Likelihood-free inference with neural compression of DES SV weak lensing map statistics

Niall Jeffrey, Justin Alsing, François Lanusse

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **501,** 954–969 (2021) doi:10.1093/mnras/staa3594
Advance Access publication 2020 November 25

Likelihood-free inference with neural compression of DES SV weak lensing map statistics

Niall Jeffrey [1,2]★ Justin Alsing[3] and François Lanusse[4]

[1]*Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, Paris, France*
[2]*Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK*
[3]*Oskar Klein Centre for Cosmoparticle Physics, Department of Physics, Stockholm University, Stockholm SE-106 91, Sweden*
[4]*AIM, CEA, CNRS, Université Paris-Saclay, Université Paris Diderot, Sorbonne Paris Cité, F-91191 Gif-sur-Yvette, France*

Accepted 2020 November 12. Received 2020 October 26; in original form 2020 August 18

ABSTRACT

In many cosmological inference problems, the likelihood (the probability of the observed data as a function of the unknown parameters) is unknown or intractable. This necessitates approximations and assumptions, which can lead to incorrect inference of cosmological parameters, including the nature of dark matter and dark energy, or create artificial model tensions. Likelihood-free inference covers a novel family of methods to rigorously estimate posterior distributions of parameters using forward modelling of mock data. We present likelihood-free cosmological parameter inference using weak lensing maps from the Dark Energy Survey (DES) Science Verification data, using neural data compression of weak lensing map summary statistics. We explore combinations of the power spectra, peak counts, and neural compressed summaries of the lensing mass map using deep convolution neural networks. We demonstrate methods to validate the inference process, for both the data modelling and the probability density estimation steps. Likelihood-free inference provides a robust and scalable alternative for rigorous large-scale cosmological inference with galaxy survey data (for DES, Euclid, and LSST). We have made our simulated lensing maps publicly available.

Key words: gravitational lensing: weak – methods: statistical – large-scale structure of Universe.

1 INTRODUCTION

Likelihood-free inference allows us to infer unknown cosmological parameters by directly comparing observed data with forward-simulated mock data. In this powerful and flexible framework, the posterior probability of unknown parameters can be estimated without resorting to simplifying, and often unjustified, likelihood assumptions (such as Gaussianity), even when data models consist of complex combinations of signal, noise, and systematic error. Incorrect assumptions can lead to incorrect and misleading conclusions for the given cosmological question, can hide new physics, or can create artificial model tensions.

The aim of observational cosmology is often to infer the cosmological parameters or models from the structure of the density field of the Universe as it evolves with time. For the late Universe, non-linear evolution has led to a highly non-Gaussian density field, which cannot be statistically characterized solely by two-point statistics (e.g. power spectra). The gravitational lensing effect on images of distant galaxies by the intervening large-scale structure provides a powerful probe of cosmology in this regime, through both structure formation and the geometry of the Universe.

In this work, we use measured statistics of the reconstructed projected density field (known as *mass maps*) from Dark Energy Survey (DES, Flaugher et al. 2015; Dark Energy Survey Collaboration 2016) weak gravitational lensing data to infer cosmological parameters of the Λ-cold dark matter (ΛCDM) model in a likelihood-free analysis. We use deep learning methods with the aim of extracting the optimal compressed statistic from our chosen data/statistic; this method is known as *neural compression*.

The physics of non-linear cosmological structure formation is included in our forward-modelled mock data using simulated (approximate N-body) density fields, with the lensing map observables calculated through ray tracing, and subsequent inclusion of (complicated) masks and non-Gaussian noise contributions corresponding to the DES data. The posterior probability densities for unknown parameters $\boldsymbol{\theta}$ are then estimated for different lensing map statistics from observed (unsimulated) data $\boldsymbol{d}_o$, without the need for an assumed analytic expression for the likelihood function $\mathcal{L}(\boldsymbol{\theta}) = p(\boldsymbol{d}_o|\boldsymbol{\theta})$.

The map-based statistics used in this likelihood-free analysis are: (1) the angular power spectrum of the map; (2) the peak statistics, also known as peak counts (Dietrich & Hartlap 2010; Kacprzak et al. 2016); (3) the joint statistic of peaks and power spectrum; and (4) convolutional neural network (CNN) compressed map statistics, aiming to compress the reconstructed mass map to optimal summary statistics using deep learning (Fluri et al. 2018, 2019; Ribli et al. 2019).

We note that the choice of mapping method, in our case Kaiser–Squires (Kaiser & Squires 1993), effectively corresponds to an initial data compression step, compressing a catalogue of images into a pixelized estimated mass map. Note that peak counts estimated from a different map reconstruction method may lead to somewhat different constraints on the cosmological parameters, owing to different information retainment associated with different reconstructions methods.

Likelihood-free inference provides an alternative inference framework for current and upcoming galaxy surveys [e.g. Euclid: Amen-

★ E-mail: niall.jeffrey@phys.ens.fr

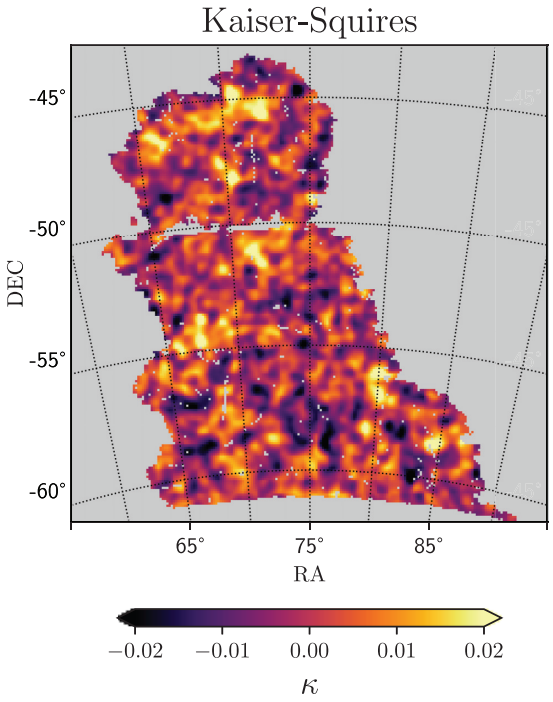

Figure 1. Convergence κ map from the DES SV weak lensing data reconstructed using Kaiser–Squires with $\sigma = 10$ arcmin Gaussian smoothing to reduce the impact of galaxy shape noise.

dola et al. 2016, or the Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST): LSST Science Collaboration 2009], relying only on our ability to forward-simulate mock data and unfettered by the need for a closed-form likelihood function. Given their appeal, it is important to understand the challenges presented by such methods, and to establish procedures to carefully validate the results.

Given current reported tensions in cosmological parameters between different data sets, such as the Hubble parameter H_0 between early- and late-Universe probes (Feeney, Mortlock & Dalmasso 2018) or the amplitude of fluctuations between galaxy surveys (Abbott et al. 2018; Joudaki et al. 2018) and the cosmic microwave background (Planck Collaboration VI 2020), likelihood-free inference offers a novel inference framework without restrictive assumptions about the likelihood or data model. Likelihood-free inference also provides a simple and robust framework for combining observed summary statistics from different data sets with forward modelling, and avoids many of the potentially difficult technical aspects of current standard analysis (such as covariance matrix estimation or sampling from high-dimensional Bayesian hierarchical models).

The mass maps in this work (Fig. 1) are generated from public DES Science Verification (SV) data (Chang et al. 2015) using the linear Kaiser–Squires method. Though the SV data cover an area of only approximately 5 per cent of the final DES sky footprint, the observations are to the approximate final depth of the full survey, so the SV data match what will be the final galaxy density and lensing signal-to-noise per pixel.

In Section 2, we introduce the formalism of likelihood-free inference and data compression methods. In Section 3, we include an overview of the relevant aspects of weak gravitational lensing with galaxies. In Section 4, we detail our map reconstruction method and the three summary statistics used: power spectrum, peak statistics, and CNN compressed map summaries. In Section 5, we discuss the DES SV data and the forward modelling of mock simulated data. In Section 6, we discuss the results and validation of the likelihood-free parameter inference using power spectra and peak statistics. In Section 7, we discuss results using likelihood-free inference with compressed summaries directly from the mass map using CNNs.

2 INFERENCE

2.1 Motivation

In Bayesian parameter inference, we aim to evaluate the posterior probability distribution

$$p(\boldsymbol{\theta}|\boldsymbol{d}_o, \mathcal{M}) = \frac{p(\boldsymbol{d}_o|\boldsymbol{\theta}, \mathcal{M})\ p(\boldsymbol{\theta}|\mathcal{M})}{p(\boldsymbol{d}_o|\mathcal{M})} \tag{1}$$

for some statistical model $\mathcal{M}$ with associated unknown model parameters $\boldsymbol{\theta}$, given some observed data (or summary statistics of the observed data) $\boldsymbol{d}_o$ (see Jaynes 2003 for details).

To evaluate the relative probabilities of different models given the observed data summaries, one may also wish to evaluate the Bayesian evidence, also known as the marginal likelihood, given by

$$p(\boldsymbol{d}_o|\mathcal{M}) = \int p(\boldsymbol{d}_o|\boldsymbol{\theta}, \mathcal{M})\ p(\boldsymbol{\theta}|\mathcal{M})\ \mathrm{d}^n\boldsymbol{\theta} \tag{2}$$

for probability densities of continuous variables $\boldsymbol{\theta}$.

For both of these tasks, knowledge of the likelihood function $\mathcal{L}(\boldsymbol{\theta}) = p(\boldsymbol{d}_o|\boldsymbol{\theta}, \mathcal{M})$ – the sampling distribution for the observations as a function of the model parameters – is required. In this paper, we will focus on parameter inference. Despite the central role of the likelihood function, in general the sampling distribution $p(\boldsymbol{d}|\boldsymbol{\theta})$ for data (or summary statistics) $\boldsymbol{d}$ is not necessarily readily available (or tractable).

For parameter inference from large cosmological surveys, especially weak lensing surveys (Kilbinger et al. 2013; Abbott et al. 2018; Joudaki et al. 2018), the conditional distribution $p(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{M})$ is not generally known exactly, owing to non-linear evolution of the underlying density field, and any number of complicated observational effects (survey masks, various systematic biases, non-Gaussian noise contributions, etc.). For two-point statistics of the weak lensing field (see Section 3), a Gaussian sampling distribution is typically assumed,[1] though the second-order moments (i.e. two-point statistics) have a skewed distribution even for an underlying Gaussian lensing field (Sellentin & Heavens 2018; Sellentin, Heymans & Harnois-Déraps 2018; Taylor et al. 2019).

For higher order statistics of the lensing field, which are necessary to extract information beyond the Gaussian component of the field, there is typically no closed-form expression for their sampling distribution (and hence likelihood function), inhibiting their use for cosmological parameter inference. Even theoretical predictions for the expectation values of higher order statistics (e.g. for peak counts or deep CNN map summary statistics) must be estimated by forward-simulated mock realizations of the data. Sampling distributions for higher order statistics are not expected to be Gaussian, with non-Gaussianity arising from non-linear combinations (e.g. counting peaks or deep CNN statistics) of an already non-Gaussian cosmological lensing field, compounded by non-Gaussian shape noise and Poissonian shot noise in the data.

[1]Exceptions are Bayesian hierarchical analyses, which account for the non-Gaussian data model (Alsing et al. 2016; Alsing, Heavens & Jaffe 2017).

Arguments that rely on the central-limit theorem or the principle of maximum entropy, both of which give the Gaussian distribution special status,[2] cannot avoid the fact that if one assumes an incorrect distribution $p(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{M})$, the resulting statistical inference may be misleading or biased.

2.2 Likelihood-free inference with forward modelling

Contrary to what the name might suggest, likelihood-free inference (also known as simulation-based inference) does not exclude a likelihood. The distribution $p(\boldsymbol{d}|\boldsymbol{\theta})$[3] is not used in closed-form, but is reconstructed from simulated mock data as part of the inference pipeline.

$\boldsymbol{d}_o$In density estimation likelihood-free inference (Papamakarios & Murray 2016; Alsing, Wandelt & Feeney 2018), the inference task is posed as a density estimation problem. One can picture the simulated mock data $\boldsymbol{d}$ realizations and their associated parameters $\boldsymbol{\theta}$ as forming a cloud of points in $\{\boldsymbol{d},\ \boldsymbol{\theta}\}$ space. In this space, we could estimate the following distributions: (A) the joint $p(\boldsymbol{d}, \boldsymbol{\theta})$; (B) the conditional $p(\boldsymbol{\theta}|\boldsymbol{d})$, which would give the posterior if evaluated for observed data $\boldsymbol{d}_o$; and (C) the conditional $p(\boldsymbol{d}|\boldsymbol{\theta})$, which would give the likelihood if evaluated for observed data . All of the above can be straightforwardly used to reconstruct the desired posterior density.

