



HAL
open science

Posterior Sampling of the Initial Conditions of the Universe from Non-linear Large Scale Structures using Score-Based Generative Models

Ronan Legin, Matthew Ho, Pablo Lemos, Laurence Perreault-Levasseur, Shirley Ho, Yashar Hezaveh, Benjamin Wandelt

► **To cite this version:**

Ronan Legin, Matthew Ho, Pablo Lemos, Laurence Perreault-Levasseur, Shirley Ho, et al.. Posterior Sampling of the Initial Conditions of the Universe from Non-linear Large Scale Structures using Score-Based Generative Models. *Monthly Notices of the Royal Astronomical Society*, 2023, 527 (1), pp.L173-L178. 10.1093/mnrasl/slad152 . hal-04084060

HAL Id: hal-04084060

<https://hal.science/hal-04084060>

Submitted on 24 Apr 2024





HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Posterior sampling of the initial conditions of the universe from non-linear large scale structures using score-based generative models

Ronan Legin ^{1,2,3}★, Matthew Ho ⁴, Pablo Lemos ^{1,2,3,5}, Laurence Perreault-Levasseur,^{1,2,3,5,6} Shirley Ho,⁵ Yashar Hezaveh ^{1,2,3,5,6} and Benjamin Wandelt^{4,5,7}

¹Department of Physics, Université de Montréal, Montréal H2V 0B3, Canada

²Mila - Quebec Artificial Intelligence Institute, Montréal H2S 3H1, Canada

³Ciela - Montreal Institute for Astrophysical Data Analysis and Machine Learning, Montréal H2V 0B3, Canada

⁴Sorbonne Université, CNRS, UMR 7095, Institut d'Astrophysique de Paris, 98 bis bd Arago, Paris 75014, France

⁵Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA

⁶Perimeter Institute for Theoretical Physics, Waterloo, Ontario, ON N2L 2Y5, Canada

⁷Sorbonne Université, Institut Lagrange de Paris, 98 bis boulevard Arago, Paris 75014, France

Accepted 2023 October 10. Received 2023 October 10; in original form 2023 June 27

ABSTRACT

Reconstructing the initial conditions of the universe is a key problem in cosmology. Methods based on simulating the forward evolution of the universe have provided a way to infer initial conditions consistent with present-day observations. However, due to the high complexity of the inference problem, these methods either fail to sample a distribution of possible initial density fields or require significant approximations in the simulation model to be tractable, potentially leading to biased results. In this work, we propose the use of score-based generative models to sample realizations of the early universe given present-day observations. We infer the initial density field of full high-resolution dark matter N -body simulations from the present-day density field and verify the quality of produced samples compared to the ground truth based on summary statistics. The proposed method is capable of providing plausible realizations of the early universe density field from the initial conditions posterior distribution marginalized over cosmological parameters and can sample orders of magnitude faster than current state-of-the-art methods.

Key words: methods: statistical – early Universe – large-scale structure of Universe.

1 INTRODUCTION

In the standard model of cosmology, structure originates from quantum fluctuations of a primordial density field, which are scaled to macroscopic distances by a physical process called inflation (Guth 1981; Albrecht & Steinhardt 1982; Linde 1982, 1983). This initial density field represents the seed to all structure seen in the Universe today. Furthermore, different models of inflation predict various levels of non-Gaussianity in the primordial density field (Acquaviva et al. 2002; Maldacena 2003; Bartolo et al. 2004). Therefore, accurate methods to reconstruct this primordial field could shed light on the unknown mechanism behind inflation and guide our search for new physics.

Beyond the study of early universe physics, knowledge of the initial conditions of the Universe can be combined with a forward model to compute predictions for any observable on our past light cone. Such predictions can – act as discovery templates for new physical effects in cross-correlation with external data, e.g. for the discovery of secondary anisotropies in the microwave sky; be used to perform posterior predictive tests of the underlying cosmological physics model; or provide insight into quantities that were hitherto

only accessible in simulations, such as the dynamical assembly history of elements of the cosmic web, such as clusters, filaments, or voids (Lavaux & Wandelt 2010; Leclercq, Jasche & Wandelt 2015; Jasche & Lavaux 2019; Feldbrugge & van de Weygaert 2022). For these reasons, substantial effort has been put in the inference of initial conditions as one of the key problems in cosmology. Currently, the primary constraints on the initial conditions come from linear reconstruction (Komatsu, Spergel & Wandelt 2005; Yadav & Wandelt 2005) applied to observations of the Cosmic Microwave Background (CMB; Planck Collaboration 2020a,b).

