Journal of Cosmology and
Astroparticle Physics

SISSA

PAPER

Reconstructing the universe with variational self-boosted sampling

To cite this article: Chirag Modi *et al* JCAP03(2023)059

View the article online for updates and enhancements.

Journal of Cosmology and Astroparticle Physics
An IOP and SISSA journal

Reconstructing the universe with variational self-boosted sampling

Chirag Modi,[a,b] **Yin Li**[a,b,d] **and David Blei**[b,c]

[a]Center for Computational Astrophysics, Flatiron Institute,
New York, NY, U.S.A.
[b]Center for Computational Mathematics, Flatiron Institute,
New York, NY, U.S.A.
[c]Columbia University,
New York, NY, U.S.A.
[d]Department of Mathematics and Theory, Peng Cheng Laboratory,
Shenzhen, Guangdong, China

E-mail: cmodi@flatironinstitute.org, eelregit@gmail.org, david.blei@columbia.edu

Received August 24, 2022
Revised December 27, 2022
Accepted March 6, 2023
Published March 29, 2023

Abstract. Forward modeling approaches in cosmology have made it possible to reconstruct the initial conditions at the beginning of the Universe from the observed survey data. However the high dimensionality of the parameter space still poses a challenge to explore the full posterior, with traditional algorithms such as Hamiltonian Monte Carlo (HMC) being computationally inefficient due to generating correlated samples and the performance of variational inference being highly dependent on the choice of divergence (loss) function. Here we develop a hybrid scheme, called variational self-boosted sampling (VBS) to mitigate the drawbacks of both these algorithms by learning a variational approximation for the proposal distribution of Monte Carlo sampling and combine it with HMC. The variational distribution is parameterized as a normalizing flow and learnt with samples generated on the fly, while proposals drawn from it reduce auto-correlation length in MCMC chains. Our normalizing flow uses Fourier space convolutions and element-wise operations to scale to high dimensions. We show that after a short initial warm-up and training phase, VBS generates better quality of samples than simple VI approaches and in the hybrid sampling phase, reduces the correlation length in the sampling phase by a factor of 10–50 over using only HMC to explore the posterior of initial conditions in 64^3 and 128^3 dimensional problems, with larger gains for high signal-to-noise data observations. Hybrid sampling with online training of the variational distribution violates Markov property, and to retain the asymptotic guarantees of HMC, in the final phase we use a fixed variational distribution as proposal distribution and propagate these samples to the posterior distribution.

Keywords: cosmological parameters from LSS, cosmological simulations, Machine learning , Statistical sampling techniques

ArXiv ePrint: 2206.15433

https://doi.org/10.1088/1475-7516/2023/03/059

Contents

1 Introduction

Forward modeling approaches for cosmological analysis is one of the most promising ways to fulfill the scientific potential of the current and upcoming cosmological surveys such as DESI [1], LSST [2], Euclid [3] and others. In this approach, we simulate the field level cosmological survey data such as galaxies, starting all the way from the cosmological parameters and the initial conditions at the beginning of the Universe. This allows one to make model comparison at the field level, thus maximizing the amount of information that can be extracted from these surveys as one no longer relies on any compressed statistics of the survey data [4, 5]. However, at the same time, it makes inference challenging due to the much larger dimensionality of the parameter space since to infer cosmological parameters, we now need to marginalize over the density field at all points in the Universe.

Recent works have investigated a number of ways to infer the Gaussian initial conditions of the Universe in a Bayesian framework. The common theme amongst them is to begin by constructing a posterior of the initial conditions by combining the Gaussian prior on them with the data likelihood at field level. To make inference tractable in high dimensions, these then use differentiable simulations [6–9] which allow one to analytically estimate response of the observed data with respect to these underlying parameters.

The simplest inference is to reconstruct a maximum-a-posterior (MAP) estimate by maximizing this posterior with traditional gradient based optimization algorithms [4, 10, 11] or by using learnt optimization [12]. While being the fastest way, this provides only a point estimate for the initial conditions and one can estimate uncertainties by making a Laplace approximation [4, 13, 14].

A more robust but expensive approach to infer the posterior is to use Markov Chain Monte Carlo (MCMC) methods [8, 9], particularly Hamiltonian Monte Carlo (HMC) [15–17]. HMC generates samples from the posterior by simulating a Markov chain following Hamiltonian dynamics for multiple steps and thus minimizes the random walk diffusive behavior between successive proposals by taking longer jumps. Despite this, successive samples generated by HMC are still correlated and these correlation lengths can be hundreds of samples long. Hence overall, the cost of this approach is prohibitively expensive for scaling up to the future cosmological surveys because one requires at least on the order of hundred independent posterior samples to ensure that the first two moments (mean and variance) are estimated correctly.

In this work, we take a step towards reducing this computational cost of sampling approaches by learning a proposal distribution and combining it with MCMC algorithms. We parameterize this proposal distribution as a normalizing flow [18] which is trained using samples from the MCMC chain itself in an initial warm-up (training) phase. We follow this with a hybrid sampling phase wherein samples are generated from both HMC kernel and the variational distribution alternately. These samples are also being used to update the variational distribution, with the motivation that the independent samples generated in this phase lead to faster exploration of the target distribution and can access the regions that might have been missed in the training phase. Thus, this online training improves the quality of the variational distribution. After training, the samples generated from the variational distribution either lie in the target distribution directly, or can be propagated to the same with short Markov chains. This is exploited in the final phase where we fix the variational distribution as the only proposal distribution and propagate samples from it to the target posterior distribution. Unlike the hybrid sampling phase, this allows us to retain the asymptotic convergence properties of HMC.

In statistics literature, similar approaches have been proposed to speed up HMC by improving or learning the geometry of the posterior distribution with a transport map [19, 20]. However our scheme most closely resembles the approach proposed in [21] which uses the variational distribution as a global sampler to facilitate mode jumping in multi-modal distributions.

Another perspective on our approach is lent by variational inference itself [22, 23]. The learnt proposal distribution for MCMC can instead be viewed as a variational approximation to the target distribution with the associated short Markov chains serving a corrections to the learnt approximation. Alternatively, turning things around yet again, we can also view this as first running MCMC to generate samples from the true distribution which can then be used for learning the variational parameters by minimizing a more constraining forward (inclusive) Kullback-Leibler (KL) divergence [24], and then using this learnt distribution to generate uncorrelated proposals and speed up MCMC.

The challenge then remains to be able to learn this high dimensional variational distribution with normalizing flows (NF). Scaling to three dimensional data and distributions with millions of parameters is still not feasible with commonly used architectures of normalizing flows. In this work we build upon a recently developed architecture that exploits rotational and translational symmetries for cosmological fields by performing convolutions in the dual Fourier space [25] instead of the usual grid (pixel) space. We show that with suitable modifications to the input base distribution of NF, our model is flexible enough to learn a good proposal kernel of HMC to explore the posterior distribution of the initial conditions.

The paper is organized as follows. We begin by setting up our cosmological inference problem formally in section 2 with the relevant details and notation. Then we briefly review the Hamiltonian Monte Carlo (HMC) and Variational Inference (VI) algorithms in section 3. We also discuss their performance on our example problem to set up a benchmark. Then we present our hybrid sampling scheme with in 4. For coherence, we have moved the detailed discussion of our normalizing flow into the appendix A. Finally we present the results in section 5 and conclude in section 6.

2 Setup

We are interested in forward modeling approach for cosmological inference wherein we begin with the cosmological parameters and the initial dark matter density field, build a forward model for the data, and use this to infer the posterior of the parameters as well as the initial conditions. In this first work, we use the final evolved dark matter density as the data. Furthermore, we focus only on inferring the posterior of the initial density field while keeping the cosmological parameters fixed to their true value since this is the challenging high-dimensional part of the problem.

2.1 Data and likelihood model

Our data ($\mathbf{y}_0$) is the dark matter density field on a cubic N^3 grid[1] where N is the number of grid points or pixels along each side of the cube. For the toy problem in this work, we consider small grids of N = 64 and 128 while analyzing future cosmological surveys will require scaling up to N = 256 or 512. We choose the boxsize between L = 500 Mpc/h and 1000 Mpc/h so as to keep the resolution of our toy model simulations same as future realistic problems.

The data has been generated from some unknown initial conditions, i.e. initial dark matter density field ($\mathbf{s}$) which is evolved under gravity with a realistic forward model (f) to simulate a final dark matter field ($\mathbf{y}$) and then corrupted with a noise model ($\mathbf{n}$). The simple forward model we use is the particle displacement predicted by the first order Lagrangian Perturbation theory (Zeldovich Approximation, ZA). We use ZA purely for computational reasons and expect our conclusions to qualitatively remain the same for more realistic forward models such as FastPM [26] or COLA [27]. We take our data noise to be Gaussian with known noise variance ($\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$) corresponding to the shot-noise of the dark matter particles. This allows us to compute the exact likelihood for our toy data.

The parameters to be inferred are the phases of the Gaussian field ($\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$) corresponding to this unknown initial dark matter density field ($\mathbf{s}$). There is a deterministic relationship between the cosmology parameters (Λ), the phases and the initial density field, $\mathbf{s} = g(\mathbf{z}, \Lambda)$. Then for our toy problem, it is identical whether we infer $\mathbf{s}$ or $\mathbf{z}$ since we keep

[1] Unless specified otherwise, all the bold-face symbols such as $\mathbf{x}$, $\mathbf{y}$, $\mathbf{s}$, $\mathbf{z}$ are N^3 vector corresponding to the cubic simulation grid. We refer to these as fields.

