while $i = 1, \ldots, N_{\rm sim}$ signify the FFP10 end-to-end CMB and noise map numbers [97,98] under Gaussian initial conditions. As described in Sec. II, these simulations are used for two purposes: first to approximate the full covariance matrix $\langle a_{\ell_1 m_1} a_{\ell_2 m_2} \rangle_{\rm MC}$ appearing in the linear term of (11), and second to form a set of $f_{\rm NL}$ estimates which is used as a null test to evaluate the statistical significance of $f_{\rm NL}$ estimated from observations.

The CMB bispectrum estimator for a single bispectrum template can then be written as

$$\begin{aligned}\widehat{f_{\rm NL}}^{(i)} = \frac{1}{F} \sum_{\ell_j, m_j} \frac{\mathcal{G}^{\ell_1\ell_2\ell_3}_{m_1m_2m_3} b_{\ell_1\ell_2\ell_3}}{6 C_{\ell_1} C_{\ell_2} C_{\ell_3}} \Bigg[a^{(i)}_{\ell_1 m_1} a^{(i)}_{\ell_2 m_2} a^{(i)}_{\ell_3 m_3} \\ - \left(\frac{1}{N_{\rm sim} - 1} \sum_{j \neq i} a^{(j)}_{\ell_1 m_1} a^{(j)}_{\ell_2 m_2} a^{(i)}_{\ell_3 m_3} + (2\,\text{perms}) \right) \Bigg].\end{aligned} \tag{23}$$

$$= \frac{1}{6F} \sum_{p_1, p_2, p_3} \alpha_{p_1p_2p_3} \left[\beta^{{\rm cub},(i)}_{p_1p_2p_3} - 3 \beta^{{\rm lin},(i)}_{p_1p_2p_3} \right], \tag{24}$$

where

$$\beta^{{\rm cub},(i)}_{p_1p_2p_3} \equiv \int dr \int d^2\hat{\mathbf{n}} M^{(i)}_{p_1}(\hat{\mathbf{n}}, r) M^{(i)}_{p_2}(\hat{\mathbf{n}}, r) M^{(i)}_{p_3}(\hat{\mathbf{n}}, r), \tag{25}$$

$$\begin{aligned}\beta^{{\rm lin},(i)}_{p_1p_2p_3} \equiv \frac{1}{N_{\rm sim} - 1} \sum_{j \neq i} \int dr \int d^2\hat{\mathbf{n}} M^{(j)}_{p_1}(\hat{\mathbf{n}}, r) \\ \times M^{(j)}_{p_2}(\hat{\mathbf{n}}, r) M^{(i)}_{p_3}(\hat{\mathbf{n}}, r).\end{aligned} \tag{26}$$

Computing the $\beta^{\rm cub}$ and $\beta^{\rm lin}$ is the most time-consuming step of CMB-BEST. Once they are computed, any given bispectrum template can be constrained instantly after a primordial basis expansion, which normally takes less than a minute on a laptop.

The normalization, directly related to the Fisher information, is obtained similarly by exploiting separability and the relation (A8) for $h^2_{\ell_1\ell_2\ell_3}$;

$$F = \frac{1}{6} \sum_{p_j, p'_j} \alpha_{p_1p_2p_3} \Gamma_{p_1p_2p_3, p'_1p'_2p'_3} \alpha_{p'_1p'_2p'_3}, \tag{27}$$

where

$$\begin{aligned}\Gamma_{p_1p_2p_3, p'_1p'_2p'_3} \equiv \int dr \int dr' \int d\mu \gamma_{p_1p'_1}(\mu, r, r') \\ \times \gamma_{p_3p'_3}(\mu, r, r') \gamma_{p_3p'_3}(\mu, r, r'),\end{aligned} \tag{28}$$

$$\gamma_{pp'}(\mu, r, r') \equiv \sum_\ell \frac{2\ell + 1}{(8\pi)^{1/3} C_\ell} \tilde{q}_p(\ell, r) \tilde{q}_{p'}(\ell, r') P_\ell(\mu). \tag{29}$$

Here, $P_\ell(\mu)$'s denote the Legendre polynomials.

Lastly, we summarize the key differences of CMB-BEST with two of the main methods used in Planck in Table I. Note that the binned bispectrum estimator [25] is omitted in the table since its core idea is different and hard to compare with CMB-BEST, but it is one of the main approaches used extensively in the Planck analyses.

The relative computational complexities shown in Table I are rough order-of-magnitude estimates. Directly comparing the computational costs between methods is difficult for two reasons. First, the relative complexity between methods depends on many different factors such as the resolution (number of basis elements) and the number of inflation models under study. Modal and CMB-BEST scale cubically with the number of mode functions used, but only need to be run once per basis set to constrain a wide class of models simultaneously. While the KSW estimator is faster, it has to be run for each individual model. Second, the computational resources required to run the codes also depend strongly on their implementational optimizations. For example, the modal code has been improved by several orders of magnitude thanks to High-Performance Computing (HPC) optimization [107], and some simple cache optimization of CMB-BEST led to a factor of 2 improvement as detailed in Appendix E.

TABLE I. Comparison of CMB-BEST with the two conventional bispectrum estimators [16,22] used in Planck analyses [10–12]. The core ideas for data/information reduction for different approaches are detailed in the text. $B_\zeta(k_1,k_2,k_3)$ and $B^{\rm th}_{\ell_1\ell_2\ell_3}$ correspond to the primordial bispectrum template and its late-time harmonic counterpart, respectively.

		Estimation accuracy		Relative computational complexity (rough estimate)
	Core idea	Separable templates	Nonseparable templates	
KSW [Komatsu *et al.* 2005]	Use separable templates	Exact	Not applicable	∼1 per model
Modal [Fergusson *et al.* 2010]	Expand $B_\zeta(k_1,k_2,k_3)$ and $B^{\rm th}_{\ell_1\ell_2\ell_3}$ using separable basis functions	As good as the $B^{\rm th}_{\ell_1\ell_2\ell_3}$ expansion	As good as the $B^{\rm th}_{\ell_1\ell_2\ell_3}$ expansion	∼30
CMB-BEST [this work]	Expand $B_\zeta(k_1,k_2,k_3)$ using separable basis functions	Exact	As good as the $B_\zeta(k_1,k_2,k_3)$ expansion	∼10,000

With these caveats in mind, we quote the number of CPU core hours required for the baseline ($p_{\rm max}=10$) and high-resolution ($p_{\rm max}=30$) runs of CMB-BEST to be 3,000 and 80,000, respectively. Note that the public code we present in Sec. V runs in a matter of minutes on a laptop since the computationally heavy parts of the pipeline are precomputed and provided as a data file.

B. Basis expansions

Computing the quantities β (25) and Γ (28) are computationally expensive but need to be performed only once per data and basis set. Afterward, CMB-BEST can rapidly constrain bispectrum shapes of interest through a two-step procedure: (1) expand the shape function with respect to a separable basis, and (2) compute $f_{\rm NL}$ and $\sigma(f_{\rm NL})$ via some simple matrix multiplications of the expansion coefficients with β and Γ. We provide the precomputed data for various basis sets together with our public code (detailed in Sec. VA) for the two steps.

We expand a given bispectrum shape $S(k_1,k_2,k_3)=(k_1k_2k_3)^2B_\zeta(k_1,k_2,k_3)$ in terms of the basis functions as follows. First, we simplify our basis functions by symmetrizing over (k_1,k_2,k_3):

$$\sum_{p_1,p_2,p_3}\alpha_{p_1p_2p_3}q_{p_1}(k_1)q_{p_2}(k_2)q_{p_3}(k_3)=\sum_{p_1,p_2,p_3}\alpha_{p_1p_2p_3}\frac{1}{6}\left[q_{p_1}(k_1)q_{p_2}(k_2)q_{p_3}(k_3)+(5\,\text{perms})\right] \tag{30}$$

$$=\sum_{n\leftrightarrow(p_1,p_2,p_3)}\alpha_nQ_n(k_1,k_2,k_3), \tag{31}$$

where n is an index mapped one-to-one with a triplet (p_1,p_2,p_3) satisfying $p_1\geq p_2\geq p_3$. Note that $\alpha_n=\Delta_{p_1p_2p_3}\alpha_{p_1p_2p_3}$, which contains an extra symmetry factor.

To obtain α_n, we solve the following linear equation:

$$\sum_{n'}\langle Q_n,Q_{n'}\rangle\alpha_{n'}=\langle Q_n,S\rangle,\quad\text{where} \tag{32}$$

$$\langle f,g\rangle\equiv\int_{\mathcal{V}_{\rm T}}d^3\mathbf{k}w(\mathbf{k})f(\mathbf{k})g(\mathbf{k}). \tag{33}$$

The inner product $\langle\cdot,\cdot\rangle$ of two functions over the tetrapyd domain is defined for some weight function $w(\mathbf{k})$. Mathematically, solving (32) for α is equivalent to finding an orthogonal projection of S into the function space spanned by the basis functions $\{Q_n\}$. The truncation error of this basis representation comes from the component of S that is perpendicular to this function space.

For small basis sizes, we can directly invert the matrix $\Gamma_{nn'}\equiv\langle Q_n,Q_{n'}\rangle$ to solve the linear equation for α. However, this can become numerically unstable for larger bases since some basis functions become more degenerate, which degrades the condition number of Γ.[4] Instead of a direct inversion, we use the conjugate gradient method [108] to obtain an approximate solution for (32) with only marginal residual errors. This iterative method is applicable

[4] Note that even though the Legendre mode functions are orthogonal in one dimension, the three-dimensional basis functions constructed are no longer orthogonal on the (nonseparable) tetrapyd domain.

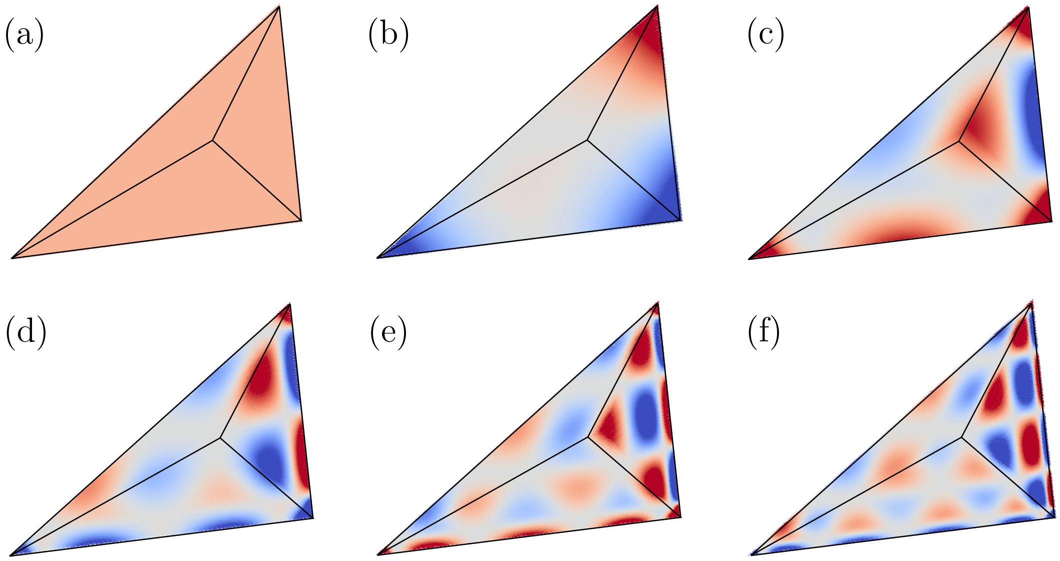


FIG. 1. Some examples of our Legendre basis functions, evaluated on a sliced tetrapyd domain with $k_1 \geq k_2 \geq k_3$. The basis functions are defined as $Q_{p_1p_2p_3}(k_1,k_2,k_3) \equiv q_{p_1}(k_1)q_{p_2}(k_2)q_{p_3}(k_3)$, where $q_p(k)$ are defined in (37). Here we plot $p_1 = p_2 = p_3 = p$, where p equals (a) 1, (b) 2, (c) 3, (d) 4, (e) 5, and (f) 6. A single color map is used across the plots: red and blue correspond to $+1$ and -1, respectively. Note that this figure has been borrowed from the first author's thesis of [110] under Open Access.

here because Γ is symmetric and positive definite by construction, and it can retrieve the exact solution when the matrix is accurately invertible. We found that the iterative algorithm implemented in the Python library SciPy [109] is numerically stable and efficient for our purposes.

C. Basis function choices

One of the greatest strengths of CMB-BEST lies in its flexibility with the choice of mode functions from which the basis set is constructed. Adopting a small set of specific mode functions provides fast and precise results for some specific bispectrum templates of interest. On the other hand, a general basis set allows us to constrain a broad range of inflationary models simultaneously but requires more computational resources upfront. In this section, we describe three types of basis sets studied in this work.

The first and simplest basis set consists of monomial mode functions of the form

$$q_p(k) = k^{p-1}, \quad \text{for } p = 0, 1, 2, 3. \tag{34}$$

The three most standard bispectrum templates—local, equilateral, and orthogonal—can be expressed as a sum of separable terms above. For example, the local template, for example, is given by

$$S^{\text{local}}(k_1k_2k_3) \equiv (k_1k_2k_3)^2 B_\Phi^{\text{local}}(k_1,k_2,k_3) = 2A^2\left[\frac{k_1^2}{k_2k_3} + \frac{k_2^2}{k_3k_1} + \frac{k_3^2}{k_1k_2}\right], \tag{35}$$

where A is the primordial gravitational potential (Φ) power spectrum amplitude, which relates to the usual scalar (curvature ζ) power spectrum amplitude through $A = 2\pi^2(3/5)^2 A_s$ due to different conventions.[5] Decomposition coefficients $\alpha_{p_1p_2p_3}$ for the local template are given exactly and have three nonzero components: $\alpha_{300} = \alpha_{030} = \alpha_{003} = 2A^2$. Coefficients for the equilateral and orthogonal templates are obtained in a similar fashion.[6]

Choosing the monomial functions and using the exact α's above render CMB-BEST to be completely equivalent to the KSW estimator. We will refer to this basis set as "Monomials."

For general analyses on a wide range of bispectrum templates, we restrict the k range to $[k_{\min}, k_{\max}]$ and define the following mode functions:

$$q_0(k) = k^{n_s-2}, \tag{36}$$

$$q_p(k) = P_p(\mu(k)), \quad p = 1, 2, \ldots, p_{\max} - 1, \quad \text{where } \mu(k) \equiv -1 + \frac{2(k - k_{\min})}{k_{\max} - k_{\min}}. \tag{37}$$

[5] The factor of $(3/5)^2$ comes from the fact that $\Phi = \frac{3}{5}\zeta$ at superhorizon scales. The other factor of $2\pi^2$ is from the relation between the power spectrum $P(k)$ and the dimensionless power spectrum $\mathcal{P}(k)$: $P(k) = (2\pi^2/k^3)\mathcal{P}(k)$.

[6] In practice, we account for a nonunit n_s by modifying the basis as $q_p(k) = k^2\left[k_*(k/k_*)^{(4-n_s)/3}\right]^{p-3}$ for $p = 0, 1, 2, 3$. The pivot scale $k_* = 0.05\ \text{Mpc}^{-1}$ so that $\mathcal{P}_\zeta(k) = A_s k^{n_s-1}$.

Many bispectrum templates have terms that depend on k^{-1} which is captured by $q_0(k)$. The number of modes, $p_{\max}$, determines the maximum order of the Legendre polynomials $P_p(\mu)$ included. Our baseline analysis uses $p_{\max} = 10$, while the high-resolution one has $p_{\max} = 30$. The latter yields 4960 basis functions in total after accounting for symmetries. We refer to this basis set as "Legendre." Some of the basis functions in the Legendre basis set are shown in Fig. 1.

As promised in the name "high-resolution CMB bispectrum estimator," CMB-BEST allows targeted analysis on complex, nonseparable bispectrum shapes with high-frequency oscillations, which has not been constrained so far due to computational challenges. Given an oscillation frequency ω_* of interest, a targeted oscillatory basis can be constructed as the following:

$$q_0(k) = k^{n_s - 2}\sin(\omega_* k), \quad q_1(k) = k^{n_s - 2}\cos(\omega_* k), \tag{38}$$

$$\begin{aligned} q_{2p}(k) &= P_p(\mu(k))\sin(\omega_* k), \\ q_{2p+1}(k) &= P_p(\mu(k))\cos(\omega_* k), \\ p &= 1, 2, \ldots, (p_{\max}/2) - 1. \end{aligned} \tag{39}$$

This basis set can be thought of as a tensor product between the Legendre basis set and $\{\sin(\omega_* k), \cos(\omega_* k)\}$. A bispectrum shape with linearly spaced oscillations and some envelope function f can be rewritten as

$$\begin{aligned} S(k_1, k_2, k_3) &= f(k_1, k_2, k_3)\sin(\omega_*(k_1 + k_2 + k_3) \\ &\quad + \phi) = \mathrm{Im}[e^{i\phi} f(k_1, k_2, k_3) e^{i\omega_* k_1} e^{i\omega_* k_2} e^{i\omega_* k_3}]. \end{aligned} \tag{40}$$

Therefore, the expansion coefficients of S with respect to the targeted oscillatory basis can be obtained by first expanding the envelope A using a Legendre basis of order $p_{\max}/2$ and then taking a tensor product with the oscillatory part: $-\alpha_{000} = \alpha_{011} = \alpha_{101} = \alpha_{110} = \cos\phi$ and $-\alpha_{001} = -\alpha_{010} = -\alpha_{100} = \alpha_{111} = \sin\phi$.

IV. NUMERICAL INTEGRATION OVER TETRAPYD

In this section, we present our novel method for a precise and efficient numerical integration over the "tetrapyd" volume, which appears frequently in bispectrum analyses including CMB-BEST. The method shows excellent performance, especially for integrands that are well approximated by polynomials, showing orders of magnitude improvement in precision with many fewer evaluation points compared with simple trapezoidal rule, as will be discussed below.

