# Bayesian Simulation-based Inference for Cosmological Initial Conditions

**Florian List**
Department of Astrophysics
University of Vienna, Austria
florian.list@univie.ac.at

**Noemi Anau Montel**
GRAPPA Institute
University of Amsterdam, The Netherlands
n.anaumontel@uva.nl

**Christoph Weniger**
GRAPPA Institute
University of Amsterdam, The Netherlands
c.weniger@uva.nl

## Abstract

Reconstructing astrophysical and cosmological fields from observations is challenging. It requires accounting for non-linear transformations, mixing of spatial structure, and noise. In contrast, forward simulators that map fields to observations are readily available for many applications. We present a versatile Bayesian field reconstruction algorithm rooted in simulation-based inference and enhanced by autoregressive modeling. The proposed technique is applicable to generic (non-differentiable) forward simulators and allows sampling from the posterior for the underlying field. We show first promising results on a proof-of-concept application: the recovery of cosmological initial conditions from late-time density fields.

## 1 Introduction

Recent developments in simulation-based machine learning are increasingly used for tackling difficult astrophysical and cosmological data analysis challenges [*e.g.*, 1–30]. While simulation-based inference (SBI) has primarily been employed to solve relatively low-dimensional ($\lesssim$ 50-dimensional) parameter estimation tasks [4, 9, 20], it has yet to cover higher-dimensionality problems like image reconstruction, which are an essential component in astrophysical and cosmological data analysis.

Here, we focus on the recovery of cosmological initial conditions from late-time density fields. This task is a challenging test case for new algorithms thanks to the non-linear, non-local mapping from the Gaussian target to the observation. Cosmic inflation predicts the density in the early Universe to be highly homogeneous, with tiny density fluctuations that are extremely well described as a Gaussian random field. These density perturbations then gradually grow over cosmic time due to gravity and eventually collapse into the non-Gaussian "Cosmic Web" structure observed today [31]. The reconstruction of the initial density field from late-time observations is an ill-posed problem (the early-to-late mapping is not injective on small scales, [*e.g.*, 32]). Therefore, there is an entire *distribution* of possible initial conditions consistent with a given late-time density field.

**Our contribution.** We frame the task of field reconstruction (or, viewed from a non-physical point of view, image reconstruction) as a parameter inference problem. We combine the power of SBI in solving parametric inverse problems together with the scalability offered by autoregressive models. Autoregressive models have established their versatility in tackling high-dimensional distribution estimation tasks by breaking down the joint distribution into a product of conditionals [33, 34], and have been successful in conditional image modelling [35]. Additionally, we employ a Gibbs sampling

algorithm based on exact data augmentation (GEDA) [36] to efficiently sample image parameter posteriors. We will formulate our method in a generic way to emphasize its applicability to a wide range of field/image reconstruction problems. Importantly, our approach accommodates arbitrary *non-differentiable* forward simulators.

**Related work.** The problem of inferring cosmological initial conditions has been studied since the late 1980s (see *e.g.* Refs. [37–42] for classical papers), for instance by applying the least-action principle or using optimal transport. In the last decade, Bayesian models have been formulated for this task [*e.g.*, 43], many of which rely on differentiable forward models [44, 45], in conjunction with Hamiltonian Monte Carlo sampling. Recently, machine learning methods such as convolutional neural networks [46, 47], variational inference [48], recurrent inference machines [49], and score-based modeling [50] have also been explored in this context.

## 2 Methodology

**Background and problem setup.** SBI methods tackle statistical inverse problems by estimating posterior distributions from model-simulations. These methods do not require explicit modeling of the data likelihood, but instead access the information within the likelihood indirectly via a stochastic simulator, which maps input parameters to simulated data. Among various SBI algorithms (see Ref. [51] for a review and Ref. [52] for benchmarks), we will focus on neural ratio estimation (NRE),[1] which rephrases posterior estimation into a binary classification problem [53, 54].

Let us assume we have a simple hierarchical simulator

$$p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z}) \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^{N \times N}$ is observed and $\boldsymbol{z} \in \mathbb{R}^{N \times N}$ are image parameters (pixel values). Here, $p(\boldsymbol{x}|\boldsymbol{z})$ can include non-linear, non-local transformations and non-Gaussian noise, and it is only implicitly defined through a forward simulator. We will discuss how we can, for a given observation $\boldsymbol{x}_o$, estimate an approximate but computationally efficient Gaussian likelihood that locally resembles $p(\boldsymbol{x}_o|\boldsymbol{z})$ for a target observation $\boldsymbol{x}_o$. A fast surrogate can then be leveraged for downstream image analysis tasks.

