

CONTROL VARIATES FROM EULERIAN AND LAGRANGIAN PERTURBATION THEORY: APPLICATION TO THE BISPECTRUM

NICKOLAS KOKRON[✉] AND SHI-FAN CHEN[✉]

[✉]School of Natural Sciences, Institute for Advanced Study, 1 Einstein Drive, Princeton, NJ 08540, USA

[✉]Department of Physics, Columbia University, New York, NY, USA 10027 and

[✉]NASA Hubble Fellowship Program, Einstein Fellow

Version October 10, 2025

Abstract

Control variates have seen recent interest as a powerful technique to reduce the variance of summary statistics measured from costly cosmological N -body simulations. Of particular interest are the class of control variates which are analytically calculable, such as the recently introduced ‘Zeldovich control variates’ for the power spectrum of matter and biased tracers. In this work we present the construction of perturbative control variates in Eulerian and Lagrangian perturbation theory, and adopt the matter bispectrum as a case study. Eulerian control variates are analytically tractable for all n -point functions, but we show that their correlation with the N -body n -point function decays at a rate proportional to the sum-of-squared wavenumbers, hampering their utility. We show that the Zeldovich approximation, while possessing an analytically calculable bispectrum, is less correlated at low- k than its Eulerian counterpart. We introduce an alternative – the ‘shifted control variate’ – which can be constructed to have the correct tree-level n -point function, is Zeldovich-resummed, and in principle has an analytically tractable bispectrum. We find that applying this shifted control variate to the $z = 0.5$ matter bispectrum is equivalent to averaging over 10^4 simulations for the lowest- k triangles considered. With a single $V = 1(\text{Gpc}/h)^3$ N -body simulation, for a binning scheme with $N \approx 1400$ triangles from $k_{\min} = 0.04 h\text{Mpc}^{-1}$ to $k_{\max} = 0.47 h\text{Mpc}^{-1}$, we obtain sub-2% precision for every triangle configuration measured. This work enables the development of accurate bispectrum emulators – a probe of cosmology well-suited to simulation-based modeling – and lays the theoretical groundwork to extend control variates for the entire n -point hierarchy.

1. INTRODUCTION

The advent of large-scale galaxy surveys has delivered precise maps of the Universe across many redshifts, revealing the non-Gaussian structure of the cosmic web at high significance. While most of the cosmological information in a galaxy survey is contained within its two-point statistics, higher-order statistics are complementary summaries which aid in constraining the fundamental parameters of the Universe by nature of being sensitive to different combinations than two-point summaries (2pt Collaboration *et al.* 2024). While there are many non-Gaussian summary statistics considered, perhaps the most natural is the bispectrum – the extension of the power spectrum to three powers of the density field – which characterizes the skewness of the cosmic matter field in question (Scoccimarro 2000).

Despite its conceptual simplicity, modeling and analyzing the bispectrum is significantly more challenging than its two-point counterpart. On the modeling side, analytic tools such as perturbation theories of large-scale structure struggle to accurately reproduce the bispectrum with the exception of triangle configurations with very small wavenumbers (although see recent developments at one loop (Angulo *et al.* 2015; Philcox *et al.* 2022; D’Amico *et al.* 2024; Bakx *et al.* 2025)). On the analysis side, measuring the bispectrum in a survey at high significance is computationally challenging. Significant care must be taken to characterize observational effects such as survey geometry (Pardede *et al.* 2022; Wang *et al.* 2025), fiber collisions (Hahn *et al.* 2017; Chudaykin *et al.*

2025, in the case of spectroscopic surveys), and even a proper characterization of the covariance of the bispectrum in an ideal simulation box is not trivial (Biagetti *et al.* 2022).

These challenges in analytically modeling the bispectrum have led to significant interest in simulation-based models. N -body simulations of large-scale structure solve for the nonlinear and non-Gaussian distribution of dark matter down to substantially smaller scales than what is analytically accessible using perturbative techniques (Angulo and Hahn 2022). Measurements of the bispectrum in simulations – significantly easier than when presented with observational challenges – can be used as models for the signals measured in galaxy surveys. By performing N -body simulations at various points in cosmological parameter space, surrogate models of the bispectrum can be built which can smoothly interpolate its signals in regions where simulations were not performed (such as the BiHaloFit model of Takahashi *et al.* (2020)).

The high precision at which the bispectrum of matter and galaxies will be measured with stage-IV surveys requires that emulators of it must be highly accurate. This is difficult to achieve without averaging over many realizations at a fixed point in parameter space, hindering the efficacy of the emulation program – simulations suffer from so-called ‘cosmic variance’ due to the randomness inherent in their initial conditions. This issue with simulation-based modeling is not inherent to the bispectrum, and indeed extends itself to all summary statistics measured from simulations.

To overcome these limitations, several techniques have been proposed to suppress the variance inherent in N -

body simulations. The technique of ‘pairing and fixing’ (Pontzen *et al.* 2016; Angulo and Pontzen 2016) is successful for the power spectrum but less-so for the simulation-based bispectrum and biased-tracer power spectra (Maion *et al.* 2022). The technique of control variates, a variance reduction tool widely adopted in Statistics (Owen 2013), has seen a particular interest in the context of cosmology (Chartier *et al.* 2020). By exploiting a cheaper-yet-correlated simulation (with shared initial conditions), the method of control variates can be used to reduce the sample variance of simulation-based observables. These summary statistics measured from cheap simulations can be categorized into two classes: cheap simulations whose mean summary statistics, μ_c , are known analytically (Tassev and Zaldarriaga 2012; Kokron *et al.* 2022; DeRose *et al.* 2023; Hadzhiyska *et al.* 2023) and those where the mean has to be determined from an ensemble average (Chartier and Wandelt 2021, 2022; Ding *et al.* 2022, 2025). The latter case requires a careful study of the convergence and uncertainty on the mean, or else the total amount of variance reduction achievable is difficult to determine.

The purpose of this paper is to carry out an in-depth investigation into the former class of models, which we dub *perturbative* control variates, focusing on the application to the matter bispectrum. Both Eulerian and Lagrangian perturbation theories of structure formation can be used to create cheap-yet-correlated simulations of the late redshift Universe whose summary statistics are analytically understandable. The Eulerian approach will be shown to have the advantage of being simpler to compute analytically and a well-defined order-by-order expansion for the control variate exists. However, the Eulerian theory will pay the cost of being exponentially decorrelated after a scale Σ (eq. (13)) characterizing the average dispersion of the motions of galaxies in linear theory. We will then turn to the Lagrangian theory, where the first-order solution (Zel’Dovich 1970, also known as the Zeldovich approximation) has an analytically tractable bispectrum. We also study the bispectrum in second-order Lagrangian Perturbation Theory (2LPT), which contains the correct tree-order bispectrum but loses analytic tractability. Finally, we introduce a class of hybrid control variate builds from the ‘shifted operator’ basis of Schmittfull *et al.* (2020), and show it possesses the correct tree-level N -point function while still being analytically resumable, outperforming both models. A visual summary of the different fields considered in this work is shown in Fig. 1, smoothed on two different scales with a Gaussian filter. The scales are chosen to be larger and smaller than the smoothing scale Σ at which Eulerian fields decorrelate from the N -body density.

This paper is structured as follows: in § 2 we explore creating a control variate for the bispectrum in a toy model where the non-linear field is generated by the Zeldovich approximation. In this toy model the ‘Eulerian’ control variates are significantly simplified, and we can explain the structure of correlations between the Zeldovich density field and order-by-order Eulerian fields, as well as the correlation of the underlying bispectra. In § 3 we turn to control variates applied to the full N -body problem. We show that the Eulerian intuition developed in § 2 holds in the N -body case and study the performance of Eulerian bispectra up to fifth order in the

density field. We then turn to bispectrum control variates in LPT, studying the Zeldovich and 2LPT bispectra and how they correlate with the N -body result. We also show how to analytically match both Eulerian and the Zeldovich control variate to lattice-based realizations at high accuracy. This section concludes with an introduction of the ‘shifted control variate’, which mixes desirable properties of both Eulerian and Lagrangian schemes. In § 4 we quantify the variance reduction achieved by each control variate as a function of scale, and compare these results to some of the past literature. We conclude with summarizing remarks and some future directions to be explored in § 5.

1.1. Conventions:

All spectra measured in this work use the fiducial **Quijote** simulations (Villaescusa-Navarro *et al.* 2020), which have $N = (512)^3$ particles in boxes of size $L = 1 h^{-1} \text{Gpc}$. Bispectra are measured using **PolyBin3D** (Philcox and Flöss 2024) whose real-space bispectrum estimator is an implementation of the estimator presented in Scoccimarro *et al.* (1999); Sefusatti *et al.* (2016). The statistical properties of these bispectra come from an ensemble of the first $N = 1000$ simulations in the fiducial **Quijote** suite, using the snapshot at $z = 0.5$. There are two binning schemes adopted, in order to highlight different aspects of our control variates. The toy model of § 2.1 where we treat the Zel’dovich approximation as the fully non-linear density field uses triangles defined by bins of width $\Delta k = 2k_f \approx 0.0125 h\text{Mpc}^{-1}$ until $k_{\text{max}} = 0.15 h\text{Mpc}^{-1}$. This corresponds to $N = 236$ bins. This is called the **Eulerian binning scheme** in the text.

Analyses where the non-linear density field is the full N -body density field use a different binning scheme which has been tailored to simultaneously keep $\Delta \log k$ constant (and small), while extending to a higher k_{max} in order to capture more decorrelation, and ensuring a manageable number of triangles are included in the measurement. The binning scheme is:

- Linearly-spaced k -bins with width $\Delta k = 3k_f$ until $k = 0.3 h\text{Mpc}^{-1}$.
- Log-spaced k -bins between $0.3 \leq k/(h\text{Mpc}^{-1}) \leq 0.5$ with width $\Delta \ln k = 0.06$.

These two binning schemes are concatenated together and then resulting triangles are computed from this full vector of bins. This results in $N = 1434$ triangles. This is called the **Zeldovich binning scheme** in the text. While $N_{\text{sims}} < N_{\text{tri}}$, we note that since we only concern ourselves with diagonal uncertainties or the diagonal part of the cross-correlation matrix, we are able to measure these at reasonably high statistical significance. We do not have sufficient statistics to resolve the full structure of the covariance matrix for this binning scheme.

2. WARM-UP: BISPECTRUM CONTROL VARIATES IN A TOY MODEL

Consider the control variates problem applied to the real-space bispectrum, where we construct the random variable y

$$y(k_1, k_2, k_3) = \hat{B}^{\text{sim}} - \beta \left[\hat{B}^{\text{CV}} - \bar{B}^{\text{CV}} \right]. \quad (1)$$

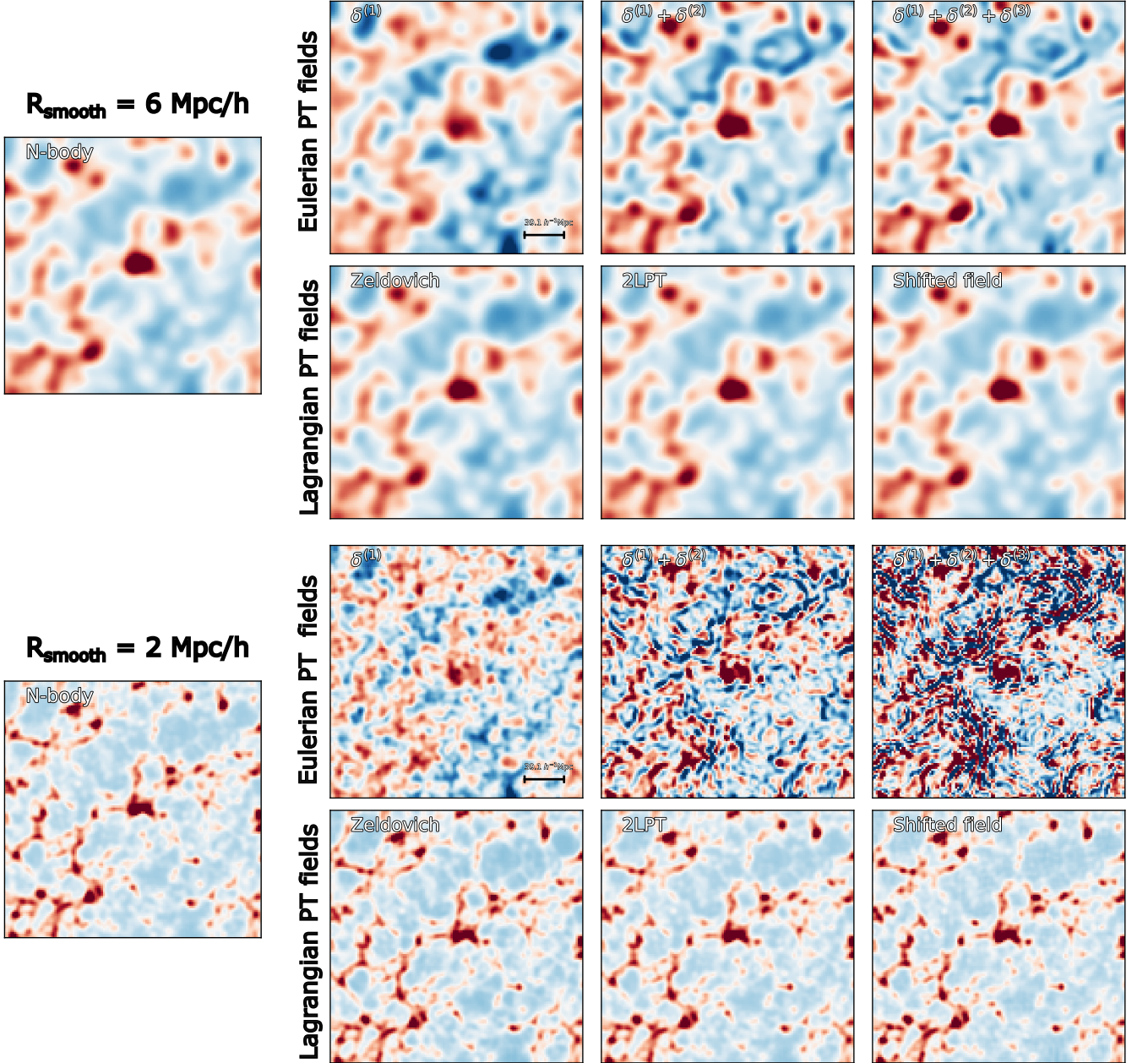


FIG. 1.— Visualization of the Eulerian and Lagrangian perturbative control variate fields considered in this work, as a function of smoothing scale. The left panels show the N -body density field centered around the largest overdensity of the simulation at $z = 0.5$, projected across $20 \text{ Mpc}/h$. When filtered on a smoothing scale with $R_{\text{smooth}} > \Sigma \sim 4.5 \text{ Mpc}/h$ (at which EPT decorrelates) the N -body, Eulerian and Lagrangian fields are all visually correlated. However, smoothing at a smaller scale reveals the rapid decorrelation of EPT, while the Lagrangian fields maintain a large degree of similarity to the N -body distribution. The color scale is chosen to be symmetric around $\pm 3\sigma_{\text{lin}}(R)$, where $\sigma_{\text{lin}}(R)$ is the standard deviation of the Gaussian panel at that smoothing scale.

β is a Lagrange multiplier, $\hat{B}^{\text{sim}}$ is the bispectrum measured in the N -body simulation, $\hat{B}^{\text{CV}}$ is the control variate bispectrum measure from the cheap-yet-correlated simulation, $\bar{B}^{\text{CV}}$ is the mean of the bispectrum, and we have suppressed the (k_1, k_2, k_3) dependence on the right-hand side. Note that $\langle y \rangle = \langle \hat{B}^{\text{sim}} \rangle$, and so y is an unbiased estimator of the simulation bispectrum. Considering the univariate problem of minimizing the triangle-by-triangle uncertainty on the bispectrum, it is known

that for the optimal β that minimizes σ_y^2 , the achieved variance suppression will be given by

$$\frac{\text{Var}[y]}{\text{Var}[\hat{B}^{\text{sim}}]} = 1 - \frac{\text{Cov}^2[\hat{B}^{\text{sim}}, \hat{B}^{\text{CV}}]}{\text{Var}[\hat{B}^{\text{sim}}]\text{Var}[\hat{B}^{\text{CV}}]} = 1 - \rho_{\text{CV}, \text{sim}}^2, \quad (2)$$

with $\rho_{\text{CV}, \text{sim}}$ the bispectrum cross-correlation coefficient. A highly correlated control variate has the potential to substantially reduce the uncertainty in simulation-based

bispectrum estimation, as has been achieved with the power spectrum. However, eq. (2) hinges on there being *no uncertainty* associated with the estimation of the mean $\bar{B}^{\text{CV}}$. In the presence of uncertainty in estimate of the mean, when N independent surrogate simulations are used to estimate $\bar{B}^{\text{CV}}$ the uncertainty will increase by

$$\sigma_y^2 \rightarrow \sigma_y^2 + \beta^2 \frac{\sigma_c^2}{N}, \quad (3)$$

Unless it holds that $\beta^2 \sigma_c^2 / N \ll \sigma_y^2 \approx \sigma_x^2 (1 - \rho_{x,c}^2)$, any potential gains from the control variate's correlation will be dwarfed by uncertainty on its mean estimate.

