Cosmological Parameter Estimation with Sequential Linear Simulation-based Inference

N. G. Mediato-Diaz
Cavendish Laboratory, University of Cambridge, JJ Thomson Avenue, Cambridge, CB3 0HE

W. J. Handley
Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge, CB3 0HA

We develop the framework of Linear Simulation-based Inference (LSBI), an application of simulation-based inference where the likelihood is approximated by a Gaussian linear function of its parameters. We obtain analytical expressions for the posterior distributions of hyper-parameters of the linear likelihood in terms of samples drawn from a simulator, for both uniform and conjugate priors. This method is applied sequentially to several toy-models and tested on emulated datasets for the Cosmic Microwave Background temperature power spectrum. We find that convergence is achieved after four or five rounds of $\mathcal{O}(10^4)$ simulations, which is competitive with state-of-the-art neural density estimation methods. Therefore, we demonstrate that it is possible to obtain significant information gain and generate posteriors that agree with the underlying parameters while maintaining explainability and intellectual oversight.

Key words: cosmology, Bayesian data analysis, simulation-based inference, CMB power spectrum

I. INTRODUCTION

In many astrophysical applications, statistical models can be simulated forward, but their likelihood functions are too complex to calculate directly. Simulation-based inference (SBI) [1] provides an alternative way to perform Bayesian analysis on these models, relying solely on forward simulations rather than likelihood estimates. However, modern cosmological models are typically expensive to simulate and datasets are often high-dimensional, so traditional methods like the Approximate Bayesian Computation (ABC) [2], which scale poorly with dimensionality, are no longer suitable for parameter estimation. Improvements to ABC, such as the inclusion of Markov-chain Monte Carlo [3] and Sequential Monte Carlo [4] methods, can also have limited effectiveness for large datasets.

In the past decade, machine learning techniques have revolutionized the field of SBI [1], enabling a reduction in the number of simulator calls required to achieve high-fidelity inference, and providing an efficient framework to analyze complex data and expensive simulators. In particular, we highlight Density Estimation Likelihood-free Inference (DELFI) [5, 6] and Truncated Marginal Neural Ratio Estimation (TMNRE) [7]. Given a set of simulated parameter-data pairs, these algorithms learn a parametric model for the joint and the likelihood-to-evidence ratio respectively, via neural density estimation (NRE) techniques [8, 9]. Furthermore, recent applications of SBI to high-dimensional cosmological and astrophysical datasets [10–17] demonstrate that these algorithms are rapidly establishing themselves as a standard machine learning technique in the field.

However, the use of neural networks presents some disadvantages, the most significant of which is their lack of *explainability*. This means that most neural networks are treated as a 'black box', where the decisions taken by the artificial intelligence in arriving at the optimized solution are not known to researchers, which can hinder intellectual oversight [18]. This problem affects the algorithms discussed above, as NRE constitutes an unsupervised learning task, where the artificial intelligence is given unlabeled input data and allowed to discover patterns in its distribution without guidance. This combines with the problem of over-fitting, where the neural network may attempt to maximize the likelihood without regard for the physical sensibility of the output. Current algorithms for simulation-based inference can produce overconfident posterior approximations [19], making them possibly unreliable for scientific use.

The question naturally arises as to whether it is possible to achieve fast and credible simulation-based inference without dependence on neural networks. Such a methodology might allow researchers to acquire control over the inference process and better assess whether their approximations produce overconfident credible regions. Moreover, disentangling ML from SBI can be pedagogically useful in explaining SBI to new audiences by separating its general principles from the details of neural networks.

This paper takes a first step in this direction. In Section II we develop the theoretical framework of Linear Simulation-based Inference (LSBI), an application of likelihood-free inference where the model is approximated by a linear function of its parameters and the noise is assumed to be Gaussian with zero mean. In Section

III, several toy models are investigated using LSBI, and in Section III B the method is applied to parameter estimation for the Cosmic Microwave Background (CMB) power spectrum.