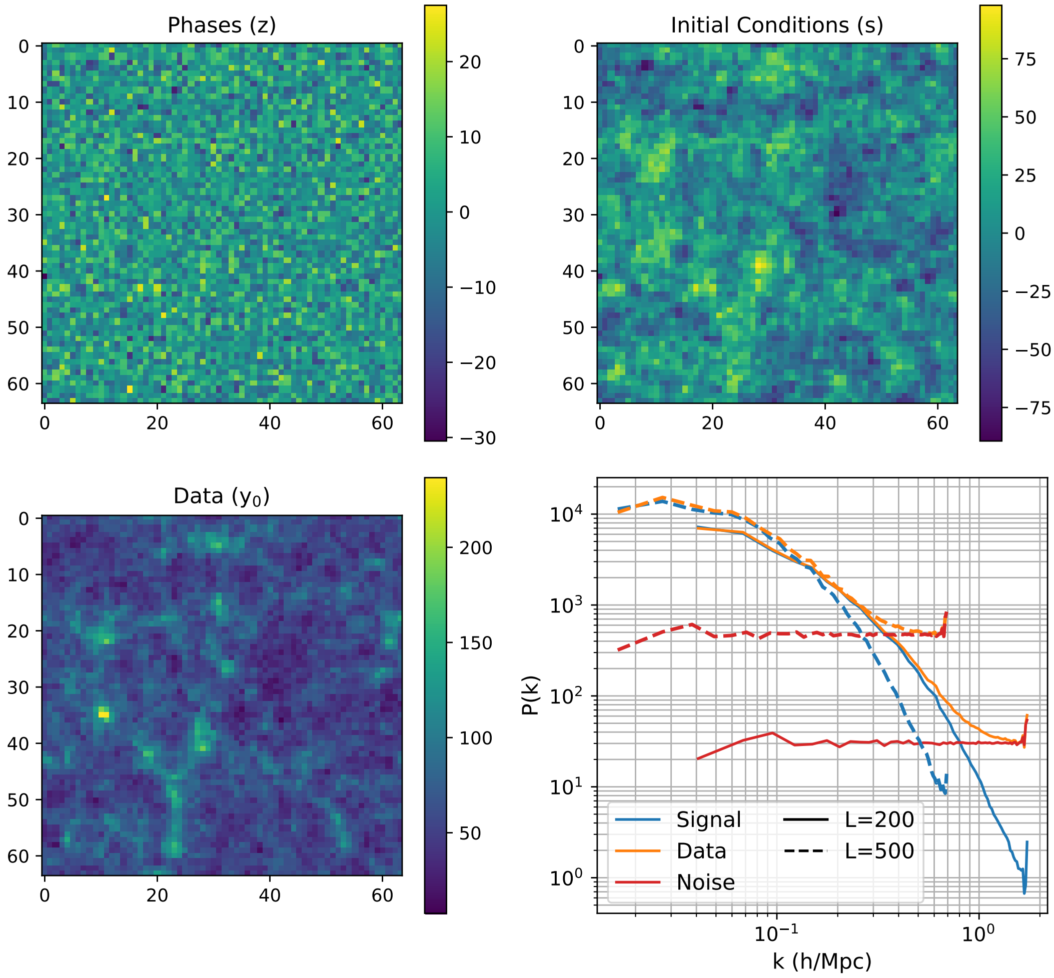


Figure 1. An example of the toy model: first three panel show different fields involved in our simulation for a L-200 Mpc/h and N = 64 grid. The fields are projected (summed) along the z-axis. The last panel shows the corresponding power spectrum for two different box sizes- L = 200 Mpc/h and 500 Mpc/h which correspond to different signal-to-noise ratio.

Λ fixed to their true value, but we choose to work with $\mathbf{z}$ here in the anticipation for follow up works when Λ needs to be inferred simultaneously with $\mathbf{z}$.

Then given the parameters $\mathbf{z}$, the likelihood model, prior, and the posterior distribution is

$$\begin{aligned} \pi(\mathbf{y}_0|\mathbf{z}) &= \mathcal{N}(\mathbf{y} = f(\mathbf{z}), \boldsymbol{\sigma}) && \text{(Gaussian likelihood)} \\ \pi(\mathbf{z}) &= \mathcal{N}(\mathbf{0}, \mathbf{1}) && \text{(Gaussian prior)} \\ \pi(\mathbf{z}, \mathbf{y}_0) &= \pi(\mathbf{y}_0|\mathbf{z})\pi(\mathbf{z}) && \text{(Unnormalized Posterior)} \end{aligned}$$

In figure 1, we show the different component fields of our problem for N = 64 grid: the phases ($\mathbf{z}$), the initial conditions ($\mathbf{s}$), and the data ($\mathbf{y}_0$, final dark matter field with noise). In the last panel, we also show the power spectra of the data signal and noise for two different box sizes which correspond to different signal to noise ratios.

2.2 Validating the posterior

Before discussing different approaches to inference, we briefly lay out the metrics that will be used to validate the posterior distribution inferred by them. We will do so by looking at the samples $\mathbf{z}_i$ (and their derived properties as described below) generated from these distributions.

2.2.1 Distribution of summary statistics

We are trying to infer the posterior of the phases of the initial conditions. This is a high dimensional density distribution, and they are notoriously hard to compare quantitatively. Thus we seek a low dimensional mapping to compare posterior samples from different algorithms. Our data model obeys rotational and translational invariance and hence the power spectrum of density fields provides a natural low dimensional candidate for this mapping. The power spectrum (P_a) of any field a measures the clustering of the overdensity field δ_a at different scales $\mathbf{k}$ and is defined as

$$\langle \tilde{\delta_a}(\mathbf{k})\tilde{\delta_a}^{*}(\mathbf{k}')\rangle = (2\pi)^3 P_a(k)\delta_D^3(\mathbf{k}-\mathbf{k}')$$

where k is the magnitude of the scale and δ_D^3 is the 3-D Dirac-delta function. We will measure the quality of the posterior distributions with the following two derived quantities of the power spectra of the posterior samples:

1. Cross correlation (r_c) of the samples from the posterior with the true initial conditions defined as
$$r_c = \frac{P_{ab}(k)}{\sqrt{P_a(k)P_b(k)}}$$
for any two fields a and b and where P_{ab} is their cross-spectra. Cross correlation between two fields effectively compares the phases of the fields i.e. if the features such as peaks and voids of the density field are physically at the same location on different scales. We expect r_c to be consistent with unity on large scales which are signal dominated and then drop to zero on the scales where our posterior is prior dominated.

2. Transfer Function (t_f) of the samples from the posterior with the true initial conditions defined as
$$t_f = \sqrt{P_b(k)/P_a(k)}$$
for any two fields a and b. Transfer function compares the amplitude of clustering at different scales. Since we use the same cosmology for data generation and inference, t_f of samples from the correct posterior should be consistent with unity on all scales.

We will show these summary statistics for the posterior samples which are the phases of the initial dark matter density field $\mathbf{z}$. Thus in our discussion, the two fields in comparisons are

$a := \mathbf{z}_0$	Phases of the true initial dark matter field corresponding to data $\mathbf{y}_0$
$b := \mathbf{z}_i$	Posterior sample $\mathbf{z}_i$

Note that when both r_c and t_f are unity, the two fields being compared are identical.

2.2.2 Auto-correlation length

Monte Carlo algorithms explore the posterior by generating samples from it instead of optimizing (learning) a parametric form of them. In this case it is important to have generated enough independent samples such that we are confident to have explored both the bulk and the tails of the posterior adequately. Thus the efficacy of such algorithms is measured with auto-correlation length which is the effective length (number of samples) between two successive independent samples.

As discussed above, due to the high dimensional nature of our problem, we will again work with low-dimensional summary statistic for quantitiative comparisons. Hence we compare the efficacy of algorithms by estimating the auto-correlation length for power spectrum of the posterior samples. Specifically, for every chain, we measure the power spectrum $P_i(k)$ for each sample $\mathbf{z}_i$ and then estimate the correlation length for each mode k_j as

$$\rho_j(t) = \frac{1}{n} \sum_{i=t+1}^{n} (P_i(k_j) - \bar{P}(k_j))(P_{i-t}(k_j) - \bar{P}(k_j)) \tag{2.1}$$

where $\bar{P}(k_j)$ is the mean power in the mode k_j across all samples of that chain and n is the total number of samples. Then the auto-correlation length (a_c) is defined as the scale where $\rho_j(a_c) \leq 0.1$. We want the auto-correlation a_c as small as possible since it implies more independent samples for the same computational cost.

3 Posterior of initial conditions

We begin by briefly reviewing the two most widely used approaches for posterior inference, which will also form the building blocks of our hybrid sampling presented in the next section. These are- i) Hamiltonian Monte Carlo (HMC) which generates samples from the posterior directly and ii) Variational Inference (VI) which learns a parametric form of the posterior distribution. For each approach, we will also discuss their merits, drawbacks, and evaluate their performance for our particular problem to infer the posterior of the initial conditions.

3.1 Hamiltonian Monte Carlo (HMC)

HMC is a widely used approach to generate samples from distributions in high dimensions. It begins by reinterpreting the parameters of interest as a position vector $\mathbf{q} \in R^d$ with the associated potential energy function $U(\mathbf{q}) = -\log \pi(\mathbf{q})$ where $\pi(\mathbf{q})$ is the target distribution (in this case, the unnormalized posterior), and introducing an auxiliary momentum vector $\mathbf{p} \in R^d$ which contributes a kinetic energy term $K(\mathbf{p}) = \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p}$, where M is a symmetric positive definite mass matrix. The choice of mass matrix can affect the performance of HMC. In this work, we take the mass matrix to be the identity matrix, $M = I$, as is often done for simplicity, and in this particular case is also the correct choice for sampling the phases of the initial conditions. With these, one can construct the Hamiltonian $H : \mathcal{R}^{2d} \to \mathcal{R}$ as the total energy function for the state $\mathbf{x} := (\mathbf{q}, \mathbf{p})$,

$$\begin{aligned} H(\mathbf{x}) \equiv H(\mathbf{q}, \mathbf{p}) &= U(\mathbf{q}) + \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} \\ &= -\log \pi(\mathbf{q}) + \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} \ . \end{aligned} \tag{3.1}$$

Procedure 1 Single step of Hamiltonian Monte Carlo Sampling

```
Input:
    · current position, $z_0$
    · target probability density, $\pi$
    · step-size, $\epsilon$
    · number of leapfrog steps, $L$
    · Hamiltonian, $H$
    · Mass matrix for momentum, $M$
Output:
    · next sample in the chain, $z_1$
$q_0 \leftarrow z_0$                                   ▷ Assign current sample as the initial position
$p_0 \sim \mathcal{N}(0,1)$                            ▷ Sample random momentum of same shape as $q_0$
$i = 0$
while $i \leq L$ do                                    ▷ Integrate Hamiltonian equations for $L$ leapfrog steps
    $q_{i+1},\ p_{i+1} \leftarrow$ LEAPFROG$(q_i,\ p_i,\ \pi, \epsilon)$
    $i \leftarrow i+1$
$H_0 \leftarrow \mathrm{H}(q_0,\ p_0, \pi, M)$          ▷ Estimate Hamiltonian using eq. (3.1)
$H_L \leftarrow \mathrm{H}(q_L,\ p_L, \pi, M)$
$\alpha \leftarrow \exp(H_0 - H_L)$                     ▷ Acceptance probability to maintain DB eq. (3.2)
if Uniform$(0,1) \geq \alpha$ then
    $z_1 \leftarrow q_0$
else
    $z_1 \leftarrow q_L$
return $z_1$
```

The goal is to simulate a Markov chain and generate samples from the target distribution. This is achieved by evolving this physical system with respect to time by following Hamiltonian dynamics. The Hamiltonian's equations are numerically evolved by integrating the ODE system using leapfrog integrator [16]. Hence there are two parameters to be tuned- the stepsize of the integration ϵ and the number of leapfrog steps (L) to take before making a proposal $\mathbf{q}_i$.