$\boldsymbol{d}$Estimating either (A) $p(\boldsymbol{d}, \boldsymbol{\theta})$ or (B) $p(\boldsymbol{\theta}|\boldsymbol{d})$, which are both densities with respect to $\boldsymbol{\theta}$, requires that the distribution of $\boldsymbol{\theta}$ in the set of simulated mocks must come from the prior distribution $p(\boldsymbol{\theta})$ (see Alsing et al. 2019). To avoid this constraint, one can take the strategy (C) of estimating the sampling distribution $p(\boldsymbol{d}|\boldsymbol{\theta})$ as a function of the model parameters, which is a probability density in rather than $\boldsymbol{\theta}$. This has the advantage that the simulated parameter values do not need to be drawn from the prior, enabling the use of various strategies for optimizing (and reducing) the set of simulations to be run for the problem at hand. It also allows for seamless subsequent analyses of alternative prior assumptions once the sampling distribution (and hence likelihood) has been learned.

Once the density $p(\boldsymbol{d}|\boldsymbol{\theta})$ is learned from our simulated mocks, it can be evaluated at the observed data $\boldsymbol{d}_o$ and treated as a likelihood $\mathcal{L}(\boldsymbol{\theta})$ in the usual way for parameter inference (equation 1). To the extent that the physics and the effects of the realistic data model are correctly included in the forward simulations, the estimated $p(\boldsymbol{d}|\boldsymbol{\theta})$ will be the correct distribution to be used for parameter inference (provided sufficiently many simulations to accurately learn the sampling distribution).

The density estimation approach to likelihood-free inference is an alternative to the more traditional Approximate Bayesian Computation (ABC, Rubin 1984; Marin et al. 2011), which uses (adaptive) rejection-based sampling (for recent applications in astronomy, see Cameron & Pettitt 2012; Schafer & Freeman 2012; Weyant, Schafer & Wood-Vasey 2013; Robin et al. 2014; Akeret et al. 2015; Ishida et al. 2015; Lin & Kilbinger 2015; Davies & Furlanetto 2016; Jennings, Wolf & Sako 2016; Carassou et al. 2017; Hahn et al. 2017; Kacprzak et al. 2018; Fagioli et al. 2020; Tortorelli et al. 2020).

In ABC, one draws parameters from some (typically adaptive) proposal density, and forward simulates data. Those simulated data are then compared to the observed data under some distance metric, and the proposed parameters accepted if the distance is below some threshold ϵ. The resulting accepted samples then constitute samples from an approximate posterior, approaching the exact posterior only in the (unattainable limit) $\epsilon \to 0$. In practice, even with sophisticated adaptive methods the use of rejection sampling means that the vast majority of samples get rejected, making for a simple but simulation inefficient approach.

Density-estimation likelihood-free methods overcomes this limitation by using all simulated data for improving the inference, and does not require a (subjective) distance metric and threshold, allowing for high-fidelity posterior inference with far fewer simulations (Papamakarios & Murray 2016; Alsing et al. 2018).

2.3 Neural density estimation

Here, we introduce the main mathematical aspects of the neural density estimators (NDEs) used in this work. For a more comprehensive discussion, see Alsing et al. (2019). Alternatively, the reader can proceed to Section 2.4 to skip the technical details of neural density estimation.

To estimate the conditional distribution $p(\boldsymbol{d}|\boldsymbol{\theta})$, we use the PYDELFI (Alsing et al. 2019) package (see Appendix B) with an ensemble of NDEs: neural network parametrizations of conditional probability densities. Specifically we use a combination of Gaussian Mixture Density Networks (MDN; Bishop 1994) and Masked Autoregressive Flows (MAF; Papamakarios, Pavlakou & Murray 2017).

These are, of course, not the only choice of NDE. For example, Diaz Rivero & Dvorkin (2020) recently demonstrated the characterization of distributions of weak lensing power spectra using the alternative FFJORD[4] NDE (Grathwohl et al. 2019, among other applications). The vibrant field of neural density estimation, Normalizing Flows in particular, will likely lead to further breakthroughs for probability density estimation and likelihood-free inference in the near future (Papamakarios et al. 2019a).

For both of our neural density estimation methods, MDN and MAF, the networks are trained to give an estimate $q(\boldsymbol{d}|\boldsymbol{\theta};\boldsymbol{\varphi})$ of the target distribution $p(\boldsymbol{d}|\boldsymbol{\theta})$, interpretable as

$$p(\boldsymbol{d}|\boldsymbol{\theta}) \approx q(\boldsymbol{d}|\boldsymbol{\theta};\boldsymbol{\varphi}), \tag{3}$$

by varying the $\boldsymbol{\varphi}$ parameters (e.g. weights and biases) of the network.[5] This is achieved by minimizing the loss function

$$U(\boldsymbol{\varphi}) = -\sum_{n=1}^{N} \log q(\boldsymbol{d}_n|\boldsymbol{\theta}_n;\boldsymbol{\varphi}) \tag{4}$$

over the N forward-modelled mock data $\boldsymbol{d}_n$. This loss corresponds to minimizing the Kullback–Leibler divergence (Kullback & Leibler 1951), a measure of difference or change going from the estimate q to the target $p(\boldsymbol{d}|\boldsymbol{\theta})$.

Gaussian MDNs represent the conditional density as a sum of K Gaussian components with mean $\boldsymbol{\mu}(\boldsymbol{\theta};\boldsymbol{\varphi})_k$, covariance $\mathbf{C}(\boldsymbol{\theta};\boldsymbol{\varphi})_k$, and component weights $\boldsymbol{r}(\boldsymbol{\theta};\boldsymbol{\varphi})_k$ all taken as unknown functions of the parameters $\boldsymbol{\theta}$, parametrized by a neural network:

$$q(\boldsymbol{d}|\boldsymbol{\theta};\boldsymbol{\varphi}) = \sum_{k=1}^{K} \boldsymbol{r}(\boldsymbol{\theta};\boldsymbol{\varphi})_k \mathcal{N}\left[\boldsymbol{d} \mid \boldsymbol{\mu}(\boldsymbol{\theta};\boldsymbol{\varphi})_k,\ \boldsymbol{C}(\boldsymbol{\theta};\boldsymbol{\varphi})_k\right]. \tag{5}$$

The second density estimation method uses Normalizing Flows. These use a series of bijective (and therefore invertible) functions to transform from simple known densities (e.g. the unit normal) to the target density (Jimenez Rezende & Mohamed 2015; Kingma et al. 2016). MAFs represent q as a transformation of a unit normal

[2]See Jaynes (2003) for details.

[3]From here on, we will drop the explicit model $\mathcal{M}$ dependence for brevity.

[4]Free-form Continuous Dynamics for Scalable Reversible Generative Models.

[5]See Goodfellow, Bengio & Courville 2016 for an introduction to neural networks.

through a series of autoregressive functions (Papamakarios et al. 2017; Papamakarios, Sterratt & Murray 2019b).

Masked Autoencoders for Distribution Estimation (MADEs, Germain et al. 2015) are autoregressive density estimators as they parametrize the estimate in terms of 1D conditionals. The density is factorized as

$$p(\boldsymbol{d}|\boldsymbol{\theta}) = \prod_{i}^{\dim(\boldsymbol{d})} p(d_i|d_{1:i-1}, \boldsymbol{\theta}) \tag{6}$$

by masking the weights in the neural network. In this way the factorized probability for d_1 may depend only on $\boldsymbol{\theta}$, d_2 may depend on d_1 and $\boldsymbol{\theta}$, d_3 may depend on d_1, d_2 and $\boldsymbol{\theta}$, and so on. In an MADE, each estimated conditional is modelled as a Gaussian whose mean and variance are functions of $(d_{1:i-1},\ \boldsymbol{\theta})$ and are given by the neural network. The resulting function is a transformation to the target distribution from a space of random samples distributed according to a unit normal, where the associated Jacobian is triangular (due to the autoregressive structure) so can be easily calculated.

MAFs are composed of a series of MADEs, where the output of the last MADE is the input for the next. This allows for the estimation of more complicated densities that are not able to be factorized into a simple product of Gaussians (which the MADE requires). The repeated MADE layers in the MAF also allow the order of factorization to be shuffled, to better estimate general densities.

As shown in Section 6, we use an ensemble of different network architectures for both MDNs and MAFs to validate the density estimation. The final density estimation is a stack of the ensemble estimates, weighted by the loss evaluated during training (see Alsing et al. 2019 for more details).

2.4 Summary statistic compression

For a given number of simulated mock data sets, density estimation likelihood-free inference is exponentially more efficient the lower the dimensionality of the data vector (i.e. summary statistics) $\boldsymbol{d}$. For likelihood-free inference to be scalable when forward simulations are expensive, it is typically necessary to compress high-dimensional data vectors down to some informative low-dimensional summaries $\boldsymbol{t}$.

To this end, we want some compression function $\boldsymbol{t} = F(\boldsymbol{d})$, that is as informative as possible with respect to the unknown parameters whilst being as low-dimensional as possible. Under certain conditions, we can find a compression of a given data vector $\boldsymbol{d}$ down to $\boldsymbol{t}$ with dimension matching the number unknown parameters, $\dim(\boldsymbol{t}) = \dim(\boldsymbol{\theta})$, that is lossless at the level of the Fisher information (Alsing & Wandelt 2018).

Different approaches exist to try and achieve this compression, which are discussed in Section 6.2. Crucially though, a poor choice of compression scheme (which we of course try to avoid) would lead to less informative summaries, but not to biased results. Provided the same compression scheme is applied to both the observed data and simulated mocks in a self-consistent way, subsequent likelihood-free inference of the parameters will be unbiased. A poor compression will however lead to added scatter in the forward-modelled samples $\boldsymbol{t}$ for a given set of parameters $\boldsymbol{\theta}$, leading to inflated parameter constraints.

In this work, our results use neural compression; using a neural network to learn the compression function. In particular, for the power spectrum and peak count summary statistics we use a regression network and for the deep CNN map compression we use both regression- and information-based training strategies for the network (see Sections 6.2 and 7 for details). We note that alternatively, neural compression can be achieved by *information maximizing neural networks* (IMNN), which aim to maximize the Fisher information (Charnock, Lavaux & Wandelt 2018). We do not adopt the IMNN framework for this work however, as it requires specific simulations allowing for finite differences estimates of the gradients of summaries with respect to unknown parameters.

In this work, we take two approaches to neural compression. In the first approach, we initially compress the mass maps down to some 'first level summaries' (in this case, power spectra and peak counts), which we then feed into a simple dense neural network for subsequent (massive) compression. In the second approach, we compress the map directly down to some informative low-dimensional summaries using deep CNNs. In Section 4.4, we introduce the details of our deep CNN map compression and show the likelihood-free results with the DES SV data in Section 7.

3 WEAK GRAVITATIONAL LENSING

Weak gravitational lensing is one of the foremost probes of cosmological large-scale structure. By using measurements of the galaxy shapes distorted by foreground matter due to gravitational lensing, we can directly infer density fluctuations in the total foreground matter (including non-visible dark matter). For convenience, here we have summarized some of relevant literature for weak gravitational lensing (see Bartelmann & Schneider 2001 and Kilbinger 2015).

The weak lensing convergence κ is given by a weighted projection of the density along the line of sight from the observer to a point with radial comoving distance χ and angular position $\vec{\phi}$ on the sky

$$\kappa(\vec{\phi}, \chi) = \frac{3H_0^2\Omega_m}{2} \int_0^{\chi} \frac{\chi'(\chi - \chi')}{\chi} \frac{\delta(\vec{\phi}, \chi')}{a(\chi')}\, \mathrm{d}\chi'. \tag{7}$$

where H_0 is the present value of the Hubble parameter, a is the cosmological scale factor, Ω_m is the matter density parameter, δ is the overdensity, and the speed of light $c = 1$. We have assumed flatness, such that the cosmological global curvature is zero, $K = 0$.

For a radial (redshift) distribution $n(\chi)$ of lensed source galaxies, the convergence is given by

$$\begin{aligned}\kappa(\vec{\phi}) &= \int_0^{\infty} n(\chi)\kappa(\vec{\phi}, \chi)\mathrm{d}\chi \\ &= \frac{3H_0^2\Omega_m}{2} \int_0^{\infty} \mathrm{d}\chi' f(\chi')\chi' \frac{\delta(\vec{\phi}, \chi')}{a(\chi')},\end{aligned} \tag{8}$$

where

$$f(\chi') = \int_{\chi'}^{\infty} \left(\frac{\chi - \chi'}{\chi}\right) n(\chi)\mathrm{d}\chi. \tag{9}$$

The convergence for the distribution of source galaxies at angular position $\vec{\phi}$ on the sky is therefore given by

$$\kappa(\vec{\phi}) = \frac{3H_0^2\Omega_m}{2} \int_0^{\infty} \left[\int_0^{\chi} \frac{\chi'(\chi - \chi')}{\chi} \frac{\delta(\vec{\phi}, \chi')}{a(\chi')}\mathrm{d}\chi'\right] n(\chi)\mathrm{d}\chi. \tag{10}$$

The shear field γ is a spin-2 (Newman & Penrose 1966; Wallis et al. 2017) field on the celestial sphere and is related to the convergence field through the lensing potential[6] ψ:

$$\kappa = \frac{1}{4}(\eth\bar{\eth} + \bar{\eth}\eth)\psi \tag{11}$$

$$\gamma = \frac{1}{2}\eth\eth\psi, \tag{12}$$

[6]See Bartelmann & Schneider (2001) and Kilbinger (2015) for full details.

where the differential operators $\eth$ and $\bar{\eth}$ are spin-weight linear operators (Castro, Heavens & Kitching 2005) defined on the sphere. It is therefore possible to determine the full-sky shear field γ from the scalar convergence κ (up to a constant of integration), for example by use of spin-weight spherical harmonic transforms. In this work, such transformations were used during the creation of the ideal shear γ fields from convergence κ fields (defined on the sphere) derived from simulations (see Section 5).

