# Differentiable Stochastic Halo Occupation Distribution

Benjamin Horowitz, Changhoon Hahn, Francois Lanusse, Chirag Modi, Simone Ferraro

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **529,** 2473–2482 (2024) https://doi.org/10.1093/mnras/stae350
Advance Access publication 2024 February 2

# Differentiable stochastic halo occupation distribution

Benjamin Horowitz,[1,2]★ ChangHoon Hahn [1], Francois Lanusse,[3] Chirag Modi [4,5] and Simone Ferraro[2]

[1]*Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA*
[2]*Lawrence Berkeley National Lab, 1 Cyclotron Road, Berkeley, CA 94720, USA*
[3]*AIM, CEA, CNRS, Université Paris-Saclay, Université Paris Diderot, Sorbonne Paris Cité, F-91191 Gif-sur-Yvette, France*
[4]*Center for Computational Astrophysics, Flatiron Institute, 162 Fifth Avenue, New York, NY 10010, USA*
[5]*Center for Computational Mathematics, Flatiron Institute, 162 Fifth Avenue, New York, NY 10010, USA*

Accepted 2024 January 18. Received 2024 January 17; in original form 2023 January 25

## ABSTRACT

In this work, we demonstrate how differentiable stochastic sampling techniques developed in the context of deep reinforcement learning can be used to perform efficient parameter inference over stochastic, simulation-based, forward models. As a particular example, we focus on the problem of estimating parameters of halo occupation distribution (HOD) models that are used to connect galaxies with their dark matter haloes. Using a combination of continuous relaxation and gradient re-parametrization techniques, we can obtain well-defined gradients with respect to HOD parameters through discrete galaxy catalogue realizations. Having access to these gradients allows us to leverage efficient sampling schemes, such as Hamiltonian Monte Carlo, and greatly speed up parameter inference. We demonstrate our technique on a mock galaxy catalogue generated from the Bolshoi simulation using a standard HOD model and find near-identical posteriors as standard Markov chain Monte Carlo techniques with an increase of $\sim 8\times$ in convergence efficiency. Our differentiable HOD model also has broad applications in full forward model approaches to cosmic structure and cosmological analysis.

**Key words:** methods: numerical – galaxies: fundamental parameters – galaxies: haloes – cosmology: theory.

## 1 INTRODUCTION

Over the past 20 yr, there has been significant observational and theoretical progress in connecting galaxies to their cosmic environments (Peacock & Smith 2000; van den Bosch, Yang & Mo 2003; Kravtsov et al. 2004; Wechsler et al. 2006; Neistein et al. 2011). Understanding this connection is critical for understanding galaxy formation/evolution (Crain et al. 2009; Zehavi et al. 2011) as well as using galaxies as bias tracers of the underlying mass density for cosmological analyses (Benson et al. 2000a; Desjacques, Jeong & Schmidt 2018). Studying this connection is a key component of many upcoming galaxy surveys, including the Prime Focus Spectrograph (Takada et al. 2014; Tamura et al. 2016), the Dark Energy Spectroscopic Instrument (DESI Collaboration 2016), and the *Nancy Grace Roman Space Telescope* (Spergel et al. 2015; Wang et al. 2022).

A key theoretical tool for these studies has been the halo occupation distribution (HOD; Lemson & Kauffmann 1999; Seljak 2000; Scoccimarro et al. 2001; Berlind & Weinberg 2002; Wechsler & Tinker 2018), a framework that specifies how collapsed dark matter haloes (Press & Schechter 1974; Bond et al. 1991; Cooray & Sheth 2002) are populated with galaxies.

This is in contrast to 'environmental' biasing schemes, such as Eulerian or Lagrangian biasing schemes (Mann, Peacock & Heavens 1998; Desjacques, Jeong & Schmidt 2018), common in cosmological analyses of galaxy survey data (Cuesta et al. 2016; Beutler et al. 2017; Ivanov, Simonović & Zaldarriaga 2020; To et al. 2021). Unlike environmental biasing schemes that only model summary statistics like power spectrum of the galaxy field, a well-formulated HOD model provides direct physical insight into galaxy formation physics through its parameters, which are related to critical mass scales in galaxy–halo relation. For example, this allows direct measurement of HOD parameters by comparing observed galaxy populations with dynamical mass measurements, such as X-ray clusters (Zheng et al. 2009; Mehrtens et al. 2016).

In standard HOD implementations (e.g. Zheng, Coil & Zehavi 2007), the HOD model specifies the probability distribution of the number of galaxies, $N$, hosted by a dark matter halo given its properties, such as halo mass: $P(N|M_{\rm halo})$. A semi-analytical halo model approach can include HOD parameters to predict two- and higher point function (Cooray & Sheth 2002), but those predictions are often not accurate enough for analysing modern data sets. Alternatively, a more precise Monte Carlo approach is often used to stochastically assign galaxies to haloes in a large simulation box following the HOD prescription, and then the galaxy power spectrum or other quantities of interest are directly measured from the simulation.

In practice, Markov chain Monte Carlo (MCMC) methods have primarily been used to fit HOD parameters from mock or actual data (e.g. White et al. 2011; Rodríguez-Torres et al. 2016; Sinha et al. 2018). However, these methods scale poorly with the number of parameters that need to be fit. As novel decorated HOD models increasingly add more assembly bias parameters to accurately cap-

★ E-mail: bhorowitz@berkeley.edu

ture small-scale observations, this exercise can become challenging, especially if there are unforeseen degeneracies in the parameter space. These challenges can be overcome more easily with parameter inference methods that rely on the gradient information, i.e. where we can estimate the response of the observations with respect to the underlying parameters of the model, such as Hamiltonian Monte Carlo (HMC; Duane et al. 1987a), variational inference (Peterson 1987; Beal 2003; Blei, Kucukelbir & McAuliffe 2016), or combinations thereof (Gabrié, Rotskoff & Vanden-Eijnden 2022; Modi, Li & Blei 2023). However, these methods are not applicable in current HOD models as the implementation of their stochastic galaxy assignment schemes makes them classically non-differentiable.

A differentiable HOD framework will enhance dynamical forward modelling frameworks that seek to reconstruct latent cosmological fields (e.g. Seljak et al. 2017), which are constrained to use gradient-based methods for optimization due to the high dimensionality of the inference problem. These frameworks generally rely on perturbative bias models that are accurate only on large scales (Modi et al. 2019; Schmidt et al. 2019) or heuristic neural network models with a large number of latent parameters (Modi, Feng & Seljak 2018). A differentiable HOD approach will allow one to push to smaller scales with a well-understood, physically developed model that has only a handful of parameters.

An alternative to HOD models that maintains the requisite differentiability is to use differentiable emulators (Kwan et al. 2015; Wibking et al. 2020) or fitting functions of the observables like the one proposed in Hearin et al. (2021) for galaxy assembly bias. However, these are efficient only for the particular summary statistics and cosmological parameters on which they are trained. Hence, they require a new training set once these are varied. Depending on the parameter space of interest, this could be of prohibitive computational cost. In addition, separate emulators must be trained for each summary statistic of interest as such methods do not match the galaxy observations at the field level.

Motivated by this, here we adopt a different approach and aim to make the HOD sampling itself differentiable. Our aim is to be able to compute gradients of any observable with respect to HOD parameters through a particular realization of a galaxy catalogue. Common wisdom states that differentiating through stochastically sampled discrete random variables, such as the number of satellites in a given halo, is not possible. However, modern reinforcement learning has spurred the development of techniques to deal with these types of categorical variables in the context of deep neural network training via back-propagation. In particular, we use the Gumbel-Softmax or CONCRETE method (Jang, Gu & Poole 2016; Maddison, Mnih & Whye Teh 2016) that utilizes continuous distributions to approximate the sampling process of discrete stochastic variables, such as galaxies, in a differentiable fashion. It relies on two insights: (1) a re-parametrization for a discrete (or categorical) distribution in terms of the Gumbel distribution (referred to as the 'Gumbel trick'; Maddison, Tarlow & Minka 2014) and (2) making the corresponding function continuous by using a continuous approximation that depends on a *temperature* parameter, which in the zero-temperature case degenerates to the discontinuous, original expression.