Upcoming galaxy surveys will provide vast amounts of information on small scales, begging the question: can we infer the small-scale initial conditions from non-linearly evolved structure? This inverse problem compounds the challenges of high dimensionality (representing the initial density field on a 3D grid of 10^7 voxels corresponds to a 10^7 -dimensional inference problem) and the complexity of computing the non-linear forward mapping between the primordial initial conditions and the present-day density field, requiring, at the very least, high-resolution N -body simulations (Springel 2005; Villaescusa-Navarro et al. 2020) and further modelling of the distribution of observable tracers of the underlying dark matter field (Somerville & Davé 2015). Although traditional methods have been used to this end (Hoffman & Ribak 1991; Nusser & Dekel 1992; Bistolas & Hoffman 1998), they typically rely on simplifying

* E-mail: ronan.legin@umontreal.ca

approximations leading to less stringent constraints and possibly biased predictions of the initial conditions.

Current state-of-the-art methods such as Bayesian Origin Reconstruction of Galaxies (BORG; Jasche & Wandelt 2013; Jasche & Lavaux 2019) use Hamiltonian Monte Carlo (HMC), a Markov Chain Monte Carlo (MCMC) algorithm for Bayesian parameter inference of the initial conditions, in which one can exploit the gradient of the likelihood to efficiently generate posterior samples. However, due to the high computational cost of full N -body simulations, BORG relies on approximate simulation methods such as second-order Lagrange Perturbation Theory and Particle-Mesh (PM) simulations, which are inaccurate at small scales. Moreover, they require fully differentiable simulators, which has so far precluded including non-differentiable operations – such as halo finding – that are part of standard modelling pipelines that map dark-matter density fields to galaxy surveys.

In the hopes of improving upon these limitations, machine learning has been used to reconstruct the initial conditions from simulations (Modi et al. 2021; Shallue & Eisenstein 2023). Unfortunately, due to the high-dimensional parameter space, these works have been limited to predicting a single point-estimate, as modelling the full multimillion-dimensional posterior distribution is intractable. Since these models do not produce samples of the early universe, they do not provide any measures of uncertainty on the reconstructions.

In this work, we propose the use of score-based generative diffusion models (Song & Ermon 2019; Ho, Jain & Abbeel 2020; Song et al. 2020) to learn the distribution of early universe density fields conditioned on the present-day matter density field and to produce samples from it. We train a neural network to predict the score of the posterior distribution using simulations from the Quijote latin-hypercube set (Villaescusa-Navarro et al. 2020). We then use the estimate of the score network to solve a reverse-diffusion stochastic differential equation (SDE) to sample the posterior distribution of the initial conditions.

2 METHODS

2.1 Problem overview

Our goal is to infer the 3D density field of the early universe \mathbf{x} given observations \mathbf{y} of the dark matter distribution at low redshift. We can define this problem within a Bayesian framework, where we are interested in sampling from the posterior distribution $p(\mathbf{x}|\mathbf{y})$. Using Bayes's theorem, the posterior distribution $p(\mathbf{x}|\mathbf{y})$ can be written as,

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}, \quad (1)$$

where $p(\mathbf{y}|\mathbf{x})$, $p(\mathbf{x})$, and $p(\mathbf{y})$ represent the likelihood, prior and evidence, respectively. The prior distribution $p(\mathbf{x})$ reflects our knowledge on the possible realizations of the early universe, while the evidence $p(\mathbf{y})$ gives the probability for a realization of the data. The likelihood distribution $p(\mathbf{y}|\mathbf{x})$ represents the distribution of possible observations \mathbf{y} given a fixed realization of the early universe \mathbf{x} . It includes our cosmological forward simulator and additional effects that resemble true observations (e.g. selection functions, galaxy shot noise). For our problem, we let \mathbf{y} be the simulated present-day comoving dark matter overdensity field. We also add low amplitude noise drawn from a normal distribution with standard deviation at a level of 1/10 the standard deviation of the overdensity field \mathbf{y} . The reason for the added noise is to make the score network (described in Section 2.4) more robust to small perturbations with respect to the input condition \mathbf{y} . As such, it is not intended to mimic real observational noise. As detailed in Section 2.5, the early and late

density fields, \mathbf{x} and \mathbf{y} , are represented on 3D mesh grids with 128^3 voxels. Therefore, the sampling space of \mathbf{x} is multimillion in dimension. Because of the high dimensionality of the inference problem, modelling the posterior using density estimation techniques such as normalizing flows is not feasible.

Instead of directly modelling the posterior $p(\mathbf{x}|\mathbf{y})$, we can model the gradient of the log posterior distribution $\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y})$. In comparison, $\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y})$ is computationally tractable, as it does not depend on the normalization of the posterior distribution. This means that it can be approximated by a simple neural network that learns a function $s(\mathbf{x}, \mathbf{y})$ mapping a set of inputs (\mathbf{x}, \mathbf{y}) to an output prediction of the score $\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y})$. In the following section, we describe how we can train a conditional neural network to estimate the score $\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y})$ and use it in the context of score-based generative models (Song & Ermon 2019; Song et al. 2020) to produce samples from the posterior distribution $p(\mathbf{x}|\mathbf{y})$.

