Reconstructing the Universe with Variational self-Boosted Sampling

Chirag Modi [1 2] **Yin Li** [1 2] **David Blei** [2 3]

Abstract

Forward modeling approaches in cosmology seek to reconstruct the initial conditions at the beginning of the Universe from the observed survey data. However the high dimensionality of the parameter space poses a challenge to explore the full posterior with traditional algorithms such as Hamiltonian Monte Carlo (HMC) and variational inference (VI). Here we develop a hybrid scheme called variational self-boosted sampling (VBS) that learns a variational approximation for the proposal distribution of HMC with samples generated on the fly, and in turn generates independent samples as proposals for MCMC chain to reduce their auto-correlation length. We use a normalizing flow with Fourier space convolutions as our variational distribution to scale to high dimensions of interest. We show that after a short initial warm-up and training phase, VBS generates better quality of samples than simple VI and reduces the correlation length in the sampling phase by a factor of 10-50 over using only HMC.

1. Introduction

Forward modeling approaches for cosmological analysis seek to infer cosmological parameters by doing a full field based analysis wherein we compare our simulation predictions with the observed data such as galaxies at the level of the individual objects. Since these approaches do not rely on any compressed summary statistics of the data, they in principle maximize the amount of information that can be extracted from cosmological surveys. The challenge however is that to simulate the survey data at field level, we need to know both, the cosmological parameters and the phases of the initial conditions[1] at the beginning of the Universe (i.e. the initial distribution of matter field in the Universe). Since both of these are unknown, we now need to infer them both from the data simultaneously. Recent works have taken a Bayesian approach to this inference (Jasche & Wandelt, 2013; Seljak et al., 2017; Modi et al., 2021). In this, we combine the prior on the phases ($\mathbf{z}$) and the cosmological parameters ($\mathbf{\Lambda}$) with the likelihood model of the data ($\mathbf{y}_0$) to write a posterior for the parameters

$$p(\mathbf{z}, \mathbf{\Lambda}|\mathbf{y}_0) \propto p(\mathbf{y}_0|\mathbf{z}, \mathbf{\Lambda})p(\mathbf{z})p(\mathbf{\Lambda}) \tag{1}$$

where we are forced to drop the evidence term $p(\mathbf{y}_0)$ which cannot be evaluated. This inference is challenging primarily for two reasons- the high dimensionality of the initial phases which can be in millions, and the expensive cosmological forward models required to evaluate the likelihood term.

Cosmologists use differentiable forward models (Modi et al., 2020; Böhm et al., 2021) to access gradient based algorithms and tackle this challenge. The simplest inference is to reconstruct a maximum-a-posterior (MAP) estimate (Modi et al., 2018; 2019) but this provides only a point estimate. A more robust approach to infer full posterior is to use Hamiltonian Monte Carlo (HMC) (Neal et al., 2011; Wang et al., 2014; Kitaura et al., 2014). However the successive samples generated by HMC are correlated and in high dimensions these correlation lengths can be hundreds of samples long. Hence HMC can be prohibitively expensive for scaling up to the future cosmological surveys.

In this work, we propose a hybrid approach to inference by learning a proposal distribution and combining it with HMC. We call it variational boosted sampling (VBS). The goal is for the proposal distribution to generate independent samples that either lie in the target distribution directly, or can propagate to the same with short Markov chains. We parameterize this proposal distribution as a normalizing flow (NF) (Kobyzev et al., 2020) which is trained on the fly using samples from the MCMC chain itself. To scale to high dimensionality of the the initial conditions, our NF uses Fourier based convolution that exploit rotational and trans-

[1] Center for Computational Astrophysics, Flatiron Institute, New York [2] Center for Computational Mathematics, Flatiron Institute, New York [3] Columbia University, New York. Correspondence to: Chirag Modi <cmodi@flatironinstitute.org>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

[1] The initial conditions are predicted to be a zero-mean Gaussian random field. Hence they can be reparameterized as an amplitude (variance), which is only a function of cosmological parameters, and a stochastic component at every point which we refer to as phase of the initial conditions (this corresponds to the particular realization draw of our Universe).

lational symmetries of cosmological fields (Dai & Seljak, 2022), see appendix for details. Similar approaches have been proposed to speed up HMC by improving the geometry of the posterior distribution with a transport map (Hoffman et al., 2019; Naesseth et al., 2020), and to learn a global sampler that assists local chains (Gabrié et al., 2021).