A. Tetrapyd quadrature

The primordial bispectrum is defined on a "tetrapyd" domain specified by triangle inequalities and the observational limits of the k range

$$\begin{aligned} \mathcal{V}_{\mathrm{T}}(k_{\min}, k_{\max}) \equiv \{&(k_1, k_2, k_3){:}\, 2\max\{k_1, k_2, k_3\} \le k_1 \\ &+ k_2 + k_3 \text{ and } k_{\min} \le k_1, k_2, k_3 \le k_{\max}\}. \end{aligned} \tag{41}$$

Volume integrals over this domain are not separable; they cannot be rewritten as three independent one-dimensional integrals, unlike the integral over the cube $[k_{\min}, k_{\max}]^3$. It is therefore difficult to simplify these integrals without introducing one or more extra integration variables.

The simplest way to numerically compute the volume integrals over the tetrapyd domain is, therefore, to approximate it with a weighted sum of the integrand evaluated at a finite number of points:

$$\int_{\mathcal{V}_{\mathrm{T}}(k_{\min}, k_{\max})} dV f(\mathbf{k}) \approx \sum_{n=1}^{N} w_n f(\mathbf{x}_n). \tag{42}$$

Such a method of numerical integration is often referred to as quadrature, or cubature in this case since it is a volume integral [111].

One of the simplest quadratures in one dimension is the trapezoidal rule with a uniformly spaced grid. Similarly, one can create a uniform grid inside the three-dimensional tetrapyd. Each grid point is weighted proportional to what fraction of its voxel (volume pixel: a cube with the grid point in the center and shares sides with neighboring grid points) intersects with the tetrapyd. This approach can yield robust results for various integrands and is sensitive to sharp local oscillations if there are any. However, it is computationally expensive to achieve high numerical precision using this method, because the number of grid points required for a three-dimensional volume scales up rapidly with the grid density.

Efficient quadrature rules for various three-dimensional volumes including spheres, tetrahedra and pyramids have been studied (e.g., [111,112]), but not for a rather complex shape of tetrapyd, to the authors' knowledge. Inspired by the Gaussian quadrature rules that yield exact results for the first M orthogonal polynomials, we seek a (approximate) tetrapyd quadrature satisfying the following conditions:

1. Polynomials $P(k_1, k_2, k_3) = k_1^p k_2^q k_3^r$ are evaluated almost exactly for $p + q + r < M$, for some M.
2. The nodes $\{\mathbf{x}_n\}$ and weights $\{w_n\}$ are invariant under the permutations of (k_1, k_2, k_3), which are the symmetries that the tetrapyd volume enjoys.
3. The weights are non-negative ($w_n \ge 0$) for numerical stability.

Note that a quadrature rule over $\mathcal{V}_{\mathrm{T}}(k_{\min}, k_{\max})$ can be easily obtained from that of $\mathcal{V}_{\mathrm{T}}(k_{\min}/k_{\max}, 1)$ after a suitable rescaling. We define $k_{\mathrm{r}} \equiv k_{\min}/k_{\max}$. Next, by symmetry,

$$\int_{\mathcal{V}_{\mathrm{T}}(k_{\mathrm{r}},1)} dV f(k_1,k_2,k_3) = \frac{1}{6}\int_{\mathcal{V}_{\mathrm{T}}(k_{\mathrm{r}},1)} dV[f(k_1,k_2,k_3) + 5\,\text{perms}] \tag{43}$$

$$= \int_{\mathcal{V}_{\mathrm{T}/6}(k_{\mathrm{r}},1)} dV[f(k_1,k_2,k_3) + 5\,\text{perms}], \tag{44}$$

where

$$\mathcal{V}_{\mathrm{T}/6}(k_{\mathrm{r}},1) \equiv \{(k_1,k_2,k_3){:}\, k_1 \le k_2 + k_3 \text{ and } k_{\mathrm{r}} \le k_3 \le k_2 \le k_1 \le 1\}. \tag{45}$$

It is therefore sufficient to find a quadrature rule for the symmetric functions $f(k_1,k_2,k_3)$ over $\mathcal{V}_{\mathrm{T}/6}(k_{\mathrm{r}},1)$. The quadrature rule can be extended symmetrically anytime to cover $\mathcal{V}_{\mathrm{T}}(k_{\mathrm{r}},1)$ and satisfy the second condition above.

B. Orthogonal polynomials

To find a quadrature rule that exactly evaluates the integrals of polynomials over a tetrapyd, we first need an analytic formula for the integrals. We make use of the following expression for the integrals of $k_1^p k_2^q k_3^r$ over a unit tetrapyd:

$$\int_{\mathcal{V}_{\mathrm{T}}(0,1)} dV k_1^p k_2^q k_3^r = \frac{1}{(p+1)(q+1)(r+1)} - \left[\frac{\Gamma(1+q)\Gamma(1+r)}{(3+p+q+r)\Gamma(3+q+r)} + (2\,\text{perms})\right], \tag{46}$$

where $\Gamma(n)$ denotes the gamma function. We have also derived (by hand) the full analytical expression for $\mathcal{V}_{\mathrm{T}}(k_{\mathrm{r}},1)$, which generalizes (46). This result can be found in Appendix B.

Next, we define an inner product over the tetrapyd as

$$\langle f,g\rangle_{\mathcal{V}_{\mathrm{T}}(k_{\mathrm{r}},1)} \equiv \int_{\mathcal{V}_{\mathrm{T}}(k_{\mathrm{r}},1)} dV f(\mathbf{k})g(\mathbf{k}). \tag{47}$$

It is possible to use different weight functions for the integral, but the analytic formulas provided here will work only if they are of form $(k_1k_2k_3)^a$. We orthogonalize and normalize the symmetric polynomials with respect to this inner product using the modified Gram-Schmidt process (MGS). In our public code TETRAQUAD, MGS has been modified further to be slower but numerically more stable. The first four orthonormalized symmetric polynomials over $\mathcal{V}_{\mathrm{T}}(0.1,1)$, for example, are given as follows:

$$P_0(k_1,k_2,k_3) = 1.4540, \tag{48}$$

$$P_1(k_1,k_2,k_3) = 9.4663\frac{k_1+k_2+k_3}{3} - 5.6334, \tag{49}$$

$$P_2(k_1,k_2,k_3) = 43.945\frac{k_1^2+k_2^2+k_3^2}{3} - 51.129\frac{k_1+k_2+k_3}{3} + 12.439, \tag{50}$$

$$P_3(k_1,k_2,k_3) = 40.363\frac{k_1k_2+k_2k_3+k_3k_1}{3} - 13.116\frac{k_1^2+k_2^2+k_3^2}{3} - 29.586\frac{k_1+k_2+k_3}{3} + 8.3660. \tag{51}$$

By orthogonality, all P_d for $d>0$ integrate to zero over the tetrapyd. Therefore, an exact quadrature rule of order M should satisfy

$$\sum_n w_n P_d(\mathbf{x}_n) = 0 \quad \text{for } 0<d<M \quad \text{and} \quad \sum_n w_n = \mathrm{vol}(\mathcal{V}_{\mathrm{T}}(k_{\mathrm{r}},1)). \tag{52}$$

Note that the above would guarantee all symmetric polynomials of total order less than that of P_M to be evaluated exactly.

C. Finding approximate quadrature

In this work, we find an approximate quadrature rule by fixing some grid points $\{\mathbf{x}_n\}$ and solving the following nonnegative least squares problem (NNLS):

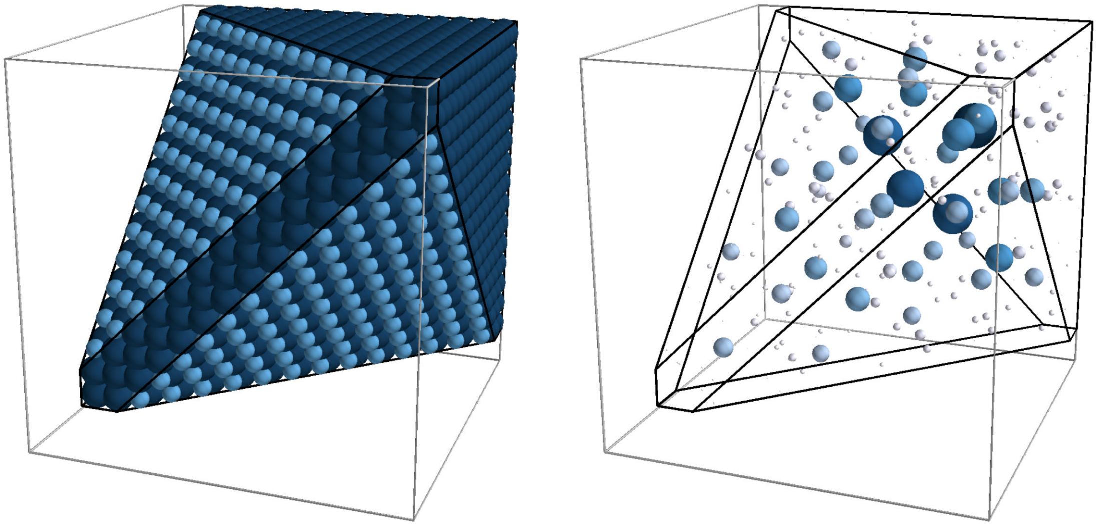


FIG. 2. Three-dimensional visualization of the numerical quadrature rules on the tetrapyd domain $\mathcal{V}_{\mathrm{T}}(0.1, 1)$ (black edges) inside a unit cube (gray edges). The position and the relative size of each sphere correspond to the given quadrature's evaluation point and its weight, respectively. Shown on the left side is the "uniform" quadrature with 2517 points which directly generalizes the trapezoidal rule. The right-hand plot is obtained TETRAQUAD ($N = 15$, $M = 40$), which achieves ~2 orders of magnitude higher precision with a smaller subset of 302 evaluation points.

$$\text{Minimize } L(\mathbf{w}) = \|P\mathbf{w} - \mathbf{a}\|^2 \text{ subject to } w_n \geq 0, \tag{53}$$

where the $M \times N$ matrix P satisfies $P_{dn} = P_d(\mathbf{x}_n)$, $\mathbf{w} = (w_1, w_2, \ldots, w_N)^{\mathrm{T}}$, and $\mathbf{a} = (\mathrm{vol}(\mathcal{V}_{\mathrm{T}}(k_{\mathrm{r}}, 1)), 0, 0, \ldots, 0)^{\mathrm{T}}$. Having $L(\mathbf{w}) = 0$ would mean that the polynomials $P_0, \ldots, P_{M-1}$ are evaluated exactly using the quadrature. In our code, we use the Python library SciPy [109] to solve NNLS using an active set method [113]. We will refer to this novel quadrature rule on tetrapyd as "Tetraquad."

The parameters M and N are free for us to choose. The numerical precision of the approximate quadrature mainly depends on M, as it dictates the maximum order of polynomials the quadrature can handle with guaranteed accuracy. On the other hand, N, the number of grid points (and weights), should be sufficiently large so that there are enough free variables w_n to minimize the error $L(\mathbf{w})$.

We have chosen to solve the optimization problem above instead of inverting directly ($\mathbf{w} = P^{-1}\mathbf{a}$) for two reasons. First, the matrix P is often singular and therefore challenging to invert. We found that this is especially problematic for grid points that are regularly spaced inside the tetrapyd volume. Second, the direct inversion does not guarantee our requirement that the weights are non-negative, which is crucial for the numerical stability.

Solving the non-negative least squares problem has another advantage; the optimal solution often leaves many of the w_n's exactly equal to zero. In fact, we found that less than 10% of the weights remained nonzero in most cases when using a grid uniformly spaced within the tetrapyd. This allows us to drop the grid points that do not contribute to the quadrature, which in turn effectively decreases N without degrading accuracy.

Figure 2 visualizes two quadrature rules over $\mathcal{V}_{\mathrm{T}}(0.1, 1)$: the simple uniform quadrature (left) and our new tetrapyd quadrature (right). The spheres are located at the points $\mathbf{x}_n$ in (42) where the integrands are evaluated, and their sizes are proportional to weights w_n. Larger spheres are also painted with darker colors for better visualization. As expected, the uniform quadrature has equally sized spheres except near the planes that enforce the triangle inequalities, which cut down the voxel volumes and reduce the weights. Starting with $N = 15$ points on each dimension, the uniform quadrature amounts to 2517 grid points in total, or 519 with symmetry. On the other hand, the tetrapyd quadrature only has 302 points or 72 with symmetry. Note that more central points in the tetrapyd tend to have larger weights. Both quadratures respect the symmetry enjoyed by $\mathcal{V}_{\mathrm{T}}$ (S_3).

D. Numerical validation

We use our public code TETRAQUAD to obtain numerical quadrature rules for $\mathcal{V}_{\mathrm{T}}(0.001, 1)$ and test how accurately they can evaluate the integral of $k_1^p k_2^q k_3^r$. We compare the result with the uniform quadrature, which has a uniformly spaced grid and weights proportional to the voxel volumes, as described in the previous section.[7] The results are shown in Fig. 3.

[7] In practice, we use a Monte Carlo method to compute the volumes of voxels that are sliced by the surfaces of tetrapyd; 100,000 sample points are drawn uniformly from the cube, and we count the fraction of samples that lie inside the tetrapyd.

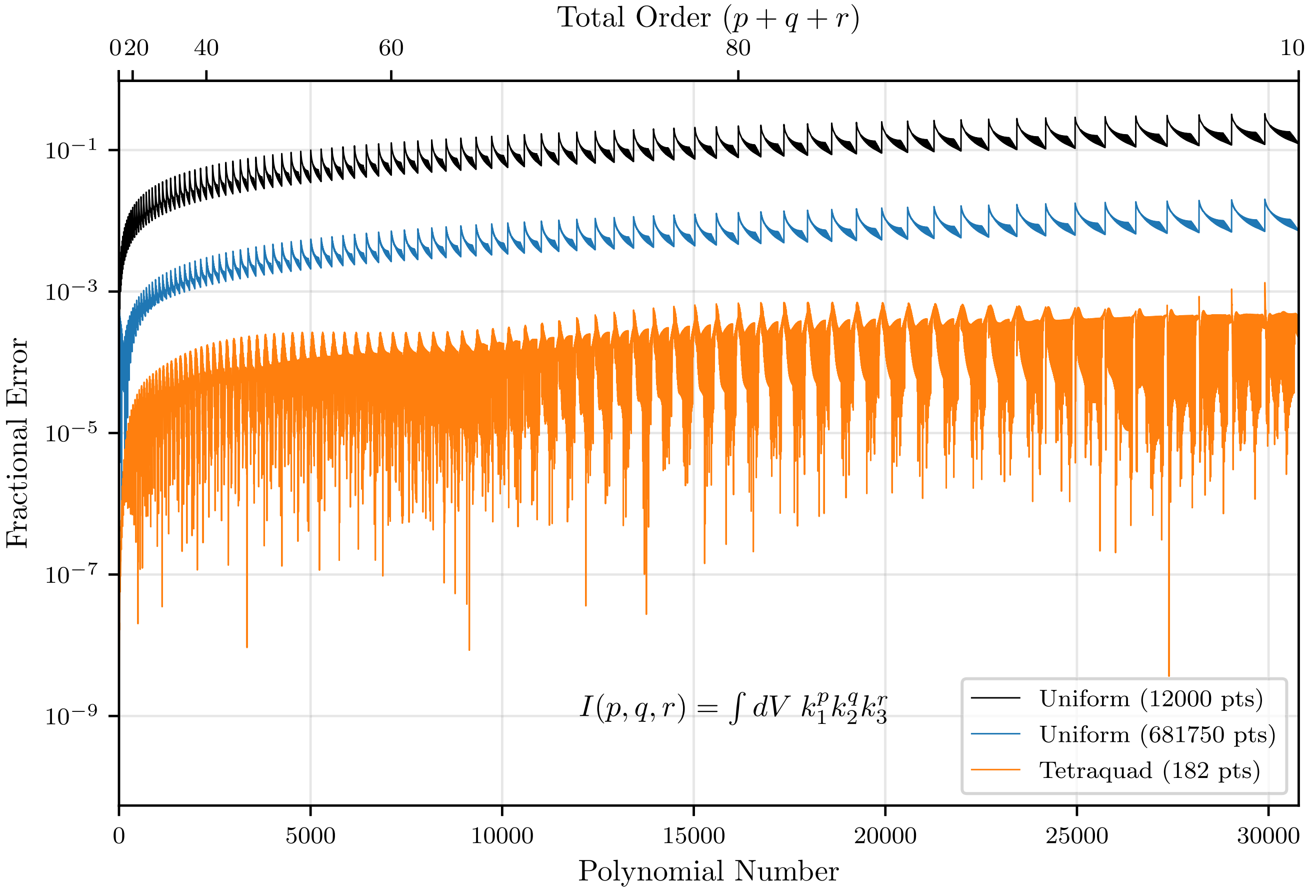


FIG. 3. Numerical accuracy of our quadrature rule on integrating $f(k_1,k_2,k_3) = k_1^p k_2^q k_3^r$ over the tetrapyd $\mathcal{V}_{\mathrm{T}}(0.001,1)$: (a) a polynomial of total order 45 and (b) a sinusoidal function of the overall scale. We compare the result with the simple three-dimensional trapezoidal rule on a uniform grid ("Uniform"). Our quadrature ("Tetraquad") achieves much lower fractional error with significantly fewer grid points and has fractional error $<10^{-3}$ for polynomials with total orders $(p+q+r)$ up to 100.

We find that the TETRAQUAD quadrature with 182 points yields orders of magnitude better precision compared with the simple uniform quadrature with 681,750 points (200 on each axis, restricted to $\mathcal{V}_{\mathrm{T}/6}$). This quadrature is obtained by minimizing the error of polynomials of total orders up to 50, but we see that the error remains small ($<10^{-3}$) for total orders up to 100. Note that the number of polynomials $k_1^p k_2^q k_3^r$ of total order d is equal to the nearest integer to $(d+6)^2/12$,[8] so there are $\sim d^3/36$ linearly independent polynomials with total order less than equal to d. Our 182 points quadrature remains accurate for more than 30,000 independent polynomials.

