Title
Priors on red galaxy stochasticity from hybrid effective field theory

Permalink
https://escholarship.org/uc/item/8bs7k87g

Journal
Monthly Notices of the Royal Astronomical Society, 514(2)

ISSN
0035-8711

Authors
Kokron, Nickolas
DeRose, Joseph
Chen, Shi-Fan
et al.

Publication Date
2022-06-16

DOI
10.1093/mnras/stac1420

Peer reviewed

Priors on red galaxy stochasticity from hybrid effective field theory

Nickolas Kokron[1,2] ⋆, Joseph DeRose[3], Shi-Fan Chen[3,4], Martin White[3,4,5], Risa H. Wechsler[1,2]
[1] *Kavli Institute for Particle Astrophysics and Cosmology and Department of Physics, Stanford University, 382 Via Pueblo Mall, Stanford, CA 94305, USA*
[2] *Kavli Institute for Particle Astrophysics and Cosmology, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA*
[3] *Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720*
[4] *Department of Physics, University of California, Berkeley, CA 94720*
[5] *Department of Astronomy, University of California, Berkeley, CA 94720*

ABSTRACT

We investigate the stochastic properties of typical red galaxy samples in a controlled numerical environment. We use Halo Occupation Distribution (HOD) modelling to create mock realizations of three separate bright red galaxy samples consistent with datasets used for clustering and lensing analyses in modern galaxy surveys. Second-order Hybrid Effective Field Theory (HEFT) is used as a field-level forward model to describe the full statistical distribution of these tracer samples, and their stochastic power spectra are directly measured and compared to the Poisson shot-noise prediction. While all of the galaxy samples we consider are hosted within haloes with sub-Poisson stochasticity, we observe that the galaxy samples themselves possess stochasticities that range from sub-Poisson to super-Poisson, in agreement with predictions from the halo model. As an application of our methodology, we place priors on the expected degree of non-Poisson stochasticity in cosmological analyses using such samples. We expect these priors will be useful in reducing the complexity of the full parameter space for future analyses using second-order Lagrangian bias models. More generally, the techniques outlined here present the first application of hybrid EFT methods to characterize models of the galaxy–halo connection at the field level, revealing new connections between once-disparate modelling frameworks.

Key words: cosmology: theory – large-scale structure of Universe – methods: statistical – methods: computational

1 INTRODUCTION

Spectroscopic galaxy redshift surveys are poised to produce some of the leading cosmological datasets of the upcoming decade. The Dark Energy Spectroscopic Instrument (DESI, Aghamousa et al. 2016), for example, will observe an order of magnitude more galaxies than the incredibly successful Sloan Digital Sky Surveys (Dawson et al. 2012, 2016). Other galaxy survey probes, such as the Vera Rubin Observatory's Legacy Survey of Space and Time (LSST, Ivezić et al. (2019); Mandelbaum et al. (2018)), will measure the shapes of roughly ten billion of galaxies and tease out the correlated weak gravitational lensing signal therein, directly measuring the gravitational effect of dark matter at unprecedented statistical power over half the sky. The upcoming increase of statistical power afforded by next-generation cosmic surveys is especially timely, as they will shed light on several 'tensions' that have crept up between cosmological datasets over the recent years. These include the tension over measurements of the Hubble constant (Di Valentino et al. 2021, or, alternatively, in the sound horizon) between early and late-Universe probes and the recently hinted S_8 tension over the amplitude of density fluctuations in the Universe compared to those predicted by observations of the cosmic microwave background (CMB) and the assumption of ΛCDM (Krolewski et al. 2021; Collaboration et al. 2021; Heymans et al. 2021; White et al. 2021). Such tensions could signify a breakdown of the standard cosmological model, ΛCDM, if validated by larger datasets.

However, the vast statistical power of future datasets simultaneously presents a significant challenge to the models we use to analyze them. Significant care must be taken to ensure that their accuracy is sub-dominant compared to statistical and systematic uncertainties; mischaracterizing the accuracy of models for describing the statistical properties of galaxy surveys could lead to biased inferences on the properties of the Universe.

In particular, models for the statistical properties of galaxy distributions must surmount two individual challenges. First, the statistical properties of the late-time dark matter distribution itself, given a cosmological model, must be well understood. This is a challenging task, as the gravitational collapse problem of the cold dark matter fluid is a non-linear process. Significant progress in this regard has been made via the numerical study of this problem, using N-body simulations of structure formation (Hockney & Eastwood 1988; Bagla 2005; Kuhlen et al. 2012; Schneider et al. 2016). Despite the large computational cost of running individual N-body simulations, computational power and statistical tools have progressed significantly, and one can now run suites of simulations that span several points in cosmological parameter space. With these suites, *emulators* have become commonplace tools in predicting the non-linear properties of structure formation. Measurements of any given observable across the simulation suite serve as inputs to models that predict non-linear statistics of the dark matter distribution rapidly, and their accuracy can be well calibrated given a suitable experimental design (Heitmann et al.

⋆ Contact e-mail: kokron@stanford.edu

2013; Garrison et al. 2018; DeRose et al. 2019; Knabenhans et al. 2019; Angulo et al. 2021)

An emulator for dark matter statistics alone, however, still cannot be used to predict signals of the clustering properties of galaxies. Modelling the *galaxy–halo connection*, or more broadly, the *tracer–matter connection* is the second challenge that must be surmounted in order to construct a model suitable for end-to-end analysis of galaxy survey data. Models for the tracer–matter connection fall into several different categories. Here, we highlight two such categories: empirical/statistical and analytic/perturbative models of the tracer–matter connection.

Empirical models include so-called halo occupation distributions (HOD), (sub)-halo abundance matching, and direct modeling of the formation histories of galaxies, among others. Empirical models attempt to infer statistical relations between halos identified in dark matter simulations and mock galaxy populations that inhabit them, in light of both observational data on the given population and properties of the host dark matter haloes. Such data includes, for example: two-and-three point correlation functions (Zheng et al. 2005; Yuan et al. 2018), luminosity functions (Yang et al. 2003; Cooray 2006), and measurements of stellar mass functions and star formation rates (Behroozi et al. 2013). Empirical models allow for deep insights into galaxy formation and evolution (Behroozi et al. 2019), the creation of realistic mock realizations of sky surveys (Wechsler et al. 2021) as well as offer potent frameworks to describe the statistical properties of galaxies down to very small scales (DeRose et al. 2021). We refer the readers to Wechsler & Tinker (2018) for a comprehensive review on empirical models and other simulation-based models of the galaxy–halo connection.

Perturbative models for the tracer–matter connection, also known as *bias models* (see Desjacques et al. 2018 for a comprehensive review), try to capture the relationship between the dark matter and a population of tracers in a different way. Instead of explicitly relating the properties of haloes to the properties of galaxy samples, bias models specify a functional form for the relation between the large-scale, smoothed, dark matter density and the density of tracers under consideration. This functional form is restricted by a set of symmetries that hold in the relation, and the given order in powers of the aforementioned density one is working in. Each term in the bias expansion is accompanied by a free coefficient that captures the response of the tracer population to that term. The flexible parameterization of bias models imply they should be able to describe, within their regime of applicability, the statistical properties of *any* tracer sample whose properties obey the imposed symmetries.

While the bias expansion captures the deterministic relation between a tracer's distribution and that of the underlying matter field, there is an additional stochastic component in this relation that decorrelates the two fields, due to small-scale processes. This noise, scatter, or *stochasticity* contribution to the tracer–matter connection is an important component that must be understood if we wish to extract the most information from our datasets. Notably, stochasticity becomes important at small scales where higher-order bias terms are also expected to be significant, and their impact on observables is partially degenerate with these bias terms. Indeed, stochasticity has previously been defined as not only this random component but also as the impact of not including higher-order bias terms in a model for the distribution of galaxies (Baldauf et al. 2013). A lack of prior understanding of the effects of stochasticity can lead to significant degradation of cosmological constraining power. Significant efforts have been undertaken to characterize the stochastic properties of galaxy samples, however not to the same extent as galaxy bias. We refer the reader to Baldauf et al. (2016); Paech et al. (2017); Ginzburg et al. (2017); Friedrich et al. (2021); Sullivan et al. (2021) for previous discussions on the interplay between stochasticity and bias modelling.

In this work we use field-level realizations of Lagrangian bias models with fully non-linear dark matter dynamics, an approach recently dubbed Hybrid Effective Field Theory (HEFT), to study the properties of specfic galaxy samples, focusing on the example red galaxy samples that are used as both clustering and lens samples in cross-correlation analyses. Specifically, we use HEFT to measure the amplitude of the stochasticity of said samples, using fits to their properties via HODs as a proxy for their statistical properties. These measurements can then be used to place priors on subsequent analyses that help reduce the computational complexity of the inference procedure. Our methodology highlights the synergies in using multiple models of the tracer–matter connection to study the same galaxy sample. The study of stochasticity we describe below also raises a number of new ways that bias models and empirical models may be combined to shed light into the ways that galaxies, or haloes, populate the broader large-scale structure of the Universe.

The paper is structured as follows: in § 2.1, 2.2 we give a brief overview of both HODs and hybrid EFT, the two statistical tools we use to characterize the tracer–matter connection in this paper. In § 2.3 we outline our procedure to use field-level realizations of the bias model to both estimate the bias parameters of HOD samples and consequently measure their stochastic power spectra. In § 2.4 we review some results on the causes of non-Poisson stochasticity in the framework of the halo model. Specifically we discuss two competing effects, one-halo *enhancement* and halo *exclusion*, which drive the large-scale stochasticity of galaxy samples to the super- and sub-Poisson regime, respectively. In § 3 we outline both the simulation suite used and the mock galaxies we populate onto these simulations using said HODs. We discuss the functional forms used and the derived parameters of the HODs for the three samples under consideration, as well as how their differences frame our results. In § 4 we report the results of our stochastic power spectrum measurement procedure applied to halo samples and to HOD samples. Specifically, in § 4.1 we look at previous results in the literature on the stochasticity of haloes within the context of our HEFT model. In § 4.2 we report the results of the same procedure on our HOD samples, as well as a discussion and some interpretation of the results that we find. In § 4.3 we address a related question of the distribution of large-scale stochasticities for all HOD models that are consistent with a given sample of DESI-like luminous red galaxies (LRGs). In § 4.4 we present priors on the allowed range of deviations from Poisson stochasticity that we expect, as a result of our experiments across different samples of galaxies and within the DESI-like sample.

2 METHODS

2.1 Halo Occupation Distributions

Halo occupation distribution modelling is an empirical parameterization of the way galaxies occupy haloes. This is done via a probabilistic mapping that specifies the average number of galaxies of a kind that are hosted within a halo. The standard HOD models (Berlind & Weinberg 2002; Zheng et al. 2005, 2007) separate these galaxies into *central* and *satellite* galaxies. A commonly used parameterization is given by

$$\langle N_{\rm cen}(M) \rangle = \frac{f_{\rm cen}}{2}\left[1+{\rm erf}\left(\frac{\log M-\log M_{\rm min}}{\sigma_{\log M}}\right)\right], \quad (1)$$

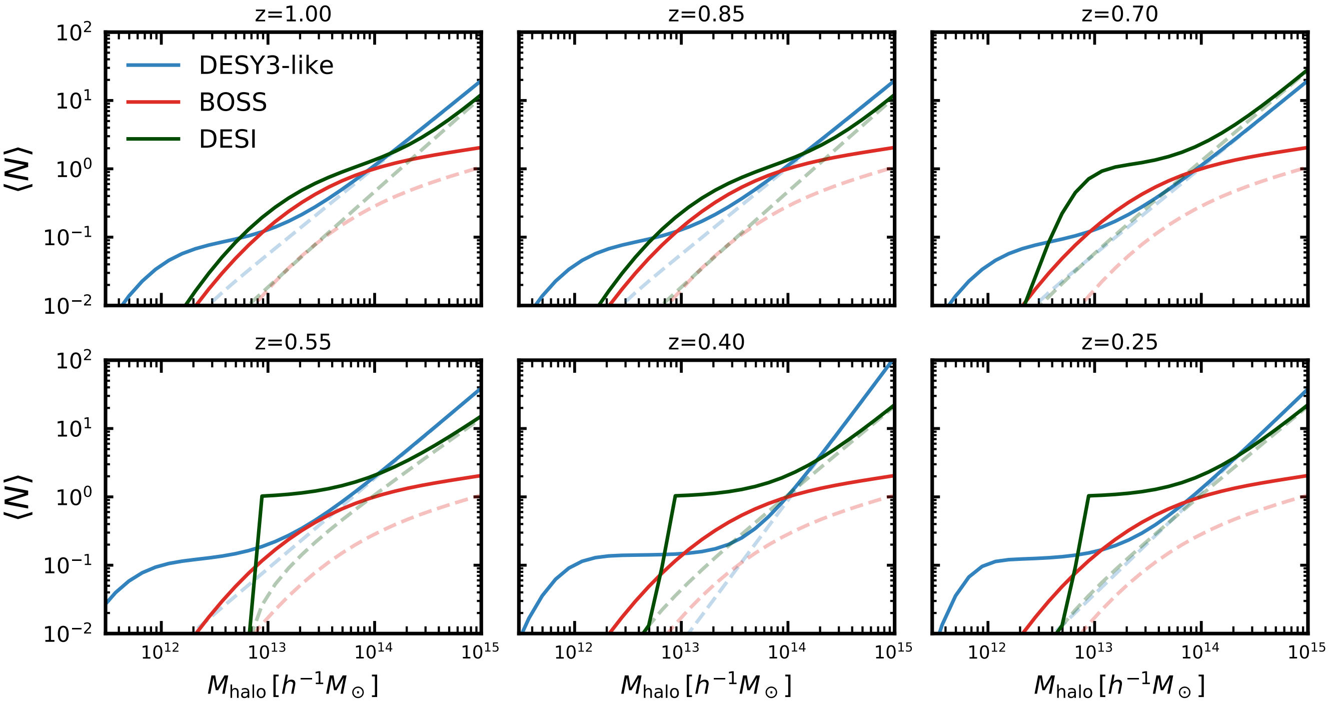

Figure 1. Mean HOD (average number of galaxies as a function of halo mass) for three red galaxy samples considered in this work. The color represents the type of HOD adopted, with solid (dashed) lines showing the number of total (satellite) galaxies per halo mass, respectively.

and

$$\langle N_{\rm sat}(M)\rangle = \left[\frac{M - M_0}{M_1}\right]^{\alpha}. \tag{2}$$

This totals six parameters, however the parameter $f_{\rm cen}$ is usually fixed to $f_{\rm cen} = 1$. Additionally, some HOD models adopting this parameterization also alternate between using $\langle N_{\rm sat}(M)\rangle$ and $\langle N_{\rm sat}(M)\rangle\langle N_{\rm cen}(M)\rangle$ for the expected number of satellites (see e.g. discussion in Reddick et al. 2013). Physically, this corresponds to down-weighting systems without centrals as also being less likely to host satellite galaxies.