In the landscape of variational inference (VI) wherein one learns a parametric form for the target distribution(Blei et al., 2017), our learnt proposal distribution can also be viewed as a variational approximation to the target distribution. The short Markov chain serve a dual purpose- they correct the samples generated from the learnt approximation, while also generating samples from the true target distribution to train the variational distribution using more powerful forward (inclusive) Kullback-Leibler (KL) divergence.

We begin by setting up our cosmological inference problem formally (section 2), present our hybrid sampling scheme (section 3) and compare its performance with HMC and VI (section 4). We conclude in section 5.

2. Setup

Our data ($\mathbf{y}_0$) is the dark matter density field on a cubic N^3 grid[2] where N is the number of grid points or pixels along each side of the cube. The mock data is 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 parameters to be inferred are the phases of the this Gaussian initial density field ($\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$). We keep the cosmology parameters ($\mathbf{\Lambda}$) fixed to their true value to focus on the challenging high-dimensional part of the problem. The forward model is the particle displacement predicted by the first order Lagrangian Perturbation theory (Zeldovich Approximation, ZA). We take our data noise to be Gaussian with known variance ($\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$) corresponding to the shot-noise of the dark matter particles. Hence we can compute the exact likelihood for our data and the posterior distribution

$$\pi(\mathbf{y}_0|\mathbf{z}) = \mathcal{N}(\mathbf{y} = f(\mathbf{z}), \boldsymbol{\sigma}) \qquad \text{(Gaussian likelihood)}$$
$$\pi(\mathbf{z}|\mathbf{y}_0) \propto \pi(\mathbf{y}_0|\mathbf{z})\pi(\mathbf{z}) \qquad \text{(Posterior)}$$

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

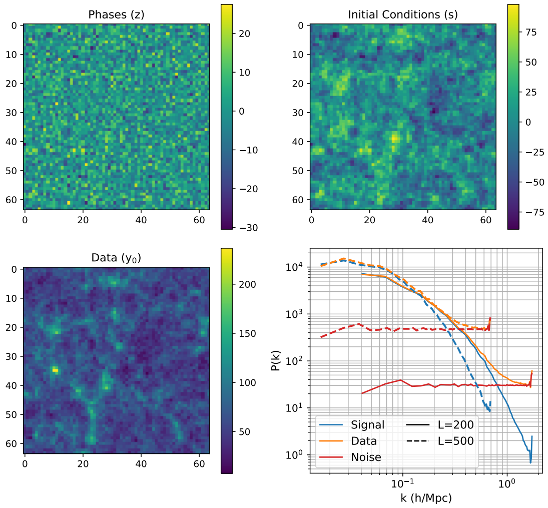

Figure 1. An example of the forward model for N=64: first three panel show component fields projected (summed) along the z-axis. The last panel shows the corresponding data and noise power spectrum for two different signal-to-noise ratios (box sizes- L=200 and 500 Mpc/h).

3. Variational self-Boosted Sampling (VBS)

In this section, we propose our hybrid scheme that combines VI with HMC. HMC draws correct samples from the target distribution but can be computationally expensive due to generating correlated samples. VI on the other hand aims to learn the target distribution by assuming it belongs to a parametric family, $q(\boldsymbol{\nu})$ and these parameters $\boldsymbol{\nu}$ are then estimated by minimizing a divergence between the variational and the target distribution. In variational self-boosted sampling (VBS), we use the samples ($\mathbf{z}_i$) 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 independent proposals from $q(\mathbf{z}; \boldsymbol{\nu})$ in MCMC chain (Gabrié et al., 2021). As the variational approximation improves over iterations, it will also become a good proposal kernel and its samples will be readily accepted, thus reducing the auto-correlation length of HMC chains.