II. THEORY

A. The Linear Approximation

Let us consider a d-dimensional dataset D described by a model $\mathcal{M}$ with n parameters $\theta = \{\theta_i\}$. We assume a Gaussian prior $\theta|\mathcal{M} \sim \mathcal{N}(\mu, \Sigma)$ with known mean and covariance. The likelihood $\mathcal{L}_D(\theta)$ for an arbitrary model is generally intractable, but in this paper, we approximate it as a homoscedastic Gaussian (thus neglecting any parameter-dependence of the covariance matrix),

$$D|\theta, \mathcal{M} \sim \mathcal{N}(\mathcal{M}(\theta), C). \tag{1}$$

Furthermore, we approximate the model linearly about a fiducial point θ_0, such that $\mathcal{M}(\theta) \approx m + M\theta$, where $M \equiv \boldsymbol{\nabla}_\theta \mathcal{M}\big|_{\theta_0}$ and $m \equiv \mathcal{M}(\theta_0) - M\theta_0$. Under these assumptions, the resulting likelihood is

$$D|\theta, \mathcal{M} \sim \mathcal{N}(m + M\theta, C) \tag{2}$$

If we knew the likelihood hyper-parameters m, M, and C, the posterior distribution could be found and would also be Gaussian

$$\theta|D, \mathcal{M} \sim \mathcal{N}(\mu_\mathcal{P}, \Sigma_\mathcal{P}) \tag{3}$$

where

$$\Sigma_\mathcal{P}^{-1} \equiv M^\mathrm{T} C^{-1} M + \Sigma^{-1}, \tag{4}$$

$$\mu_\mathcal{P} \equiv \mu + \Sigma_\mathcal{P} M^\mathrm{T} C^{-1}(D - m - M\mu). \tag{5}$$

Similarly, the evidence would given by

$$D|\mathcal{M} \sim \mathcal{N}(m + M\mu, C + M\Sigma M^\mathrm{T}). \tag{6}$$

Nevertheless, m, M, and C are unknown, so we must obtain their distributions before computing the posterior.

B. Linear Simulation-based Inference

The goal of SBI is to obtain an approximate form of the likelihood through the simulator itself. For the purposes of this paper, the simulator $\mathcal{S}_\mathcal{M}$ is understood to be a stochastic function that takes the model parameters as input and yields a set of noisy simulated data

$$\theta \xrightarrow{\quad\mathcal{S}_\mathcal{M}\quad} D_\mathrm{sim}. \tag{7}$$

Hence, upon fixing a set of parameter samples $\{\theta^{(i)}\}$, we can run the simulator on each to obtain a set of simulated data samples $\{D^{(i)}_\mathrm{sim}\}$. Crucially, these are distributed as the true likelihood $D^{(i)}_\mathrm{sim} \sim \mathcal{L}(\cdot\,|\theta)$. Thus, one may obtain a numerical estimation of the likelihood through the simulator by learning the probability density of the pairs $\{\theta^{(i)}, D^{(i)}_\mathrm{sim}\}$ for a sufficient number of simulator runs. Here, we follow this strategy, except that the assumption of linearity in Eq. 2 avoids the need for machine-learning tools. This linear analysis applied to SBI is not available in the literature, although some recent works are similar, such as SELFI [20] or MOPED [21].