**Likelihood estimation.** Firstly, in order to obtain locally optimal data summaries $\boldsymbol{s}(\boldsymbol{x})$ for the image reconstruction task, we use NRE to estimate the marginal, pixel-wise ratios[2]

$$\tilde{r}(s_i(\boldsymbol{x}); z_i) \equiv \frac{p(s_i, z_i)}{p(s_i)p(z_i)} \ . \tag{2}$$

We assume here that both the joint and marginal distributions can be approximated as Gaussians, whereas the mapping $s_i(\boldsymbol{x})$ is an arbitrary learnable function (usually a neural network). Means and covariances are estimated on-the-fly during training (similarly to batch normalization [55]) and are not represented as learnable parameters. Secondly, in order to obtain an estimate of the joint likelihood $p(\boldsymbol{x}|\boldsymbol{z})$, we proceed as follows. We split the problem auto-regressively along the observation axis, and we use NRE to estimate

$$
\begin{aligned}
\frac{p(\boldsymbol{x}|\boldsymbol{z})}{p(\boldsymbol{x})} &\simeq \frac{p(\boldsymbol{s}(\boldsymbol{x})|\boldsymbol{z})}{p(\boldsymbol{s}(\boldsymbol{x}))} = \prod_{i=1}^{N^2} \frac{p(s_i|\boldsymbol{s}_{1:i-1}, \boldsymbol{z})}{p(s_i|\boldsymbol{s}_{1:i-1})} \\
&\simeq \prod_{i=1}^{N^2} \tilde{r}(s_i; l_i, z_i) = \prod_{i=1}^{N^2} \frac{p(s_i, l_i, z_i)}{p(s_i)p(l_i, z_i)} \ ,
\end{aligned}
\tag{3}
$$

where we have introduced $l_i = (\boldsymbol{L}(\boldsymbol{s}))_i$ with $\boldsymbol{L}$ a (generally non-linear) autoregressive function. Again, we assume that the individual functions $p(s_i, l_i, z_i)$ are (three-dimensional multivariate) Gaussians.

---

[1] We use `swyft` NRE implementation that can be found at https://github.com/undark-lab/swyft.

[2] In this work we use the following notation for ratio estimators $\tilde{r}(a; b) = \frac{p(a,b)}{p(a)p(b)} = \frac{p(a|b)}{p(a)}$. If necessary, multiple variables are comma separated, for example $\tilde{r}(a; b, c) = \frac{p(a,b,c)}{p(a)p(b,c)}$.

By rewriting the above components (for a complete derivation and definitions of $\boldsymbol{Q}_{\text{like}}$ and $\boldsymbol{b}$ see Appendix A), we can obtain the likelihood function in the so-called *information form*,

$$\ln p(\boldsymbol{x}|\boldsymbol{z}) = -\frac{1}{2}\boldsymbol{z}^T\boldsymbol{Q}_{\text{like}}\boldsymbol{z} + \boldsymbol{b}\boldsymbol{z} + C(\boldsymbol{s}) \ . \tag{4}$$

In this paper, we assume that the precision matrix $\boldsymbol{Q}_{\text{like}}$ is diagonal. However, correlations between data summaries $\boldsymbol{s}$ are accounted for through the autoregressive function mentioned above; also, each component $s_i(\boldsymbol{x})$ of the data summary may depend on all components of $\boldsymbol{x}$ and thus accounts for cross-pixel information.

To enhance the robustness of Bayesian approaches in data analysis, frequently likelihood tempering techniques are employed that result in conservative estimates [56, 57]. Tempering the likelihood amounts to raising it to a fractional power $\gamma \in [0,1]$, leading to progressively coarser posterior samples, with the extreme case of $\gamma = 0$ ($\gamma = 1$) corresponding to samples drawn from the prior (posterior) distribution.

**Posterior sampling.**    In our Bayesian framework and assuming Gaussian distributions, we combine the estimated likelihood with the known prior to derive the posterior in the information form

$$\ln p(\boldsymbol{z}|\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{z}^T\underbrace{\left(\boldsymbol{Q}_{\text{like}} + \boldsymbol{Q}_{\text{prior}}\right)}_{\boldsymbol{Q}_{\text{post}}}\boldsymbol{z} + \boldsymbol{b}\boldsymbol{z} + C'(\boldsymbol{s}) \ , \tag{5}$$

where we have assumed a zero-mean prior, consistent with the physical problem we study in Sec. 3. We use the conjugate gradient (CG) algorithm[3] to compute the maximum-a-posteriori (MAP) estimate $\boldsymbol{z}_{\text{MAP}}$ of the image parameters $\boldsymbol{z}$ by solving the linear system $\boldsymbol{Q}_{\text{post}}\boldsymbol{z} = \boldsymbol{b}$. The surrogate Gaussian posterior distribution is then given by $p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}_{\text{MAP}}, \boldsymbol{Q}_{\text{post}}^{-1})$. Once we have the posterior distribution in the above form, we simply have to sample from it. Various techniques have been presented over time to tackle the problem of efficient sampling from a high-dimensional Gaussian distribution (for a recent review, see Ref. [58]). To obtain the Gaussian posterior samples, we use a Gibbs sampler based on the generalized exact data augmentation algorithm (GEDA) [36]. GEDA solves the problem of high-dimensional Gaussian sampling specifically for distributions whose precision matrix can be expressed as $\boldsymbol{Q} = \boldsymbol{Q}_1 + \boldsymbol{Q}_2$ by exploiting specific properties of $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ (for details see Appendix B). Crucially, $\boldsymbol{Q}_{\text{post}}$ satisfies these constraints.