We can analytically estimate how large N has to be in order to not dilute the gain in precision. We define r to the ratio of variances with and without an empirically determined mean for the control variate

$$r = 1 + \frac{\beta^2 \sigma_c^2}{N \sigma_x^2 (1 - \rho_{x,c}^2)}, \quad (4)$$

where we have switched to a shorthand notation where x is $\hat{B}^{\text{sim}}$ and c is the control variate $\hat{B}^{\text{CV}}$. In the case of the optimal Lagrange multiplier, β^* ,

$$\beta^* \equiv \frac{\sigma_{xc}}{\sigma_c^2} = \rho_{xc} \frac{\sigma_x}{\sigma_c}, \quad (5)$$

and so we can recast r in a way that depends solely on ρ

$$r = 1 + \frac{\rho_{x,c}^2}{N(1 - \rho_{x,c}^2)}. \quad (6)$$

eq. (6) makes it clear what are the requirements imposed on a control variate if its mean is to be sampled empirically – the more correlated the control variate, the more realizations of it will be needed to not have the uncertainty on the mean saturate the variance reduction. The variance reduction is halved when

$$N = 1/(1 - \rho^2) - 1. \quad (7)$$

In the regime of large cross-correlation, we see that N is simply the ratio of the N -body variance to the control variate's variance – also called the ‘volume multiplier’. Empirical control variates require the uncertainty on the mean to be reduced by a factor equivalent to the volume multiplier achieved by the control variate. For a volume multiplier of 10^4 , the equivalent number of surrogate simulations and bispectra measurements would have to be carried out. Emulation suites which sample the parameter space with $\mathcal{O}(100)$ simulations would require, then, 10^6 simulation-bispectrum runs.

It is evidently desirable that a control variate have its mean be determined as precisely as possible, and the most precision one can achieve is through an analytic calculation of its signal. The suitability of different analytically calculable bispectra as a control variate is one of the main aims of this work. In the next session we will introduce an analytic toy model of bispectrum control variates where many quantities are calculable, in order to understand what to expect in the full N -body problem.

2.1. A control variate with no mean

While for the power spectrum dramatic reduction in variance can be achieved with a highly correlated observable with analytically tractable means (such as the power

spectrum in the Zeldovich approximation) (Kokron *et al.* 2022; DeRose *et al.* 2023), this is less feasible for higher N -point functions such as the bispectrum. Analytic predictions of higher N -point functions are more computationally intensive than the power spectrum: for example, the bispectrum in the Zeldovich approximation has recently been calculated by (Chen *et al.* 2024), but evaluation times for a single triangle remain on the order of 1 second. Furthermore, in order to obtain sufficiently precise predictions of these N -point functions, it is important to take into account the discreteness of Fourier modes in simulations boxes beyond the usual continuous approximations for binning. Bispectrum estimators involve averaging over many triangle configurations that fall into a bin, with an analytic estimate for the number of triangles in a bin of width Δk , given by (Sefusatti *et al.* 2010)

$$N_{123} \approx 8\pi k_1 k_2 k_3 (\Delta k)^3 \frac{V^2}{(2\pi)^6}. \quad (8)$$

The number of triangles is a rapidly growing function of k , and for volumes comparable to the fiducial simulation suites we use in this simulation, we must average over millions of configurations – this naive average is unfeasible even with rapid evaluations on the order of a second. That said, much work has been done in the direction of simplifying this bin-averaging and we will return on approximations to improve the applicability of the ZA bispectrum in § 3.2.2.

However, consider the bispectrum calculated from a linear, Gaussian density field which has seeded the full N -body simulation whose variance we wish to cancel. In this case, we can write an estimator for its *linear* bispectrum

$$\hat{B}^{111}(k_1, k_2, k_3) = \frac{1}{N_{123} V} \sum_{k_1, k_2, k_3} \delta_1 \delta_2 \delta_3 \delta_{123}^D, \quad (9)$$

where N_{123} is the number of triangles in a given bin. For any given set of initial conditions, $\hat{B}^{111}$ doesn't have to be zero. At the same time, it's immediately clear that $\langle \hat{B}^{111} \rangle = 0$. Nevertheless, this trivial bispectrum could still be a useful control variate. This is because despite possessing a zero mean, its covariance is non-zero

$$\text{Cov}[\hat{B}^{\text{sim}}, \hat{B}^{111}] \neq 0. \quad (10)$$

That this covariance is non-zero can be readily seen from considering the Gaussian disconnected contribution to the bispectrum covariance. Indeed, many analyses of the bispectrum rely on this Gaussian disconnected covariance in lieu of difficulties in estimating its full form. Thus, we expect the linear field's correlation with the N -body field to also be useful in computing a trivial control variate.

Let's elucidate the structure of this zero-mean Gaussian control variate. Consider the following toy model: take the N -body field, δ^N , to be the density field after being displaced by the Zeldovich approximation:

$$\delta^N(\mathbf{k}) = \int d^3 q e^{i\mathbf{k} \cdot \mathbf{q}} \left[e^{i\mathbf{k} \cdot \Psi^{\text{ZA}}(\mathbf{q})} - 1 \right], \quad (11)$$

where $\Psi^{\text{ZA}}(\mathbf{q})$ is the Zeldovich displacement (later defined in eq. (34)). The Gaussian disconnected part of

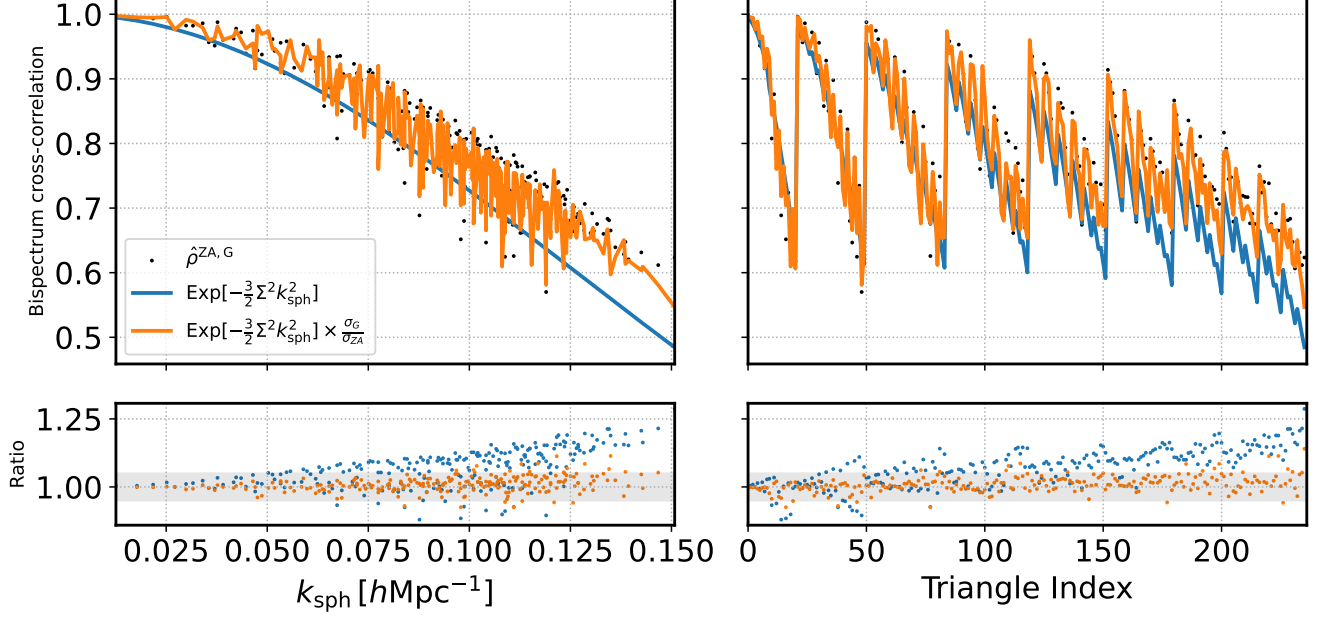


FIG. 2.— *Top*: Cross-correlation coefficient between bispectra in Zeldovich boxes and those measured from the same Gaussian initial conditions. The shape of this cross-correlation coefficient is in close agreement with the prediction from eq. (14). *Bottom*: Ratio of the empirically measured cross-correlation coefficients and the predictions from eq. (15), in blue and eq. (14), in orange. The grey bands denote $\pm 5\%$ residuals. The left and right panels show the same data, but as a function of $k_{\text{sph}} = \sqrt{(k_1^2 + k_2^2 + k_3^2)}/3$ on left and the *triangle index* representation on the right. The dominant behavior comes from the derived k_{sph} dependence. The gray band in the residuals denotes $\pm 5\%$ deviations from unity.

this covariance is given schematically¹ by

$$\begin{aligned} \text{Cov}[\hat{B}^{\text{sim}}, \hat{B}^{111}] &\propto \langle \delta_1^N \delta_2^N \delta_3^N | \delta_4 \delta_5 \delta_6 \rangle, \\ &\approx \langle \delta^N \delta \rangle \langle \delta^N \delta \rangle \langle \delta^N \delta \rangle + \dots \\ &\propto e^{-\frac{1}{2} \Sigma^2 (k_1^2 + k_2^2 + k_3^2)} P_{\text{lin}}(k_1) P_{\text{lin}}(k_2) P_{\text{lin}}(k_3), \end{aligned} \quad (12)$$

where we have used $\langle \delta^N \delta \rangle(k) = e^{-\frac{1}{2} \Sigma^2 k^2} P_{\text{lin}}(k)$ and the displacement dispersion is

$$\Sigma^2 = \frac{1}{3} \langle |\Psi(\mathbf{q})|^2 \rangle = \frac{1}{6\pi^2} \int dk P_{\text{lin}}(k). \quad (13)$$

This relation is exact in the Zeldovich approximation; it is also a good approximation in the general case given that most of the decorrelation between initial conditions and the final density field are due to the bulk linear motions (Chisari and Pontzen 2019). From the above we can write the bispectrum cross-correlation coefficient as

$$\rho(k_1, k_2, k_3) = e^{-\frac{1}{2} \Sigma^2 (k_1^2 + k_2^2 + k_3^2)} \frac{\sigma_G}{\sigma_{\text{ZA}}}, \quad (14)$$

where $\sigma_{G,\text{ZA}}$ is the standard deviation of the Gaussian / ZA bispectra respectively, for that given triangle configuration. If the Gaussian and ZA bispectra have covariances which are dominated by disconnected terms, and assuming that $P^{\text{ZA}} \approx P_{\text{lin}}$ at large scales (up to smoothing of any BAO wiggles) we find the cross-correlation coefficient between the two bispectra is given purely by the exponential damping term:

$$\rho(k_1, k_2, k_3) \approx e^{-\frac{1}{2} \Sigma^2 (k_1^2 + k_2^2 + k_3^2)}, \quad (15)$$

and thus, using the Gaussian bispectrum of an N -body simulation as a control variate should provide variance

¹ Neglecting geometric factors arising from the number of triangles or degeneracies depending on specific triangle shape.

cancellation on large scales compared to Σ . Notice that the decorrelation is exponential in the variable $k_1^2 + k_2^2 + k_3^2$. Thus, we will interchangeably show our results in terms of either *triangle index*, i , or the ‘spherical wavenumber’

$$k_{\text{sph}} \equiv \sqrt{(k_1^2 + k_2^2 + k_3^2)}/3$$

defined in Tomlinson and Jeong (2023). Any scatter at fixed k_{sph} indicates dependence on the ‘azimuthal’ and ‘polar’ angles of the spherical bispectrum ($\phi_{\text{sph}} = \arctan(k_2/k_1)$ and $\theta_{\text{sph}} = \arctan(\sqrt{k_1^2 + k_2^2}/k_3)$), which we expect to be sub-dominant when exploring the decorrelation for a control variate.

As an explicit check of this decorrelation, we calculate the bispectrum in the Gaussian initial conditions of N -body simulations, as well as from the Zeldovich-displaced field seeded by these same initial conditions, at $z = 0.5$. Computing these bispectra for $N = 1000$ boxes we can measure the cross-correlation coefficient empirically, and this is shown in Fig. 2. The curves in the figure also show the two approximations for this correlation coefficient, eq. (15) and eq. (14). The measured correlation is in close agreement with that of eq. (14), demonstrating that even a bispectrum with zero mean can serve as a control variate.

The advantage of this Gaussian control variate is that one does not have to be concerned with subtle aspects of comparing a numerically-measured bispectrum to an analytic prediction, or ensuring the uncertainty on its mean is negligible. The mean, being zero, will remain zero even when considering averaging over various triangles within a bin. While clearly convenient, the utility of this Gaussian control variate is somewhat limited. The

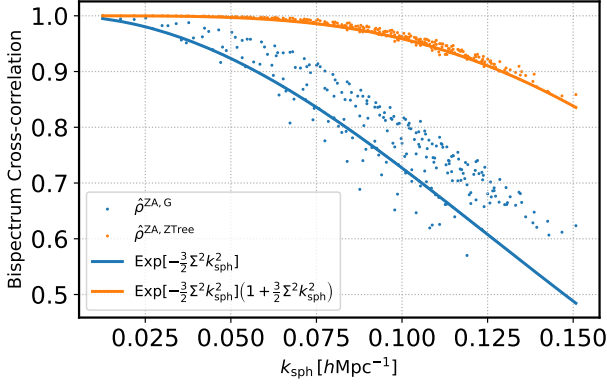


FIG. 3.— Bispectrum cross-correlation coefficient between the Zeldovich density field and Eulerian realizations, including the ‘Zeldovich tree level’ bispectrum of eq. (18). The cross-correlation coefficient is enhanced by a term $\propto k_{\text{sph}}^2$, as argued analytically. We also see a significant tightening of the cross-correlation coefficients for different triangle configurations

exponential decorrelation that affects ‘linear control variates’ (discussed in Hadzhiyska *et al.* (2023)) is even more dramatic here, since each leg of the triangle contributes its own suppression. Is it possible to do better while still retaining analytic control over the bispectrum?

2.2. Perturbative control variates in the Zeldovich approximation

In this toy model where the ‘expensive’ simulation is the Zeldovich density field, we can perturbatively expand eq. (11) to find ‘Eulerian’ PT kernels for the Zeldovich approximation (Grinstein and Wise 1987):

$$\delta^N \approx \int d^3q e^{i\mathbf{k}\cdot\mathbf{q}} \sum_n \frac{(i\mathbf{k} \cdot \boldsymbol{\Psi}(\mathbf{q}))^n}{n!} \quad (16)$$

$$\approx \sum_{n=1}^{\infty} \int_{\mathbf{k}_1 \dots \mathbf{k}_n} \delta^D \left(\mathbf{k} - \sum_{i=1}^n \mathbf{k}_i \right) \underbrace{\frac{1}{n!} \prod_{i=1}^n \frac{\mathbf{k} \cdot \mathbf{k}_i}{k_i^2}}_{\equiv Z_n(\mathbf{k}_1, \dots, \mathbf{k}_n)} \delta_{\mathbf{k}_i}, \quad (17)$$

$$\langle \delta^N \delta^{(2)} \rangle' = \int d^3q e^{i\mathbf{k}\cdot\mathbf{r}} \left\langle e^{i\mathbf{k}\cdot\boldsymbol{\Psi}(\mathbf{q})} \left[\frac{2}{3} \delta^2(\mathbf{x}) - \Psi_a(\mathbf{x}) \partial_a \delta(\mathbf{x}) + \frac{1}{2} s^2(\mathbf{x}) \right] \right\rangle_{\mathbf{r}=\mathbf{x}-\mathbf{q}}. \quad (23)$$

Each of these terms can be evaluated exactly using the cumulant expansion theorem. The spectrum is given by²

$$\langle \delta^N \delta^{(2)} \rangle' = -e^{-\frac{1}{2}k^2 \Sigma^2} \int d^3q e^{i\mathbf{k}\cdot\mathbf{r}} k_i k_j \left[\frac{2}{3} \langle \Psi_i \delta \rangle \langle \Psi_j \delta \rangle - \langle \partial_a \Psi_i \rangle \langle \Psi_a \Psi_j \rangle + \frac{1}{2} \langle \Psi_i s_{ab} \rangle \langle \Psi_j s_{ab} \rangle \right] \quad (24)$$

The Fourier transform in eq. (24) can be done to re-express the sub-spectra in the form of mode-coupling kernels in Eulerian PT; indeed, this yields the simple expression $\langle \delta^N \delta^{(2)} \rangle = \langle \delta^{(2)} \delta^{(2)} \rangle \exp(-k^2 \Sigma^2 / 2)$. Taking

² There are also contributions at zero-lag for individual correlators but they cancel when all terms are included.