In this paper, we will implement the Gumbel-Softmax method in the context of HOD models and apply it to mock data sets. In Section 2, we will describe our HOD model and the methods used to allow differentiability of its categorical outputs. In Section 3, we apply this technique to a Monte Carlo analysis of a mock galaxy catalogue constructed from the Planck–Bolshoi simulation. In Section 4, we compare the differentiable HOD model to that from a standard approach and discuss its applications.

## 2 METHOD

In this section, we provide some background on the various components of our HOD model, and detail our strategy to make this model differentiable. We implement our model using TensorFlow Probability (Tran et al. 2016; Morgan 2018), particularly the TensorFlow Distribution package (Dillon et al. 2017).

### 2.1 HOD model

To describe the population of galaxies in our haloes, we use the standard Zheng et al. (2007) HOD model. In the Zheng et al. (2007) model, the probability of a given halo hosting $N$ galaxies is dictated solely by its mass – $P(N|M)$. The model separately populates central and satellite galaxies, motivated by theoretical studies (Kravtsov et al. 2004; Zheng et al. 2005), and has five free parameters with some physical significance that can be related back to well-studied mass–luminosity relationships.

#### *2.1.1 Central occupation*

For central galaxies, the mean occupation function is step-like with a soft cut-off at high mass to account for natural scatter between galaxy luminosity the halo host mass. There are two free parameters controlling this function, the characteristic minimum mass of haloes hosting central galaxies above some luminosity threshold, $M_{\rm min}$, and the width of the cut-off profile, $\sigma_{\log M}$:

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

where erf is the standard error function and $M$ is the halo mass. Given the mean occupation for a halo of a given mass, central galaxies are assigned to haloes by sampling a Bernoulli distribution:

$$N_{\rm cen} \sim {\rm Bernoulli}\big(p = \langle N_{\rm cen}(M)\rangle\big). \tag{2}$$

#### 2.1.2 SATELLITE OCCUPATION

Simulations suggest that satellites follow an approximately power-law distribution with a slope close to unity at the high-mass end. At lower masses, the shape of the distribution changes and the overall distribution can be parametrized as

$$\langle N_{\rm sat}(M)\rangle = \langle N_{\rm cen}(M)\rangle \left(\frac{M - M_0}{M_1'}\right)^{\alpha}, \tag{3}$$

where $\alpha$ is the power-law slope at high masses, $M_0$ is the characteristic mass of the change-over, and $M_1'$ is the characteristic amplitude. This mean number of satellites for a given mass is then used to define the intensity $\lambda$ of a Poisson distribution, from which a particular number of satellites are drawn for each halo.

$$N_{\rm sat} \sim {\rm Poisson}\big(\lambda = \langle N_{\rm sat}(M)\rangle\big). \tag{4}$$

#### *2.1.3 Spatial satellite distribution*

In the Zheng et al. (2007) HOD, central galaxies are located at the centre of its host halo and satellite galaxies are distributed according to a Navarro, Frenk & White (1997) profile (hereafter NFW). To sample the satellite galaxy positions, we utilize the Robotham & Howlett (2018) implementation, which constructs an efficient mapping from a random sample and the full NFW profile via

the calculation of the quantile function, i.e. the inverse cumulative distribution function (CDF). This can be written analytically as

$$q(p; c, M_{\rm vir}) = -\frac{1}{c}\left[1 + \frac{1}{W_0(-e^{-pM_{\rm vir}-1})}\right], \quad (5)$$

where $W_0$ is the Lambert-W function, $M_{\rm vir}$ is the virial mass, and $c$ is the concentration parameter. Using this inverse CDF, we can now randomly draw radial distances for satellites by sampling $p$ from $U[0, 1]$ and mapping it to radii as $r/r_{\rm vir} = q(p, c)$. The angular position of the satellite is sampled uniformly in this isotropic model.

### 2.2 Differentiable stochastic sampling

In this section, we review the key ideas behind differentiable stochastic sampling, which will form the building blocks of DIFFHOD.

#### 2.2.1 *Stochastic backpropagation by reparametrization*

One of the most common approaches for backpropagation through stochastic sampling is the so-called reparametrization trick, extensively used for instance in Variational Auto-Encoders (Kingma & Welling 2013; Jimenez Rezende, Mohamed & Wierstra 2014).

The key idea of this approach is to rewrite samples $z$ from a given parametric distribution $\mathbb{P}_\theta$ as a deterministic and differentiable transformation $f$ applied to a fixed distribution $\mathbb{P}_\epsilon$:

$$z = f(\theta, \epsilon) \text{ with } \epsilon \sim \mathbb{P}_\epsilon. \quad (6)$$

This reparametrization of the samples allows to side-step having to take derivatives of the stochastic variable $\epsilon$ when computing derivatives of some downstream function $h$ with respect to the distribution parameters $\theta$. This can be expressed in terms of the expectation value with respect to random variable, $\mathbb{E}_z$, as

$$\frac{\partial}{\partial\theta}\mathbb{E}_{z\sim\mathbb{P}_\theta}[h(z)] = \mathbb{E}_{\epsilon\sim\mathbb{P}_\epsilon}\left[\frac{\partial}{\partial\theta}h(f(\theta,\epsilon))\right]. \quad (7)$$

In the right-hand side of this expression, the derivative now only involves taking gradients of a deterministic function of $\theta$ since $\epsilon$ is treated as an input to the function.

To provide a simple concrete example of such reparametrization, let us consider a Gaussian distribution of mean $\mu$ and standard deviation $\sigma$. This is the standard example used in the case of variational auto-encoders. One can express a sample $z \sim \mathcal{N}(\mu, \sigma^2)$ as $z = \mu + \sigma\epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, making it trivial to take derivatives of the samples with respect to the parameters of the distribution ($\mu$ and $\sigma$). We can then use this form for an optimization problem, for example minimizing $\mathbb{E}_{z\sim p_\mu}[z^2]$ with respect to parameter $\mu$, via calculating the derivative as follows:

$$\frac{\partial}{\partial\mu}\mathbb{E}_{z\sim p_\mu}\left[z^2\right] = \frac{\partial}{\partial\mu}\mathbb{E}_{\epsilon\sim q_\epsilon}\left[(\mu+\sigma\epsilon)^2\right] = \mathbb{E}_{\epsilon\sim q_\epsilon}[2(\mu+\sigma\epsilon)]. \quad (8)$$

Setting this equal to zero, we find $\mu = 0$, the expected result.

#### 2.2.2 *Gumbel-Softmax trick for categorical variables*

The reparametrization trick as presented above requires the samples to be expressible as a deterministic and differentiable function of a random variable. While this can often be achieved for continuous distributions, it is typically not directly possible for discrete categorical variables. To overcome this limitation, the Gumbel-Softmax trick (Jang, Gu & Poole 2016; Maddison, Mnih & Whye Teh 2016) introduces a relaxation of a categorical distribution to a continuous distribution, which can then be handled with the reparametrization trick.

Let $z$ be a categorical variable with class probabilities $\pi_1, \pi_2, ... \pi_j$ that we wish to sample. We assume that categorical samples are encoded as $N$-dimensional one-hot vectors, i.e. they are $1 \times N$ vectors with all elements 0 except the element corresponding to the sampled class which is 1. The simplest way to sample $z$ is by

$$z = \text{onehot}(\max i | \pi_1 + ... + \pi_{i-1} \le U), \quad U \sim \text{Uniform}(0, 1), \quad (9)$$

where we use the onehot function to express a onehot vector embedding. A first step towards making these samples differentiable is to use the Gumbel-Max trick (Gumbel 1954; Maddison, Tarlow & Minka 2014), which reparametrizes categorical sampling as

$$z = \text{onehot}\left(\text{argmax}_i[g_i + \log(\pi_i)]\right), \quad (10)$$

where $g_i$ are i.i.d. random variables drawn from the Gumbel distribution between 0 and 1, Gumbel(0, 1).[1] This reparametrization trick refactors the sampling of $z$ into a deterministic function of the parameters ($\pi$) and some independent noise with a fixed distribution.