2.2 Score-based generative models

Score-based generative modelling is a framework designed to learn the distribution of variables from a data set, by approximating the score $\nabla_{\mathbf{x}} \log p(\mathbf{x})$, which is typically modelled by a neural network (Song et al. 2020). The procedure also provides the possibility to generate new samples from the learned distributions. They have been used to great success across a wide variety of domains (e.g. Popov et al. 2021; Song et al. 2021; Adam et al. 2022; Anand & Achim 2022; Gnaneshwar et al. 2022; Mudur & Finkbeiner 2022; Legin et al. 2023), and have surpassed previous state-of-the-art methods such as generative adversarial networks (GANs, e.g. Dhariwal & Nichol 2021; Müller-Franzes et al. 2022). These achievements have been made possible due to a number of technical improvements in regard to the sampling strategy used to generate samples. In part, one important breakthrough consisted of framing the sampling method as a reverse-diffusion process, where the data \mathbf{x} is perturbed at various noise levels, and sampling new data points \mathbf{x} consists of iteratively reversing this process starting from pure noise. In Song et al. (2020), this process is defined as the SDE:

$$d\mathbf{x} = f(\mathbf{x}, t)dt + g(t)d\mathbf{w}, \quad (2)$$

where $f(\mathbf{x}, t)$ is called the *drift* term, $d\mathbf{w}$ is a Wiener process characterizing the random noise, and $g(t)$ is a scalar function that determines the level of added noise. The key to generating samples \mathbf{x} from $p(\mathbf{x})$ lies in reversing the diffusion process by solving the reverse-SDE:

$$d\mathbf{x} = (f(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})) dt + g(t)d\mathbf{w}, \quad (3)$$

where $p_t(\mathbf{x})$ is the probability distribution of \mathbf{x} at time t . Alternatively, we can extend the previous equation to reverse a conditional diffusion process by solving backward in time

$$d\mathbf{x} = (f(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x}|\mathbf{y})) dt + g(t)d\mathbf{w}, \quad (4)$$

which requires the conditional score $\nabla_{\mathbf{x}} \log p_t(\mathbf{x}|\mathbf{y})$ at time t .

In this work, we train a neural network conditioned on the observation \mathbf{y} and time t , denoted $s(\mathbf{x}, \mathbf{y}, t)$ to learn the score of the posterior, $\nabla_{\mathbf{x}} \log p_t(\mathbf{x}|\mathbf{y})$, via denoising score matching (Hyvärinen 2005; Vincent 2011; Song et al. 2020). We then solve equation (4) to sample from the posterior distribution of initial conditions $p(\mathbf{x}|\mathbf{y})$ by replacing the score $\nabla_{\mathbf{x}} \log p_t(\mathbf{x}|\mathbf{y})$ by its approximation from the score network, $s(\mathbf{x}, \mathbf{y}, t)$.

2.3 Solving the reverse-SDE

There exist many numerical methods to solve the reverse-SDE from equation (4). The ordinary differential equation corresponding to the reverse-SDE can also be solved, as highlighted by Song et al. (2020). Although more efficient, the absence of a noise term makes the solution more sensitive to small errors in the learned score function, which could lead to biased samples. In this work, we solve equation (4) using the Euler-Maruyama method in which discretizes the reverse-SDE as

$$\mathbf{x}_{t+\Delta t} = \mathbf{x}_t + \left(f(\mathbf{x}_t, t) - g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | \mathbf{y}) \right) \Delta t + g(t) \mathbf{z}_t \sqrt{-\Delta t}, \quad (5)$$

where $\Delta t = -1/N$ is the step size, N is the number of steps and \mathbf{z}_t is sampled from a standard normal distribution with the same dimensions as \mathbf{x}_t . We perturb the initial conditions \mathbf{x} with noise following the Variance Exploding SDE (VESDE) proposed in Song et al. (2020). In VESDE, $f(\mathbf{x}_t, t) = 0$ and $g(t) = \sqrt{\frac{d[\sigma^2(t)]}{dt}}$, where $\sigma^2(t)$ is the variance of the noise as a function of time. We then solve the discretized reverse-SDE from equation (5) to sample from the posterior distribution $p(\mathbf{x} | \mathbf{y})$. We let $\sigma(t) = \sigma_{\min} \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)^t$ and based on the geometric interpretation from Song & Ermon (2020), we choose a maximum and minimum diffusion noise level of $\sigma_{\max} = 100$ and $\sigma_{\min} = 0.01$, respectively, and discretize the SDE uniformly across time with $N = 1000$ steps.

2.4 Score network architecture and training

The score network architecture is based on the PYTORCH implementation by Song et al. (2020) available on GitHub.¹ The network follows the RefineNet architecture (Lin et al. 2016) with five downsampling and upsampling levels each with two *ResNet* blocks from the *BigGAN* model (Brock, Donahue & Simonyan 2018) and with outputs of 32 feature maps for the first, fourth, and fifth level and 64 feature maps for the second and third level. Each *ResNet* block contains a *Dropout* layer (Srivastava et al. 2014) with a dropout rate of 10 per cent during training. The main difference with the implementation by Song et al. (2020) is that \mathbf{y} is concatenated to \mathbf{x}_t along the channel dimension and fed as input to the network. A proof showing the validity of this input scheme to learn the score of conditional probability distributions can be found in Batzolis et al. (2021). We train the network for approximately 400 epochs with a batch size of eight split across four NVIDIA H100 80GB Graphical Processing Units (GPU). We use the Adam optimizer with a learning rate of 2×10^{-4} and clip the gradient of the weights to a maximum gradient norm of 1. The duration of training under these settings is approximately 24 h.