While the parameterization of Eqns. 1 and 2 are standard, it is by no means an exhaustive list of occupation prescriptions adopted in the literature. First due to only depending on halo mass they are inevitably incomplete, as it is known that properties other than halo mass (such as concentration, spin, and local environment) influence summary statistics (Wechsler et al. 2002; Gao et al. 2005; Wechsler et al. 2006; Dalal et al. 2008; Mao et al. 2018; Salcedo et al. 2018; Chue et al. 2018; Mansfield & Kravtsov 2020). This phenomenon, known as assembly bias, has motivated extensions to the standard HOD (Hearin et al. 2016; Yuan et al. 2018). These extensions can capture the dependence of halo occupation on additional properties beyond mass, but at the cost of introducing more free parameters.

Nevertheless, it seems that for samples of LRGs observed by current and upcoming surveys this parameterization might be sufficient (Zacharegkas et al. 2021, however, see Yuan et al. (2021) for further discussion on this topic). Understanding the applicability of the standard HOD parameterization is an active subject of research (Hadzhiyska et al. 2020), with recent results indicating that alternative samples of galaxies such as emission line galaxies observed by DESI will require alternative parameterizations (Hadzhiyska et al. 2021b).

2.2 Hybrid Effective Field Theory

Lagrangian biasing theory establishes a statistical relationship between the smoothed, large-scale properties of any given tracer sample and the underlying matter distribution at very high redshift (Matsubara 2008b). The functional form of this expression is given by a functional series expansion in the quantities allowed by the equivalence principle, rotational symmetry and translational symmetry at the Lagrangian coordinates $\boldsymbol{q}$:

$$\delta_h(\boldsymbol{q}) = F[\delta(\boldsymbol{q}), s_{ij}(\boldsymbol{q})] + \epsilon(\boldsymbol{q}), \tag{3}$$

where δ_h is the proto-tracer density contrast, δ is the matter density contrast, and s_{ij} is the traceless tidal tensor field. The field $\epsilon(\boldsymbol{q})$ is a stochastic field that captures the fact the process of tracer formation is not purely deterministic when considering the large-scale smoothed fields that this expansion is applicable to. Including all terms to second order, the expansion of $F[\delta(\boldsymbol{q}), s_{ij}(\boldsymbol{q})]$ is given by (Vlah et al. 2016)

$$\begin{aligned} F[\delta(\boldsymbol{q}), s_{ij}(\boldsymbol{q})] \approx\, & 1 + b_1\delta(\boldsymbol{q}) + b_2(\delta^2(\boldsymbol{q}) - \langle\delta^2\rangle) + \\ & b_{s^2}(s^2(\boldsymbol{q}) - \langle s^2\rangle) + b_{\nabla^2}\nabla^2\delta(\boldsymbol{q}). \end{aligned} \tag{4}$$

At low redshifts, the statistical properties of these tracer fields depend on the combination of the initial relation of Eqn. 3 and the time evolution of its ingredients under the influence of gravity. This time evolution is captured by the advection process from the Lagrangian coordinates to the late-time positions of tracer particles of the matter density

$$\boldsymbol{x} = \boldsymbol{q} + \boldsymbol{\Psi}(\boldsymbol{q}), \tag{5}$$

where $\boldsymbol{\Psi}(\boldsymbol{q})$ is the Lagrangian displacement vector. Under number density conservation, the distribution of tracers at late times is then given by

$$1+\delta_h(\boldsymbol{x}) = \int d^3q\,[F[\delta(\boldsymbol{q}), s_{ij}(\boldsymbol{q})] + \epsilon(\boldsymbol{q})] \\ \times \delta^D\left(\boldsymbol{x}-\boldsymbol{q}-\boldsymbol{\Psi}(\boldsymbol{q})\right). \tag{6}$$

Using Lagrangian Perturbation Theory (LPT), one can analytically determine order-by-order the properties of $\boldsymbol{\Psi}(\boldsymbol{q})$ (Matsubara 2008a; Carlson et al. 2012). These can then be combined with the expansion in Eqn. 4 to create a model for the summary statistics of $\delta_h(\boldsymbol{q})$ (see Chen et al. 2020, 2021 and references therein for an overview of the state of LPT).

It has recently been pointed out (Modi et al. 2020) that N-body simulations of structure formation similarly solve for $\boldsymbol{\Psi}(\boldsymbol{q})$, however in a non-perturbative way. This then implies that the ingredients of the bias expansion can be combined with the numerical displacements from an N-body simulation to create late-time representations of the basis fields that compose the expansion.

At late times, a tracer field under this second-order hybrid expansion is explicitly written as

$$\delta_h(\boldsymbol{x}) = \delta_m(\boldsymbol{x}) + b_1\mathcal{O}_\delta(\boldsymbol{x}) + b_{\nabla^2}\mathcal{O}_{\nabla^2\delta}(\boldsymbol{x}) + \tag{7}$$

$$b_2\mathcal{O}_{\delta^2}(\boldsymbol{x}) + b_{s^2}\mathcal{O}_{s^2}(\boldsymbol{x}) + \epsilon(\boldsymbol{x}), \tag{8}$$

where $\delta_m(\boldsymbol{x})$ is the dark matter density contrast from the simulation. The operators $\mathcal{O}_i$ are constructed by advecting each operator to late-times using the $\boldsymbol{\Psi}(\boldsymbol{q})$ obtained from simulations.

This combination of analytic bias and N-body displacements forms the model known as Hybrid Effective Field Theory (HEFT) and it is the main tool used in this paper to explore the properties of the $\epsilon(\boldsymbol{x})$ stochastic field. We refer the reader to Modi et al. (2020); Kokron et al. (2021); Zennaro et al. (2021a) for further discussions of HEFT and to Hadzhiyska et al. (2021a) for an application of HEFT to survey data.

2.3 Bias parameters and stochastic spectra

From our construction of the field-level model for the tracer overdensity (Eqn. 7), we can rearrange the expression to provide an estimate of the stochastic field $\epsilon(\mathbf{k})$ for a given set of bias parameters b_i

$$\epsilon(\mathbf{k}) = \delta_h(\mathbf{k}) - \delta_m(\mathbf{k}) - \sum_i b_i\mathcal{O}_i(\mathbf{k}), \tag{9}$$

where the variable $\mathbf{k}$ indicates that we are treating the fields in Fourier space. The *stochastic power spectrum* is then defined as the power spectrum of this residual field for a given realization[1]

$$P_{\rm err}(k) \equiv \langle\epsilon(\mathbf{k})\epsilon(-\mathbf{k})\rangle. \tag{10}$$

We can use Eqn. 9 and the standard estimator for the expectation value in the power spectrum to find an explicit expression for $P_{\rm err}(k)$ as a function of the bias parameters

$$P_{\rm err}(k) = \frac{1}{N(k)}\sum_{\mathbf{k}\in\mathcal{S}(k)}\left\|\delta_h(\mathbf{k}) - \delta_m(\mathbf{k}) - \sum_i b_i\mathcal{O}_i(\mathbf{k})\right\|^2, \tag{11}$$

where $\mathcal{S}(k)$ is the spherical shell of radius k and width dk and $N(k)$ is the number of Fourier modes that fall within $\mathcal{S}(k)$.

Previous studies have estimated the best-fit bias parameters at a given scale by solving the least-squares problem of minimizing the error power spectrum (Schmittfull et al. 2019), leading to the so-called *bias transfer functions*:

$$\hat{b}_i(k) = \langle\mathcal{O}_i\mathcal{O}_j\rangle^{-1}(k)\,\langle\mathcal{O}_j(-\mathbf{k})\,[\delta_h(\mathbf{k}) - \delta_m(\mathbf{k})]\rangle\,. \tag{12}$$

If our model for the tracer field is sufficiently accurate, then $P_{\rm err}(k)$ should correspond solely to the power spectrum of stochastic contributions. However, the determinations at a given k-scale are independent and one could find that the estimate of $\hat{b}_i(k_*)$ with k_* a small-scale mode could degrade the fit to the error power spectrum at large-scales. Instead, we seek a comparable estimator for these bias parameters that appropriately penalizes over-fitting at small scales.

We first apply a low-pass sharp-k filter to $\epsilon(\boldsymbol{x})$ to remove the influence of very small scale modes in finding the optimal bias parameters. We represent these smoothed fields as $[\epsilon(\boldsymbol{x})]_{k_{\rm max}}$. If we then choose to minimize the average configuration-space stochastic field squared

$$S = \langle[\epsilon(\boldsymbol{x})]^2_{k_{\rm max}}\rangle, \tag{13}$$

we find a loss function that is very similar to the *EFT likelihood* of Schmidt et al. (2019) and Cabass & Schmidt (2020):

$$S \approx \int_{|\mathbf{k}|<k_{\rm max}}\frac{d^3k}{(2\pi)^3}\left\|\delta_h(\mathbf{k}) - \delta_m(\mathbf{k}) - \sum_i b_i\mathcal{O}_i(\mathbf{k})\right\|^2. \tag{14}$$

Minimizing S with respect to bias parameters leads to an estimator $\hat{b}_i$ comparable to Eqn. 12 but that includes information from *all* modes until a maximum $k_{\rm max}$

$$\hat{b}_i = M^{-1}_{ij}A_j, \tag{15}$$

where A_j and M_{ij} are defined as

$$A_j = \langle[\mathcal{O}_j(\boldsymbol{x})\,(\delta_h(\boldsymbol{x}) - \delta_m(\boldsymbol{x}))]_{k_{\rm max}}\rangle, \tag{16}$$

$$= \int_{|\mathbf{k}|<k_{\rm max}}\frac{d^3k}{(2\pi)^3}\mathcal{O}_j(\mathbf{k})[\delta_h - \delta_m]^*(\mathbf{k}), \tag{17}$$

and

$$M_{ij} = \langle[\mathcal{O}_i(\boldsymbol{x})\mathcal{O}_j(\boldsymbol{x})]_{k_{\rm max}}\rangle, \tag{18}$$

$$= \int_{|\mathbf{k}|<k_{\rm max}}\frac{d^3k}{(2\pi)^3}\mathcal{O}_i(\mathbf{k})\mathcal{O}^*_j(\mathbf{k}). \tag{19}$$

The procedure we adopt to estimate $P_{\rm err}(k)$ is then as follows. We obtain estimates for the bias parameters using Eqn. 15. We proceed to use these $\hat{b}_i$ to create realizations of the tracer fields and subtract these from the tracer sample, realizing Eqn. 9 and our estimate $\hat{\epsilon}(\mathbf{k})$

$$\hat{\epsilon}(\mathbf{k}) = \delta_h(\mathbf{k}) - \delta_m(\mathbf{k}) - \sum_i \hat{b}_i\mathcal{O}_i(\mathbf{k}). \tag{20}$$

The stochastic power spectrum is then estimated directly from the fields constructed with Eqn. 20. Our fiducial figures are made adopting $k_{\rm max} = 0.4\,h{\rm Mpc}^{-1}$, but in Appendix B we show the impact of varying $k_{\rm max}$ on a subset of our results.

The standard parameterization for this solely stochastic contribution can be informed by the symmetries of the occupation procedure and has a form broadly given by the power series (Desjacques et al. 2018; Cabass & Schmidt 2020)

$$P_{\rm err}(k) = \frac{1}{\bar{n}}\left[a_1 + a_2k^2 + \cdots\right], \tag{21}$$

where $\bar{n}$ is the number density of the tracer sample in question. If the stochasticity of the sample arises solely due to Poisson statistics, then this corresponds to $a_1 = 1$ and $a_{2,3,\ldots} = 0$. This is the standard

[1] A similar quantity was the object of study in Hamaus et al. (2010) and Baldauf et al. (2013), however there the ‘stochastic field’ was defined in the Eulerian frame explicitly as $\epsilon(\mathbf{k}) = \delta_h(\mathbf{k}) - b_1\delta_m(\mathbf{k})$.

Poisson shot-noise form of the stochastic power spectrum. As noted in Baldauf et al. (2013) the observed form can differ from the standard Poisson form of Eq. 21 due to halo exclusion and non-linearities in clustering. We will return to this point shortly.