In the weak lensing limit, the observed ellipticity of a galaxy $\epsilon_{\rm obs}$ is composed of both the intrinsic ellipiticity of the source galaxy ϵ_s plus the gravitational lensing shear γ. We therefore treat the measured ellipticity of a galaxy as an estimator for the 'shear', where the measurement is degraded by 'shape noise' caused by the intrinsic ellipticity. In matrix notation, we can express a linear model with a data vector of observed shear measurements

$$\boldsymbol{\gamma} = \boldsymbol{A}\boldsymbol{\kappa} + \boldsymbol{n}, \tag{13}$$

$\boldsymbol{n}$where $\boldsymbol{A}$ is the linear transformation between shear $\boldsymbol{\gamma}$ and convergence $\boldsymbol{\kappa}$ and is a vector of noise per pixel. In our formulation, the elements of the data vector are the galaxy shear measurements binned into pixels depending on their sky position.

4 SUMMARY STATISTICS

4.1 Kaiser–Squires map reconstruction

For smaller patches of sky, in the *flat-sky approximation* the $\eth$ operators on the sphere may be reduced to ∂ derivatives with respect to the two sky angles ϕ_1 and ϕ_2, so that shear γ may be related to convergence κ as

$$\tilde{\gamma}(\boldsymbol{k}) = \frac{k_1^2 - k_2^2 + 2ik_1k_2}{k_1^2 + k_2^2}\tilde{\kappa}(\boldsymbol{k}), \tag{14}$$

where k_1 and k_2 are elements of the 2D Fourier vector $\boldsymbol{k}$, so that

$$\gamma(\vec{\phi}) = \frac{1}{\pi}\int_{\mathbb{R}^2} \mathrm{d}^2\phi' \Gamma(\vec{\phi} - \vec{\phi}')\kappa(\vec{\phi}') \quad \text{where } \Gamma(\vec{\phi}) = -(\phi_1 - i\phi_2)^{-2}. \tag{15}$$

The Kaiser & Squires (1993) reconstruction method uses the pixelized observed $\boldsymbol{\gamma}$ map to reconstruct the unknown $\boldsymbol{\kappa}$ by inverting equation (14). This procedure to infer $\boldsymbol{\kappa}$ takes no account of varying shape noise in the shear map or masks, which introduce artefacts into the recovered convergence *mass map*. Fig. 1 shows the reconstruction for the DES SV data (as described in Section 5).

The Kaiser–Squires reconstruction, in addition to not accounting for spatially-varying noise, includes no explicit prior information about the signal.[7] This will not lead to incorrect inferences in the likelihood-free inference framework, as the anisotropic noise and mask will be forward modelled in the mock simulated data. Different reconstruction methods that use prior information about the signal have been shown to more accurately reconstruct the convergence field κ, either in closed-form (Marshall et al. 2002; Alsing et al. 2016, 2017; Lanusse et al. 2016; Jeffrey et al. 2018; Price et al. 2019) or implicitly learned using samples from the prior (Shirasaki, Yoshida & Ikeda 2019; Jeffrey et al. 2020). More accurate mapping methods would likely increase the signal-to-noise ratio of summary statistics, and therefore improve constraining power, and a study of this in the context of likelihood-free inference would merit future work.

[7]Implicitly, the reconstruction is the *maximum a posteriori* estimate under the assumption of Gaussian noise and a uniform prior $p(\kappa)$.

4.2 Power spectrum

The angular power spectrum for the convergence field κ on the celestial sphere is given by

$$\langle a_{\ell m} a^*_{\ell' m'} \rangle = C_\kappa(\ell)\delta_{mm'}\delta_{\ell\ell'}, \tag{16}$$

with spherical harmonic coefficients $a_{\ell m}$ of the convergence κ field, where we have used Kronecker delta $\delta_{mm'}$. The expectation, $\langle\rangle$, is with respect to random realizations.

For a given field on the sphere, an unbiased estimate of the power spectrum is given by

$$\hat{C}_\kappa(\ell) = \frac{1}{2\ell+1}\sum_{m=-\ell}^{m=+\ell} |a_{\ell m} a^*_{\ell m}|. \tag{17}$$

For a simple contiguous masked region (e.g. an octant of the sky), the measured power spectrum can be rescaled by the fraction of sky observed $f_{\rm sky}$ to give an approximate unbiased estimate; in this approximation $C_\ell \approx \langle f_{\rm sky}^{-1}\hat{C}_\ell\rangle$ (Dodelson 2003). The variance due to cosmic variance (i.e. finite m modes per ℓ) and additional sample variance due to finite sky coverage is then given by

$$\sigma^2_{C_\ell} = \frac{1}{f_{\rm sky}}\frac{2}{2\ell+1}C_\ell^2. \tag{18}$$

Fig. 2 shows the theoretical convergence power spectrum and the power spectrum measured from the L-PICOLA simulations (Howlett, Manera & Percival 2015). The source of the discrepancy between the two power spectra is discussed further in Section 5.2.1. However, the 1σ and 2σ confidence intervals in Fig. 2 correspond to a full sky cosmic variance uncertainty. For the 139 deg^2 area of the DES SV data, the actual 1σ region would be a factor of more than 10 greater.

As the data cover a small sky area we do not use spherical harmonics (similarly to the map making). Instead we measure the power spectrum of our complex Kaiser–Squires maps (i.e. combining the E- and B-modes) using fast Fourier transforms (Cooley & Tukey 1965). We bin the power spectrum into 18 Fourier band powers, with centres of $k_{\rm centre}/(10^{-3}\,\mathrm{arcmin}^{-1})$:

0.434	1.183	2.044	2.997	3.939	5.177
6.805	8.943	11.75	15.44	20.30	26.68
35.07	46.10	60.59	79.63	104.6	137.5

Since we are in the likelihood-free framework, we do not need to correct for this flat-sky approximation, as we perform the same operations self-consistently to the simulated mock data and the actual observed data.

Equivalently with the mask and noise in the power spectrum, we measure the 'raw' power spectrum from the maps in the same manner for both the simulated mock data and the actual observed data. Subtracting and rescaling the measured spectra to account for the mask and noise effects, essentially the same as a scalar *pseudo-*C_ℓ estimate (Hivon et al. 2002), is implicitly taken care of during the likelihood-free inference.

In this work, we have measured the unsmoothed Kaiser–Squires map $\kappa_{\rm KS} = \kappa_{\rm KS,E} + i\kappa_{\rm KS,B}$. This was found to perform marginally better with simulations than the $\kappa_{\rm KS,E}$ map alone. One could measure the power spectrum of both separately and combine them, thus keeping the maximum amount of information; this would increase the size of the summary statistic data vector, making the compression step more difficult, yet with little added constraining power. It is therefore possible that a spin *pseudo-*C_ℓ estimate of the E-mode power using both contributions would provide a more constraining

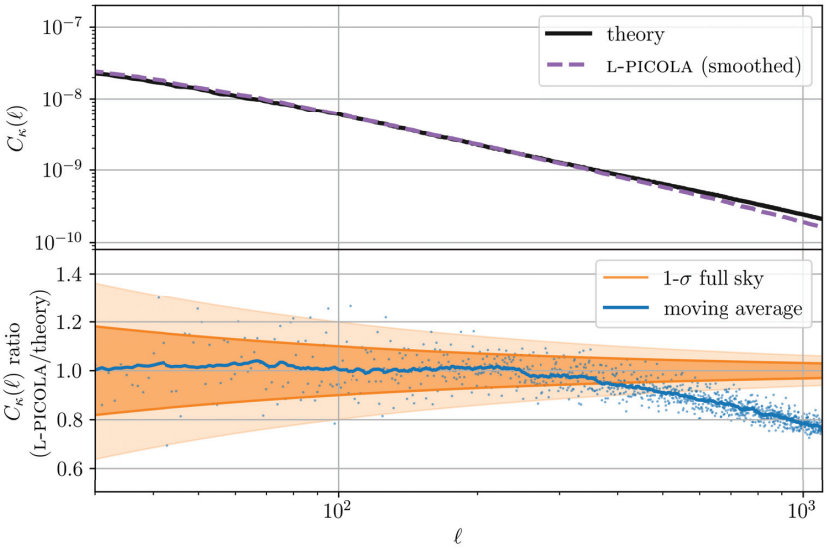

Figure 2. Example of a full-sky power spectrum of convergence $C_\kappa(\ell)$ from theory (NICAEA) and a single L-PICOLA simulation realization. The shaded 1σ and 2σ regions correspond to full-sky measurement uncertainty due to cosmic variance. For our case, the sky fraction of the observed data is less than 5×10^{-3}; the resulting sample variance will in fact give a confidence region that is a factor of more than 10 greater than the confidence region from full-sky cosmic variance alone. Nevertheless, we still correct the power spectra of the simulated convergence fields by rescaling the spherical harmonic coefficients $a_{\ell m}$ as described in Section 5.

summary statistic (Taylor et al. 2019), while keeping the size of the to-be-compressed data vector low. Similarly, a minimum-variance power spectrum estimate (e.g. Tegmark 1997) would provide the optimally-constraining power spectrum without increasing the size of the data vector in likelihood-free inference.

4.3 Peak statistics

For an isotropic Gaussian random field, the mean (zero by definition for our κ field) and the power spectrum are sufficient to entirely characterize the field. Of course, in the late-Universe, non-Gaussianity arises due to non-linear structure formation, for which summary statistics beyond the power spectrum are suited to provide additional information to constrain cosmological parameters.

Counts of the number of peaks in a mass map are particularly promising, as peaks in the density field probe the non-Gaussian structure directly in a manner that is sensitive to the changes in the cosmological parameters (Dietrich & Hartlap 2010; Peel et al. 2017; Martinet et al. 2018; Shan et al. 2018). In previous work, Lin & Kilbinger (2015) used a likelihood-free inference method, the ABC rejection sampling technique (rather than density estimation as in this work) applied to simulated data, in which the forward model was a fast halo approximation and, therefore, cannot easily be used for joint power spectrum and peak constraints without considerable adaptation.

We define our peak summary statistic as the number of pixels in the smoothed Kaiser–Squires map reconstruction that are of a value greater than all of their neighbours, which we bin according the convergence κ value of the peak, $n(\kappa)$. The smoothing scale of 10 arcmin was shown in Jeffrey et al. (2018) to give a map reconstruction closest to the underlying convergence κ field.

As discussed in Section 4.1, there are reconstruction methods beyond Kaiser–Squires that can be more accurate and would therefore likely give more informative peak statistic summaries, and these deserve future investigation. The choice of a given reconstruction method with a certain smoothing scale and definition of peak corresponds to a specific choice of summary statistic. The peak statistic analysis in this work is, therefore, not equivalent to Kacprzak et al. (2016), where different choices and definitions were used.

Fig. 3 shows the peak statistics measured from simulated mock data, showing our 11 bins between $\kappa = 0$ and 0.028. It was noted by Martinet et al. (2018) that adding peaks with negative κ did not lead to more informative summaries, ostensibly because they are strongly correlated with high-valued peaks. The validation of our simulated peak statistic summaries, as shown in Fig. 3, is discussed in Section 5.

4.4 Deep convolutional features

As another way to access the non-Gaussian information (beyond two-point) of the weak lensing field, deep convolutional networks have recently attracted significant attention. Instead of relying on crafted non-Gaussian statistics, like peak counts, which are based on a priori understanding of the physics, CNNs can be seen as flexible non-linear feature extractors; they can be optimized as to find maximally relevant summary statistics from the mass map. CNN outputs have been first applied as a cosmological summary statistic by Fluri et al. (2018, 2019) and Ribli et al. (2019).