## 3   Experiment

To demonstrate the efficiency of our method, we apply it to the task of reconstructing the initial conditions of the Universe. In this proof-of-concept study, we consider the two-dimensional case and assume Einstein–de Sitter cosmology (i.e. non-relativistic, collisionless matter only). While our framework readily supports marginalizing over image parameters such as the power spectra of the target fields, we use a fixed power-law power spectrum for the initial density contrast $\boldsymbol{z} = \boldsymbol{\delta}_{\text{ini}} \in \mathbb{R}^{128 \times 128}$, which is the target of our inference. As a forward model, we use second-order Lagrangian perturbation theory (2LPT, see Refs. [59, 60] and Appendix C) and evolve the initial density to a time when non-linear structures have formed. The observation is then given by $\boldsymbol{x} = \log_{10}[1.1 + \boldsymbol{\delta}_{\text{final}}] + \boldsymbol{\varepsilon}$, where $\boldsymbol{\delta}_{\text{final}}$ is the density contrast at final time, the logarithm is applied element-wise, and we add $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ with $\sigma = 0.15$ as a simplistic model for observational noise.

To obtain pixel-wise summaries $\boldsymbol{s}(\boldsymbol{x})$, we use a standard U-Net [61]. In our current implementation, we use an autoregressive convolution [*e.g.*, 35] as the autoregressive function $\boldsymbol{L}$. In our experiments, we observed that overly small or large kernel sizes result in somewhat biased posteriors. Interestingly, we obtain the best results with a physically motivated kernel size for the convolution, i.e. when choosing it in such as way that the "radius of influence" of the autoregressive convolution (i.e. roughly half the kernel size) matches the typical distance that particles (or fluid elements) have traveled by the end time – a quantity known as displacement. Specifically, the mean displacement in our case is $\sim 5$ pixels, so we take the kernel size to be $11 \times 11$ pixels. While we find this choice to be optimal, slight bias still occurs occasionally, potentially due to the spatial heterogeneity of the

---

[3] We use a slightly modified implementation of the preconditioned CG algorithm from https://github.com/sbarratt/torch_cg.
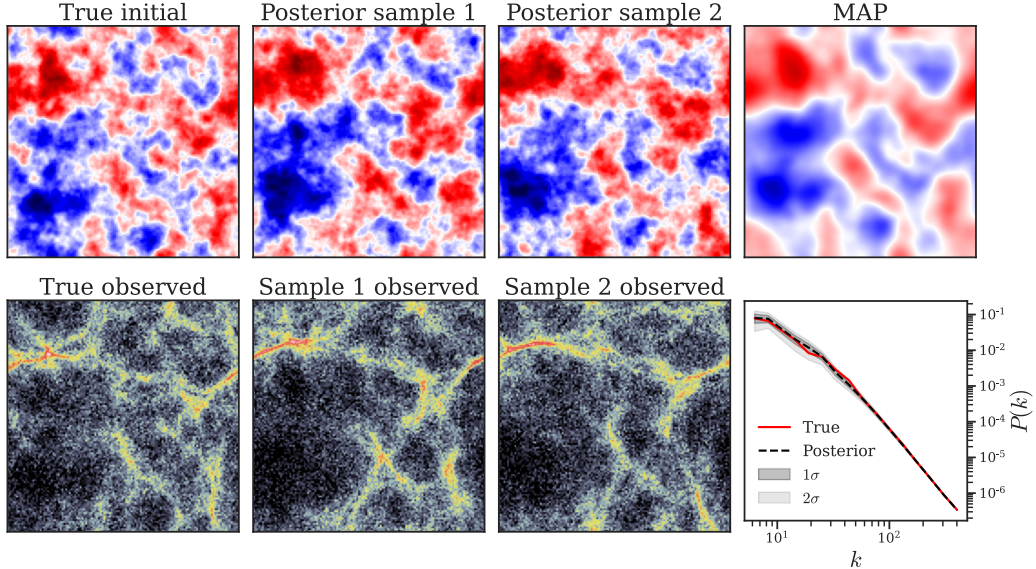
Figure 1: Reconstruction of cosmological initial conditions in 2D. *Top left:* True initial density $\boldsymbol{z}_o$. *Bottom left:* True observation $\boldsymbol{x}_o$, i.e. the (logarithm of the) late-time density evolved from the initial conditions $\boldsymbol{z}_o$, corrupted by uncorrelated Gaussian noise. *Top center:* Two samples drawn from the posterior $p(\boldsymbol{z}|\boldsymbol{x}_o)$. *Bottom center:* Observations computed from the posterior samples shown above. *Upper right:* Maximum-a-posteriori probability (MAP) estimate $\boldsymbol{z}_{\text{MAP}}$ of the initial conditions. *Bottom right:* Distribution of the reconstructed initial power spectrum.