We first draw k parameter vectors $\{\theta^{(i)}\}$; since we are estimating the likelihood, we may draw these from an arbitrary distribution that does not need to be the model prior $\pi(\theta)$. Then, for each $\theta^{(i)}$ the simulator is run to produce the corresponding data vector $D^{(i)}$. We define the first- and second-order statistics

$$\overline{\theta} = \frac{1}{k}\sum_{i=1}^{k}\theta^{(i)}, \qquad \overline{D} = \frac{1}{k}\sum_{i=1}^{k} D^{(i)}, \tag{8}$$

and

$$\Theta = \frac{1}{k}\sum_{i=1}^{k}(\theta^{(i)} - \overline{\theta})(\theta^{(i)} - \overline{\theta})^\mathrm{T}, \tag{9}$$

$$\Delta = \frac{1}{k}\sum_{i=1}^{k}(D^{(i)} - \overline{D})(D^{(i)} - \overline{D})^\mathrm{T}, \tag{10}$$

$$\Psi = \frac{1}{k}\sum_{i=1}^{k}(D^{(i)} - \overline{D})(\theta^{(i)} - \overline{\theta})^\mathrm{T}, \tag{11}$$

Then, by expanding the joint likelihood,

$$\begin{aligned} &p(\{D^{(i)}\}|\{\theta^{(i)}\}, m, M, C) \\ &\qquad \equiv \textstyle\prod_{i=1}^{k} p(D^{(i)}|\theta^{(i)}, m, M, C), \end{aligned} \tag{12}$$

From this result, and a choice of broad uniform priors for m, M and C, we find the distributions

$$m|M, C, S \sim \mathcal{N}\left(\overline{D} - M\overline{\theta}, \tfrac{C}{k}\right) \tag{13}$$

$$M|C, S \sim \mathcal{MN}\left(\Psi\Theta^{-1}, \tfrac{C}{k}, \Theta^{-1}\right) \tag{14}$$

$$C|S \sim \mathcal{W}^{-1}\left(k(\Delta - \Psi\Theta^{-1}\Psi^\mathrm{T}), \nu\right) \tag{15}$$

where $S = \{(\theta^{(i)}, D^{(i)})\}$ are the simulated parameter-data pairs, and $\nu = k - d - n - 2$. $\mathcal{MN}(M, U, V)$ stands for the matrix normal distribution with mean M and covariance $U \otimes V$, and $\mathcal{W}^{-1}(\Lambda, \nu)$ stands for the inverse Wishart distribution with scale matrix Λ and ν degrees of freedom [22]. Note that $\nu > d - 1$, so there is a minimum number of simulations $k_\mathrm{min} = n + 2d + 2$ required to yield well-defined distributions. Appendix A gives the details of the equivalent result for a choice of conjugate priors for m, M, and C,

$$m|M, C, \{\theta^{(i)}\} \sim \mathcal{N}(0, C), \tag{16}$$

$$M|C, \{\theta^{(i)}\} \sim \mathcal{MN}(0, C, \Theta^{-1}), \tag{17}$$

$$C|\{\theta^{(i)}\} \sim \mathcal{W}^{-1}(C_0, \nu_0). \tag{18}$$

The posterior can finally be estimated

$$\mathcal{P}_D(\theta) = \int \mathcal{P}_D(\theta|m,M,C) \times p(m,M,C)\ \mathrm{d}m\,\mathrm{d}M\,\mathrm{d}C \approx \Big\langle \mathcal{P}_D(\theta|m,M,C) \Big\rangle_{m,M,C}. \quad (19)$$

The average can be computed by drawing N exact samples $(m^{(I)}, M^{(I)}, C^{(I)})$ from Eqs. 13 – 15, where N is large enough to guarantee convergence. For $N > 1$, the resulting posterior is a Gaussian mixture of N components. Since each sample is independent, this calculation can be made significantly faster by parallelization, allowing a large N without much effect on the computational efficiency of the calculation.

C. Sequential LSBI

As discussed in Section II A, the linear approximation is only applicable within a localized region surrounding the fiducial point θ_0. Given that the prior distribution is typically broad, this condition is not often satisfied. Consequently, when dealing with non-linear models, LSBI will truncate the model to its linear expansion, thereby computing the posterior distribution corresponding to an 'effective' linear likelihood function. This truncation results in a less constraining model, leading to a broader 'effective' posterior distribution relative to the 'true' posterior.