Next, we investigate how the accuracy of our quadrature scales with the number of grid points. Figure 4 shows how the fractional errors of integrating $k_1^{15}k_2^{15}k_3^{15}$ and $\cos(2\pi(k_1+k_2+k_3))$ depend on the number of grid points. The latter was chosen to test how robust the quadrature is when applied to nonpolynomial functions. An analytic expression for the integrals of $\sin(\omega(k_1+k_2+k_3)+\phi)$ over tetrapyd was used to evaluate the numerical accuracy.

For grid sizes less than 50, Tetraquad shows comparable or sometimes worse performance compared with the uniform quadrature of similar grid size. However, it improves much faster with the grid size and becomes a few orders of magnitude more precise than the uniform quadrature by 100+ grid points for both of the test functions. It is also worth noting that Tetraquad remains accurate for functions that are not polynomials. We have not increased the number of grid points of Tetraquad beyond 150 in the plot due to increased computational cost, but also as we expect the errors to become even smaller. Only ≈ 10% of the initial grid points remain nonzero after the computation of Tetraquad, so the largest Tetraquad quadrature in Fig. 4 started with 1500 points. The memory and computational cost hence scales rapidly with the number of grid points. We found the Tetraquad rule shown in Fig. 3 to be sufficient for our purposes.

[8] This is equal to the number of partitions of $n+3$ into exactly three parts, or equivalently, the number of partitions where the maximum partition size exactly equal to 3. The proof is given in [114].

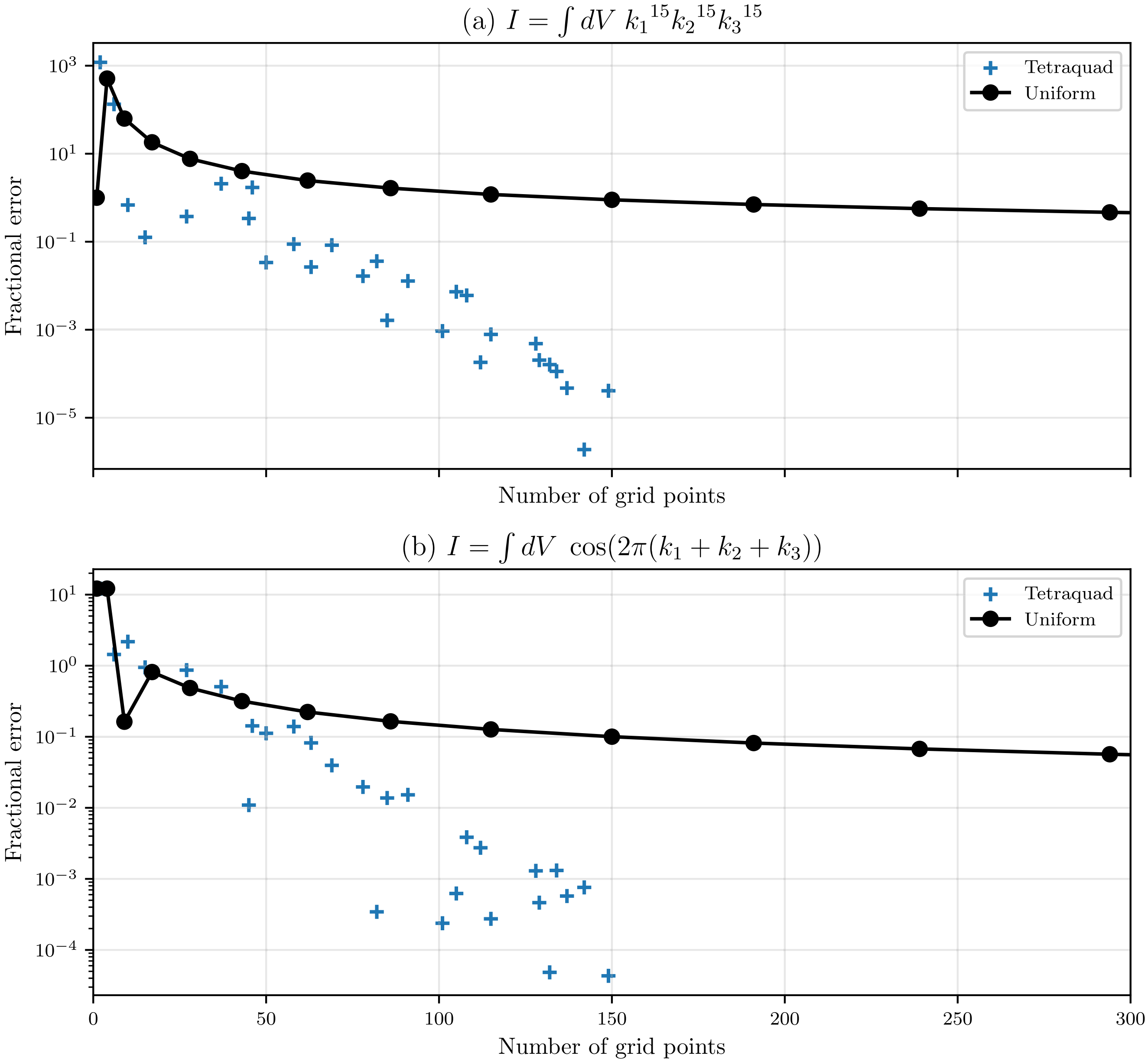


FIG. 4. Numerical accuracy of our quadrature rules with varying grid sizes on integrating two test functions over the tetrapyd $\mathcal{V}_{\mathrm{T}}(0.001, 1)$. We compare the result with the simple three-dimensional trapezoidal rule on a uniform grid ("Uniform") as before. For both test functions, we see that our quadrature rule performs and scales better; its precision is several orders of magnitude better with the same number of grid points. Note that analytical forms of the two integrals were used to compute the errors. Tetraquad rules with a larger number of grid points are not shown due to increased computational cost.

V. RESULTS AND DISCUSSION

We implemented CMB-BEST using C and Python independently from other existing pipelines. Details about the implementation and high-performance computing optimization can be found in Appendix C.

A. Public release of the code

We release the public code for obtaining the Planck 2018 CMB constraints on given bispectrum shapes of interest—effectively the CMB bispectrum likelihood—as a Python package CMBBEST, which can be found at Ref. [85] together with the installation guidelines and an example notebook. The code requires a set of precomputed data for some basis functions. Provided with the code is an HDF5 data file that contains the precomputed results for the Planck 2018 CMB temperature and polarization dataset under the monomials and Legendre basis functions.

1. Description

Speaking in the language of the CMB-BEST formalism, CMBBEST contains the Cython code for the basis expansion given in (19) to compute the expansion coefficients α from a given bispectrum shape. The data file provided contains the precomputed values of $\beta^{\rm cub}$ in (25), $\beta^{\rm lin}$ in (26), and Γ in (28). The $f_{\rm NL}$ estimate then follows from simple matrix multiplications given by (24). The error $\sigma(f_{\rm NL})$ is computed from the sample variance of the $f_{\rm NL}$ estimates for Gaussian simulations. The Fisher errors in (24) are also computed for reference.

Planck 2018 best-fit parameters [12] were used to compute the CMB transfer functions, with the exception of the scalar amplitude $A_{\rm s}$ and tilt $n_{\rm s}$, which can be varied in the code. The Planck 2018 CMB map from SMICA component separation [99], together with the 160 FFP10 end-to-end simulated maps [97,98] with Gaussian initial conditions were used.

2. Outputs

CMBBEST outputs the $f_{\rm NL}$ constraints for the given set of models as $f_{\rm NL}^{(i)} \pm \sigma(f_{\rm NL}^{(i)})$, together with the expansion coefficients $\alpha^{(i)}$ and the Fisher matrix F_{ij}. If the Fisher matrix is invertible, the marginalized constraints from a joint analysis are provided as well. The package provides some utility functions to save the results in various formats including `pandas` data frames and LATEX tables.

3. Conventions

Our conventions follow the Planck PNG conventions closely [10–12]. The "Model" instance of CMBBEST is specified by the shape function S_Φ defined as

$$S_\Phi(k_1, k_2, k_3) \equiv (k_1 k_2 k_3)^2 B_\Phi(k_1, k_2, k_3), \qquad (54)$$

where the primordial potential bispectrum B_Φ satisfies

$$\langle \Phi(\mathbf{k}_1)\Phi(\mathbf{k}_2)\Phi(\mathbf{k}_3)\rangle = (2\pi)^3 \delta^{(3)}(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3) \times B_\Phi(k_1, k_2, k_3). \qquad (55)$$

The Fourier convention with $\int dx$ and $\int dk/(2\pi)$ is used throughout this work. Under this convention, the local shape function is given by [reiterating (35) but with a nonunit $n_{\rm s}$]

$$S_\Phi^{\rm local}(k_1, k_2, k_3) \equiv 6A^2 \frac{1}{3}\left[\frac{k_1^2}{k_2^{2-n_{\rm s}} k_3^{2-n_{\rm s}}} + \frac{k_2^2}{k_3^{2-n_{\rm s}} k_1^{2-n_{\rm s}}} + \frac{k_3^2}{k_1^{2-n_{\rm s}} k_2^{2-n_{\rm s}}}\right], \qquad (56)$$

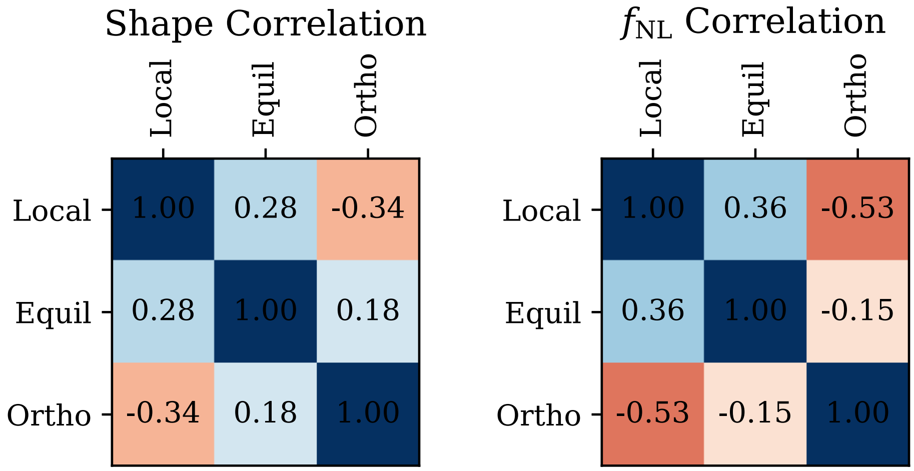
where the power spectrum amplitude A follows the convention in [12]: $P_\Phi(k) = Ak^{n_{\rm s}-4}$. Note that this is different but closely related to the scalar power spectrum amplitude $A_{\rm s}$ that appears in the power spectrum analyses such as [12] via $A = 2\pi^2 (\frac{3}{5})^2 k_*^{1-n_{\rm s}} A_{\rm s}$, where the pivot scale $k_* = 0.05\ {\rm Mpc}^{-1}$. A detailed comparison between these somewhat confusing conventions is given in Appendix D. Lastly, note that the shape functions are often normalized with a factor of $6A^2$ so that, in ζ space, $S_\zeta(k_*, k_*, k_*) = \frac{9}{10}(2\pi)^4 {A_{\rm s}}^2$.

4. Caveats

The CMB bispectrum constraints obtained from CMBBEST are as accurate as the basis expansion. While we have extensively tested that our basis set can accurately handle most classes of models studied in Planck, we advise the users to keep an eye on the convergence statistics included in the resulting constraints. The "convergence correlation" is defined as $r \equiv \langle S, S^{\rm proj}\rangle / \sqrt{\langle S, S\rangle \langle S^{\rm proj}, S^{\rm proj}\rangle}$ and measures the "cosine" between the original and projected (basis-expanded) shape functions. This value should be close to 1. The "convergence epsilon" $\epsilon \equiv \sqrt{2(1-r^2)}$ is a rough estimate of the expected level of the offset in $f_{\rm NL}$ caused by the inaccurate expansion, as discussed in the appendix of [10]. We currently provide two different resolutions for the Legendre basis: $p_{\rm max} = 10$ and $p_{\rm max} = 30$. For most smooth shape functions, the standard $p_{\rm max} = 10$ basis should be sufficient. The higher resolution setting with $p_{\rm max} = 30$ is desired for more oscillatory shape functions. A rule of thumb is that the Legendre basis would struggle to decompose a shape function with more than $p_{\rm max}$ oscillations in the k range of $[2 \times 10^{-4}, 2 \times 10^{-1}]$.

TABLE II. Comparison between $f_{\rm NL}$ constraints for the standard templates by CMB-BEST and Planck using the Planck 2018 SMICA maps. The Planck results are quoted from [12]. For CMB-BEST, the monomials (34) and Legendre basis (37) have been used, while for Planck, the KSW, binned, and modal 2 [24] estimator results are shown. The results are based on independent single-shape analyses with the lensing-ISW bias subtracted and errors corresponding to 68% confidence levels. SMICA foreground-cleaned maps have been used.

			$f_{\rm NL} \pm \sigma(f_{\rm NL})$	
Template	Analysis	Method	T	T + E
Local	CMB-BEST	KSW	-1.7 ± 6.0	-1.1 ± 5.1
		Legendre	-1.4 ± 6.4	-1.1 ± 5.3
	Planck 2018	KSW	-0.5 ± 5.6	-0.9 ± 5.1
		Binned	-0.1 ± 5.6	-2.5 ± 5.0
		Modal	-0.6 ± 6.4	-2.0 ± 5.0
Equilateral	CMB-BEST	KSW	14 ± 66	-22 ± 49
		Legendre	15 ± 66	-22 ± 49
	Planck 2018	KSW	7 ± 66	-18 ± 47
		Binned	26 ± 69	-19 ± 48
		Modal	34 ± 67	-4 ± 43
Orthogonal	CMB-BEST	KSW	-9 ± 40	-31 ± 24
		Legendre	-9 ± 40	-32 ± 24
	Planck 2018	KSW	-15 ± 36	-37 ± 23
		Binned	-11 ± 39	-34 ± 24
		Modal	-26 ± 43	-40 ± 24



Shape Correlation	Local	Equil	Ortho
Local	1.00	0.28	-0.34
Equil	0.28	1.00	0.18
Ortho	-0.34	0.18	1.00

$f_{\rm NL}$ Correlation	Local	Equil	Ortho
Local	1.00	0.36	-0.53
Equil	0.36	1.00	-0.15
Ortho	-0.53	-0.15	1.00

FIG. 5. Correlation matrix for the standard templates: local, equilateral and orthogonal. On the left is the correlation (“cosine”) between templates in the primordial (k_1, k_2, k_3) space, while on the right is the correlation between single $f_{\rm NL}$ estimates obtained from 160 simulated maps. Both temperature and polarization maps were used for the latter. We find that the standard templates are mostly orthogonal to each other in both the primordial and late-time spaces, with the exception of local and orthogonal which are more correlated late time.

B. Planck 2018 constraints

CMB-BEST has been thoroughly tested on internal consistency and validated against Planck’s modal estimator. In this section, we reproduce some results from the Planck 2018 analysis [12] on the standard templates: local, equilateral and orthogonal. We note that the analyses shown in this part are mostly not new and are based heavily on [10–12]. The following part (Sec. V C) shows some new constraints on highly oscillatory templates not studied by Planck.

A comparison between the constraints from CMB-BEST and those of Planck is shown in Table II. Both the monomials basis set (34), which is equivalent to Planck’s KSW estimator, and the Legendre basis (37) have been used for consistency checks. All results shown assume the ΛCDM baseline cosmology with the background parameters fixed to the best-fit values for Planck 2018 power spectrum TTTEEE + lowE + lensing [115], although it has been shown that the bispectrum constraints are not very sensitive to these choices [11]. Following Planck, we use the multipole range of $2 \le \ell \le 2500$ for temperature and $4 \le \ell \le 2000$ for E-mode polarization.

We find that the constraints from CMB-BEST are consistent with Planck and the differences are within the scatter between Planck’s own estimators, which is explained in the appendix of [10]. The results between the “KSW” and “Legendre” basis sets of CMB-BEST are entirely consistent as well. The small differences in the estimated $f_{\rm NL}$ and error bars, mainly noticeable in the local template, come from the fact that the Legendre basis has a truncated k range, which slightly reduces the power at very low ℓ’s. This indeed affects the local constraints the most as it reduces the power of squeezed configurations (e.g. $b_{2\ell\ell}$) and hence increases the error bars by up to 6%. We thoroughly investigated both internal and external consistency by a map-by-map validation of CMB-BEST bases against the modal estimator using 160 Gaussian simulations, which is detailed in Appendix F, Sec. XII A. We found the results from the two independent methods to be consistent on these simulated maps. Note also that the mean squared error (MSE) in the basis expansion for each of these models is kept less than 10^{-8} in all cases.

The constraints can be understood as hypothesis tests for the individual templates with $\hat{f}_{\rm NL}$ as a test statistic: testing the alternative hypothesis H_1: $f_{\rm NL} \neq 0$ against the null hypothesis H_0: $f_{\rm NL} = 0$. The $f_{\rm NL}$ estimates of the 160 simulated maps follow a Gaussian distribution with zero mean under the null hypothesis. If the p value of the observation $\hat{f}_{\rm NL}$ with respect to this distribution is sufficiently low (e.g. <0.01),[9] we would reject the null hypothesis. As can be seen from Table II, we do not have sufficient evidence to reject the null hypotheses of $f_{\rm NL} = 0$ for the three standard templates.

The constraints above are from independent single-shape analyses; for each template and the corresponding $B^{\rm temp}_{\ell_1\ell_2\ell_3}$, we consider the model $B^{\rm th}_{\ell_1\ell_2\ell_3} = f_{\rm NL} B^{\rm temp}_{\ell_1\ell_2\ell_3}$. If the model is correct and the template is an accurate representation of the primordial bispectrum, then the result is an estimate of the amplitude of PNG in the given shape. Various inflationary models that predict different amounts of $f_{\rm NL}$ can be constrained by this estimate.

