Nested sampling with any prior you like

Justin Alsing\textsuperscript{1,2,*} and Will Handley\textsuperscript{3,4}
\textsuperscript{1}Oskar Klein Centre for Cosmoparticle Physics, Department of Physics, Stockholm University, Stockholm SE-106 91, Sweden
\textsuperscript{2}Imperial Centre for Inference and Cosmology, Department of Physics, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK
\textsuperscript{3}Astrophysics Group, Cavendish Laboratory, JJ Thomson Avenue, Cambridge CB3 0HA, UK
\textsuperscript{4}Kavli Institute for Cosmology, Madingley Road, Cambridge CB3 0HA, UK

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

Nested sampling is an important tool for conducting Bayesian analysis in Astronomy and other fields, both for sampling complicated posterior distributions for parameter inference, and for computing marginal likelihoods for model comparison. One technical obstacle to using nested sampling in practice is the requirement (for most common implementations) that prior distributions be provided in the form of transformations from the unit hyper-cube to the target prior density. For many applications – particularly when using the posterior from one experiment as the prior for another – such a transformation is not readily available.

In this letter we show that parametric bijectors trained on samples from a desired prior density provide a general-purpose method for constructing transformations from the uniform base density to a target prior, enabling the practical use of nested sampling under arbitrary priors. We demonstrate the use of trained bijectors in conjunction with nested sampling on a number of examples from cosmology.

Key words: data analysis: methods

1 INTRODUCTION

Nested sampling has become one of the pillars of Bayesian computation in Astronomy (and other fields), enabling efficient sampling of complicated and high-dimensional posterior distributions for parameter inference, and estimation of marginal likelihoods for model comparison.

One technical obstacle to using nested sampling in practice is the requirement (for most common implementations) that prior distributions be provided in the form of transformations from the unit hyper-cube to the target prior density. For many applications – particularly when using the posterior from one experiment as the prior for another – such a transformation is not readily available.

In this letter we show that parametric bijectors trained on samples from a desired prior density provide a general-purpose method for constructing transformations from the uniform base density to a target prior, enabling the practical use of nested sampling under arbitrary priors. We demonstrate the use of trained bijectors in conjunction with nested sampling on a number of examples from cosmology.

Key words: data analysis: methods

2 NESTED SAMPLING AND PRIOR SPECIFICATION

Given a likelihood function $L(\theta)$ and prior distribution $\pi(\theta)$, nested sampling was incepted by Skilling (2006) as a Monte Carlo tool to...
compute the Bayesian evidence
\[ Z = \int \mathcal{L}(\theta) \pi(\theta) d\theta, \]  
(1)
whilst simultaneously generating samples drawn from the normalised posterior distribution \( \mathcal{P}(\theta) = \mathcal{L}(\theta) \pi(\theta)/Z. \)

Nested sampling performs these dual processes by evolving an ensemble of \( n_{\text{live}} \) live points by initially drawing them from the prior distribution \( \pi \), and then at each subsequent iteration: (a) discarding the lowest likelihood live point and (b) replacing it by a live point drawn from the prior \( \pi \) subject to hard constraint that the new point is at a higher likelihood than the discarded point. The set of discarded dead points and likelihoods can be used to compute a statistical estimate of the Bayesian evidence (1), and be weighted to form a set of samples from the posterior distribution.

The nested sampling procedure therefore differs substantially from traditional strategies such as Gibbs sampling, Metropolis-Hastings or Hamiltonian Monte Carlo in that it is acts as a likelihood scanner rather than a posterior sampler. It scans athermally, so in fact is capable of producing samples at any temperature and computing the full partition function in addition to the Bayesian evidence (Handley et al. 2019b).

Implementations of the nested sampling meta-algorithm e.g. MultiNest (Feroz et al. 2009) and PolyChord (Handley et al. 2015a,b) differ primarily in their choice of method by which to draw a new live point from the prior subject to a hard likelihood constraint. Nested sampling algorithms typically assume that each parameter has a prior that is uniform over the prior range \([0, 1]\) i.e. the unit hypercube. As we shall see in the next section, this is not as restrictive as it sounds, since any prior model may be transformed onto the unit hypercube via a bijective transformation.