Proposals generated at the end of each iteration are accepted or rejected to maintain a detailed balance (DB) condition which guarantees that the samples are generated from a stationary target distribution. As per detailed balance, the probability of accepting a proposal $\mathbf{x}_0 \rightarrow \mathbf{x}_1$ is

$$\alpha = \min(1, \exp(H(\mathbf{x}_0) - H(\mathbf{x}_1)) \tag{3.2}$$

The complete algorithm for generating proposals is described in algorithm 1. A more in-depth discussion of HMC can be found in [15–17].

3.1.1 Cost of HMC

We set up HMC for our toy problem with the parameters of interest, $\mathbf{z}$ corresponding to the position vector in HMC $\mathbf{q}$. We fit ϵ by dual averaging scheme [28] and based on some preliminary experiments, we randomly choose the number of leapfrog steps L uniformly between 25 and 50 for every proposal. We run 4 independent chains for robustness, generating 5000 samples in each chain and then thinning them by a factor of 20. Figure 2 shows the

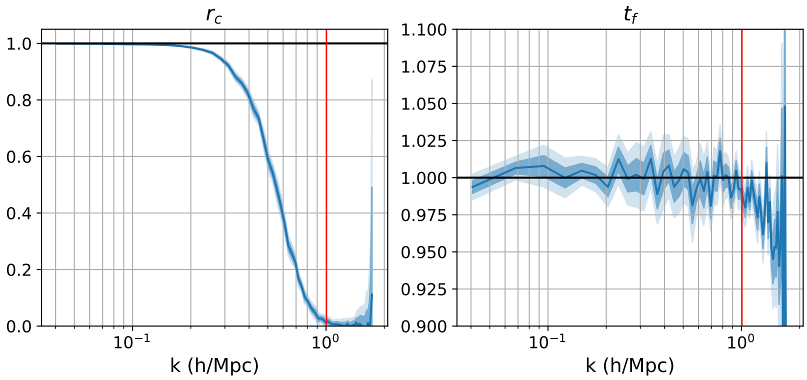


Figure 2. Posterior explored with HMC: we show the distribution of summary statistics for samples from posterior generated with HMC after sub-sampling by a factor of 20. The results are shown for our fiducial configuration of $\mathrm{L} = 200\,\mathrm{Mpc/h}$ and $\mathrm{N} = 64$ simulation. The vertical red line represents nyquist frequency. Note both cross correlation and transfer function being 1 implies a perfect inference.

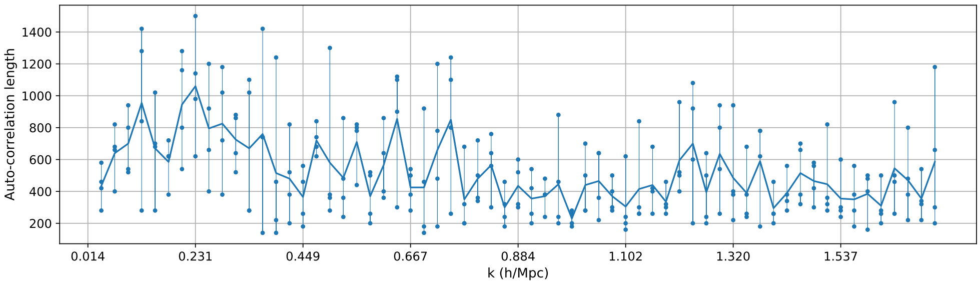


Figure 3. Cost of HMC: we show auto-correlation length of HMC samples as estimated for the power in different modes for our fiducial configuration of $\mathrm{L} = 200\,\mathrm{Mpc/h}$ and $\mathrm{N} = 64$. Different points along the same vertical (k-mode) are four different chains.

summary statistics for the posterior samples generated by HMC for the configuration with $\mathrm{N} = 64$ and $\mathrm{L} = 200\,\mathrm{Mpc/h}$ simulation. Both the cross correlation and transfer function follow the expected behavior — the cross correlation is one on large scales and falls to zero on small scales while the transfer function is unity across all scales upto the nyquist frequency.

While HMC samples the distribution correctly, it is only true in asymptotic sense and consecutive samples are still correlated which makes the algorithm computationally expensive. To gauge it's efficacy, we estimate the auto-correlation length of the chains in terms of their power spectra as described in eq. (2.1). This is shown in figure 3 as function of different scales. The correlation length is larger on the largest scales which are more signal dominated than the noisy small scales. On the largest scales, the correlation length reach upto a 1000 samples long. Taking into account the fact that each sample is generated after taking $L \in [25, 50]$ leapfrog steps, the cost of a single independent sample on these largest scales can easily be up to ∼10000 forward simulations. This makes HMC prohibitively expensive for scaling to problems of larger size and higher signal-to-noise since one expects the correlation lengths in these cases to be longer still. However due to the algorithm's guaranteed asymptotic correctness, we will use HMC samples as benchmark to compare other posteriors throughout this work.

3.2 Variational inference

Variational inference [22] takes a different approach from sampling and instead approximates the target distribution π with a distribution belonging to a parametric family, $q(\boldsymbol{\nu})$. The parameters $\boldsymbol{\nu}$ are estimated so as to minimize a divergence between the variational distribution $q(\boldsymbol{\nu})$ and the target distribution π. Since this minimization is an optimization, VI is generally much faster than HMC but does not enjoy the guarantees of asymptotic correctness same as HMC. In fact, as we will show in this section, the quality of VI depends significantly on the choice of parametric family and the objective function which is used for this optimization.

3.2.1 Normalizing flow

In this work, we approximate the posterior of the phases of the initial conditions $\pi(\mathbf{z}|\mathbf{y}_0)$ with a normalizing flow (NF) [18]. Normalizing flows consist of a series of invertible, bijective mappings that successively transform (transport) a simple base distribution to a complex distribution with non-trivial correlations. Let this transport map be given by T_θ with parameters θ and the base distribution be given by q_B. Then the variational family of our posterior as parameterized by the normalizing flow is

$$q(\mathbf{z};\boldsymbol{\nu}) = q_B(T_\theta^{-1}(\mathbf{z});\boldsymbol{\nu}_B)|\det \nabla_{\mathbf{z}} T_\theta^{-1}| \tag{3.3}$$

The set of variational parameters consists of both the parameters of the base distribution as well as the transport map $\boldsymbol{\nu} = \{\boldsymbol{\nu}_B, \theta\}$.

For our base distribution, we choose the mean-field normal i.e. $q(\mathbf{z};\boldsymbol{\nu}_B) := \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with a diagonal covariance matrix. Our transport map is inspired by the fact that the distribution of cosmological fields is rotationally and translationally invariant [25]. Building upon these invariances, we can model the convolutions as simple products in Fourier space. This maintains bijectivity and easy Jacobian evaluations which allows us to learn the high dimensional distributions of interest. We alternate these Fourier convolutions with other element-wise bijective operations. Further details of our normalizing flow are given in the appendix A where we also demonstrate that this variational distribution is flexible enough to learn a good approximation for the target posterior distribution.

3.2.2 Backward or exclusive KL divergence

The other component of variational inference is the choice of divergence to be minimized between the variational distribution and the target distribution. The most commonly used divergence is Kullback-Leibler (KL) divergence with the variational distribution as the reference distribution, In this case its called backward or exclusive KL divergence and is defined as

$$\begin{aligned} D_{\mathrm{KL}}(q||p) &= \mathbb{E}_q(\log q - \log p) \\ &= \mathbb{E}_q(\log q(\mathbf{z};\boldsymbol{\nu}) - \log \pi(\mathbf{z}|\mathbf{y}_0)) \\ &\approx \sum_{\mathbf{z}_i\sim q(\mathbf{z})} [\log q(\mathbf{z}_i;\boldsymbol{\nu}) - \log \pi(\mathbf{z}_i|\mathbf{y}_0)] \\ &\leq \sum_{\mathbf{z}_i\sim q(\mathbf{z})} [\log q(\mathbf{z}_i;\boldsymbol{\nu}) - \log \pi(\mathbf{y}_0|\mathbf{z}_i) - \log \pi(\mathbf{z}_i)] \end{aligned} \tag{3.4}$$

where in the third line we have approximated the expectation with empirical expectation as estimated by the samples $\mathbf{z}_i \sim q(\mathbf{z};\boldsymbol{\nu})$ from the variational family. In the last line, we expand the posterior distribution in terms of the likelihood and the prior while dropping the

Procedure 2 Backward/Exclusive Variational Inference

```
Input:
    · variational family with parameters, $\boldsymbol{\nu}$ $q(\mathbf{z};\boldsymbol{\nu})$;
    · likelihood function, $\pi(\mathbf{y}_0|\mathbf{z})$
    · prior, $\pi(\mathbf{z})$
    · step-size for optimizer, $\epsilon$
    · maximum number of iterations, $N$
    · number of samples per iteration, $n$
Output:
    · Trained variational distribution, $q(\mathbf{z};\boldsymbol{\nu}^*)$
$i = 0$
while $i \leq N$ do
    $\{\mathbf{z}_i \dots \mathbf{z}_n\} \sim q(\mathbf{z};\boldsymbol{\nu})$        ▷ Generate $n$ samples from variational distribution
    ELBO $= \sum_{\mathbf{z}_i} \log \pi(\mathbf{y}_0|\mathbf{z}_i) + \log \pi(\mathbf{z}_i) - \log q(\mathbf{z}_i;\boldsymbol{\nu})$
    $\boldsymbol{\nu} \leftarrow \boldsymbol{\nu} - \epsilon \nabla_{\boldsymbol{\nu}}$ELBO        ▷ Optimization
$\boldsymbol{\nu}^* \leftarrow \boldsymbol{\nu}$
return $q(\mathbf{z};\boldsymbol{\nu}^*)$
```

evidence term which is a negative constant with respect to the variational parameters. This is also called the evidence lower bound (ELBO)

$$\text{ELBO} := \sum_{\mathbf{z}_i \sim q(\mathbf{z})} \log \pi(\mathbf{y}_0|\mathbf{z}_i) + \log \pi(\mathbf{z}_i) - \log q(\mathbf{z}_i;\boldsymbol{\nu}) \tag{3.5}$$

For inferring the posterior, backward VI maximizes the ELBO with respect to the variational parameters.

$$\boldsymbol{\nu}^* = \text{argmax}_{\nu}\text{ELBO} \tag{3.6}$$

The full algorithm for this is given in algorithm 2.