Alternatively, we could do a joint analysis where our theoretical bispectrum is given by a linear combination of either all or some of the templates under consideration. Here, for demonstration purposes, we assume that a model

[9] In principle, the p-value threshold should be set before seeing the data.

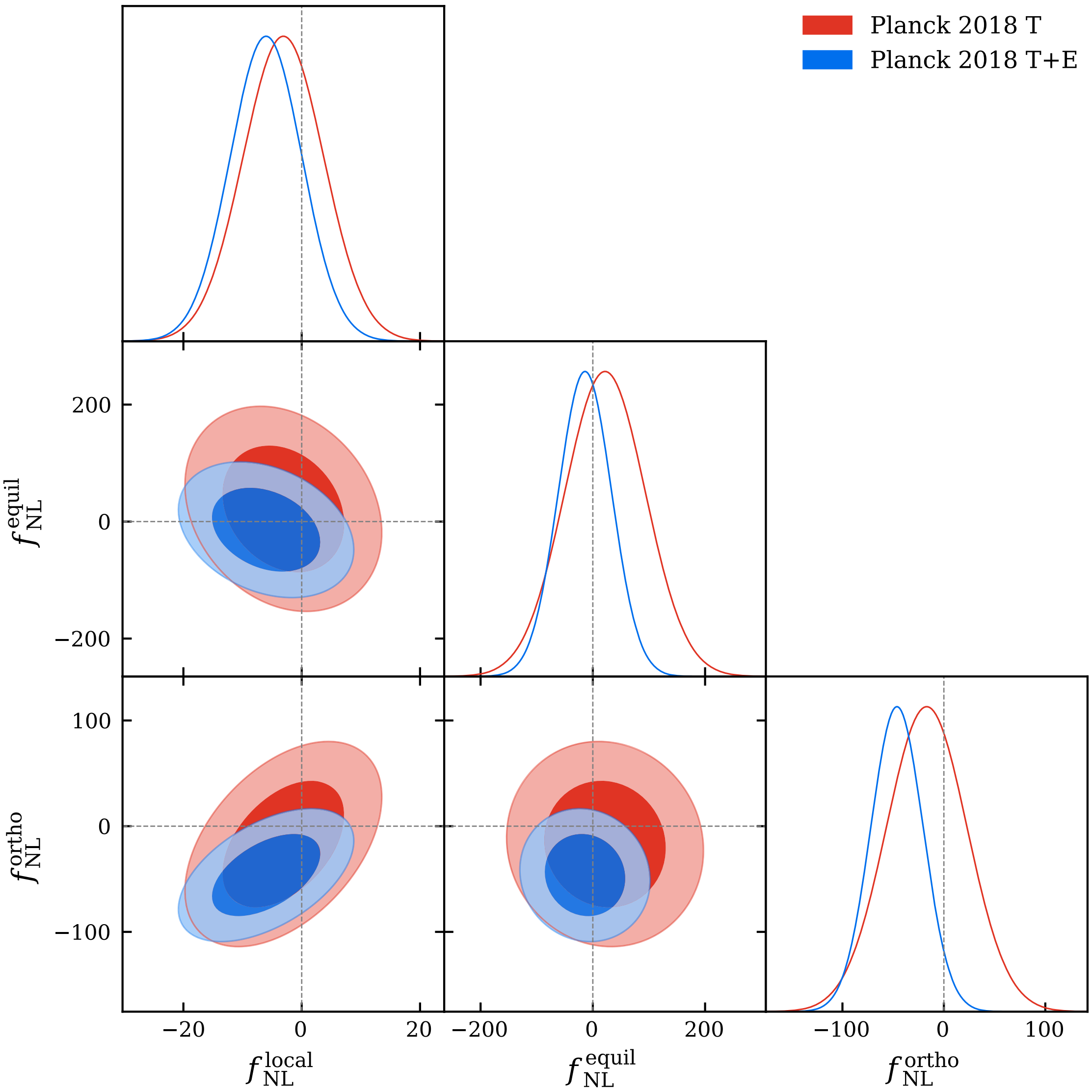


FIG. 6. The CMB bispectrum likelihood for a joint analysis on the local, equilateral and orthogonal templates using CMB-BEST. The contours and line plots show marginalized likelihoods obtained from the temperature-only (T) and temperature + polarization (T + E) data. Note that the Fisher matrix was used in the likelihood, which is an accurate description with only a slight underestimation of the errors (see Appendix F, Sec. XII B, for details). The lensing-ISW bias has been taken into account. The dashed lines correspond to $f_{\rm NL} = 0$ in each plot. CMB-BEST's Legendre basis set was used for this analysis.

predicts a bispectrum shape that is represented as a linear combination of all three standard templates: $B^{\rm th}_{\ell_1\ell_2\ell_3} = f_{\rm NL}^{\rm local} B^{\rm local}_{\ell_1\ell_2\ell_3} + f_{\rm NL}^{\rm equil} B^{\rm equil}_{\ell_1\ell_2\ell_3} + f_{\rm NL}^{\rm ortho} B^{\rm ortho}_{\ell_1\ell_2\ell_3}$. The $f_{\rm NL}$ estimate is given by (14) in this case. For templates that are orthogonal in shape [with respect to the inner product (33)], the Fisher matrix tends to be diagonal, and the constraints are nearly identical to those of single-shape analyses. Figure 5 shows correlations between the three standard templates, both in the primordial space with the inner product (33) and in the $f_{\rm NL}$ estimates. The templates are weakly correlated as expected, with some noticeable correlations between the local and orthogonal templates.

A joint analysis was performed to obtain the likelihood $\mathcal{L}(\mathbf{B}^{\rm th}|\hat{\mathbf{B}}) = \mathcal{L}(f_{\rm NL}^{\rm local}, f_{\rm NL}^{\rm equil}, f_{\rm NL}^{\rm ortho}|\hat{\mathbf{B}})$ as defined in (5). The marginalized likelihood of the three $f_{\rm NL}$ parameters is shown in Fig. 6, plotted using the Python library GetDist [116]. Overall, there is no evidence for a nonzero $f_{\rm NL}$.

The likelihood in Fig. 6 is plotted under the assumption that $\{\hat{B}_{\ell_1\ell_2\ell_3}\}$ is multivariate normal distributed with diagonal covariances equal to $(6/\Delta_{\ell_1\ell_2\ell_3})C_{\ell_1}C_{\ell_2}C_{\ell_3}$, as we discussed when writing down the likelihood in (5). The validity of this assumption for this analysis is tested in Appendix F, Sec. XII B, using the simulated maps. The Fisher matrix slightly underestimates the error (up to 10% in orthogonal), but the estimator remains nearly optimal in all cases.

Assuming that the likelihood is an accurate representation of the CMB bispectrum statistics, there can be multiple interpretations of the results depending on one's statistical alignments. A frequentist would interpret the contours as the distribution of $f_{\rm NL}$ estimates we could have obtained under different realizations of the Universe and independent observations. They would conclude that there is no significant evidence for rejecting the null hypothesis of $\hat{\mathbf{f}}_{\rm NL} = \mathbf{0}$. When testing an early Universe model that predicts $\mathbf{f}_{\rm NL} = \mathbf{x}$ for some $\mathbf{x}$, the likelihood can similarly be used to construct a p-value test for a rejection/nonrejection of the model.

On the other hand, a Bayesian-minded person may interpret the contours as the posterior distributions for $\mathbf{f}_{\rm NL}$ when the prior is flat and wide. The posterior distribution represents the probability distribution of each $f_{\rm NL}$ *given* the model and the data. Alternatively, the constraints from the large-scale structure surveys (LSS) can be used as priors. As the current constraints from LSS are weaker in comparison, the posteriors would appear similar to the plots in Fig. 6. The Bayesian model comparison methods such as Akaike information criteria or Bayes factors could be used to compare different models from various templates against the model without PNG. Note that unfortunately the constant term appearing in (9) is not computed directly in CMB-BEST, KSW or modal estimator, so the likelihood is determined only up to a constant multiplicative factor.

C. Targeted oscillatory basis

Having validated the pipeline internally and against Planck, we present some new constraints on $f_{\rm NL}$ for templates that have not been studied before in Planck analysis [12]. Motivated by various inflationary models which predict linearly spaced oscillations in the bispectrum, we consider the bispectrum shapes of form

$$(k_1k_2k_3)^2 B(k_1,k_2,k_3) = f(k_1,k_2,k_3) \times \sin(\omega(k_1+k_2+k_3)+\phi), \tag{57}$$

where $f(k_1,k_2,k_3)$ is some envelope function of choice and ω and ϕ parametrize the overall linearly spaced oscillations.

The case where $f(k_1,k_2,k_3) = 1$ was thoroughly studied using a KSW-like estimator [12,72] up to oscillation frequencies $0 \le \omega \le 3000$ in units of Mpc. This was viable due to the separability of linear oscillations. However, such a trick does not immediately apply to a more complex type of envelope function f, and it is challenging to go to a highly oscillatory regime; the polynomial basis of the modal estimator often lacks resolution, and the oscillations are washed out by binning in the binned estimator. Planck's modal studied general oscillatory templates up to $\omega \lesssim 350$ Mpc.

Using the flexibility of mode function choices in CMB-BEST, we can construct a targeted basis with a fixed value of ω as shown in (39). This allows us to study general shape functions $f(k_1,k_2,k_3)$ with a fixed oscillation frequency ω. As a part of the proof of concept, we performed a targeted analysis on $\omega = 1000$ Mpc. This frequency was chosen somewhat arbitrarily for demonstration purposes, but in future analyses, noticeable oscillatory signals found in the power spectrum or bispectrum could guide this choice. The results are summarized in Table III. Overall, we find no significant evidence from the CMB bispectrum for nonzero PNG in these templates.

TABLE III. Constraints on some highly oscillatory bispectra from CMB-BEST's targeted basis. Templates are products of the standard templates with linear oscillations $S(k_1,k_2,k_3) = f(k_1,k_2,k_3)\sin(\omega(k_1+k_2+k_3)+\phi)$ for $\omega = 1000$ Mpc, where f is the local, equilateral, or orthogonal template. The constant and exponential envelopes are defined in the text. Overall, we do not find evidence for PNG with oscillations with $\omega = 1000$ Mpc.

Envelope shape	Phase	$f_{\rm NL} \pm \sigma(f_{\rm NL})$	$f_{\rm NL}/\sigma(f_{\rm NL})$
Local	sin	-0.39 ± 0.98	-0.40
	cos	-1.4 ± 1.1	-1.3
Equilateral	sin	0.29 ± 0.32	0.91
	cos	0.48 ± 0.46	1.1
Orthogonal	sin	0.11 ± 0.12	0.91
	cos	0.16 ± 0.17	0.98
Constant	sin	-17 ± 14	-1.2
	cos	12 ± 15	0.76
Exponential	sin	-14 ± 27	-0.51
	cos	-16 ± 24	-0.66

The templates are normalized so that the envelope part $f(k_1,k_2,k_3)$ follows the conventions from Planck [10]. The "constant" envelope is equivalent to the constant feature models studied in Planck [12]. For the "exponential" envelope, we set $f(k_1,k_2,k_3) \propto \exp[-((1/3)\times(k_1+k_2+k_3)-k_*)^2/2d^2]$, where $k_* = 0.05\ \mathrm{Mpc}^{-1}$ and $d = 0.02\ \mathrm{Mpc}^{-1}$ were chosen as an example. All constraints shown are independent single-shape analyses.

In order to test the ability of the CMB-BEST to differentiate these oscillatory templates, we plot the correlation matrix between the $f_{\rm NL}$ estimates obtained from 160 Gaussian simulations. The results are shown in Fig. 7.

We find that the sine and cosine templates with out-of-phase oscillations always remain relatively uncorrelated, while within the same phases of oscillations, the local, equilateral and orthogonal templates all show strong correlations (≥ 0.88) in the $f_{\rm NL}$ values they retrieve from the simulations. The in-phase templates are as uncorrelated in the primordial space as the correlation between the envelopes shown in Fig. 5 since our basis expansion

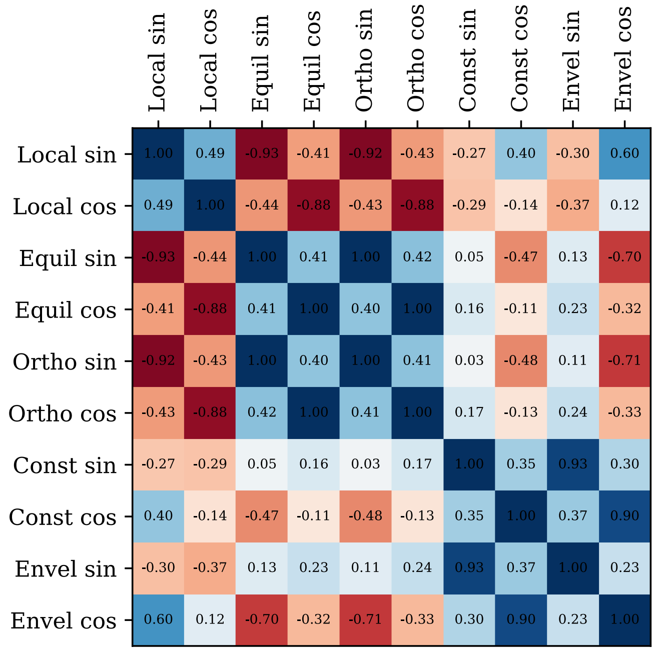


FIG. 7. Correlation matrix for highly oscillatory bispectrum templates obtained using CMB-BEST's targeted basis. The models considered are the same as in Table III and are described in the text.

for the envelope is accurate (MSE less than 10^{-8}). The differences, however, appear to get washed out by the projection effects; the templates predict nearly identical CMB bispectrum despite having different envelopes.

The constant and exponential envelopes are found to be mostly uncorrelated with the templates with local, equilateral and orthogonal envelopes. The two of them are highly correlated themselves, which suggests that most of the bispectrum signatures from oscillations come from the enveloped region around $K = 0.15\ \mathrm{Mpc}^{-1}$.

A joint analysis between these enveloped shapes can cover models which predict different oscillation phases at different limits. For example, considering the local and equilateral envelopes only, the best-fit bispectrum for the 4-parameter model is given by

$$\begin{aligned} B^{\mathrm{bf}}(k_1,k_2,k_3) &= 5.46B^{\mathrm{Local}}(k_1,k_2,k_3)\sin(\omega K + 38°) \\ &\quad + 1.46B^{\mathrm{Equil}}(k_1,k_2,k_3)\sin(\omega K + 30°), \end{aligned} \tag{58}$$

where $K \equiv k_1 + k_2 + k_3$ and $\omega = 1000$ Mpc. Note that oscillations have similar but different phases in the equilateral and squeezed limits. A single-shape analysis on this template gives the constraint 1.00 ± 0.49 with a 2.1σ-level significance, which is not very high especially given that we had three free parameters to fit for. We conclude that there is no significant evidence for these highly oscillatory ($\omega = 1000$ Mpc) templates with the Planck data.

VI. CONCLUSION

In this work, we presented the formalism and implementation of our new high-resolution bispectrum estimator CMB-BEST. We publicly release the frontend of our code as CMBBEST [85], together with a data file containing precomputed results from the computation-heavy parts of the pipeline, so that researchers can obtain Planck CMB bispectrum constraints for their theoretical bispectrum predictions using our code.

CMB-BEST is formulated and coded for general choices of mode functions and can be used for various purposes. When using a small set of monomials as mode functions, the code is equivalent to the KSW estimator and gives quick constraints on the standard bispectrum templates. Using a larger set of Legendre polynomials allows for an extensive analysis covering a broad range of models. Lastly, a targeted oscillatory basis enables an in-depth, high-resolution analysis of models with a specific oscillation frequency.

We have thoroughly optimized the code, tested the consistency of the results, and showed a proof-of-concept example of a high-resolution bispectrum analysis. In doing so, we came up with some methodological advances in algorithmic optimization and numerical computation. In particular, TETRAQUAD is a code for computing numerical quadrature rules for integrating functions over a tetrapyd domain. The quadrature rule is efficient, accurate and completely general, and is expected to benefit various bispectrum analyses in CMB and LSS.

We plan on studying various inflationary models using this code, and extend the methodology and code to existing and future CMB data, including the upcoming Simons observatory.

Code and data availability. Our codes CMBBEST and TETRAQUAD are publicly available and can be found in [85] and [86], respectively.

ACKNOWLEDGMENTS

W. S. would like to thank Philip Clarke for all his support throughout the development and validation process of CMB-BEST, Arman Shafieloo, David Parkinson and Rodrigo Calderon for many valuable discussions, Kushal Lodha, Dong-Gang Wang and Carlos Duaso Pueyo for their practical feedback on CMBBEST in its early stages, and Kacper Kornet for the high-performance computing support. W. S. was supported by the project Understanding Dark Universe Using Large Scale Structure of the Universe funded by the Ministry of Science. W. S. was also supported by Internal Graduate Studentship from Trinity College, Cambridge during his PhD when a part of this work was undertaken.

APPENDIX A: SPHERICAL HARMONICS RELATED IDENTITIES

In this section, we quote useful identities related to spherical harmonics that appear in the bispectrum formalism. The equations are quoted from various sources including [3,103,110].