It should be noted that normalising flows and bijectors have been used by Moss (2020) & Williams et al. (2021) in the context of techniques for generating new live points, but the approach in this letter for using them for prior specification is applicable to all existing nested sampling algorithms.

3 REPRESENTING PRIORS AS BIJECTORS

A bijector \( f : x \rightarrow y \) is an invertible function between the elements of two sets, with a one-to-one correspondence between elements, i.e., each element of one set is paired with precisely one element of the other set and there are no unpaired elements. In the context of probability theory, bijections are convenient for defining mappings from one probability distribution to another. If a random variate \( x \) with probability density \( p(x) \) is transformed under a bijector \( f : x \rightarrow y = f(x) \), then samples and densities transform as:
\[ y = f(x), \quad x = f^{-1}(y) \]
\[ p(y) = |J| p(x), \quad p(x) = |J^{-1}| p(y) \]

where \( J_{ij} = \frac{\partial f_j}{\partial x_i} \) is the Jacobian corresponding to the bijector function \( f \). Any probability density can be represented as a bijection from a base density, or equivalently, any target density can be completely specified by a base density and bijector from that base density to the target. Note also that any composition of bijections is also a bijector.

In the context of nested sampling, priors \( \pi(\theta) \) over model parameters \( \theta \) should be provided as a bijector that transforms from a base (unit) uniform density \( u \sim \mathcal{U}(0, 1) \) to the desired prior \( \pi(\theta) \), i.e.,
\[ \theta = f(u), \]  
(3)
where
\[ \pi(\theta) = |J| \mathcal{U}(f^{-1}(\theta)) \]

In practice, all that is required for specifying nested sampling priors is the bijector function \( f(u) \).

In many use cases, such a bijector is readily available. For example, for any univariate prior density, the bijector from the unit uniform is simply the inverse cumulative density function (CDF) of the target prior,
\[ f(u) = \text{CDF}^{-1}(u), \]
\[ \text{CDF}(u) = \int_0^u \pi(\theta) d\theta, \]  
(5)
where the CDF can be interpolated and inverted numerically if the integral is intractable.

However, for the general case of correlated multivariate priors, the requisite bijector is far more involved to compute analytically (see 5.1 of Feroz et al. 2009) and in general numerically requires obtaining by other means. One particularly common scenario where this arises is when one wants to use the (sampled) posterior from one experiment as the prior for another; in these cases, all that is available are samples from the desired prior and un-normalized values of its density at those points.

The solution is to fit a parametric model for the required bijector \( f(u, w) \) (with parameters \( w \)) to the target prior. In the following sections we describe how to define expressive parametric bijector models, and fit them to target (prior) densities.

3.1 Parametric bijectors

Parametric bijector models \( f(u, w) \) have recently been gaining popularity for solving complex density estimation tasks, probabilistic learning and variational inference, with a plethora of models available. In this section we give a brief overview of the key methods (in order of increasing complexity). For a more in-depth review, see Papamakarios et al. (2019).

Compositions of simple invertible functions

Since any composition of bijections is also a bijector, it is possible to construct flexible parametric bijectors by composing even simple invertible functions, such as affine transformations, exponentials, sigmoids, hyperbolic functions, etc.

Some such transformations applied to a base density lead to already well-studied distribution families, such as the “sinh-arsinh” distributions (Jones & Pewsey 2009) which are generated by applying a \( a \sinh(b \sinh^{-1}(x) + c) \) bijector (with some parameters \( a, b \) and \( c \)) to a base normal random variate. Chaining a number of such transformations together, interleaved with linear transformations of the parameter space, can already lead to rather expressive families of distributions.