We learn the posterior for the same configuration of our toy problem as before, i.e. N = 64 and L = 200 Mpc/h simulation, with the aforementioned normalizing flow and by minimizing the backward KL loss. In figure 4, we show the summary statistics for the samples generated by the learnt variational distribution. Note that while the cross-correlation of the samples is similar to the correct samples generated by HMC, the transfer function is wrong. There can be two potential reasons for this- i) our parametric family i.e. the combination of mean-field Gaussian and the normalizing flow is not flexible enough to model the correct posterior, ii) our choice of divergence is not appropriate to learn the correct distribution. However as we show in appendix A, our normalizing flow is indeed powerful enough to learn the distribution of samples generated from HMC (atleast at the level of the summary statistics). This means that in our case, the backward KL loss is not constraining enough to learn the correct variational distribution. This is not entirely unexpected since backward KL loss is known to have mode-seeking behavior and does not guarantee coverage of the full target distribution.

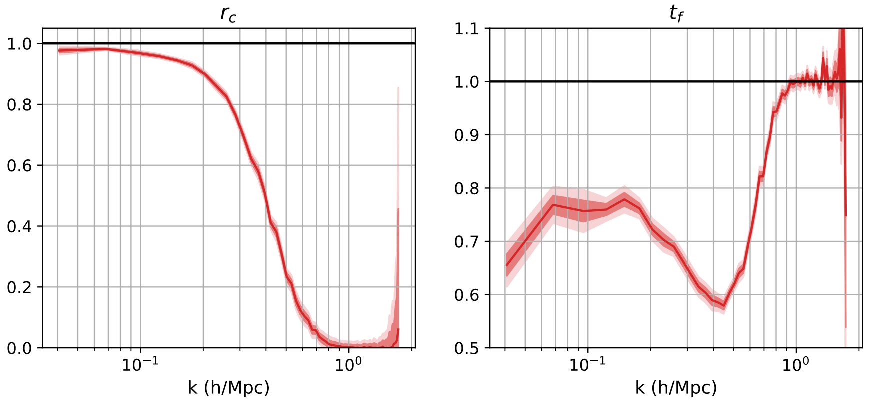


Figure 4. Posterior learnt with backwards VI: we show the cross correlation and transfer function for samples from the variational distribution fit by maximizing ELBO as in algorithm 2. The result is shown for our fiducial configuration of L = 200 Mpc/h and N = 64 simulation. The solid lines and shaded regions show the mean and 1,2σ variations of 100 samples generated from the variational distribution after fitting.

3.2.3 Forward or inclusive KL divergence

An alternative to backward KL divergence is the forward KL divergence which uses the target distribution as the reference. Then

$$D_{\mathrm{KL}}(p||q) = \mathbb{E}_p(\log p - \log q) \tag{3.7}$$

$$= \mathbb{E}_{\pi(\mathbf{z}|\mathbf{y}_0)}(\log \pi(\mathbf{z}|\mathbf{y}_0) - \log q(\mathbf{z};\boldsymbol{\nu})) \tag{3.8}$$

$$\approx \sum_{\mathbf{z}_i \sim \pi(\mathbf{z}|\mathbf{y}_0)} (\log \pi(\mathbf{z}|\mathbf{y}_0) - \log q(\mathbf{z};\boldsymbol{\nu})) \tag{3.9}$$

where we have again approximated the expectation with empirical expectation. Note that since the samples are generated from the target distribution itself, the first term is independent of the variational parameters. Thus minimizing this divergence for variational inference is achieved by maximizing the log-probability of the samples under the variational distribution

$$\boldsymbol{\nu}^* = \mathrm{argmax}_\nu \sum_{\mathbf{z}_i \sim \pi(\mathbf{z}|\mathbf{y}_0)} \log q(\mathbf{z};\boldsymbol{\nu}) \tag{3.10}$$

Looking at this equation, we can see the chicken-and-egg problem of the forward KL loss — we need samples $\mathbf{z}_i$ from the target distribution (e.g., as generated by HMC) to learn the variational distribution, but if we had an easy access to such samples, we would not need to learn a variational distribution in the first place. Recent works have investigated some ways to get around this, such as with importance weighing the samples generated from the variational distribution [20, 29]. However we find that none of these approaches work well in our case. Hence in the next section, we turn back to HMC and VI to combine them in a hybrid approach that benefits from the complementary advantages of both the algorithms.

Procedure 3 Variational (self-)Boosted Sampling

```
Input:
    · Initial sample from the target distribution, $\mathbf{z}_0$
    · variational family, $q(\mathbf{z};\boldsymbol{\nu})$
    · target distribution (posterior), $\pi(\mathbf{z}|\mathbf{y}_0)$
    · annealed target distribution for VI, $\pi^*(\mathbf{z}|\mathbf{y}_0)$
    · step-size for HMC, $\epsilon$
    · step-size for training, $\epsilon_q$
    · number of leapfrog steps, $L$
    · mass matrix, $M$
    · number of HMC iterations for training, $N_1$
    · number of samples to generate in hybrid sampling, $N_2$
    · probability of generating proposal from VI distribution, $p_{\text{jump}}$
    · number of samples to generate with fixed variational distribution, $N_3$
    · termination criterion for propagating chains from fixed variational distribution, $g$
Output:
    · Samples from phase IIA, $\{\mathbf{z}_1 \ldots \mathbf{z}_{N_1+N_2}\}$
    · Samples from phase IIB, $\{\mathbf{z}^{(0)}, \ldots, \mathbf{z}^{(N_3)}\}$
    · Learnt variational distribution, $q(\mathbf{z};\boldsymbol{\nu}^*)$
$i = 0$
while $i \leq N_1$ do                                                ▷ Phase I, Learning
    $\mathbf{z}_{i+1} \leftarrow \text{HMC step}(\mathbf{z}_i, \pi, \epsilon, L, H, M)$
    Sample batch $\mathcal{B} = \{\mathbf{z}_{(1)} \ldots \mathbf{z}_{(B)}\}$ uniformly from $\{\mathbf{z}_1 \ldots \mathbf{z}_i\}$
    $\mathcal{L} = -\sum_B \log q(\mathbf{z}_{(i)};\boldsymbol{\nu})$
    $\boldsymbol{\nu} \leftarrow \boldsymbol{\nu} - \epsilon \nabla_{\boldsymbol{\nu}} \mathcal{L}$                                  ▷ Optimization
while $i \leq N_1 + N_2$ do                                   ▷ Phase IIA, Hybrid Sampling
    if Uniform$(0,1) \geq p_{\text{jump}}$ then
        $\mathbf{z}_{i+1} \leftarrow \text{HMC step}(\mathbf{z}_i, \pi, \epsilon, L, H, M)$
    else
        $\mathbf{z} \sim \log q(\mathbf{z};\boldsymbol{\nu})$
        $\alpha = \frac{\pi^*(\mathbf{z}) q(\mathbf{z}_i;\boldsymbol{\nu})}{\pi^*(\mathbf{z}_i) q(\mathbf{z};\boldsymbol{\nu})}$
        $\mathbf{z}_{i+1} \leftarrow \mathbf{z}$ with probability $\alpha$, otherwise $\mathbf{z}_{i+1} \leftarrow \mathbf{z}_i$
    Sample batch $\mathcal{B} = \{\mathbf{z}_{(1)} \ldots \mathbf{z}_{(B)}\}$ uniformly from $\{\mathbf{z}_1 \ldots \mathbf{z}_i\}$
    $\mathcal{L} = -\sum_B \log q(\mathbf{z}_{(i)};\boldsymbol{\nu})$
    $\boldsymbol{\nu} \leftarrow \boldsymbol{\nu} - \epsilon_q \nabla_{\boldsymbol{\nu}} \mathcal{L}$                                 ▷ Optimization
$i = 0$
for $j \leq N_3$ do                                                ▷ Phase IIB, Sampling
    $\mathbf{z}^{(j)} \sim \log q(\mathbf{z};\boldsymbol{\nu})$
while $g\big(\mathbf{z}^{(0)}, \ldots, \mathbf{z}^{(N_3)}\big)$ do
    for $j \leq N_3$ do
        $\mathbf{z}^{(j)} \leftarrow \text{HMC step}(\mathbf{z}^{(j)}, \pi, \epsilon, L, H, M)$
return $\{\mathbf{z}_1 \ldots \mathbf{z}_{N_1+N_2}\}, \{\mathbf{z}^{(0)}, \ldots, \mathbf{z}^{(N_3)}\}, q(\mathbf{z};\boldsymbol{\nu}^*)$
```

4 Variational self-boosted sampling (VBS)

In the previous section, we considered two approaches to infer posterior distribution— i) HMC which is accurate but prohibitively expensive and ii) VI which is fast but inaccurate when using backward KL divergence. In this section we propose a hybrid approach which combines VI with sampling [19–21]: we use the samples generated from HMC to train a variational approximation $q(\mathbf{z};\boldsymbol{\nu})$ to the target distribution on the fly and in turn simultaneously make proposals from the variational distribution to break correlations in successive samples generated in MCMC [21]. As the variational approximation improves over iterations, it will also become a good proposal kernel for the Monte Carlo chain and its samples will be readily accepted, thus reducing the correlation length of HMC. We call this scheme variational self-boosted sampling (VBS).