However, the reparametrized function is still non-differentiable due to the argmax function. A continuous, differentiable approximation to this is given by a softmax function,

$$\text{softmax}(\mathbf{z}, \tau)_i = \frac{e^{z_i/\tau}}{\sum_{j=1}^k e^{z_j/\tau}}, \quad \mathbf{z} = (z_1, z_2, ..., z_k) \quad (11)$$

where $\tau$ is a free parameter sometimes referred to as the 'temperature.' This is also known as a relaxation of the distribution, since it can be viewed as a smoothed version of the original distribution.

Using this approximation relaxes the discreteness of the Gumbel-Max trick and generates a $k$-dimensional vector $\mathbf{z}$

$$\hat{z}_i = \frac{\exp((\log(\pi_i) + g_i)/\tau)}{\sum_j \exp((\log(\pi_j) + g_j)/\tau)}. \quad (12)$$

We recover the true discrete function in the limit of $\tau \to 0$. As this function is analytical in class probabilities $\pi$ for $\tau > 0$, we can estimate the gradients of the observed samples $\mathbf{z}$ with respect to the parameters parametrizing $\pi$. Note that as $\tau \to \infty$, the function goes to a constant value.

In the remaining of this work, we will make use of the special case when the number of classes is 2, i.e. when the categorical distribution reduces to a Bernoulli distribution. In this special binary case, equation (12) can be simplified. Maddison, Mnih & Whye Teh (2016) refer to the resulting distribution as BinConcrete; we will refer to it in this work as a Relaxed Bernoulli distribution. Using the fact that the difference of two Gumbel variables follows a Logistic distribution,[2] Maddison, Mnih & Whye Teh (2016) show the Relaxed Bernoulli can be reparametrized as

$$z = \frac{1}{1 + \exp(-(\log\pi + \epsilon)/\tau)} \text{ with } \epsilon \sim \text{Logistic}(0, 1), \quad (13)$$

where in this expression $\pi$ is the odds ratio $\pi = p/(1-p)$ if $p$ is the probability of corresponding Bernoulli distribution.

[1] The standard Gumbel distribution is defined by cumulative distribution function $\text{CDF}(x) = \exp[-\exp(x)]$ and probability density function, $\text{PDF}(x) = \exp\{-[x + \exp(-x)]\}$.

[2] The standard logistic distribution follows the following probability density function: $p(x) = \frac{\exp(-x)}{[1-\exp(-x)]^2}$.

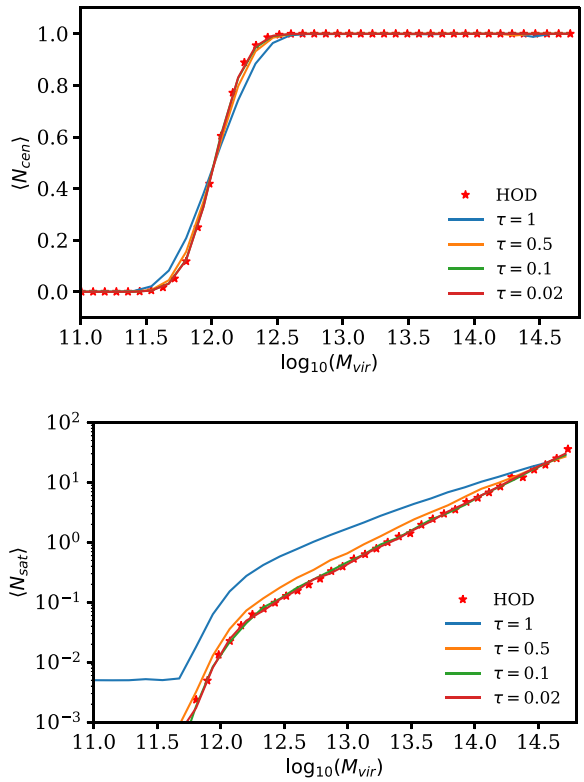


**Figure 1.** HOD for central (top) and satellite galaxies (bottom) as a function of halo virial mass of our differentiable HOD model (DIFFHOD; solid). Lines indicate decreasing temperature values used in the Gumbel-Softmax approximation (see equation 12). We include the occupation distribution from the standard Zheng et al. (2007) HOD model for reference (star). In this work, we use DIFFHOD with $\tau = 0.1$, which is in good agreement with the standard HOD model throughout the full halo mass range (https://github.com/DifferentiableUniverseInitiative/DHOD/blob/master/nb/Plots_for_Paper.ipynb).

### 2.3 DIFFHOD implementation

We now have all the elements needed to build a differentiable HOD (hereafter DIFFHOD) model. We describe in this section our strategy for sampling central and satellites occupation, and satellites positions.

#### 2.3.1 *Differentiable central occupation sampling*

As described in Section 2.1, the central occupation is defined by a Bernoulli distribution, with a parameter $p = \langle N_{\rm cen}(M|M_{\rm min}, \sigma_{\log M})\rangle$ defined as a deterministic function.

Here, we can directly apply the Gumbel-Softmax trick introduced above in the special case of a binary variable. We therefore sample the central occupation of each halo using the Relaxed Bernoulli distribution:

$$N_{\rm cen} = \frac{1}{1+\exp\left(-\left(\log\left(\frac{p}{1-p}\right)+\epsilon\right)/\tau\right)} \text{ with } \epsilon \sim \text{Logistic}(0,1), \tag{14}$$

at given temperature $\tau$. Fig. 1 illustrates the high agreement between the halo occupation obtained by sampling centrals with this relaxed distribution compared to the analytical expectation at our fiducial choice of $\tau = 0.1$.

#### 2.3.2 *Differentiable satellite occupation sampling*

For satellites, we aim to define a differentiable approach to sampling from a Poisson distribution with intensity $\lambda = \langle N_{\rm sat}(M|M_0, M_1', \alpha)\rangle$, also a deterministic and differentiable function. To build on the Gumbel-Softmax trick, we propose to replace conventional Poisson sampling of the total number of satellites by sampling each satellite individually from a Bernoulli distribution.

Let us consider a halo with a Poisson rate $\lambda$ for its satellite occupation. We assume the halo can have a maximum of $N$ satellites, then for each potential satellite we sample from a Bernoulli distribution with probability $p = \lambda/N$ whether this satellite will be included in the halo. The resulting statistics of the number of satellites with this procedure will be Binomial (as $N$ draws from i.i.d. Bernoulli).

More formally, we propose to approximate the Poisson distribution with intensity $\lambda$ of a standard HOD by a Binomial distribution with $N$ trials and probability $p = \lambda/N$:

$$N_{\rm sat} \sim \text{Binomial}\left(N, p = \frac{\langle N_{\rm sat}\rangle}{N}\right). \tag{15}$$

By construction, this Binomial distribution will yield the same mean number of satellites as the Poisson distribution, however the variance of both distributions is different:

$$\text{Var}(N_{\rm sat}^{\rm Pois.}) = \langle N_{\rm sat}\rangle\,, \tag{16}$$

$$\text{Var}(N_{\rm sat}^{\rm Bin.}) = \langle N_{\rm sat}\rangle * \left(1 - \frac{\langle N_{sat}\rangle}{N}\right). \tag{17}$$

From this expression, one can foresee that the Binomial distribution will be a close approximation to a Poisson distribution when the ratio $\frac{\langle N_{\rm sat}\rangle}{N}$ is small, i.e. when the number of trials is large compared to the expected number of satellites. This is actually known as the *law of rare events*, and at fixed $N*p$ the Binomial distribution Binomial($N$, $p$) converges to a Poisson distribution when the number of trials $N \to \infty$.