displacements. Therefore, to be conservative, we use a tempered likelihood with $\gamma = 0.5$, trading some of the statistical constraining power of our method for increased unbiasedness. The importance of the specific choice of $\boldsymbol{L}$ (and in particular its radius of influence) indicates that a locally adaptive or multiscale approach for $\boldsymbol{L}$ is a promising avenue for future exploration. We train our model on 1000 $(\boldsymbol{x}, \boldsymbol{z})$-pairs for 15 epochs, which only takes $\sim 3$ minutes on a laptop GPU for this experiment. Obtaining the MAP estimate with the CG algorithm takes less than 3 seconds. To obtain posterior samples with GEDA, we perform 300 sampling steps, which takes $< 1$ second per sample (and multiple samples can be drawn in batches).

The left column of Fig. 1 shows a target initial density contrast $\boldsymbol{z}_o$ (*top*), together with a resulting noisy observation at late time $\boldsymbol{x}_o$ (*bottom*). In the two center panels, we plot samples drawn from the (tempered) posterior $p(\boldsymbol{z}|\boldsymbol{x}_o)$, together with an observation of each sample. The upper right panel shows the MAP estimate $\boldsymbol{z}_{\text{MAP}}$. The initial density field is faithfully reconstructed on large scales whereas, as expected, small-scale information remains unconstrained. Finally, the lower right panel shows that the power spectra of the posterior samples are consistent with the target field. The excellent agreement between reconstructed and true power spectra on small scales is aided by the fact that the power spectrum (which is fixed in our example) directly enters the GEDA sampling (see Eq. 9 in Appendix B, where the prior precision matrix $\boldsymbol{Q}_2$ is diagonalized by the Fourier transform, with the power spectrum on the diagonal of $\boldsymbol{D}_2$). We will present a detailed quantitative analysis of the results obtained with our framework in a separate publication.

## 4    Discussion and Conclusion

We have introduced a framework for Bayesian field/image reconstruction by combining SBI, autoregressive modeling, and a Gibbs sampling algorithm based on exact data augmentation (GEDA). We presented promising results for a toy example related to reconstructing the initial conditions of the Universe. In view of its *remarkable speed, low simulation costs, and the fact that it works for general non-differentiable simulators*, we expect our method to be capable of handling significantly higher-dimensional problems, including 3D cosmological simulations.

There are multiple promising avenues to be explored in future work. For many applications—such as the reconstruction of cosmological initial conditions—formulating the problem in Fourier space can be expected to produce significantly tighter posteriors, as the input-to-output mapping is exactly invertible on large to intermediate scales, and the stochasticity of the reconstruction only affects small scales. Alternatively, wavelets [*e.g.*, 62] or other multiscale techniques could be exploited. In this context (but also more generally), other choices for the autoregressive function $\boldsymbol{L}$ are worthwhile exploring, which could, for instance, implement the hierarchy between the different scales. Another promising enhancement involves harnessing the sequential aspect of SBI techniques. In principle, a viable strategy is to employ an adaptive scheduler to control the value of $\gamma$ to draw new training samples for sequential inference rounds. Moreover, being SBI-based, our proposed method is capable of marginalizing over cosmological parameters or, alternatively, infer them in addition to the phases of the initial random field.

Let us also comment on the current limitations of our method: while the assumption of Gaussianity for the prior is justified for reconstructing cosmological initial conditions, using a Gaussian likelihood is an approximation whose justification depends on the specific task at hand. We take cross-pixel information into account through the summary statistics and an autoregressive function, but we currently model the precision matrix of the likelihood as being diagonal. In addition, the susceptibility of our framework to issues encountered in the context of autoregressive models such as exposure bias [63] merits further investigation.

Finally, we remark that the forward model generating "observations" $\boldsymbol{x}$ does not necessarily need to be a physical one, and our framework also holds great promise for generic image reconstruction problems consisting in inferring one image from another.

## Acknowledgments and Disclosure of Funding

## References

[1] Mishra-Sharma S., Cranmer K., 2022, Physical Review D, 105, 063017

[2] Cole A., Miller B. K., Witte S. J., Cai M. X., Grootes M. W., Nattino F., Weniger C., 2022, Journal of Cosmology and Astroparticle Physics, 09, 004

[3] Alvey J., Bhardwaj U., Domcke V., Pieroni M., Weniger C., 2023 (`arXiv:2309.07954`)

[4] Anau Montel N., Alvey J., Weniger C., 2023 (`arXiv:2308.08597`)

[5] Anau Montel N., Weniger C., 2022, in 36th Conference on Neural Information Processing Systems. (`arXiv:2211.04291`)

[6] Anau Montel N., Coogan A., Correa C., Karchev K., Weniger C., 2022, Monthly Notices of the Royal Astronomical Society, 518, 2746