Autoregressive flows

More sophisticated bijector models typically involve constructing normalizing flows parameterized by neural networks. Of these neural network based models, Inverse Autoregressive Flows (IAFs: Kingma et al. 2016) have been demonstrated to be effective at representing complex densities, are fast to train, and fast to evaluate once trained.
Figure 1. Demonstration of consistency between sampling using traditional and bijector-based priors. All plots show 66.7% and 95% credibility regions of marginal probability distributions projected into the $\Omega_K$-$H_0$ plane, and colours are chosen to be consistent across the grid. The first row indicates the prior placed on the $\Omega_K$-$H_0$ parameter space by CAMB ruling certain portions of the parameter space to be unphysical (e.g. for Universes with too much negative curvature collapsing, or unphysical lensing amplitudes), and it should be noted that it is banana-like rather than flat. The second row indicates the prior from the first row, and the posterior distributions derived by using CMB data alone, CMB+BAO data and BAO data alone respectively for each column, with appropriately coloured arrows indicating the application of data to the prior to produce a corresponding posterior distribution. The final row demonstrates the comparison between approaches. In the first and final columns, we see the resultant posterior derived by using the corresponding posterior from the second row as a prior for the cosmology run. The central bottom panel compares the three routes to a combined data posterior and demonstrates them to be in general agreement to within sampling error. Dashed lines indicate the corresponding relationship between limits of the plots. Plot produced under anesthetic (Handley 2019b).
IAFs define a series of $n$ bijectors composed into a normalizing flow, as follows:

$$
\begin{align*}
    z^{(0)} &= n \sim N(0, 1) \\
    z^{(1)} &= \mu^{(1)}[z^{(0)}] + \sigma^{(1)}[z^{(0)}] \odot z^{(0)} \\
    &\vdots \\
    y &= z^{(n)} = \mu^{(n)}[z^{(n-1)}] + \sigma^{(n)}[z^{(n-1)}] \odot z^{(n-1)}
\end{align*}
$$

where the shifts $\mu$ and scales $\sigma$ of the subsequent affine transforms are autoregressive and are parameterized by neural networks, i.e.,

$$
\begin{align*}
    \mu^{(k)}_i &= \mu^{(k)}_i[z_{1:k-1}; w], \\
    \sigma^{(k)}_i &= \sigma^{(k)}_i[z_{1:k-1}; w],
\end{align*}
$$

where $w$ denote the weights and biases of the neural network(s). For more details see Papamakarios et al. (2017).

### Continuous normalizing flows

In order to take a further step forward in expressivity, one can replace the discrete set of transforms with an integral of continuous-time dynamics (Chen et al. 2018; Grathwohl et al. 2018), i.e., defining the bijector as the solution to an initial value problem:

$$
\begin{align*}
    z(0) &\sim N(0, 1) \\
    y &= z(t) = \int_0^t f(z, t'; w) dt',
\end{align*}
$$

where the dynamics $f(z, t; w) = dz/dt$ are parameterized by a neural network with weights $w$. The elevation from discrete to continuous normalizing flows comes with a significant increase in expressivity, making such models appropriate for the most complex bijector representation tasks. See Chen et al. (2018) and Grathwohl et al. (2018) for more details.

### 3.2 Fitting bijectors to target densities

Fitting a bijector model to the target can be achieved by minimizing the Kullback-Leibler (KL) divergence between the target $p(\theta)$ and the (model) bijective density (with respect to the model parameters $w$). Since the KL divergence will not be analytically tractable in general, one can instead minimize a Monte Carlo estimate of the KL divergence in certain cases (Seljak & Yu 2019) (and can be readily extended to exploit gradient information about the target density if it is available; see Seljak & Yu 2019 for details).