The estimate of $\hat{b}_i$, along with our assumption that the fundamental field-level parameters are constant, may also be used to assess the validity of the model. At the scale k_M where the bias expansion is assumed to break down we would expect the estimates to begin running strongly with scale, as well as using $k_{\rm max} > k_M$ in estimating $P_{\rm err}(k)$ leading to significant changes in its measurement. However, we caution that running of the bias parameters with scale by itself is not necessarily indicative of a breakdown in the bias expansion. The operators as defined in Eqn. 7 are correlated, and our procedure to estimate the bias parameters could be selecting a set $\hat{b}_i$ that runs with scale within a flat sub-region of the likelihood along the principal components of the Hessian, M_{ij}, with subsequent components that are poorly determined due to the the statistical uncertanties arising from a finite box volume. We report measurements carried out with subsets of the whole parameter set $\hat{b}_i$ and discuss the validity of second-order HEFT with constant bias parameters down to small scales in Appendices A and D. The covariance of our bias parameter estimator $\hat{b}_i$ and correlations between bias parameters as a function of $k_{\rm max}$ are quantified and discussed in Appendix C.

2.4 HODs, the halo model, and stochasticity

When combined with the halo model of structure formation (Cooray & Sheth 2002), HOD modelling provides analytic expressions for galaxy observables. The density contrast field is modelled as a mixture of the density of *central* and *satellite* galaxies,

$$\delta_g(\mathbf{k}) = (1-f_{\rm sat})\delta_c(\mathbf{k}) + f_{\rm sat}\delta_s(\mathbf{k}). \tag{22}$$

The power spectrum of galaxies is thus decomposed into contributions that arise from central–central, central–satellite and satellite–satellite correlations

$$P_{gg}(k) = (1-f_{\rm sat})^2 P_{\rm cc}(k) + 2f_{\rm sat}(1-f_{\rm sat})P_{\rm cs}(k) + f_{\rm sat}^2 P_{\rm ss}(k). \tag{23}$$

Each of these spectra, in turn, can be decomposed into contributions that arise from correlations in the *one-halo* regime (that is, galaxies occupying the same halo) and the *two-halo* regime (correlations between galaxies in separate halos). The one-halo contributions to each term can be written as

$$P_{\rm cc}^{(1h)} = \frac{1}{\bar{n}_c} \tag{24}$$

$$P_{\rm cs}^{(1h)} = \frac{1}{\bar{n}_c\bar{n}_s}\int dMn(M)\langle N_{\rm sat}\rangle(M)\langle N_{\rm cen}\rangle(M) \times u(k|M)\theta(\langle N_{\rm sat}\rangle(M)-1) \tag{25}$$

$$P_{\rm ss}^{(1h)} = \frac{1}{\bar{n}_s} + \frac{1}{\bar{n}_s^2}\int dMn(M)\langle N_{\rm sat}\rangle(M)\left[\langle N_{\rm sat}\rangle(M)-1\right] \times u^2(k|M)\theta(\langle N_{\rm sat}\rangle(M)-1), \tag{26}$$

where $n(M)$ is the halo mass function and $u(k|M)$ is the density profile of the halo. The specific functional form of the density profile does not matter for the purposes of investigating stochasticity on the scales considered in this work, however we assume that $\lim_{k\to 0} u(k|M) = 1$ and $\lim_{k\to\infty} u(k|M) = 0$. In the $k \to \infty$ limit the one-halo spectra predict the standard Poisson term

$$P_{gg}^{(1h)}(k) \underset{k\to\infty}{=} \frac{(1-f_{\rm sat})^2}{\bar{n}_c} + \frac{f_{\rm sat}^2}{\bar{n}_s} \tag{27}$$

$$\underset{k\to\infty}{=} \frac{1}{\bar{n}_g}. \tag{28}$$

In the $k \to 0$ limit, the terms that depend on the halo density profile will contribute to the observed spectrum. In this limit, Eqns. 25, 26 respectively contribute to the total power spectrum as

$$2(1-f_{\rm sat})f_{\rm sat}P_{\rm cs}^{(1h)} \underset{k\to 0}{=} \frac{2}{\bar{n}_g^2}\int dMn(M)\langle N_{\rm sat}\rangle(M) \times \langle N_{\rm cen}\rangle(M)\theta(\langle N_{\rm sat}\rangle(M)-1), \tag{29}$$

$$f_{\rm sat}^2 P_{\rm ss}^{(1h)} \underset{k\to 0}{=} \frac{f_{\rm sat}}{\bar{n}_g} + \frac{1}{\bar{n}_g^2}\int dMn(M)\langle N_{\rm sat}\rangle(M) \times \left[\langle N_{\rm sat}\rangle(M)-1\right]\theta(\langle N_{\rm sat}\rangle(M)-1). \tag{30}$$

Including all terms together, the full $k \to 0$ halo model + HOD spectrum is

$$P_{gg}(k) \underset{k\to 0}{=} \frac{1}{\bar{n}_g} + \frac{1}{\bar{n}_g^2}\int dMn(M)\langle N_{\rm sat}(M)\rangle \times \theta(\langle N_{\rm sat}\rangle(M)-1)\left[\langle N_{\rm sat}(M)\rangle + 2\langle N_{\rm cen}\rangle(M) - 1\right]. \tag{31}$$

The second term is always positive, and can be thought of as being related to the variance of the satellite–satellite (or central–satellite) occupation relative to the average density. HODs that have most of their satellite occupation sourced from the high-mass tail of the mass function could then be expected to have this term comparable to the original Poisson prediction, sourcing a considerable amount of super-Poisson stochasticity at large scales.

Eqn. 31 is particularly illuminating in the limit of a monochromatic mass function at a mass M_h, with the expected occupations satisfying $\langle N_{\rm cen}(M_h)\rangle = 1,\ \langle N_{\rm sat}(M_h)\rangle \geq 1:$

$$n(M) = \bar{n}_h\delta^D(M-M_h), \tag{32}$$

where $\delta^D(M-M_h)$ is a Dirac delta function at a fixed halo mass M_h and $\bar{n}_h$ is the number density of haloes at this mass. The integral over the delta function results in the simplified expression

$$P_{gg}(k) \underset{k\to 0}{=} \frac{1}{\bar{n}_g} + \frac{\bar{n}_h\langle N_{\rm sat}\rangle}{\bar{n}_g^2}\left[\langle N_{\rm sat}\rangle(M_h)+1\right]$$

$$\underset{k\to 0}{=} \frac{1}{\bar{n}_g}\left(1 + f_{\rm sat}\left[\langle N_{\rm sat}\rangle(M_h)+1\right]\right), \tag{33}$$

where we used that $\bar{n}_h\langle N_{\rm sat}\rangle/\bar{n}_g = f_{\rm sat}$. Note that using $\bar{n}_g = \bar{n}_h(1+\langle N_{\rm sat}\rangle)$, Eqn. 33 is equivalent to the shot-noise prediction from the number density of haloes $1/\bar{n}_h$. However, we choose to express it in the above form to make the connection to the HOD parameters clearer, as the equivalence is only formally true in the monochromatic limit where all haloes host a central galaxy.

The above expression clarifies the analysis, for example, of Baldauf et al. (2013) where it is noted that increasing the satellite fraction of an HOD can push the observed stochasticity to the super-Poisson regime. However, we see that it is not *solely* the satellite fraction that controls this but instead the interplay between the steepness of the occupation and the satellite fraction that controls the amplitude of this super-Poisson contribution. That is, one could find samples with lower satellite fractions but larger deviations from Poisson shot noise than others, as we show is the case for the `redMaGiC` sample described in § 4.2.

It is clear then that one can observe super-Poisson stochasticity in

$P_{\rm err}$ even when all of the non-linear clustering contributions are taken into account in the model. This same one-halo term was previously noted to enhance the stochasticity of signals expected from line intensity mapping surveys (Schaan & White 2021; Dizgah et al. 2021), and Dizgah et al. (2021) have additionally explored how higher-order bias operators contribute to non-Poissonian noise in the context of line intensity mapping.

In the previous sub-section we also alluded to the importance of including the effect of *halo exclusion* in the analysis of stochastic power spectra. In a sense, halo exclusion is the opposite effect to the previously discussed one-halo enhancement of stochasticity. Whereas the enhancement comes from multiple satellites contributing to self pairs at the same pixel, halo exclusion leads to a suppression of stochasticity due to the minimum distance scale imposed on halo correlations. In the simplified $k \to 0$ case with monochromatic mass function, the effect of exclusion is to decrease large-scale stochasticity.

Following Baldauf et al. (2013), one can construct a toy model for exclusion by imposing a break in the correlation function at a radius $R_{\rm exc}$:

$$\xi^{(d)}(r) = \begin{cases} \xi^{(c)}(r) \text{ if } r \geq R_{\rm exc}, \\ -1 \text{ if } r < R_{\rm exc}, \end{cases} \tag{34}$$

where $\xi^{(c)}(r)$ is the non-excluded two-point correlation function of the sample. The power spectrum under the presence of exclusion is then

$$P^{(d)}(k) = -\int_0^R d^3r j_0(kr) + \int_R^\infty d^3r \xi^{(c)}(r) j_0(kr), \tag{35}$$

$$P^{(d)}(k) = P^{(c)}(k) - V_{\rm exc}\Bigg(W_R(k) + \int \frac{d^3q}{(2\pi)^3} P(k) W_R(|\mathbf{k} - \boldsymbol{q}|)\Bigg), \tag{36}$$

where $j_0(kr)$ zero-order spherical Bessel function, $W_R(k)$ is the Fourier transform of the top-hat window function and $V_{\rm exc}$ is the exclusion volume. For illustrative purposes here we take the exclusion volume to be $V_{\rm exc} = 4\pi R_h^3/3$ where R_h is the spherical radius of a halo of mass M_h. In the $k \to 0$ limit both the window function and convolution term contribute with an overall negative amplitude to the total signal.

Adding both the enhancement and exclusion terms together we find

$$P_{gg}(k) \underset{k\to 0}{=} \frac{1}{\bar{n}_g}\left(1 + f_{\rm sat}\left[\langle N_{\rm sat}\rangle(M_h) + 1\right]\right) - \frac{458}{1+\delta}\left(\frac{M_h}{10^{13}{\rm M}_\odot h^{-1}}\right)\left(\frac{\rm Mpc}{h}\right)^3, \tag{37}$$

where δ represents the overdensity of a halo relative to the background density of the Universe, as we have re-written the exclusion volume in terms of the halo mass instead. While it is standard to take $1 + \delta = 200$ when treating haloes, Baldauf et al. (2013) have shown that for exclusion radii at late times the exclusion term is better fit by assuming $\delta \approx 30$[2]. Nevertheless we may infer that for the case of an HOD, the properties of the galaxy sample control the super-Poissonianity, whereas the host haloes set the degree of sub-Poissonianity. This interplay between enhancement and exclusion will be crucial in interpreting the results of our subsequent numerical experiments.

3 SIMULATIONS AND SAMPLES

We use the `Aemulus` (DeRose et al. 2019) suite of N-body simulations to populate three different samples of red galaxies using HODs. `Aemulus` is composed of 40+35 dark-matter-only N-body simulations with size $L_{\rm box} = 1050$ Mpc h^{-1} and $N = (1400)^3$ particles. They were designed to serve as a training set for the construction of emulators of non-linear summary statistics that could be readily applied to analysis of modern cosmological data sets. To date, these include emulators for the mass function (McClintock et al. 2019b), halo bias (McClintock et al. 2019a), and the correlation function of HOD galaxies (Zhai et al. 2019). We specifically choose a subset of the `Aemulus` suite that strikes a balance between being close to Planck's ΛCDM constraints but also offering repeated realizations so we may make statistical assessments of our results. Specifically, we pick the cosmology from the test suite that is closest to ΛCDM by relative differences. The parameters of this cosmology are $\Omega_c = 0.27$, $\Omega_b = 0.05$, $h = 0.6673$, $n_s = 0.973$, $\sigma_8 = 0.798$, $N_{\rm eff} = 3.2$, $w = -0.9$.

We use the public halo catalogs associated with these boxes and apply an HOD prescription to populate them with galaxies across the redshift range $z = [0.25, 1.0]$, which corresponds to most of the redshift range spanned by the samples of galaxies we wish to consider[3]. We choose previously published HODs that describe three samples of red galaxies that are used for lensing and clustering analyses by current and next-generation surveys. These are: the `redMaGiC` sample used in the Dark Energy Survey (Rozo et al. 2016; Clampitt et al. 2016; Zacharegkas et al. 2021); luminous red galaxies from BOSS and eBOSS (Zhai et al. 2017); and DESI-like luminous red galaxies from Zhou et al. (2020). The parametric forms adopted are all variations of the standard Zheng et al. (2005, 2007) HOD discussed in § 2.1 and displayed in Eqns. 1 and 2. For all of our samples in question, we note that their auto-power spectra become dominated by shot noise at similar scales, at around $k \sim 0.4\,h{\rm Mpc}^{-1}$. Since we use field-level information and not just P_{gg} in our analysis, we believe this justifies our fiducial choice of $k_{\rm max}$ in the analysis below.

The samples we mock up are representative of those observed by leading galaxy surveys. While the HOD parameters adopted are not calibrated to the specific cosmology we use, the impact of this difference is sub-leading for the kind of analysis we wish to carry out for the BOSS and DESI samples, as their derived parameters we obtain are statistically consistent with those reported in the original publications. For the case of the DES Y3 HOD presented in Zacharegkas et al. (2021), when applied to our fiducial cosmology, we find significantly different satellite fractions, mean halo masses, and number densities compared to the published results.

The differences between these results and ours arise from differences in the high-mass tail of the halo mass function of our simulations compared to the Tinker et al. (2008) mass function adopted in the original publication at their fiducial cosmology. The large satellite slopes ($\alpha > 1.6$ for all bins) and suppression of occupation until high masses leads to significant differences in derived parameters for the galaxy samples in our simulations.