CNNs are particularly suited for 2D image or 1D time-series data with translation invariant features in the underlying signal. Mathematically, they are a sequence of iteratively computed layers. At a given layer j, the signal $\boldsymbol{x}_j$ is computed from the previous layer

$$\boldsymbol{x}_j = \rho \boldsymbol{M}_j \boldsymbol{x}_{j-1} \tag{19}$$

with linear operators (i.e. convolutions) $\boldsymbol{M}_j$ and non-linear *activation function* ρ (LeCun et al. 1990; Mallat 2016). Deep architectures,

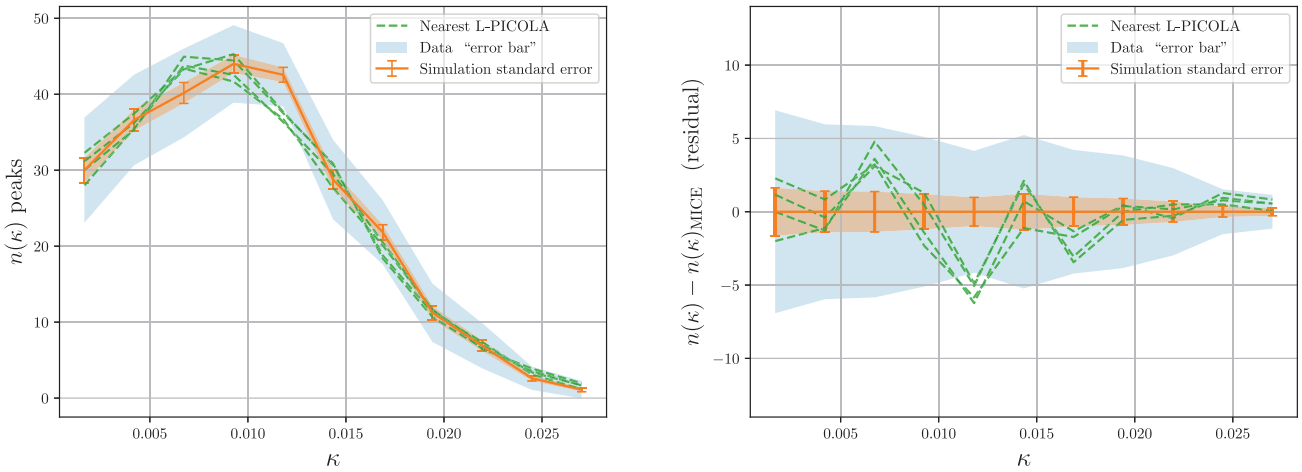

Figure 3. Validation of peak count summary statistic prediction from L-PICOLA simulations with respect to the prediction from MICE simulations. This shows the mean predicted L-PICOLA peak statistic from four points in parameter space closest to the MICE cosmological parameters. The left-hand panel shows the absolute peak count summary statistic and the right-hand panel shows the difference between L-PICOLA and MICE. Given the noise level of this data (corresponding to the larger blue confidence interval), the discrepancies will not lead to a significant parameter shift, though the discrepancy could have some impact for a key result from a cosmological survey.

with a series of additional layers, are often able to learn features with greater complexity than shallow architectures. For a general overview of deep learning and neural networks, see Goodfellow et al. (2016).

Recent works have demonstrated that the statistics extracted by a CNN are more powerful than both the two-point functions and peak counts, hinting that such models are accessing additional cosmological information present in the data.

This should not be surprising as CNNs have the capacity to be sensitive to both global scales and local features, and as such, should be able to be sensitive to both the power spectrum and peak count signals (Ribli et al. 2019; Cheng et al. 2020).

In contrast to most other works which adopt simple CNN architectures, in this work we opt for a state-of-the-art deep residual network model (He et al. 2015b). In recent years, *ResNets* have become an established standard architecture for image classification and regression tasks, significantly outperforming simpler CNNs (He et al. 2016).

Our ResNet model, which we will refer to as a *deep compressor*, accepts the noisy Kaiser–Squires map as an input, and is tasked with returning two numbers, which will constitute our summary statistics. The details of the model are presented in Section 7.

5 DATA AND FORWARD MODEL

5.1 DES SV data

DES is a ground-based photometric galaxy survey, which observed in the southern sky from the 4-m Blanco telescope in Chile with five photometric filters (Flaugher et al. 2015). The SV (A1) data[8] come from an initial run of 139 deg^2, but with depth approximately that of the full 6 yr survey (Chang et al. 2015). Data selection choices match Jeffrey et al. (2018).

We make a redshift cut of $0.6 < z_{\rm mean} < 1.2$, where $z_{\rm mean}$ is the mean of the z posterior for each galaxy. In our analysis, we use a single tomographic redshift bin for all selections, matching the shear peak analysis for this data set performed in Kacprzak et al. (2016). By using more bins, one could generate multiple maps that probed different redshifts through a range of peak lensing kernels, giving the possibility of more constraining power.

5.2 Dark matter simulations

We use 74 independent dark matter simulations, each with a different pair of cosmological parameters Ω_m and σ_8, and each covering an octant of the celestial sphere. All simulations used a standard flat ΛCDM cosmological model with Hubble parameter $H_0 = 70\,{\rm km\,Mpc^{-1}s^{-1}}$ and fixed values of the scalar spectral index and baryon density, $n_s = 0.95$ and $\Omega_b = 0.044$, respectively. Fixing these parameters can be interpreted as using a Dirac delta as the prior probability distribution for these parameters during inference.

The dark matter simulations are generated using the L-PICOLA code (Howlett et al. 2015), which is based on the COLA (Tassev, Zaldarriaga & Eisenstein 2013) algorithm. This uses a combination of second-order Lagrangian perturbation theory (2LPT) and a particle mesh (PM) which requires fewer time steps than 'full' N-body (e.g. Gadget Springel 2005) and which therefore can generate simulations more quickly.

We used a 1250 Mpc h^{-1} comoving simulation box, 768^3 particles, and a 1536^3 grid. A $z < 1.6$ light-cone was generated with up to four box replicates, using 30 time-steps from $z = 20$. The initial conditions used the linear matter power spectrum from Eisenstein & Hu (1999).

5.2.1 *Mock shear maps*

To generate a convergence map from a simulation, the matter particles were binned using the HEALPIX (Górski et al. 2005) pixelization of the sphere with NSIDE = 2048 in comoving radial shells of 50 Mpc h^{-1}. The particle density ρ map in a given redshift was converted into an overdensity $\delta = \rho/\bar{\rho} - 1$ using the average density in the shell $\bar{\rho}$. The convergence was calculated per pixel using equation (10), under the Born approximation (see Appendix B). We use the $n(z)$ in the lensing kernel matching the DES SV data, which we approximate by summing the individual posterior redshift distributions per galaxy from the BPZ photometric redshift code (Coe et al. 2006), matching the original analysis of this dataset (Abbott

[8] http://des.ncsa.illinois.edu

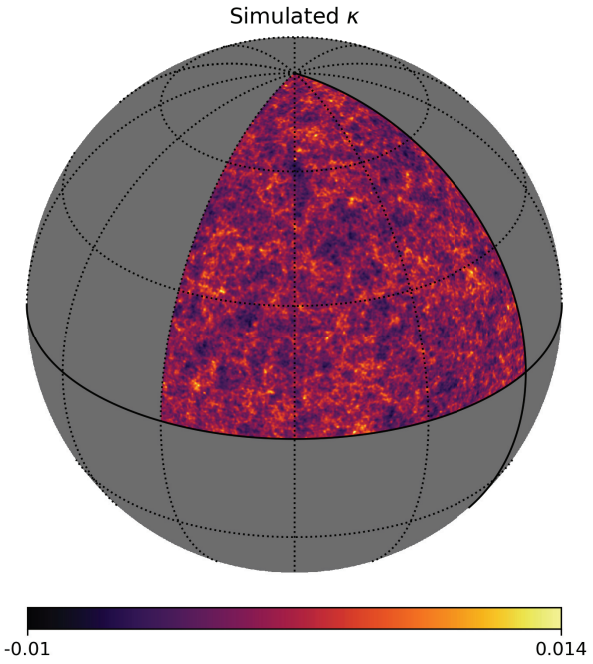

Figure 4. Example octant convergence κ field calculated by ray tracing an L-PICOLA dark matter simulation. From such a field we generate 18 non-overlapping DES SV mock data sets. The convergence field has been smoothed with a $\sigma = 10$ arcmin Gaussian kernel for visualization purposes.

et al. 2016). The convergence maps were downgraded to NSIDE = 1024. Fig. 4 shows an example of the resulting convergence map.

The drawback of the COLA approach is the accuracy of the dark matter distribution. The finite spatial resolution and fewer time-steps used by the COLA method particularly affect small distance scales. Fig. 2 shows a suppression of the L-PICOLA power spectrum at scales of $\ell > 700$ of order 10 per cent (relative to NICAEA[9], Kilbinger et al. 2009 theory), as is expected with COLA methods. We correct the power of the L-PICOLA convergence using the NICAEA predictions with halofit (Smith et al. 2003).

We estimate the smooth part of the $C_\kappa(\ell)$ from a convergence map realization using a polynomial order 1 Savgol filter with window size 91 for each convergence map and reweighting spherical harmonics. We rescale the spherical harmonic coefficients of the map by the ratio of the NICAEA theory and only the smooth part of the measured simulation power spectrum, which ensures that the natural fluctuations inherent in $C_\kappa(\ell)$ for a given realization are preserved.

The octant convergence fields are transformed to shear fields using spherical transformations as described in Section 3. The mask introduces small errors in the large-scale shear field, which are negligible for the much smaller data patches, especially once the non-smooth data mask and noise are included.

To generate mock DES SV shear maps, square patches of 256 × 256 pixels with 4.5 arcmin resolution are extracted from the spherical map with a gnonomic projection. The 139 deg^2 DES SV mask is then applied by excluding pixel data outside the observed area. From an octant we extract 18 non-overlapping DES SV mock data sets.

Noise realizations are generated from the data by randomly exchanging the ellipticity values between galaxies in the catalogue to remove the lensing signal and leave the shape noise remaining; this standard process assumes the variance due to lensing is negligible contribution to the ellipticity variance per galaxy. This method generates realistic, non-Gaussian shape noise in our mock catalogues.

Of the effects excluded from our forward modelling, the uncertainty in the redshift distribution and the intrinsic alignments of galaxies (Kirk et al. 2015) are likely to have the largest impact. In future work, these effects can be included in the mock data with a chosen data model, and any associated nuisance parameters included in the inference (for treatment of nuisance parameters while keeping low-dimensional summary statistics for likelihood-free analyses, see Alsing & Wandelt 2019).

5.2.2 *Peak count validation*

By construction the power spectra of the mock data match those predicted by theory (see Section 5.2.1). For the peak statistics, we validate our mock data against higher resolution and more accurate *N*-body predictions.

We use the public MICE lensing simulations (Fosalba et al. 2015) as a higher resolution 'truth', against which we can validate our forward-modelled data that used the approximate L-PICOLA *N*-body method. We generate 18 non-overlapping MICE maps, using the same source galaxy distribution, mask and shape noise generation method, to match the L-PICOLA mock data. From these mock maps, the peak count summary statistics are calculated identically in both cases.

All cosmological parameters in our forward-model pipeline, other than those to be inferred from the data, are fixed to the MICE values. We take four points in parameter space $[\Omega_m, \sigma_8]$ closest to the MICE values, such that the four points form the corners of a quadrilateral with the MICE parameter values inside. At each of these four parameter points, we take the average of the 18 mock peak count summary statistics; in Fig. 3 these are labelled 'Nearest L-PICOLA'.

From the 18 MICE realizations, we calculate the mean and the standard deviation for each element. The standard deviation corresponds to the larger blue shaded confidence interval in Fig. 3 labelled 'Data 'error bar'. This standard deviation corresponds to the square root of the diagonal of the data covariance (as used in a standard likelihood analysis), and so tells us whether a discrepancy between MICE and L-PICOLA is larger than the expected scatter in the observed data. The smaller orange confidence interval with plotted error bars are the standard errors on the mean (the larger confidence interval scaled by $18^{-\frac{1}{2}}$). In the left-hand panel of Fig. 3, there is agreement with the overall shape of the peak count distribution between MICE and L-PICOLA. The right-hand panel shows the difference between L-PICOLA and MICE, in which shows that there are some differences between all four L-PICOLA curves for certain κ bins, in particular the bins either side of the $n(\kappa)$ peak and the two highest κ bins.

To quantify the discrepancy and ascertain whether it is significant, we evaluate the reduced chi-squared statistic. As a conservative approximation, we take the mean of the four L-PICOLA curves as the prediction $\boldsymbol{\mu}$ and calculate the reduced chi-squared

$$\chi^2_\nu = \frac{1}{\nu} \sum_{ij} (d_i - \mu_i)^T \Sigma^{-1}_{ij} (d_j - \mu_j), \tag{20}$$

for the different elements of the data vector d_i with degrees of freedom $\nu = 11$ (as the unknown parameters are fixed) and covariance Σ estimated from the four L-PICOLA sets (a total of 72 realizations).

For χ^2_ν using the uncertainty due to the finite number of MICE simulations, we scale Σ by 18, resulting in $\chi_\nu \approx 3.4$. For the

[9] nicaea.readthedocs.io

realistic data covariance, we calculate $\chi_\nu \approx 0.2$. There is therefore a measurable difference between MICE and L-PICOLA with a poor fit of $\chi_\nu > 1$ with simulation errors alone. However, with realistic data noise, a $\chi_\nu \ll 1$ implies that this discrepancy is subdominant in comparison to other uncertainty contributions.