In practice, we will need to limit the number of trials $N$ to some finite value and Fig. 2 compares the shapes of satellites distributions with a Poisson model versus a Binomial model for two different choices of $N$, and for different halo masses. A higher value of $N$ can improve accuracy but will also increases memory costs, which scales linearly with the maximum number of satellite galaxies encoded in our one-hot embedding. Meanwhile, too low of a $N$ can bias results by artificially reducing the variance of the satellite occupation distribution, or worse, truncating the satellite galaxies of the most massive haloes.

Our fiducial choice in this work is $N = 48$. In Fig. 1 (bottom), we can see that the satellite population reaches a maximum of $\sim$40 galaxies for the most massive $\sim 10^{15}\,M_\odot$ haloes. For the mass range considered here, we find $N > 40$ does not significantly improve the statistical match in summary statistics (correlation function, power spectra, etc.) of the resulting galaxy fields and larger $N$ further will increase memory requirements and computational time. If the end statistic of interest is particularly sensitive to galaxy populations in the most massive clusters, a higher $N$ might be needed. If one is limited by memory or implementing HOD for haloes with a broader mass range, then it may be more efficient to have multiple halo mass bins with different maximum number of allowed satellites $N$.

This Binomial assumption for the sampling of satellites brings two concrete advantages:

(i) Having restated satellite sampling as draws from Bernoulli distributions, we can make the procedure differentiable by using the Relaxed Bernoulli, similarly to centrals.

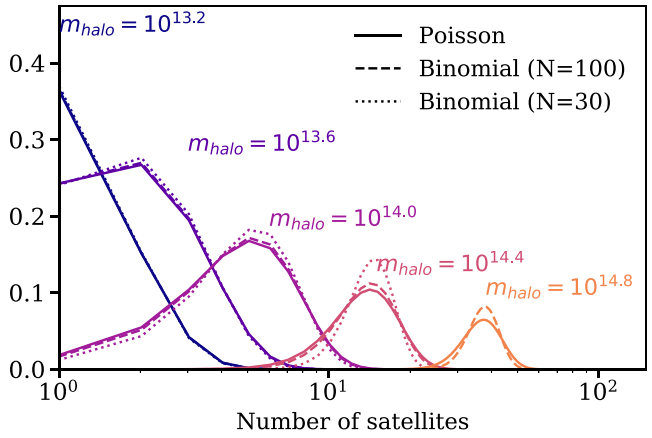


**Figure 2.** Comparison of satellite occupation distributions for different halo masses, under various assumed distributions: Poisson (solid line), Binomial with 100 trials (dashed line), Binomial with 30 trials (dotted line). By construction, the Binomial approximation recovers the mean number of satellites, but for massive haloes limiting the number of samples N will lead to departure in the spread of the distribution compared to a Poisson distribution. Note that for the $m_{\rm halo} = 10^{14.8}$ halo the mean number of satellites is above 30, so the Binomial distribution is not well defined (https://github.com/DifferentiableUniverseInitiative/DHOD/blob/master/nb/PoissonVSBinomial.ipynb).

(ii) Using a fixed number $N$ of potential satellites gives us a practical way to handle varying number of satellites per haloes.

Concretely, for each candidate satellite $i \in [\![1, N]\!]$ of a halo, we sample whether the satellite will be included in the halo using

$$z_i = \frac{1}{1 + \exp\left(-\left(\log\left(\frac{p}{1-p}\right) + \epsilon_i\right)/\tau\right)} \text{ with } \epsilon \sim \text{Logistic}(0, 1), \tag{18}$$

where $p = \frac{\langle N_{sat}\rangle}{N}$ and $\tau$ is the temperature. As a result, for each halo we obtain a vector $\boldsymbol{z}$ of size $N$ which encodes active satellites for the halo.

In all downstream computations, this vector $\boldsymbol{z}$ can be interpreted as a weight between 0 and 1 to apply to each of the $N$ satellites of each halo, for instance in the computation of two-point correlation functions.

### 2.3.3 *Differentiable sampling from NFW distribution*

The last step to complete our HOD implementation is to sample the position of satellites based on an NFW profile centred at the halo position. There are two difficulties here: 1. sampling positions for a varying number of satellites, 2. making the sampled positions differentiable with respect to the NFW parameters.

The first question of dealing with varying number of satellites is solved by our Binomial model with fixed number of trials $N$. For each halo, we will sample the same number of $N$ sets of coordinates, one for each potential satellite. Whether these coordinates will actually contribute in downstream computations will depend on the per halo satellite occupation vector $\boldsymbol{z}$ introduced above.

The second question, of the differentiability of stochastic coordinates, is again solved by applying the re-parametrization trick to the NFW profile. From equation (5), we know the CDF of the NFW profile, meaning that for each satellite $i \in [\![1, N]\!]$ of a given halo we can sample the halo-centric satellite radial distance as

$$r_i = r_{\rm vir}\, q(\epsilon, c) \text{ with } \epsilon \sim \text{Uniform}(0, 1), \tag{19}$$

where $q$ is a differentiable function of parameters $c$.

In our adaptation of the Zheng et al. (2007) model, we assume an isotropic NFW distribution for satellite, so to retrieve halo-centric cartesian coordinates $x_i$, $y_i$, $z_i$ of a given satellite, we first sample $x_i$, $y_i$, $z_i$ on the unit sphere and then multiply these coordinates by the $r_i$ value sampled above. This is nothing more than an another reparametrization step and the resulting Cartesian coordinates remain fully differentiable with respect to the NFW parameters.

We note that contrary to the sampling of central of satellites which are differentiable approximation to a standard HOD (due to the discrete variables involved), this differentiable implementation of NFW sampling is exact.

### 2.3.4 *Impact of temperature parameter $\tau$*

In the DIFFHOD model, we introduce temperature, $\tau$, as a free parameter (equation 14). Depending on the context/implementation, it may be beneficial to anneal (i.e. reduce) $\tau$ over the course of the optimization such that $\tau \to 0$. For instance, in the original papers (Jang, Gu & Poole 2016; Maddison, Mnih & Whye Teh 2016), the Gumbel-Softmax trick was used in the context of training a neural network where only the optimal network weights were of interest and hence $\tau$ was reduced to nearly zero over the course of the training. However in our case, we are interested in stochastic sampling, and not optimization, so we will pick a single temperature that well approximates the target distribution (i.e. unbiased) while maintaining reasonable derivative properties (i.e. less noisy).

In Fig. 1, we present the central (top) and satellite (bottom) occupation distributions of our DIFFHOD model for different temperatures: $\tau = 0.02$ (red) 0.1 (green), 0.5 (orange), and 1 (blue). We include the occupation distributions for the standard HOD model for comparison (star). For this work, we take an experimental approach for determining $\tau$, as advocated in Maddison, Mnih & Whye Teh (2016): the temperature should set as high as possible while maintaining the desired accuracy of the target distribution. We find that a fixed $\tau = 0.1$ provides high accuracy while maintaining stable gradients for both sampling the number of galaxies in each halo as well as the positions of the satellites from the NFW profile (see later Fig. 3).

In general, an end user should confirm that the temperature parameter chosen is appropriate for their application both in terms of suitably smooth derivatives and in distribution accuracy. While there are some 'principled' approaches to setting the temperature parameter discussed in machine learning literature (Abid, Balin & Zou 2019), they are in the context of training deep neural networks and are not necessarily useful for the differentiable model context. The quantitative requirements for accuracy and differentiability depend entirely on the overall workflow (i.e. optimization/sampling algorithm, observable of interest, etc.) and are difficult to set *ab initio*.