4.1 Algorithm

Starting from a sample point in the target distribution, our algorithm can broadly be divided into two phases-[2] i) learning phase and ii) sampling phase. The full algorithm is presented in algorithm 3 but briefly, the two phases are:

- **Phase I, Learning Phase**:
  In this phase we only run vanilla HMC chains to generate samples from the exact posterior while simultaneously using these samples to learn the variational parameters using eq. (3.10). We do not thin HMC samples i.e. we use all the samples which can be correlated. This phase lasts until the variational approximation learns the distribution of the current samples. The computational cost of this phase is practically the same as HMC since the training cost is sub-dominant.

- **Phase IIA, Hybrid Sampling**:
  In this phase we alternate between (with some pre-chosen probability, p_{jump}) making proposals from HMC kernel and the variational distribution. We need a criterion to accept these proposals to ensure that we sample from the correct target distribution. For HMC proposals, this is satisfied with the same detailed balance condition as before, eq. (3.2). For variational proposals, we discuss this criterion in detail below. At the same time, we continue to update the variational distribution with both the new and the old samples from the learning phase. This phase lasts until we have the requisite number of independent samples, or the quality of the variational samples stops improving. We will call the samples generated in this phase *VBS samples.*

- **Phase IIB, Sampling**:
  In this phase, we fix the variational distribution. We generate initial samples from this distribution, which by construction are now independent, and propagate these with HMC until we reach a sample from the target distribution. We need a termination criterion for these chains, and we note that it is the same as the termination criterion for warm-up phase in standard HMC. In our example, we monitor the true posterior probability ($\log p$) of the samples and terminate when these values stop changing more

[2] Here we have assumed that we begin with an initial sample from the target distribution. If instead we do not have such a sample, there is a warmup or burn-in phase to initialize from a random point and reach such a sample from the target distribution. However since this is identical to HMC, we do not include it explicitly as a part of the algorithm.

than a pre-defined threshold (see figure 7, left panel). Other termination criterion can include using convergence statistics like $\hat{R}$ [30].

How to choose the length of Phase I, IIA and IIB depends on the particular problem setup i.e. on factors such as the complexity of the target distribution, the flexibility of the normalizing flow, and the inductive biases imposed on the architecture of the flow to name a few. If the variational distribution is able to model the target distribution accurately, one may not need to generate samples from Phase IIB. However, one reason to include Phase IIB is that we continuously adapt the variational distribution in Phase IIA, hence it is not Markovian. Phase IIB is needed to enjoy the theoretical guarantees of HMC.

On the other hand, if the variational distribution is not flexible enough to capture the target distribution completely, then the procedure is likely not ergodic with only the first two phases and Phase IIB is required. There are different ways to combine Phase IIA and Phase IIB. In algorithm 3, we outline the sequential approach in which we generate a fixed, pre-determined number of samples in Phase IIA. One can also smoothly transition from Phase IIA to IIB by decreasing the adaptation rate with iterations, or simply skip Phase IIA and directly move to Phase IIB after the learning phase. As we show later, the last approach turns out to be the most efficient for our toy problem.

In Phase IIA, we still need a detailed balance condition for the acceptance of variational proposals similar to HMC. Let $\mathbf{z}_1$ be the current sample and $\mathbf{z}_2$ be the proposal from the variational distribution $q(\mathbf{z}, \boldsymbol{\nu})$. Then the detailed balance condition is met if the acceptance probability α of making the transition $\mathbf{z}_1 \rightarrow \mathbf{z}_2$ is

$$\alpha = \min\left(1, \frac{\pi^*(\mathbf{z}_2)q(\mathbf{z}_1; \boldsymbol{\nu})}{\pi^*(\mathbf{z}_1)q(\mathbf{z}_2; \boldsymbol{\nu})}\right) \tag{4.1}$$

This is the balance condition to correctly sample from the distribution $\pi^*(\mathbf{z})$, which should ideally be the target unnormalized posterior probability of the sample, $\pi(\mathbf{z}, \mathbf{y}_0)$.

However we find in our experiments that due to the high dimensional nature of our posterior distribution and incomplete exploration of this distribution in the learning phase with the correlated samples, the variance in acceptance probability is quite high and makes the scheme inefficient. To reduce this variance, we re-scale the likelihood with the number of grid points while keeping the prior unchanged, $\pi^*(\mathbf{z}) = \pi(\mathbf{y}_0|\mathbf{z})^{1/N^3}\pi(\mathbf{z})$. In this view, we then consider the learnt variational distribution to be *a proposal distribution for MCMC wherein we can quickly reach samples from the target by running a short chain starting from this proposed point.* This is why we alternate between variational proposal and HMC proposal with a pre-set probability p_{jump}. We find that $p_{\text{jump}} \sim 0.2$ gives a good balance between the quality of samples and the acceptance rate of proposals generated from the variational distribution.

5 Results

In this section, we present results for our proposed VBS scheme. Based on the discussion of merits and issues for HMC and VI in section 3.1 and 3.2 respectively, as well as our motivations for the VBS scheme, our goal here will be three-fold: i) to verify that the samples generated by our scheme are correct and follow the same distribution as HMC samples, ii) to establish that the learnt variational distribution with forward KL is still insufficient and is improved upon with short HMC chains, and finally iii) to gauge the gains of our hybrid scheme over HMC as measured with the auto-correlation length of the chains.

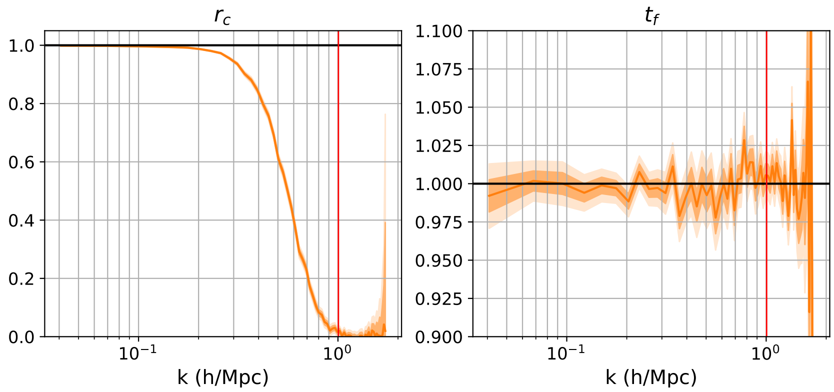


Figure 5. Posterior explored with VBS: we show the cross correlation and transfer function for 1500 consecutive samples from Phase II of hybrid sampling. The results are shown for our fiducial configuration of L = 200 Mpc/h and N = 64 simulation. The vertical red line represents nyquist frequency.

5.1 Verifying VBS posterior

We begin with verifying the posterior of the hybrid approach. As for HMC, we run 4 chains in parallel. However unlike HMC, these chains are coupled since the samples from all the chains are used to train the NF at each update step and then proposals from the NF are generated for each chain in the hybrid sampling phase. Based on our experiments, we set $p_{\text{jump}} = 0.2$ and the number of samples in training phase N_1=500. Note that this value of N_1 leads to only 2 independent samples or less on the largest scales (see figure 3) and hence we are initially training the NF with mostly correlated samples. We find the performance of our scheme to be quite robust to M and p_{jump} within reasonable limits. We use the same parameters as before for the HMC step of the hybrid scheme i.e. ϵ fit by dual averaging scheme [28] and the number of leapfrog steps L is chosen uniformly between 25 and 50 for every proposal.

Figure 5 shows the distribution of the summary statistics for the samples generated with VBS scheme at the end of Phase IIA for the baseline case we have been considering, L = 200 Mpc/h and N = 64 simulations. We find that both the transfer function and the cross correlation match the expected behavior with the former being unity on all scales and the latter only on the signal dominated scales. Moreover, the scatter in the transfer function across scales is similar to that of HMC samples, pointing towards the two distributions being consistent.

5.2 Importance of short MCMC chains

Our next goal is to show that the hybrid nature of our scheme is indeed important and simply using HMC samples to train the variational distribution with a forward KL loss as proposed in section 3.2.3 is insufficient.

To this end, we take the NF trained at the end of Phase I and Phase IIA and generate samples from them. In figure 6, we compare the distribution of the transfer function (left column) of these samples with the HMC and VBS samples at the end of Phase IIA On the largest scales, the samples generated from either NF show a much larger variance than either of the sampling schemes.

In the right panel, we also show distribution of the unnormalized log p values, i.e. the target posterior probabilities of these samples. The samples from NF after Phase I (learning