[7] Alsing J., Wandelt B., 2018, Monthly Notices of the Royal Astronomical Society, 476, L60

[8] Alsing J., Wandelt B., 2019, Monthly Notices of the Royal Astronomical Society, 488, 5093

[9] Alsing J., Charnock T., Feeney S., Wandelt B., 2019, Monthly Notices of the Royal Astronomical Society, 488, 4440

[10] Makinen T. L., Charnock T., Alsing J., Wandelt B. D., 2021, Journal of Cosmology and Astroparticle Physics, 11, 049

[11] Modi C., Philcox O. H. E., 2023 (`arXiv:2309.10270`)

[12] Modi C., Pandey S., Ho M., Hahn C., Blancard B. R.-S., Wandelt B., 2023 (`arXiv:2309.15071`)

[13] Barrué R., Conde-Muíño P., Dao V., Santos R., 2023 (`arXiv:2308.02882`)

[14] Heinrich L., Mishra-Sharma S., Pollard C., Windischhofer P., 2023 (`arXiv:2306.12584`)

[15] Lin K., von Wietersheim-Kramsta M., Joachimi B., Feeney S., 2023, Monthly Notices of the Royal Astronomical Society, 524, 6167

[16] Gagnon-Hartman S., Ruan J., Haggard D., 2023, Monthly Notices of the Royal Astronomical Society, 520, 1

[17] Poh J., Samudre A., Ćiprijanović A., Nord B., Khullar G., Tanoglidis D., Frieman J. A., 2022, in 36th Conference on Neural Information Processing Systems. (`arXiv:2211.05836`)

[18] Lemos P., Cranmer M., Abidi M., Hahn C., Eickenberg M., Massara E., Yallup D., Ho S., 2023, Machine Learning: Science and Technology, 4, 01LT01

[19] Brehmer J., Mishra-Sharma S., Hermans J., Louppe G., Cranmer K., 2019, The Astrophysical Journal, 886, 49

[20] Alvey J., Bhardwaj U., Nissanke S., Weniger C., 2023 (`arXiv:2308.06318`)

[21] Bhardwaj U., Alvey J., Miller B. K., Nissanke S., Weniger C., 2023, Physical Review D, 108, 042004

[22] Alvey J., Gerdes M., Weniger C., 2023, Monthly Notices of the Royal Astronomical Society, 525, 3662

[23] Karchev K., Trotta R., Weniger C., 2023, Monthly Notices of the Royal Astronomical Society, 520, 2209.06733

[24] Crisostomi M., Dey K., Barausse E., Trotta R., 2023, Physical Review D, 108, 044029

[25] Campeau-Poirier È., Perreault-Levasseur L., Coogan A., Hezaveh Y., 2023 (`arXiv:2309.16063`)

[26] Coogan A., Anau Montel N., Karchev K., Grootes M. W., Nattino F., Weniger C., 2023, Monthly Notices of the Royal Astronomical Society

[27] Hahn C., et al., 2023, Journal of Cosmology and Astroparticle Physics, 04, 010

[28] Jeffrey N., Alsing J., Lanusse F., 2020, Monthly Notices of the Royal Astronomical Society, 501, 954

[29] Mishra-Sharma S., Cranmer K., 2022, Physical Review D, 105

[30] Saxena A., Cole A., Gazagnes S., Meerburg P. D., Weniger C., Witte S. J., 2023, Monthly Notices of the Royal Astronomical Society, 525, 6097

[31] Bond J. R., Kofman L., Pogosyan D., 1996, Nature, 380, 603

[32] Brenier Y., Frisch U., Hénon M., Loeper G., Matarrese S., Mohayaee R., Sobolevskiĭ A., 2003, Monthly Notices of the Royal Astronomical Society, 346, 501

[33] Papamakarios G., Pavlakou T., Murray I., 2018 (`arXiv:1705.07057`)

[34] Uria B., Côté M.-A., Gregor K., Murray I., Larochelle H., 2016 (`arXiv:1605.02226`)

[35] Van Den Oord A., Kalchbrenner N., Kavukcuoglu K., 2016, in International conference on machine learning. pp 1747–1756 (`arXiv:1601.06759`)

[36] Marnissi Y., Abboud D., Chouzenoux E., Pesquet J.-C., El-Badaoui M., Benazza-Benyahia A., 2019, in 2019 27th European Signal Processing Conference (EUSIPCO). pp 1–5, doi:10.23919/EUSIPCO.2019.8902496

[37] Croft R. A., Gaztanaga E., 1997, Monthly Notices of the Royal Astronomical Society, 285, 793

[38] Frisch U., Matarrese S., Mohayaee R., Sobolevski A., 2002, Nature, 417, 260

[39] Gramann M., 1993, The Astrophysical Journal, 405, 449

[40] Nusser A., Dekel A., 1992, The Astrophysical Journal, 391, 443

[41] Peebles P. J. E., 1989, Astrophysical Journal Letters, 344, L53

[42] Weinberg D. H., 1992, Monthly Notices of the Royal Astronomical Society, 254, 315