The Gaunt integral is defined as an integral of spherical harmonics $Y_{\ell m}$ over the sphere S^2:

$$\mathcal{G}^{\ell_1\ell_2\ell_3}_{m_1m_2m_3} \equiv \int d^2\hat{\mathbf{n}} Y_{\ell_1 m_1}(\hat{\mathbf{n}}) Y_{\ell_2 m_2}(\hat{\mathbf{n}}) Y_{\ell_3 m_3}(\hat{\mathbf{n}}). \tag{A1}$$

It can be rewritten in terms of the Wigner 3j symbols as follows:

$$\mathcal{G}^{\ell_1\ell_2\ell_3}_{m_1m_2m_3} = \sqrt{\frac{(2\ell_1+1)(2\ell_2+1)(2\ell_3+1)}{4\pi}} \times \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix} \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ 0 & 0 & 0 \end{pmatrix} \tag{A2}$$

$$= \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix} h_{\ell_1\ell_2\ell_3}. \tag{A3}$$

The Wigner 3j symbols are normalized so that

$$\sum_{m_j} \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix}^2 = 1, \tag{A4}$$

and therefore

$$\sum_{m_j} \begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3 \end{pmatrix} \mathcal{G}^{\ell_1\ell_2\ell_3}_{m_1m_2m_3} = h_{\ell_1\ell_2\ell_3}, \tag{A5}$$

$$\sum_{m_j} (\mathcal{G}^{\ell_1\ell_2\ell_3}_{m_1m_2m_3})^2 = h^2_{\ell_1\ell_2\ell_3}, \tag{A6}$$

as long as (ℓ_1, ℓ_2, ℓ_3) form a triangle.

When $m_1 = m_2 = m_3 = 0$, a useful identity relates Wigner 3j symbols with Legendre polynomials:

$$\begin{pmatrix} \ell_1 & \ell_2 & \ell_3 \\ 0 & 0 & 0 \end{pmatrix}^2 = \frac{1}{2}\int_{-1}^{1} d\mu P_{\ell_1}(\mu) P_{\ell_2}(\mu) P_{\ell_3}(\mu). \tag{A7}$$

It immediately follows that

$$h^2_{\ell_1\ell_2\ell_3} = \frac{2\ell_1+1)(2\ell_2+1)(2\ell_3+1)}{8\pi} \times \int_{-1}^{1} d\mu P_{\ell_1}(\mu) P_{\ell_2}(\mu) P_{\ell_3}(\mu). \tag{A8}$$

APPENDIX B: ANALYTIC FORMULA FOR INTEGRALS OF POLYNOMIALS OVER A GENERAL TETRAPYD

In this section, we write down the analytic formula for the integral of polynomials over a general tetrapyd defined as

$$\mathcal{V}_\mathrm{T}(\alpha, 1) \equiv \{(k_1, k_2, k_3) : 2\max\{k_1, k_2, k_3\} \le k_1 + k_2 + k_3 \text{ and } \alpha \le k_1, k_2, k_3 \le 1\}, \tag{B1}$$

for $0 \le \alpha \le 1/2$. Note that the integral over $\mathcal{V}_\mathrm{T}(k_\mathrm{min}, k_\mathrm{max})$ can be derived from that of $\mathcal{V}_\mathrm{T}(k_\mathrm{min}/k_\mathrm{max}, 1)$. We have derived the formula given here by hand with occasional aid from Mathematica [117].

Instead of the tetrapyd with a rather complicated shape, we consider a cube $[\alpha, 1]^3$ and the volumes inside this cube excluded by the tetrapyd

$$V_\mathrm{cube} = \mathcal{V}_\mathrm{T}(\alpha, 1) \cup V_1 \cup V_2 \cup V_3, \tag{B2}$$

$$V_1 \equiv \{(k_1, k_2, k_3) : k_2 + k_3 \le k_1 \text{ and } \alpha \le k_2, k_3 \le k_1 \le 1\}, \tag{B3}$$

$$V_2 \equiv \{(k_1, k_2, k_3) : k_3 + k_1 \le k_2 \text{ and } \alpha \le k_3, k_1 \le k_2 \le 1\}, \tag{B4}$$

$$V_3 \equiv \{(k_1, k_2, k_3) : k_1 + k_2 \le k_3 \text{ and } \alpha \le k_3, k_1 \le k_3 \le 1\}. \tag{B5}$$

Since each set on the right-hand side of (B2) is disjoint, we have

$$\int_{\mathcal{V}_\mathrm{T}(\alpha,1)} dV k_1^p k_2^q k_3^r = I_\mathrm{cube} - I_1 - I_2 - I_3. \tag{B6}$$

The integral over the cube is straightforward due to separability:

$$I_\mathrm{cube} = \int_{[\alpha,1]^3} dV k_1^p k_2^q k_3^r = \frac{(1-\alpha^{p+1})(1-\alpha^{q+1})(1-\alpha^{r+1})}{(p+1)(q+1)(r+1)}. \tag{B7}$$

On the other hand,

$$I_1 = \int_{V_1} dk_1 dk_2 dk_3 k_1^p k_2^q k_3^r \tag{B8}$$

$$= \int_\alpha^{1-\alpha} dk_2 \int_\alpha^{1-k_2} dk_3 \int_{k_2+k_3}^{1} dk_1 k_1^p k_2^q k_3^r \tag{B9}$$

$$= \frac{1}{p+1}\int_{\alpha}^{1-\alpha} dk_2 \int_{\alpha}^{1-k_2} dk_3 k_2^q k_3^r [1-(k_2+k_3)^{p+1}] \tag{B10}$$

$$\vdots$$

$$= \frac{1}{p+1}\int_{2\alpha}^{1} ds s^{q+r+1}(1-s^{p+1})\mathrm{Beta}\left(\frac{\alpha}{s}, 1-\frac{\alpha}{s}; q+1, r+1\right) \tag{B11}$$

$$\vdots$$

$$\begin{aligned} &= \frac{\mathrm{Beta}(\alpha, 1-\alpha; q+1, r+1)}{(p+q+r+3)(q+r+2)} \\ &\quad - \frac{1}{(p+1)(q+1)(r+1)}\left[\frac{(r+1)\alpha^{r+1}(1-\alpha)^{q+1}}{q+r+2} + \frac{(q+1)\alpha^{q+1}(1-\alpha)^{r+1}}{q+r+2} - \alpha^{q+r+2}\right] \\ &\quad + \frac{\alpha^{p+q+r+3}}{(p+1)(p+q+r+3)}\left[(-1)^q \mathrm{Beta}\left(2, \frac{1}{\alpha}; p+2, q+1\right) + (-1)^r \mathrm{Beta}\left(2, \frac{1}{\alpha}; p+2, r+1\right)\right]. \end{aligned} \tag{B12}$$

Above, we used a change of variables $(k_2, k_3) = (st, s(1-t))$ and integration by parts with $du = s^{q+r+1}(1-s^{p+1})ds$ and $v = \mathrm{Beta}(\frac{\alpha}{s}, 1-\frac{\alpha}{s}; q+1, r+1)$. The incomplete beta function is defined as

$$\mathrm{Beta}(x, y; p, q) \equiv \int_x^y dt\, t^{p-1}(1-t)^{q-1}, \tag{B13}$$

so that $\mathrm{Beta}(0, 1; p, q)$ corresponds to the usual (complete) beta function. The integrals I_2 and I_3 are obtained similarly by cycling (p, q, r) around. Although the definition originally restricts x and y to [0, 1], the function can be analytically continued to the region outside. We compute these by using the ordinary hypergeometric function ${}_2F_1(a, b; c; z)$ and its relation to the incomplete beta function [118]:

$$\mathrm{Beta}(x, y; p, q) = \frac{y^p(1-y)^q}{p}{}_2F_1(p+q, 1; p+1; y) - \frac{x^p(1-x)^q}{p}{}_2F_1(p+q, 1; p+1; x). \tag{B14}$$

In our public code TETRAQUAD, the hypergeometric function is computed using `mpmath` [119], a Python library for arbitrary-precision arithmetic. Note that the formula above holds from general p, q, r that are not necessarily integers.

APPENDIX C: BISPECTRUM ESTIMATOR INCLUDING POLARIZATION

In this section, we generalize the CMB bispectrum likelihood (5) to include both CMB temperature and E-mode polarization.

For simplicity, we assume that the covariance matrix is diagonal:

$$\left\langle a^{X_1}_{\ell_1 m_1} a^{X_2}_{\ell_2 m_2}\right\rangle = C^{X_1 X_2}_{\ell_1}\delta_{\ell_1\ell_2}\delta_{m_1 -m_2}, \tag{C1}$$

where $X = \mathrm{T}, \mathrm{E}$ denotes temperature and E-mode polarization, respectively.

The CMB bispectrum now consists of multiple parts: TTT, TTE, TEE and EEE. The bispectrum estimate (2) can be generalized as

$$\begin{aligned}\hat{B}^{X_1X_2X_3}_{\ell_1\ell_2\ell_3} \equiv \sum_{m_j}\begin{pmatrix}\ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3\end{pmatrix}[a^{X_1}_{\ell_1 m_1}a^{X_2}_{\ell_2 m_2}a^{X_3}_{\ell_3 m_3} \\ - \left[\langle a^{X_1}_{\ell_1 m_1}a^{X_2}_{\ell_2 m_2}\rangle a^{X_3}_{\ell_3 m_3} + (2\,\mathrm{cyc})\right]\right].\end{aligned} \tag{C2}$$

$$= \sum_{m_j}\begin{pmatrix}\ell_1 & \ell_2 & \ell_3 \\ m_1 & m_2 & m_3\end{pmatrix}(B^{X_1X_2X_3})^{\ell_1\ell_2\ell_3}_{m_1m_2m_3}. \tag{C3}$$

As discussed in [24,120], it is convenient to work with linear combinations of $X = \mathrm{T}, \mathrm{E}$ in which the covariance matrix $C^{X_1X_2}_\ell$ at each multipole ℓ is diagonal. Considering the Cholesky decomposition of $C^{-1}_\ell = L^{\mathrm{T}}_\ell L_\ell$, where

$$L_\ell \equiv \begin{pmatrix} \frac{1}{\sqrt{C^{\mathrm{TT}}_\ell}} & 0 \\ \frac{-C^{\mathrm{TE}}_\ell}{\sqrt{C^{\mathrm{TT}}_\ell}\sqrt{C^{\mathrm{TT}}_\ell C^{\mathrm{EE}}_\ell - (C^{\mathrm{TE}}_\ell)^2}} & \frac{C^{\mathrm{TT}}_\ell}{\sqrt{C^{\mathrm{TT}}_\ell}\sqrt{C^{\mathrm{TT}}_\ell C^{\mathrm{EE}}_\ell - (C^{\mathrm{TE}}_\ell)^2}} \end{pmatrix}. \tag{C4}$$

The C_ℓ's here include the corrections from the beam and noise. We now work with $\tilde{X} = \tilde{\mathrm{T}}, \tilde{\mathrm{E}}$ obtained by transforming $X = \mathrm{T}, \mathrm{E}$ through L. The CMB transfer function $T_\ell(k)$ and observed $a_{\ell m}$'s transform as

$$T_\ell^{\tilde{X}}(k) = \sum_X L_\ell^{\tilde{X}X} T_\ell^X, \quad \text{and} \quad a_{\ell m}^{\tilde{X}} = \sum_X L_\ell^{\tilde{X}X} a_{\ell m}^X. \tag{C5}$$

It follows that $\left\langle a_{\ell_1 m_1}^{\tilde{X}_1} a_{\ell_2 m_2}^{\tilde{X}_2} \right\rangle = \delta_{\ell_1 \ell_2} \delta_{m_1 - m_2}$.

Analogously to the temperature-only analysis, we assume that $\{\hat{B}_{\ell_1\ell_2\ell_3}\}_{\ell_1 \leq \ell_2 \leq \ell_3}$ have a diagonal covariance matrix with diagonal entries $(6/\Delta_{\ell_1\ell_2\ell_3}) C_{\ell_1} C_{\ell_2} C_{\ell_3}$, but now with 2×2 matrices $C_\ell^{X_1 X_2}$. Under this assumption, the CMB bispectrum likelihood can be written as

$$\mathcal{L}(\mathbf{B}^{\mathrm{th}} | \hat{\mathbf{B}}) \propto \exp\left[-\frac{1}{2} \sum_{\ell_1 \leq \ell_2 \leq \ell_3} \sum_{X_j, X_j'} \left[\frac{\Delta_{\ell_1\ell_2\ell_3}}{6} \left(\hat{B}_{\ell_1\ell_2\ell_3}^{X_1 X_2 X_3} - B_{\ell_1\ell_2\ell_3}^{\mathrm{th}, X_1 X_2 X_3} \right) \right.\right.$$
$$\left.\left. (C_{\ell_1}^{-1})^{X_1 X_1'} (C_{\ell_2}^{-1})^{X_2 X_2'} (C_{\ell_3}^{-1})^{X_3 X_3'} \left(\hat{B}_{\ell_1\ell_2\ell_3}^{X_1' X_2' X_3'} - B_{\ell_1\ell_2\ell_3}^{\mathrm{th}, X_1' X_2' X_3'} \right) \right]\right]. \tag{C6}$$

After the transformation from X to $\tilde{X}$, the likelihood (C6) simplifies to

$$\mathcal{L}(\mathbf{B}^{\mathrm{th}} | \hat{\mathbf{B}}) \propto \exp\left[-\frac{1}{2} \sum_{\ell_1 \leq \ell_2 \leq \ell_3} \sum_{\tilde{X}_j} \frac{\Delta_{\ell_1\ell_2\ell_3}}{6} \left(\hat{B}_{\ell_1\ell_2\ell_3}^{\tilde{X}_1 \tilde{X}_2 \tilde{X}_3} - B_{\ell_1\ell_2\ell_3}^{\mathrm{th}, \tilde{X}_1 \tilde{X}_2 \tilde{X}_3} \right)^2 \right]. \tag{C7}$$

Given a set of bispectrum templates and their amplitude parametrized by f_{NL},

$$B_{\ell_1\ell_2\ell_3}^{\mathrm{th}, X_1 X_2 X_3} = h_{\ell_1\ell_2\ell_3} \sum_i f_{\mathrm{NL}}^{(i)} b_{\ell_1\ell_2\ell_3}^{(i), X_1 X_2 X_3}. \tag{C8}$$

The maximum likelihood estimator for f_{NL} is then given by

$$\widehat{f_{\mathrm{NL}}}^{(i)} = \sum_j (F^{-1})_{ij} S_j, \tag{C9}$$

where

$$S_i \equiv \frac{1}{6} \sum_{\ell_j, m_j, \tilde{X}_j} \mathcal{G}_{m_1 m_2 m_3}^{\ell_1 \ell_2 \ell_3} b_{\ell_1\ell_2\ell_3}^{(i), \tilde{X}_1 \tilde{X}_2 \tilde{X}_3} \left[a_{\ell_1 m_1}^{\tilde{X}_1} a_{\ell_2 m_2}^{\tilde{X}_2} a_{\ell_3 m_3}^{\tilde{X}_3} - \left[\langle a_{\ell_1 m_1}^{\tilde{X}_1} a_{\ell_2 m_2}^{\tilde{X}_2} \rangle a_{\ell_3 m_3}^{\tilde{X}_3} + (2\ \mathrm{cyc}) \right] \right], \tag{C10}$$

$$F_{ij} \equiv \frac{1}{6} \sum_{\ell_j, \tilde{X}_j} h_{\ell_1\ell_2\ell_3}^2 b_{\ell_1\ell_2\ell_3}^{(i), \tilde{X}_1 \tilde{X}_2 \tilde{X}_3} b_{\ell_1\ell_2\ell_3}^{(j), \tilde{X}_1 \tilde{X}_2 \tilde{X}_3}. \tag{C11}$$

The CMB-BEST formalism can effectively exploit the separability of the summation over $\tilde{X}$. The projected modes defined in (21) are extended to polarization:

$$\tilde{q}_p^{\tilde{X}}(\ell, r) \equiv \frac{2r^{\frac{2}{3}}}{\pi} \int dk q_p(k) T_\ell^{\tilde{X}}(k) j_\ell(kr). \tag{C12}$$

The filtered maps (22) for the T + E analyses have a simple summation over $\tilde{X}$ as follows:

$$M_p^{(i)}(\hat{\mathbf{n}}, r) \equiv \sum_{\ell, m, \tilde{X}} \tilde{q}_p^{\tilde{X}}(l, r) a_{\ell m}^{\tilde{X}} Y_{\ell m}(\hat{\mathbf{n}}). \tag{C13}$$

The rest of the CMB-BEST formalism is identical to those described in the main text. Therefore, adding polarization does not significantly affect the computational complexity; it only doubles the number of spherical harmonic transforms (SHT) needed to obtain the filtered maps.

APPENDIX D: PRIMORDIAL POWER SPECTRUM AMPLITUDE CONVENTIONS

Here, we clarify and relate the different notations used in (a) Planck 2018 constraints on PNG [12], (b) Planck 2018 cosmological parameters [115], and (c) Xingang Chen's review on PNG [2] for the primordial power spectrum amplitudes.

In [12], the primordial power spectrum $P_\Phi(k)$ is defined as

$$P_\Phi(k) = Ak^{n_s-4}, \tag{D1}$$

so that

$$\langle \Phi(\mathbf{k}_1)\Phi(\mathbf{k}_2)\rangle = P_\Phi(k_1)(2\pi)^3\delta^{(3)}(\mathbf{k}_1+\mathbf{k}_2). \tag{D2}$$

This is the *dimensionful* primordial power spectrum. Note that A is written as Δ_Φ in [23].

The *dimensionless* primordial power spectrum $\mathcal{P}_\zeta$ (denoted P_ζ in [2]) is defined from

$$\langle \Phi(\mathbf{k}_1)\Phi(\mathbf{k}_2)\rangle = \frac{\mathcal{P}_\Phi(k_1)}{2k_1^3}(2\pi)^5\delta^{(3)}(\mathbf{k}_1+\mathbf{k}_2), \tag{D3}$$

so that

$$\mathcal{P}_\Phi(k) = \frac{k^3}{2\pi^2}P_\Phi(k). \tag{D4}$$

At superhorizon scales, the Bardeen potential Φ and the curvature perturbation ζ are related via $\Phi = (3/5)\zeta$. Hence,

$$P_\Phi(k) = \left(\frac{3}{5}\right)^2 P_\zeta(k), \qquad \mathcal{P}_\Phi(k) = \left(\frac{3}{5}\right)^2 \mathcal{P}_\zeta(k). \tag{D5}$$

In [115], the scalar amplitude A_s and the spectral index n_s are defined through the *dimensionless* curvature power spectrum:

$$\mathcal{P}_\zeta(k) = A_s\left(\frac{k}{k_*}\right)^{n_s-1}, \tag{D6}$$

for some pivot scale $k_* = 0.05\ \mathrm{Mpc}^{-1}$ so that $\mathcal{P}(k_*) = A_s$. This A_s is identical to $\tilde{P}_\zeta$ in [2].