The discrepancy would not give a significant parameter shift, as it is within the noise level of the data. Some shift is nevertheless still there, so this discrepancy may be too large for key results from cosmological surveys. Additionally, for data sets more constraining than the DES SV data, this discrepancy would become more significant.

By using a simulation with only dark matter particles to validate our forward model, we are excluding tests for baryonic effects. These effects should appear at small scales, and we are likely to have suppressed them for DES data due to our 10 arcmin smoothing (Weiss et al. 2019). Further validation tests with high-resolution simulations with baryonic effects included would validate this.

6 RESULTS: POWER AND PEAK SUMMARY STATISTICS

6.1 Overview

In this section, we present the likelihood-free inference results using the power spectra and peak count summary statistics measured from the DES SV weak gravitational lensing mass maps. These are crafted summaries that we expect to be informative with respect to our unknown cosmological parameters $\boldsymbol{\theta}$. We leave the results from the deep compressor, extracted informative summaries from the lensing mass maps using deep CNNs, to the next section.

For the power and peak summary statistics, we first use a neural compressor to reduce our high-dimensional measured summaries $\boldsymbol{d}$ to low-dimensional compressed summaries $\boldsymbol{t}$. We then estimate the density $p(\boldsymbol{t}|\boldsymbol{\theta})$ using PYDELFI. We validate this density using an ensemble of NDEs.

The likelihood is then evaluated for the compressed summaries of the observed (compressed) data $\boldsymbol{t}_o$ and combined with the prior $p(\boldsymbol{\theta})$ to give our posterior probability for the parameters given our compressed power spectra and peak count summary statistics.

6.2 Data compression

Using PYDELFI, we aim to estimate the conditional distribution $p(\boldsymbol{t}|\boldsymbol{\theta})$, with compressed summary statistic $\boldsymbol{t}$. As discussed in Section 2.4, the compression of the summary statistics down to the same dimension as the parameters $\boldsymbol{\theta}$ (in this case two) is done to improve the density estimation with finite simulated summaries; the density estimation is done in $\{\boldsymbol{t}, \boldsymbol{\theta}\}$ space, rather than $\{\boldsymbol{d}, \boldsymbol{\theta}\}$.

In Alsing & Wandelt (2018), it was shown that for an underlying true likelihood, the optimal compression is lossless for a fiducial value of the parameter $\boldsymbol{\theta}$. For an underlying Gaussian likelihood, this *score compression* is linear (assuming the covariance is parameter-independent) and corresponds to the MOPED compression (Heavens, Jimenez & Lahav 2000) often used in astrophysical data compression. In general, the optimal score compression reproduces summaries that are transformations of, and have the same constraining power as, the maximum-likelihood estimate under the assumed likelihood parameter estimates, $\boldsymbol{t} = F(\boldsymbol{d}) = \boldsymbol{\theta}_{\rm MLE}$.

Rather than using an assumed likelihood, our main result uses neural compression, where $\boldsymbol{t} = F_\varphi(\boldsymbol{d})$ is approximated by a neural network. This is achieved by training over an augmented training set of $\{\boldsymbol{d}^+, \boldsymbol{\theta}^+\}$, aiming to minimize the following loss function with respect to the neural network parameters $\boldsymbol{\varphi}$:

$$J(\boldsymbol{\varphi}) = ||F_\varphi(\boldsymbol{d}^+) - [\boldsymbol{A}\boldsymbol{\theta}^+ + b]||_2^2, \tag{21}$$

which is averaged over simulated data $\boldsymbol{d}^+$ and parameter pairs $\boldsymbol{\theta}^+$ from the augmented training set, with fixed rescaling $\boldsymbol{A}$ and shift b to normalize for efficient training depending on network architecture. In this approach, unlike for the density estimation, the mock data $\boldsymbol{d}^+$ need not be independent, and we therefore generate 2500 data realization per simulated convergence octant to create the augmented training set.

In comparison to the score compression having an equivalence with maximum-likelihood estimation for an assumed likelihood, this choice of loss $J(\boldsymbol{\varphi})$ has an equivalence with a mean posterior estimate (Jaynes 2003), but without an assumed likelihood and with an implicit prior given by the distribution of the training labels $\boldsymbol{\theta}^+$.

An alternative commonly used loss is the L_1 norm, for example, $||F(Y) - X||_1$, corresponding to minimizing the least absolute deviation (LED) or mean absolute error (MAE), which would have an equivalence to a median posterior estimate. This was similarly tested, but resulted in more lossy compression when tested with simulated data (see Appendix A).

However, once the neural compression is trained for a given summary statistic using the augmented set $\{\boldsymbol{d}^+, \boldsymbol{\theta}^+\}$, the compressor is fixed and the density estimation of $p(\boldsymbol{t}|\boldsymbol{\theta})$ is performed with PYDELFI with the unsimulated observed data. A suboptimal data compression will lead to larger scatter in $\{\boldsymbol{t}, \boldsymbol{\theta}\}$ spaces, leading to inflated parameter constraints, but not to incorrect inference.[10]

The chosen network architecture and training scheme are described in Appendix A.

6.3 PYDELFI density estimation of $p(\boldsymbol{t}|\boldsymbol{\theta})$

To robustly estimate the distributions $p(\boldsymbol{t}|\boldsymbol{\theta})$, we used an ensemble of NDEs with PYDELFI.

We used two (full-rank) Gaussian MDNs and two MAFs. The two MDNs had two and three Gaussian components respectively, and both had two dense hidden layers with 30 neurons per layer. The two MAFs had four and five MADE layers respectively, each with two dense hidden layers with 50 neurons per layer. We used tan$h$ activation functions throughout. The final reconstructed density $p(\boldsymbol{t}|\boldsymbol{\theta})$ is then taken as a weighted average of the individual NDEs, weighted by the value of the loss achieved during training (i.e. model averaged relative to their individual performances).

The density estimation was carried out using the transformed parameters $\boldsymbol{\theta}' = [\Omega_m, S_8 = \sigma_8(\Omega_m/0.3)^{0.5}]$. This coordinate transformation simplifies the density estimation task, and the densities can then be transformed back to $\boldsymbol{\theta} = [\Omega_m, \sigma_8]$. We restrict the density estimation procedure to our eventual prior range of $0.1 < \Omega_m < 0.8$ and $0.45 < S_8 < 1.05$.

Beyond the usual training-validation split during training implemented in PYDELFI, combined with early-stopping to avoid overfitting, the individual estimates from the neural density ensemble can be used as a further (visual) validation step. Fig. 5 shows the marginal posterior probabilities (see the next section for details) for the parameters Ω_m and $S_8 = \sigma_8(\Omega_m/0.3)^{0.5}$ using the joint peak and power spectrum summary statistics. If the marginal distributions for each independent density estimation were in disagreement, this would be evidence that we had insufficient forward-modelled simulations

[10] As with all data analysis, poor data compression or cuts can lead to loss of information and less ability to infer the unknown parameters.

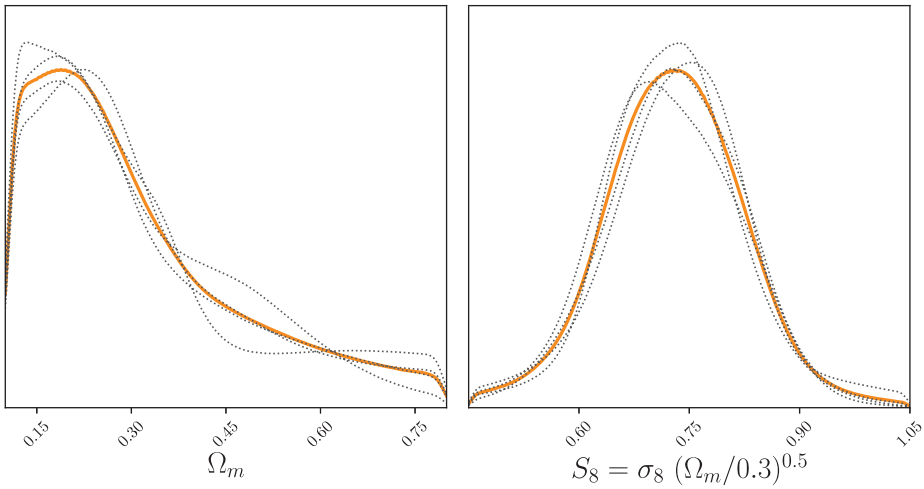

Figure 5. Validation of probability density estimation: marginal posterior distributions for the different NDEs in the PYDELFI ensemble with the (neural compressed) joint peak and power spectrum data. The orange line is the marginal posterior using the final stacked density. Though the individual density estimates are stacked with weights corresponding to their loss during training, there is nevertheless general agreement between the estimates. In future surveys, these distributions would give an estimate for where new simulations should be run in parameter space during *active learning*.

to constrain the density. The marginal distributions are generally in good agreement, albeit with some scatter.

If we were to run more simulations, we could use the scatter between the NDEs in the ensemble to make decisions about where to run new simulations, a process known as *active learning*. Such an acquisition procedure can reduce the number of simulations necessary for robust likelihood-free inference (Alsing et al. 2019) and will be an important tool for large-scale implementation in current and upcoming galaxy survey analysis.

6.4 Posterior probabilities

The chosen prior is uniform with respect to the physical parameters Ω_m and σ_8, but still bounded by $0.1 < \Omega_m < 0.8$ and $0.45 < \sigma_8(\Omega_m/0.3)^{0.5} < 1.05$. This prior range is the unshaded region in Fig. 6. To ensure this prior is used, evaluating in terms of Ω_m and $S_8 = \sigma_8(\Omega_m/0.3)^{0.5}$ and applying a prior $p(S_8) \propto (\Omega_m/0.3)^{-0.5}$ gives the same posterior as evaluation of the learned likelihood in terms of Ω_m and σ_8 directly.

As the posterior is in low dimension, evaluation on a grid (96×96) is much simpler than Markov Chain Monte Carlo sampling. The final smooth posterior distributions use CHAINCONSUMER (Hinton 2016) Kernel Density Estimation with the evaluated posterior grid points.

The left-hand panel of Fig. 7 shows the posterior probability for the two unknown parameters from the compressed power spectrum using the weak lensing map from DES SV data. The right-hand panel shows the posterior probability for the unknown parameters from the compressed peak count summary statistic using the DES SV weak lensing map. The central panel shows the parameter posterior probability distribution from the compressed joint power spectrum and peak count summary statistic.

The 2D posterior for our peak $n(\kappa)$ statistic (right-hand panel) from our observed data is centred with low Ω_m and high σ_8, and is therefore sharply cut by the lower limit of the prior $p(\Omega_m)$. The resulting marginal posterior distribution for Ω_m is ostensibly more sharply peaked, but this is due to the prior boundary rather than the data constraints.

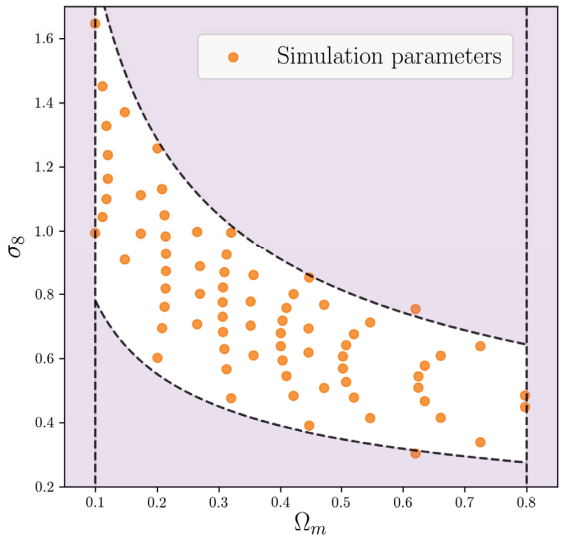

Figure 6. The 74 parameter pairs corresponding to the input cosmology of the forward-modelled data. The dashed lines shows the limits of the prior range (as described in Section 6) and the shaded region is therefore excluded by the prior during parameter inference.

The parameter constraints for the combined summary statistics (central panel) are modestly improved relative to the power spectrum alone. The change in marginal posterior with respect to the parameter combination $S_8 = \sigma_8\,(\Omega_m/0.3)^{0.5}$ is shown in Fig. 10.