3.1. Algorithm

Starting from a sample point in the target distribution, our algorithm can broadly be divided into two phases[3]- i) learning phase and ii) sampling phase. The full algorithm is presented in Algorithm 1 but briefly, the two phases are:

- **Phase I, Learning Phase**:

[2] 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

[3] We have assumed that we have access to a sample from the target distribution to initialize from. If this is not the case, there is a burn-in phase to initialize from a random point and reach such a sample. However since this is identical to HMC, we do not include it explicitly as a part of the algorithm.

Algorithm 1 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 after training $N_2$; probability of generating proposal from VI distribution $p_{\text{jump}}$

set i=0                                              {Phase 1, Learning}
  for $i = 0$ to $N_1$ do
    $\mathbf{z}_{i+1} \leftarrow \text{HMC step}(\mathbf{z}_i, \pi, \epsilon, L, H, M)$
    Sample $\mathcal{B} = \{\mathbf{z}_{(1)}...\mathbf{z}_{(B)}\}$ uniformly from $\{\mathbf{z}_1...\mathbf{z}_i\}$
    $\mathcal{L} = -\sum_{\mathcal{B}} \log q(\mathbf{z}_{(i)};\boldsymbol{\nu})$
    $\boldsymbol{\nu} \leftarrow \boldsymbol{\nu} - \epsilon\nabla_{\boldsymbol{\nu}}\mathcal{L}$                 {Optimization}
  end for
                                                     {Phase 2, Sampling}
  for $i = N_1$ to $N_2$ do
    if $\text{Uniform}(0,1) \geq p_{\text{jump}}$ 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 $\mathbf{z}_{i+1} \leftarrow \mathbf{z}_i$
    end if
    Sample $\mathcal{B} = \{\mathbf{z}_{(1)}...\mathbf{z}_{(B)}\}$ uniformly from $\{\mathbf{z}_1...\mathbf{z}_i\}$
    $\mathcal{L} = -\sum_{\mathcal{B}} \log q(\mathbf{z}_{(i)};\boldsymbol{\nu})$
    $\boldsymbol{\nu} \leftarrow \boldsymbol{\nu} - \epsilon_q\nabla_{\boldsymbol{\nu}}\mathcal{L}$                {Optimization}
  end for
output : $\{\mathbf{z}_1...\mathbf{z}_{M+N}\}$, $q(\mathbf{z};\boldsymbol{\nu}^*)$
```

In this phase we only run vanilla HMC chains to generate samples ($\mathbf{z}_i$) from the true posterior. We simultaneously use these samples to learn the variational distribution by maximizing the log-probability of these samples: $\boldsymbol{\nu}^* = \text{argmax}_\nu \sum_{\mathbf{z}_i \sim \pi(\mathbf{z}|\mathbf{y}_0)} \log q(\mathbf{z};\boldsymbol{\nu})$. We do not thin the chain i.e. we use all the samples which are correlated. This phase lasts until the variational approximation learns the distribution of the current samples. Until now, the computational cost of this phase is practically the same as HMC.

- **Phase II, Hybrid Sampling**:
  In this phase we alternate between (with some pre-chosen probability, p_{jump}) making proposals from HMC kernel and the variational distribution. 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.

Note that since we continuously adapt our variational distribution, our approach is not strictly Markovian. However if the adaptation decreases with iterations, our approach also becomes Markovian asymptotically. Then for our algorithm to enjoy asymptotic correctness of MCMC algorithms, we need a detailed balance (DB) condition for the acceptance of proposals[4]. For HMC proposal step of VBS, it is the same as DB for HMC. For variational proposals, let $\mathbf{z}_1$ be the current sample and $\mathbf{z}_2$ be the proposal made from $q(\mathbf{z},\boldsymbol{\nu})$. Then DB is met if the acceptance probability α of making the transition $\mathbf{z}_1 \rightarrow \mathbf{z}_2$ is

$$\alpha = \min\left(1, \frac{\pi^*(\mathbf{z}_2)q(\mathbf{z}_1;\boldsymbol{\nu})}{\pi^*(\mathbf{z}_1)q(\mathbf{z}_2;\boldsymbol{\nu})}\right) \qquad (2)$$