![Figure 2. In the same three paths of Figure 1 the Bayesian evidences $\log Z$ for each of the three runs combine approximately consistently to within error, demonstrating that the bijective prior methodology can be used to recover reasonably accurate evidences in addition to posteriors. It should be noted that in contrast with Figure 1, since the error bars on this plot represent the nested sampling errors in inferring the evidence (in comparison with the posterior probability density spread), one should only expect these plots to overlap to within the histogram width, in analogue with repeated independent measurements of a fixed quantity. We note that typically error estimates on the log-evidence are typically underestimated due to implementation-specific errors Higson et al. (2019) Nonetheless, the inferred evidences are statistically consistent to within $3 \sigma$, indicating that there may be a small amount of unquantified error from the use of bijectors, but that the evidences are sufficiently accurate for model comparison purposes (where $\Delta \log Z \sim O(1)$). Plot produced under anesthetic (Handley 2019b).](image)

### 4 COSMOLOGICAL POSTERIORS AS BIJECTORS

As a concrete and non-trivial cosmological example we choose the concordance $\Lambda$CDM model of the universe (Scott 2018) with an additional spatial curvature parameter $\Omega_k$. Such models have been debated recently in the literature (Handley 2019a; Di Valentino et al. 2019; Efstathiou & Gratton 2020). We choose this model not for its controversy, but merely for the fact that it yields non-trivial banana-like priors and posteriors and therefore proves a more challenging cosmological example for a bijector to learn than models without curvature.

For likelihoods we choose three datasets: (1) The Planck baseline TT,TE,EE+lowE likelihood (without lensing) Planck Collabo-
ration (2020a,b) (hereafter CMB), (2) Baryon Acoustic Oscillation and Redshift Space Distortion measurements from the Baryon Oscillation Spectroscopic Survey (BOSS) DR12 Alam et al. (2017); Beutler et al. (2011); Ross et al. (2015) (hereafter BAO), and (3) local cosmic distance ladder measurements of the expansion rate, using type Ia SNe calibrated by variable Cepheid stars from the SH0ES collaboration SH0ES Collaboration (2018) (hereafter SH0ES).

The bijector code is implemented as an extension to CosmoChord using forpy to call tensorflow bijectors from FORTRAN.

Figure 1 shows our key results for CMB and BAO data. In this case we compare the three routes to a combined CMB and BAO constraint on a $\Lambda$CDM universe. The simplest approach is to run nested sampling with the default CosmoMC prior with both likelihoods turned on $\pi \rightarrow$CMB+BAO. The bijector approach first runs the CMB dataset to update the prior $\pi \rightarrow$CMB, then after using the posterior samples from this run to train a CMB bijector constructs a distribution to use as the prior for the input for the Bayesian update with the next dataset CMB$\rightarrow$BAO. In Figure 1 we show that doing this in either order recovers the same posterior.

One critical advantage of this approach is that since much of the cost of a nested sampling run is in compressing from prior to posterior, whilst the first run $\pi \rightarrow$CMB is expensive, the second update CMB$\rightarrow$BAO requires considerably fewer likelihood calls. Furthermore, since the journey from prior to posterior is shorter, less poisson error accumulates over nested sampling iterations, making for more accurate posteriors and evidence estimates. For expensive models like those involving cosmic curvature, this is a significant advantage.

It should be noted that as CMB and BAO are in mild tension, it has been argued that these combined constraints should be viewed with some scepticism (Handley 2019a; Vagnozzi et al. 2020), and this tension is much stronger between CMB and SH0ES data (Verde et al. 2019). Whilst this is a general cause for simultaneous concern and excitement scientifically speaking, in the context of this work this also serves to highlight the expressivity of these bijector models when combined with nested sampling chains. The tail behaviour of these bijective transformations is sufficiently powerful to still recover the correct answer even when combining likelihoods which are in strong tension. Figure 2 demonstrates that the evidence estimates extracted by this run from anesthetic (Handley 2019b) are also consistent within the errors on nested sampling.

5 CONCLUSIONS

Nested sampling implementations have been burdened by a practical limitation that priors need to be specified as bijective transforms from the unit hyper-cube, and in many cases such a representation is not readily available. We have shown that nested sampling priors can be constructed by fitting flexible parametric bijectors to samples from the target prior, providing a general-purpose approach for constructing generic priors for nested sampling. This is of particular importance for use cases where an interim posterior from one experiment is to be used as the prior for another, where a parametric bijector can then simply be trained on samples from the interim posterior. We have demonstrated the use of parametric bijectors on some typical use-cases from cosmology.