However, since the simulation parameters $\{\theta^{(i)}\}$ can be drawn from any non-singular distribution, independent of the prior, LSBI can be performed on a set of samples generated by simulations that are proximal to the observed data, i.e., a narrow distribution with θ_0 near the true parameter values. A natural choice for this distribution is the 'effective' LSBI posterior. This leads to the concept of Sequential LSBI, wherein LSBI is iteratively applied to simulation samples drawn from the posterior distribution of the previous iteration, with the initial iteration corresponding to the prior distribution.

It is worth noting that this method suffers from two disadvantages compared to plain LSBI. Firstly, the algorithm is no longer amortized, as subsequent iterations depend on the choice of $D_{\rm obs}$. Secondly, as the sampling distribution becomes narrower, Θ becomes smaller, resulting in a broader distribution for M. Thus, the value of N may need to be increased accordingly.

The evidence may be evaluated similarly. Thus, if the procedure is repeated for a different model $\mathcal{M}'$, the Bayes' ratio between the two models may be calculated,

$$\mathcal{B} = \frac{\langle p(D_{\rm obs}|\mathcal{M})\rangle_{m,M,C}}{\langle p(D_{\rm obs}|\mathcal{M}')\rangle_{m',M',C'}} \quad (20)$$

Nevertheless, this calculation is inefficient for large datasets, so a data compression algorithm is proposed in Appendix B, although it is not investigated further in this paper.

III. RESULTS

In the development of LSBI, we have made two foundational assumptions about the nature of the underlying likelihood and model:

- the model $\mathcal{M}(\theta)$ is approximated as a linear function of the parameters.
- the likelihood $\mathcal{L}_D(\theta)$ can be accurately approximated by a homoscedastic Gaussian distribution (Eq. 1),

In this section, we test the resilience of LSBI against deviations from these assumptions by applying the procedure to several toy models, as well as the CMB temperature power spectrum. These toy models were implemented with the help of the Python package `lsbi`, currently under development by W.J. Handley, and tools from the `scipy` library. To simulate the cosmological power spectrum data, we use the `cmb_tt` neural emulator from CosmoPowerJAX [23, 24].

In addition to LSBI, the parameter posteriors are also calculated via nested sampling with `dynesty` [25–27] for comparison. The plots were made with the software `getdist` [28]. Unless otherwise stated, the calculations in this section use broad uniform priors.

A. Toy Models

For simplicity, the prior on the parameters is a standard normal $\theta \sim \mathcal{N}(0, I_n)$, and the samples for the simulations are taken directly from this prior. To generate the model, we draw the entries of m from a standard normal, whereas the entries of M have mean 0 and standard deviation $1/d$. The covariance C is drawn from $\mathcal{W}^{-1}(\sigma^2 I_d, d+2)$, where $\sigma = 0.5$. The number of samples N taken from posterior distributions of m, M, and C depends on the dataset's dimensionality d; generally, we choose the highest value allowing for manageable computation time.

Our starting point is a 50-dimensional dataset with a quadratic model of $n = 4$ parameters,

$$\mathcal{M}(\theta) = m + M\theta + \theta^{\rm T} Q\,\theta \quad (21)$$

and Gaussian likelihood, where m and M are as above, and Q is a $n \times d \times n$ matrix with entries drawn form $\mathcal{N}(0, 1/d)$. The noise is now Gaussian with covariance C. At each round, we perform LSBI with $k = 2500$ simulations to obtain an estimate for the posterior distribution, where the sampling distribution of the parameter sets $\{\theta^{(i)}\}$ is the posterior of the previous round. The posterior is calculated for a set of 'observed' data, which are determined by applying the model and noise to a set of 'real' parameters θ^*. We also calculate the KL divergence ($\mathcal{D}_{\rm KL}$) between the prior and each posterior, which will help us determine the number of rounds of LSBI required

to obtain convergent results. The posterior distribution is also computed using nested sampling for comparison.