[43] Jasche J., Wandelt B. D., 2013, Monthly Notices of the Royal Astronomical Society, 432, 894

[44] Li Y., Modi C., Jamieson D., Zhang Y., Lu L., Feng Y., Lanusse F., Greengard L., 2022 (arXiv:2211.09815)

[45] Modi C., Lanusse F., Seljak U., 2021, Astronomy and Computing, 37, 100505

[46] Shallue C. J., Eisenstein D. J., 2023, Monthly Notices of the Royal Astronomical Society, 520, 6256

[47] Chen X., Zhu F., Gaines S., Padmanabhan N., 2023, Monthly Notices of the Royal Astronomical Society, 523, 6272

[48] Modi C., Li Y., Blei D., 2023, Journal of Cosmology and Astroparticle Physics, 2023, 059

[49] Modi C., Lanusse F., Seljak U., Spergel D. N., Perreault-Levasseur L., 2021 (arXiv:2104.12864)

[50] Legin R., Ho M., Lemos P., Perreault-Levasseur L., Ho S., Hezaveh Y., Wandelt B., 2023 (arXiv:2304.03788)

[51] Cranmer K., Brehmer J., Louppe G., 2020, Proceedings of the National Academy of Sciences, 117, 30055

[52] Lueckmann J.-M., Boelts J., Greenberg D. S., Gonçalves P. J., Macke J. H., 2021 (arXiv:2101.04653)

[53] Hermans J., Begy V., Louppe G., 2020 (arXiv:1903.04057)

[54] Miller B. K., Cole A., Forré P., Louppe G., Weniger C., 2021, in 35th Conference on Neural Information Processing Systems. (arXiv:2107.01214)

[55] Ioffe S., Szegedy C., 2015, 32nd International Conference on Machine Learning, ICML 2015, 1, 448

[56] Holmes C., Walker S., 2017 (arXiv:1701.08515)

[57] Jasche J., Lavaux G., 2019, Astronomy & Astrophysics, 625, A64

[58] Vono M., Dobigeon N., Chainais P., 2020 (arXiv:2010.01510)

[59] Bouchet F. R., Juszkiewicz R., Colombi S., Pellat R., 1992, Astrophysical Journal Letters, 394, L5

[60] Buchert T., Ehlers J., 1993, Monthly Notices of the Royal Astronomical Society, 264, 375

[61] Ronneberger O., Fischer P., Brox T., 2015, in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp 234–241, doi:10.48550/arXiv.1505.04597

[62] Graps A., 1995, IEEE computational science and engineering, 2, 50

[63] Bengio S., Vinyals O., Jaitly N., Shazeer N., 2015, Advances in neural information processing systems, 28

[64] Peebles P. J. E., 2020, The large-scale structure of the universe. Vol. 98, Princeton university press

[65] Rampf C., Schobesberger S. O., Hahn O., 2022, Monthly Notices of the Royal Astronomical Society, 516, 2840

[66] Hockney R. W., Eastwood J. W., 1988, Computer Simulation Using Particles (1st ed.). CRC Press, doi:10.1201/9780367806934

## A   Appendix: Derivation of the likelihood in the information form

In this appendix we show how to derive the quadratic term $Q_{\text{like}}$ and the linear term $b$ of the likelihood in the information form (Eq. 4) starting from Eq. 3. Having approximated $p(s_i, l_i, z_i)$ with a 3-dimensional Gaussian distribution, we can write the log-likelihood as

$$\ln p(\boldsymbol{s}(\boldsymbol{x})|\boldsymbol{z}) = \sum_{i=1}^{N^2} \ln \frac{p(s_i, l_i, z_i)}{p(l_i, z_i)} = \sum_{i=1}^{N^2} \ln \frac{\mathcal{N}(\mu_{(s_i, l_i, z_i)}, \Sigma_{(s_i, l_i, z_i)})}{\mathcal{N}(\mu_{(l_i, z_i)}, \Sigma_{(l_i, z_i)})} \ . \tag{6}$$

Let us consider just one element $i$ and drop the index $i$ for simplicity of notation. Expanding the above expression for one component, we obtain

$$-\frac{1}{2} \begin{pmatrix} s - \mu_s \\ l - \mu_l \\ z - \mu_z \end{pmatrix}^T \underbrace{\begin{pmatrix} \Sigma_{ss} & \Sigma_{sl} & \Sigma_{sz} \\ \Sigma_{ls} & \Sigma_{ll} & \Sigma_{lz} \\ \Sigma_{zs} & \Sigma_{zl} & \Sigma_{zz} \end{pmatrix}^{-1}}_{\equiv Q' = \begin{pmatrix} Q'_{ss} & Q'_{sl} & Q'_{sz} \\ Q'_{ls} & Q'_{ll} & Q'_{lz} \\ Q'_{zs} & Q'_{zl} & Q'_{zz} \end{pmatrix}} \begin{pmatrix} s - \mu_s \\ l - \mu_l \\ z - \mu_z \end{pmatrix} + \frac{1}{2} \begin{pmatrix} l - \mu_l \\ z - \mu_z \end{pmatrix}^T \underbrace{\begin{pmatrix} \Sigma_{ll} & \Sigma_{lz} \\ \Sigma_{zl} & \Sigma_{zz} \end{pmatrix}^{-1}}_{\equiv Q'' = \begin{pmatrix} Q''_{ll} & Q''_{lz} \\ Q''_{zl} & Q''_{zz} \end{pmatrix}} \begin{pmatrix} l - \mu_l \\ z - \mu_z \end{pmatrix} \ . \tag{7}$$