Putting these all together, we relate A from [12] and A_s from [115] as follows:

$$Ak^{n_s-4} = P_\Phi(k) = \frac{2\pi^2}{k^3}\left(\frac{3}{5}\right)^2\mathcal{P}_\zeta(k) = \frac{2\pi^2}{k^3}\left(\frac{3}{5}\right)^2 A_s\left(\frac{k}{k_*}\right)^{n_s-1}, \tag{D7}$$

and therefore

$$A = 2\pi^2\left(\frac{3}{5}\right)^2 k_*^{1-n_s}A_s. \tag{D8}$$

Above is the final conversion relation between the amplitude Planck PNG paper's [12] amplitude A and the Planck parameter estimation paper [115].

APPENDIX E: IMPLEMENTATION AND OPTIMIZATION

The CMB-BEST formalism significantly reduces the computational complexity of CMB bispectrum estimation, but obtaining the linear term β^{lin} in (26) can still be prohibitively expensive unless thoroughly optimized. Better performance also means higher resolution, as we may include mode functions in our basis.

Here, we detail our optimization process: algorithm design, parallel computing, and data locality improvements. Throughout this section, the tensors and functions are treated as discrete multidimensional arrays. Our notation for indices and their limits are summarized in Table IV. A simple trapezoidal rule is used for most numerical integrals except the μ integral in (28), where the Gauss-Legendre quadrature computed from the public code Quadpts [121] is used for the highly oscillatory integrand. Multidimensional arrays are stored in the row-major order following the C convention. The HEALPix library [100] is used for pixelization of the sky, and its component library Libsharp [122] for the SHTs. Note that a large portion of the material in this Appendix is adapted from [110].

A. Algorithm

There are three key quantities to be precomputed and stored for CMB-BEST: Γ (28), β^{cub} (25), and β^{lin} (26). The matrix Γ directly relates to the Fisher information matrix for the estimator, while the two β's provide the amplitude of f_{NL} for each of the CMB observations and simulations.

TABLE IV. Our index conventions for discretized arrays and their range.

Index	Range	Description
r	$[0, N_r)$	Line-of-sight integral r grid index.
p, p_j	$[0, p_{\max})$	Mode number. p_j is a shorthand for (p_1, p_2, p_3).
$i,\ j$	$[0, N_{\mathrm{sim}}]$	Map number. Index $i = 0$ corresponds to the observed map, while $i > 0$ corresponds to different FFP10 simulated maps [97].
n	$[0, N_{\mathrm{pix}})$	Map pixel number.
ℓ, m	$[\ell_{\min}, \ell_{\max}]$	Spherical harmonic multipole moments. Note $-\ell \le m \le \ell$.
μ	$[0, N_\mu)$	Grid index for the Gauss-Legendre quadrature.

Algorithm 1. Computing β's: Fast and memory efficient.

```
Allocate $m(p, n)$                                              ▷ Memory $\sim p_{\max} \cdot N_{\rm pix}$
Allocate $C(p_1, p_2, n)$                                       ▷ Memory $\sim p_{\max}^2 \cdot N_{\rm pix}$

for each map $i$ do
    for each mode $p$ do                                        ▷$O(N_{\rm sims} \cdot p_{\max} \cdot N_{\rm pix}^{3/2})$
        compute $M(i, p, n)$ by SHT and store in $m(p, n)$
    end for

    for each pair of modes $(p_1, p_2)$ do
        for each pixel $n$ do                                   ▷ $O(N_{\rm sim} \cdot p_{\max}^2 \cdot N_{\rm pix})$
            $C(p_1, p_2, n) += m(p_1, n) \cdot m(p_2, n)$
        end for
    end for
end for                                                         ▷ $C(p_1, p_2, n)$ ready

for each map $i$ do
    for each of mode $p$ do                                     ▷ $O(N_{\rm sim} \cdot p_{\max} \cdot N_{\rm pix}^{3/2})$
        compute $M(i, p, n)$ by SHT and store in $m(p, n)$
    end for

    for each set of modes $(p_1, p_2, p_3)$ do
        for each pixel $n$ do                                   ▷ $O(N_{\rm sim} \cdot p_{\max}^3 \cdot N_{\rm pix})$
            $\beta^{\rm cub}(i, p_1, p_2, p_3) += m(p_1, n) \cdot m(p_2, n) \cdot m(p_3, n)$
            $\beta^{\rm lin}(i, p_1, p_2, p_3) += C(p_1, p_2, n) \cdot m(p_3, n)$
        end for
    end for
end for
```

In most cases, the bottleneck point of our pipeline is computing the linear term $\beta^{\rm lin}$. Even though the Γ matrix scales more rapidly with the number of basis functions ($\propto p_{\max}^6$) than the β's ($\propto p_{\max}^3$), it does not involve operations with high-resolution maps and remains subdominant in terms of the total cost in the regime we consider ($p_{\max} \leq 30$).

The discretized versions of (25) and (26) are given by

$$\beta^{\rm cub}(i, p_1, p_2, p_3) = \sum_r \sum_n M(r, i, p_1, n) \cdot M(r, i, p_2, n) \cdot M(r, i, p_3, n), \quad \text{(E1)}$$

$$\beta^{\rm lin}(i, p_1, p_2, p_3) = \sum_r \sum_{j \neq i} \sum_n M(r, j, p_1, n) \cdot M(r, j, p_2, n) \cdot M(r, i, p_3, n), \quad \text{(E2)}$$

respectively. The order of indices is chosen such that the following calculations have optimal memory layouts. Some integral weights and factors are absorbed into arrays for brevity.

First, note that the data arrays for different values of r are completely independent of each other. This provides us with a natural way to split tasks. We compute and save contributions to β's for each r separately and summed over at the end with minimal overhead. Therefore, throughout the rest of this chapter, we assume that r is fixed and drop the r dependence in the descriptions of our algorithms.

The filtered map arrays $M(i, p, n)$ are obtained as follows. A given map i is first transformed into spherical harmonic coefficients $a^{(i)}(\ell, m)$'s via SHT. We then compute $\tilde{q}_(p, l) * a(i, \ell, m)/C(\ell)$ from (22) through inverse SHT to synthesize the filtered maps.

As a rough guide to the size of each summation, we typically have $N_{\rm sim} = 160$ simulations, $p_{\max} = 30$ modes, and $N_{\rm pix} = 50,331,648$ pixels.[10] Considering the fact that one double-precision array of size ~50 million pixels takes about 400 MB of memory space, this is indeed a task for supercomputers.

Finding the optimal algorithm for β computation was about reducing the computational complexity while keeping the memory footage in check. Computing and saving some intermediate quantities often reduce the computational complexity but can often be infeasible due to maximum memory restraints. We describe our final algorithm below as Algorithm 1. A more detailed reasoning can be found in [110].

The trick is to reduce the amount of memory required at the cost of doubling the SHTs for computing $M(i, p, n)$'s

[10] This value corresponds to $N_{\rm side} = 2048$ in HEALPix: $N_{\rm pix} = 12 N_{\rm side}^2$.

which are subdominant. The quantity $C(p_1, p_2, n)$ is precomputed during the first round of SHTs. This quantity is then used during the second round of SHTs to compute the ß s for each map. Algorithm 1 requires memory of size $\approx p_{\max}^2 N_{\rm pix}$ and has leading computational complexity of $O(N_{\rm sim} p_{\max}^3 N_{\rm pix})$.

SHTs have a subdominant contribution to the total computation time even after becoming doubled in number. One of the main strengths of Algorithm 1 is that both the memory and computation time scale linearly with the number of simulations used, $N_{\rm sims}$. In the future when a larger number of Gaussian simulations is required to acquire a more accurate estimate of the linear term, it is straightforward to adapt our method accordingly.

B. Parallelism

In order to fully benefit from modern computer architecture, we introduce parallelism in multiple levels of CMB-BEST for optimal performance on supercomputers.

Following [107], we discuss parallelization at three different levels. They mostly correspond to nodes, cores, and registers/cache in modern computer clusters. Each level has distinct characteristics which make them ideal for different parallelization techniques. We make full use of each level in our methodology.

The first level of parallelism relates to dividing the main work into many computation-heavy tasks with limited data communication between them. Since the tasks are largely independent, each of them can be assigned to independent nodes or to the message passing interface (MPI) [123] ranks. We exploit this level of parallelism in CMB-BEST by splitting the line-of-sight integration over r; almost no data are shared between different r's despite the heavy operations within each of them. CMB-BEST scales well with the number of computing units assigned for each r point (or several r points).

The second level of parallelism is for multiple computational subtasks on a single set of data. Most modern supercomputers use multicore processors. Each core, or processing unit, can run one or more threads, executing instructions independently from each other while sharing memory space. MPI would not be as effective here due to the large amount of data sharing required; ranks would have to either continuously communicate with each other, or store duplicate copies of the data. This type of parallelism is required in the SHTs and map operations of CMB-BEST. We use open multiprocessing (OpenMP) [124] for multithreading in C for this purpose.

The third level of parallelism applies to the identical arithmetic operations applied to multiple data items, ideally adjacent in memory space: single instruction, multiple data. Many procedures in CMB-BEST involve large array operations which benefit from this type of parallelism. We use advanced vector extensions (AVX) supported by Intel processors to exploit this, especially for operations involving large filtered map arrays. In particular, Intel's Xeon Phi series' AVX-512 implementation, where the 512-bit registers hold up to eight double-precision floating numbers [107], provided a major boost to the computation speed.

C. Data locality

We implemented Algorithm 1 and profiled it using the Intel VTune Amplifier. Our program was found to be memory bound, meaning its speed is limited mainly by the speed of memory access. The CPU speed, rate of the file I/O, and MPI communication all have subdominant contributions in comparison. This is somewhat expected since our method deals with large map arrays. The number of operations on each data element is small compared with the size of data, causing the CPUs to be "starved" for data to work on most of the time. Our final set of optimization focuses on improving memory access patterns.

CPUs of most modern computers contain a small amount of memory attached to them called a *cache*. Recently used data and instructions are stored in cache memory so that reusing them is more efficient; accessing them is much faster than loading from the main, larger memory often shared with other CPUs. A cache is often divided into multiple levels. The smallest and fastest is the L1 cache, which is the first level a CPU checks for data. When the required data are not stored in the L1 cache, a *cache miss* occurs. The system then has to look further down the cache levels to fetch the desired data, incurring a large time loss. As the cache "caches" memory locations in units of cache lines, or chunks of memory containing multiple data elements, accessing memory locally significantly increases the chance of cache hits and boosts overall performance.

CMB-BEST has been modified in two ways to improve data locality and memory performance. The first one was simple yet effective; we made sure to initialize large arrays within the same OpenMP construct as the main computation loop. This guarantees the physical memory of array elements to be allocated near the cores where they are going to be used. Memory access during the main computation loop is therefore much faster than it would be otherwise. Systems with nonuniform memory access especially benefit from this method. For CMB-BEST, we gained a two-times speedup compared with when a single master thread initialized the entire array.

The second optimization centers around *cache blocking*, a technique used to maximize data reuse. The most expensive loop from Algorithm 1 is:

```
**for** each set of modes $(p_1, p_2, p_3)$ **do**
  **for** each pixel $n$ **do**
    $\beta^{\rm cub}(i, p_1, p_2, p_3) += m(p_1, n) \cdot m(p_2, n) \cdot m(p_3, n)$
    $\beta^{\rm lin}(i, p_1, p_2, p_3) += C(p_1, p_2, n) \cdot m(p_3, n)$
  **end for**
**end for**
```

In every outermost loop, four large arrays are read from memory: $m(p_1,\cdot)$, $m(p_2,\cdot)$, $m(p_3,\cdot)$, and $C(p_1,p_2,\cdot)$. Each of them takes up around 400 MB of memory. Since their size is greater than the cache storage capacity, data in front of them are gone from the cache by the time a loop over n completes. All four arrays will then have to be loaded from the main memory again when the next iteration starts.

To allow different parts of the array to be reused before they are lost in cache, we divide the large arrays into equally sized blocks that fit inside the cache memory. We restructure the loop as follows:

```
for each block $b$ do
    for each set of modes $(p_1, p_2, p_3)$ do
        for each pixel $n'$ in block do
            $\beta^{\mathrm{cub}}(i, p_1, p_2, p_3) += m(p_1, n') \cdot m(p_2, n') \cdot m(p_3, n')$
            $\beta^{\mathrm{lin}}(i, p_1, p_2, p_3) += C(p_1, p_2, n') \cdot m(p_3, n')$
        end for
    end for
end for
```

New pixel numbers are calculated as $n' = B \cdot b + n$, where B is the size of each block and $0 \leq n < B$. We have not changed the total number of arithmetic operations required, so the computational complexity remains the same. Meanwhile, data locality within each block is greatly improved, as each of the blocked arrays is now small enough to fit in the cache. Each data element is accessed in closer succession temporarily as well.

One caveat here is that having too many blocks may degrade the overall performance. There exists a non-negligible overhead coming from an extra *for* loop and the OpenMP construct used over n'. The block size divided by the number of cores should not be smaller than the size of cache lines either. The optimal size of cache blocks depends on the memory architecture of the processor used. This often needs to be found empirically. For our implementation, we found the optimal number of blocks to be 128, yielding a three-times speedup compared with the original code without cache blocking.

APPENDIX F: VALIDATION

1. Map-by-map validation of CMB-BEST

In this section, we show map-by-map validation of the CMB-BEST pipeline against the modal estimator used in Planck analyses.

Figures 8 and 9 show comparisons between the Monomials and Legendre basis sets of CMB-BEST. A total of 160 Planck's FFP10 SMICA-component-separated simulations [97,99] were used for this analysis. For each simulated map, we compare the f_{NL}'s of the standard templates that have been computed using two different bases. The two of them agree almost exactly, with a map-by-map correlation of more than 0.99 in all cases.

Despite using identical datasets, we see a small scatter between the Monomials and Legendre basis sets in the local f_{NL}. This is mainly due to the limited k range of the Legendre basis: $[2.08 \times 10^{-4}, 2.08 \times 10^{-1}]$ Mpc^{-1}. The range has been set wide enough to cover the scales where CMB information comes from, but also narrow enough to have enough resolution for expanding oscillatory templates with polynomials of maximum order $p_{\max}$. Loss of large-scale information with $k < 2.08 \times 10^{-4}$ has a small effect on equilateral and orthogonal shapes, but affects the local shape through the squeezed limit. The error bar $\sigma(f_{\mathrm{NL}})$ is about 7% smaller for the Monomials basis for this reason. Choosing a wider k range for the Legendre basis has been shown to remove this scatter [110].

We also perform a map-by-map comparison of the CMB-BEST pipeline with Planck's modal estimator. The results are shown in Fig. 10. Overall, we find the two methods to be consistent and have f_{NL} correlations varying between 0.93 and 0.97. The level of scatter shown here is within the bounds seen between the different estimators used in Planck [10]. The local template shows a slightly higher level of discrepancy due to the limited k range in the Legendre basis mentioned above. For a detailed analysis of the reason why different estimators are not perfectly correlated, we refer to the appendix of [10].

2. Optimality of the bispectrum estimator

In this section, we test the optimality of CMB-BEST estimation to evaluate the validity of the Fisher matrix approximation for writing down the CMB bispectrum likelihood (5).

The error given by the Fisher matrix in the likelihood is a rather theoretical quantity; it measures the expected scatter of the estimated f_{NL} values across different realizations of the Universe. Unfortunately, we only have one universe to observe from, so simulated maps are used instead. As before, 160 CMB maps from FFP10 simulations after SMICA component separation [97–99] are plugged into CMB-BEST pipeline to compute the f_{NL} estimates. The simulations are based on Gaussian initial conditions.

The analysis would benefit from having more simulations, but the publicly available maps from the Planck Legacy Archive are only usable up to simulation number 160,[11] since the ones after then are currently erroneous. Their temperature and polarization maps yield angular

[11] The files are named dx12_v3_smica_cmb_mc_00xxx_raw.fits, where xxx corresponds to the simulation number.