## 3 EXPERIMENTATION

To test our implementation, we construct a fiducial mock galaxy catalogue from the Planck Bolshoi simulation halo catalogue (Klypin, Trujillo-Gomez & Primack 2011) at $z = 0$. We treat this catalogue as our mock observation. This simulation has a side length of 250 $h^{-1}$ Mpc, and contains 1367 493 unique haloes ranging in mass from $1.1 \times 10^{15}\,{\rm M}_\odot$ down to $2.7 \times 10^{8}\,{\rm M}_\odot$. We use halotools (Hearin

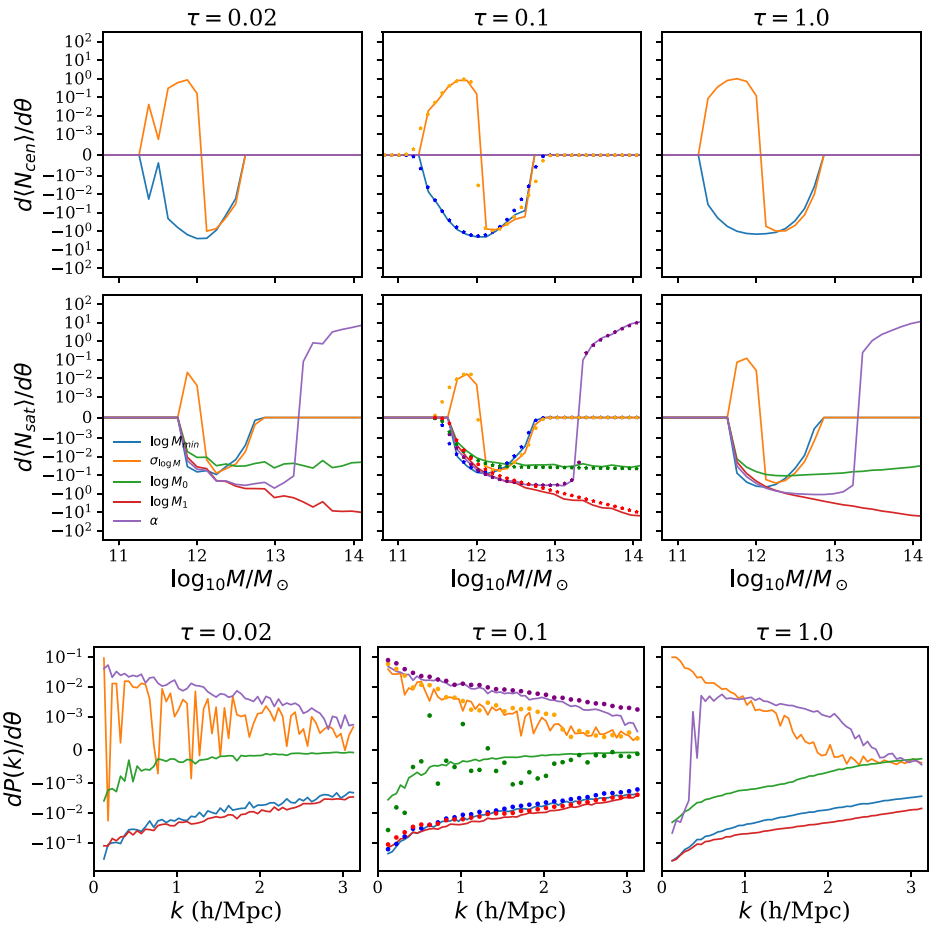


**Figure 3.** *Top*: Derivatives of the central and satellite occupation functions with respect to the HOD parameters for various temperatures: $\tau = 0.02$ (left) 0.1 (centre), and 1 (right). The derivatives are calculated at the fiducial HOD parameter values. We include derivatives analytically derived from equations (1) and (3) for comparison (star). *Bottom*: Derivatives of the galaxy power spectrum with respect to the HOD parameters, evaluated at the fiducial HOD parameter values. We include derivatives calculated using the standard HOD with finite difference for comparison (star). Our DIFFHOD model with $\tau = 0.1$ provides sufficiently smooth derivatives that are in good agreement with analytical derivatives for the occupation function and with standard HOD derivatives for the power spectrum (https://github.com/DifferentiableUniverseInitiative/DHOD/blob/master/nb/Plots_for_Paper.ipynb).

et al. 2016) to generate our fiducial catalogue with HOD parameter values:

$$\log M_{\rm min} = 12.02, \sigma_{\log M} = 0.26, \log M_0 = 11.38,$$
$$\log M_1 = 13.31, \alpha = 1.06.$$

These parameters correspond to the best-fitting values from Zheng et al. (2007) using an SDSS-based galaxy catalogue (Zehavi et al. 2005) with $r$-band absolute magnitude threshold of $M_r < -20$.

### 3.1 Summary statistics and derivatives with DIFFHOD

Next, we compare the mock observations to parallel galaxy catalogues constructed using DIFFHOD (Section 2). The galaxy catalogues are then painted on to a grid using a differentiable cloud-in-cell (CIC) painting method (Modi, Lanusse & Seljak 2021) and its real-space power spectra calculated via a differentiable TensorFlow power spectrum implementation (Horowitz et al. 2021). For DIFFHOD, all steps are differentiable so the overall mapping from original halo catalogue to end power spectrum is also differentiable via the chain rule. In Fig. 3, we present the DIFFHOD derivatives of the central and satellite occupation distribution functions (top) and resulting power spectra (bottom) with respect to the HOD parameters for different temperatures: $\tau = 0.02$ (left), 0.1 (middle), and 1 (right). The derivatives are evaluated at the fiducial HOD parameter values. In the centre panels, for comparison, we include derivatives of the central and satellite occupation distribution functions derived analytically and derivatives of the power spectrum derived using the standard HOD with finite differences (star).

With $\tau = 0.02$, the power spectrum derivatives have significant numerical noise. The $\tau = 1.0$ derivatives are smoother but we find inaccurate occupation distributions (Fig. 1) and a significantly biased power spectrum. Meanwhile, with $\tau = 0.1$ there is still some noticeable numerical noise in the derivatives, however, this is sufficiently smooth for our application and for our optimization to be well behaved. We also find that the $\tau = 0.1$ DIFFHOD derivatives are in good agreement with analytical derivatives for the occupation function and with standard HOD derivatives estimated using finite differences for the power spectrum.

We compare the differentiable power spectrum from DIFFHOD to the DIFFHOD power spectrum from the standard HOD model in Fig. 4. We use the fiducial HOD parameter values and estimate the error bars from 500 realizations of the DIFFHOD model. We find good

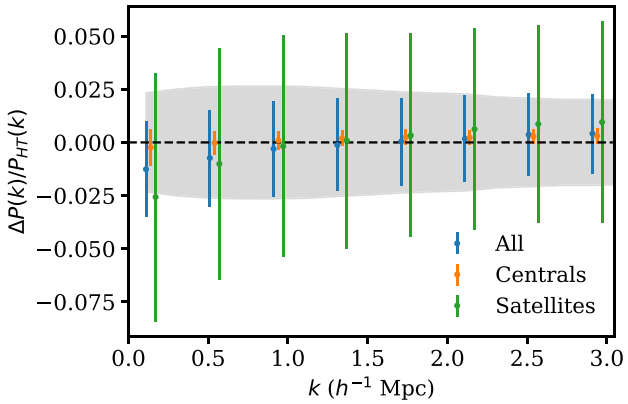


**Figure 4.** Comparison of power spectrum derived using a standard HOD model ($P_{\rm HT}$) versus our DIFFHOD model. We plot the ratios for central and satellite galaxies, as well as for all galaxies. Error bars represent the standard deviation of the DIFFHOD model estimated from 500 realizations of the galaxy sampling. Grey band represents the standard deviation of $P_{\rm HT}$ for all galaxies. We find good agreement between the differentiable power spectrum and $P_{\rm HT}$ across all scales ($k < 3\,h\,{\rm Mpc}^{-1}$), as well as similar distributional properties (https://github.com/DifferentiableUniverseInitiative/DHOD/blob/master/nb/power_spectra_comparison.ipynb).

agreement (better than 1 per cent) across all scales, with particularly good agreement for central galaxies, which are not sampled from the NFW profile. Low $k$ modes are most sensitive to the most massive haloes whose satellite galaxy populations are truncated by our one-hot distribution ($N = 48$; Section 2.3); those who are interested in this regime can further increase $N$. However, even at $k < 0.15\,h\,{\rm Mpc}^{-1}$, we find good agreement between the power spectra.