We can compare the resulting marginal in Fig. 10 with the main result from the original shear two-point correlation DES analysis (Abbott et al. 2016), which gave a marginal $S_8 = 0.81 \pm 0.06$ and is, therefore, completely consistent. Fig. 2 of Abbott et al. (2016) shows the 2D posterior distribution, which appears in agreement to our power-spectrum result (left-hand panel, Fig. 7), but cannot be directly compared as the summary statistics, including data selection

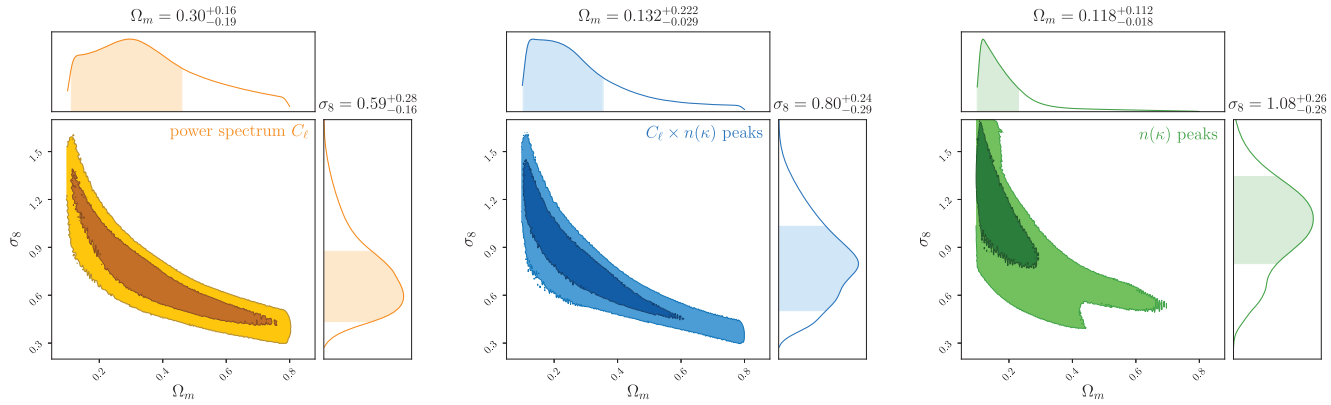

Figure 7. Posterior probability distributions from neural compressed summary statistics. The left-hand panel shows the posterior probability distribution from the power spectrum, the centre panel shows the joint peaks and power spectrum, and the right-hand panel shows the posterior from peaks alone. The resulting marginal posterior distribution for Ω_m from peaks alone is particularly peaked due to the prior lower bound for Ω_m, not due to the data constraints.

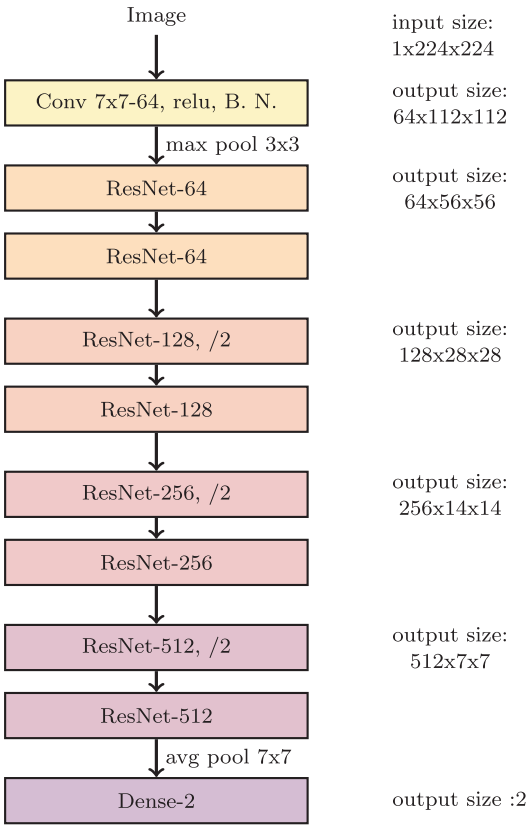

Figure 8. Architecture of the ResNet deep compressor. This network compresses the Kaiser–Squires weak lensing map into a 2D summary statistic, which is constructed to be informative with respect to our two unknown cosmological parameters. These informative summaries are evaluated for all the many simulated mock data maps and the single observed data maps to be used for the likelihood-free inference step.

effects (e.g. scale cuts) and modelling choices, are different. The joint posterior from original DES shear peak paper (Kacprzak et al. 2016) is even less directly comparable, as there are significant differences with error modelling, map making, and the definition of the peak count summary statistic, though the inferred parameters are not discrepant with our result or with Abbott et al. (2016).

7 RESULTS: MASS MAP DEEP COMPRESSOR

7.1 Overview

In this section, we present the likelihood-free inference results using compressed summary statistics $\boldsymbol{t}$ directly extracted from the DES SV weak lensing mass map using deep CNNs, the deep compressor. As described 4.4, CNNs are flexible feature extractors that can be optimized as to find maximally informative summary statistics from the mass map.

We first describe our CNN architecture, the ResNet-18 model. This acts as the function that takes the noisy Kaiser–Squires mass map as input and returns the compressed summaries. This network is included in the SSELFI implementation.

This implementation uses the MSE loss function to train the network (as described in Section 6.2) and also the Variational Mutual Information Maximization (VMIM) as an optimization objective (described below) to extract optimally informative summaries of the mass map.

As in the previous section, we use PYDELFI estimate the density $p(\boldsymbol{t}|\boldsymbol{\theta})$. The likelihood is then evaluated for the compressed summaries of the observed data $\boldsymbol{t}_o$ and combined with the prior probability distribution $p(\boldsymbol{\theta})$ to give our posterior probability for the parameters given deep compressor summary statistics extracted from the DES SV weak lensing mass map.

7.2 Convolutional neural network architecture

As discussed in Section 4.4, we base the deep compressor method on a standard deep ResNet architecture (He et al. 2015b). Specifically, we adopt a ResNet-18 model. The main component of this architecture is the *ResNet block*, in which a shortcut connection directly connects the input of the block to its output while on a parallel second branch the input is processed through several convolution layers (with associated non-linear activation functions) before being merged with the shortcut branch at the output of the block. This CNN architecture has proven extremely efficient, and enables the training of extremely deep models, with over 1000 layers.

The residual structure of ResNet alleviates one of the main limitations that prevents very deep neural networks training efficiently: vanishing gradients (Bengio, Simard & Frasconi 1994; Hochreiter & Schmidhuber 1997). Gradient backpropagation (see Goodfellow et al. 2016) is hindered by the convolutional and non-linearity layers,

to the point that in a standard CNN architecture deeper than about 10 layers, the upper layers of the model (close to the input) may not receive enough gradient signal to properly train. A residual network does not face this issue as gradients always have an almost unhindered path to reach any layer of the network model through this residual connection. Another aspect that helps explain the superiority of ResNet in practice is an easier initialization of the network parameters (He et al. 2015a). As each individual layer of a ResNet typically has to model small residual changes between input and outputs of the residual block, a random zero-mean initialization is already an appropriate choice.

The ResNet-18 model begins with an initial single convolution layer with a larger 7×7 pixel kernel with a rectified linear unit (ReLU) activation and batch normalization. The following convolutional layers have a 3×3 pixel kernel size; again, each with ReLU activation and batch normalization. The input Kaiser–Squires map is 224×224, with a border region removed as a simple solution to deal with edge effects. The full network is shown in Fig. 8.

Once the network is trained, the 2D output of the model acts as our CNN compressed summary statistic.

Our implementation is publicly available (see Appendix B) and is based on the official TensorFlow ResNet implementation[11] for Google's Tensor Processing Units (TPUs). Training such a ResNet on TPU allows us to reach a high throughput of over 5500 examples per second, bringing training time to just under 4 h for 110 000 training steps on the free Google Colab service.

7.3 Optimization objective

Training the compressor to yield informative statistics can be done in several ways. In the simplest approach, we could train the network to perform regression of the cosmological parameters under either an L_2 norm (MSE) or L_1 norm loss. As explained in Section 6.2, these correspond to training the model to estimate the mean and median of the posterior distribution respectively. This corresponds to the approach followed in a number of previous works (described in Section 4.4) of using CNNs for convergence maps analysis. This does not necessarily ensure the recovery of maximally informative summary statistics in general, though for fixed fiducial parameter values and assumptions of Gaussianity this may be true (Section 6.2).

Another approach to train the deep compressor model is to attempt to maximize the mutual information $I(\boldsymbol{t}, \boldsymbol{\theta})$ between the output summary statistics and cosmological parameters. The mutual information quantifies how much knowledge about $\boldsymbol{\theta}$ is obtained by an observation $\boldsymbol{t}$. The mutual information can be formally defined with respect to the Kullback–Leibler divergence (Kullback & Leibler 1951) as:

$$\begin{aligned} I(\boldsymbol{t}, \boldsymbol{\theta}) &= D_{\mathrm{KL}}(p(\boldsymbol{t}, \boldsymbol{\theta}) \parallel p(\boldsymbol{t})p(\boldsymbol{\theta})) \\ &= \int \mathrm{d}^n\boldsymbol{\theta}\, \mathrm{d}^n\boldsymbol{t}\; p(\boldsymbol{t}, \boldsymbol{\theta}) \log\left(\frac{p(\boldsymbol{t}, \boldsymbol{\theta})}{p(\boldsymbol{t})p(\boldsymbol{\theta})}\right) \\ &= \int \mathrm{d}^n\boldsymbol{\theta}\, \mathrm{d}^n\boldsymbol{t}\; p(\boldsymbol{t}, \boldsymbol{\theta}) \log\left(\frac{p(\boldsymbol{\theta}|\boldsymbol{t})}{p(\boldsymbol{\theta})}\right) \\ &= \int \mathrm{d}^n\boldsymbol{\theta}\, \mathrm{d}^n\boldsymbol{t}\; p(\boldsymbol{t}, \boldsymbol{\theta}) \log p(\boldsymbol{\theta}|\boldsymbol{t}) - \int \mathrm{d}^n\boldsymbol{\theta}\; p(\boldsymbol{\theta}) \log p(\boldsymbol{\theta}) \\ &= \mathbb{E}_{p(\mathbf{t},\boldsymbol{\theta})}\left[\log p(\boldsymbol{\theta}|\boldsymbol{t})\right] - \mathbb{E}_{p(\boldsymbol{\theta})}\left[\log p(\boldsymbol{\theta})\right] \\ &= \mathbb{E}_{p(\boldsymbol{t},\boldsymbol{\theta})}\left[\log p(\boldsymbol{\theta}|\boldsymbol{t})\right] - H(\boldsymbol{\theta}), \end{aligned} \tag{22}$$

where $p(\boldsymbol{t}, \boldsymbol{\theta})$ is the joint distribution of summary statistics and cosmological parameters, which is sampled by the simulations, and the expectation value is with respect to samples $\boldsymbol{\theta}$ and $\mathbf{t}$. On the right-hand side of the second expression, we recognize the entropy $H(\boldsymbol{\theta})$ of the distribution of cosmological parameters in the set of simulations.

In the context of data compression, a compressed $\boldsymbol{t}$ is obtained from a realization $\mathbf{d}$ of the high-dimensional signal (in this case the full lensing mass map). In this case we parametrize this mapping as $\boldsymbol{t} = F_{\boldsymbol{\varphi}}(\boldsymbol{d})$ using our ResNet-18 model. Our goal is therefore to find the parameters $\boldsymbol{\varphi}$ of the deep compressor that maximize the mutual information between summary statistics and cosmological parameters, given by

$$\boldsymbol{\varphi}^* = \arg\max_{\boldsymbol{\varphi}}\; I(F_{\boldsymbol{\varphi}}(\boldsymbol{d}), \boldsymbol{\theta}). \tag{23}$$

Unfortunately, the mutual information as expressed in equation (22) is not tractable, and estimation of this quantity remains an open problem in statistics and machine learning. However, the topic of mutual information estimation has attracted significant attention in the machine learning literature recently (e.g. Tishby & Zaslavsky 2015; Alemi et al. 2016; Chen et al. 2016), due in part to its potential for representation learning. These recent work have in common that they rely on tractable bounds on the mutual information, which allows for the mutual information to be optimized for instance in the context of training a deep neural network.

A variety of bounds exist with various properties, and we direct the interested reader to a recent review (Poole et al. 2019), but in this work, we adopt the Barber & Agakov (2003) variational lower bound. This is given by

$$I(\boldsymbol{t}, \boldsymbol{\theta}) \geq \mathbb{E}_{p(\boldsymbol{t},\boldsymbol{\theta})}\left[\log q(\boldsymbol{\theta}|\boldsymbol{t}; \boldsymbol{\varphi}')\right] - H(\boldsymbol{\theta}), \tag{24}$$

where $q(\boldsymbol{\theta}|\boldsymbol{t}; \boldsymbol{\varphi}')$ is a variational conditional distribution, which aims to model the posterior $p(\boldsymbol{\theta}|\boldsymbol{t})$. This lower bound becomes an equality when $q(\boldsymbol{\theta}|\boldsymbol{t}; \boldsymbol{\varphi}')$ matches the true posterior $p(\boldsymbol{\theta}|\boldsymbol{t})$. Using this bound, and taking advantage of the fact that $H(\boldsymbol{\theta})$ is constant for a given training set, we can restate the optimization problem as:

$$\arg\max_{\boldsymbol{\varphi},\boldsymbol{\varphi}'} \mathbb{E}_{p(\boldsymbol{d},\boldsymbol{\theta})}\left[\log q\left(\boldsymbol{\theta}|F_{\boldsymbol{\varphi}}(\boldsymbol{d}); \boldsymbol{\varphi}'\right)\right] \tag{25}$$

This procedure is known as VMIM, and the optimization problem can be solved by gradient descent over the weights of the neural network $F_{\boldsymbol{\varphi}}$, and parameters of the variational distribution $q(\boldsymbol{\varphi}')$.