This is the balance condition to correctly sample from the distribution $\pi^*(\mathbf{z})$. Ideally this should be the true posterior distribution of interest, $\pi(\mathbf{z})$. However we find that using this can lead to high variance in the acceptance probability for our high dimensional posterior distribution. To reduce this variance, we re-scale the target posterior probability with the number of grid points $\pi^*(\mathbf{z}) = \pi(\mathbf{z})^{1/N^3}$. 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*. Hence 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 from the variational distribution.

4. Results

In this section, we compare our proposed VBS scheme with HMC and VI, with HMC samples serving as a benchmark due to their guaranteed correctness. For our experiments, we consider cosmological simulations with configurations of box size (L) and the mesh (N) as: (L, N) = (200 Mpc/h, 64), (500 Mpc/h, 64), and (1000 Mpc/h, 128). The first two of these have different noise levels to compare the effect of signal-to-noise ratio (SNR) in our data, while the third allows us to see how well VBS scales to larger problems (N). For both, VBS and HMC scheme, we run 4 chains for robustness. We use the same stepsize ϵ for both and it is fit by dual averaging scheme (Hoffman & Gelman, 2011). The number of leapfrog steps L is chosen uniformly between 25 and 50 for every proposal in HMC and HMC steps in VBS.

Based on a few 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 found

[4] We assume here that the learnt variational distribution is ergodic, which is the second required condition

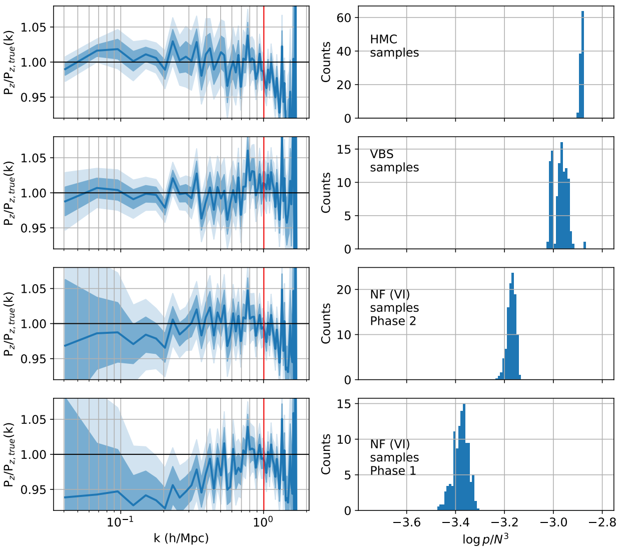

Figure 2. Comparing the posterior for different approaches: (left) We show the mean (solid lines), one- and two-standard deviations (shaded regions) of transfer function for samples of the phase field ($\mathbf{z}$) from different approaches for L=200 Mpc/h and N=64 simulations. The red vertical line is the nyquist frequency. (right) We show the distribution of unnormalized log posterior probabilities, $\log\, p$, for the samples generated by different approaches. (first row) Vanilla HMC samples that act as benchmark, (second row) VBS samples, (third row) Samples from the variational distribution (NF) at the end of the second phase and (fourth row) at the end of the first phase respectively.

that the qualitative performance of our scheme was quite robust to N_1 and p_{jump} within reasonable limits, hence not requiring much fine tuning. However we note that the actual quantitative gains can vary as decreasing p_{jump} decreases the frequency of NF proposals which break auto-correlation. Similarly we did not fine-tune our neural network architecture for normalizing flows. We used a single layer of Fourier convolutions with global and mean-field affine transformations (see Appendix C) for N=64 and N=128 experiments respectively. Optimizing our architecture further can likely lead to quantitative improvements in our results.

We begin by verifying the posterior sampled by VBS. Since it is hard to quantitatively compare the distribution of high dimensional distributions, we focus on low dimensional summary statistics- in this case the *transfer function* which is the ratio of the power spectrum of samples from the posterior with the power spectrum of the true initial conditions. We expect the mean of this ratio to be unity on all scales for samples from the true posterior. Figure 2 shows this distribution for samples from VBS, HMC as well as samples generated from the normalizing flow (VI approximation)[5] at the end of phase I and II of VBS. The distribution of VBS samples is consistent with HMC upto the nyquist frequency (red vertical) while NF samples have much larger scatter.