### 3.2 Monte Carlo analysis with DIFFHOD

Lastly, we demonstrate that we can derive unbiased inference using DIFFHOD, by comparing the posteriors on HOD parameters derived using DIFFHOD to posteriors derived using standard methods with the same mock observations and likelihood. We use the power spectrum measured from the fiducial galaxy catalogue as our mock observations, and construct a covariance matrix using 100 galaxy catalogue realizations at the fiducial HOD catalogues. To this covariance, we added a small constant diagonal term ($8.0 \times 10^{-5}$) to improve numerical stability. We limit our comparison to $k < 1.0$. In both DIFFHOD and standard cases, we use the same halo catalogue used to construct the mock observations, so that the only source of error is variation caused by the HOD model. For this analysis we impose wide Gaussian priors, $\mathcal{N}(\mu, \sigma^2)$, where $\mu$ is the mean, and $\sigma^2$ is the variance, on our parameter values as follows:

$$\log M_{\rm min} \sim \mathcal{N}(12.0, 0.5), \sigma_{\log M} \sim \mathcal{N}(0.25, 0.2),$$
$$\log M_0 \sim \mathcal{N}(11.25, 0.5), \log M_1 \sim \mathcal{N}(13.20, 0.5),$$
$$\alpha \sim \mathcal{N}(1.0, 0.2).$$

To derive the posteriors using DIFFHOD, we sample over the HOD parameter using HMC (Duane et al. ; Neal et al. 2011). We use the NoUTurn HMC implementation (Hoffman, Gelman et al. 2014) in TensorFlow Probability. We use three chains initialized around our fiducial HOD parameters with over 1000 steps (300 steps of burn-in).

For the standard approach, we run the same analysis using the standard HOD. However, since this implementation does not allow easy differentiation, we cannot use HMC instead use an MCMC analysis. We use the emcee (Foreman-Mackey et al. 2013)

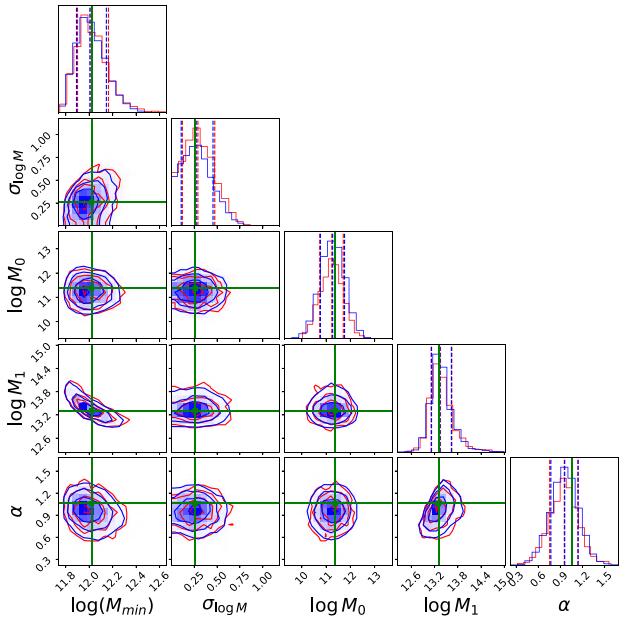


**Figure 5.** Posterior distributions of the Zheng et al. (2007) HOD parameters derived using DIFFHOD with HMC (red) and the standard HOD with MCMC (blue). We use a single mock galaxy catalogue realization at the fiducial HOD parameter values (green) as our observations (https://github.com/DifferentiableUniverseInitiative/DHOD/blob/master/nb/chain_scripts/chain_analysis_5d.ipynb).

**Table 1.** Posterior values from the HOD analyses using the DIFFHOD and standard HOD model. Uncertainties are estimate from the 16 and 84 per cent quantiles.

| | DIFFHOD | Standard HOD |
|---|---|---|
| $\log(M_{\rm min})$ | $12.03^{+0.15}_{-0.03}$ | $12.01^{+0.14}_{-0.02}$ |
| $\sigma_{\log M}$ | $0.28^{+0.19}_{-0.16}$ | $0.27^{+0.18}_{-0.16}$ |
| $\log M_0$ | $11.25^{+0.38}_{-0.43}$ | $11.27^{+0.49}_{-0.50}$ |
| $\log M_1$ | $13.32^{+0.23}_{-0.23}$ | $13.34^{+0.29}_{-0.22}$ |
| $\alpha$ | $0.96^{+0.19}_{-0.19}$ | $0.96^{+0.18}_{-0.19}$ |

implementation with 10 walkers and 6000 steps. We present the posteriors on the HOD parameters for the DIFFHOD (black) and standard (blue) analyses in Fig. 5 and list the median posterior values in Table 1 with associated errors calculated from percentiles. We mark the fiducial ('true') HOD values in green. As we are using a single galaxy realization for our mock observations, we expect some variation between the best-fitting parameters and the true values. The posteriors derived using DIFFHOD and HMC is in excellent agreement with the posteriors derived using the standard HOD and MCMC.

Our HMC analysis takes approximately 10 h on a single Tesla V100-PCIE-32Gb GPU. Meanwhile, the standard approach takes substantially more time to get comparable results: ∼200 h on 1 CPU – ∼20× slower than our DIFFHOD analysis. Some of this improvement is due to the fact that our DIFFHOD implementation is faster per iteration than the standard HOD implementation (∼1 and ∼4 s per iteration, respectively). Most of the improvement, however, comes from the fact that DIFFHOD allows us to exploit a more efficient gradient-based method to derive the posterior.

We compare the DIFFHOD and standard approaches in more detail by comparing the effective sample size of each chain per function

evaluation (Gelman et al. 2013). This is roughly equivalent to the number of independent samples per evaluation and this measure has the advantage of being independent of particular implementation speed and platform (i.e. GPU versus CPU). The effective sample size incorporates information about the autocorrelations within a chain; i.e. it accounts for the dependent relationships between the samples. We calculate it from the output of the Markov chain:

$$N_{\rm eff} = \frac{N}{1 + \sum_{t=1}^{\infty} \rho_t}, \tag{20}$$

where $N$ is the number of samples in the chain and $\rho_t$ is the autocorrelation of length $t$. Averaging over all HOD parameters, we find a mean effective sample size of 524.2 for our DIFFHOD HMC evaluation and 403.9 for the standard MCMC evaluation. This corresponds to an effective sample of 0.05 per evaluation for the HMC and 0.006 per evaluation for the MCMC.

## 4 CONCLUSIONS

In this work, we have constructed a differentiable stochastic HOD model going from a halo catalogue to an observed galaxy power spectrum. This allows us to use derivative-based optimization methods to quickly optimize for the underlying model parameters. This is the first time that differentiable stochastic models have been used in the astrophysics literature. We find that the DIFFHOD model provides a 4× increase of speed versus the same analysis performed via MCMC with the standard HOD implementation. DIFFHOD is an alternative to a number of recent works focusing on emulating galaxy clustering statistics (Kwan et al. 2015; Wibking et al. 2019; Kobayashi et al. 2020; Wibking et al. 2020; Hearin et al. 2022). Unlike emulator-based methods, DIFFHOD works at the level of the halo catalogue and allows fast, differentiable, generation of any summary statistic with respect to the HOD parameters.