In practice, to train the neural compressor under VMIM, we use a conditional Normalizing Flow to model $q(\boldsymbol{\theta}|\boldsymbol{t}; \boldsymbol{\varphi}')$. We adopt a four-stage MAF, each stage containing two hidden layers of size 128, and we train jointly the concatenation of the ResNet-18 and this Normalizing Flow model under the loss:

$$J_{\mathrm{VMIM}}(\boldsymbol{\varphi}, \boldsymbol{\varphi}') = -\sum_{n=1}^{N} \log q\left(\boldsymbol{\theta}_n^+|F_{\boldsymbol{\varphi}}(\boldsymbol{d}_n^+); \boldsymbol{\varphi}'\right), \tag{26}$$

where the sum is over the samples $\{\boldsymbol{\theta}_n^+,\ \mathbf{d}_n^+\}$ from the augmented training set.

After training, we discard the trained density estimator $q(\boldsymbol{\varphi}')$ and only export the neural compressor $F_{\boldsymbol{\varphi}}$. This is used to then compress the Kaiser–Squires maps to summary statistics which are used in the PYDELFI likelihood-free framework described in previous sections.

[11] https://github.com/tensorflow/tpu

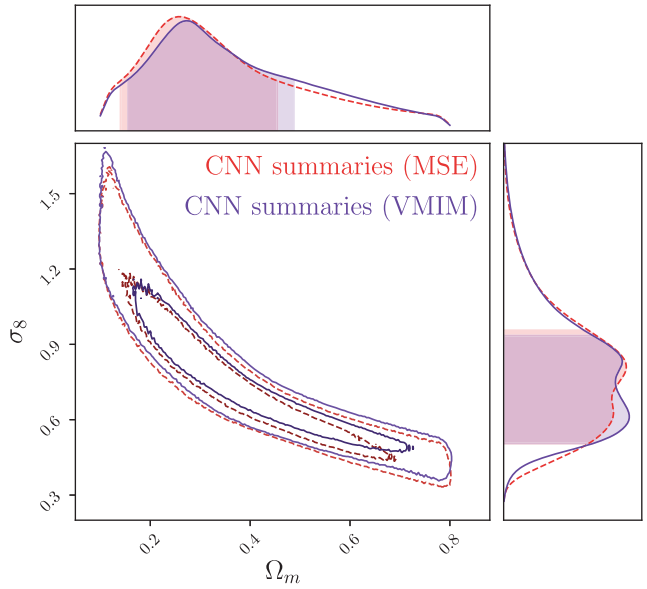

Figure 9. The posterior probability distributions from the two CNN map compressed statistics with the MSE (red dashed line) and VMIM (purple solid line) loss function.

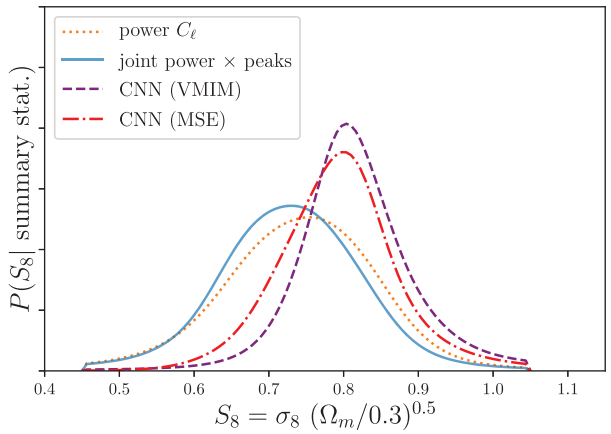

Figure 10. The marginal posterior probability densities for the parameter combination $S_8 = \sigma_8\,(\Omega_m/0.3)^{0.5}$ from the power spectrum compressed summary (dotted orange line), the joint power spectrum and peak compressed summary statistic (solid blue line), and from the two CNN map compressed summary statistics with MSE loss (dashed purple line) and VMIM loss (dashed–dotted red line).

7.4 Posterior constraints from deep CNN compression of the DES SV mass maps

Fig. 9 shows the posterior probability distributions from the DES SV weak lensing mass map using two CNN compressed map summary statistics. Each CNN had been trained using the same augmented training data but using different loss functions: the MSE and the VMIM.

As with the neural compression of the power spectrum and peak summary statistics, the network was trained and its weights fixed before application to data. Freezing the network architecture and weights before using the observed data removes the opportunity to take advantage of randomness in training, where one could, consciously or not, retrain in the hope of getting a different result due to scatter (a form of confirmation bias). This procedure of fixing the weights is a form of *blinding*.

The CNN map compressed summary statistics (both MSE and VMIM) have slightly tighter constraints in the marginal posterior shown in Fig. 10 than the other summary statistics. These deep compressor summaries result in a higher posterior mean (e.g. comparing the joint power × peaks with VMIM) for the $S_8 = \sigma_8\,(\Omega_m/0.3)^{0.5}$ parameter combination, but the posteriors are clearly not in tension in either the S_8 marginal (where the joint power×peak 2σ credible interval contains the value of the VMIM mean posterior) or in the full 2D posterior (compare Figs 7 and 9).

8 DISCUSSION AND CONCLUSIONS

In this work, we have used likelihood-free inference to estimate the posterior probability distributions of unknown cosmological parameters given observed DES SV weak lensing map statistics. This likelihood-free framework used a forward modelling approach to include the underlying physics and the combined effects of multiple sources of data and measurement noise.

The mock data summary statistics, labelled with their associated cosmological parameters, were compressed and compared with the observed data (also compressed) to estimate the likelihood $p(\boldsymbol{d_o}|\boldsymbol{\theta})$ for the given compressed summary statistic. With this simulation-based likelihood reconstruction for the DES SV data, the posterior distributions for the following summary statistics were evaluated: weak lensing map power spectrum (Fig. 7, left-hand panel), weak lensing map peak count summary statistic (Fig. 7, right-hand panel), the joint power×peak summary statistic (Fig. 7, centre panel), and a deep learning compressed summary statistic of the weak lensing map using CNNs (Fig. 9).

We use the PYDELFI (Alsing et al. 2019) package to perform the density estimation likelihood-free inference. To improve the efficiency of the density estimation, we aim to find some compression function taking the data summaries $\boldsymbol{d}$ and giving low-dimensional compressed summaries $\boldsymbol{t} = F(\boldsymbol{d})$ that retains the information about the unknown parameters. We aimed to learn such a compression function using deep neural networks (Section 6.2). Compressing the weak lensing map directly, rather than compressing summaries (e.g. power spectra), is an extension of this, which was implemented using deep CNNs (Section 7).

In this work, we have implemented a series of validation steps in our likelihood-free inference pipeline. Some of these tests aim to validate the forward model, which ensures the reliability of the physics and the data modelling. We also validate the density estimation step, which ensures the reliability of the resulting posterior distributions given the simulated data.

To validate the forward model we compare the measured summary statistics from our simulations with summary statistics measured from higher resolution simulations. In our case, we show that, up to the noise level of our data, the L-PICOLA simulations give power spectra and peak count summary statistics that are consistent with the MICE simulations. In this framework, we can deal with *known unknowns*. For example, we are aware that approximations in the L-PICOLA algorithm impact the small scales in the matter density, and therefore satisfied that the discrepancy in the power spectrum at high ℓ exists and is below the noise level.

There may be *unknown unknowns* in the forward model that are more difficult to validate. For example, one could construct a generator that created mock data with the power spectra matching theory perfectly, but still had a theoretically incorrect power spectrum

covariance. Here, we must rely on the physics encoded in the simulator to reflect the true behaviour of our system (the evolution of the cosmic matter density), and test for inaccuracies that could be reasonably expected (e.g. finite-resolution, *N*-body approximations, and baryonic effects). In all parameter or model inference, one should be able to describe the data model, and the same concerns about *unknown unknowns* should apply.

The clarity afforded by the ability to forward model mock data is a strength of likelihood-free inference. In the likelihood-free framework, the individual elements of the forward model are distinct and testable, so are less easily hidden.

In this work, our forward model includes or implicitly accounts for: non-Gaussian shape noise, the non-Gaussian density field, all mask and spatially-varying shape noise effects, resulting higher order moments of the sampling distribution, projection and flat-sky/Limber effects.

Effects that are often included as nuisance parameters, but which we have omitted, are the intrinsic alignments and the uncertainty in the photometric redshift distribution of the source galaxies. With the addition of extra nuisance parameters, we could have used different models to include the intrinsic alignments of galaxies. This is something that can be done in future work. Kacprzak et al. (2016) found that parameter constraints from peak counts can indeed be shifted by the effects of intrinsic alignments.

We could additionally include uncertainties in the photometric redshift distribution, either by simply marginalizing over nuisance parameters that re-calibrate or transform the distribution (e.g. Abbott et al. 2016) at the expense of additional parameters or, more generally, including the redshift analysis in the forward model.

To validate the density estimation step, we trained an ensemble of NDEs with different architectures. Fig. 5 shows marginal posterior distributions for the different NDEs (from the compressed joint power×peak data). The results from the different density estimators are in general agreement, though additional simulations with certain parameter combinations would decrease the variance of the ensemble. The model averaged stack of NDEs that is used for deriving the final posterior is more robust than any individual NDE in the ensemble.

By running additional simulations 'on the fly', one could update the density ensemble after each new batch of simulations to learn where in parameter space to best run the next batch. *Active learning* has been shown to be far more efficient in terms of the number of simulations needed (Leclercq 2018; Alsing et al. 2019). For likelihood-free inference using *N*-body simulations from data from current and upcoming surveys, with much larger volumes and a larger parameter space, this approach may be vital.

The likelihood-free approach can not only account for non-Gaussianity (which is not just a problem for the higher order statistics in weak lensing), but it also provides an efficient and alternative analysis pipeline that avoids many of the troublesome aspects of the standard cosmological inference pipelines. For example, with the same number of simulations that may be used to estimate the inverse covariance matrix for a Gaussian likelihood, we have estimated the full sampling distribution of the compressed summaries.

The likelihood-free framework improves the reliability of inference from standard summary statistics (e.g. power spectra) by providing a flexible and robust way forward by using direct comparison of simulations with data. It can also allow us, without resorting to misleading likelihood approximations, to use non-standard summary statistics of our data (including peaks and deep CNN map statistics) that can extract information beyond the standard summaries. As well as opening up non-standard summary statistics, likelihood-free inference for weak lensing surveys may also enable us to extract information from non-standard weak lensing observables (e.g. magnification, Hildebrandt, van Waerbeke & Erben 2009; van Waerbeke 2010; Duncan et al. 2013; Heavens, Alsing & Jaffe 2013; Hildebrandt et al. 2013; Alsing et al. 2015) that have been inhibited by complicated systematics effects that could (in principle) be included in forward simulations.

ACKNOWLEDGEMENTS

The authors thank Lorne Whiteway and Pablo Lemos for helpful comments on the manuscript, and thank Ofer Lahav and Benjamin Wandelt for useful discussions. NJ acknowledges funding from the École Normale Supérieure (ENS) and also acknowledges support from Science and Technology Facilities Council (STFC) grant ST/R000476/1.

Some of the results in this paper have been derived using the HEALPY package (Zonca et al. 2019). We also acknowledge use of MATPLOTIB (Hunter 2007), KERAS (Chollet et al. 2015), TENSORFLOW (Abadi et al. 2015), and CHAINCONSUMER (Hinton 2016).

DATA AVAILABILITY

The data underlying this article are publicly available from the DES Data Management system as part of the SVA1 Gold Release: https://des.ncsa.illinois.edu/releases/sva1

We have made the simulations used, along with associated code, publicly available: https://github.com/NiallJeffrey/Likelihood-free_DES_SV. Links to packages used in this paper are including in Appendix B.