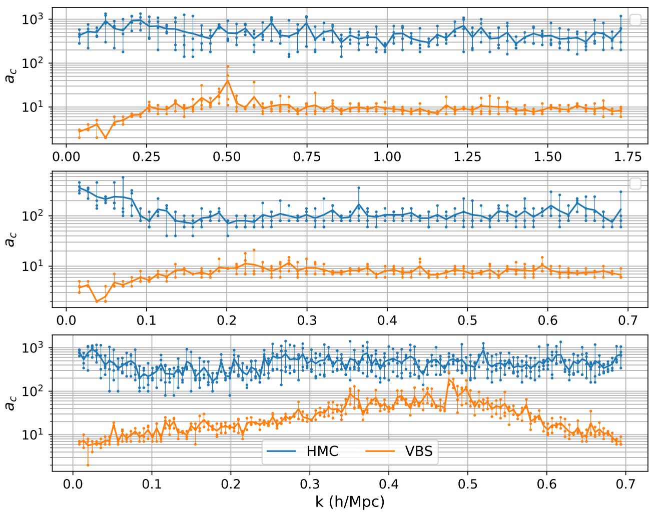

Figure 3. Auto-correlation length for power in different modes in HMC and VBS samples. Different points along the same vertical (k-mode) are four different chains. Different panels show different experiments with (L, N)=(200, 64), (500, 64), and (1000, 128).

We also compared the cross-correlation coefficient r_c of these samples with the true initial conditions conditions. However we do not show it here since the distribution for HMC and VBS samples was indistinguishable from each other with both having r_c unity on the signal dominated large scales and zero on the noise dominated small scales.

In the right panel of Fig. 2, we show distribution of the unnormalized $\log\, p$ values, i.e. the true posterior probabilities of these samples. Samples generated from both the NF (VI) are of poor quality but the short HMC chains do markedly improve the quality of these samples and hence VBS samples much closer to the HMC samples. 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.

Having established that VBS explores posterior correctly and generates higher quality samples than VI, we next compare VBS with HMC in terms of their efficacy. We do so by estimating the auto-correlation length (a_c) for the power in different modes/scales (k) in the power spectrum of the posterior samples in each chain. These are shown in Figure 3 for different configurations. For N=64, a_c of HMC samples is $\mathcal{O}(1000)$ for high SNR case and $\mathcal{O}(100)$ 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. On the other hand, the auto-correlation length of hybrid samples is $\mathcal{O}(10-100)$ in both the cases. N=128 case is more challenging for both

[5] These NF correspond to VI with forward KL loss. VI with backward KL performed significantly worse and hence not shown.

the algorithms, however VBS still gains a factor of at least 5-50x over HMC across scales. The auto-correlation length for VBS in this case also shows an interesting feature of increasing until the scale where the SNR$\sim$1 and then dropping again. In the future work, we will investigate how this affects the inference of cosmological parameters.

5. Conclusions

Forward modeling approaches face the challenging task of doing inference in high dimensions. In this work, we have proposed a hybrid scheme called variational self-boosted sampling (VBS) that combines VI and HMC to reap the benefits of both. Our approach can be seen as learning the proposal kernel for HMC on the fly, or alternatively as a variational approximation to the target distribution with short chains to correct the learnt approximation. We show that for different configurations of box size and mesh, corresponding to different SNRs in the data and different scales of the problem, VBS reduces the auto-correlation length of samples by a factor of 5-50x over HMC while generating higher quality samples than VI.

References

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Vanessa Böhm, Yu Feng, Max E Lee, and Biwei Dai. Madlens, a python package for fast and differentiable non-gaussian lensing simulations. *Astronomy and Computing*, 36:100490, 2021.

Jörg Bornschein and Yoshua Bengio. Reweighted wake-sleep. *arXiv preprint arXiv:1406.2751*, 2014.

Biwei Dai and Uros Seljak. Translation and rotation equivariant normalizing flow (trenf) for optimal cosmological analysis. *arXiv preprint arXiv:2202.05282*, 2022.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Marylou Gabrié, Grant M Rotskoff, and Eric Vanden-Eijnden. Efficient bayesian sampling using normalizing flows to assist markov chain monte carlo methods. *arXiv preprint arXiv:2107.08001*, 2021.