2.5 Simulations

We use density fields from the 512³ resolution Quijote latin-hypercube set of N -body simulations (Villaescusa-Navarro et al. 2020) to train and test the score network. Specifically, we use the set of 2000 (1 Gpc h⁻¹)³ dark matter N -body simulations without massive neutrinos and train the score network on the overdensity fields defined as $\rho/\bar{\rho} - 1$, where $\bar{\rho}$ is the average of the density field ρ . The simulations initialized with different random seeds run with different values for the cosmological parameters in the range $\Omega_m \in [0.1, 0.5]$, $\Omega_b \in [0.03, 0.07]$, $h \in [0.5, 0.9]$, $n_s \in [0.8, 1.2]$, and $\sigma_8 \in$

[0.6, 1.0]. We use 1900 N -body simulations for training and 100 for testing with the \mathbf{x} and \mathbf{y} overdensity fields computed on a 128³ grid at redshift $z = 127$ and redshift $z = 0$, respectively. The training and testing losses for the score network, plotted as a function of training epochs, can be found in the supplementary material.

In this work, the redshift $z = 127$ density fields we train the score network with are individually normalized by their own variance, resulting in generated posterior samples of the early universe density field with variance of one. We found that this improves training convergence and in practice, the generated samples can be rescaled with the true variance predicted from the redshift $z = 0$ density field power spectrum from e.g. CAMB (Lewis, Challinor & Lasenby 2000; Villaescusa-Navarro et al. 2020). Note that the samples predicted at redshift $z = 127$ are within the linear regime on the scales we consider and, therefore, can be related directly to an earlier density field (e.g. $z \sim 1000$) using linear theory.

3 RESULTS

We show results using our score network on density fields from the fiducial Quijote set of simulations using the fiducial Planck cosmology with $\Omega_m = 0.3175$, $\Omega_b = 0.049$, $h = 0.6711$, $n_s = 0.9624$, and $\sigma_8 = 0.834$, and from the Quijote latin-hypercube set of simulations from our test set with different cosmological parameter values. Note that our score network was neither trained on density fields from fiducial cosmology simulations nor with the same cosmological parameter values as the test set simulations. For the fiducial cosmology simulation and six simulations from the test set, we solve equation (4) to produce 100 samples from the posterior distribution $p(\mathbf{x} | \mathbf{y})$ using our trained score network $s(\mathbf{x}, \mathbf{y}, t)$. We then compute the power spectrum, cross-correlation and transfer function of these samples, which are then compared to the ground truth in Fig. 3 and in the supplementary material. In Fig. 1 and in the supplementary material, we show examples of sampled initial conditions for the fiducial cosmology simulation. In Fig. 2, we show a map of the posterior sample variance for the fiducial cosmology simulation from Fig. 1.

Additionally, we conduct a test to verify the mean and standard deviation of produced initial condition samples \mathbf{x} as a function of overdensity amplitude \mathbf{y} . For each simulation in our test set, we produce a single sample of initial conditions \mathbf{x} from the posterior $p(\mathbf{x} | \mathbf{y})$. We then bin the voxel values of \mathbf{x} based on the values of \mathbf{y} at the same corresponding voxel position and compute the mean and standard deviation of the binned \mathbf{x} values. We do the same for the true initial conditions \mathbf{x}_{true} for each simulation in the test set. In Fig. 4, the mean $\mu_{\mathbf{x}}$ and standard deviation $\sigma_{\mathbf{x}}$ as a function of different bins in \mathbf{y} is shown for both the sampled and true initial conditions. This test verifies that the score network is capable of reproducing the mean and standard deviation of the true posterior $p(\mathbf{x} | \mathbf{y})$ over a broad range of present-day overdensity values \mathbf{y} .

4 DISCUSSION

In this work, sampling possible initial conditions from the multimillion-dimensional posterior distribution is done by solving a diffusion process backwards in time. This requires the score $\nabla_{\mathbf{x}} \log p_t(\mathbf{x} | \mathbf{y})$, which guides the diffusion towards the posterior distribution while random noise diffuses the samples in order to explore the sampling space. Compared to the posterior probability $p(\mathbf{x} | \mathbf{y})$, modelling $\nabla_{\mathbf{x}} \log p_t(\mathbf{x} | \mathbf{y})$ is computationally tractable as it is independent of the normalization of $p(\mathbf{x} | \mathbf{y})$, which is intractable given the high dimensionality of the inference problem. Therefore,