To alleviate these discrepancies, we have slightly tuned the parameters of the DESY3 `redMaGiC` HOD in a way that recovers the satellite fractions, mean host halo masses, and number densities reported in Zacharegkas et al. (2021) in our simulations. We show the resulting HOD compared to the original in Fig. D2. The largest

[2] The expression as presented in Eqn. 37 neglects the amplitude of the $k \to 0$ contribution from the convolution in Eqn. 36, and a lower δ would correspond to higher amplitude from this contribution.

[3] This is done using `simplehod`, available at https://github.com/bccp/simplehod.

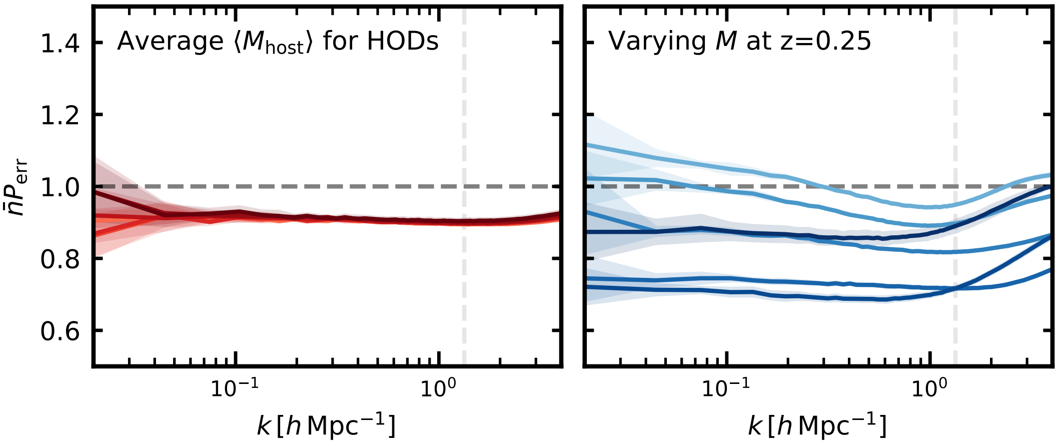

Figure 2. Error power spectra for two sets of halo samples. The power spectra are divided by the Poisson prediction, $\frac{1}{\bar{n}}$, to highlight deviations from the standard expectation. The shaded regions correspond to one standard deviation from repeating this estimate for the 5 `Aemulus` boxes that belong to our reference cosmology. The vertical dashed line corresponds to the inverse grid size, $L_{\rm cell}^{-1} \approx 1.33\,h{\rm Mpc}^{-1}$. *Left panel:* halo mass bins, as described in Eqn. 38, that encompass the average host halo mass for the HOD samples we consider in this paper in each snapshot, with the lightest (darkest) shade corresponding to the highest (lowest) redshift snapshots respectively. *Right panel:* Varying halo mass in bins of 0.5 dex width for the snapshot at redshift $z = 0.25$. The lightest (darkest) shade correspond to $\log M \in [12, 12.5]$ ($\log M \in [14.5, 15.0]$) respectively.

change is a reduction in α across all redshift bins, as well as a slight boost to the halo occupations at lower masses.

We interpret our results on galaxy stochasticity by analyzing the dependence of stochastic power spectra on three derived HOD parameters: satellite fraction $f_{\rm sat}$, galaxy number density $\bar{n}$ and average host halo mass $\log_{10}\langle M_{\rm host}\rangle$. The results we find at each redshift are reported in Tables 1, 2 and 3. The corresponding galaxy catalogs span an order of magnitude in density and halo mass, with $\bar{n} \in [1.39, 9.75] \times 10^{-4}\,[h{\rm Mpc}^{-1}]^3$ and $\log_{10}\langle M_{\rm host}/(h^{-1}M_{\odot})\rangle \in [12.95, 13.64]$. We plot the galaxy occupations of our HOD samples for each snapshot in Fig. 1. This figure highlights a few key differences in how our mock galaxies populate their haloes, which we use to interpret our findings on stochasticity. Most notably, the Y3-like `redMaGiC` HOD has significantly lower occupations than the other samples until higher masses. The `redMaGiC` galaxies have high number density, and this deficiency at low mass is made up by having a larger slope in the satellite occupation. This is noticeable in the bottom three panels of Fig. 1. On the other hand, the BOSS and DESI HODs share similar derived parameters with the largest difference being the significantly higher number density of DESI galaxies.

Both the DESI-like LRGs and `redMaGiC` galaxies have constrained the redshift-dependence of the fundamental HOD parameters. For the BOSS sample, we fix its parameters as a function of redshift. However, as the DESI LRG sample infers little redshift dependence in its own parameters we believe this is a suitable approximation for the intent of our analysis.

4 RESULTS

4.1 Sub-poisson stochasticity of massive halos

The stochastic power spectrum of dark matter haloes has been the subject of considerable past study (Hamaus et al. 2010; Baldauf et al. 2013). Notably, Schmittfull et al. (2019) measured these statistics for halo samples that collectively spanned four orders of magnitude in halo mass. They observed that less massive haloes tend to exhibit super-Poisson stochasticity[4], which trends toward sub-Poisson with increasing mass. These results were obtained using a different model for the field-level tracer density, and so in this section we report our results for the halo samples that host the galaxies using the second-order HEFT model that is the subject of study of this paper. While Schmittfull et al. (2019) used a third-order Eulerian bias model, we work with the second-order Lagrangian bias expansion of Eqn. 4, which has been shown in the past to capture the clustering statistics of halos in the mass ranges under consideration (Abidi & Baldauf 2018). Another key difference is that we use the fully non-linear Lagrangian displacement field as determined from the N-body simulation, whereas Schmittfull et al. (2019) only includes Zel'dovich displacements in their shifted operators.

From Tables 1, 2, 3 we can infer that for any given snapshot the average host halo mass spans approximately 0.2 dex for all of the HOD samples we have constructed. At a given snapshot we select for halos in the mass bin given by

$$13.1 + \frac{2}{3}(1-z) \leq \log M < 13.3 + \frac{2}{3}(1-z), \tag{38}$$

where z is the redshift of the snapshot. These mass ranges encompass the average halo masses of our HOD samples.

We proceed to measure $P_{\rm err}(k)$ for these halo samples using the procedure outlined in § 2.3, and report our results in the left panel of Fig. 2. For all snapshots under consideration we observe a slight but significant sub-Poisson signal, in concordance with previous results in the literature for comparable halo masses. For our second-order basis, the amplitude of the deviation from the Poisson expectation is approximately 10 per-cent, staying approximately constant with redshift. Qualitatively, this can be understood by noting that we simultaneously increase the mean halo mass across snapshots, which increases the expected exclusion signal, while simultaneously going down the mass function to less dense halo samples, which increases the one-halo enhancement expected.

We also observe the $P_{\rm err}(k)$ we have measured are approximately

[4] We define super and sub-Poisson stochasticity as the regimes where $\bar{n}P_{\rm err} > 1$ and $\bar{n}P_{\rm err} < 1$ respectively.

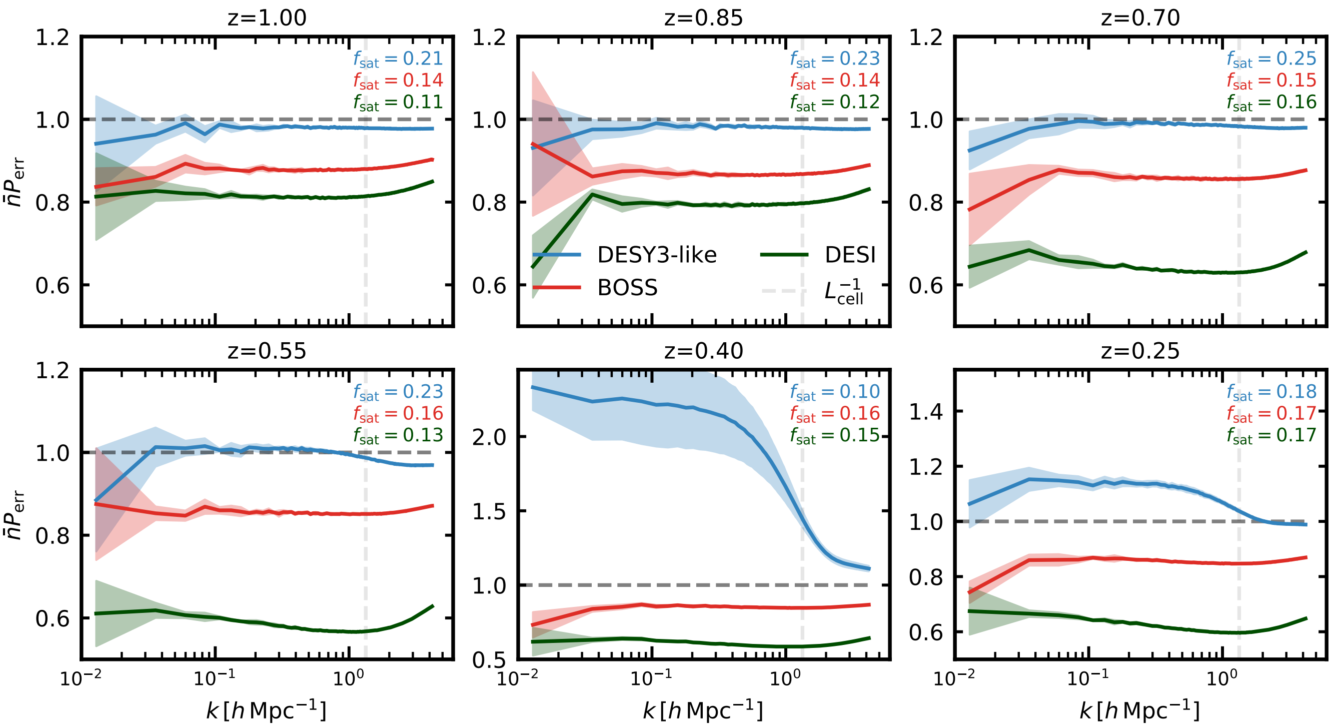

Figure 3. Error power spectra for the different samples, for the standard Lagrangian bias basis, using bias parameters inferred from our variance-minimization procedure assuming $k_{\rm max} = 0.4\,h{\rm Mpc}^{-1}$. The envelopes are the scatter found from the five independent realizations from the Aemulus suite at this cosmology. The vertical dashed line corresponds to the inverse grid size, $L_{\rm cell}^{-1} \approx 1.33\,h{\rm Mpc}^{-1}$. Note that at low redshifts the y-axis ranges are altered to accommodate for the large amount of super-Poisson stochasticity observed in the `redMaGiC` sample.

scale independent out to scales comparable to the inverse of the grid spacing, at which point we observe an uptick due to the impact of deconvolving with the window function used to assign objects to the grid. The scale independence of the $P_{\rm err}(k)$ indicate second-order HEFT is a suitable forward model for the halo sample under consideration. These initial results highlight consistency between hybrid EFT approaches to estimating $P_{\rm err}(k)$ and others in the literature, such as the shifted-Eulerian operator basis of Schmittfull et al. (2019).

As an additional test of second-order HEFT, we also look at the evolution of halo stochasticity as a function of mass, at fixed redshift. Specifically, we select six broad bins in halo mass, from $10^{12}h^{-1}M_\odot$ to $10^{15}h^{-1}M_\odot$ with width of 0.5 dex. We measure their stochastic power spectra and show them in the right panel of Fig. 2. The purpose of these measurements are twofold: they help us verify the validity of the approximate exclusion treatment as described in § 2.4 and also how well can second-order HEFT describe different halo samples of varying mass bins. If halo exclusion scales roughly linearly with halo mass, as shown in Eqn. 37, then in the power-law regime of the mass function, the exclusion signal should get larger with mass bin. However, once the mass bin reaches the exponential cut-off of the mass function, the number density should fall off faster than M and the stochastic power spectrum should begin to revert to Poisson. This is precisely the behavior we observe, where the line corresponding to the highest mass bin at $M \in [10^{14.5}, 10^{15}]\,h^{-1}M_\odot$ has its stochasticity closer to Poisson than the previous bin.

We also note that the lowest mass bins in the right panel of Fig. 2 show a slight tendency toward super-Poissonianity but also with significant scale dependence. These results ostensibly show that lower-mass haloes are potentially more sensitive to neglected higher-order bias contributions that manifest themselves as super-Poissonianities atop the exclusion signal. However, a quantitative assessment of the impact of higher-order operators is postponed to future work, because for the halo masses relevant to our HODs the stochastic power spectra of host haloes are relatively scale independent.

4.2 The stochasticity of red survey galaxies

Having established a baseline for the degree of deviation from Poisson stochasticity in our host halo samples, we can now turn to analyzing the mock galaxy samples we have created. In Fig. 3 we report the redshift-dependent product $\bar{n}(z)P_{\rm err}(k, z)$ for the three red galaxy samples we consider, again obtained using the procedure outlined in § 2.3. If the impact of enhancement and exclusion on these galaxies were negligible, we would expect $\bar{n}P_{\rm err}(k) = 1$.