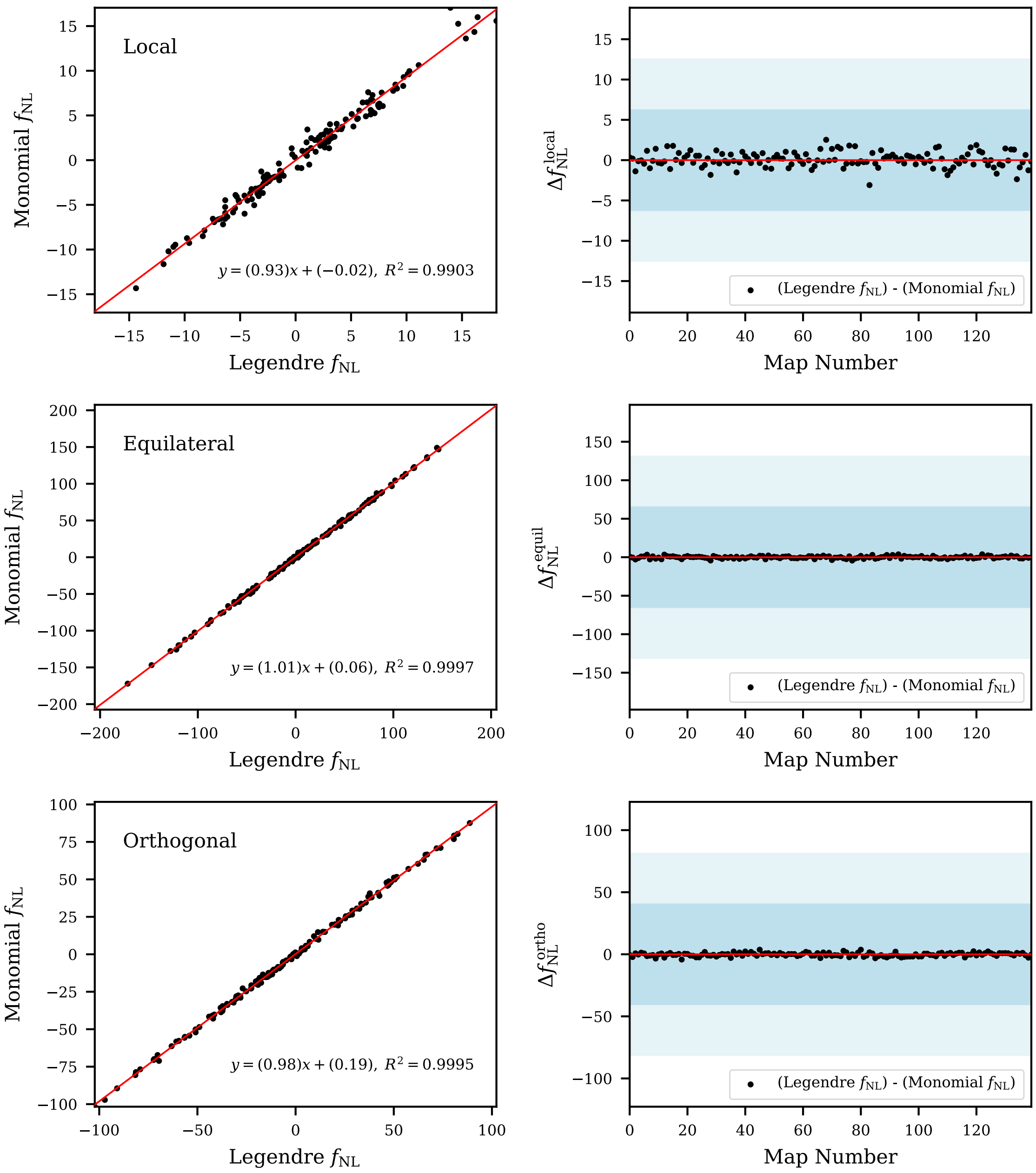


FIG. 8. A map-by-map comparison of the f_{NL} estimates evaluated using the Monomials (KSW) and Legendre basis sets for three standard templates. The Planck 2018 CMB temperature map and 160 FFP10 simulations have been used, each representing a single point on the scatter plot (left). Details of the linear regression to data (red) are annotated below. Shown on the right-hand side are plots of the differences in the f_{NL} values for each map, together with the 1σ and 2σ intervals shaded in blue. In the ideal case where the two basis sets yield identical results, we should see all the points lie on the line $y = x$ for the left plot and $y = 0$ for the right plot. For more information on each of the three theoretical templates used, see e.g., [10].

power spectra that are consistent with the background cosmology, but the cross spectra (C_ℓ^{TE}) are completely inconsistent.

Figures 11 and 12 show comparisons between the distribution of the f_{NL} estimates from simulations and the expected distribution computed using the Fisher matrix. The Python library GetDist [116] was used to plot the sample points as contours. The irregular shapes of the contours are mainly due to the limited number of simulations used. Overall, we find that the Fisher matrix well approximates the error except for a slight underestimation between the 4% and 10% level.

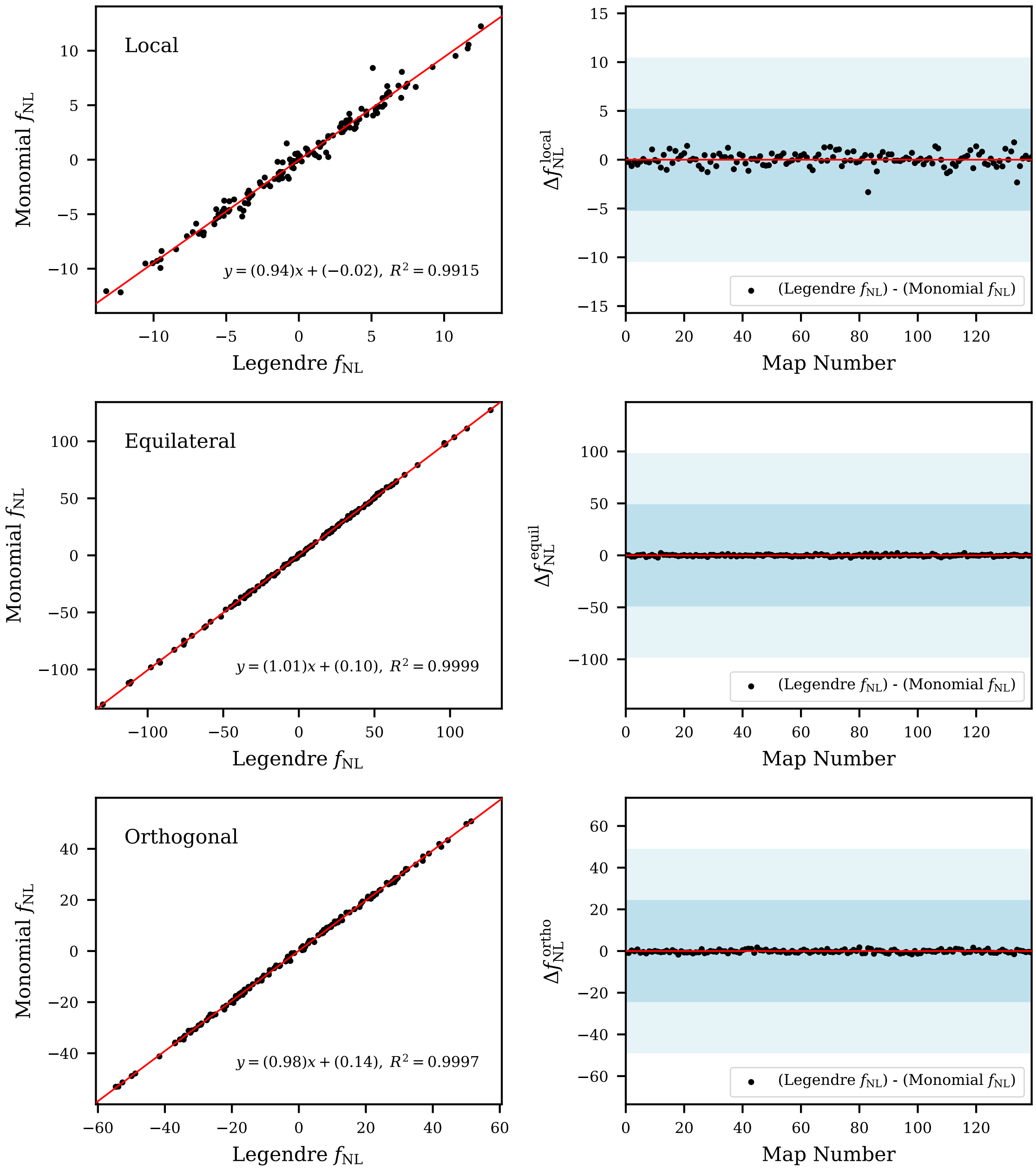


FIG. 9. A map-by-map comparison of the $f_{\rm NL}$ estimates evaluated using the Monomials (KSW) and Legendre basis sets for three standard templates. The same as Fig. 8 but using both temperature and E-mode polarization.

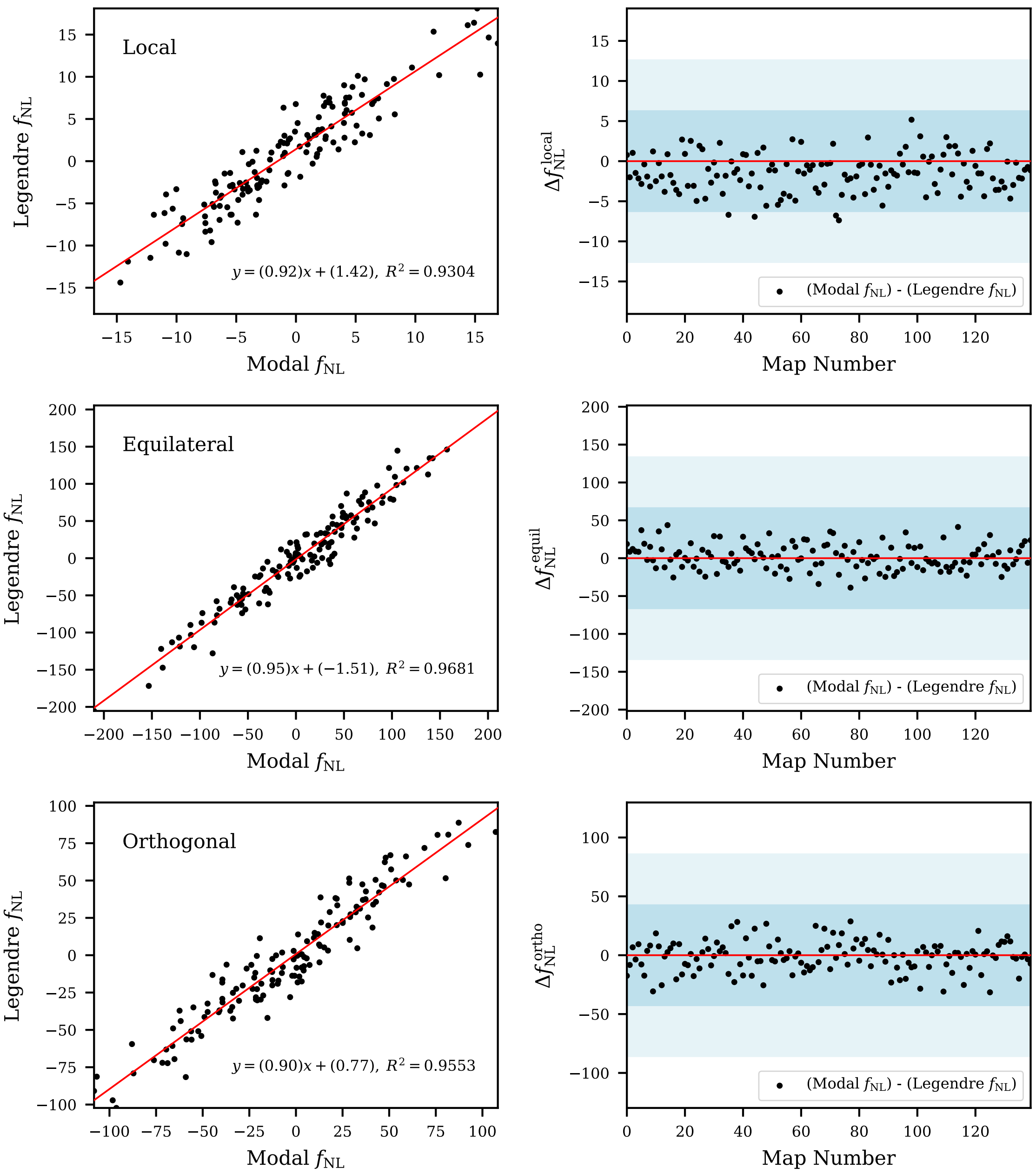


FIG. 10. A map-by-map comparison of the $f_{\rm NL}$ estimates obtained from the CMB-BEST's Legendre basis set against the modal estimator results of the Planck 2018 analysis [12]. The first 160 FFP10 simulations are used here. On the left-hand side are scatter plots where each simulation is represented by a point according to $f_{\rm NL}$ estimates of standard templates. Their linear best-fit lines are shown in red. Differences in the estimates from the two routines are shown map-by-map on the right-hand side, together with the 1σ and 2σ levels shaded in blue. Overall, CMB-BEST and modal are in good agreement without any significant systematic errors.

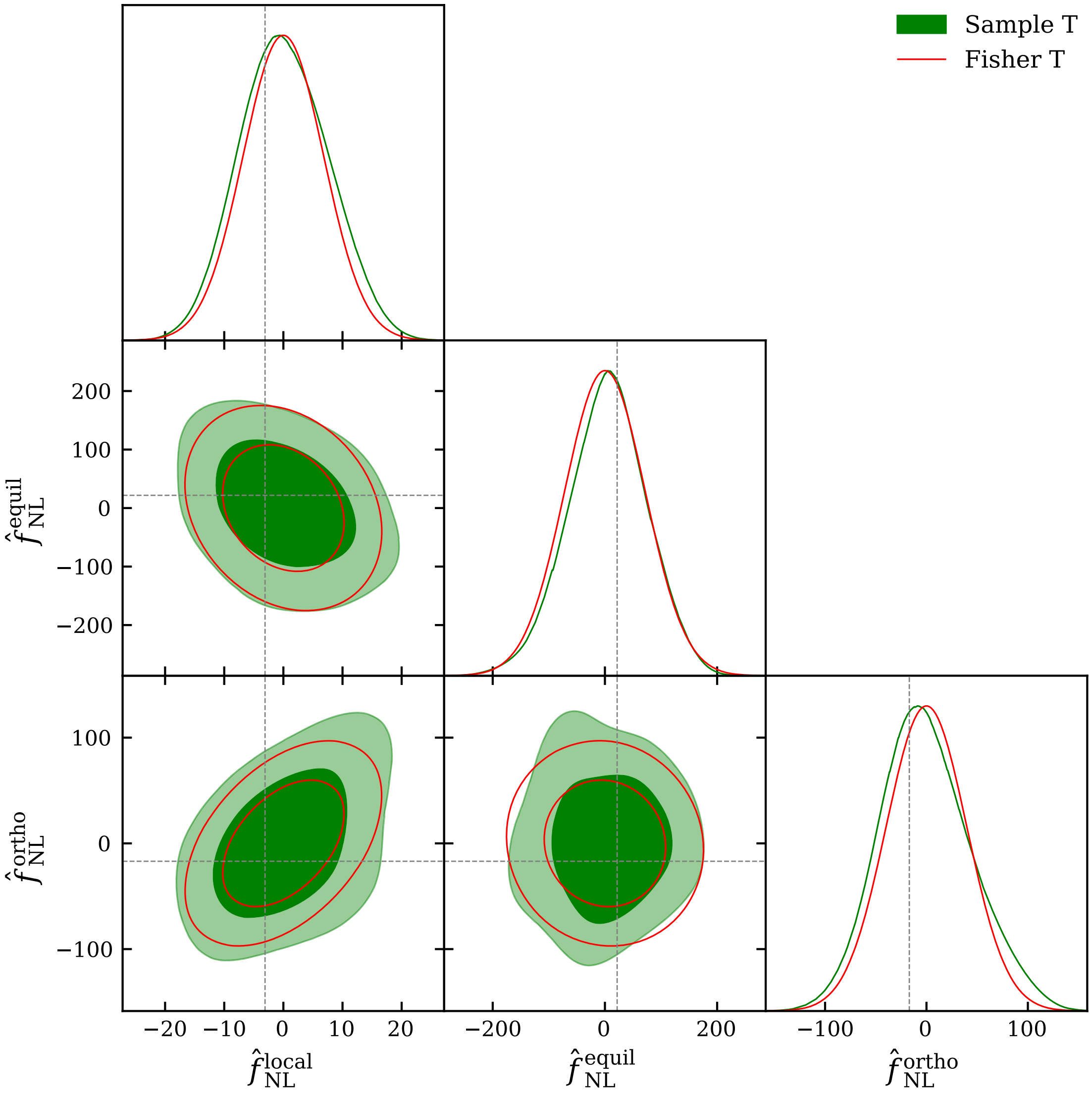


FIG. 11. Comparison between the likelihoods estimated from the Fisher matrix and from $f_{\rm NL}$ estimates of 160 FFP10 simulated temperature maps, under Gaussian initial conditions. The dashed lines indicate the marginal $f_{\rm NL}$ estimates from observations. The two likelihoods are consistent with each other, even though the Fisher one underestimates the marginalized error by up to 10%. Note that the contour looks slightly irregular due to the limited number of samples.