REFERENCES

Abadi M. et al., 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Available at: https://www.tensorflow.org/ (accessed December 20)
Abbott T. et al., 2016, Phys. Rev. D, 94, 022001
Abbott T. M. C. et al., 2018, Phys. Rev. D, 98, 043526
Akeret J., Refregier A., Amara A., Seehars S., Hasner C., 2015, J. Cosmol. Astropart. Phys., 2015, 043
Alemi A. A., Fischer I., Dillon J. V., Murphy K., 2016, CoRR, preprint (arXiv:abs/1612.00410)
Alsing J., Wandelt B., 2018, MNRAS, 476, L60
Alsing J., Wandelt B., 2019, MNRAS, 488, 5093
Alsing J., Kirk D., Heavens A., Jaffe A. H., 2015, MNRAS, 452, 1202
Alsing J., Heavens A., Jaffe A. H., Kiessling A., Wandelt B., Hoffmann T., 2016, MNRAS, 455, 4452
Alsing J., Heavens A., Jaffe A. H., 2017, MNRAS, 466, 3272
Alsing J., Wandelt B., Feeney S., 2018, MNRAS, 477, 2874
Alsing J., Charnock T., Feeney S., Wand elt B., 2019, MNRAS, 488, 4440
Amendola L. et al., 2016, Living Rev. Relativ., 21, 2
Barber D., Agakov F., 2003, in Proc. 16th International Conference on Neural Information Processing Systems. NIPS'03. MIT Press, Cambridge, MA, USA, p. 201
Bartelmann M., Schneider P., 2001, Phys. Rep., 340, 291
Bengio Y., Simard P., Frasconi P., 1994, IEEE Trans. Neural Netw., 5, 157
Bishop C., 1994, Working Paper, Mixture Density Networks. Aston University, England
Cameron E., Pettitt A., 2012, MNRAS, 425, 44
Carassou S., de Lapparent V., Bertin E., Borgne D. L., 2017, A&A, 605, A9
Castro P. G., Heavens A. F., Kitching T. D., 2005, Phys. Rev. D, 72, 023516

Chang C. et al., 2015, Phys. Rev. Lett., 115, 051301
Charnock T., Lavaux G., Wandelt B. D., 2018, Phys. Rev. D, 97, 083004
Chen X., Duan Y., Houthooft R., Schulman J., Sutskever I., Abbeel P., 2016, CoRR, preprint (arXiv:1606.03657)
Cheng S., Ting Y.-S., Ménard B., Bruna J., 2020, MNRAS, 499, 5902
Chollet F. et al., 2015, Keras. Available at: https://keras.io (accessed December 20)
Coe D., Benítez N., Sánchez S. F., Jee M., Bouwens R., Ford H., 2006, AJ, 132, 926
Cooley J. W., Tukey J. W., 1965, Math. Comput., 19, 297
Dark Energy Survey Collaboration, 2016, MNRAS, 460, 1270
Davies F. B., Furlanetto S. R., 2016, MNRAS, 460, 1328
Diaz Rivero A., Dvorkin C., 2020, Phys. Rev. D, 102, 103507
Dietrich J. P., Hartlap J., 2010, MNRAS, 402, 1049
Dodelson S., 2003, Modern Cosmology. Academic Press, Amsterdam
Duncan C. A. J., Joachimi B., Heavens A. F., Heymans C., Hildebrandt H., 2013, MNRAS, 437, 2471
Eisenstein D. J., Hu W., 1999, ApJ, 511, 5
Fagioli M., Tortorelli L., Herbel J., Zürcher D., Refregier A., Amara A., 2020, J. Cosmol. Astropart. Phys., 2020, 050
Feeney S. M., Mortlock D. J., Dalmasso N., 2018, MNRAS, 476, 3861
Flaugher B. et al., 2015, AJ, 150, 150
Fluri J., Kacprzak T., Refregier A., Amara A., Lucchi A., Hofmann T., 2018, Phys. Rev. D, 98, 123518
Fluri J., Kacprzak T., Lucchi A., Refregier A., Amara A., Hofmann T., Schneider A., 2019, Phys. Rev. D, 100, 063514
Fosalba P., Gaztañaga E., Castander F. J., Crocce M., 2015, MNRAS, 447, 1319
Germain M., Gregor K., Murray I., Larochelle H., 2015, in Francis B., Blei D., eds, International Conference on Machine Learning. PMLR, Lille, France, p. 881
Goodfellow I. J., Bengio Y., Courville A., 2016, Deep Learning. MIT Press, Cambridge, MA, USA
Górski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759
Grathwohl W., Chen R. T., Bettencourt J., Sutskever I., Duvenaud D., 2019, Proc. International Conference on Learning Representations
Hahn C., Vakili M., Walsh K., Hearin A. P., Hogg D. W., Campbell D., 2017, MNRAS, 469, 2791
He K., Zhang X., Ren S., Sun J., 2015a, in Proc. IEEE International Conference on Computer Vision. p. 1026
He K., Zhang X., Ren S., Sun J., 2015b, CoRR, preprint (arXiv:1512.03385)
He K., Zhang X., Ren S., Sun J., 2016, in European Conference on Computer Vision. p. 630
Heavens A. F., Jimenez R., Lahav O., 2000, MNRAS, 317, 965
Heavens A., Alsing J., Jaffe A. H., 2013, MNRAS, 433, L6
Hildebrandt H., van Waerbeke L., Erben T., 2009, A&A, 507, 683
Hildebrandt H. et al., 2013, MNRAS, 429, 3230
Hinton S. R., 2016, J. Open Source Softw., 1, 00045
Hivon E., Górski K. M., Netterfield C. B., Crill B. P., Prunet S., Hansen F., 2002, ApJ, 567, 2
Hochreiter S., Schmidhuber J., 1997, Neural Comput., 9, 1735
Howlett C., Manera M., Percival W. J., 2015, A&C, 12, 109
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Ishida E. et al., 2015, Astron. Comput., 13, 1
Jaynes E. T., 2003, Probability Theory: The Logic of Science. Cambridge Univ. Press, Cambridge
Jeffrey N. et al., 2018, MNRAS, 479, 2871
Jeffrey N., Lanusse F., Lahav O., Starck J.-L., 2020, MNRAS, 492, 5023
Jennings E., Wolf R., Sako M., 2016, preprint (arXiv:1611.03087)
Jimenez Rezende D., Mohamed S., 2015, preprint (arXiv:1505.05770)
Joudaki S. et al., 2018, MNRAS, 474, 4894
Kacprzak T. et al., 2016, MNRAS, 463, 3653
Kacprzak T., Herbel J., Amara A., Réfrégier A., 2018, J. Cosmol. Astropart. Phys., 2018, 042
Kaiser N., Squires G., 1993, ApJ, 404, 441
Kilbinger M., 2015, Rep. Prog. Phys., 78, 086901
Kilbinger M. et al., 2009, A&A, 497, 677
Kilbinger M. et al., 2013, MNRAS, 430, 2200
Kingma D. P., Ba J., 2014, preprint (arXiv:1412.6980)
Kingma D. P., Salimans T., Jozefowicz R., Chen X., Sutskever I., Welling M., 2016, in Advances in Neural Information Processing Systems. p. 4743
Kirk D. et al., 2015, Space Sci. Rev., 193, 139
Kullback S., Leibler R. A., 1951, Ann. Math. Stat., 22, 79
Lanusse F., Starck J.-L., Leonard A., Pires S., 2016, A&A, 591, A2
Leclercq F., 2018, Phys. Rev. D, 98, 063511
LeCun Y., Boser B. E., Denker J. S., Henderson D., Howard R. E., Hubbard W. E., Jackel L. D., 1990, in Advances in Neural Information Processing Systems. p. 396
Lin C.-A., Kilbinger M., 2015, A&A, 583, A70
LSST Science Collaboration, 2009, preprint (arXiv:0912.0201)
Mallat S., 2016, Philos. Trans. R. Soc. Lond. Ser. A, 374, 20150203
Marin J.-M., Pudlo P., Robert C. P., Ryder R., 2011, preprint (arXiv:1101.0955)
Marshall P. J., Hobson M. P., Gull S. F., Bridle S. L., 2002, MNRAS, 335, 1037
Martinet N. et al., 2018, MNRAS, 474, 712
Newman E. T., Penrose R., 1966, J. Math. Phys., 7, 863
Papamakarios G., Murray I., 2016, Advances in Neural Information Processing Systems. p. 1028
Papamakarios G., Pavlakou T., Murray I., 2017, Advances in Neural Information Processing Systems. p. 2338
Papamakarios G., Nalisnick E., Rezende D. J., Mohamed S., Lakshminarayanan B., 2019a, preprint (arXiv:1912.02762)
Papamakarios G., Sterratt D., Murray I., 2019b, in Chaudhuri K., Sugiyama M., eds, Proceedings of Machine Learning Research Vol. 89, Proceedings of Machine Learning Research. PMLR, p. 837
Peel A., Lin C.-A., Lanusse F., Leonard A., Starck J.-L., Kilbinger M., 2017, A&A, 599, A79
Planck Collaboration VI, 2020, A&A, 641, A6
Poole B., Ozair S., van den Oord A., Alemi A. A., Tucker G., 2019, CoRR, preprint (arXiv:1905.06922)
Price M. A., Cai X., McEwen J. D., Pereyra M., Kitching T. D., 2019, MNRAS, 492, 394
Ribli D., Pataki B. Á., Zorrilla Matilla J. M., Hsu D., Haiman Z., Csabai I., 2019, MNRAS, 490, 1843
Robin A., Reylé C., Fliri J., Czekaj M., Robert C., Martins A., 2014, A&A, 569, A13
Rubin D. B., 1984, The Annals of Statistics. Instit. Mathematical Statistics, USA, p. 1151
Schafer C. M., Freeman P. E., 2012, in Statistical Challenges in Modern Astronomy V. Springer, New York, p. 3
Sellentin E., Heavens A. F., 2018, MNRAS, 473, 2355
Sellentin E., Heymans C., Harnois-Déraps J., 2018, MNRAS, 477, 4879
Shan H. et al., 2018, MNRAS, 474, 1116
Shirasaki M., Yoshida N., Ikeda S., 2019, Phys. Rev. D, 100, 043527
Smith R. E. et al., 2003, MNRAS, 341, 1311
Springel V., 2005, MNRAS, 364, 1105
Tassev S., Zaldarriaga M., Eisenstein D. J., 2013, J. Cosmol. Astropart. Phys., 2013, 036
Taylor P. L., Kitching T. D., Alsing J., Wandelt B. D., Feeney S. M., McEwen J. D., 2019, Phys. Rev. D, 100, 023519
Tegmark M., 1997, Phys. Rev. D, 55, 5895
Tishby N., Zaslavsky N., 2015, CoRR, preprint (arXiv:1503.02406)
Tortorelli L., Fagioli M., Herbel J., Amara A., Kacprzak T., Refregier A., 2020, J. Cosmol. Astropart. Phys., 2020, 048
van Waerbeke L., 2010, MNRAS, 401, 2093
Wallis C. G. R., McEwen J. D., Kitching T. D., Leistedt B., Plouviez A., 2017, preprint (arXiv:1703.09233)
Weiss A. J., Schneider A., Sgier R., Kacprzak T., Amara A., Refregier A., 2019, J. Cosmol. Astropart. Phys., 2019, 011
Weyant A., Schafer C., Wood-Vasey W. M., 2013, ApJ, 764, 116
Zonca A., Singer L., Lenz D., Reinecke M., Rosset C., Hivon E., Gorski K., 2019, J. Open Source Softw., 4, 1298

APPENDIX A: SUMMARY STATISTIC NEURAL COMPRESSION TRAINING

For the neural compression of the power and peak summary statistics, we tested a series of architectures with varying activation functions, loss functions, and learning rates. The final (best performing) architecture used three dense hidden layers with 100 nodes, each followed by a ReLU activation function, and a final dense layer (to a two-element output) without an activation function. The network was trained using the stochastic optimizer *Adam* (Kingma & Ba 2014) with 130 000 training samples and 55 000 validation samples (generated using the data augmentation method described in Section 6.2). The network, which showed no evidence of significant overfitting, was trained for 20 epochs.

For the loss function we considered two options: the L_1 loss and the L_2 MSE loss (e.g. equation 21). As discussed in Section 6.2, the L_2 minimization corresponds to a point estimate of the posterior mean and the L_1 minimization corresponds to a point estimate of the posterior mode. From the augmented training data, a small sample of four realizations were taken as mock observations and each compressed in two different ways (with L_1 and L_2 trained networks). The compressed summaries were used in our PYDELFI pipeline with the result that the L_2 MSE loss consistently gave slightly tighter constraints (with the four mock observations) than the L_1 loss. Any differences between the two choices of loss function were nevertheless extremely small.

All of these tests were performed before the final likelihood-free inference step using data, to avoid misleading results due to post-hoc analysis (analogous to 'p-hacking').

APPENDIX B: PUBLIC CODE

In this work, we have used code to generate mock simulations, compress the observed summary statistics of the data, and estimate the likelihood-free posterior probability densities.

(i) To generate the mock simulations, we use the L-PICOLA dark matter simulation code (Howlett et al. 2015): https://cullanhowlett.github.io/l-picola/.

(ii) To convert the dark matter overdensity shells into convergence maps, we use a Born approximation ray tracing code, which we have made available: https://github.com/NiallJeffrey/born_raytrace.

(iii) For density estimation likelihood-free inference, we use the PYDELFI code (Alsing et al. 2019): https://github.com/justinalsing/pydelfi.

(iv) For the deep compressor CNN summary statistic extraction, we use SSELFI: https://github.com/EiffL/SSELFI.

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