We find that for most of the above galaxy samples, the stochastic power spectra are approximately scale independent across the range of scales considered. Even though all of our mock galaxies populate comparable dark matter haloes, whose stochasticity is significantly sub-Poisson, we in fact observe that the three HOD parameterizations adopted exhibit markedly different levels of stochasticity. Once again, this points to the suitability of second-order Lagrangian bias in describing these samples[5]. At high redshifts, the `redMaGiC` sample exhibits a slight sub-Poisson trend but over time has $\bar{n}P_{\rm err}$ evolve to

[5] Additionally, we note that for the case of `redMaGiC` galaxies at $z = 0.55$ the inferred bias parameters from the field-level fitting procedure are consistent with those obtained by fitting galaxy–galaxy and galaxy–matter spectra in Kokron et al. (2021), providing additional validation of our methodology.

strongly super-Poisson, with a strong scale-dependence that reverts to close to the Poisson measure at small scales. This result can be explained by the analysis carried out in § 2.4, combined with knowledge of the HOD of the Y3 `redMaGiC` sample. From Eqn. 33, the steeper the satellite occupation at a given mass, the larger its contribution is to the one-halo enhancement that drives the stochasticity to be super-Poisson. From Fig. 1 and Table 1 the galaxy samples at $z = 0.4$ and $z = 0.25$ have both simultaneously suppressed occupation until high masses, and then significantly larger slopes than the other samples to try and fit to the observed number density. The panels that exhibit this super-Poisson stochasticity have the highest slopes relative to the other bins, and the reversion to close to $nP \sim 1$ is indicative of the halo density profile decaying at small scales as expected. Thus, we can conclude that the lower-redshift `redMaGiC` samples exhibit a super-Poisson signal.

In fact, this super-Poissonian aspect of `redMaGiC` stochasticity has been previously observed in the literature (Friedrich et al. 2018; Gruen et al. 2018; Friedrich et al. 2021). We also note the snapshot at $z = 0.40$, corresponding to the second `redMaGiC` lens bin, is anomalous in its behavior. The degree of super-Poisson stochasticity is markedly higher, and the measurements themselves are quite noisy. The second lens bin in DES Y3 has by far the highest slope α in its satellite distribution. This means that a small number of very massive haloes host satellites. The number of these high mass haloes varies significantly between our five realizations, which explains the increased scatter we observe only for this snapshot for the `redMaGiC` HOD.

Turning to the sample of BOSS-like LRGs, we observe mild disagreement with the Poisson prediction across a wide range of scales for all of the snapshots under consideration. The snapshots possess a slightly sub-Poisson stochastic power spectrum, however to a mild degree when compared to the other two samples – on the order of 10 per-cent. In this case, numerical estimates of the amplitude of one-halo enhancement for BOSS galaxies through Eqn. 31 show that it is entirely negligible, as at no masses in our halo catalog does this HOD have $\langle N_{\rm sat}\rangle \geq 1$. Given that the satellite occupation is sub-leading at all masses, this is not surprising. The mild sub-Poissonianity we observe corresponds solely to the imprint of halo exclusion on this sample.

The DESI LRGs, on the other hand, exhibit significant sub-Poisson stochasticity. This is despite the fact that their derived parameters are quite similar to that of BOSS; both satellite fractions and host halo masses are very comparable on a per-snapshot basis. In fact, the occupations for both samples are quite similar except for the fact that the DESI occupation is almost multiplicatively offset from the BOSS one. If both samples are hosted within similar haloes, then, we expect their exclusion signatures to be comparable. However, as the DESI galaxies are significantly denser, the relative size of the exclusion signal compared to the galaxy density is larger (see Eqn. 37) and we observe a more sub-Poisson stochasticity as a result.

The significantly higher number density of the DESI HOD could imply additional bias terms not included in our model that could contaminate our estimate of $P_{\rm err}(k)$. This, then, would explain the stronger deviations from a flat noise curve at large scales for the DESI sample relative to other samples. However, as third-order bias models introduce a significant number of new parameters and our simulations are of limited volume, we defer the investigation of the impact of including higher-order HEFT operators to future work.

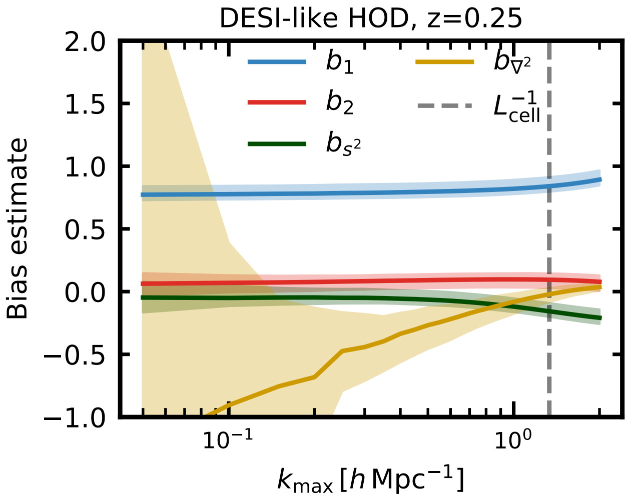

Figure 4. Estimates of second-order Lagrangian bias parameters obtained from $N = 1500$ mock galaxy samples from the DESI-like LRG posteriors, as a function of maximum wavenumber $k_{\rm max}$ used in the fit. The solid line shows the median bias parameter and the shaded regions the 98% quantiles. The vertical dashed line corresponds to the inverse grid size, $L_{\rm cell}^{-1} \approx 1.33\,h{\rm Mpc}^{-1}$.

4.3 What range of deviations from Poisson stochasticity is allowed when conditioned on an HOD?

The numerical experiments carried out in the previous sub-section are concerned with the expected degree of stochasticity subject to a specific set of HOD parameters. In principle, one would expect that degeneracies in the HOD lead to a population of HOD parameters that can equally capture the statistical properties of a galaxy sample. However, for any given set of parameters we will find a value of stochasticity, and thus we should also expect that as a result any given sample is consistent with a range of large-scale stochastic behavior.

To assess this, we take a deeper look at the set of parameters consistent with the DESI-like HOD of Zhou et al. (2020) for their lowest redshift bin. We take 300 random samples from the post burn-in MCMC chains run in the publication and use the procedure laid out in § 2.3 to estimate the low-k stochasticity for that point in the chain. For each sample, we populate galaxies across all five boxes at our fiducial cosmology, leading to a total of $N = 1500$ sets of bias parameters and measurements of $\bar{n}P_{\rm err}$ that to some extent also includes a contribution from cosmic variance.

In Fig. 4 we report the estimated bias parameters that we obtain from this procedure. Interestingly, the first three parameters b_1, b_2, b_{s^2} are approximately constant out to very small scales, beyond the fiducial $k_{\rm max} = 0.4\,h{\rm Mpc}^{-1}$ value we use, and their 98% quantiles show very little spread in the estimates. For b_{∇^2} we find a significantly larger scatter, that decreased as we increase $k_{\rm max}$, and a median value that also seems to run more steeply with scale than other parameters. The larger scatter in b_{∇^2} seems to indicate this field possesses a larger correlation between realization-to-realization noise. This is somewhat expected given the origin of this operator, as it arises from integrating out the untameable effects of small scales.

As an additional test, we investigate the posterior distribution of stochasticities consistent with DESI-like LRGs, $P(\bar{n}P_{\rm err})$ sliced through different scales k_*. All of the $\bar{n}P_{\rm err}$ are measured with bias parameters estimated at $k_{\rm max} = 0.4\,h{\rm Mpc}^{-1}$, our fiducial choice for this publication. The resulting distributions are shown in Fig. 5.

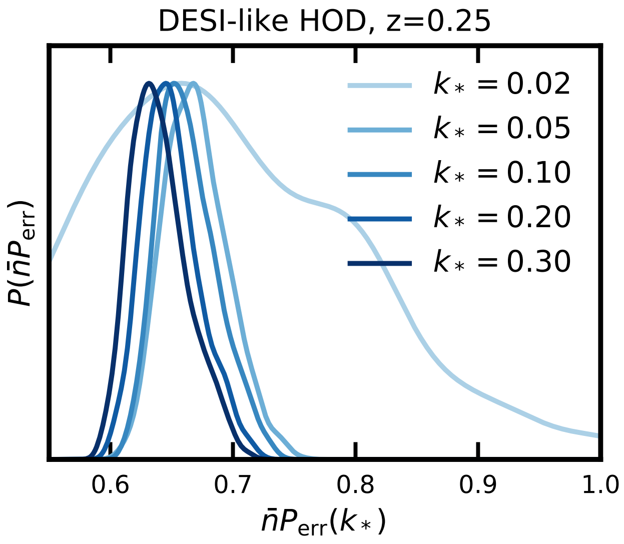

Figure 5. Distributions of stochastic power spectra $\bar{n}P_{\rm err}(k)$ obtained from $N = 1500$ mock galaxy samples from the DESI-like LRG posteriors. Each curve represents a different slice of the stochastic spectrum evaluated at a different wavenumber, in units of $h\mathrm{Mpc}^{-1}$.

Slicing through the stochastic spectra at large scales shows a wide distribution of values. However, this is not surprising as our boxes have limited volume and scales such as $k \sim 0.02\,h\mathrm{Mpc}^{-1}$ will still be affected by cosmic variance. As we probe the distribution toward quasi-linear scales we see that the allowed range of stochasticity remains relatively peaked around a single value, while the value of the peak itself shifts very slightly.

Thus, we can expect that the range of HODs consistent with a given data set will have relatively similar large-scale stochasticities, more tightly spread than what is observed as we look across HODs that describe different galaxy samples.

4.4 Priors on red galaxy stochasticity

The results presented in this section highlight that analyses of red galaxy samples using second-order Lagrangian bias models can safely control the impact of stochastic contributions to model power spectra. The detections of deviations from Poisson stochasticity are significant, and their amplitudes span the range of being 60–140 per-cent of the Poisson expectation, with the exception of the anomalous $z = 0.4$ `redMaGiC` sample. Cosmological parameter inference carried out with comparable models and samples of galaxies can adopt informative priors on the degree of stochasticity given certain knowledge about the selection of the sample. Priors can be adopted as being of the form

$$P(a_1) = \mathcal{N}(1, 0.4)\,, \qquad (39)$$

where a_1 is the scale-independent stochastic amplitude from Eqn. 21. This prior encompasses the whole range of deviations from Poisson stochasticity observed across the redshift range $0.25 \leq z \leq 1$ within $1-\sigma$, with the exception of the $z = 0.40$ bin of the `redMaGiC` sample, which possesses an anomalously high super-Poisson stochasticity. If the sample is constrained to higher redshifts $z > 0.55$ then setting $\sigma_{a_1} \approx 0.3$ gives a tighter prior that still captures the observed deviations from Poisson stochasticity to within $1 - \sigma$.

z	$f_{\rm sat}$	$10^4\bar{n}$	$\log_{10}\langle M_{\rm host}\rangle$
1.00	0.21	3.66	12.95
0.85	0.23	3.99	13.06
0.70	0.25	4.33	13.18
0.55	0.23	9.70	13.24
0.40	0.10	9.22	13.51
0.25	0.18	9.75	13.51

Table 1. Derived parameters for the `redMaGiC` HOD.

z	$f_{\rm sat}$	$10^4\bar{n}$	$\log_{10}\langle M_{\rm host}\rangle$
1.00	0.14	1.39	13.32
0.85	0.14	1.62	13.37
0.70	0.15	1.88	13.43
0.55	0.16	2.15	13.49
0.40	0.16	2.44	13.55
0.25	0.17	2.74	13.61

Table 2. Derived parameters for the `BOSS` LRG HOD.

z	$f_{\rm sat}$	$10^4\bar{n}$	$\log_{10}\langle M_{\rm host}\rangle$
1.00	0.11	2.17	13.29
0.85	0.12	2.53	13.36
0.70	0.16	6.69	13.39
0.55	0.13	5.88	13.50
0.40	0.15	7.57	13.55
0.25	0.17	8.50	13.64

Table 3. Derived parameters for the `DESI` LRG HOD.

5 CONCLUSIONS

In this work we have applied a field-level model for biased tracers to study the stochastic contribution to the tracer–matter connection. Specifically, we focused on the stochasticity of both halo samples and example samples of red galaxies that are hosted in these halos, using three different forms of halo occupation distributions. The HODs we adopted have been previously used to describe the clustering statistics of bright red galaxy samples from three different galaxy surveys.

We use second-order Hybrid Effective Field Theory to model the distribution of these galaxies at the field level. We developed an estimator that can reliably infer the maximum-likelihood bias parameters from the field-level information and applied it to infer the stochastic power spectra of these galaxy samples. We proceeded to compare our results to the commonly used assumption of Poisson stochasticity, framing our findings within the context of the well-established halo model of large-scale structure. Our findings can be summarized as follows:

(i) Almost all of the HODs used deviate from the constant, Poisson prediction of shot-noise by at most 40%. This implies tight priors can be used on the expected stochasticity from these kinds of samples. This will reduce degeneracies between stochasticity, bias parameters, and cosmological parameters.

(ii) The form of the HOD is connected to the degree of non-Poisson stochasticity. Specifically, HODs with a high variance in their satellite occupation, or equivalently large slopes, will have super-Poisson stochasticity due to one-halo enhancements to the signal.

(iii) Very dense galaxy samples with negligible one-halo enhancements will instead have their large-scale stochasticity dominated by the effects of halo exclusion. As a result they will tend to have sub-Poisson stochasticity.

(iv) The stochastic power spectra of galaxy samples can be either sub- or super-Poisson, despite being hosted in similar halo populations whose own stochasticities are consistently sub-Poisson.

These findings showcase the synergistic gains to be obtained from jointly studying models of the galaxy–halo connection with different models for modeling large-scale structure and bias. The combination of an empirical parameterization with a Lagrangian bias model allowed us to quantify the degree of stochasticity of these galaxies and place informative priors on this that will help future analyses of galaxy survey data.