In this work, we have focused on the Zheng et al. (2007) HOD model, but our methods can be easily extended to a broad class of models. While standard HODs are based only on halo mass, in general various properties of the haloes' environment and formation history could affect the galaxy properties (Zhu et al. 2006; Croton, Gao & White 2007). Galaxy assembly bias has been argued (Feldmann & Mayer 2015; Hadzhiyska et al. 2020) to cause significant deviations between predictions of standard HOD models and those from hydrodynamical simulations. Decorated HOD models have been introduced to account for assembly bias (Hearin et al. 2016) and have been extended to include other possible effects (Yuan, Eisenstein & Garrison 2018). These models still rely on stochastic discrete sampling for assigning centrals and satellites, so they can be modelled in a differentiable way using the techniques described in this work. In practice, since our code separates out the stochastic element of the occupation via the Gumbel-Softmax trick, any analytical formula for the underlying occupation distribution that depends on a calculable (sub)halo property can be trivially included. At the level of our implementation in TensorFlow, extended HOD models are as easy to implement as changing the defined $N_{\rm sat}$ and $N_{\rm cen}$ statistics to the chosen formula.

As the dimensionality of our problem increases, either with extended HOD models or with joint analysis with cosmological parameters, we expect the relative performance of derivative-based methods, like HMC, over pure sampling-based methods to further improve (Neal et al. 2011).

Differentiable HOD models have even more apparent applications in the case of dynamical forward model large-scale reconstructions (Seljak et al. 2017) when paired with efficient differentiable halo finding methods (Modi et al. 2018, 2021; Kodi Ramanah, Charnock & Lavaux 2019). While it is possible to perform these reconstructions by interpreting the galaxy field as a biased version of the dark matter field (i.e. in Horowitz et al. 2021), inaccuracies in this prescription will result in biases that would be difficult to account for in cosmological constraints. Through joint inference of the HOD parameters with the initial density field, these astrophysical uncertainties can be rigorously marginalized out. Differentiable models are critical for this application as the optimization is highly multidimensional (approximately number of particles in the simulation) and would be computationally infeasible without gradient-based methods.

While in this work we have highlighted using our DIFFHOD model inside an HMC framework, one can exploit its automatic differentiation for a variety of first-order optimization and parameter inference methods. For example, standard variational inference relies on having well-defined derivatives for the optimization of latent space parameters describing the likelihood surface (Peterson 1987; Beal 2003; Blei et al. 2016). Variational inference could further accelerate parameter inference when compared to HMC or nested sampling methods (Gunapati et al. 2018).

## ACKNOWLEDGEMENTS

We thank Andrew Hearin and Matt Becker for very useful discussions and suggestions. BH and CH were supported by the AI Accelerator programme of the Schmidt Futures Foundation. SF was supported by the Physics Division of Lawrence Berkeley National Laboratory. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DEC02-05CH11231.

## DATA AVAILABILITY

No new data were generated or analysed in support of this research. All code related to this work, including scripts to generate figures, is available at https://github.com/DifferentiableUniverseInitiative/DHOD.

## REFERENCES

Abid A., Balin M. F., Zou J., 2019, preprint (arXiv:1901.09346)
Beal M. J., 2003, PhD thesis, Univ. London
Benson A. J., Cole S., Frenk C. S., Baugh C. M., Lacey C. G., 2000a, MNRAS, 311, 793
Berlind A. A., Weinberg D. H., 2002, ApJ, 575, 587
Beutler F. et al., 2017, MNRAS, 466, 2242
Blei D. M., Kucukelbir A., McAuliffe J. D., 2016, preprint (arXiv:1601.00670)
Bond J. R., Cole S., Efstathiou G., Kaiser N., 1991, ApJ, 379, 440
Cooray A., Sheth R., 2002, Phys. Rep., 372, 1
Crain R. A. et al., 2009, MNRAS, 399, 1773
Croton D. J., Gao L., White S. D. M., 2007, MNRAS, 374, 1303
Cuesta A. J. et al., 2016, MNRAS, 457, 1770
DESI Collaboration, 2016, preprint (arXiv:1611.00036)
Desjacques V., Jeong D., Schmidt F., 2018, Phys. Rep., 733, 1
Dillon J. V. et al., 2017, preprint (arXiv:1711.10604)
Duane S., Kennedy A. D., Pendleton B. J., Roweth D., 1987a, Phys. Lett. B, 195, 216
Feldmann R., Mayer L., 2015, MNRAS, 446, 1939
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306
Gabrié M., Rotskoff G. M., Vanden-Eijnden E., 2022, Proc. Natl. Acad. Sci., 119, e2109420119

Gelman A., Carlin J. B., Stern H. S., Dunson D. B., Vehtari A., Rubin D. B., 2013, Bayesian Data Analysis. CRC Press
Gumbel E. J., 1954, Statistical Theory of Extreme Values and Some Practical Applications: A Series of Lectures, Vol. 33. US Government Printing Office
Gunapati G., Jain A., Srijith P. K., Desai S., 2018, preprint (arXiv:1803.06473)
Hadzhiyska B., Bose S., Eisenstein D., Hernquist L., Spergel D. N., 2020, MNRAS, 493, 5506
Hearin A. P., Chaves-Montero J., Becker M. R., Alarcon A., 2021, preprint (arXiv:2105.05859)
Hearin A. P., Ramachandra N., Becker M. R., DeRose J., 2022, Open J. Astrophys., 5, 3
Hearin A. P., Zentner A. R., van den Bosch F. C., Campbell D., Tollerud E., 2016, MNRAS, 460, 2552
Hoffman M. D., Gelman A. et al., 2014, J. Mach. Learn. Res., 15, 1593
Horowitz B., Zhang B., Lee K.-G., Kooistra R., 2021, ApJ, 906, 110
Ivanov M. M., Simonović M., Zaldarriaga M., 2020, J. Cosmol. Astropart. Phys., 2020, 042
Jang E., Gu S., Poole B., 2016, preprint (arXiv:1611.01144)
Jimenez Rezende D., Mohamed S., Wierstra D., 2014, preprint (arXiv:1401.4082)
Kingma D. P., Welling M., 2013, preprint (arXiv:1312.6114)
Klypin A. A., Trujillo-Gomez S., Primack J., 2011, ApJ, 740, 102
Kobayashi Y., Nishimichi T., Takada M., Takahashi R., Osato K., 2020, Phys. Rev. D, 102, 063504
Kodi Ramanah D., Charnock T., Lavaux G., 2019, Phys. Rev. D, 100, 043515
Kravtsov A. V., Berlind A. A., Wechsler R. H., Klypin A. A., Gottlöber S., Allgood B., Primack J. R., 2004, ApJ, 609, 35
Kwan J., Heitmann K., Habib S., Padmanabhan N., Lawrence E., Finkel H., Frontiere N., Pope A., 2015, ApJ, 810, 35
Lemson G., Kauffmann G., 1999, MNRAS, 302, 111
Maddison C. J., Mnih A., Whye Teh Y., 2016, preprint (arXiv:1611.00712)
Maddison C. J., Tarlow D., Minka T., 2014, preprint (arXiv:1411.0030)
Mann R. G., Peacock J. A., Heavens A. F., 1998, MNRAS, 293, 209
Mehrtens N. et al., 2016, MNRAS, 463, 1929
Modi C., Feng Y., Seljak U., 2018, J. Cosmol. Astropart. Phys., 2018, 028
Modi C., Lanusse F., Seljak U., 2021, Astronomy and Computing, 37, 100505
Modi C., Li Y., Blei D., 2023, Journal of Cosmology and Astroparticle Physics, 03, 059
Modi C., White M., Slosar A., Castorina E., 2019, Journal of Cosmology and Astroparticle Physics, 11, 023
Morgan P., 2018, Probabilistic Programming in TensorFlow
Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493
Neal R. M. et al., 2011, Handbook of Markov Chain Monte Carlo, 2, 2
Neistein E., Weinmann S. M., Li C., Boylan-Kolchin M., 2011, MNRAS, 414, 1405
Peacock J. A., Smith R. E., 2000, MNRAS, 318, 1144
Peterson C., 1987, Complex Syst., 1, 995
Press W. H., Schechter P., 1974, ApJ, 187, 425
Robotham A. S. G., Howlett C., 2018, Res. Notes Am. Astron. Soc., 2, 55
Rodríguez-Torres S. A. et al., 2016, MNRAS, 460, 1173
Schmidt F., Elsner F., Jasche J., Nguyen N. M., Lavaux G., 2019, J. Cosmol. Astropart. Phys., 2019, 042
Scoccimarro R., Sheth R. K., Hui L., Jain B., 2001, ApJ, 546, 20
Seljak U., 2000, MNRAS, 318, 203
Seljak U., Aslanyan G., Feng Y., Modi C., 2017, J. Cosmol. Astropart. Phys., 2017, 009
Sinha M., Berlind A. A., McBride C. K., Scoccimarro R., Piscionere J. A., Wibking B. D., 2018, MNRAS, 478, 1042
Spergel D. et al., 2015, preprint (arXiv:1503.03757)
Takada M. et al., 2014, PASJ, 66, R1
Tamura N. et al., 2016, in Evans C. J., Simard L., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 9908, Ground-Based and Airborne Instrumentation for Astronomy VI. SPIE, Bellingham, p. 99081M
To C. et al., 2021, Phys. Rev. Lett., 126, 141301
Tran D., Kucukelbir A., Dieng A. B., Rudolph M., Liang D., Blei D. M., 2016, preprint (arXiv:1610.09787)
van den Bosch F. C., Yang X., Mo H. J., 2003, MNRAS, 340, 771
Wang Y. et al., 2022, ApJ, 928, 1
Wechsler R. H., Tinker J. L., 2018, ARA&A, 56, 435
Wechsler R. H., Zentner A. R., Bullock J. S., Kravtsov A. V., Allgood B., 2006, ApJ, 652, 71
White M. et al., 2011, ApJ, 728, 126
Wibking B. D. et al., 2019, MNRAS, 484, 989
Wibking B. D., Weinberg D. H., Salcedo A. N., Wu H.-Y., Singh S., Rodríguez-Torres S., Garrison L. H., Eisenstein D. J., 2020, MNRAS, 492, 2872
Yuan S., Eisenstein D. J., Garrison L. H., 2018, MNRAS, 478, 2019
Zehavi I. et al., 2005, ApJ, 630, 1
Zehavi I. et al., 2011, ApJ, 736, 59
Zheng Z. et al., 2005, ApJ, 633, 791
Zheng Z., Coil A. L., Zehavi I., 2007, ApJ, 667, 760
Zheng Z., Zehavi I., Eisenstein D. J., Weinberg D. H., Jing Y. P., 2009, ApJ, 707, 554
Zhu G., Zheng Z., Lin W. P., Jing Y. P., Kang X., Gao L., 2006, ApJ, 639, L5