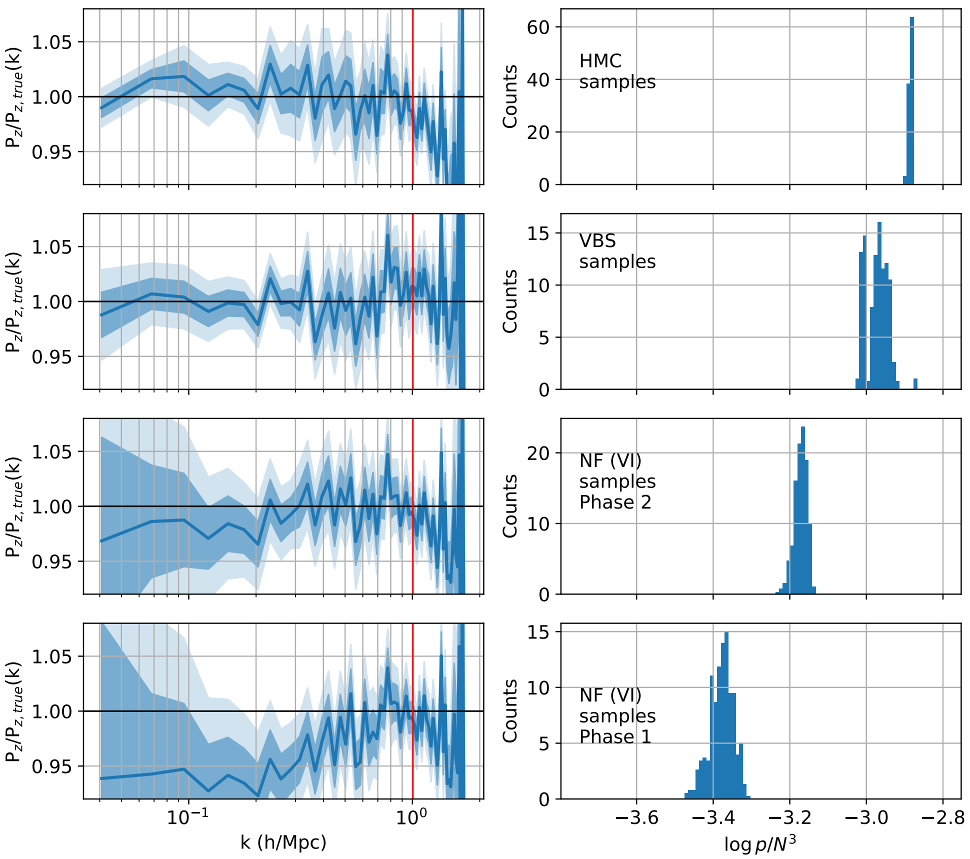


Figure 6. Comparing the posterior with different approaches: (left) We show the mean (solids), one- and two-standard deviation (shaded regions) of transfer function for samples of the phase field (**z**) from different approaches for our fiducial configuration of L = 200 Mpc/h and N = 64. The vertical red line represents nyquist frequency. (right) We show the distribution of log posterior probabilities, log p, for the samples generated by different approaches. (first row) Vanilla HMC where we have used 800 samples (4 chains, 200 samples each after thinning by factor 20). This acts as the benchmark. (second row) VBS samples at the end of Phase IIA where we use 4000 samples (4 chains, 1000 samples each without thinning). (third row) Samples generated by the variational distribution (NF) at the end of the Phase IIA. (fourth row) Samples generated by the variational distribution (NF) at the end of the Phase I. This corresponds to a variational distribution fit with forward KL loss, eq. (3.10). Though both VBS samples and VI samples after the second phase have transfer function consistent with unity, VBS samples after Phase IIA are of higher quality since the distribution of their log p is much closer to the benchmark HMC samples.

phase) are of the poorest quality since the training samples generated so far are correlated and explore only a tiny region in the distribution. Samples generated from NF after Phase IIA are better since now the NF has been trained over more independent samples from a larger region in the distribution. However the VBS samples are still markedly better than the either and much closer to the HMC samples. HMC samples do still have slightly higher probability than VBS samples because our DB criterion (eq. (4.1)) of accepting NF proposals in Phase IIA is based on a re-scaled version of the target distribution and not the exact target itself.

This shows that while it is not completely accurate to interpret the variational distribution as having learnt the target distribution, it can still serve as a good proposal distribution

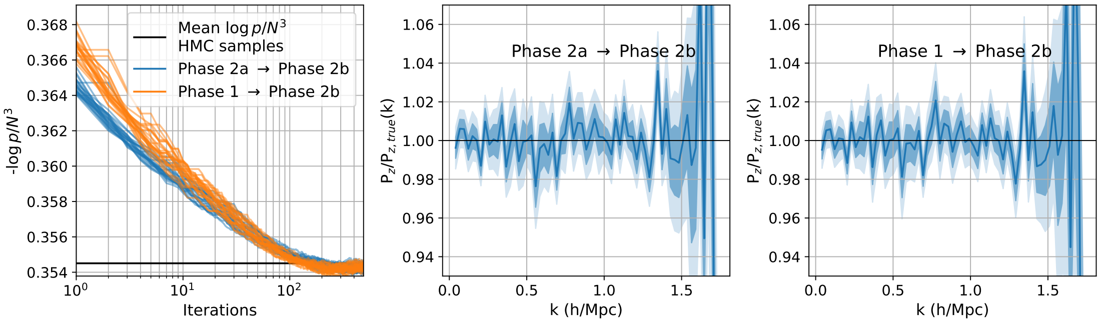


Figure 7. Phase IIB for our fiducial configuration of L = 200 Mpc/h and N = 64: (left) Evolution of log p for samples generated from the variational distribution fixed at the end of Phase I (orange) and Phase IIA (blue) as they are propagated with standard HMC. The trajectories reach the typical set, here defined by the range of log p values explored by HMC (shown in grey) after around 100 samples. (middle panel) Distribution of transfer function same as figure 6 but for the samples at the 100th iteration of Phase IIB when starting from the variational distribution at the end of Phase IIA (blue trajectories on left). (right panel) Same distribution of transfer function but when starting from the variational distribution at the end of Phase I.

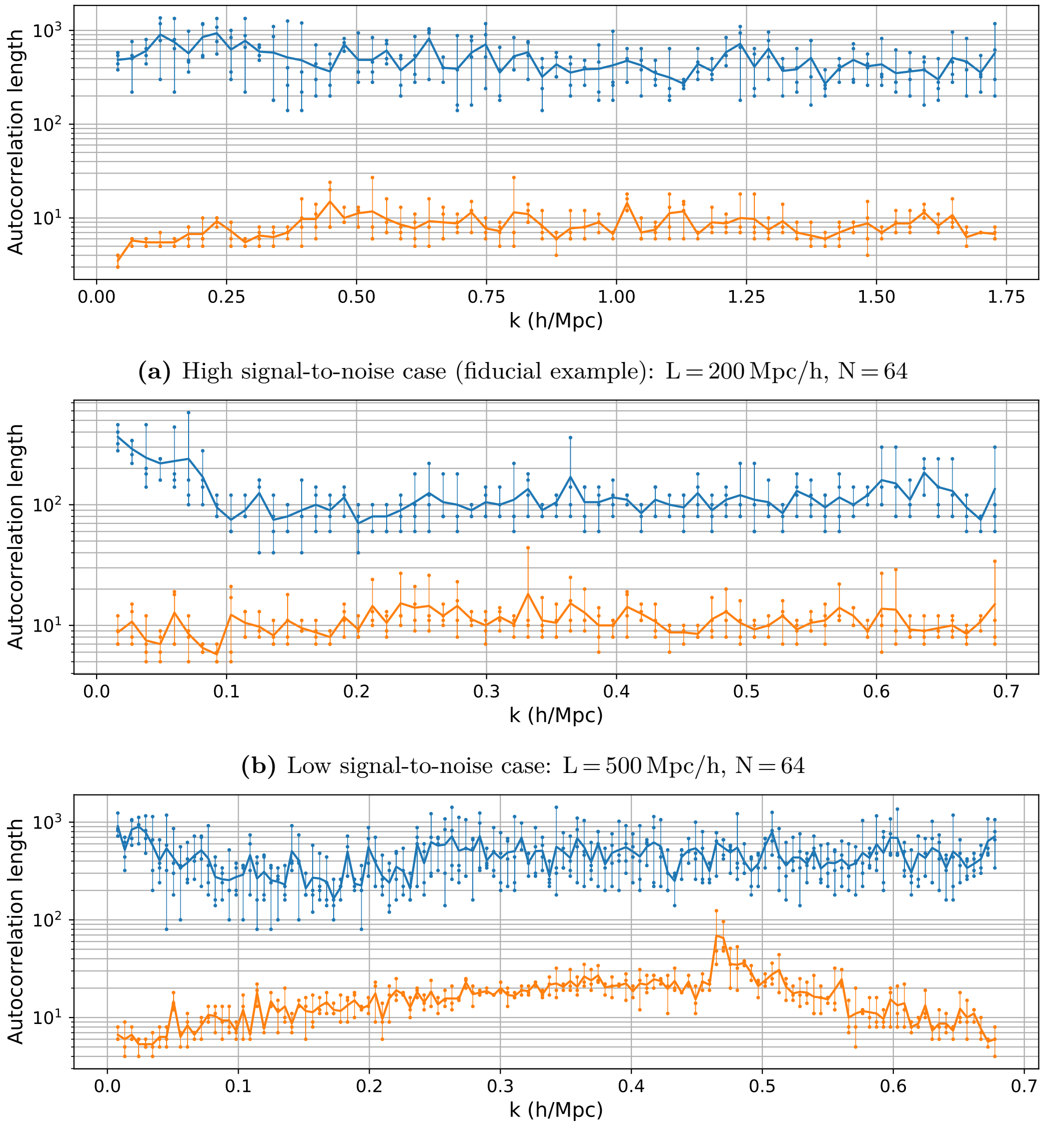
and the short HMC chains do improve the quality of our inference. We demonstrate this in figure 7 for our fiducial configuration with Phase IIB of our algorithm. Here we generate independent initial samples from the variational distribution fixed at the end of Phase I and Phase IIB, and propagate them with standard HMC. As shown in the left panel, these trajectories generate the samples of similar to the HMC samples in terms of target probability log p distribution after around 100 iterations. The other two panels show that these samples at the end of Phase IIB also have consistent distribution of the transfer function.

5.3 Cost of VBS

We have established that the samples generated by VBS at the end of Phase IIA have same distribution of the transfer function as HMC samples, and are of higher quality than VI. These can further be corrected in Phase IIB with longer chains of standard HMC. Next we compare the cost of VBS algorithm with HMC for problem. We begin by comparing the efficacy of hybrid sampling in Phase IIA with HMC sampling. This is shown in figure 8 in terms of the auto-correlation length. We note that this is not an ideal comparison since sampling in Phase IIA is non-Markovian and thus comparing the auto-correlation lengths can be misleading. However we still find this exercise instructive for two reasons- first, this is the most commonly used metric for HMC algorithm, and second, if the VBS samples had a systematic drift due to the continuous adaptation of the proposal kernel, then we expect the auto-correlation length to be of the same order as the length of the total chain.

In figure 8, we show the auto-correlation length of samples for following configurations of the box size (L) and the mesh (N): (L, N) = (200 Mpc/h, 64), (500 Mpc/h, 64), and (1000 Mpc/h, 128). The first two of these experiments have different shot noise levels that allow us to compare the effect of signal-to-noise ratio (SNR) in our data. The second and the third case have the same shot noise but allow us to see how well our approach scales to larger problems (N).

For N = 64, the auto-correlation length of HMC samples is order hundreds for high SNR case and order tens for low SNR case. This is consistent with the expectation that the posterior distribution is more complex in high signal regime and hence harder to sample.



(a) High signal-to-noise case (fiducial example): L = 200 Mpc/h, N = 64

(b) Low signal-to-noise case: L = 500 Mpc/h, N = 64

(c) Scaling up the low signal-to-noise case: L = 1000 Mpc/h, N = 128

Figure 8. Auto-correlation length for different experiments (panels) as measured with the power in different modes for HMC chains and hybrid samples. Different points along the same vertical (k-mode) are four different chains.