¹https://github.com/yang-song/score_sde_pytorch

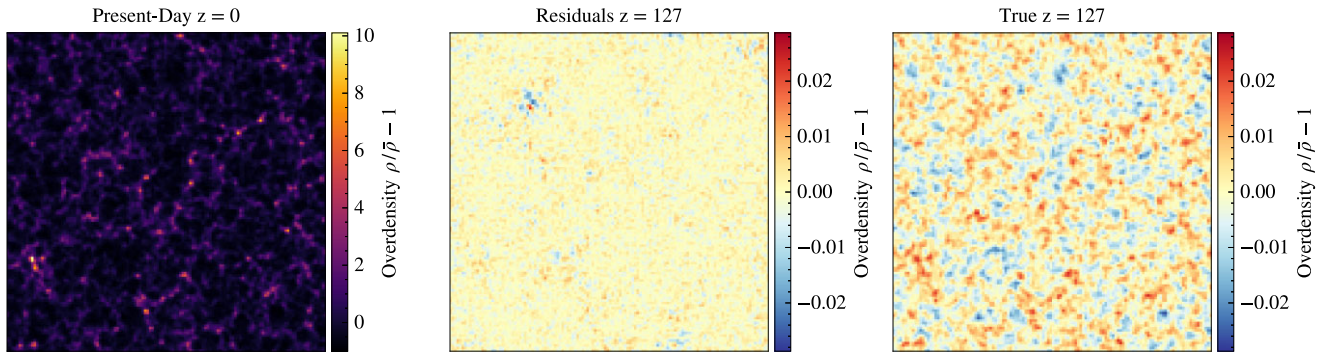


Figure 1. Left: The density field at redshift $z = 0$ for the fiducial Planck cosmology. Centre: Residuals $\mathbf{x}_{\text{sample}} - \mathbf{x}_{\text{true}}$ of the initial conditions between a sample $\mathbf{x}_{\text{sample}}$ generated from the posterior $p(\mathbf{x}|\mathbf{y})$ and the ground truth \mathbf{x}_{true} . Right: The true initial conditions. All three fields span a $1000 \times 1000 \times 15.625$ (Mpc h^{-1})³ region averaged over the third axis. This example demonstrates the capability of score-based generative models to sample highly detailed initial conditions consistent with the ground truth. See Fig. 2 for quantification of uncertainty.

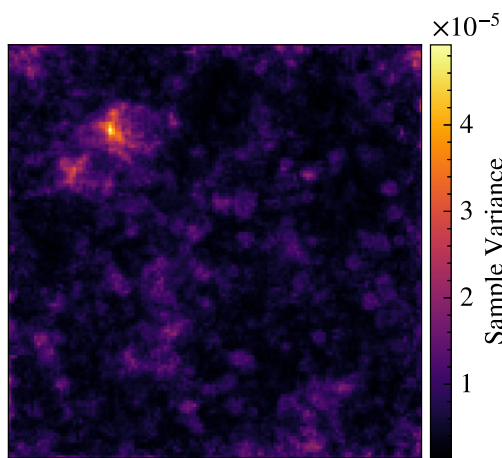


Figure 2. Variance of initial condition samples from the fiducial Planck cosmology simulation per voxel averaged over the depth of a volume of $1000 \times 1000 \times 15.625$ (Mpc h^{-1})³. Reconstruction variance is high in patches that will collapse into large haloes at the present day (see Fig. 1). Moreover, the variance also increases near the boundaries of the data volume. An example of this is the high variance region at the top left corner of the field, which only occurs near the boundary of the data cube.

methods that directly learn the posterior density using models such as masked autoregressive flows (Papamakarios, Pavlakou & Murray 2017) are not suitable for this task.

Efficient MCMC sampling methods in high dimensions such as HMC, as implemented in BORG (Jasche & Wandelt 2013; Jasche & Lavaux 2019), also do not need a normalized posterior distribution, but they do require a differentiable forward model that must be run for each step in the integration of the Hamiltonian trajectory. For our problem, each forward step amounts to running an N -body simulation from initial conditions until the present day. Each such run must be combined with an adjoint run to compute the gradient. Therefore, the generation of each new sample from the posterior requires 10s of simulations. As in all MCMC techniques, accepted samples are correlated leading to a burn-in stage that limits the parallelism of the approach since each running chain must first converge to the posterior distribution before generating useful samples. Moreover, the effective number of samples in the chain is smaller than the number of accepted samples by a factor proportional to the correlation length of the chain. As a result BORG must run