Matthew Hoffman, Pavel Sountsov, Joshua V Dillon, Ian Langmore, Dustin Tran, and Srinivas Vasudevan. Neutralizing bad geometry in hamiltonian monte carlo using neural transport. *arXiv preprint arXiv:1903.03704*, 2019.

Matthew D. Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *arXiv e-prints*, art. arXiv:1111.4246, November 2011.

Jens Jasche and Benjamin D. Wandelt. Bayesian physical reconstruction of initial conditions from large-scale structure surveys. *Monthly Notices of the Royal Astronomical Society*, 432(2):894–913, June 2013. doi: 10.1093/mnras/stt449.

F. S. Kitaura, G. Yepes, and F. Prada. Modelling baryon acoustic oscillations with perturbation theory and stochastic halo biasing. *Monthly Notices of the Royal Astronomical Society*, 439:L21–L25, March 2014. doi: 10.1093/mnrasl/slt172.

Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.

C. Modi, Y. Feng, and U. Seljak. Cosmological reconstruction from galaxy light: neural network based light-matter connection. *Journal of Cosmology and Astro-Particle Physics*, 10:028, October 2018. doi: 10.1088/1475-7516/2018/10/028.

C. Modi, M. White, A. Slosar, and E. Castorina. Reconstructing large-scale structure with neutral hydrogen surveys. *arXiv e-prints*, art. arXiv:1907.02330, Jul 2019.

Chirag Modi, Francois Lanusse, and Uros Seljak. FlowPM: Distributed TensorFlow Implementation of the FastPM Cosmological N-body Solver. *arXiv e-prints*, art. arXiv:2010.11847, October 2020.

Chirag Modi, François Lanusse, Uroš Seljak, David N Spergel, and Laurence Perreault-Levasseur. Cosmicrim: Reconstructing early universe by combining differentiable simulations with recurrent inference machines. *arXiv preprint arXiv:2104.12864*, 2021.

Christian Naesseth, Fredrik Lindsten, and David Blei. Markovian score climbing: Variational inference with kl (p—— q). *Advances in Neural Information Processing Systems*, 33:15499–15510, 2020.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

U. Seljak, G. Aslanyan, Y. Feng, and C. Modi. Towards optimal extraction of cosmological information from nonlinear data. *Journal of Cosmology and Astro-Particle Physics*, 2017(12):009, Dec 2017. doi: 10.1088/1475-7516/2017/12/009.

Huiyuan Wang, H. J. Mo, Xiaohu Yang, Y. P. Jing, and W. P. Lin. ELUCID—Exploring the Local Universe with the Reconstructed Initial Density Field. I. Hamiltonian Markov Chain Monte Carlo Method with Particle Mesh Dynamics. ApJ, 794(1):94, Oct 2014. doi: 10.1088/0004-637X/794/1/94.

A. Inference algorithms

In this appendix, we briefly review the two most widely used approaches for posterior inference, which also form the building blocks of our hybrid sampling algorithm. 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.

A.1. Hamiltonian Monte Carlo (HMC)

Algorithm 2 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 $M$
$q_0 \leftarrow z_0$ {Assign current sample as the initial position}
$p_0 \sim \mathcal{N}(0, 1)$ {Sample random momentum of same shape as $q_0$}
$i = 0$
for $i = 0$ to $L$ do
    {Integrate Hamiltonian equations for $L$ leapfrog steps}
    $q_{i+1},\ p_{i+1} \leftarrow$ LEAPFROG$(q_i,\ p_i,\ \pi, \epsilon)$
    $i \leftarrow i + 1$
end for
$H_0 \leftarrow \mathrm{H}(q_0,\ p_0, \pi, M)$ {Estimate Hamiltonian Eq. 3}
$H_L \leftarrow \mathrm{H}(q_L,\ p_L, \pi, M)$
$\alpha \leftarrow \exp(H_0 - H_L)$ {Maintain DB Eq. 4}
if $\mathrm{Uniform}(0, 1) \geq \alpha$ then
    $z_1 \leftarrow q_0$
else
    $z_1 \leftarrow q_L$
end if
output : $z_1$
```