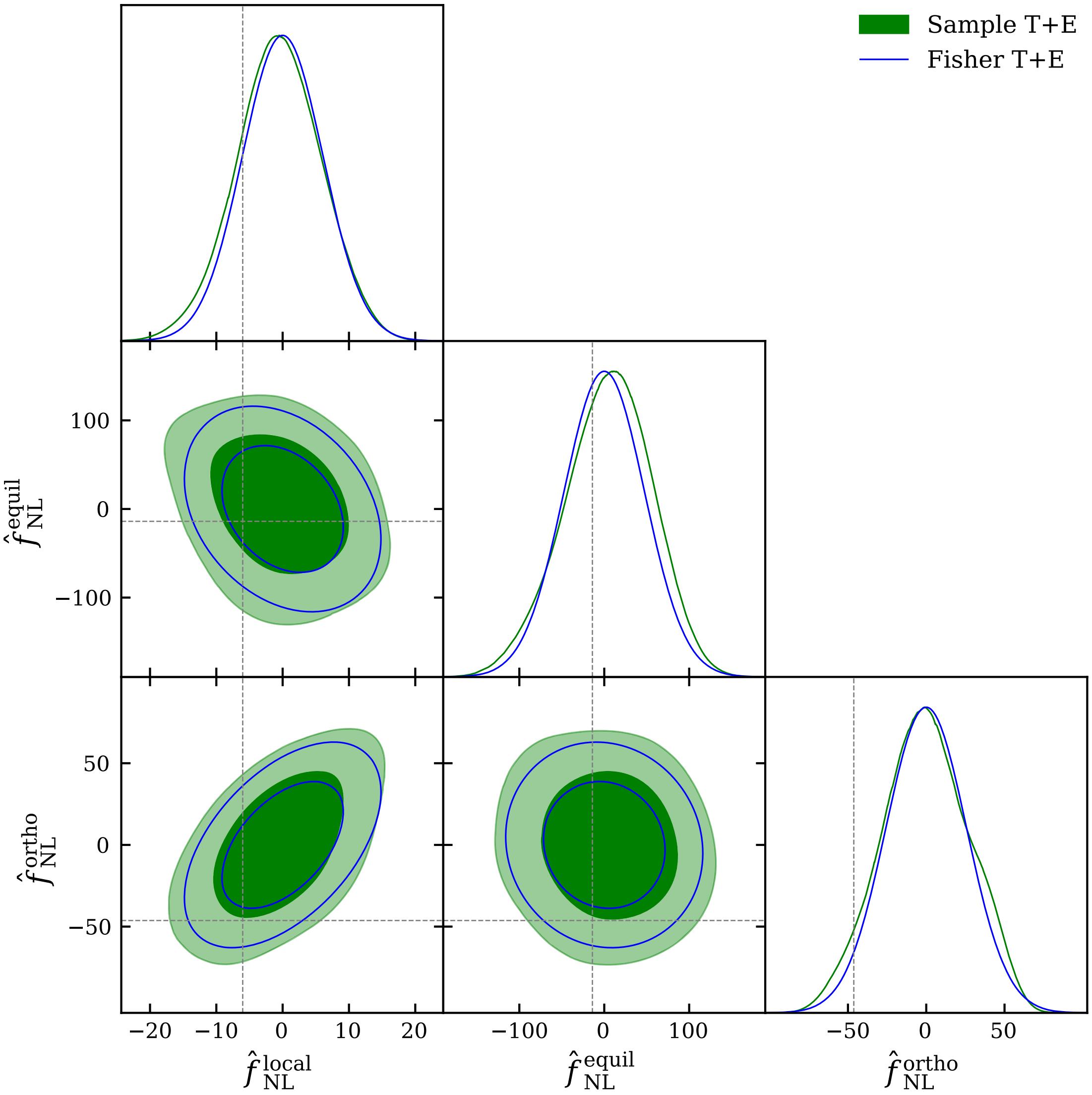


FIG. 12. Comparison between the likelihoods estimated from the Fisher matrix and from $f_{\rm NL}$ estimates of 160 FFP10 simulated temperature and polarization maps, under Gaussian initial conditions. Same as Fig. 11 but with both temperature and polarization included. We again find that the two likelihoods are consistent with each other despite the underestimation of the marginalized error by up to 9%.

[1] P. D. Meerburg *et al.*, Astro2020 Science White Paper (2019), arXiv:1903.04409.

[2] X. Chen, Adv. Astron. **2010**, 638979 (2010).

[3] E. Komatsu, Classical Quantum Gravity **27**, 124010 (2010).

[4] M. Liguori, E. Sefusatti, J. R. Fergusson, and E. P. Shellard, Adv. Astron. **2010**, 980523 (2010).

[5] V. Desjacques and U. Seljak, Classical Quantum Gravity **27**, 124011 (2010).

[6] S. Renaux-Petel, C. R. Phys. **16**, 969 (2015).

[7] N. Bartolo, S. Matarrese, and A. Riotto, Phys. Rev. D **65**, 103505 (2002).

[8] J. Maldacena, J. High Energy Phys. 05 (2003) 013.

[9] N. Bartolo, E. Komatsu, S. Matarrese, and A. Riotto, Phys. Rep. **402**, 103 (2004).

[10] Planck Collaboration, Astron. Astrophys. **571**, A24 (2014).

[11] Planck Collaboration, Astron. Astrophys. **594**, A17 (2016).

[12] Planck Collaboration, Astron. Astrophys. **641**, A9 (2020).

[13] C. L. Bennett, D. Larson, J. L. Weiland, N. Jarosik, G. Hinshaw, N. Odegard, K. Smith, R. Hill, B. Gold, M. Halpern *et al.*, Astrophys. J. Suppl. Ser. **208**, 20 (2013).

[14] E.-M. Mueller, M. Rezaie, W. J. Percival, A. J. Ross, R. Ruggeri, H.-J. Seo, H. Gil-Marín, J. Bautista, J. R. Brownstein, K. Dawson, A. de la Macorra, N. Palanque-Delabrouille, G. Rossi, D. P. Schneider, and C. Yéche, Mon. Not. R. Astron. Soc. **514**, 3396 (2022).

[15] G. Cabass, M. M. Ivanov, O. H. E. Philcox, M. Simonović, and M. Zaldarriaga, Phys. Rev. Lett. **129**, 021301 (2022).

[16] E. Komatsu, D. N. Spergel, and B. D. Wandelt, Astrophys. J. **634**, 14 (2005).

[17] P. Creminelli, A. Nicolis, L. Senatore, M. Tegmark, and M. Zaldarriaga, J. Cosmol. Astropart. Phys. 05 (2006) 004.

[18] A. P. S. Yadav, E. Komatsu, and B. D. Wandelt, Astrophys. J. **664**, 680 (2007).

[19] L. Senatore, K. M. Smith, and M. Zaldarriaga, J. Cosmol. Astropart. Phys. 01 (2010) 028.

[20] K. M. Smith and M. Zaldarriaga, Mon. Not. R. Astron. Soc. **417**, 2 (2011).

[21] D. Munshi and A. Heavens, Mon. Not. R. Astron. Soc. **401**, 2406 (2010).

[22] J. R. Fergusson, M. Liguori, and E. P. S. Shellard, Phys. Rev. D **82**, 023502 (2010).

[23] J. R. Fergusson, M. Liguori, and E. P. S. Shellard, J. Cosmol. Astropart. Phys. 12 (2012) 032.

[24] J. R. Fergusson, Phys. Rev. D **90**, 043533 (2014).

[25] M. Bucher, B. van Tent, and C. S. Carvalho, Mon. Not. R. Astron. Soc. **407**, 2193 (2010).

[26] M. Bucher, B. Racine, and B. Van Tent, J. Cosmol. Astropart. Phys. 05 (2016) 055.

[27] A. A. Starobinskij, JETP Lett. **55**, 489 (1992).

[28] J. Adams, B. Cresswell, and R. Easther, Phys. Rev. D **64**, 123514 (2001).

[29] L. Covi, J. Hamann, A. Melchiorri, A. Slosar, and I. Sorbera, Phys. Rev. D **74**, 083509 (2006).

[30] X. Chen, R. Easther, and E. A. Lim, J. Cosmol. Astropart. Phys. 06 (2007) 023.

[31] X. Chen, R. Easther, and E. A. Lim, J. Cosmol. Astropart. Phys. 04 (2008) 010.

[32] P. Adshead, W. Hu, C. Dvorkin, and H. V. Peiris, Phys. Rev. D **84**, 043519 (2011).

[33] P. Adshead, C. Dvorkin, W. Hu, and E. A. Lim, Phys. Rev. D **85**, 023531 (2012).

[34] D. K. Hazra, A. Shafieloo, G. F. Smoot, and A. A. Starobinsky, Phys. Rev. Lett. **113**, 071301 (2014).

[35] A. Achúcarro, J. O. Gong, S. Hardeman, G. A. Palma, and S. P. Patil, J. Cosmol. Astropart. Phys. 01 (2011) 030.

[36] M. Nakashima, R. Saito, Y.-i. Takamizu, and J. Yokoyama, Prog. Theor. Phys. **125**, 1035 (2011).

[37] M. Park and L. Sorbo, Phys. Rev. D **85**, 083520 (2012).

[38] A. Achucarro, V. Atal, P. Ortiz, and J. Torrado, Phys. Rev. D **89**, 103006 (2014).

[39] A. Achucarro, V. Atal, B. Hu, P. Ortiz, and J. Torrado, Phys. Rev. D **90**, 023511 (2014).

[40] X. Gao, D. Langlois, and S. Mizuno, J. Cosmol. Astropart. Phys. 10 (2012) 040.

[41] X. Gao, D. Langlois, and S. Mizuno, J. Cosmol. Astropart. Phys. 10 (2013) 023.

[42] T. Noumi, M. Yamaguchi, and D. Yokoyama, J. High Energy Phys. 06 (2013) 051.

[43] T. Noumi and M. Yamaguchi, J. Cosmol. Astropart. Phys. 12 (2013) 038.

[44] E. Silverstein and A. Westphal, Phys. Rev. D **78**, 106003 (2008).

[45] L. McAllister, E. Silverstein, and A. Westphal, Phys. Rev. D **82**, 046003 (2010).

[46] R. Flauger, L. McAllister, E. Pajer, A. Westphal, and G. Xu, J. Cosmol. Astropart. Phys. 06 (2010) 009.

[47] M. Berg, E. Pajer, and S. Sjörs, Phys. Rev. D **81**, 103535 (2010).

[48] R. Flauger and E. Pajer, J. Cosmol. Astropart. Phys. 01 (2011) 017.

[49] M. Aich, D. K. Hazra, L. Sriramkumar, and T. Souradeep, Phys. Rev. D **87**, 083526 (2013).

[50] S. R. Behbahani and D. Green, J. Cosmol. Astropart. Phys. 11 (2012) 056.

[51] J. Chluba, J. Hamann, and S. P. Patil, Int. J. Mod. Phys. D **24**, 1530023 (2015).

[52] A. Slosar, X. Chen, C. Dvorkin, D. Green, P. D. Meerburg, E. Silverstein, and B. Wallisch, arXiv:1903.09883.

[53] A. Achúcarro, M. Biagetti, M. Braglia, G. Cabass, R. Caldwell, E. Castorina, X. Chen, W. Coulton, R. Flauger, J. Fumagalli *et al.*, arXiv:2203.08128.

[54] J. Martin and C. Ringeval, Phys. Rev. D **69**, 083515 (2004).

[55] C. Pahud, M. Kamionkowski, and A. R. Liddle, Phys. Rev. D **79**, 083503 (2009).

[56] D. K. Hazra, M. Aich, R. K. Jain, L. Sriramkumar, and T. Souradeep, J. Cosmol. Astropart. Phys. 10 (2010) 008.

[57] C. Dvorkin and W. Hu, Phys. Rev. D **84**, 063515 (2011).

[58] M. Benetti, M. Lattanzi, E. Calabrese, and A. Melchiorri, Phys. Rev. D **84**, 063509 (2011).

[59] P. D. Meerburg, R. A. Wijers, and J. P. van der Schaar, Mon. Not. R. Astron. Soc. **421**, 369 (2012).

[60] X. Chen and C. Ringeval, J. Cosmol. Astropart. Phys. 08 (2012) 014.

[61] M. Benetti, Phys. Rev. D **88**, 087302 (2013).

[62] D. K. Hazra, A. Shafieloo, and T. Souradeep, J. Cosmol. Astropart. Phys. 07 (2013) 031.

[63] P. D. Meerburg, D. N. Spergel, and B. D. Wandelt, Phys. Rev. D **89**, 063536 (2014).

[64] P. D. Meerburg, D. N. Spergel, and B. D. Wandelt, Phys. Rev. D **89**, 063537 (2014).

[65] Planck Collaboration, Astron. Astrophys. **571**, A15 (2014).

[66] X. Chen and M. H. Namjoo, Phys. Lett. B **739**, 285 (2014).

[67] Planck Collaboration, Astron. Astrophys. **594**, A11 (2016).

[68] Planck Collaboration, Astron. Astrophys. **594**, A16 (2016).

[69] Planck Collaboration, Astron. Astrophys. **641**, A5 (2020).

[70] G. Cañas-Herrera, J. Torrado, and A. Achúcarro, Phys. Rev. D **103**, 123531 (2021).

[71] P. D. Meerburg, Phys. Rev. D **82**, 063517 (2010).

[72] M. Munchmeyer, F. Bouchet, M. G. Jackson, and B. Wandelt, Astron. Astrophys. **570**, A94 (2014).

[73] M. Munchmeyer, P. D. Meerburg, and B. D. Wandelt, Phys. Rev. D **91**, 043534 (2015).
[74] P. D. Meerburg and M. Munchmeyer, Phys. Rev. D **92**, 063527 (2015).
[75] J. R. Fergusson, H. F. Gruetjen, E. P. S. Shellard, and M. Liguori, Phys. Rev. D **91**, 023502 (2015).
[76] J. R. Fergusson, H. F. Gruetjen, E. P. S. Shellard, and B. Wallisch, Phys. Rev. D **91**, 123506 (2015).
[77] P. D. Meerburg, M. Munchmeyer, and B. Wandelt, Phys. Rev. D **93**, 043536 (2016).
[78] J. Hamann, L. Covi, A. Melchiorri, and A. Slosar, Phys. Rev. D **76**, 023503 (2007).
[79] B. Hu and J. Torrado, Phys. Rev. D **91**, 064039 (2015).
[80] M. Benetti and J. S. Alcaniz, Phys. Rev. D **94**, 023526 (2016).
[81] F. Beutler, M. Biagetti, D. Green, A. c. v. Slosar, and B. Wallisch, Phys. Rev. Res. **1**, 033209 (2019).
[82] T. Mergulhão, F. Beutler, and J. A. Peacock, arXiv: 2303.13946.
[83] A. Duivenvoorden, KSW, https://github.com/AdriJD/ksw.
[84] K. Fornazier, Bispectrum calc, https://github.com/kfornaz/Bispectrum-calc.
[85] W. Sohn, CMB-BEST: A code for CMB bispectrum estimation of primordial non-Gaussianity (2023), https://github.com/Wuhyun/CMB-BEST.
[86] W. Sohn, Tetraquad: A code for computing numerical quadrature rule for integrals over a tetrapyd domain (2023), https://github.com/Wuhyun/Tetraquad.
[87] The Simons Observatory Collaboration, J. Cosmol. Astropart. Phys. 02 (2019) 056.
[88] K. N. Abazajian *et al.*, arXiv:1610.02743.
[89] K. Abazajian, G. Addison, P. Adshead, Z. Ahmed, S. W. Allen, D. Alonso, M. Alvarez, A. Anderson, K. S. Arnold, C. Baccigalupi *et al.*, arXiv:1907.04473.
[90] A. Becker and D. Huterer, Phys. Rev. Lett. **109**, 121302 (2012).
[91] A. Heavens, Mon. Not. R. Astron. Soc. **299**, 805 (1998).
[92] E. Komatsu and D. N. Spergel, Phys. Rev. D **63**, 063002 (2001).
[93] P. Creminelli, L. Senatore, and M. Zaldarriaga, J. Cosmol. Astropart. Phys. 03 (2007) 019.
[94] W. R. Coulton, P. D. Meerburg, D. G. Baker, S. Hotinli, A. J. Duivenvoorden, and A. van Engelen, Phys. Rev. D **101**, 123504 (2020).
[95] A. Lewis, A. Challinor, and A. Lasenby, Astrophys. J. **538**, 473 (2000).
[96] D. Blas, J. Lesgourgues, and T. Tram, J. Cosmol. Astropart. Phys. 07 (2011) 034.
[97] Planck Collaboration, Astron. Astrophys. **594**, A12 (2016).
[98] Planck Collaboration, Astron. Astrophys. **641**, A3 (2020).
[99] Planck Collaboration, Astron. Astrophys. **641**, A4 (2020).
[100] K. M. Gorski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann, Astrophys. J. **622**, 759 (2005).
[101] A. Zonca, L. Singer, D. Lenz, M. Reinecke, C. Rosset, E. Hivon, and K. Gorski, J. Open Source Softwaare **4**, 1298 (2019).
[102] H. F. Gruetjen, J. R. Fergusson, M. Liguori, and E. P. S. Shellard, Phys. Rev. D **95**, 043532 (2017).
[103] S. Dodelson, *Modern Cosmology* (Elsevier, New York, 2003).
[104] W. Hu, Phys. Rev. D **62**, 043007 (2000).
[105] D. Hanson, K. M. Smith, A. Challinor, and M. Liguori, Phys. Rev. D **80**, 083004 (2009).
[106] A. Lewis, A. Challinor, and D. Hanson, J. Cosmol. Astropart. Phys. 03 (2011) 018.
[107] J. Jeffers, J. Reinders, and A. Sodani, *Intel Xeon Phi Processor High Performance Programming: Knights Landing Edition* (Elsevier Science, New York, 2016).
[108] M. R. Hestenes, E. Stiefel *et al.*, J. Res. Natl. Bur. Stand. **49**, 409 (1952).
[109] P. Virtanen *et al.* (SciPy 1.0 Contributors), Nat. Methods **17**, 261 (2020).
[110] W. H. Sohn, High-resolution CMB bispectrum estimator, Ph.D. thesis, University of Cambridge, 2022.
[111] C. W. Ueberhuber, *Numerical Computation 1: Methods, Software, and Analysis* (Springer Science & Business Media, New York, 1997), Vol. 16.
[112] J. Jaśkowiec and N. Sukumar, Int. J. Numer. Methods Eng. **122**, 148 (2021).
[113] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems* (SIAM, USA, 1995).
[114] G. H. Hardy, *Some Famous Problems of the Theory of Numbers and in Particular Waring's Problem* (Clarendon Press, 1920).
[115] Planck Collaboration, Astron. Astrophys. **641**, A6 (2020).
[116] A. Lewis, arXiv:1910.13970.
[117] W. R. Inc., *Mathematica*, Version 13.2, Champaign, IL, 2022.
[118] DLMF, NIST Digital Library of Mathematical Functions, http://dlmf.nist.gov/, Release 1.1.8 of 2022-12-15, edited by f. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain.
[119] The mpmath development team, mpmath: A PYTHON library for arbitrary-precision floating-point arithmetic (version 1.3.0) (2023), http://mpmath.org/.
[120] W. Sohn and J. R. Fergusson, Phys. Rev. D **100**, 063536 (2019).
[121] N. Hale and A. Townsend, SIAM J. Sci. Comput. **35**, A652 (2013).
[122] M. Reinecke and D. S. Seljebotn, Astron. Astrophys. **554**, A112 (2013).
[123] A. S. William Gropp and Ewing Lusk, *Using MPI: Portable Parallel Programming with the Message-Passing Interface* (MIT Press, Cambridge, MA, 1999), Vol. 1.
[124] L. Dagum and R. Menon, IEEE Comput. Sci. Eng. **5**, 46 (1998).