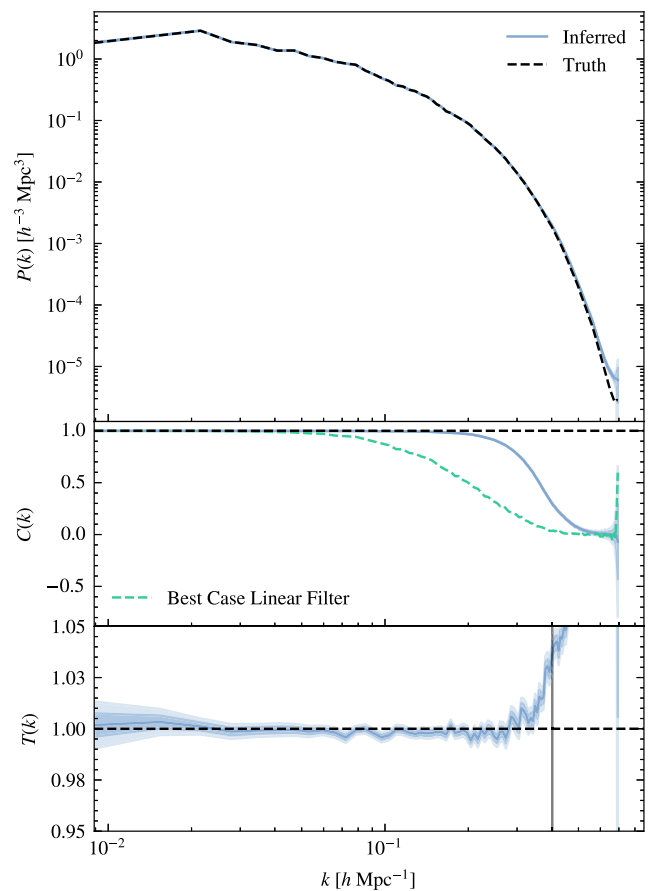


Figure 3. Test statistics to verify the accuracy of samples from the posterior distribution of initial conditions against the ground truth for the fiducial Planck cosmology simulation. Note that this cosmology was not used for training the score network. Top: The power spectrum of the redshift $z = 127$ density field versus the true density field. Middle: The cross-correlation $C(k)$ between every posterior sample and the true field. For comparison we show some for the data and the true field, corresponding to the best achievable $C(k)$ for an optimal linear filter (Wiener filter, green dashed). The posterior samples contain significantly more information. Bottom: The transfer function between the posterior samples and the true field with the Nyquist frequency shown as the vertical black line. The shaded regions represent 1σ and 2σ errors. The results illustrate the high level of accuracy of the inferred initial conditions.

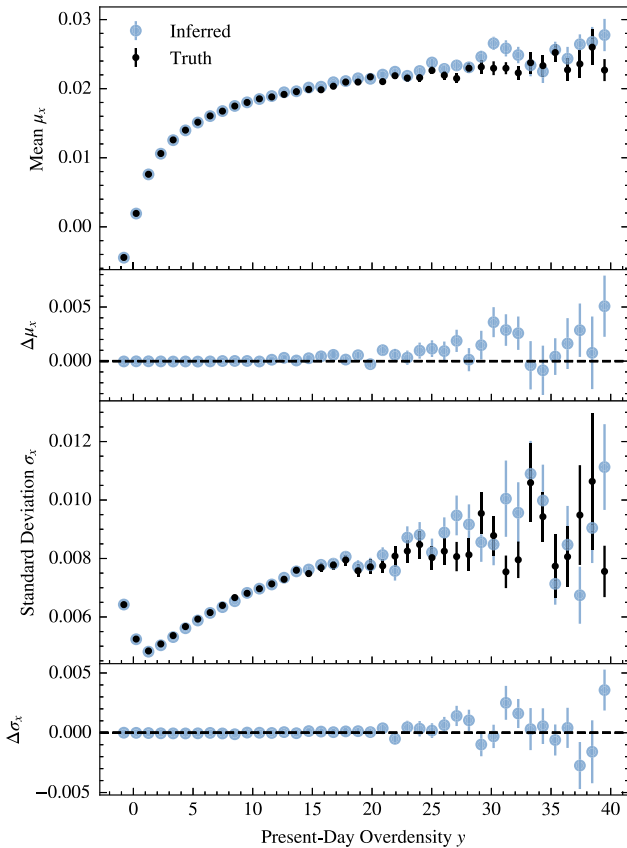


Figure 4. The mean μ_x and standard deviation σ_x of sampled and true initial conditions \mathbf{x} binned as a function of different ranges in present-day overdensity values y . The error bars on μ_x and σ_x were computed using the standard error of the mean and bootstrapping, respectively. Note that the test set consists entirely of simulations with cosmological parameter values not seen during training.

on the order of 100 N -body simulations for each approximately independent sample drawn from the posterior distribution, although this could be improved by fitting a variational approximation for the MCMC proposal distribution (Modi, Li & Blei 2023). Because of this, BORG uses approximate N -body simulators such as PM, which sacrifice accuracy for speed, as they are typically orders of magnitude faster than full N -body simulations. Even so, it can require upwards of 1000 CPU hours to generate a single independent posterior sample of initial conditions. Given this, using a full N -body simulator as the forward model for BORG is not computationally feasible.

In contrast, the method used in this work performs inference using full N -body simulations and can generate independent samples of initial conditions from the posterior distribution within minutes on a single GPU. The sampling procedure can also be trivially parallelized by independently solving equation (4) across multiple devices. Furthermore, our approach samples from the posterior distribution of initial conditions marginalized over cosmological parameters. There are straightforward generalizations to score-based sampling that will allow sampling from the joint posterior distribution of initial conditions and cosmological parameters. This will be explored in future work once we move to more realistic models of the observables.