For a specific observation $\boldsymbol{x}$ we can compute $\boldsymbol{s}$ and therefore $\boldsymbol{l}$, hence individual components $s$ and $l$ (where we have dropped the index $i$). We can then read off the expression in Eq. 7 the terms that depend on $z$. We group the rest into an irrelevant constant $C(s)$. In this way, the *quadratic term* in $z$ in Eq. 7 is

$$-\frac{1}{2} Q_{zz} z^2 \quad \text{with} \quad Q_{zz} \equiv Q'_{zz} - Q''_{zz} \ .$$

Each $Q_{zz}$ is a diagonal entry of $Q_{\text{like}}$ in Eq. 4, while non-diagonal components are 0 by construction. In principle, correlations between pixels in the data $\boldsymbol{x}$ can be modeled and we plan to explore it in future work. The *linear term* in $z$ in Eq. 7 is

$$b \equiv [-(s - \mu_s)Q'_{sz} - (l - \mu_l)Q_{lz} + \mu_z Q_{zz}] \ . \tag{8}$$

Each component $b$ composes the likelihood linear term $\boldsymbol{b}$.

## B   Appendix: Generalized Exact Data Augmentation (GEDA)

In general, data augmentation approaches target precision matrices of the form $\boldsymbol{Q} = \boldsymbol{Q}_1 + \boldsymbol{Q}_2$, which naturally arise from the statistical model under investigation. Taking advantage of potential specific structures of $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$, data augmentation strategies introduce one (or several) auxiliary variable $\boldsymbol{u} \in \mathbb{R}^d$ such that the joint distribution of the pair $(\boldsymbol{u}, \boldsymbol{z})$ yields simple conditional distributions, thus sampling steps for a Gibbs sampler. One can recover the target distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$ via marginalization of the auxiliary variable $\boldsymbol{u}$, either exactly (as in exact data augmentation schemes, like GEDA) or in an asymptotic regime.

In GEDA, as described in Ref. [36], the underlying assumption is that the precision matrix $\boldsymbol{Q}$ can be split as follows:

$$\begin{aligned} \boldsymbol{Q} &= \boldsymbol{Q}_1 + \boldsymbol{Q}_2 \\ \boldsymbol{Q}_1 &= \boldsymbol{G}_1^T \boldsymbol{D}_1 \boldsymbol{G}_1 \\ \boldsymbol{Q}_2 &= \boldsymbol{U}_2^T \boldsymbol{D}_2 \boldsymbol{U}_2 \, , \end{aligned} \tag{9}$$

where $\boldsymbol{G}_1$ is arbitrary, $\boldsymbol{D}_1$ is diagonal and positive definite, $\boldsymbol{U}_2$ is unitary, and $\boldsymbol{D}_2$ is diagonal. GEDA introduces two auxiliary variables, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, such that the joint distribution is

$$
\begin{aligned}
p(\boldsymbol{z}, \boldsymbol{u}_1, \boldsymbol{u}_2) \propto \exp & \left( -\frac{1}{2} \left[ (\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{Q}(\boldsymbol{z} - \boldsymbol{\mu}) + (\boldsymbol{u}_1 - \boldsymbol{z})^T \boldsymbol{R}(\boldsymbol{u}_1 - \boldsymbol{z}) \right] \right) \\
& \times \exp \left( -\frac{1}{2}(\boldsymbol{u}_2 - \boldsymbol{G}_1\boldsymbol{u}_1)^T \boldsymbol{D}_1(\boldsymbol{u}_2 - \boldsymbol{G}_1\boldsymbol{u}_1) \right) ,
\end{aligned}
\tag{10}
$$

where $\boldsymbol{R} = \omega^{-1}\boldsymbol{I}_d - \boldsymbol{Q}_1$, and $\omega$ is a positive hyper-parameter of the algorithm that must obey $1 < \omega < 1/|\boldsymbol{Q}_1|$, meaning that $1/\omega$ should be larger than the largest singular value of $\boldsymbol{Q}_1$.