Figure 1 illustrates the outcomes of these simulations. The first iteration of LSBI indeed produces an excessively broad posterior, which subsequent iterations rapidly improve upon. Figure 2 confirms that after four iterations, the Kullback-Leibler divergence between the prior and the LSBI posterior converges to that calculated via nested sampling, with no appreciable discrepancy as expected for Gaussian noise.

In addition, we test the performance of LSBI for non-Gaussian noise shapes. The cases considered are uniform, Student-t with $d+2$ degrees of freedom, and asymmetric Laplacian noise. The model is also given by Eq. 21, and the uncertainty in these distributions is defined directly in terms of the model covariance C. The posterior distribution is also computed using nested sampling for the Laplacian and Student-t cases, whereas that of the uniform likelihood is obtained through an Approximate Bayesian Computation (rejection ABC).

The one- and two-dimensional LSBI posteriors for the models with non-Gaussian error are shown in Figures 3 and 4. The results demonstrate that the posteriors converge to a stable solution after approximately 4 rounds of sequential LSBI. Furthermore, the final $\mathcal{D}_{\rm KL}$ between the prior and posterior distributions approaches the values obtained using nested sampling / rejection ABC methods. Nevertheless, the distributions show some discrepancy, illustrating the fact that non-Gaussian noise can affect the accuracy of LSBI. The results are less satisfactory for Laplacian noise, as the $\mathcal{D}_{\rm KL}$ does not converge to a value within the error bars of the nested-sampling estimation. On the other hand, the lower value of the $\mathcal{D}_{\rm KL}$ estimated via rejection ABC for uniform noise compared to the LSBI values after round 3 is probably due to the inaccuracy of ABC as a posterior estimation method.

We note that the distributions considered here have well-defined first and second moments, so they can be approximated by a Gaussian. Although not considered in this paper, there exist distributions with an undefined covariance, such as the Cauchy distribution (Student-t with one degree of freedom). In these cases, it has been checked that LSBI fails to predict a posterior, instead returning the original prior.

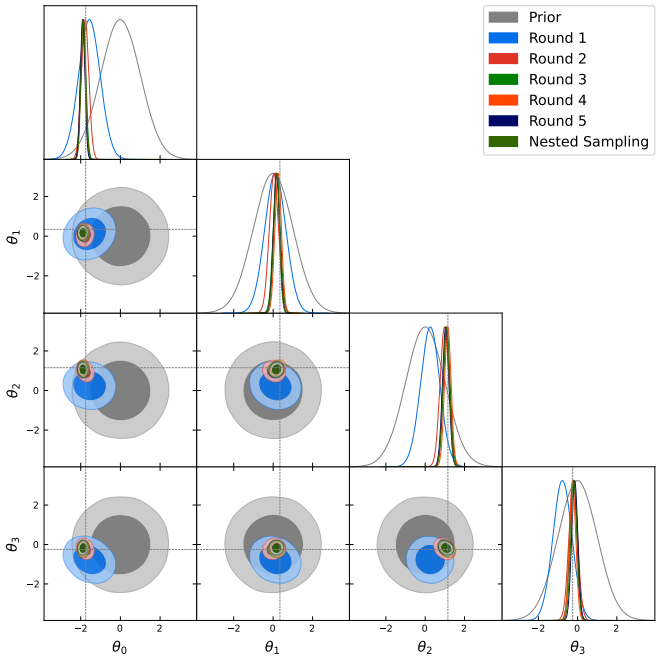

FIG. 1. Prior and posterior distributions on the parameters for a 50-dimensional dataset described by a non-linear 4-parameter model with Gaussian error. Each round corresponds to the output of LSBI after $k = 2500$ simulations, where the sampling distribution of the parameter sets $\{\theta^{(i)}\}$ is the posterior of the previous round. The result of nested sampling is also shown. The dashed lines indicate the values of the 'real' parameters θ^*.