The corresponding tree-level bispectrum is

$$B_{\text{ZA}}^{211}(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) = \frac{\mathbf{k}_1 \cdot \mathbf{k}_2}{k_2^2} \frac{\mathbf{k}_1 \cdot \mathbf{k}_3}{k_3^2} P(k_2) P(k_3) + \text{cyclic}, \quad (18)$$

where $\mathbf{k}_1 = -(\mathbf{k}_2 + \mathbf{k}_3)$. The disconnected covariance of the tree-level bispectrum with the Zeldovich spectrum will involve contractions

$$\text{Cov}[\hat{B}^{\text{sim}}, \hat{B}^{211}] \sim \langle \delta_1^N \delta_2^N \delta_3^N | \delta_1^{(2)} \delta_2 \delta_3 \rangle \quad (19)$$

$$\sim \langle \delta^N \delta^{(2)} \rangle \langle \delta^N \delta \rangle \langle \delta^N \delta \rangle + \dots \quad (20)$$

The inclusion of the tree-level bispectrum (beyond the Gaussian term) now has exponential suppression in two of the legs as well as a new correlator – of the Zeldovich field with its second-order Eulerian counterpart, $\delta^{(2)}$. We can calculate this correlator exactly and see its effect on the cross-correlation. To see this, note that the second-order ‘Eulerian Zeldovich’ density field, which is given by

$$\delta^{(2)}(\mathbf{x}) = \frac{1}{2} \partial_{x,i} \partial_{x,j} [\Psi_i(\mathbf{x}) \Psi_j(\mathbf{x})], \quad (21)$$

can be re-written as

$$\delta^{(2)}(\mathbf{x}) = \frac{2}{3} \delta_0^2(\mathbf{x}) - \Psi_i(\mathbf{x}) \partial_i \delta_0(\mathbf{x}) + \frac{1}{2} s_0^2(\mathbf{x}), \quad (22)$$

where s_0^2 is $s_{0,ij} s_{0,ij}$ and $s_{0,ij} = (\partial^{-2} \partial_i \partial_j - \delta_{ij}/3) \delta_0$ is linear tidal field tensor. Note that the initial fields and their derivatives are evaluated in Eulerian coordinates above. This expression possesses the same degrees of freedom as the standard $F_2(\mathbf{k})$ kernel in Eulerian perturbation theory, but with a different weighting of the δ_0^2 and s_0^2 coefficients. We write the $\langle \delta^N \delta^{(2)} \rangle$ power spectrum as a correlator in LPT

the asymptotic limit of P_{22} yields the IR contribution

$$\langle \delta^N \delta^{(2)} \rangle'(k) \supset k^2 \Sigma^2 e^{-\frac{1}{2}k^2 \Sigma^2} P_{\text{lin}}(k). \quad (25)$$

This IR contribution is partially canceled by the IR contribution to the damping exponent in $\langle NL^{(1)} \rangle$, leading to an overall IR contribution to the second-order field equal

to

$$(1 + k^2 \Sigma^2) e^{-\frac{1}{2} k^2 \Sigma^2} P_{\text{lin}}(k) \approx \left(1 + \frac{1}{2} k^2 \Sigma^2\right) P_{\text{lin}}(k) \quad (26)$$

at large scales. This large-scale enhancement of the cross correlation is in turn canceled by the IR contributions to P_{22} in the denominator when computing the correlation between the two fields, leading to damping due to bulk displacements beginning only at quartic order in k . In other words, while the cross-covariance between the fully nonlinear and quadratic fields are artificially enhanced by long-wavelength modes, their cross-correlation is suppressed by an asymptotically equally large enhancement in the variance of the quadratic field by those modes. Computing the matter field up to cubic order doesn't further reduce the leading-order k^2 decorrelation—since it is already zeroed in the quadratic cross correlation—but increases the correlation towards smaller scales due to better matching the mode coupling. Regardless, while the leading IR contributions in the form of polynomials of $k\Sigma$ cancel the decorrelation due to bulk modes to some extent, they are eventually overcome by the exponential in e.g. Equation 25. We thus see that including the tree-level control variate improves the correlation, but does not cancel the exponential decay at $k \gtrsim \Sigma^{-1}$. The interested reader is referred to Appendix A for more explicit calculations of the correlation coefficients and comparisons to simulations.

In Fig. 3 we show the cross-correlation coefficient between the full Zeldovich bispectrum and the ‘gaussian + tree-level’ subset. We also plot the expected increase in correlation derived in eq. (25). The inclusion of the tree-level bispectrum boosts the cross-correlation coefficient from $\sim 60\%$ at $k_{\text{sph}} = k_{\text{max}} = 0.15 h\text{Mpc}^{-1}$, to $\sim 85\%$, and the scale-dependence of the cross-correlation reflects the leading-order cancellation of the IR contribution derived above. We also observe the correlation coefficient is a very tight function of k_{sph} when this additional term is

included – variations in covariance from non-equilateral triangles are significantly more well-captured with the inclusion of the tree level spectrum.

Given the success of the tree-level Zeldovich bispectrum in this simplified toy model, we are motivated to investigate the performance of perturbative control variates in ‘real-world’ applications where the N -body bispectrum is the object whose variance we wish to cancel.

3. PERTURBATIVE CONTROL VARIATES AND N-BODY SIMULATIONS

In the previous section we showed, in an analytic toy model, that while Gaussian control variates (also referred to as ‘linear control variates’ in the context of the power spectrum) can be beneficial, their variance suppression in a polyspectrum decays as an exponential in the sum-of-squared wavenumbers. However, we also saw that Eulerian perturbation theory provides an order-by-order method to construct a control variate which improves variance suppression to smaller scales. In this section we shall go beyond the Zeldovich approximation and use bispectra measured from N -body simulations as the object of interest.

We will begin with a discussion of bispectra measured from Eulerian PT on the lattice, up to bispectra of order $\mathcal{O}(\delta^5)$. In § 3.2 we turn to bispectra in Lagrangian perturbation theory, where we evaluate the Zeldovich and 2LPT bispectra. This includes a discussion on how to compare grid-based and analytic Zeldovich bispectra to high accuracy. § 3.3 introduces *shifted control variates*, which use the shifted operator basis of Schmittfull *et al.* (2020) to construct a control variate which combines optimal characteristics of Eulerian and Lagrangian theory.

3.1. Eulerian Control Variates

In Eulerian perturbation theory, a recursion relation can be used to generate the n -th order Eulerian density and velocity fields. This recursion relation is given in matrix form by (Bernardeau *et al.* 2002; Taruya *et al.* 2018)

$$\begin{pmatrix} \delta_n(\mathbf{x}) \\ \theta_n(\mathbf{x}) \end{pmatrix} = \frac{2}{(2n+3)(n-1)} \begin{pmatrix} n+1/2 & 1 \\ 3/2 & n \end{pmatrix} \sum_{m=1}^{n-1} \begin{pmatrix} (\nabla \delta_m) \cdot \mathbf{u}_{n-m} + \delta_m \theta_{n-m} \\ [\partial_j (\mathbf{u}_m)_k] [\partial_k (\mathbf{u}_{n-m})_k] + \mathbf{u}_m \cdot (\nabla \theta_{n-m}) \end{pmatrix}. \quad (27)$$

Given a linear density field $\delta_{\text{lin}}(\mathbf{x}) = \theta_{\text{lin}}(\mathbf{x})$ that seeds an N -body simulation, evaluating this recursion relation produces field-level realizations of Eulerian PT.

3.1.1. Tree-level Eulerian control variate

The second-order Eulerian density field can be expressed as

$$\delta^{(2)}(\mathbf{x}) = \frac{17}{21} \delta^2(\mathbf{x}) - \Psi(\mathbf{x}) \cdot \nabla \delta(\mathbf{x}) + \frac{2}{7} s^2(\mathbf{x}). \quad (28)$$

This density field has the same analytic structure as the second-order ZA density field in eq. (21), with different coefficients for the growth and tidal contributions. We then measure the bispectrum of the full second-order density field

$$\delta(\mathbf{x}) = \delta^{\text{lin}}(\mathbf{x}) + \delta^{(2)}(\mathbf{x}). \quad (29)$$

The bispectrum of the density field in eq. (29) is not an ideal control variate – it contains a one-loop contribution from $B^{222} = \langle \delta_1^{(2)} \delta_2^{(2)} \delta_3^{(2)} \rangle$ – which is difficult to evaluate analytically. To isolate diagrams order-by-order, we multiply the second-order density field by a counting parameter ϵ , $\delta^{(2)} \rightarrow \epsilon \delta^{(2)}$. A linear combination of bispectra from the Gaussian field and from this new field can be used to extract the tree-level, $\mathcal{O}(\epsilon)$, piece. The precise procedure, including its extensions to higher order diagrams, is detailed in Appendix B.

The tree-level bispectrum, as measured in simulations, must also be computed to high accuracy in order to ensure the mean is unbiased. The leading theoretical uncertainty that has to be managed in order to achieve agreement between the tree level bispectrum measured on the lattice, and its analytic prediction, is to correct

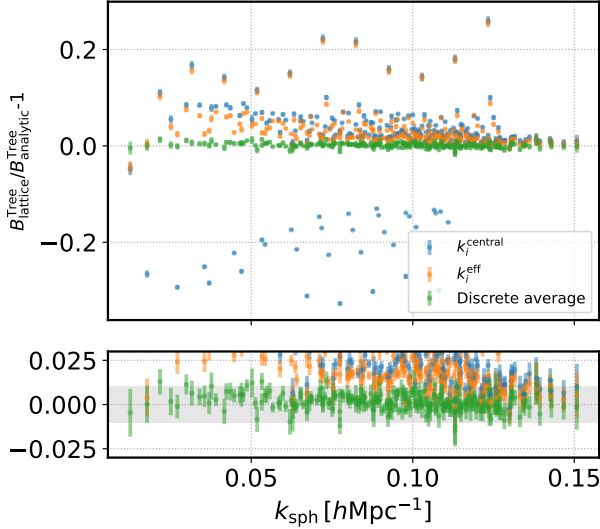


FIG. 4.— *Top panel:* Relative deviation of tree level bispectrum averaged over $N = 1000$ lattice-based realizations of the $\delta^{(2)}$ field compared to analytic predictions. The blue points show the tree level bispectrum when evaluated at the central (k_1, k_2, k_3) of the bin, the orange curve shows the prediction when using the ‘effective triangle’ in eq. (31) and the green points show the result from averaging over all discrete triangles. The error induced due to not bin-averaging can reach 30% for triangles in this scheme where $\Delta k = 2k_f$. *Bottom panel:* The same panel, but zoomed to show the size of spread when the discrete bin average is made. The gray bands indicate 1% scatter around zero. Error bars correspond to error bars on the mean of $N = 1000$ realizations.

for the effect of the bin size that contributes to triangle estimation. In principle, for a given bin, the bispectrum measured in that bin is given by

$$B(k_1, k_2, k_3) = \frac{1}{N_{123}} \sum_{\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3 \in \text{bin}} B^{\text{tree}}(k'_1, k'_2, k'_3) \delta_K(\mathbf{k}'_{123}), \quad (30)$$

and N_{123} is the number of triangles that contribute to a bin, which grows rapidly with wavenumber (c.f. eq. (8)). In the Eulerian binning scheme, the equilateral bin with central value at $k_{\text{equi}} \approx 0.151 h\text{Mpc}^{-1}$ and width $\Delta k = 2k_f \approx 0.0123 h\text{Mpc}^{-1}$ contains $N_{123} \approx 8.6 \times 10^6$ triangles. While for the tree level spectrum the discrete triangle average can be evaluated, alternatives have been proposed such as approximating the sum as a continuous integral and applying ‘discreteness weights’ (Ivanov *et al.* 2022), or evaluating the bispectrum at an ‘effective triangle’ (Sefusatti *et al.* 2010)

$$k_i^{\text{eff}} = \langle |\mathbf{k}_i| \rangle_{\mathbf{k}_i \in \text{bin}}, \quad (31)$$

where the average is taken over all triangle configurations that land in a bin. In Fig. 4 we show the relative deviation between the tree-level bispectrum computed on the lattice and the analytic predictions from different averaging schemes. Evaluating the bispectrum at the center of the bin incurs large errors reaching 30%. The effective wavenumber prescription reduces the bias in the calculation, but we find it is still well above the statistical uncertainties of the measurement. Performing the full discrete average brings the uncertainties to sub-1%, within the statistical errors on the mean bispectra for nearly all points. While averaging over discrete triangles produces accurate results, at $\sim 10^7$ triangles in a bin it

is clearly infeasible for all but tree-level predictions. We will return to this point in § 3.2.2, but for now consider that we can suitably predict the mean control variate for the tree-level Eulerian case.

3.1.2. Higher-order contributions with no mean

Given the second-order density field, we can also construct the combination B^{221} , which is 5th-order in the density. Being comprised of an odd number of density fields, the B^{221} bispectrum has zero mean. Nevertheless, its addition should improve the performance of the Eulerian control variate. In order to include it, we should include all diagrams that enter at 5th order, which includes the B^{311} bispectrum. In the ϵ expansion described in Appendix B, these terms contribute at $O(\epsilon^2)$. An appealing aspect of including these ϵ^2 terms is that, since their mean is zero, their inclusion is ‘free’ from the perspective of building a control variate whose mean is known analytically. Since we previously established that the binned tree level bispectrum can, in principle, be calculated accurately, it is worth investigating potential gains from including the ϵ^2 bispectrum. We expect these zero-mean diagrams to improve the cross-correlation because they possess non-zero covariances with the N -body field

$$\begin{aligned} \langle \delta^N \delta^N \delta^N | \delta^{(2)} \delta^{(2)} \delta \rangle &\sim \langle \delta^N \delta^{(2)} \rangle \langle \delta^N \delta^{(2)} \rangle \langle \delta^N \delta \rangle, \\ \langle \delta^N \delta^N \delta^N | \delta^{(3)} \delta \delta \rangle &\sim \langle \delta^N \delta^{(3)} \rangle \langle \delta^N \delta \rangle \langle \delta^N \delta \rangle, \end{aligned}$$

while having zero expectation value.

To compute the $\delta^{(3)}$ contribution to the density field we return to the recursive algorithm for EPT on the grid from Ref. (Taruya *et al.* 2018), previously shown in eq. (27). In order to control for the effect of aliasing, we follow the prescription of Ref. (Taruya *et al.* 2018) and apply a spherical Fourier top-hat filter with $k_{\text{cut}} = 4/3 h\text{Mpc}^{-1}$ to the second order solutions, before they are used in the recursion relation to compute terms cubic in δ . Since the cubic field will only contribute in a mean-zero form to the control variate we do not have to concern ourselves with the subtleties of implementing this filtering analytically, but extensions of Eulerian control variates to the one loop bispectrum (including up to the diagram B^{411}) would.

The left panel of Fig. 5 shows the ratio between the N -body bispectrum and the three forms of Eulerian bispectra we have considered so far – Gaussian, $O(\epsilon)$, and $O(\epsilon^2)$. The Gaussian bispectrum clearly averages to zero, and including the ϵ^2 corrections imperceptibly alters the ratio of means, as expected from analytic considerations and the assumption of a Gaussian initial field. The right panel of Fig. 5 shows the cross-correlation coefficient between the N -body bispectrum and the three Eulerian control variates discussed in this section. Confirming our intuition developed for the Zeldovich approximation, we see that the inclusion of the tree-level bispectrum on top of the standard Gaussian contribution significantly increases the cross-correlation coefficient with the bispectrum measured from an N -body simulation. Displaying the correlation coefficients as a function of k_{sph} also shows that the including the tree-level bispectrum *tightens* the scatter in the other triangle coordinates. At $k_{\text{sph}} \sim 0.15 h^{-1}\text{Mpc}$ the correlation coefficient is 80% for the tree level control variate, 90% for the B^{ϵ^2} and has dropped to near 40% for the strictly Gaussian case. The

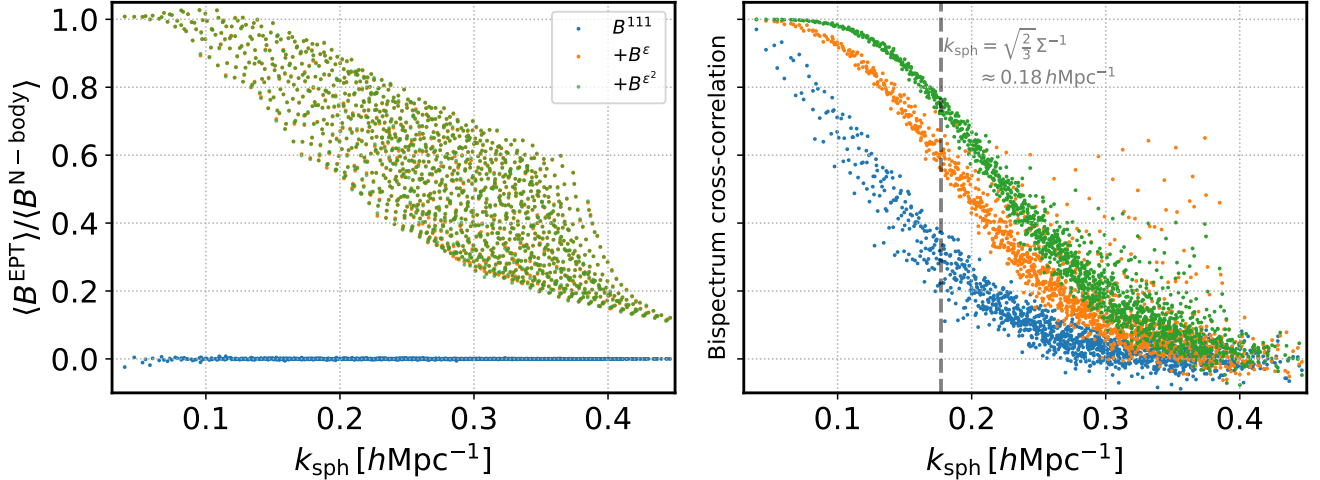


FIG. 5.— *Left*: Ratio of bispectra in Eulerian perturbation theory, to order $\mathcal{O}(\epsilon^2)$ divided by the N -body bispectrum, as a function of $k_{\text{sph}}^2 = (k_1^2 + k_2^2 + k_3^2)/3$. Each bispectrum is averaged over $N = 1000$ realizations. The blue points indicate the Gaussian bispectrum, orange points include the tree-level contribution, and the green points include terms up to the bispectrum diagrams B^{221} and B^{311} , which have zero mean. *Right*: Cross-correlation coefficients between the full N -body bispectrum and different realizations of the bispectrum measured in lattice Eulerian perturbation theory. Despite their zero-mean, the Gaussian term, as well as the fifth-order-in- δ terms, both contribute to increase the cross-correlation coefficient with the N -body bispectrum, and thus the total variance reduction achievable with EPT. The dashed vertical line in the right panel corresponds to the k_{sph} mode at which one e-folding of decorrelation is expected, according to eq. (15).

inclusion of the fifth-order bispectrum extends the range of triangles over which appreciable cross-correlation is observed.