HMC (Neal et al., 2011) is the most 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, 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. Generally the mass matrix is taken to be the indentity matrix, $M = I$. With these, one can construct the Hamiltonian $H : \mathcal{R}^{2d} \rightarrow \mathcal{R}$ as the total energy function for the state $\mathbf{x} := (\mathbf{q}, \mathbf{p})$,

$$H(\mathbf{x}) = H(\mathbf{q}, \mathbf{p}) = U(\mathbf{q}) + \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p}\ . \tag{3}$$

To simulate a Markov chain and generate samples from the target distribution, this physical system is evolved with respect to time by following Hamiltonian dynamics. The Hamiltonian's equations are numerically evolved by integrating the ODE system using leapfrog integrator. 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 condition which guarantees that the samples are generated from the correct distribution. As per detailed balance, the probability of accepting a proposal $\mathbf{x}_0 \rightarrow \mathbf{x}_1$ is

$$\alpha = \min(1, \exp(H(\mathbf{x}_0) - H(\mathbf{x}_1)) \tag{4}$$

The complete algorithm for generating proposals is described in Algorithm 2.

A.2. Variational Inference

Variational inference (Blei et al., 2017) takes a different approach from sampling and instead aims to directly learn the distribution of interest. It assumes that the target distribution π belongs to a parametric family, $q(\boldsymbol{\nu})$ and these parameters $\boldsymbol{\nu}$ are then estimated by minimizing 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, the quality of VI depends significantly on the the choice of parametric family and the divergence function.