We now make a qualitative assessment of whether our posterior samples exploit all the available information. As our work is the

first to sample from the posterior distribution of initial conditions using full N -body simulations, it opens the door to using the full information content of future sky surveys for inference. Fig. 3 demonstrates that the posterior samples reproduce the true power spectrum to an accuracy of better than 1 per cent for all k except within <10 per cent of the Nyquist frequency of the grid. The average of the Pearson correlation coefficient between the sampled and true initial conditions is equal to 0.92, signifying that the reconstructed fields account for 92 per cent of the variance of the true initial field. This level is expected given that gravitational collapse destroys information within Lagrangian patches that fall into haloes. Using the Press–Schechter formalism and cutting off the power spectrum at the grid frequency gives an estimated collapse fraction of 12 per cent in resolved haloes. This is close to the measured fraction in the simulation of 16 per cent. These calculations suggest that our samples nearly saturate the information that is physically available given non-linear gravitational collapse and coarse-graining on the grid scale.

It bears mentioning that our analysis of DM density fields represents a more stringent test of the network capability than what will be required for the analysis of realistic observations, since these are far more sparsely sampled and hence noisier than the DM density field. However, the systematic errors inherent in realistic observations may complicate the shape of the posterior, making it more difficult to learn. Investigating the score-based approach in the context of these realistic observations is a topic for future research.

A possible application of our methodology would be to run N -body simulations from the samples of the inferred initial conditions; for existing samplers, such as BORG, this is a natural by-product (Leclercq et al. 2017). These constrained realisations would open the possibility towards unlocking important cosmological questions; we would obtain samples from the posterior distribution of possible N -body simulations that result in the final conditions. This would allow us to sample the posterior over properties of galaxy clusters, such as the positions, masses, and velocities of all dark matter haloes in surveys. Furthermore, the possibility to sample constrained high-resolution N -body simulations would provide us with the means to uncover the set of possible halo assembly histories. For example, this could be used to sample possible histories of the Local Group and the Milky Way, as in McAlpine et al. (2022). Initial condition inference with full N -body simulations would help ensure accuracy since galaxy assembly histories are sensitive to scales that are deeply in the non-linear regime today.

5 CONCLUSIONS

This work proposes score-based generative models for efficient sampling of the posterior distribution of initial conditions, a problem that has up until now been intractable using high-resolution N -body simulations. The key is to sample by solving a reverse-diffusion process requiring only the gradient of the noise-perturbed log posterior $\nabla_x \log p_i(\mathbf{x}|\mathbf{y})$, which can be learned using neural networks. The results show that we can perform accurate inference of the initial conditions marginalized over cosmological parameters at a fraction of the cost of state-of-the-art methods, generating samples from the posterior within minutes on a single GPU. Our tests indicate that the samples have the correct statistics at a level of better than 1 per cent across the relevant range of scales. In future work, we aim to expand the proposed approach to infer initial conditions from simulated halo and galaxy catalogues and ultimately apply it to real data.

ACKNOWLEDGEMENTS

This work is supported by the Simons Collaboration on ‘Learning the Universe’. This work is partially supported by Schmidt Futures, a philanthropic initiative founded by Eric and Wendy Schmidt as part of the Virtual Institute for Astrophysics (VIA). The Flatiron Institute is supported by the Simons Foundation. This work is also supported by the NASA Research Opportunities in Space and Earth Science (ROSES) program through grant number 12-EUCLID12-0004. The work is in part supported by computational resources provided by Calcul Quebec and the Digital Research Alliance of Canada. Yashar Hezaveh and Laurence Perreault-Levasseur acknowledge support from the Canada Research Chairs Program, the Natural Sciences and Engineering Research Council of Canada through grants RGPIN-2020-05073 and 05102, and the Fonds de recherche du Québec through grants 2022-NC-301305 and 300397. Benjamin Wandelt acknowledges support by the ANR BIG4 project, grant ANR-16-CE23-0002 of the French Agence Nationale de la Recherche (ANR); and the Labex Institut Lagrange de Paris (ILP) under the reference ANR-10-LABX-63 part of the Idex SUPER, and received financial state aid managed by the Agence Nationale de la Recherche, as part of the programme Investissements d’avenir under the reference ANR-11-IDEX-0004-02. Ronan Legin thanks the Flatiron Institute for their hospitality and the Centre for Research in Astrophysics of Quebec for their support. Pablo Lemos acknowledges support from the Simons Foundation. We thank Alexandre Adam for reading the manuscript and providing important feedback.

DATA AVAILABILITY

The data used in this article (the Quijote simulations) can be accessed via the Globus server, following the instructions provided in the documentation at <https://quijote-simulations.readthedocs.io/en/latest/access.html>. The source code to replicate this study is open to the public and can be found in the following GitHub repository: <https://github.com/RonanLegin/ICdiffusion>.