Despite the improvements from pushing to higher order in Eulerian PT, there exists an unavoidable exponential decay in the cross-correlation coefficient stemming from the absence of large-scale displacements in Eulerian theory.

3.2. Lagrangian Control Variates – Zeldovich Approximation and 2LPT

The previous sections have shown that Eulerian control variates provide a systematic form of producing correlated surrogates of N -body summary statistics. However, there is also the immediately clear limitation of an exponential decorrelation, as a function of k_{sph} , between the surrogate and the non-linear bispectrum. The reason for this decorrelation is tied to lack of large-scale bulk flows in EPT. The Zeldovich approximation includes these large-scale bulk flows and this drives the success of the ‘Zeldovich control variates’ technique introduced in Kokron *et al.* (2022); DeRose *et al.* (2023). We now turn to a study of the bispectrum in lattice realizations of LPT, and its suitability as a control variate compared to EPT.

Matter density fields in Lagrangian perturbation theory can be generated by displacing particles, located at a position $\mathbf{q}$, by their Lagrangian displacement $\Psi(\mathbf{q})$ to late-time positions $\mathbf{x} = \mathbf{q} + \Psi(\mathbf{q})$. From mass conservation, the late-time density contrast field in LPT is given exactly by the relation

$$1 + \delta(\mathbf{x}) = \int d^3q \delta^D(\mathbf{x} - \mathbf{q} - \Psi(\mathbf{q})). \quad (32)$$

While N -body simulations solve for this displacement exactly, Lagrangian Perturbation Theory provides series

solutions to this displacement of the form

$$\Psi(\mathbf{q}) \approx \Psi^{\text{ZA}}(\mathbf{q}) + \Psi^{2\text{LPT}}(\mathbf{q}) + \dots, \quad (33)$$

where the Fourier-space representations of the Zeldovich and 2LPT displacements are

$$\begin{aligned} \Psi_{\mathbf{k}}^{\text{ZA}} &= \frac{i\mathbf{k}}{k^2} \delta_{\mathbf{k}}, \\ \Psi_{\mathbf{k}}^{2\text{LPT}} &= \int_{\mathbf{k}_1, \mathbf{k}_2} \delta^D(\mathbf{k} - \mathbf{k}_{12}) \frac{i\mathbf{k}}{k^2} \frac{3}{14} \left[1 - \frac{(\mathbf{k}_1 \cdot \mathbf{k}_2)^2}{k_1^2 k_2^2} \right] \delta_{\mathbf{k}_1} \delta_{\mathbf{k}_2}, \end{aligned} \quad (34)$$

and the $\Psi^{n\text{LPT}}$ kernels can be generated to all orders using known recursion relations, as in EPT (Matsubara 2015).

The LPT density fields are generated by displacing particles sampled from the same pre-initial condition grid and linear density field as their corresponding Quijote box. The displaced particles are deposited using cloud-in-cell deposition and the grid is corrected for this smoothing (Sefusatti *et al.* 2016). We generate three sets of Lagrangian displacements from which we compute their bispectra:

- Zeldovich displacements sampled from the initial density field smoothed by an explicit Gaussian cut-off $e^{-(k/k_{\text{cut}})^2}$ with $k_{\text{cut}} = 0.5 \text{ hMpc}^{-1}$. These will always be referred to as the *damped ZA* sample.
- Zeldovich displacements sampled from the full initial conditions without damping.
- 2LPT displacements sampled from the full undamped initial conditions.

Damping the initial power spectrum before sampling displacements will be important to match the analytic calculation of observables in ZA with the lattice representation, which we discuss shortly. This damping was also

required in the case of the power spectrum (Kokron *et al.* 2022).

3.2.1. Empirical LPT bispectra

We measure LPT bispectra for the three sets of LPT displacements discussed previously. Figure 6 is the Lagrangian analog to Fig. 5: the ratio of LPT bispectra relative to the N -body bispectrum is shown in the left panel, while the right panel shows the corresponding cross-correlation coefficients as a function of k_{sph} .

As expected from the discussion in § 2.2, the tree-level bispectrum in the Zel’dovich approximation differs from the EPT prediction. As a result, even at low k_{sph} the ZA bispectra disagree from the N -body bispectra, by up to 40%. Damping the initial conditions only has a mild effect on the resulting spectra. Turning to the cross-correlation coefficient between the N -body bispectrum and the Zel’dovich bispectrum, we observe two interesting trends. That the two bispectra disagree at very low k_{sph} implies that the cross-correlation coefficient does not asymptote to $\rho_{x,c} \rightarrow 1$. At arbitrarily low- k the Zel’dovich bispectrum should not yield as much variance reduction as the EPT bispectrum. However, by virtue of including large-scale bulk flows we also observe that for the smallest-scale triangles the cross-correlation coefficient is still on the order of 50%, whereas the Eulerian bispectra have fully decorrelated. While the damped ZA bispectrum is a worse fit to the N -body amplitude, we also observe that the cross-correlation coefficient at high- k is slightly larger for the damped simulations. At low redshifts, past shell-crossing, the Zel’dovich approximation is known to over-predict the displacement of particles compared to the N -body result. By smoothing the initial conditions, the displacements are slightly smaller in magnitude, and this results in a smaller overshoot and better cross-correlation with the N -body bispectrum.

That the damped ZA bispectrum is more correlated with the N -body result is interesting: it points to the possibility of engineering a filtering scale for the linear

density field which maximizes the ZA correlation with the N -body result, without paying a price in calculating the mean. However, this optimization is probably quantitatively minor and we leave the investigation of how to determine the optimal filtering scale to future work.

Turning to the 2LPT bispectrum, Fig. 6 shows a better agreement with the N -body result, especially for the lowest k -binned triangles, but a similarly quick disagreement as k_{sph} increases. The 2LPT result is only 20% of the N -body amplitude for the smallest scales considered. Unlike for the Zel’dovich approximation, the cross-correlation coefficient between the 2LPT bispectrum and the N -body bispectrum asymptotes to 1 at large scales, and decays similarly but with a larger amplitude for all triangles. This is attributed to the presence of large-scale displacements, while containing the tree-level bispectrum.

Quantitatively, the 2LPT bispectrum has $(1 - \rho^2)^{-1} \sim 10^4$ for the lowest k_{sph} bins. Since the full 2LPT bispectrum cannot at present be easily computed analytically this also implies, from eq. (7), that at least 10^4 2LPT boxes at similar volume would be needed to ensure a sufficiently precise determination of the mean.

3.2.2. The analytic Zel’dovich bispectrum

The large number of simulations needed to measure the 2LPT bispectrum to sufficient precision for it to be a suitable control variate motivate studying the Zel’dovich bispectrum further. Despite not fully correlating with the matter bispectrum at low- k , the decorrelation is slower as a function of k_{sph} . Another key advantage of the Zel’dovich bispectrum is that it is analytically calculable. Simultaneously employing an Eulerian and Zel’dovich bispectrum as control variates (and extending the problem to two Lagrange multipliers $\beta_{\epsilon^2}, \beta_{\text{ZA}}$, for example) would allow for a set of control variates which are correlated over all scales and analytically calculable, for any triangle configuration.

In Chen *et al.* (2024) it was shown that the Zel’dovich matter bispectrum can be written in a closed form

$$B_{\text{ZA}}(\mathbf{k}_1, \mathbf{k}_2) = \int_{\mathbf{q}, \mathbf{r}} e^{-i\mathbf{k}_1 \cdot \mathbf{q} - i\mathbf{k}_2 \cdot \mathbf{r}} \exp \left[\frac{1}{2} (k_{1,i} k_{3,j} A_{ij}(\mathbf{q}) + k_{2,i} k_{3,j} A_{ij}(\mathbf{r}) + k_{1,i} k_{2,j} A_{ij}(\mathbf{q} - \mathbf{r})) \right], \quad (35)$$

where $A_{ij}(\mathbf{q})$ is the standard Zel’dovich pairwise displacement correlator

$$A_{ij}(\mathbf{q}) = 2 \int_{\mathbf{k}} (1 - e^{i\mathbf{k} \cdot \mathbf{q}}) \left(\frac{k_i k_j}{k_4} \right) P_{\text{lin}}(k).$$

eq. (35) can be cast into a convolution of three scalar functions

$$\mathcal{E}(\mathbf{k}_i, \mathbf{k}_j, \mathbf{p}) \equiv \int_{\mathbf{q}} e^{-i\mathbf{p} \cdot \mathbf{q}} e^{\frac{1}{2} \mathbf{k}_i \cdot \mathbf{a} \mathbf{k}_j \cdot \mathbf{b} A_{ab}(\mathbf{q})}$$

which can be efficiently numerically evaluated using fast Fourier transforms.³ The scalar $\mathcal{E}(\mathbf{k}_i, \mathbf{k}_j, \mathbf{p})$ functions couple long- and short-wavelength modes to arbitrary order, in addition to the coupling implied by their convolution; by damping the linear spectrum, we are more

immune to the effects of specific regularization schemes and can thus more easily match analytic predictions to the lattice-based realizations.

The main challenge with using the analytic Zel’dovich bispectrum, then, is to compute the bin-averaged bispectrum that is measured in N -body simulations. As discussed in the case of the tree-level Eulerian bispectrum in § 3.1.1, evaluating the bispectrum over either the central triangle or the effective triangle incurs unacceptably large errors between the empirical measurement and its analytic prediction. In the case of the tree-level bispectrum we were able to compute the average over all triangles that contributed to the bin (although note this becomes significantly more difficult for the Zel’dovich binning scheme which extends to $k_{\text{sph}}^{\text{max}} \approx 0.5 h \text{Mpc}^{-1}$). In the case of the Zel’dovich bispectrum this is no longer feasible. While being substantially faster than direct in-

³ This implementation is publicly available in the python package *triceratops*.

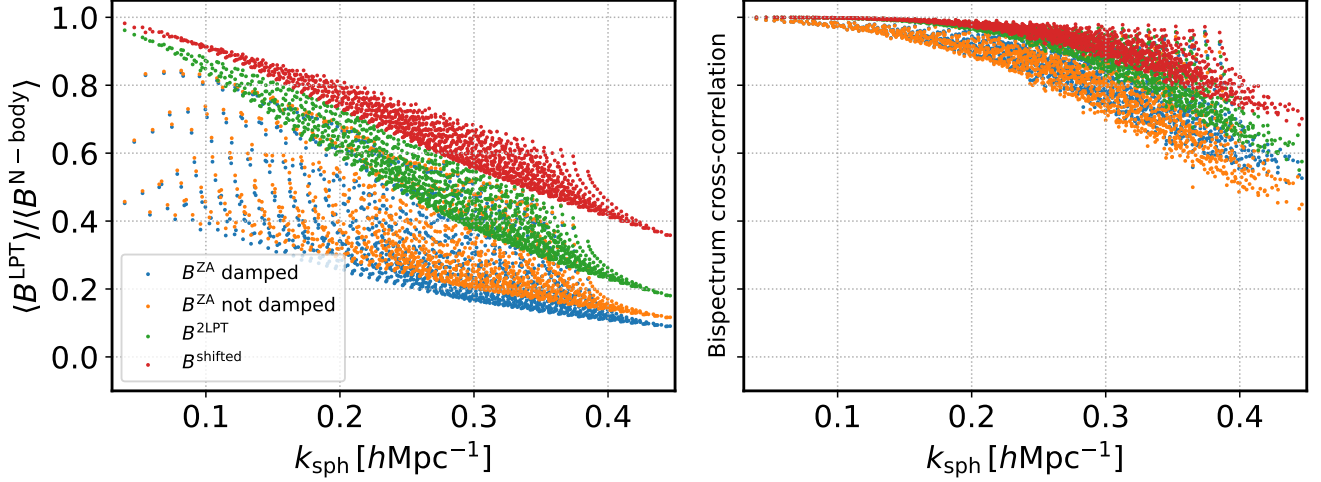


FIG. 6.— *Left*: Ratio of bispectra in Lagrangian perturbation theory, to 2nd order in displacements, divided by the N -body bispectrum. Blue points correspond to the Zeldovich bispectrum, orange points to the Zeldovich bispectrum of undamped initial conditions, and the green points to the 2LPT bispectrum. Red points correspond to the bispectrum of the ‘shifted control variate’ introduced in § 3.3. *Right*: Cross-correlation coefficients between the full N -body bispectrum and different realizations of the bispectrum from lattice Lagrangian perturbation theory. We observe a high degree of cross-correlation across all triangles, especially when contrasted to the EPT case. However, note the Zeldovich bispectrum cross-correlation deviates slightly from $\rho = 1$ at low- k_{sph} . Despite the higher degree of correlation it is also clear the LPT bispectrum is a worse approximation to the N -body bispectrum than the tree-level prediction.

tegration, discrete evaluation over large numbers of k modes is still prohibitively expensive for the FFT-based predictions of the Zeldovich bispectrum, which takes on the order of ~ 1 second on modern computers.

To compute the Zeldovich bispectrum at sufficient accuracy, we devise weights to correct for binning and discreteness effects. If the Zeldovich bispectrum is similar in amplitude to a bispectrum that can be exactly bin-averaged, the effective triangle prediction can be re-weighted by the deviation of this second bispectrum. For example, assuming we assess triangles where $B_{\text{Zel}} \approx B_{\text{tree}}$ (note that ‘tree’ denotes the Zeldovich tree-level bispectrum of eq. (18)), we can write the bin-average of the Zeldovich bispectrum as

$$\begin{aligned} \langle B_{\text{Zel}} \rangle &= \left(\frac{\langle B_{\text{Zel}} \rangle}{B_{\text{Zel}}^{\text{eff}}} \right) B_{\text{Zel}}^{\text{eff}} \\ &= \left(\frac{\langle B_{\text{tree}} \rangle + \langle \Delta B \rangle}{B_{\text{tree}}^{\text{eff}} + \Delta B^{\text{eff}}} \right) B_{\text{Zel}}^{\text{eff}}, \quad \Delta B = B_{\text{Zel}} - B_{\text{tree}} \\ &= \left(\frac{\langle B_{\text{tree}} \rangle}{B_{\text{tree}}^{\text{eff}}} \right) B_{\text{Zel}}^{\text{eff}} \left(1 + \frac{\langle \Delta B \rangle - \Delta B^{\text{eff}}}{\langle B_{\text{tree}} \rangle} + \dots \right) \\ &= \left(\frac{\langle B_{\text{tree}} \rangle}{B_{\text{tree}}^{\text{eff}}} \right) B_{\text{Zel}}^{\text{eff}} \left(1 + \mathcal{O}(\Delta \ln k)^2 \mathcal{O}(P) \right), \end{aligned}$$

where we’ve used that $\langle B_{\text{tree}} \rangle / B_{\text{tree}}^{\text{eff}} - 1 = \mathcal{O}(\Delta \ln k^2)$ in the third line, since the effective wavenumber approximation cancels any errors linear in the relative bin widths.

The relative deviation between the bin-average and the effective triangle approximation, for the calculable tree-level case, can be used to define weights

$$w(k_1, k_2, k_3) = \langle B_{\text{tree}} \rangle / B_{\text{tree}}^{\text{eff}} \quad (36)$$

that correct for binning effects, similar to the discreteness weights employed in Ivanov *et al.* (2022) except that in this case the nonlinear structure of the bispectrum is known with no free coefficients. Corrections to this

weighting scheme are suppressed by the closeness of the tree-level and fully nonlinear calculations – governed by the size of loop corrections of order $\mathcal{O}(P)$ – along with the narrowness of the bin $\Delta \ln k$. In practice, we use undamped linear power spectra to compute these weights rather than the damped versions used in the simulations, finding that this leads to slightly better performance. This is due to the damping dominating the scale dependence on small scales, where mode coupling contributes significantly to B_{Zel} .

Figure 7 shows the bispectrum measured from lattice realizations of the Zeldovich density field compared to analytic predictions employing various approximations, using the Eulerian binning scheme. The fully nonlinear prediction computed using *triceratops* with tree-level binning weights in Equation 36 (green) are in excellent agreement with the simulated bispectra, with a combined χ^2 across all measurements – assuming a diagonal covariance measured from 1000 simulations – very close to 1 per bin. In comparison, the Zeldovich tree level predictions of eq. (18), shown in orange, agree with the fully nonlinear prediction at very large scales ($k < 0.05 h \text{ Mpc}^{-1}$) but rise to roughly 10% by $k < 0.10 h \text{ Mpc}^{-1}$. The unweighted but fully nonlinear predictions shown in blue are consistently different from the green points at the 10% level on all scales, reaching nearly 50% discrepancies for the most squeezed triangles where the longest leg has $k_1 = 2k_f$. These results demonstrate that the tree level weights are sufficient to restore concordance with lattice measurements without performing the costly averages over triangle configurations.