In the longer term, we plan to release a library of cosmological bijectors for use as nested sampling priors in order to save users computational resources and standardise cosmological prior choice.

ACKNOWLEDGEMENTS

We thank Johannes Buchner for useful comments. Justin Alsing was supported by research project grant Fundamental Physics from Cosmological Surveys funded by the Swedish Research Council (VR) under Dnr 2017-04212. Will Handley was supported by a Gonville & Caius Research Fellowship, STFC grant number ST/T001054/1 and a Royal Society University Research Fellowship.

DATA AVAILABILITY

All the nested sampling runs, code and details for reproducing the results in this letter can be found on Zenodo (Handley & Alsing 2020).

REFERENCES

Alam S., et al. 2017, MNRAS, 470, 2617
Beutler F., et al., 2011, MNRAS, 416, 3017
Brewer B. J., Foreman-Mackey D., 2016, arXiv preprint arXiv:1606.03757
Chen R. T., Rubanova Y., Bettencourt J., Duvenaud D. K., 2018, in Advances in neural information processing systems. pp 6571–6583
Di Valentino E., Melchiorri A., Silk J. 2019, Nature Astronomy, 4, 196–203
Efstathiou G., Gratton S., 2020, Monthly Notices of the Royal Astronomical Society: Letters, 496, L91–L95
Feroz F., Hobson M. P., Bridges M., 2009, Monthly Notices of the Royal Astronomical Society, 398, 1601–1614
Grathwohl W., Chen R. T., Bettencourt J., Sutskever I., Duvenaud D., 2018, arXiv preprint arXiv:1810.01367
Handley W., 2019a, Curvature tension: evidence for a closed universe (arXiv:1908.09139)
Handley W., 2019b, Journal of Open Source Software, 4, 1414
Handley W., Alsing J., 2020, Nested sampling with any prior you like (Supplementary inference products), doi:10.5281/zenodo.4247552, https://doi.org/10.5281/zenodo.4247552
Handley W. J., Hobson M. P., Lasenby A. N., 2015a, Monthly Notices of the Royal Astronomical Society: Letters, 450, L61–L65
Handley W. J., Hobson M. P., Lasenby A. N. 2015b, Monthly Notices of the Royal Astronomical Society, 453, 4385–4399
Higson E., Handley W., Hobson M., Lasenby A., 2019, MNRAS, 483, 2044
Jones M. C., Pewsey A., 2009, Biometrika, 96, 761
Kingma D. P., Salimans T., Jozefowicz R., Chen X., Sutskever I., Welling M., 2015, in Advances in neural information processing systems. pp 4743–4751
Moss A., 2020, MNRAS, 496, 328
Papamakarios G., Pavlakou T., Murray I., 2018, arXiv preprint arXiv:1705.07057
Papamakarios G., Nalisnick E., Rezende D. J., Mohamed S., Lakshminarayanan B., 2019, arXiv preprint arXiv:1912.02762
Planck Collaboration 2020a, A&A, 641, A3
Planck Collaboration 2020b, A&A, 641, A6
Ross A. J., Samushia L., Howlett C., Percival W. J., Burden A., Manera M., 2015, MNRAS, 449, 835
SH0ES Collaboration 2018, ApJ, 855, 136
Scott D., 2018, The Standard Model of Cosmology: A Skeptic’s Guide (arXiv:1804.01318)
Seljak U., Yu B., 2019, arXiv preprint arXiv:1901.04454
Skilling J., 2006, Bayesian Anal., 1, 833
Vagnozzi S., Di Valentino E., Garaizzo S., Melchiorri A., Mena O., Silk J., 2020, arXiv e-prints, p. arXiv:2010.02230
Verde L., Treu T., Riess A. G., 2019, arXiv e-prints, p. arXiv:1907.10625
Williams M. J., Veitch J., Messenger C., 2021, arXiv e-prints, p. arXiv:2102.11056

This paper has been typeset from a \LaTeX\ file prepared by the author.