The 40% priors highlighted in this publication highlight the wealth of stochasticities that red galaxies samples can exhibit. They are directly related to the fundamental scatter in allowed stochasticities of the physical process underlying galaxy formation. This implies narrowing the priors without making further assumptions about the sample will be challenging. Efforts to tighten these priors would require going beyond our analysis and carrying out more systematic studies for an intended sample. For example, the analysis of § 4.3, where we found that conditioned on HODs consistent with DESI-like red galaxies, the spread of stochasticities was significantly tighter than 40%.

The techniques to estimate field-level residual maps of tracer stochasticity developed here can be extended further. While we have limited ourselves to studying the auto-spectra of our residual maps, there are several avenues of investigation that can be pursued and are highly relevant to modern galaxy surveys. For example, one could characterize the cross-spectra of stochasticity between different tracer samples such as galaxies and clusters, whose cross-correlations are a promising future probe of cosmology (To et al. 2021) but whose cross-stochasticity is poorly understood. As analyses of higher N-point correlation functions become more prominent, the residual maps created here could shed new insights into the significantly more complicated case of stochasticity in higher N-point functions as well as be used to better understand when a second-order hybrid EFT model is no longer sufficient in describing a sample.

This study also points to the feasibility of a broader program of learning about the galaxy–halo connection using bias models and empirical models in tandem, at the field level, to gain insights into how tracers relate to the distribution of dark matter at large. The analysis carried out here can be extended to different forms of tracer samples. For example, one could study the relationship between assembly bias and the Lagrangian bias parameters within HEFT. First steps in this direction, albeit for a different forward model, were carried out in Lazeyras et al. (2021). HEFT could also be used to study HODs of different samples of galaxies that are not captured by the standard Zheng et al. (2005) form, and place priors on the bias parameters expected in that case. Similar work, for galaxies in `IllustrisTNG` (Nelson et al. 2021), has been explored recently (Barreira et al. 2021). As this work was being finalized, Zennaro et al. (2021b) used hybrid EFT models precisely in this way to study the distributions of second-order Lagrangian bias for samples of galaxies populated using an extended sub-halo abundance matching scheme. Connecting efforts to measure bias parameters within different bias models to each other is another important goal that will aid in characterizing the relationship between empirical models, bias models, and their applicability to optimizing analyses of data collected by galaxy surveys.

ACKNOWLEDGEMENTS

We would like to thank Rongpu Zhou for making the DESI-like LRG posteriors available, Georgios Zacharegkas for helpful discussions about the Y3 `redMaGiC` HOD, and Sihan Yuan for comments on a draft of this work. We are grateful to the `Aemulus` collaboration for making the simulation suite used here publicly available. This work was supported in part by U.S. Department of Energy contracts to SLAC (DE-AC02- 76SF00515). NK acknowledges support from the Gerald J. Lieberman fellowship. SC and MW acknowledge support from the National Science Foundation and Department of Energy. JD is supported by the Chamberlain fellowship at Lawrence Berkeley National Laboratory. This research has made use of NASA's Astrophysics Data System and the arXiv preprint server.

Some of the computing for this project was performed on the Sherlock cluster at Stanford. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results.

Calculations and figures in this work have been made using `nbodykit` (Hand et al. 2018), `GetDist` (Lewis 2019), and the SciPy Stack (Harris et al. 2020; Virtanen et al. 2020; Hunter 2007).

DATA AVAILABILITY

The data underlying this article are available in the `Aemulus` Project's website. Access to the HEFT products is available on reasonable request.

REFERENCES

Abidi M. M., Baldauf T., 2018, JCAP, 07, 029

Aghamousa A., et al., 2016, arXiv e-prints

Angulo R. E., Zennaro M., Contreras S., Aricò G., Pellejero-Ibañez M., Stücker J., 2021, The BACCO Simulation Project: Exploiting the full power of large-scale structure for cosmology (`arXiv:2004.06245`)

Bagla J. S., 2005, Curr. Sci., 88, 1088

Baldauf T., Seljak U., Smith R. E., Hamaus N., Desjacques V., 2013, Physical Review D, 88

Baldauf T., Codis S., Desjacques V., Pichon C., 2016, Monthly Notices of the Royal Astronomical Society, 456, 3985–4000

Banerjee A., Kokron N., Abel T., 2021, Modeling Nearest Neighbor distributions of biased tracers using Hybrid Effective Field Theory (`arXiv:2107.10287`)

Barreira A., Lazeyras T., Schmidt F., 2021, Galaxy bias from forward models: linear and second-order bias of IllustrisTNG galaxies (`arXiv:2105.02876`)

Behroozi P. S., Wechsler R. H., Conroy C., 2013, The Astrophysical Journal, 770, 57

Behroozi P., Wechsler R. H., Hearin A. P., Conroy C., 2019, Monthly Notices of the Royal Astronomical Society, 488, 3143–3194

Berlind A. A., Weinberg D. H., 2002, The Astrophysical Journal, 575, 587–616

Cabass G., Schmidt F., 2020, Journal of Cosmology and Astroparticle Physics, 2020, 051–051

Carlson J., Reid B., White M., 2012, Monthly Notices of the Royal Astronomical Society, 429, 1674–1685

Chen S.-F., Vlah Z., White M., 2020, Journal of Cosmology and Astroparticle Physics, 2020, 062–062

Chen S.-F., Vlah Z., Castorina E., White M., 2021, Journal of Cosmology and Astroparticle Physics, 2021, 100

Chue C. Y. R., Dalal N., White M., 2018, Journal of Cosmology and Astroparticle Physics, 2018, 012–012

Clampitt J., et al., 2016, Monthly Notices of the Royal Astronomical Society, 465, 4204–4218

Collaboration D., et al., 2021, Dark Energy Survey Year 3 Results: Cosmological Constraints from Galaxy Clustering and Weak Lensing (`arXiv:2105.13549`)

Cooray A., 2006, Monthly Notices of the Royal Astronomical Society, 365, 842–866

Cooray A., Sheth R., 2002, Physics Reports, 372, 1–129

Dalal N., White M., Bond J. R., Shirokov A., 2008, Astrophys. J., 687, 12

Dawson K. S., et al., 2012, The Astronomical Journal, 145, 10

Dawson K. S., et al., 2016, The Astronomical Journal, 151, 44

DeRose J., et al., 2019, Astrophys. J., 875, 69

DeRose J., Becker M. R., Wechsler R. H., 2021, Modeling Redshift-Space Clustering with Abundance Matching (arXiv:2105.12104)

Desjacques V., Jeong D., Schmidt F., 2018, Phys. Rept., 733, 1

Di Valentino E., et al., 2021, Classical and Quantum Gravity, 38, 153001

Dizgah A. M., Nikakhtar F., Keating G. K., Castorina E., 2021, Precision Tests of CO and [CII] Power Spectra Models against Simulated Intensity Maps (arXiv:2111.03717)

Friedrich O., et al., 2018, Physical Review D, 98

Friedrich O., Halder A., Boyle A., Uhlemann C., Britt D., Codis S., Gruen D., Hahn C., 2021, The PDF perspective on the tracer-matter connection: Lagrangian bias and non-Poissonian shot noise (arXiv:2107.02300)

Gao L., Springel V., White S. D., 2005, Mon. Not. Roy. Astron. Soc., 363, L66

Garrison L. H., Eisenstein D. J., Ferrer D., Tinker J. L., Pinto P. A., Weinberg D. H., 2018, The Astrophysical Journal Supplement Series, 236, 43

Ginzburg D., Desjacques V., Chan K. C., 2017, Physical Review D, 96

Gruen D., et al., 2018, Physical Review D, 98

Hadzhiyska B., Bose S., Eisenstein D., Hernquist L., Spergel D. N., 2020, Monthly Notices of the Royal Astronomical Society, 493, 5506–5519

Hadzhiyska B., García-García C., Alonso D., Nicola A., Slosar A., 2021a, Hefty enhancement of cosmological constraints from the DES Y1 data using a Hybrid Effective Field Theory approach to galaxy bias (arXiv:2103.09820)

Hadzhiyska B., Tacchella S., Bose S., Eisenstein D. J., 2021b, Monthly Notices of the Royal Astronomical Society, 502, 3599–3617

Hamaus N., Seljak U., Desjacques V., Smith R. E., Baldauf T., 2010, Physical Review D, 82

Hand N., Feng Y., Beutler F., Li Y., Modi C., Seljak U., Slepian Z., 2018, The Astronomical Journal, 156, 160

Harris C. R., et al., 2020, Nature, 585, 357–362

Hearin A. P., Zentner A. R., van den Bosch F. C., Campbell D., Tollerud E., 2016, Monthly Notices of the Royal Astronomical Society, 460, 2552–2570

Heitmann K., Lawrence E., Kwan J., Habib S., Higdon D., 2013, The Astrophysical Journal, 780, 111

Heymans C., et al., 2021, Astronomy & Astrophysics, 646, A140

Hockney R., Eastwood J., 1988, Computer Simulation Using Particles. CRC Press, https://books.google.com/books?id=nTOFkmnCQuIC

Hunter J. D., 2007, Computing in Science Engineering, 9, 90

Ivezić v., et al., 2019, Astrophys. J., 873, 111

Knabenhans M., et al., 2019, Mon. Not. Roy. Astron. Soc., 484, 5509

Kokron N., DeRose J., Chen S.-F., White M., Wechsler R. H., 2021, Monthly Notices of the Royal Astronomical Society, 505, 1422–1440

Krolewski A., Ferraro S., White M., 2021, Cosmological constraints from unWISE and Planck CMB lensing tomography (arXiv:2105.03421)

Kuhlen M., Vogelsberger M., Angulo R., 2012, Phys. Dark Univ., 1, 50

Lazeyras T., Barreira A., Schmidt F., 2021, Assembly bias in quadratic bias parameters of dark matter halos from forward modeling (arXiv:2106.14713)

Lewis A., 2019, GetDist: a Python package for analysing Monte Carlo samples (arXiv:1910.13970)

Mandelbaum R., et al., 2018, The LSST Dark Energy Science Collaboration (DESC) Science Requirements Document (arXiv:1809.01669)

Mansfield P., Kravtsov A. V., 2020, Mon. Not. Roy. Astron. Soc., 493, 4763

Mao Y.-Y., Zentner A. R., Wechsler R. H., 2018, Mon. Not. Roy. Astron. Soc., 474, 5143

Matsubara T., 2008a, Physical Review D, 77

Matsubara T., 2008b, Phys. Rev. D, 78, 083519

McClintock T., et al., 2019a, The Aemulus Project IV: Emulating Halo Bias (arXiv:1907.13167)

McClintock T., et al., 2019b, Astrophys. J., 872, 53

Modi C., Chen S.-F., White M., 2020, Mon. Not. Roy. Astron. Soc., 492, 5754

Nelson D., et al., 2021, The IllustrisTNG Simulations: Public Data Release (arXiv:1812.05609)

Paech K., Hamaus N., Hoyle B., Costanzi M., Giannantonio T., Hagstotz S., Sauerwein G., Weller J., 2017, Monthly Notices of the Royal Astronomical Society, 470, 2566–2577

Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S., 2013, The Astrophysical Journal, 771, 30

Rozo E., et al., 2016, Mon. Not. Roy. Astron. Soc., 461, 1431

Salcedo A. N., Maller A. H., Berlind A. A., Sinha M., McBride C. K., Behroozi P. S., Wechsler R. H., Weinberg D. H., 2018, Monthly Notices of the Royal Astronomical Society, 475, 4411–4423

Schaan E., White M., 2021, Journal of Cosmology and Astroparticle Physics, 2021, 068

Schmidt F., Elsner F., Jasche J., Nguyen N. M., Lavaux G., 2019, Journal of Cosmology and Astroparticle Physics, 2019, 042–042

Schmittfull M., Simonović M., Assassi V., Zaldarriaga M., 2019, Physical Review D, 100

Schneider A., et al., 2016, Journal of Cosmology and Astroparticle Physics, 2016, 047–047

Sullivan J. M., Seljak U., Singh S., 2021, arXiv e-prints, p. arXiv:2104.10676

Tinker J. L., Kravtsov A. V., Klypin A., Abazajian K., Warren M. S., Yepes G., Gottlober S., Holz D. E., 2008, Astrophys. J., 688, 709

To C., et al., 2021, Physical Review Letters, 126

Virtanen P., et al., 2020, Nature Methods, 17, 261

Vlah Z., Castorina E., White M., 2016, Journal of Cosmology and Astroparticle Physics, 2016, 007–007

Wechsler R. H., Tinker J. L., 2018, Ann. Rev. Astron. Astrophys., 56, 435

Wechsler R. H., Bullock J. S., Primack J. R., Kravtsov A. V., Dekel A., 2002, Astrophys. J., 568, 52

Wechsler R. H., Zentner A. R., Bullock J. S., Kravtsov A. V., 2006, Astrophys. J., 652, 71

Wechsler R. H., DeRose J., Busha M. T., Becker M. R., Rykoff E., Evrard A., 2021, ADDGALS: Simulated Sky Catalogs for Wide Field Galaxy Surveys (arXiv:2105.12105)

White M., et al., 2021, Cosmological constraints from the tomographic cross-correlation of DESI Luminous Red Galaxies and Planck CMB lensing (arXiv:2111.09898)

Yang X., Mo H. J., Bosch F. C. v. d., 2003, Monthly Notices of the Royal Astronomical Society, 339, 1057–1080

Yuan S., Eisenstein D. J., Garrison L. H., 2018, Mon. Not. Roy. Astron. Soc., 478, 2019

Yuan S., Garrison L. H., Hadzhiyska B., Bose S., Eisenstein D. J., 2021, AbacusHOD: A highly efficient extended multi-tracer HOD framework and its application to BOSS and eBOSS data (arXiv:2110.11412)

Zacharegkas G., et al., 2021, Dark Energy Survey Year 3 results: Galaxy-halo connection from galaxy-galaxy lensing (arXiv:2106.08438)

Zennaro M., Angulo R. E., Pellejero-Ibáñez M., Stücker J., Contreras S., Aricò G., 2021a, The BACCO simulation project: biased tracers in real space (arXiv:2101.12187)

Zennaro M., Angulo R. E., Contreras S., Pellejero-Ibáñez M., Maion F., 2021b, Priors on Lagrangian bias parameters from galaxy formation modelling (arXiv:2110.05408)

Zhai Z., et al., 2017, The Astrophysical Journal, 848, 76

Zhai Z., et al., 2019, Astrophys. J., 874, 95

Zheng Z., et al., 2005, Astrophys. J., 633, 791

Zheng Z., Coil A. L., Zehavi I., 2007, Astrophys. J., 667, 760

Zhou R., et al., 2020, Monthly Notices of the Royal Astronomical Society

APPENDIX A: THE SCALE-DEPENDENCE OF BIAS PARAMETERS

In this appendix we report our results on the estimates of the bias parameters $\hat{b}_i$. For each snapshot and HOD form we estimate the $\hat{b}_i$ following Eqn. 15. The assumption that second-order HEFT describes the galaxy sample well can then be tied to the range of scales at which the coefficients remain scale independent. We show these results in Fig. A1. We show the results for the DESI sample at $z = [1.0, 0.7, 0.25]$. The results are broadly consistent with the expectations from biasing theory – the amplitudes of the bias coefficients decrease as we arrive at later times and consequently less-biased samples. We also note that with the exception of the quadratic bias parameter b_{∇^2}, even using $k_{\rm max} = 0.1$ leads to highly precise determinations of the bias parameters despite the limited volume of our boxes. This is due to the field-level nature of the fit, which includes significantly more cosmological information than simply finding bias parameters by fitting summary statistics such as the clustering and galaxy–matter power spectra.