As previously argued, an advantage of using Zeldovich control variates compared to Eulerian ones is their ability to remain correlated until smaller scales, or larger k . At these larger k it is computationally more efficient to evaluate bispectra over wider bins; however, for sufficiently wide k bins, even computing the tree-level prediction summing over discrete k modes is computa-

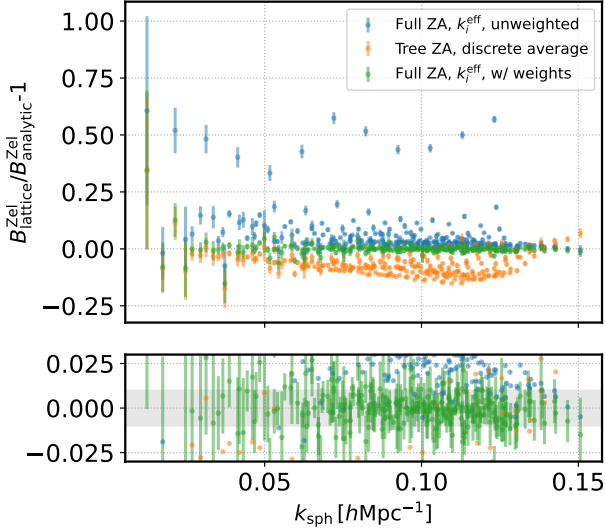


FIG. 7.— Analytic predictions for the Zeldovich bispectrum, including the impact of the weights in eq. (36), compared to lattice-based realizations of the Zeldovich density field. The weighted fully nonlinear predictions from *triceratops* (green) agree well with the simulations on all scales shown. The tree-level Zeldovich predictions only agree at $k \lesssim 0.05 h \text{ Mpc}^{-1}$ (orange). The effective triangle approximation is inadequate for the full Zeldovich bispectrum (blue), which disagrees at $\sim 10\%$ for nearly all triangles and reaches $\sim 50\%$ for squeezed triangles.

tionally intensive, making it nontrivial to compute the tree-level weights in Equation 36. For a very wide bin, it is unlikely that any individual triangle configuration significantly impacts the results of the bin average and a Monte-Carlo sampling of valid triangles should provide a good approximation to the full average over discrete pairs.

Assuming that the bispectrum is reasonably smooth within a triangle bin, the number of discrete pairs required to achieve an accuracy ϵ is roughly $N_{\text{toler}} = \epsilon^{-2}$. We can therefore choose to evaluate the bispectrum a total number of $N = \min(N_{\text{toler}}, N_{123})$ times.⁴ We thus downsample the number of wavevectors in bins $k_1 \pm \Delta k, k_2 \pm \Delta k$ via

$$N_{1,2}^{\text{downsample}} = \sqrt{f_{\text{mc}} N_1 N_2}, \quad f_{\text{mc}} = \frac{N}{N_{\text{tri}}}, \quad (37)$$

where N_i is the original numbers of wavevectors in that bin. As in the un-downsampled case, not all of the $N_1^{\text{downsample}} \times N_2^{\text{downsample}}$ triangles will produce a triangle that falls into the given bispectrum bin—rejecting these results in a total of $N_{\text{tri}}^{\text{downsample}}$ bispectrum configurations. The estimate of the binned bispectrum is the average value of the bispectrum evaluated at the remaining $\approx f_{\text{mc}} N_{\text{tri}}$ points. An alternative scheme is to instead downsample pairs $(\mathbf{k}_1, \mathbf{k}_2)$ that satisfy the geometric constraint of the bin ($|\mathbf{k}_1 + \mathbf{k}_2| \approx k_3$) by a factor of f_{mc} ; we have checked that this method returns a very comparable degree of accuracy but is significantly more memory-intensive to run efficiently due to having to sample the product space of wavevectors rather than the wavevectors themselves.

Figure 8 shows predictions for the Zeldovich bis-

⁴ In practice we estimate N_{123} using the continuous approximation in eq. (8).

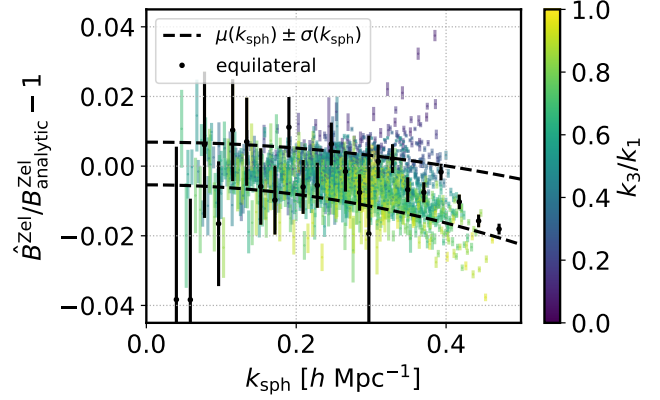


FIG. 8.— Similar to Figure 7 except for the wider Zeldovich binning scheme, where the tree-level weights have to be computed by Monte-Carlo. The predictions come with a theoretical error slightly smaller than one percent, shown here as a confidence interval dependent on k_{sph} in the black dashed lines, as is expected given the bin width and the slope of the linear power spectrum; however, the theoretical errors are quite dependent on triangle configuration (colorbar), with the squeezed configurations most deviant while equilateral triangles (black) are statistically consistent at all but the smallest scales shown.

pectrum computed using *triceratops* and tree-level weights in the wider Zeldovich binning scheme. In contrast to the finer binning scheme used in Figure 7, the tree-level weights in this case have been computed by Monte-Carlo, as the bins contain too many triangles to evaluate the weights. Unlike in the previous binning, the analytic predictions in this case differ from the simulations in a statistically significant way. The dashed lines in Fig. 8 show the band spanned by 1σ around a quadratic fit to the mean relative error, assuming it only depends on k_{sph} .⁵ The differences are consistent with a roughly 0.7% theoretical error on our analytic predictions, though the theoretical error at low $k_{\text{sph}} \lesssim 0.2 h \text{ Mpc}^{-1}$ is subdominant to statistical uncertainties.

This is inline with expectations that the error is of order $(\Delta \ln k)^2$, taking into account the slope of the linear power spectrum. This isotropic summary of the theoretical error only partially captures the story, however. The deviations are very strongly dependent on the orientation (shown here using k_3/k_1 as a proxy), with squeezed triangles as clear outliers in their deviation. In contrast, equilateral bins, highlighted in the black points, remain statistically consistent between the simulations and predictions until $k_{\text{sph}} \approx 0.3 h \text{ Mpc}^{-1}$, where the effects of grid-level smoothing also becomes rather significant.

One more numerically intensive but potentially useful solution is to construct interpolation tables of the Zeldovich bispectrum computed by *triceratops* as a func-

⁵ In particular, we maximize the log-likelihood

$$\ln L = \sum_i -\frac{1}{2} \frac{(\Delta_i - \mu_{\text{th}}(k_{\text{sph},i}))^2}{\sigma_{\text{th}}^2(k_{\text{sph},i}) + \sigma_i^2} - \frac{1}{2} \ln(\sigma_{\text{th}}^2(k_{\text{sph},i}) + \sigma_i^2) \quad (38)$$

where the relative error in each bispectrum bin is $\Delta_i = (B_{\text{pred},i} - \hat{B}_i)/\hat{B}_i$ and σ_i is its standard deviation, assuming that $\mu_{\text{th}}, \sigma_{\text{th}}$ are polynomials of a given order in k_{sph} . We estimate the theoretical and total (theoretical and statistical) uncertainty as a function of k_{sph} by fitting this likelihood with and without the addition of σ_i^2 , finding that the theoretical error reaches about 50% of the total variance at $k_{\text{sph}} \approx 0.2 h \text{ Mpc}^{-1}$ fairly independently of polynomial order.

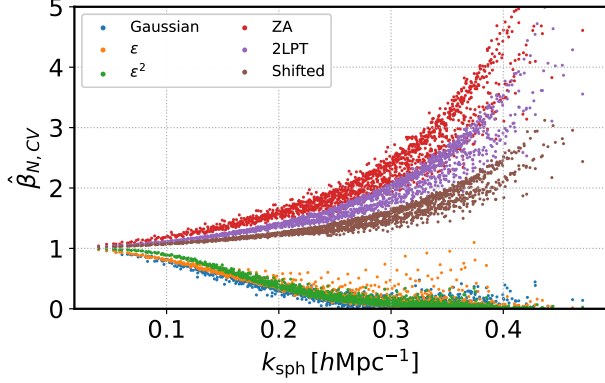


FIG. 9.— Estimates of the Lagrange multiplier, β , from the estimators of eq. (42) for the various control variates considered in this work. The solid lines show the median and 95% quantiles along bins of k_{sph} .

tion of k_1 , k_2/k_1 and k_3/k_1 . The latter two are dimensionless quantities with ranges $[0, 1]$, making them more intuitive targets for interpolation.

Nevertheless, we have shown that the Zeldovich bispectrum can in principle be evaluated analytically at the accuracy required for it to be a successful control variate, especially for large-scale triangles where the bispectrum variance is significant.

$$\delta^{CV,(2)}(\mathbf{k}) = \int d^3q e^{-i\mathbf{k}\cdot\mathbf{q}} e^{-i\mathbf{k}\cdot\mathbf{\Psi}^{\text{ZA}}(\mathbf{q})} [1 + c_1\delta(\mathbf{q}) + c_2\delta^2(\mathbf{q}) + c_s s^2(\mathbf{q})]. \quad (39)$$

Expanding linearly in the Zeldovich displacement, the Eulerian version of this field is

$$\delta^{CV,(2)}(\mathbf{x}) \approx (1 + c_1)\delta(\mathbf{x}) + \left(\frac{2}{3} + c_1 + c_2\right)\delta^2(\mathbf{x}) - (1 + c_1)\mathbf{\Psi} \cdot \nabla\delta + \left(\frac{1}{2} + c_s\right)s^2(\mathbf{x}) + O(\delta^3; c_1, c_2, c_s), \quad (40)$$

where the $2/3$ and $1/2$ terms come from the Eulerian expansion of the Zel'dovich kernel, eq. (22). Setting $\{c_1, c_2, c_s\} = \{0, 1/7, -3/14\}$ results in a surrogate density field which is in agreement with the second order Eulerian PT solution. Its bispectrum at tree level is given by the Eulerian tree level bispectrum⁶. However, the field is fundamentally Lagrangian and so it will not pay the price of exponential decorrelation.

An advantage of constructing an operator in this fashion is that since only Zeldovich displacements are exponentiated, their bispectra can in principle be calculated analytically to all orders. The expressions will involve integrals similar to eq. (35)⁷

$$\int_{\mathbf{q}, \mathbf{r}} e^{-i\mathbf{k}_1\cdot\mathbf{q} - i\mathbf{k}_2\cdot\mathbf{r}} \left\langle \mathcal{O}_i(\mathbf{q}_1) \mathcal{O}_j(\mathbf{q}_2) \mathcal{O}_k(\mathbf{q}_3) e^{-i\mathbf{k}_1\cdot\mathbf{\Psi}(\mathbf{q}_1) - \mathbf{k}_2\cdot\mathbf{\Psi}(\mathbf{q}_2) + i(\mathbf{k}_1 + \mathbf{k}_2)\cdot\mathbf{\Psi}(\mathbf{q}_3)} \right\rangle, \quad (41)$$

which can be evaluated using the cumulant expansion theorem. The resulting expressions for the basis fields of the LPT bispectrum can be found in Appendix E of Chen *et al.* (2024). While explicitly written down, the efficient numerical implementation of these bispectra is still an open task. We will thus explore the applicability of this hybrid scheme, dubbed ‘shifted control variates’, by numerically estimating their bispectra. A calculation of the mean of the control variate to high accuracy is only needed when computing the unbiased average. An assess-

⁶ Note that we could also re-write the ‘1+’ term in eq. (39) a sum of shifted operators with well-defined coefficients, which will change the value of the $\{c_i\}$ that recover the correct tree-level expression. The resulting expression will be different at 3rd order compared to using the ‘1+’ term.

⁷ The expression in eq. (35) is equivalent to eq. (41) if we set all $\mathcal{O}_i(\mathbf{q}) = 1$.

3.3. Shifted control variates

Previous sections considered bispectrum control variates in Eulerian and Lagrangian theory. Eulerian control variates captured the tree-level correlation but decayed exponentially due to their lack of resummed displacements. The Zeldovich bispectrum maintained a high correlation with the N -body bispectrum but lacked the correct low- k form and would not optimally cancel variance if used as a control variate. Is there a way to engineer a control variate which only resums linear displacements, but possesses the correct tree-level bispectrum (or N -point function more generally)?

The ‘shifted operator’ scheme for perturbation theory of Schmittfull *et al.* (2019) can be used to this end. Shifted operators keep only the Zeldovich displacement exponentiated and in Fourier-space, the shifted operator of a field $\mathcal{O}(\mathbf{q})$ is defined as

$$\tilde{\mathcal{O}}(\mathbf{k}) = \int_{\mathbf{q}} e^{-i\mathbf{k}\cdot(\mathbf{q} + \mathbf{\Psi}^{\text{ZA}}(\mathbf{q}))} \mathcal{O}(\mathbf{q}).$$

Contributions from higher-order displacements can be expanded from the exponential, as they are small compared to the Zeldovich displacement. The tree-level bispectrum contains the quadratic bias fields δ^2 and s^2 . Consider then, instead of the Eulerian $\delta^{(2)}$ field, using the bispectrum measured from a Lagrangian ‘biased tracer’

ment of the total amount of variance reduction achieved by a control variate does not require this. Thus, we will study the shifted control variate to motivate future development of calculations of the analytic Zeldovich bispectrum for biased tracers. The left and right panels of fig. 6 show the shifted control variate compared to the other Lagrangian models assessed in this work. We clearly see that not only is the shifted control variate the closest approximation to the N -body bispectrum, it is also the most correlated and decays at a slower rate than even the 2LPT bispectrum. For the highest k_{sph} triangles the shifted control variate is found to be $\sim 70\%$ correlated while the 2LPT coefficient is near $\sim 60\%$.

4. RESULTS AND DISCUSSION

Having characterized our ability to analytically estimate the binned perturbative bispectrum, as well as the correlation coefficient of several Eulerian and Lagrangian control variate candidates, we turn to the study of the variance reduction offered by each candidate.

4.1. Estimates of β

We estimate the Lagrange multiplier, β , for each control variate from our set of $N_{\text{sims}} = 1000$ simulations. In contrast with the methodology developed in Kokron *et al.* (2022), we will estimate the Lagrange multiplier only in the univariate approximation

$$\hat{\beta}_{N,\text{CV}}(k_1, k_2, k_3) = \frac{\text{diag}(\text{Cov}(\hat{B}_N, \hat{B}_{\text{CV}}))}{\text{diag}(\text{Var}(\hat{B}_{\text{CV}}))}, \quad (42)$$

for the full set of triangles. We compare the performance of the univariate estimator with the full multivariate estimator

$$\hat{\beta}_{N,\text{CV}}(k_1, k_2, k_3) = (\Sigma_{N,\text{CV}} \cdot \Sigma_{\text{CV}}^+), \quad (43)$$

in Appendix C, as well as the diagonal approximation of Kokron *et al.* (2022). In this appendix we also show that the observed ‘damping’ of β discussed in Kokron *et al.* (2022) was spuriously driven by the under-determination of the covariance matrix in the univariate problem, and usage of the pseudoinverse to invert the control variate covariance.

We report the measurements of the Lagrange multiplier, for the different classes of control variates considered, in Fig. 9, as a function of k_{sph} . For the Eulerian control variates we observe a monotonic and steep decrease for β as a function of k_{sph} . This can be understood from the fact that any cross-covariance between an Eulerian control variate and an N -body summary statistic has to decay as $\sim \exp(-\Sigma^2 k_{123}^2)$ (as discussed in § 2.1). Since the Lagrange multiplier can be similarly expressed as

$$\beta = \rho_{x,c} \frac{\sigma_x}{\sigma_c}, \quad (44)$$

for an Eulerian control variate we expect exponential decorrelation. For the Lagrangian control variates shown in the right panel of Fig. 6, the cross-correlation coefficient has not fallen off as aggressively, reaching $\rho \sim 0.5$ for the smallest-scale triangles considered. The left panel of Fig. 6 shows that $B^{\text{N-body}} \approx 10B^{\text{ZA}}$ at those scales, and assuming that the bispectrum covariance at these scales is dominated by the BB term⁸

$$\frac{\sigma(B^{\text{N-body}})}{\sigma(B^{\text{ZA}})} \sim \frac{B^{\text{N-body}}}{B^{\text{ZA}}} \approx 10,$$

then we expect $\beta \sim 5$ which is close to the observed value for the ZA spectrum. In concordance with the rough scalings presented here, we also expect to find a lower value of β for the 2LPT bispectrum (since the degree of correlation is similar but $B^{2\text{LPT}} \sim 2B^{\text{ZA}}$), and we see this in Fig. 9. The shifted control variate’s bispectrum has a Lagrange multiplier closest to unity for all triangles, compared to other methods.

⁸ This is a very approximate scaling, but in Fig. 11 we see that for the highest- k triangles the disconnected ‘PPP’ contribution underestimates the covariance by a significant amount.

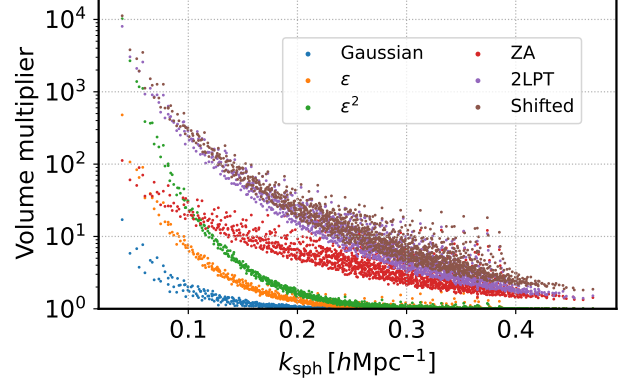


FIG. 10.— Effective ‘volume multiplier’ from applying the various bispectrum control variates considered in this work, after combining with the optimal β . The shifted operator and 2LPT control variates perform the best across most triangle configurations.