Algorithm 3 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 $n$
set $i = 0$
for $i = 0$ to $N$ do
    $\{\mathbf{z}_i ... \mathbf{z}_n\} \sim q(\mathbf{z};\boldsymbol{\nu})$    {Generate $n$ samples from variational distribution}
    $\text{ELBO} = \sum_{\mathbf{z}_i} \log\pi(\mathbf{y}_0|\mathbf{z}_i) + \log\pi(\mathbf{z}_i) - \log q(\mathbf{z}_i;\boldsymbol{\nu})$
    $\boldsymbol{\nu} \leftarrow \boldsymbol{\nu} - \epsilon\nabla_{\boldsymbol{\nu}}\text{ELBO}$    {Optimization}
end for
$\boldsymbol{\nu}^* \leftarrow \boldsymbol{\nu}$
output : $q(\mathbf{z};\boldsymbol{\nu}^*)$
```

A.2.1. 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

$$\begin{aligned} D_{\text{KL}}(q||p) &= \mathbb{E}_q(\log q - \log p) \\ &= \mathbb{E}_q(\log q(\mathbf{z};\boldsymbol{\nu}) - \log\pi(\mathbf{z}|\mathbf{y}_0)) \\ &\approx \sum_{\mathbf{z}_i \sim q(\mathbf{z})} \left[\log q(\mathbf{z}_i;\boldsymbol{\nu}) - \log\pi(\mathbf{z}_i|\mathbf{y}_0)\right] \\ &\leq \sum_{\mathbf{z}_i \sim q(\mathbf{z})} \left[\log q(\mathbf{z}_i;\boldsymbol{\nu}) - \log\pi(\mathbf{y}_0|\mathbf{z}_i) - \log\pi(\mathbf{z}_i)\right] \end{aligned} \tag{5}$$

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

$$\text{ELBO} := \sum_{\mathbf{z}_i \sim q(\mathbf{z})} \log\pi(\mathbf{y}_0|\mathbf{z}_i) + \log\pi(\mathbf{z}_i) - \log q(\mathbf{z}_i;\boldsymbol{\nu}) \tag{6}$$

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

$$\boldsymbol{\nu}^* = \text{argmax}_{\nu}\text{ELBO} \tag{7}$$

The full algorithm for this is given in Algorithm 3

A.2.2. 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

$$D_{\text{KL}}(p||q) = \mathbb{E}_p(\log p - \log q) \tag{8}$$

$$= \mathbb{E}_{\pi(\mathbf{z}|\mathbf{y}_0)}(\log\pi(\mathbf{z}|\mathbf{y}_0) - \log q(\mathbf{z};\boldsymbol{\nu})) \tag{9}$$

$$\approx \sum_{\mathbf{z}_i \sim \pi(\mathbf{z}|\mathbf{y}_0)} (\log\pi(\mathbf{z}|\mathbf{y}_0) - \log q(\mathbf{z};\boldsymbol{\nu})) \tag{10}$$

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

$$\boldsymbol{\nu}^* = \text{argmax}_{\nu} \sum_{\mathbf{z}_i \sim \pi(\mathbf{z}|\mathbf{y}_0)} \log q(\mathbf{z};\boldsymbol{\nu}) \tag{11}$$

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 true 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 (Naesseth et al., 2020; Bornschein & Bengio, 2014). However we find that none of these approaches work well in our case.

B. Validating the posterior

In this appendix, we give the details of the metrics we have used to validate and compare the posterior samples of VBS and HMC.

B.1. Distribution of summary statistics

High dimensional distributions are hard to compare quantitatively and so we seek a low dimensional mapping to compare samples from different algorithms. Since our data model obeys rotational and translational invariance, the power spectrum of density fields provides a natural low dimensional candidate. 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 compare the quality of the posterior distributions by measuring the distribution of the transfer function of the posterior samples. Transfer Function (t_f) of these samples is defined as the ratio of power of these samples with the true initial conditions

$$t_{f,i} = \sqrt{P_{\mathbf{z}_i}(k)/P_{\mathbf{z}_{true}}(k)}$$

Thus it 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.

Though not shown in the text, we also compare the cross-correlation coefficient, r_c, measured in terms of cross power spectrum, of the posterior samples with the true initial conditions. However the differences in this were not significant for different algorithms. Note that when both r_c and t_f are unity, the two fields being compared are identical.

B.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

$$\rho_j(t) = \frac{1}{n}\sum_{i=t+1}^{n}(P_i(k_j)-\bar{P}(k_j))(P_{i-t}(k_j)-\bar{P}(k_j)) \tag{12}$$

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.

C. Normalizing Flow

In this section we describe the architecture of our normalizing flow (NF). 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 (Kobyzev et al., 2020). We use NF to parameterize our variational family such that it is flexible enough to capture the target distribution

$$q(\mathbf{z};\boldsymbol{\nu}) = q_B(T_\theta^{-1}(\mathbf{z});\boldsymbol{\nu}_B)|\det\nabla_{\mathbf{z}}T_\theta^{-1}| \tag{13}$$

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

C.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.

C.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 (Dinh et al., 2016; Papamakarios et al., 2017). 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 (Dai & Seljak, 2022). Motivated by the fact that the cosmological fields are rotationally and translationally invariant, (Dai & Seljak, 2022) propose constructing transport maps using Fourier-space convolutions.

C.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

$$\mathbf{x}' = \mathcal{F}^{-1}(t_\theta(k)\mathcal{F}(\mathbf{x})) \tag{14}$$

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.

C.2.2. Element-wise Transformations

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

The simplest Ψ_ϕ's are affine (scale and shift) transformations

$$\mathbf{x}' = \alpha\mathbf{x} + \beta \tag{15}$$

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 N=64, global affine transformations sufficed but for 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, (Dai & Seljak, 2022) 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.

C.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 $f = \mathbf{x}_0 \to \mathbf{x}_1$:

$$\mathbf{x}_1 = \Psi_\phi(\mathcal{F}^{-1}(t_{\theta_1}(k)\mathcal{F}(\mathbf{x}_0))) \tag{16}$$

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