We also find, again with the exception of b_{∇^2}, that the inferred parameters remain relatively stable out to very high $k_{\rm max}$ across all of our redshift bins. We also find some slight running for the tidal bias b_{s^2} at $k \gtrsim 0.4\,h{\rm Mpc}^{-1}$ for some bins, which motivates our fiducial choice of $k_{\rm max} = 0.4\,h{\rm Mpc}^{-1}$ in this publication. However, we note that the procedure selects a single set of biases that minimizes $P_{\rm err}(k)$. As the operators are correlated, there could exist a separate set of $\hat{b}'_i$ that are scale independent and result in an equally acceptable $P_{\rm err}(k)$. Indeed, the fact that at $k_* = 0.3\,h{\rm Mpc}^{-1}$ and $k_* = 0.5\,h{\rm Mpc}^{-1}$ we find statistically indistinguishable stochastic power spectra supports this argument.

The results of Fig. A1 indicate that our estimation procedure for bias parameters from field-level data is both precise and robust. While in the main publication we concern ourselves mainly with $\bar{n}P_{\rm err}$ there is considerable interest in applying these techniques to study the actual parameters for a varying class of tracer–matter connection models.

We also note that the fact that the bias parameters begin to run at $k_{\rm max} \sim 0.4\,h{\rm Mpc}^{-1}$ is not inconsistent with the results of the work of Kokron et al. (2021) which fit power spectra to $k_{\rm max} = 0.6\,h{\rm Mpc}^{-1}$ (and recently Zennaro et al. (2021b) which go to even smaller scales). Those publications were concerned with fitting the clustering and lensing power spectra, while in this publication we concern ourselves with the significantly more difficult problem of describing the full statistical properties of the field that encapsulates information from *all* N-point functions. Indeed, similar tests were carried out in Banerjee et al. (2021) with statistics that also encode higher N-point information and HEFT was found to be a good fit to them at slightly more conservative scales than what was found for power spectra.

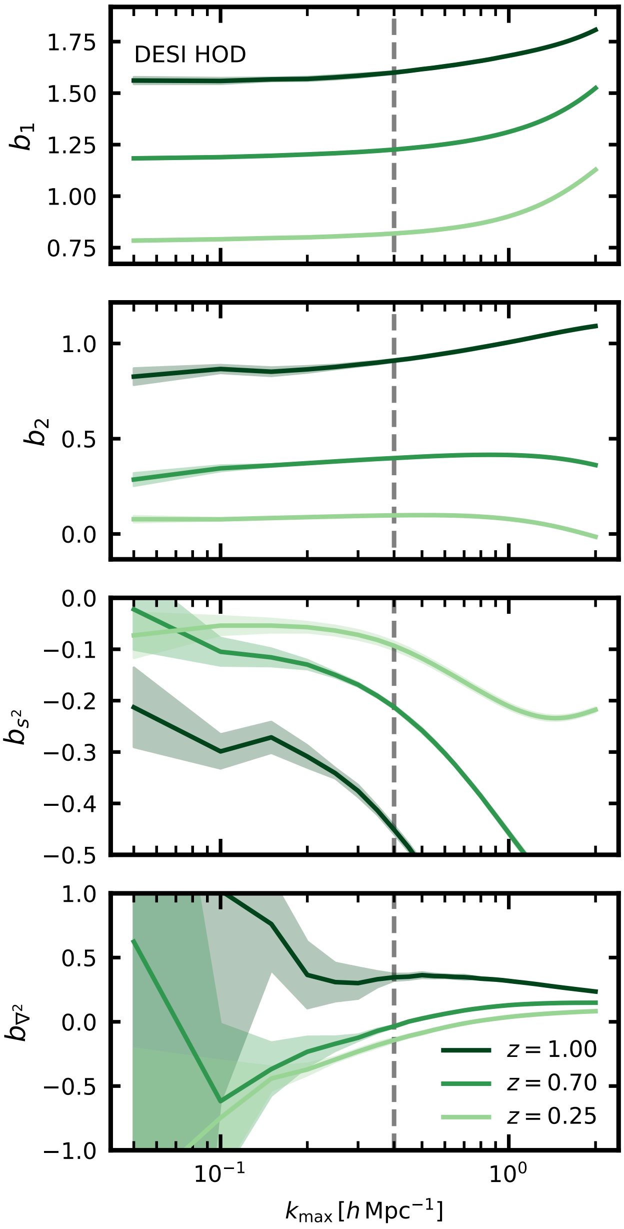

Figure A1. Scale-dependent estimates of Lagrangian bias parameters for the DESI HOD, for three snapshots used in this analysis. The figure shows both the regimes over which we can trust the bias model and also the redshift-dependence of the estimated bias parameters.

APPENDIX B: THE DEPENDENCE OF $\bar{N}P_{\rm ERR}$ ON $K_{\rm MAX}$

In our procedure to obtain the error power spectrum $P_{\rm err}(k)$ we must choose a cut-off scale $k_{\rm max}$ that indicates the smallest scales used to estimate $\hat{b}_i$ and construct a field-level realization of that galaxy sample, under the assumption of *constant bias parameters*. Within the realm of applicability of the bias model, we should then find comparable $P_{\rm err}(k)$ among different choices of $k_{\rm max}$. The purpose of this appendix is to investigate how our results change if we diverge from the fiducial scale $k_{\rm max} = 0.4\,h{\rm Mpc}^{-1}$ adopted throughout the work.

We take a subset of our HOD samples, the `redMaGiC` DES sample that possesses the highest degree of super-Poisson stochasticity and the DESI sample that showed the most sub-Poisson behavior. We pick $z = 0.55$ and $z = 0.25$ as benchmarks. This allows us to assess both super-/sub-Poissonianity at low-redshifts where we expect linear bias to be sufficient and intermediate-redshifts where higher-order operators become more important. We perform measurements of the error power spectra using several different $k_{\rm max}$, namely $k_{\rm max} = [0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.6, 0.7, 0.75, 1.0]$ which spans quasi-linear scales down to deeply nonlinear scales beyond the expected smallest scale where a perturbative bias expansion is applicable.

We show our results in the two panels of Fig. B1. For the DES sample we observe that for both redshifts under consideration, the large-scale spectra are very similar independent of $k_{\rm max}$. However, when lower $k_{\rm max}$ cutoffs are adopted we see a large amount of sample variance in the measured error power spectra at $k \gtrsim 0.4$ which is subsequently reduced when smaller-scale information is incorporated. Notably, including smaller-scale information leads to

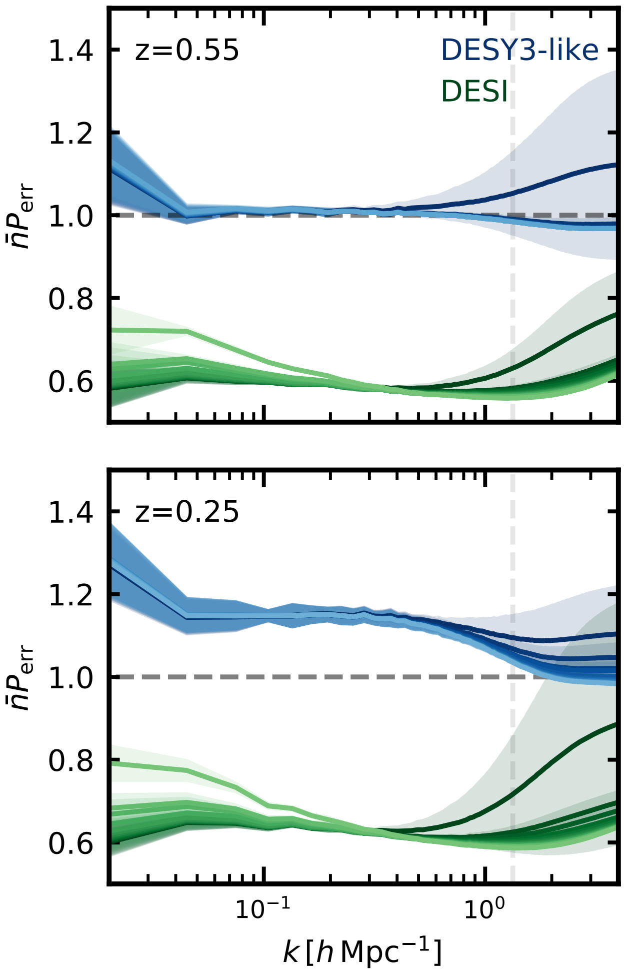

Figure B1. The change in the error power spectrum of the DESI and `redMaGiC` -like samples as we vary the cut-off $k_{\rm max}$ at used to estimate the bias parameters. Cutoffs correspond from $k_{\rm max} = 0.1\,h{\rm Mpc}^{-1}$ to $k_{\rm max} = 1.0\,h{\rm Mpc}^{-1}$. The lightest (darkest) shade of the color corresponds to the largest (smallest) $k_{\rm max}$ used.

small-scale power spectra that asymptote to a value close to the Poisson shot-noise prediction.

For the case of the DESI curves, we observe a similarly large scatter at low $k_{\rm max}$. However, unlike the DESY3-like sample as we include smaller scale information we eventually reach a regime where the large-scale fit is degraded and the error power spectra are no longer consistent with a constant at large scales, indicating a potential breakdown of the bias model.

Nevertheless, we note that the observed $\bar{n}P_{\rm err}$ are quite stable to the choice of $k_{\rm max}$ at large scales until extreme values are chosen. This indicates that our fiducial choice of $k_{\rm max} = 0.4\,h{\rm Mpc}^{-1}$ is adequate for the analysis carried out in this work.

APPENDIX C: COVARIANCE BETWEEN BIAS PARAMETERS

In § 2.3 we derived an estimator for the bias parameters at the field level by minimizing the variance of the residual field $\epsilon(\boldsymbol{x}) = \delta_h(\boldsymbol{x}) - \delta(\boldsymbol{x}) - \sum_i b_i\mathcal{O}_i(\boldsymbol{x})$ and commented on the fact the parameters are correlated. In this appendix we derive the co-variance of the bias parameters and report the structure of the correlation matrix as a function of $k_{\rm max}$.