4.2. Variance reduction

With β in hand, we construct variance-reduced estimates of the bispectrum for each of the control variates considered in this work, for the $N = 1000$ simulations in the Quijote ensemble. We measure the resulting variance from this ensemble, and compare it to the variance of the N -body bispectrum. In the Fig. 10 we report the ‘volume multiplier’

$$\text{Volume multiplier}(k_1, k_2, k_3) \equiv \left(\frac{\sigma_N^2}{\sigma_{\text{CV}}^2} \right) (k_1, k_2, k_3),$$

which corresponds to the effective number of simulations needed to be averaged over to achieve the same variance. For all triangle configurations with $k_{\text{sph}} \lesssim 0.25 h\text{Mpc}^{-1}$, the shifted operator and 2LPT control variates achieve a 10-fold increase in volume, nearing a 10^4 -fold increase for the largest-scale triangle configurations we have considered. For the Eulerian control variates, the $\mathcal{O}(\epsilon^2)$ control variate achieves comparable volume increase to the shifted operator at the largest scales, but this increase decays and by $k_{\text{sph}} = 0.2 h\text{Mpc}^{-1}$ the improvement is negligible.

We present another view of the results of applying different control variates to the bispectrum in Fig. 11. Focusing on equilateral triangles (although the trends observed hold for all triangles), we show the precision with which the bispectrum is measured for several different forms of control variate. We also show, for comparison, the ‘true’ precision obtained using the full nonlinear covariance and the precision inferred by assuming an approximate form for the covariance. The approximation we make is to use the ‘nonlinear disconnected covariance’. Specifically, for these equilateral triangle configurations we measure the bispectrum covariance from the Gaussian simulations with matched initial conditions. The covariance in this Gaussian disconnected case is

$$\hat{\sigma}_{\text{gauss}}^2(B) \sim \frac{1}{N_k} P_{\text{lin}}^3(k_{\text{equi}}),$$

where N_k is the number of k -modes that enter that equilateral triangle bin. We compute the nonlinear disconnected covariance (also called the ‘PPP’ covariance) by rescaling the empirical estimate of the Gaussian bispec-

trum covariance with the ratio of power spectra

$$\hat{\sigma}_{\text{gauss}}^2(B) \rightarrow \hat{\sigma}_{\text{gauss}}^2(B) \times \left(\frac{P_{\text{nonlin}}}{P_{\text{lin}}} \right)^3 (k_{\text{equi}}). \quad (45)$$

Any difference between the full covariance and eq. (45), then, will arise from the importance of terms such as the bispectrum-bispectrum, power spectrum-trispectrum, and pentaspectrum terms in the covariance. This comparison reveals an interesting feature of numerical bispectrum estimation: even for the relatively broad bins considered in this work, the precision to which the bispectrum is measured at scales of $k_{\text{equi}} \gtrsim 0.4 \, h\text{Mpc}^{-1}$ saturates at a level of 4% for our volume of $V = 1(\text{Gpc}/h)^3$. Using the nonlinear ‘PPP’ covariance would lead to $O(1)$ error on the precision with which the bispectrum is measured.

Turning to the performance of the different control variates we have considered, we find that employing the Gaussian control variate does not significantly increase the precision of the measurement – there is some improvement for bins with $k_{\text{equi}} \leq 0.1 \, h\text{Mpc}^{-1}$ but the measurement uncertainties are still around 15%. The best performing Eulerian control variate, the bispectrum to $O(\epsilon^2)$, achieves below 1% for the first equilateral bin, but rapidly decorrelates. By $k_{\text{equi}} \sim 0.2 \, h\text{Mpc}^{-1}$ the measurement uncertainties are comparable to the N -body values. Finally, we see that the shifted control variate leads to sub-1% measurements of the bispectrum for the longest triangles, and better precision than what can be achieved at the smallest scales. As the shifted bispectrum decorrelates from the N -body result we approach the N -body precision but there is still an improvement even for the highest- k triangle. We see, thus, that the shifted control variate can essentially eliminate large-scale sample variance as a concern for bispectrum estimation, achieving sub 2-% precision across all triangle bins in a single $V = 1(\text{Gpc}/h)^3$ box. Simulation-based modeling of the bispectrum is now gated by the precision to which small-scale triangle configurations can be measured.

4.3. Comparison to past results

Having assessed the performance of our different perturbative control variates, we compare them to past applications of variance reduction in the published literature.

The first paper to consider the performance of a control variate in the bispectrum was [Chartier et al. \(2020\)](#). Two sets of triangles are considered – ‘squeezed isosceles triangles’ with $k_1 = k_2$ from $k_3/k_1 \in [0.025, 0.2]$, and a set of equilateral triangles with $k_{\text{max}} = 0.75 \, h\text{Mpc}^{-1}$. Their analysis uses as a surrogate a cheaper N -body simulation from COLA ([Tassev and Zaldarriaga 2012](#)), and they average their control variate estimators over five N -body – COLA pairs to report their variance reduction. The effective volume multiplier they report for isosceles triangles is between $40\times$ and $\sim 5\times$ in this scenario, when considering a single pair of simulations. Since this squeezed isosceles analysis bins all triangles together, it is not clear how to map these volume reductions, but we note that a volume multiplier of 5 is achieved for $k_{\text{sph}} \sim 0.3 \, h\text{Mpc}^{-1}$ for our shifted control variate. Their equilateral numbers are more translatable – focusing on their equilateral triangle with $k_{\text{min}} \sim 0.04 \, h\text{Mpc}^{-1}$ they report a per-pair

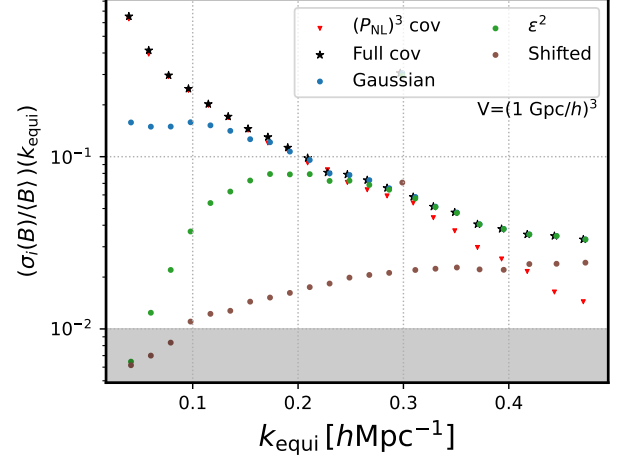


FIG. 11.— Effective precision of the bispectrum for equilateral triangles, for a $1 \, (\text{Gpc}/h)^3$ box, after applying the Gaussian, best-performing Eulerian, and best-performing Lagrangian control variates. The red triangles denote the bispectrum precision inferred when assuming the ‘PPP’ disconnected term of the covariance only, and the black stars show the precision inferred from using the full numerical covariance, without applying any control variate. There is a numerical artifact in the measurement for the point slightly below $k_{\text{equi}} = 0.3 \, h\text{Mpc}^{-1}$, caused by the heterogeneous binning scheme we have adopted. No other triangle is affected.

volume multiplier of $200\times$, and their equilateral triangle with $k_{\text{sph}} \approx 0.5 \, h\text{Mpc}^{-1}$ has a per-pair volume multiplier of 1.4. The shifted control variate has a volume multiplier of ~ 1.8 for the $k_{\text{sph}} = 0.47 \, h\text{Mpc}^{-1}$ equilateral triangle of our Lagrangian binning scheme. Thus, we find performance comparable to the original investigation of [Chartier et al. \(2020\)](#) with a control variate that is analytically tractable.

The other work which has investigated control variates for the bispectrum is [Ding et al. \(2022\)](#), which investigates the use of the FastPM ([Feng et al. 2016](#)) as a control variate. Additionally, [Ding et al. \(2022\)](#) investigate the case of the halo bispectrum which is not immediately comparable to the analysis we have carried out in this work. Nevertheless, we note that their analysis corresponds to triangles with $k_1 = 0.1 \, h\text{Mpc}^{-1}$, $k_2 = 0.2 \, h\text{Mpc}^{-1}$ and an angle of $\theta = [0, \pi]$, which are triangles with $k_{\text{sph}} \in [0.14, 0.21] \, h\text{Mpc}^{-1}$. At these scales they note no improvement from pairing and fixing, and an improvement in volume that corresponds to a per-pair volume multiplier of around $20\times$, under the assumption of no uncertainty in μ_c . It is also interesting that their volume multiplier is somewhat flat across the triangles they consider. Across the equivalent range of scales we find the volume multiplier sharply varies from $100\times$ at $k_{\text{sph}} \sim 0.14 \, h\text{Mpc}^{-1}$ to $20\times$ at $k_{\text{sph}} \sim 0.21 \, h\text{Mpc}^{-1}$. It could be that the presence of shot noise in the case of the halo bispectrum has set an upper bound to the volume multiplier that can be obtained, but we leave an explicit comparison to halo bispectra for future work.

While no past work has investigated variance reduction for the bispectrum as we have here, for cases where a comparison is possible we find the shifted control variate performs as well as approximate N -body solvers investigated in past works at a fraction of the computational cost.

5. CONCLUSIONS

We have investigated the problem of reduction of variance in the empirical, simulation-based bispectrum using control variates inspired by Eulerian and Lagrangian perturbation theory. We used the Zeldovich approximation as a toy model for the full nonlinear problem, where the exact solution is known and the corresponding ‘Eulerian perturbation theory’ is highly simplified, to investigate the structure of exponential decorrelation for any Eulerian control variate as well as how perturbative corrections restore some of the correlation structure. Eulerian control variates always decorrelate as $\sim \exp[-\Sigma^2 k_{123}^2]$, with second and third order corrections serving to eliminate the leading-order suppression of the cross-correlation coefficient. Additionally, we showed that bispectra of mean-zero can still be useful control variates, suggesting that including diagrams which are naively zero in the Eulerian case can be beneficial since there is a non-zero covariance between the N -body and Gaussian case.

We then turned to a study of Eulerian and Lagrangian control variates in the non-linear problem, where our Eulerian intuition developed in the preceding toy model holds. We find the Zeldovich approximation is not an optimal control variate in the case of the bispectrum – unlike the power spectrum – and show this is due to the differing tree-level structure of the bispectrum in ZA and in EPT. Still, the Zeldovich bispectrum possesses the advantage of being significantly correlated with the N -body case out to small scales, as in the case of the power spectrum. The 2LPT bispectrum is shown to be more optimally correlated at the cost of not being analytically calculable. We introduce the ‘shifted control variate’ as an optimal solution – being in principle analytically calculable while simultaneously possessing maximal correlation with the N -body case. Indeed, the shifted control variate is shown to outperform all other perturbative control variates in this work. The use of a single N -body / shifted control variate pair is shown to reduce the variance of N -body simulations by factors ranging from 10^4 to 1.8 for triangles between $k_{\text{sph}} = [0.036, 0.48] h\text{Mpc}^{-1}$. Equivalently, with a single $V = 1(\text{Gpc}/h)^3$ box we can measure the matter bispectrum at $z = 0.5$ to sub-2% precision for all triangle configurations in question. This implies that accurate simulation-based bispectrum emulators can be devised, extending the conclusions of past work on the power spectrum to this domain.

There exist clear directions to continue this work. We have focused on the case of the matter bispectrum, the scales of $k \lesssim 0.5 h\text{Mpc}^{-1}$ are currently being probed by galaxy surveys and it is of great interest to extend this technique to the bispectrum of biased tracers. We note that the methodology laid out here is fully sufficient – the biased tracer bispectrum is given by eq. (41) and

is a mild generalization of the shifted operator control variate. The remaining work would be to select samples representative of the clustering samples of upcoming surveys such as DESI, Euclid and Rubin to study their bispectrum signatures, as has been investigated (using non-perturbative control variates and HODs) in [Ding et al. \(2025\)](#). The other direction would be to extend this work to redshift space. The extension of Zeldovich control variates to redshift space is known for the power spectrum ([DeRose et al. 2023](#)). As PolyBin3D possesses the functionality to measure bispectrum multipoles, extending the empirical results of this paper to redshift space is readily achievable. The biggest challenge comes in estimating the mean: while the Zeldovich matter bispectrum can be computed in an identical way to the real space one, as shown in [Chen et al. \(2024\)](#), in order to capture nonlinearities beyond the Zeldovich approximation one would either have to formulate redshift space lattice-based EPT and match its bispectrum to high accuracy (as in [Taruya et al. \(2021a\)](#)), or extend the LPT bispectrum to redshift space including quadratic nonlinearities in order to produce analytic predictions of the mean of the redshift space bispectrum control variate.

Finally, the construction of the shifted control variate points to the possibility of engineering a control variate whose n -point is correct at tree level. The extension of shifted control variates to the trispectrum could have implications in the problem of covariance matrix estimation for galaxy surveys. We plan to return to this topic, and the others mentioned above, in future work.

ACKNOWLEDGMENTS

The authors thank Matias Zaldarriaga for helpful discussions. NK acknowledges support from the Bershadsky Fund and the Fund for Natural Sciences of the Institute for Advanced Study. SC acknowledges support from the National Science Foundation at the IAS through NSF/PHY 2207583. Support for this work was provided by NASA through the NASA Hubble Fellowship grant #HST-HF2-51572.001 awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS5-26555. This work was performed in part at the Aspen Center for Physics, which is supported by National Science Foundation grant PHY-2210452 and a grant from the Simons Foundation (1161654, Troyer). SC thanks the Galileo Galilei Institute for Theoretical Physics for the hospitality and the INFN for partial support during the completion of this work. Calculations and figures in this work have been made using `nbodykit` ([Hand et al. 2018](#)) and the SciPy Stack ([Harris et al. 2020](#); [Virtanen et al. 2020](#); [Hunter 2007](#)). This research has made use of NASA’s Astrophysics Data System and the arXiv preprint server.

APPENDIX

A. CROSS CORRELATION OF NONLINEAR AND EULERIAN FIELDS: PERTURBATION THEORY AND IR-ENHANCED DIAGRAMS

In this appendix we will explore the cross correlation of the nonlinear Zeldovich and matter density fields with the predictions of Eulerian perturbation theory (EPT) on the lattice, in particular to understand the correlation coefficient as a function of perturbative order and the role of long-wavelength (IR) displacements.

Let us begin with the cross correlation of nonlinear and Eulerian fields in the Zeldovich approximation. We want to consider the case where an arbitrarily nonlinear vertex of the fully nonlinear Zeldovich field, Z_{m+2n} , is cross-correlated

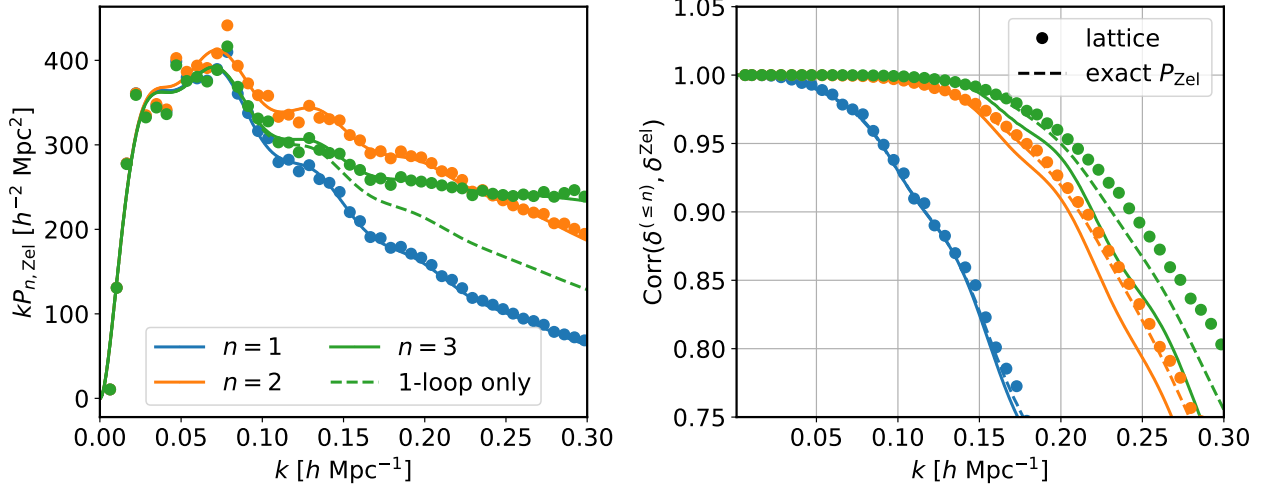


FIG. A1.— Cross-correlation of the nonlinear and Eulerian Zeldovich density fields on the lattice (dots) compared to analytic predictions. (Left) The cross power spectrum between the Eulerian field summed to n^{th} order, computed to 1-loop in perturbation theory, with large-IR contributions at higher loops resummed via an exponential. For the third-order cross correlation there is also an IR-enhanced 2-loop diagram, and we show the prediction without it as a dashed line. (Right) The correlation coefficient between the two fields. The analytic predictions break down around $k \sim 0.15h \text{Mpc}^{-1}$ reflecting the onset of 2-loop mode coupling. For $n = 1, 2$ the resummed expressions at 1-loop are exact except for the autospectrum of the nonlinear Zeldovich field in the operator, such that swapping out the 1-loop expression for a fully nonlinear one (dashed) is enough to achieve good agreement to small scales.

to a lower order Eulerian vertex Z_m . In this case all the unpaired vertices must be paired amongst themselves, leading to an integral

$$\int_{\mathbf{p}} Z_{m+2n}(\mathbf{k}_1, \dots, \mathbf{k}_m, \mathbf{p}_1, -\mathbf{p}_1, \dots, \mathbf{p}_n, -\mathbf{p}_n) P_{\text{lin}}(p_1) \dots P_{\text{lin}}(p_n) \sim Z_m(\mathbf{k}_1, \dots, \mathbf{k}_m) \left(\int_{\mathbf{p}} \frac{(\mathbf{k} \cdot \mathbf{p})^2}{p^4} P_{\text{lin}}(p) \right)^n \supset e^{-\frac{1}{2}k^2\Sigma^2} Z_m(\mathbf{k}_1, \dots, \mathbf{k}_m) \quad (\text{A1})$$

where the final expression follows from summing up all n and accounting for combinatorial factors for the number of pairing possibilities out of the original $m + 2n$ momenta. This implies that the correlation of the nonlinear and Eulerian Zeldovich fields at each order in the latter can be computed exactly by enumerating diagrams excluding bubbles ($k^2\Sigma^2$) in the former, then resumming the bubbles as an exponential. We therefore have that, for example, $\langle \delta^N \delta^{(1)} \rangle' = P_{\text{lin}}(k) e^{-\frac{1}{2}k^2\Sigma^2}$ and $\langle \delta^N \delta^{(2)} \rangle' = P_{22}(k) e^{-\frac{1}{2}k^2\Sigma^2}$, as also derived in Equation 24.