On the other hand, the auto-correlation length of hybrid samples is order tens in both the cases. Also note that while the average gains are 40x for low noise case and 10x for high noise case, they are larger on the largest scales which are the most correlated for HMC samples. N = 128 case, as expected, is more challenging than N = 64 case for both the algorithms. However VBS still gains a factor of at least 5–50x over HMC depending on the scale under consideration. The auto-correlation length in hybrid scheme also shows an interesting feature of increasing until the scale where the SNR$\sim$1 and then dropping again.

Next, we turn to Phase IIB of VBS which further improves the quality of VBS samples to match the HMC samples. As shown in the left panel of figure 7, we need HMC chains of around 100 samples long to propagate VBS samples into the 'typical set' where they have the same log p distribution as HMC samples. We find that though the starting distribution of log p is different for samples generated from the variational distribution at the end of Phase I and Phase IIA, HMC trajectories from both these reaches the typical set in about 100 iterations. This suggests that in the current problem, we can altogether skip Phase IIA. Then the cost to generate a single independent sample from the target distribution of the same quality as HMC is ∼100 samples. This is larger than the auto-correlation of VBS Phase IIA, but still 4–10x cheaper than HMC where the auto-correlation length for this configuration can be 400–1000 samples long, as shown in figure 8. However at the same time, we caution that skipping Phase IIA might not always be the best choice, and this should be informed instead by monitoring the quality of the variational samples with the continuous training in hybrid sampling phase.

6 Conclusions

Forward modeling approaches face the challenging task of doing inference in high dimensions where we need to marginalize over millions of latent parameters which are the phases of the Gaussian initial conditions. Hamiltonian Monte Carlo (HMC) approaches generate correct samples from the posterior distribution but are prohibitively expensive due to long auto-correlation lengths in the sample chains (figure 3). On the other hand, variational inference (VI) is fast but the backward KL loss can sometimes not be constraining enough to learn the correct target distribution (figure 4). In this work, we build upon recent works in statistics and machine learning literature to develop variational self-boosted sampling (VBS) — a hybrid scheme that combines HMC with VI to reap the benefits of both the approaches. Thus our approach can be seen as learning the proposal kernel for HMC or alternatively as a variational approximation to the target distribution with short chains serving to correct the learnt approximation.

We parameterize our variational distribution with a normalizing flow (NF) which has a learnable mean field Gaussian as the base distribution and transport layers of alternating Fourier convolutions and element-wise operations. We show that this architecture is flexible enough to learn a low-dimensional distribution of the HMC samples from the posterior which have correct cross-correlation and transfer function (figure 9), making it a promising proposal kernel.

We run experiments on a toy model with the dark matter density corrupted with Gaussian noise as the data. We run three configurations of the box size (L) and the mesh (N), with (L, N) = (200 Mpc/h, 64), (500 Mpc/h, 64), and (1000 Mpc/h, 128), to explore the effect of the noise in the data as well as the scaling of our normalizing flow and hybrid scheme. We verify that the transfer function and cross correlation of VBS samples follow a similar distribution to the HMC samples. We also show that the VBS samples are of a higher quality than the trained NF samples (trained with forward KL) thus demonstrating the benefits of short MCMC chains.

As compared HMC, we find that in the hybrid sampling phase, the auto-correlation length is reduced by a factor of ∼50 for the low noise case and factor of ∼10 for high noise case, primarily due to the longer auto-correlation lengths of HMC in the former. At the same time, both the schemes find it more challenging to $\mathrm{N} = 128$ but the hybrid scheme still sees gains of order tens over HMC (figure 8). However, the quality of VBS samples in the hybrid sampling phase can be worse than HMC samples, but this can be corrected by propagating independent samples from the variational distribution with longer HMC chains in the final sampling phase. For our fiducial configuration, the length of these chains is still ∼4–10x less than the auto-correlation length of HMC.

Finally, we note that while training the variational distribution in the hybrid sampling phase improves the quality of the proposals generated, it does not reduce the number of iterations required in the final sampling phase to propagate them to the target distribution. This suggests more work is required to understand the trade-offs between hybrid sampling phase and the final sampling phase. We will pursue this in the follow-up works when we will extend this framework to simultaneously infer the cosmology parameters and the phases of the initial conditions. In the current work, we have also used simple forward models such as ZA for computational ease, and dark matter density as the data observable but we expect our conclusions to qualitatively remain the same for more realistic gravity models as well as tracers such as halos and weak lensing maps.

Acknowledgments

CM would like to thank Francois Lanusse for the initial discussions at the conception of the project and help with VI code, Marylou Gabrie for sharing insights on hybrid sampling approach, Biwei Dai for sharing his TRENF code. CM would also like to thank Uros Seljak, Vanessa Boehm, Ben Wandelt, Shirley Ho, Kaze Wong, Wenda Zhou for useful discussions and Vanessa Boehm again for detailed comments on the manuscript. We would also like to thank the referee, Florent Leclercq, for the detailed review including the suggestion to explore the final sampling phase.

A Normalizing flow

Normalizing flows transform a simple base distribution q_B with a transport map T_θ consisting of a series of invertible, bijective mappings into more complex distributions of interest [18]. We use NF to parameterize our variational family such that it is flexible enough to capture the target distribution

$$q(\mathbf{z};\boldsymbol{\nu}) = q_B(T_\theta^{-1}(\mathbf{z});\boldsymbol{\nu}_B)|\det \nabla_{\mathbf{z}} T_\theta^{-1}| \tag{A.1}$$

where the parameters of the base distribution and the transport map compose our variational parameters $\boldsymbol{\nu} = \{\boldsymbol{\nu}_B, \theta\}$.

A.1 Base distribution

Traditionally when NF are used to learn generative models, the base distribution consists of a simple distribution with few or no trainable parameters, such as a standard normal. However in our case, the target is the posterior of a specific data realization and this breaks the symmetry of the target distribution. Hence for our base distribution, we choose the mean-field normal i.e. $q(\mathbf{z};\boldsymbol{\nu}_B) := \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are now the same shape and size as the phase field, i.e., N^3 grids. In our experiments, we find that while fixing $\boldsymbol{\Sigma} = 1$ does not affect our posterior accuracy significantly, however keeping $\boldsymbol{\mu}$ trainable is crucial for any meaningful inference.

A.2 Transport map

The transport map consists of a series of invertible transformations such that the log-determinant of their Jacobain can be estimated quickly. Hence NF typically use specialized coupling layers or autoregressive layers [31, 32]. However these NF scale poorly to three dimensional data and large (millions) parameter spaces.

We take an alternate approach for our transport map that was recently shown to accurately learn the high dimensional data likelihood of cosmological fields in [25]. Motivated by the fact that the cosmological fields are rotationally and translationally invariant, [25] propose constructing transport maps using Fourier-space convolutions.

A.2.1 Fourier space convolutions

A convolution in configuration space can be performed as a product with a transfer function $t(\mathbf{k})$ in the Fourier space. This transfer function can be element-wise and hence of the same dimensionality N^3 as the parameters. However for rotational and translational invariant fields, the transfer function becomes only a function of scales, $t(k)$, which can be parameterized by a few tens of parameters. Moreover since the transformation consists of simply multiplying be a scalar function, the Jacobian is straightforward to estimate.

Thus the overall transformation for a configuration space field $\mathbf{x}$ is

$$\mathbf{x}' = \mathcal{F}^{-1}(t_\theta(k)\mathcal{F}(\mathbf{x})) \tag{A.2}$$

where θ are the learnable (variational) parameters and $\mathcal{F}$ is the Fourier transform operation. The transfer function can be any interpolation function and we model it as a Cubic Hermite polynomial. Then, the knots values and slopes at knot positions constitute θ.

A.2.2 Element-wise transformations

We alternate the Fourier space convolutions with learnable element-wise transformations Ψ_ϕ in the configuration space.

The simplest such transformations are affine (scale and shift) transformations

$$\mathbf{x}' = \alpha\mathbf{x} + \beta \tag{A.3}$$

with $\phi = \{\alpha, \beta\}$ as the scale and shift variational parameters. We consider two cases- i) global affine transformations wherein α and β are scalars and the entire field is shifted and scaled uniformly, or ii) mean-field affine transformations wherein $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are now N^3 grids, same as the parameters $\mathbf{x}$. While ii) increases the number of parameters of our NF by a lot, it allows us break the constraint of rotational and translational invariances in our transport map that is made by using Fourier space convolutions. We find that for $\mathrm{N} = 64$, global affine transformations sufficed but for $\mathrm{N} = 128$ using mean-field affine transformations markedly improved the quality of inference.

Affine transformations are linear but the element-wise transformations can also be made non-linear. For instance, [25] used monotonic rational-quadratic splines as non-linear transformations. However in our experiments, using splines instead of linear transformations did not seem to significantly affect the quality of posteriors for our toy model and hence we did not use them for the current experiments.

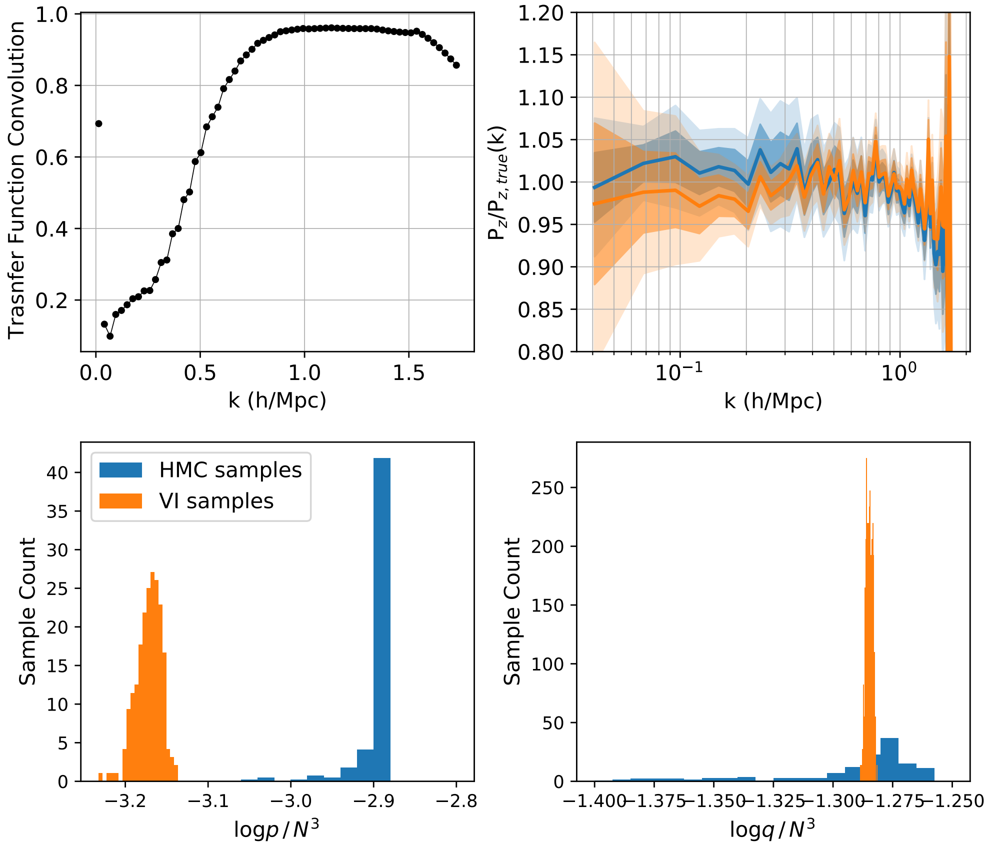


Figure 9. Learning of normalizing flow: (top left) The transfer function learnt by the Fourier convolution layer of NF. (top right) Distribution of the transfer function for samples from HMC on which NF is trained (blue) and the samples generated by the normalizing flow (orange). (bottom left) Distribution of unnormalized $\log p$, the true posterior probability of HMC samples (blue) and samples from the trained NF (orange). (bottom right) Distribution of $\log q$, the variational posterior probability as estimated by NF of HMC samples (blue) and samples from the trained NF (orange).