The covariance ${\rm Cov}(\hat{b}_i, \hat{b}_j)$ will be given by

$$\mathrm{Cov}(\hat{b}_i,\hat{b}_j) = \langle(\hat{b}_i - b_i)(\hat{b}_j - b_j)\rangle \tag{C1}$$

$$= \langle\hat{b}_i\hat{b}_j\rangle - b_ib_j, \tag{C2}$$

where've used the fact that $\langle\hat{b}_i\rangle = b_i$[6]. Recall that from our estimator $\hat{b}_i = M^{-1}_{ij}A_j$ where M_{ij} and A_j are given by Eqns. 18 and 16 respectively. First, we note that we may re-write A_j as

$$A_j = \int\limits_{|\mathbf{k}|<k_{\rm max}} \frac{d^3k}{(2\pi)^3}\mathcal{O}_j(\mathbf{k})[\delta_h - \delta_m]^*(\mathbf{k}), \tag{C3}$$

$$= \int\limits_{|\mathbf{k}|<k_{\rm max}} \frac{d^3k}{(2\pi)^3}\mathcal{O}_j(\mathbf{k})\left[\epsilon + \sum_i b_i\mathcal{O}_i\right]^*(\mathbf{k}), \tag{C4}$$

$$\equiv \int_k \mathcal{O}_j\,[\epsilon + b_i\mathcal{O}_i]^*\,, \tag{C5}$$

where in the last line we have introduced a notational convenience to not clutter subsequent equations. The co-variance term is then given by

$$\langle\hat{b}_i\hat{b}_j\rangle = \langle M^{-1}_{ik}M^{-1}_{jn}A_kA_n\rangle \tag{C6}$$

$$= \left\langle M^{-1}_{ik}M^{-1}_{jn}\int_{k,k'}\mathcal{O}_k\mathcal{O}_n\,[\epsilon + b_i\mathcal{O}_i]^*\,[\epsilon + b_i\mathcal{O}_i]^*\right\rangle \tag{C7}$$

$$= b_ib_j + \left\langle M^{-1}_{ik}M^{-1}_{jn}\int_{k,k'}\mathcal{O}_k\mathcal{O}_n\epsilon\epsilon\right\rangle. \tag{C8}$$

The Hessian M_{ij} is fixed for a given realization, and thus we can remove it from the expectation value. The b_ib_j terms will cancel, and we are left with

$$\mathrm{Cov}(\hat{b}_i,\hat{b}_j) = M^{-1}_{ik}M^{-1}_{jn}\int_{k,k'}\left\langle\mathcal{O}_k(\mathbf{k})\mathcal{O}^*_n(-\mathbf{k}')\epsilon(\mathbf{k})\epsilon^*(-\mathbf{k}')\right\rangle. \tag{C9}$$

Once again, the component fields $\mathcal{O}_k$ are deterministic given a dark matter realization and so we find that our end result is

$$\mathrm{Cov}(\hat{b}_i,\hat{b}_j) = M^{-1}_{ik}M^{-1}_{jn}\int\limits_{|\mathbf{k}|<k_{\rm max}}\frac{d^3k}{(2\pi)^3}\mathcal{O}_k(\mathbf{k})\mathcal{O}^*_n(\mathbf{k})P_{\rm err}(k). \tag{C10}$$

Under the approximation the next-to-leading non-Poisson corrections to $P_{\rm err}(k)$ will be negligible up to $k_{\rm max}$ we then find

$$\mathrm{Cov}(\hat{b}_i,\hat{b}_j) \approx \frac{M^{-1}_{ij}(k_{\rm max})}{\bar{n}}\left[a_1 + O\left(\int_k k^2\mathcal{O}_k\mathcal{O}_n\right)\right]. \tag{C11}$$

The correlation coefficient of bias parameters, ρ_{b_i,b_j} is thus given by

$$\rho_{\hat{b}_i,\hat{b}_j}(k_{\rm max}) = \frac{M^{-1}_{ij}(k_{\rm max})}{\sqrt{M^{-1}_{ii}(k_{\rm max})M^{-1}_{jj}(k_{\rm max})}}. \tag{C12}$$

In Fig. C1 we show the bias correlation coefficients as computed in Eqn. C12 for the DESI HOD across three different redshifts. We observe a pattern of correlations that evolves significantly with $k_{\rm max}$ and with redshift. As mentioned in § 2.3, these correlations can be partially responsible for the running observed in the bias parameters with $k_{\rm max}$. Better quantifying these correlations and their impact in

[6] This is true under the assumption the stochastic residual field ϵ and the HEFT component fields $\mathcal{O}_i$ are uncorrelated.

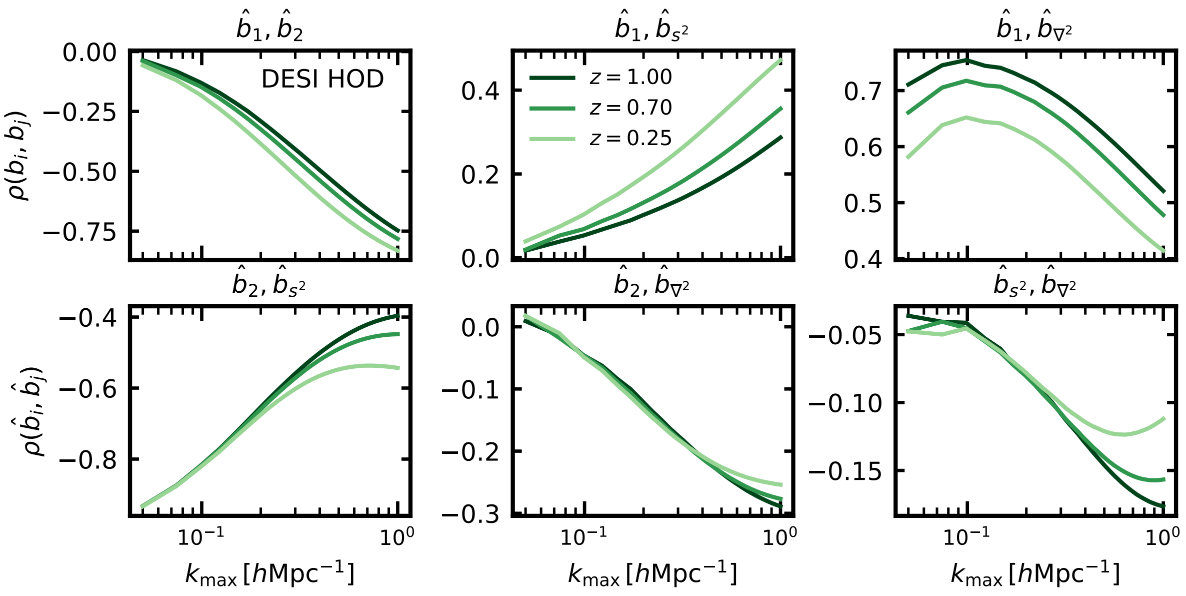

Figure C1. Cross-correlation coefficients for Lagrangian bias parameters from the field-level estimator developed in this work, as a function of $k_{\rm max}$, for the case of the DESI HOD across three snapshots. We show the mean curve for the five simulations at our fiducial cosmology, but note that the scatter in the correlation coefficients is negligible.

field-level bias parameter estimation is an important step in using the techniques developed in this work to also place priors in the bias parameters themselves, and not just the stochasticity of galaxy samples.

APPENDIX D: SUBSETS OF OPERATORS AND $\bar{N}P_{\rm ERR}$

Previous studies of halo stochasticity have resorted to studying the error power spectrum under the assumption of only including linear bias (Hamaus et al. 2010; Baldauf et al. 2013). The impact of not including higher-order bias operators then manifests itself as a super-Poisson contribution beyond the one-halo enhancement we have previously discussed.

We investigate the dependence of our $\bar{n}P_{\rm err}$ measurements on including the full set of second-order Lagrangian bias fields. We fit bias parameters to the DESI and redMaGiC HOD fields at redshifts $z = 0.25$ and $z = 0.55$, which allows us to probe the impact of including subsequent operators as a function of redshift. These measurements are reported in Fig. D1, where we show the impact of including subsequent operators in the error power spectra.

For the DESI sample, we see that the including additional bias operators has the largest effect at large scales. The biggest impact comes from including the quadratic bias operator $\mathcal{O}_{\delta^2}$, which eliminates a significant portion of the excess super-Poissonian stochasticity compared to the expected asymptotic value coming only from exclusion. While the impact of including the tidal shear and non-local bias operators is more modest, their inclusion trends toward flattening the large-scale power spectrum. We also note that the higher redshift snapshot has a significantly larger impact from including additional bias operators. This is consistent with the fact that galaxy samples are more biased tracers of the matter density field at higher redshifts. At very small scales, including additional bias parameters has little effect.

In the case of redMaGiC , we find that the impact of including subsequent bias operators is smaller than for the DESI sample. At large scales the trend is to flatten the spectra, but the impact of any given operator is seemingly more modest. We additionally find, for the low-redshift snapshot, that including the higher-derivative bias operator leads to high-k spectra that are closer to the Poisson expectation. A potential explanation is that the excess one-halo enhancement seen in redMaGiC can contribute with an amplitude comparable to higher-order operators, making their impact less significant. However, we caution this is a potential explanation and a more cautious investigation is reserved for future work.

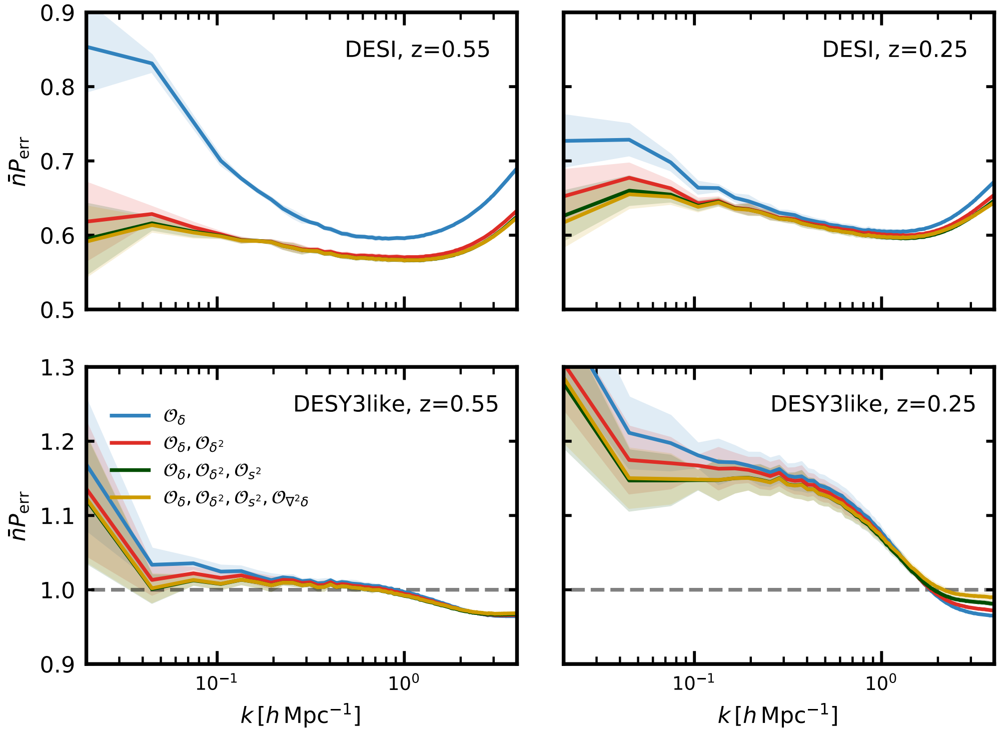

Figure D1. The impact of including higher-order bias operators in the field-level description of different HOD samples, at different snapshots. Each row designates a different type of mock galaxy, while each column indicates the simulation snapshot at which we have carried out these measurements. Note the reduced y-axes compared to other error power spectra presented in this publication. All fits are made using the same fiducial cut-off $k_{\rm max} = 0.4\,h{\rm Mpc}^{-1}$.

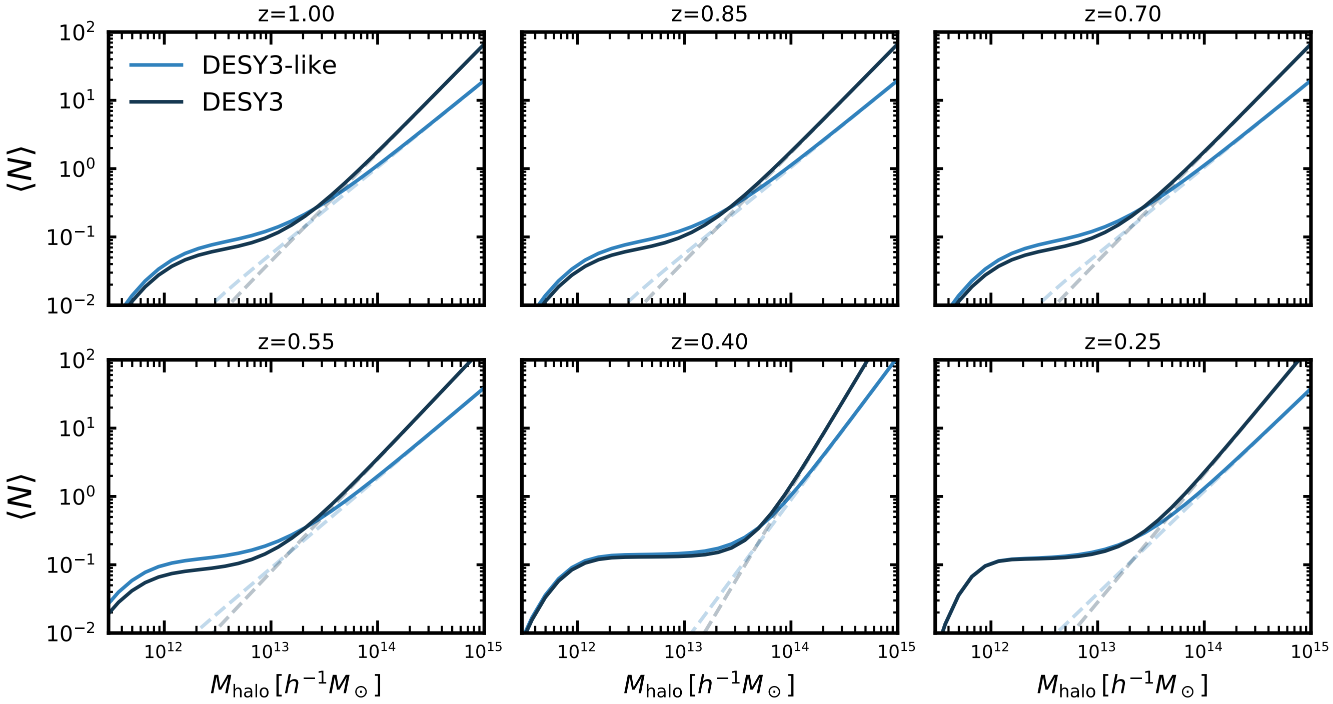

Figure D2. Visualizations of the halo occupation distributions for the DES Y3 `redMaGiC` galaxy sample. The light blue shade corresponds to the 'DESY3-like' sample we adopt as our fiducial in this publication. The darker blue shade corresponds to the occupation as constrained in Zacharegkas et al. (2021). The solid (dashed) lines show the number of total (satellite) galaxies per halo mass, respectively.