In order to accurately predict cross correlations of this sort, as well as the autocorrelations of the Eulerian fields, it will be important to account for the large parameter $k\Sigma \gtrsim 1$ beyond the loop order considered. This is, for example, why we have isolated out the exponential in the paragraph above, since expanding its argument to a given order would lead to known higher-order corrections that are significantly larger than naive expectations. Many of these terms are already included in the full 1-loop calculation, e.g. $P_{22} \supset (k^2\Sigma^2)P_{\text{lin}}$ and $P_{13} \supset -\frac{1}{2}(k^2\Sigma^2)P_{\text{lin}}$, where they in addition cancel when properly combined. However, when considering also the cubic lattice EPT prediction, we also generate a subset of 2-loop diagrams whose IR contributions do not cancel, and are thus artificially enhanced. This comes from P_{33} , which is given by

$$P_{33}^{\text{Zel,IR}}(k) = \frac{3}{4}(k^2\Sigma^2)^2 P_{\text{lin}}(k) + (k^2\Sigma^2) P_{1\text{-loop}}^{\text{Zel}}(k) \quad (\text{A2})$$

while the contribution to the cross correlation gives the total spectrum

$$P_{3,\text{Zel}}^{\text{IR-enhanced}}(k) = \left(P_{13}^{\text{Zel}}(k) + \frac{1}{2}(k^2\Sigma^2)^2 P_{\text{lin}}(k) + (k^2\Sigma^2) P_{1\text{-loop}}^{\text{Zel}}(k) \right) e^{-\frac{1}{2}k^2\Sigma^2} \quad (\text{A3})$$

where we note that the latter has a 13 contribution due to the linear term in the Zeldovich field. In fact, the 2-loop terms in these two expressions are equal, since $P_{13}^{\text{Zel}} = -\frac{1}{2}k^2\Sigma^2 P_{\text{lin}}$. The left panel in Figure A1 shows this cross-spectrum compared to the measured cross-spectra in simulations. These are in excellent agreement, though for the third-order Eulerian field the IR-enhanced 2-loop diagram is critical in establishing this agreement, which we emphasize is achieved without any free parameters.

We are now in a position to better understand the cross correlation of the fully nonlinear and perturbative fields order-by-order. These cross-correlation coefficients are given by

$$r_{n,\text{Zel}}(k) = \frac{\langle \delta^N | \sum_i^n \delta^{(i)} \rangle(k)}{\sqrt{P_{\text{Zel}}(k) \sum_{i,j}^n \langle \delta^{(i)} | \delta^{(j)} \rangle(k)}}. \quad (\text{A4})$$

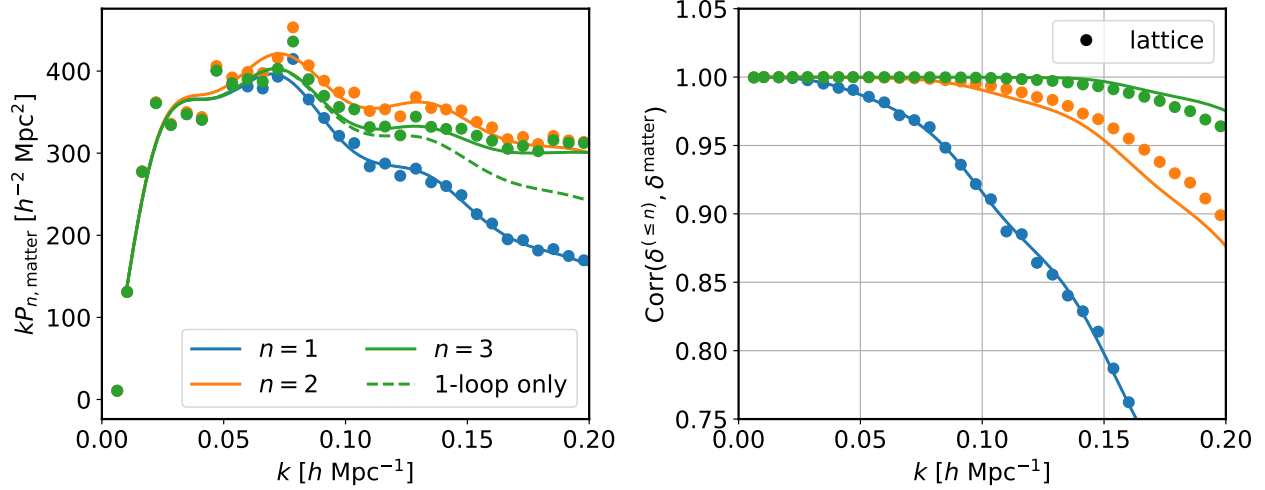


FIG. A2.— Cross-correlation of the nonlinear and Eulerian matter density fields on the lattice (dots) compared to analytic predictions. (Left) The cross power spectrum between the Eulerian field summed to n^{th} order, computed to 1-loop in perturbation theory, with large-IR contributions at higher loops resummed via an exponential. For the third-order cross correlation there is also an IR-enhanced 2-loop diagram, and we show the prediction without it as a dashed line. (Right) The correlation coefficient between the two fields. The analytic predictions break down around $k \sim 0.15 h \text{Mpc}^{-1}$ reflecting the onset of 2-loop mode coupling.

The right panel of Figure A1 shows the measured and predicted correlation coefficients. These are also in excellent agreement until roughly $k \sim 0.15 h \text{Mpc}^{-1}$, where nonlinear mode coupling departs from the 1-loop prediction. To elucidate these cross-correlation coefficients we can write $P_{22} = (k\Sigma)^2 P_{\text{lin}} + P_{1\text{-loop}}$, where $P_{1\text{-loop}}$ is the mode-coupling integral which is left un-cancelled by $2P_{13}$. At one-loop we have $P_{\text{Zel}}(k) \approx P_{\text{lin}}(k) + P_{1\text{-loop}}(k)$, and we may Taylor expand in the large displacement $k\Sigma$ to have

$$\begin{aligned} r_{1,\text{Zel}}(k) &= \frac{1}{\sqrt{1+\lambda}} - \frac{1}{2\sqrt{1+\lambda}} k^2 \Sigma^2 + \mathcal{O}(k^4 \Sigma^4), \\ r_{2,\text{Zel}}(k) &= 1 - \frac{\lambda}{2(1+\lambda)} k^2 \Sigma^2 + \mathcal{O}(k^4 \Sigma^4), \\ r_{3,\text{Zel}}(k) &= 1 - \mathcal{O}(k^4 \Sigma^4), \end{aligned} \quad (\text{A5})$$

where we have defined $\lambda = P_{1\text{-loop}}/P_{\text{lin}}$. These expanded correlation coefficients have a few interesting features: First, even though the second order field by itself has an uncanceled IR divergence $k^2 \Sigma^2 P_{\text{lin}}$, this divergence is cancelled in the numerator and denominator of the correlation coefficient between the nonlinear and quadratic fields. In the absence of other nonlinearities ($\lambda = 0$), this cancels the leading de-correlation of the two fields, since the IR divergence increases the cross correlation but also the noise in the denominator. Second, differences in the higher-order mode coupling can affect the correlation coefficient as well: in the correlation coefficient between the nonlinear and linear fields, for example, this is almost entirely captured by the factor $1/\sqrt{1+\lambda}$. Since $\lambda < 0$, the lack of mode coupling in the linear field enhances its correlation with the nonlinear field. Similarly, the quadratic field has its leading decorrelation cancelled but retains an order $k^2 \Sigma^2$ decorrelation due to mode coupling, which disappears entirely when cubic operators are added.

The above arguments carry straightforwardly to the cross correlation between the nonlinear matter field and its lattice EPT predictions. Here, Equation A1 does not exactly hold, though the infrared contributions derived from it remain the same due to the structure of the EPT mode-coupling kernels (Bernardeau *et al.* 2002). In this case we can write

$$\begin{aligned} P_{1,\text{matter}}(k) &= \left((1 + \alpha k^2) P_{\text{lin}}(k) + P_{13}(k) + \frac{1}{2} k^2 \Sigma^2 P_{\text{lin}} \right) e^{-\frac{1}{2} k^2 \Sigma^2} \\ P_{2,\text{matter}}(k) &= \left((1 + \alpha k^2) P_{\text{lin}}(k) + P_{13}(k) + \frac{1}{2} k^2 \Sigma^2 P_{\text{lin}} + P_{22}(k) \right) e^{-\frac{1}{2} k^2 \Sigma^2} \\ P_{3,\text{matter}}(k) &= \left((1 + \alpha k^2) P_{\text{lin}}(k) + P_{13}(k) + \frac{1}{2} k^2 \Sigma^2 P_{\text{lin}} + P_{22}(k) \right) e^{-\frac{1}{2} k^2 \Sigma^2} + P_{3,\text{matter}}^{\text{IR-enhanced}} \end{aligned} \quad (\text{A6})$$

where α is the counterterm fit to the matter autospectrum and $P_{3,\text{matter}}^{\text{IR-enhanced}}$ is given by swapping the Zeldovich 1-loop terms in Equation A3 for the corresponding matter ones. As can be seen in the left panel of Figure A2, retaining the IR-enhanced terms beyond 1-loop order is sufficient to obtain very good agreement with the cross spectrum of these two types of fields measured in N-body simulations, with no additional parameters (see Taruya *et al.* (2018) for

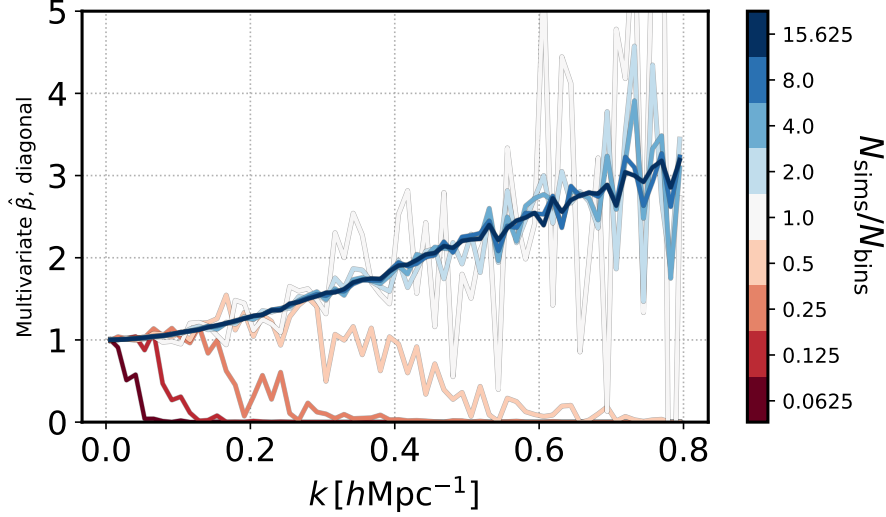


FIG. C1.— Impact of underdetermination of the control variate covariance matrix on the estimate of β due to the use of the Moore-Penrose pseudoinverse. The color indicates how under(over)-determined the covariance matrix estimation is through the parameter $N_{\text{sims}}/N_{\text{bins}}$. When $N_{\text{sims}}/N_{\text{bins}} \geq 1$ the control variate covariance is invertible, and we clearly see there is no ‘damping’ of the Lagrange multiplier in question. When the control variate covariance matrix is under-determined and the pseudoinverse is used, a damping is induced which depends on the degree of underdetermination.

a comparison to the full 2-loop prediction, though without the EFT corrections we have employed here). The right panel of the same figure shows the thus-predicted correlation coefficient, where we have also retained the corresponding 2-loop enhanced IR terms in P_{33} , again with very good agreement up to $k \sim 0.15 h \text{ Mpc}^{-1}$ where nonlinear mode coupling sets in.

B. ORDER-BY-ORDER EULERIAN BISPECTRA ON THE LATTICE

In order to measure the tree-level bispectrum from our lattice realizations we create an artificially modulated second-order density field

$$\delta^{(E)}(\mathbf{x}; \epsilon) = \delta^{(1)}(\mathbf{x}) + \epsilon \delta^{(2)}(\mathbf{x}) + \epsilon^2 \delta^{(3)}(\mathbf{x}), \quad (\text{B1})$$

where ϵ is a free parameter and $\delta^{(2)}(\mathbf{k})$ is the second-order EPT density field

$$\delta^{(2)}(\mathbf{x}) = \frac{17}{21} \delta^{(1)}(\mathbf{x}) - \Psi^{(1)}(\mathbf{x}) \cdot \nabla \delta^{(1)}(\mathbf{x}) + \frac{2}{7} s^2(\mathbf{x}). \quad (\text{B2})$$

The bispectrum of $\delta^{(E)}$ in a given box will be given by the terms

$$B^{EEE}(k_1, k_2, k_3; \epsilon) = B^{111} + \epsilon(B^{211} + B^{121} + B^{112}) + \epsilon^2(B^{221} + B^{122} + B^{212} + B^{311} + B^{131} + B^{113}) + \mathcal{O}(\epsilon^3). \quad (\text{B3})$$

We know exactly B^{111} for an individual simulation, and thus an estimator for B^{Tree} is given by

$$\hat{B}^{\text{Tree}}(k_1, k_2, k_3) = \frac{B^{EEE} - B^{111}}{\epsilon}, \quad (\text{B4})$$

which is accurate to $\mathcal{O}(\epsilon)$. We use the fiducial value of $\epsilon = 10^{-2}$ in this work, from which we find a tree-level spectrum converged to within 0.1% relative to using $\epsilon = 10^{-3}$.

We can also extract quintic bispectra contributions by considering a pair of quadratic fields $\delta^{(E)}(\mathbf{x}; \epsilon), \delta^{(E)}(\mathbf{x}; -\epsilon)$. The combination

$$\frac{(B^{EEE}(k_1, k_2, k_3; \epsilon) + B^{EEE}(k_1, k_2, k_3; -\epsilon)) - 2B^{111}}{2\epsilon^2}$$

estimates the quintic contribution directly. Notice that even though the ϵ^2 terms average to zero, their inclusion as a control variate will contribute to the cross-correlation coefficient.

By modulating higher-order Eulerian contributions with additional powers of ϵ we maintain strict control over the Eulerian order of the control variate we wish to construct. To extend to $\mathcal{O}(\epsilon^3)$, equivalent to the one-loop matter bispectrum, we would also have to construct the $\delta^{(4)}$ Eulerian field. There is no opposition to doing this in principle – the recursion relations are known and [Taruya et al. \(2018, 2021b\)](#) have explored this – but we elect not to since we do not expect qualitative improvements in Eulerian control variates from extending to one more power in the density field.