This joint distribution yields conditional Gaussian distributions with diagonal covariance matrices for both $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ that can be sampled efficiently by a Gibbs sampler with the following steps:

1. $\boldsymbol{u}_2 \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{u}_2}, \boldsymbol{Q}_{\boldsymbol{u}_2}^{-1})$ with $\boldsymbol{\mu}_{\boldsymbol{u}_2} = \boldsymbol{G}_1\boldsymbol{u}_1$ and $\boldsymbol{Q}_{\boldsymbol{u}_2} = \boldsymbol{D}_1$,
2. $\boldsymbol{u}_1 \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{u}_1}, \boldsymbol{Q}_{\boldsymbol{u}_1}^{-1})$ with $\boldsymbol{\mu}_{\boldsymbol{u}_1} = \boldsymbol{z} - \omega(\boldsymbol{Q}_1\boldsymbol{z} - \boldsymbol{G}_1^T\boldsymbol{D}_1^{-1}\boldsymbol{u}_2)$ and $\boldsymbol{Q}_{\boldsymbol{u}_1} = \omega^{-1}\boldsymbol{I}_d$, $\quad$ (11)
3. $\boldsymbol{u}_1 \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{u}_1}, \boldsymbol{Q}_{\boldsymbol{u}_1}^{-1})$ with $\boldsymbol{\mu} = \boldsymbol{Q}(\boldsymbol{R}\boldsymbol{u}_1 + \boldsymbol{Q}\boldsymbol{\mu})$ and $\boldsymbol{Q} = \omega^{-1}\boldsymbol{I}_d + \boldsymbol{Q}_2$ .

As detailed in Sec. 2, in our setup the dimensionality is equivalent to the number of pixels $d = N^2$, the covariance matrix is $\boldsymbol{Q} = \boldsymbol{Q}_{\text{post}} = \boldsymbol{Q}_{\text{like}} + \boldsymbol{Q}_{\text{prior}}$, and $\boldsymbol{G}_1 = \boldsymbol{I}$.

## C   Appendix: Second-order Lagrangian perturbation theory (2LPT)

Second-order Lagrangian perturbation theory (2LPT) describes the cosmological dynamics of a cold, collisionless medium. It is based on a fluid description of the Vlasov–Poisson system (*e.g.* [64]) and therefore ceases to be valid as soon as particle trajectories cross (so-called "shell-crossing"). Although the density fields we consider herein are well in the post-shell-crossing regime, 2LPT still serves as a suitable forward model for validating our image reconstruction method.

The central quantity in 2LPT is the displacement $\boldsymbol{\psi}(\boldsymbol{q}, a) = \boldsymbol{p}(\boldsymbol{q}, a) - \boldsymbol{q}$, i.e. the vector pointing from each Lagrangian ("initial") position $\boldsymbol{q}$ to the associated Eulerian position $\boldsymbol{p}(\boldsymbol{q}, a)$ at scale-factor time $a$ on the characteristic curve originating at $\boldsymbol{q}$. LPT expands the displacement as a Taylor series w.r.t. $a$, with a purely space-dependent coefficient $\boldsymbol{\psi}^{(n)}(\boldsymbol{q})$ at the $n$-th order. For 2LPT, this yields

$$
\boldsymbol{\psi}(\boldsymbol{q}, a) = a\boldsymbol{\psi}^{(1)}(\boldsymbol{q}) + a^2\boldsymbol{\psi}^{(2)}(\boldsymbol{q}) ,
\tag{12}
$$

where

$$
\boldsymbol{\psi}^{(1)} = -\boldsymbol{\nabla}_{\boldsymbol{q}}\phi^0 ,
\tag{13a}
$$

$$
\boldsymbol{\psi}^{(2)} = -\frac{3}{7}\boldsymbol{\nabla}_{\boldsymbol{q}}\Delta_{\boldsymbol{q}}^{-1} \left[ (\partial_{q_1,q_1}\phi^0)(\partial_{q_2,q_2}\phi^0) - (\partial_{q_1,q_2}\phi^0)^2 \right] .
\tag{13b}
$$

Here, $\phi^0$ is the initial gravitational potential, and the scale factor $a$ simply acts as a time variable and has no physical meaning in 2D. Note that we consider Einstein–de Sitter cosmology ($\Omega_m = 1$); for extensions to the $\Lambda$CDM case and higher-order LPT, see *e.g.* Ref. [65] and references therein.

Given the 2LPT displacement $\psi$, the density contrast $\delta$ is given by

$$
1 + \delta(\boldsymbol{p}(\boldsymbol{q}, a)) = \int \delta_{\mathrm{D}}(\boldsymbol{p}(\boldsymbol{q}, a) - \boldsymbol{q} - \boldsymbol{\psi}(\boldsymbol{q}, a)) \, \mathrm{d}^2q ,
\tag{14}
$$

where $\delta_{\mathrm{D}}$ is the Dirac $\delta$-distribution. In practice, we use cloud-in-cell interpolation [*e.g.* 66] to compute the density contrast on a Cartesian grid. Finally, let us remark that since we view the discrete density contrast as an $N \times N$-dimensional image in the main part of this work, we use the bold symbol $\boldsymbol{\delta}$ in that context.