REFERENCES

- Acquaviva V., Bartolo N., Matarrese S., Riotto A., 2002, *Nucl. Phys. B*, 667, 119
- Adam A., Coogan A., Malkin N., Legin R., Perreault-Levasseur L., Hezaveh Y., Bengio Y., 2022, preprint (arXiv:2211.03812)
- Albrecht A., Steinhardt P. J., 1982, *Phys. Rev. Lett.*, 48, 1220
- Anand N., Achim T., 2022, preprint (arXiv:2205.15019)
- Bartolo N., Komatsu E., Matarrese S., Riotto A., 2004, *Phys. Rep.*, 402, 103
- Batzolis G., Stanczuk J., Schönlieb C.-B., Etmann C., 2021, preprint (arXiv:2111.13606)
- Bistolas V., Hoffman Y., 1998, *ApJ*, 492, 439
- Brock A., Donahue J., Simonyan K., 2018, preprint (arXiv:1809.11096)
- Dhariwal P., Nichol A., 2021, preprint (arXiv:2105.05233)
- Feldbrugge J., van de Weygaert R., 2022, *J. Cosmol. Astropart. Phys.*, 2023, 67
- Gnaneshwar D., Ramsundar B., Gandhi D., Kurchin R., Viswanathan V., 2022, preprint (arXiv:2203.04698)

- Guth A. H., 1981, *Phys. Rev. D*, 23, 347
- Ho J., Jain A., Abbeel P., 2020, preprint (arXiv:2006.11239)
- Hoffman Y., Ribak E., 1991, *ApJ*, 380, L5
- Hyvärinen A., 2005, *J. Mach. Learn. Res.*, 6, 695
- Jasche J., Lavaux G., 2019, *A&A*, 625, A64
- Jasche J., Wandelt B. D., 2013, *MNRAS*, 432, 894
- Komatsu E., Spergel D. N., Wandelt B. D., 2005, *AJ*, 634, 14
- Lavaux G., Wandelt B. D., 2010, *MNRAS*, 403, 1392
- Leclercq F., Jasche J., Wandelt B., 2015, *J. Cosmol. Astropart. Phys.*, 2015, 015
- Leclercq F., Jasche J., Lavaux G., Wandelt B., Percival W., 2017, *J. Cosmol. Astropart. Phys.*, 2017, 049
- Legin R., Adam A., Hezaveh Y., Perreault Levasseur L., 2023, *ApJ*, 949, L41
- Lewis A., Challinor A., Lasenby A., 2000, *ApJ*, 538, 473
- Lin G., Milan A., Shen C., Reid I., 2016, preprint (arXiv:1611.06612)
- Linde A. D., 1982, *Phys. Lett. B*, 108, 389
- Linde A. D., 1983, *Phys. Lett. B*, 129, 177
- Maldacena J., 2003, *J. High Energy Phys.*, 2003, 013
- McAlpine S. et al., 2022, *MNRAS*, 512, 5823
- Modi C., Lanusse F., Seljak U., Spergel D. N., Perreault-Levasseur L., 2021, preprint (arXiv:2104.12864)
- Modi C., Li Y., Blei D., 2023, *J. Cosmol. Astropart. Phys.*, 2023, 059
- Mudur N., Finkbeiner D. P., 2022, preprint (arXiv:2211.12444)
- Müller-Franzes G. et al., 2022, *Sci. Rep.*, 13, 12098
- Nusser A., Dekel A., 1992, *ApJ*, 391, 443
- Papamakarios G., Pavlakou T., Murray I., 2017, preprint (arXiv:1705.07057)
- Planck Collaboration I, 2020a, *A&A*, 641, A1
- Planck Collaboration IX, 2020b, *A&A*, 641, A9
- Popov V., Vovk I., Gogoryan V., Sadekova T., Kudinov M., 2021, preprint (arXiv:2105.06337)
- Shallue C. J., Eisenstein D. J., 2023, *MNRAS*, 520, 6256
- Somerville R. S., Davé R., 2015, *ARA&A*, 53, 51
- Song Y., Ermon S., 2019, preprint (arXiv:1907.05600)
- Song Y., Ermon S., 2020, *Adv. neural inf. process. syst.*, 33, 12438
- Song Y., Sohl-Dickstein J., Kingma D. P., Kumar A., Ermon S., Poole B., 2020, preprint (arXiv:2011.13456)
- Song Y., Shen L., Xing L., Ermon S., 2021, preprint (arXiv:2111.08005)
- Springel V., 2005, *MNRAS*, 364, 1105
- Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R., 2014, *J. Mach. Learn. Res.*, 15, 1929
- Villaescusa-Navarro F. et al., 2020, *ApJS*, 250, 2
- Vincent P., 2011, *Neural Comput.*, 23, 1661
- Yadav A. P., Wandelt B. D., 2005, *Phys. Rev. D*, 71, 123004

SUPPORTING INFORMATION

Supplementary data are available at *MNRASL* online.

suppl_data

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

This paper has been typeset from a $\text{\TeX}/\text{\LaTeX}$ file prepared by the author.