A.3 Learnt distribution

Every layer of our NF consists of a Fourier space convolution followed by an element-wise operation to construct a unit transformation $f = \mathbf{x}_0 \rightarrow \mathbf{x}_1$:

$$\mathbf{x}_1 = \Psi_\phi(\mathcal{F}^{-1}(t_{\theta_1}(k)\mathcal{F}(\mathbf{x}_0))) \tag{A.4}$$

These layers can be stacked and are combined with the base distribution to parameterize our target distribution.

It is crucial for variational inference to ensure that the parametric family is flexible enough to capture the target distribution. To verify this, we train our NF on the independent samples generated from HMC. and show different metrics to gauge its performance in figure 9. We show the results for L = 200 Mpc/h box and N = 64 grid. The first panel shows the transfer spectra learnt by the Fourier convolution layer and it seems that the convolution layer acts as a high pass filter. The second panel shows the distribution of the power spectra of HMC samples and samples generated from the trained normalizing flow. The NF samples have higher variance than HMC samples but are still consistent with unity on all scales. Note that this is similar to figure 6 except that the NF there is trained on VBS samples while in this case it is trained on HMC samples. While not shown here, we have verified that the distribution of the cross-correlation of the NF samples with the true initial conditions is similar to that of HMC samples. This, combined with the visual consistency of all samples, can lead one to conclude that NF learns the high dimensional target distribution.

In the bottom panel of figure 9, we examine the learnt NF in greater detail. The two panels show the histogram of the unnormalized logarithm of probability under the target distribution (posterior, $\log p$) and the learnt variational or NF distribution ($\log q$) for HMC and NF samples.[3] It's immediately apparent that the two distributions are different and in more ways than can be explained by simple normalization. This, coupled with the fact that the posterior distribution of the transfer function and cross-correlations for the two sample sets is somewhat consistent leads us to conclude that while the NF does not learn the full high dimensional target distribution, it does correctly learn a lower-dimensional manifold. Hence instead of using NF to learn the posterior of the phase fields completely, we instead use the variational distribution as a proposal kernel in the hybrid scheme where short HMC chains propagate the generated samples to the target distribution.

References

[1] DESI collaboration, *The DESI Experiment Part I: Science, Targeting, and Survey Design*, arXiv:1611.00036 [INSPIRE].

[2] LSST Science, LSST Project collaboration, *LSST Science Book, Version 2.0*, arXiv:0912.0201 [INSPIRE].

[3] L. Amendola et al., *Cosmology and fundamental physics with the Euclid satellite*, Living Rev. Rel. **21** (2018) 2 [arXiv:1606.00180] [INSPIRE].

[4] U. Seljak, G. Aslanyan, Y. Feng and C. Modi, *Towards optimal extraction of cosmological information from nonlinear data*, JCAP **12** (2017) 009 [arXiv:1706.06645] [INSPIRE].

[5] F. Leclercq, W. Enzi, J. Jasche and A. Heavens, *Primordial power spectrum and cosmology from black-box galaxy surveys*, Mon. Not. Roy. Astron. Soc. **490** (2019) 4237 [arXiv:1902.10149] [INSPIRE].

[6] C. Modi, F. Lanusse and U. Seljak, *FlowPM: Distributed TensorFlow implementation of the FastPM cosmological N-body solver*, Astron. Comput. **37** (2021) 100505 [arXiv:2010.11847] [INSPIRE].

[7] V. Böhm, Y. Feng, M.E. Lee and B. Dai, *MADLens, a python package for fast and differentiable non-Gaussian lensing simulations*, Astron. Comput. **36** (2021) 100490 [arXiv:2012.07266] [INSPIRE].

[8] J. Jasche and G. Lavaux, *Physical Bayesian modelling of the non-linear matter distribution: new insights into the Nearby Universe*, Astron. Astrophys. **625** (2019) A64 [arXiv:1806.11117] [INSPIRE].

[9] H. Wang, H.J. Mo, X. Yang, Y.P. Jing and W.P. Lin, *ELUCID — Exploring the Local Universe with reConstructed Initial Density field I: Hamiltonian Markov Chain Monte Carlo Method with Particle Mesh Dynamics*, Astrophys. J. **794** (2014) 94 [arXiv:1407.3451] [INSPIRE].

[10] C. Modi, Y. Feng and U. Seljak, *Cosmological Reconstruction From Galaxy Light: Neural Network Based Light-Matter Connection*, JCAP **10** (2018) 028 [arXiv:1805.02247] [INSPIRE].

[3] Note that since these probabilities are not normalized by the evidence, these should not be compared directly across panels.

[11] C. Modi, M. White, A. Slosar and E. Castorina, *Reconstructing large-scale structure with neutral hydrogen surveys*, *JCAP* **11** (2019) 023 [arXiv:1907.02330] [INSPIRE].

[12] C. Modi, F. Lanusse, U. Seljak, D.N. Spergel and L. Perreault-Levasseur, *CosmicRIM: Reconstructing Early Universe by Combining Differentiable Simulations with Recurrent Inference Machines*, arXiv:2104.12864 [INSPIRE].

[13] B. Horowitz, U. Seljak and G. Aslanyan, *Efficient Optimal Reconstruction of Linear Fields and Band-powers from Cosmological Data*, *JCAP* **10** (2019) 035 [arXiv:1810.00503] [INSPIRE].

[14] M. Millea and U. Seljak, *Marginal unbiased score expansion and application to CMB lensing*, *Phys. Rev. D* **105** (2022) 103531 [arXiv:2112.09354] [INSPIRE].

[15] S. Duane, A.D. Kennedy, B.J. Pendleton and D. Roweth, *Hybrid Monte Carlo*, *Phys. Lett. B* **195** (1987) 216 [INSPIRE].

[16] R.M. Neal, *Mcmc using hamiltonian dynamics*, in *Handbook of Markov chain Monte Carlo*, Chapman & Hall/CRC (2011), chapter 5, pg. 113.

[17] M. Betancourt, *A Conceptual Introduction to Hamiltonian Monte Carlo*, arXiv:1701.02434.

[18] I. Kobyzev, S. J. Prince and M. A. Brubaker, *Normalizing flows: An introduction and review of current methods*, *IEEE Trans.Pattern Anal. Mach. Iintell.* **43** (2020) 3964.

[19] M. Hoffman, P. Sountsov, J.V. Dillon, I. Langmore, D. Tran and S. Vasudevan, *NeuTra-lizing Bad Geometry in Hamiltonian Monte Carlo Using Neural Transport*, arXiv:1903.03704.

[20] C. Naesseth, F. Lindsten and D. Blei, *Markovian score climbing: Variational inference with $KL(p||q)$*, *Adv. NIPS* **33** (2020) 15499 [arXiv:2003.10374].

[21] M. Gabrié, G.M. Rotskoff and E. Vanden-Eijnden, *Efficient bayesian sampling using normalizing flows to assist markov chain monte carlo methods*, arXiv:2107.08001.

[22] D.M. Blei, A. Kucukelbir and J.D. McAuliffe, *Variational Inference: A Review for Statisticians*, *J. Am. Statist. Assoc.* **112** (2017) 859 [arXiv:1601.00670].

[23] P. Frank, R. Leike and T.A. Enßlin, *Geometric variational inference*, *Entropy* **23** (2021) 853.

[24] S. Kullback and R.A. Leibler, *On Information and Sufficiency*, *Ann. Math. Stat.* **22** (1951) 79 [INSPIRE].

[25] B. Dai and U. Seljak, *Translation and rotation equivariant normalizing flow (TRENF) for optimal cosmological analysis*, *Mon. Not. Roy. Astron. Soc.* **516** (2022) 2363 [arXiv:2202.05282] [INSPIRE].

[26] Y. Feng, M.-Y. Chu, U. Seljak and P. McDonald, *FastPM: a new scheme for fast simulations of dark matter and haloes*, *Mon. Not. Roy. Astron. Soc.* **463** (2016) 2273 [arXiv:1603.00476] [INSPIRE].

[27] S. Tassev, M. Zaldarriaga and D. Eisenstein, *Solving Large Scale Structure in Ten Easy Steps with COLA*, *JCAP* **06** (2013) 036 [arXiv:1301.0322] [INSPIRE].

[28] M.D. Hoffman and A. Gelman, *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo*, arXiv:1111.4246.

[29] J. Bornschein and Y. Bengio, *Reweighted wake-sleep*, arXiv:1406.2751.

[30] A. Vehtari, A. Gelman, D. Simpson, B. Carpenter and P.-C. Bürkner, *Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC*, arXiv:1903.08008.

[31] L. Dinh, J. Sohl-Dickstein and S. Bengio, *Density estimation using Real NVP*, arXiv:1605.08803.

[32] G. Papamakarios, T. Pavlakou and I. Murray, *Masked Autoregressive Flow for Density Estimation*, *Adv. NIPS* **30** (2017) 2338 [arXiv:1705.07057].