C. ESTIMATING THE LAGRANGE MULTIPLIER β

In this appendix we discuss in more detail the suitability of a number of approximations to estimate the Lagrange multiplier, β , which provides optimal variance reduction. In the univariate problem it is given by (c.f. eq. (42) for the

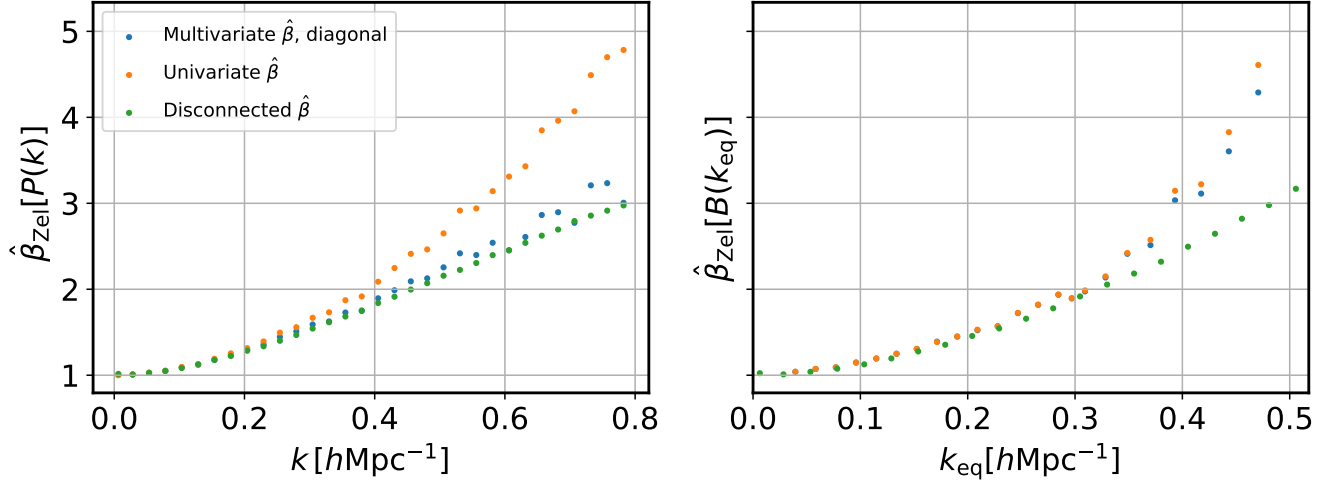


FIG. C2.— Estimate of the Lagrange multiplier, $\hat{\beta}$ for the Zeldovich control variate in the case of the power spectrum (left) and bispectrum (right), for equilateral triangle configurations. The blue points show the diagonal contribution from the full multivariate estimate of β , the orange points show the univariate estimate from the bin-by-bin problem, and the green points show the disconnected approximation to the Lagrange multiplier.

bispectrum-specific case)

$$\hat{\beta}_{\text{uni}} = \frac{\text{Cov}(\mathbf{x}, \mathbf{c})}{\text{diag}(\text{Var}(\mathbf{c}))}. \quad (\text{C1})$$

In the multi-variate problem (which minimizes the determinant of the covariance matrix of $\mathbf{y}$) the equivalent estimate is

$$\hat{\beta}_{\text{multi}} = \text{Cov}(\mathbf{x}, \mathbf{c}) \cdot [\text{Cov}(\mathbf{c})]^{-1}. \quad (\text{C2})$$

Finally, in Kokron *et al.* (2022) we introduced a disconnected approximation to the univariate Lagrange multiplier, which in the case of the matter power spectrum reduces to

$$\hat{\beta}_{\text{disc}} = \left(\frac{P_{N,Z}(k)}{P_N(k)} \right)^2. \quad (\text{C3})$$

In Kokron *et al.* (2022), the authors checked the validity of eq. (C3) against an estimate of eq. (C2) where the number of simulations $N_{\text{sims}} = 100$ was smaller than the number of power spectrum bins $N_k = 512$. In order to invert the control variate covariance, the Moore-Penrose pseudoinverse was used to estimate the inverse. A "damping" of the Lagrange multiplier was observed and attributed to being physical in nature, being well fit by a tanh function.

This damping is not physical and was spuriously driven by the usage of the pseudoinverse. From the fiducial set of $N = 1000$ boxes, we compute the Zeldovich and nonlinear matter N -body power spectra at $z = 0.5$ in $N_{\text{bins}} = 64$, up to $k_{\text{max}} \sim 0.8 h\text{Mpc}^{-1}$. We then compute the multivariate form of β using subsets with $N_{\text{sims}} = [4, 8, 16, 32, 64, 128, 256, 512, 1000]$, numerically inverting the Zeldovich power spectrum covariance matrix. In Fig. C1 we show the resulting diagonal of the Lagrange multiplier matrix as a function of $N_{\text{sims}}/N_{\text{bins}}$, the degree of determination of the control variate covariance matrix $\text{Cov}[\mathbf{c}]$. There is clear evidence for the damping being a function of how well-determined this matrix is, and thus a numerical artifact. This damping is a consequence of the construction of the pseudoinverse, where in the singular value decomposition $\text{Cov}[\mathbf{c}] = U\Sigma V$, the diagonal matrix Σ containing eigenvalues of $\text{Cov}[\mathbf{c}]$ is inverted. Poorly-determined eigenvalues are set to zero, leading to a matrix with a strictly lower trace than the 'true' inverse. The diagonal of the β matrix is consequently smaller as a result. The original analysis of Kokron *et al.* (2022) was in this underdetermined regime, with $N_{\text{sims}}/N_{\text{bins}} \sim 0.2$.

Using this larger sample we now turn to a comparison between the three different estimates eqs. (C1) to (C3) for both the power spectrum and the bispectrum. To compress the dimensionality of the bispectrum analysis we restrict ourselves to equilateral triangles, for which there are $N_{\text{bins}} = 23$ in the Zeldovich binning scheme. Additionally, for the bispectrum, the disconnected Gaussian covariance is proportional to $P(k_1)P(k_2)P(k_3)$ and so the appropriate disconnected Lagrange multiplier is

$$\hat{\beta}_{\text{disc}}(k_1, k_2, k_3) = \frac{P_{N,Z}(k_1)P_{N,Z}(k_2)P_{N,Z}(k_3)}{P_N(k_1)P_N(k_2)P_N(k_3)}. \quad (\text{C4})$$

Fig. C2 shows the three different estimates for β . In the left panel, which shows estimates for the power spectrum, we can see that all three estimates agree well until $k \sim 0.2 h\text{Mpc}^{-1}$ at which point they begin to diverge. The disconnected approximation closely tracks the diagonal of the multivariate term through all scales considered, while the univariate β grows at a faster rate. We stress there is no reason a priori that these different estimates of β should

closely agree across all scales – the multivariate control variate problem optimizes over a fundamentally different objective function than the univariate bin-by-bin problem. In the case of the power spectrum we see the disconnected approximation is a close match to the multivariate estimator. In the case of the equilateral bispectrum the picture is somewhat different – the univariate and multivariate estimates of β agree closely for all triangles considered, while the disconnected approximation underestimates the covariance-based values of β starting at $k_{\text{equi}} = 0.3 \, h\text{Mpc}^{-1}$. This can be understood by studying the behavior of the bispectrum precision, shown in Fig. 11. $k \approx 0.3 \, h\text{Mpc}^{-1}$ is precisely the triangle scale at which the SNR estimated from using the disconnected nonlinear covariance begins to disagree from the full empirical estimate – a proxy for the importance of connected terms in the covariance.

How much variance reduction is lost from adopting a slightly sub-optimal value of β ? Suppose one uses $\beta = \beta^*(1 + \epsilon)$ where β^* is the optimal multiplier for the univariate control variate problem. In this case, a straightforward calculation shows the variance reduction goes to

$$\frac{\sigma_y^2}{\sigma_x^2} = (1 - \rho^2) + \epsilon^2 \rho^2. \quad (\text{C5})$$

For the disconnected approximations shown in Fig. C2, we find that the largest amount observed is $\epsilon^2 \approx 0.15$ at the smallest scales considered, where the cross-correlation coefficient ρ is already suppressed.

REFERENCES

- 2pt Collaboration, E. Krause, Y. Kobayashi, A. N. Salcedo, M. M. Ivanov, T. Abel, K. Akitsu, R. E. Angulo, G. Cabass, S. Contarini, C. Cuesta-Lazaro, C. Hahn, N. Hamaus, D. Jeong, C. Modi, N.-M. Nguyen, T. Nishimichi, E. Paillas, M. P. Ibañez, O. H. E. Philcox, A. Pisani, F. Schmidt, S. Tanaka, G. Verza, S. Yuan, and M. Zennaro, “A parameter-masked mock data challenge for beyond-two-point galaxy clustering statistics,” (2024), [arXiv:2405.02252 \[astro-ph.CO\]](#).
- R. Scoccimarro, *The Astrophysical Journal* **544**, 597–615 (2000).
- R. E. Angulo, S. Foreman, M. Schmittfull, and L. Senatore, *Journal of Cosmology and Astroparticle Physics* **2015**, 039–039 (2015).
- O. H. Philcox, M. M. Ivanov, G. Cabass, M. Simonović, M. Zaldarriaga, and T. Nishimichi, *Physical Review D* **106** (2022), [10.1103/physrevd.106.043530](#).
- G. D’Amico, Y. Donath, M. Lewandowski, L. Senatore, and P. Zhang, *Journal of Cosmology and Astroparticle Physics* **2024**, 041 (2024).
- T. Bakx, M. M. Ivanov, O. H. E. Philcox, and Z. Vlah, “One-loop galaxy bispectrum: Consistent theory, efficient analysis with cobra, and implications for cosmological parameters,” (2025), [arXiv:2507.22110 \[astro-ph.CO\]](#).
- K. Pardede, F. Rizzo, M. Biagetti, E. Castorina, E. Sefusatti, and P. Monaco, *Journal of Cosmology and Astroparticle Physics* **2022**, 066 (2022).
- M. Wang, F. Beutler, J. Aguilar, S. Ahlen, D. Bianchi, D. Brooks, T. Claybaugh, A. de la Macorra, P. Doel, A. Font-Ribera, E. Gaztañaga, G. Gutierrez, K. Honscheid, C. Howlett, D. Kirkby, A. Lambert, M. Landriau, R. Miquel, G. Niz, F. Prada, I. Pérez-Ràfols, G. Rossi, E. Sanchez, D. Schlegel, M. Schubnell, D. Sprayberry, G. Tarlé, and B. Weaver, *Journal of Cosmology and Astroparticle Physics* **2025**, 031 (2025).
- C. Hahn, R. Scoccimarro, M. R. Blanton, J. L. Tinker, and S. Rodríguez-Torres, *Monthly Notices of the Royal Astronomical Society*, [stx185](#) (2017).
- A. Chudaykin, M. M. Ivanov, and O. H. E. Philcox, “Reanalyzing desi dr1: 1. Λ cdm constraints from the power spectrum and bispectrum,” (2025), [arXiv:2507.13433 \[astro-ph.CO\]](#).
- M. Biagetti, L. Castiblanco, J. Noreña, and E. Sefusatti, *Journal of Cosmology and Astroparticle Physics* **2022**, 009 (2022).
- R. E. Angulo and O. Hahn, *Living Reviews in Computational Astrophysics* **8** (2022), [10.1007/s41115-021-00013-z](#).
- R. Takahashi, T. Nishimichi, T. Namikawa, A. Taruya, I. Kayo, K. Osato, Y. Kobayashi, and M. Shirasaki, *The Astrophysical Journal* **895**, 113 (2020).
- A. Pontzen, A. Slosar, N. Roth, and H. V. Peiris, *Physical Review D* **93** (2016), [10.1103/physrevd.93.103519](#).
- R. E. Angulo and A. Pontzen, *Mon. Not. Roy. Astron. Soc.* **462**, L1 (2016), [arXiv:1603.05253 \[astro-ph.CO\]](#).
- F. Maion, R. E. Angulo, and M. Zennaro, *arXiv e-prints*, [arXiv:2204.03868](#) (2022), [arXiv:2204.03868 \[astro-ph.CO\]](#).
- A. B. Owen, *Monte Carlo theory, methods and examples* (2013).
- N. Chartier, B. Wandelt, Y. Akrami, and F. Villaescusa-Navarro, *arXiv e-prints*, [arXiv:2009.08970](#) (2020), [arXiv:2009.08970 \[astro-ph.CO\]](#).
- S. Tassev and M. Zaldarriaga, *Journal of Cosmology and Astroparticle Physics* **2012**, 013–013 (2012).
- N. Kokron, S.-F. Chen, M. White, J. DeRose, and M. Maus, *Journal of Cosmology and Astroparticle Physics* **2022**, 059 (2022).
- J. DeRose, S.-F. Chen, N. Kokron, and M. White, *Journal of Cosmology and Astroparticle Physics* **2023**, 008 (2023).
- B. Hadzhiyska, M. J. White, X. Chen, L. H. Garrison, J. DeRose, N. Padmanabhan, C. Garcia-Quintero, J. Mena-Fernández, S.-F. Chen, H.-J. Seo, P. McDonald, J. Aguilar, S. Ahlen, D. Brooks, T. Claybaugh, A. de la Macorra, P. Doel, A. Font-Ribera, J. E. Forero-Romero, S. G. A. Gontcho, K. Honscheid, A. Kremin, M. Landriau, M. Manera, R. Miquel, J. Nie, N. Palanque-Delabrouille, M. Rezaie, G. Rossi, E. Sanchez, M. Schubnell, G. Tarlé, and Z. Zhou, *The Open Journal of Astrophysics* **6** (2023), [10.21105/astro.2308.12343](#).
- N. Chartier and B. D. Wandelt, *Monthly Notices of the Royal Astronomical Society* (2021), [10.1093/mnras/stab3097](#).
- N. Chartier and B. D. Wandelt, (2022), [arXiv:2204.03070 \[astro-ph.CO\]](#).
- Z. Ding *et al.*, (2022), [arXiv:2202.06074 \[astro-ph.CO\]](#).
- Z. Ding, A. Variu, S. Alam, Y. Yu, C. Chuang, E. Paillas, C. Garcia-Quintero, X. Chen, J. Mena-Fernández, J. Aguilar, S. Ahlen, D. Brooks, T. Claybaugh, A. de la Macorra, P. Doel, K. Fanning, J. E. Forero-Romero, E. Gaztañaga, S. G. A. Gontcho, G. Gutierrez, C. Hahn, K. Honscheid, C. Howlett, S. Juneau, R. Kehoe, T. Kisner, A. Kremin, A. Lambert, M. Landriau, L. L. Guillou, M. Manera, R. Miquel, E. Mueller, A. D. Myers, J. Nie, G. Niz, C. Poppett, M. Rezaie, G. Rossi, E. Sanchez, M. Schubnell, H. Seo, J. Silber, D. Sprayberry, G. Tarlé, M. Vargas-Magaña, and H. Zou, “Suppressing the sample variance of desi-like galaxy clustering with fast simulations,” (2025), [arXiv:2404.03117 \[astro-ph.CO\]](#).
- Y. B. Zel’Dovich, *Astronomy & Astrophysics* **500**, 13 (1970).
- M. Schmittfull, M. Simonović, M. M. Ivanov, O. H. E. Philcox, and M. Zaldarriaga, “Modeling galaxies in redshift space at the field level,” (2020), [arXiv:2012.03334 \[astro-ph.CO\]](#).
- F. Villaescusa-Navarro, C. Hahn, E. Massara, A. Banerjee, A. M. Delgado, D. K. Ramanah, T. Charnock, E. Giusarma, Y. Li, E. Allys, A. Brochard, C. Uhlemann, C.-T. Chiang, S. He, A. Pisani, A. Obuljen, Y. Feng, E. Castorina, G. Contardo, C. D. Kreisch, A. Nicola, J. Alsing, R. Scoccimarro, L. Verde, M. Viel, S. Ho, S. Mallat, B. Wandelt, and D. N. Spergel, *The Astrophysical Journal Supplement Series* **250**, 2 (2020).
- O. H. E. Philcox and T. Flöss, “Polybin3d: A suite of optimal and efficient power spectrum and bispectrum estimators for large-scale structure,” (2024), [arXiv:2404.07249 \[astro-ph.CO\]](#).
- R. Scoccimarro, M. Zaldarriaga, and L. Hui, *The Astrophysical Journal* **527**, 1–15 (1999).

- E. Sefusatti, M. Crocce, R. Scoccimarro, and H. M. P. Couchman, *Monthly Notices of the Royal Astronomical Society* **460**, 3624–3636 (2016).
- S.-F. Chen, Z. Vlah, and M. White, *J. Cosmology Astropart. Phys.* **2024**, 012 (2024), arXiv:2406.00103 [astro-ph.CO].
- E. Sefusatti, M. Crocce, and V. Desjacques, *MNRAS* **406**, 1014 (2010), arXiv:1003.0007 [astro-ph.CO].
- N. E. Chisari and A. Pontzen, *Phys. Rev. D* **100**, 023543 (2019), arXiv:1905.02078 [astro-ph.CO].
- J. Tomlinson and D. Jeong, *JCAP* **08**, 040 (2023), arXiv:2204.00668 [astro-ph.CO].
- B. Grinstein and M. B. Wise, *ApJ* **320**, 448 (1987).
- F. Bernardeau, S. Colombi, E. Gaztanaga, and R. Scoccimarro, *Phys. Rept.* **367**, 1 (2002), arXiv:astro-ph/0112551.
- A. Taruya, T. Nishimichi, and D. Jeong, *Physical Review D* **98** (2018), 10.1103/physrevd.98.103532.
- M. M. Ivanov, O. H. Philcox, T. Nishimichi, M. Simonović, M. Takada, and M. Zaldarriaga, *Physical Review D* **105** (2022), 10.1103/physrevd.105.063512.
- T. Matsubara, *Physical Review D* **92** (2015), 10.1103/physrevd.92.023534.
- M. Schmittfull, M. Simonović, V. Assassi, and M. Zaldarriaga, *Physical Review D* **100** (2019), 10.1103/physrevd.100.043514.
- Y. Feng, M.-Y. Chu, U. Seljak, and P. McDonald, *MNRAS* **463**, 2273 (2016), arXiv:1603.00476 [astro-ph.CO].
- A. Taruya, T. Nishimichi, and D. Jeong, “Grid-based calculations of redshift-space matter fluctuations from perturbation theory: U_V sensitivity and convergence at the field level.” (2021a).
- N. Hand, Y. Feng, F. Beutler, Y. Li, C. Modi, U. Seljak, and Z. Slepian, *The Astronomical Journal* **156**, 160 (2018).
- C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, *Nature* **585**, 357–362 (2020).
- P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Í. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, *Nature Methods* **17**, 261 (2020).
- J. D. Hunter, *Computing in Science Engineering* **9**, 90 (2007).
- A. Taruya, T. Nishimichi, and D. Jeong, *Phys. Rev. D* **103**, 023501 (2021b), arXiv:2007.05504 [astro-ph.CO].

This paper was built using the Open Journal of Astrophysics L^AT_EX template. The OJA is a journal which

provides fast and easy peer review for new papers in the **astro-ph** section of the arXiv, making the reviewing process simpler for authors and referees alike. Learn more at <http://astro.theoj.org>.