## APPENDIX A: DISTRIBUTIONAL PROPERTIES OF DIFFHOD MODEL

In the main body of the work, we sampled from a Bernoulli Distribution rather than a pure Poisson Distribution due to the existing analytical tools to relax the Bernoulli Distribution via the Gumbel-Softmax trick. This was demonstrated to be a valid approximation at the level of various summary statistics, such as HOD functions and resulting galaxy power spectrum. In this section, we show the HOD as a function of halo mass.

We sample our satellite DIFFHOD, at $\tau = 0.1$, and the standard satellite HOD using the parameters in the main text 100 times in order to attain reasonable number statistics at the high mass bins. We show our results in Fig. A1, finding quantitative good agreement between the models as calculated by their modal and variance properties. We calculate for each mass bin the 32, 50, and 68 per cent percentiles. Since DIFFHOD uses a relaxed distribution instead of sampling, it is possible to get non-integer number of satellites while for the standard HOD model all sampling is discretized. Qualitatively, we see a slight broadening of the distributions at the extreme high-mass end; however, this does not noticeably impact any of our resulting analysis due to the very small population of these extreme high mass haloes. Additional optimizations in terms of maximum satellite population and choice of temperature could be performed if this populations is of high interest.

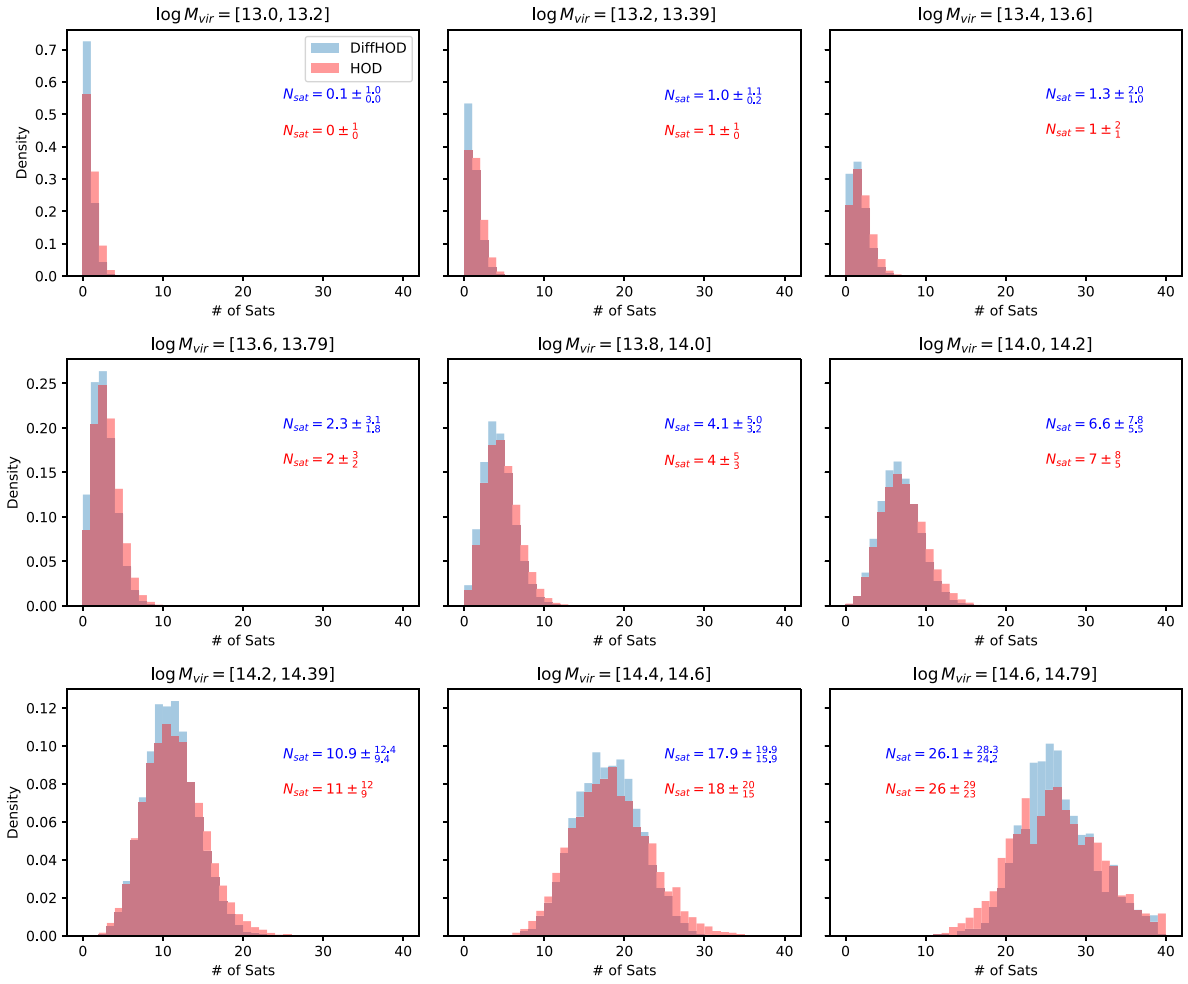


**Figure A1.** Histogram showing the HOD from 100 independent samples from the complete halo catalogue. We compare haloes populated by our DIFFHOD model as well as a standard HOD implementation. We show the median satellite occupation number, and quote error bars representing the 32 and 68 per cent quartiles. We find excellent quantitative and qualitative agreement between the two distributions. `fa fa-file-code-o`.

This paper has been typeset from a TEX/LATEX file prepared by the author.