B. The CMB Power Spectrum

In this section, we test the performance of LSBI on a pseudo-realistic cosmological dataset. In the absence of generative Planck likelihoods, we produce the simulations through CosmoPowerJAX [23, 24], a library of machine learning-based inference tools and emulators. In particular, we use the `cmb_tt` probe, which takes as input the six ΛCDM parameters: the Hubble constant H_0, the densities of baryonic matter $\Omega_{\rm b}h^2$ and cold dark matter $\Omega_{\rm c}h^2$, where $h = H_0/100{\rm km\,s^{-1}\ Mpc^{-1}}$, the re-ionization optical depth τ, and the amplitude $A_{\rm s}$ and slope $n_{\rm s}$ of the initial power spectrum of density perturbations. The out-

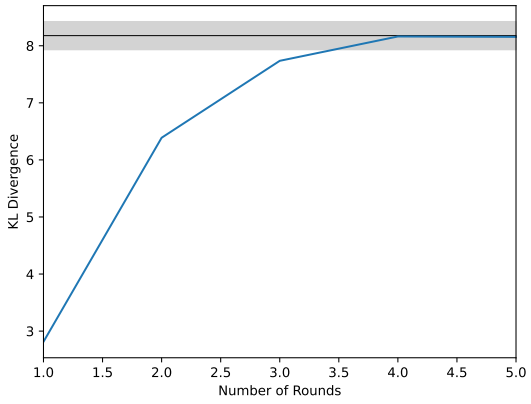

FIG. 2. $\mathcal{D}_{\rm KL}$ between the prior and posterior for each round of Sequential LSBI for the data displayed in Figure 1 The black line corresponds to the value computed via nested sampling; the estimated error is also shown as a gray band.

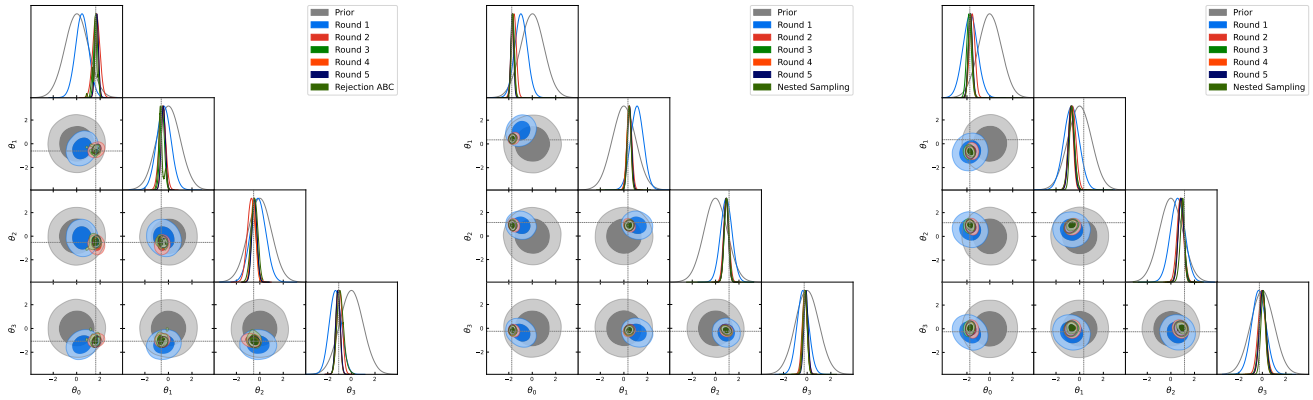

FIG. 3. Prior and posterior distributions on the parameters for a 50-dimensional dataset described by a linear 4-parameter model and non-Gaussian error with $\sigma \approx 0.5$. The posteriors are computed from $k = 106$, 500, 2500, and 10000 samples drawn from the simulated likelihood. The dashed lines indicate the values of the underlying parameters θ^*; (left) uniform noise; (centre) Student-t noise; (right) asymmetric Laplacian noise. The posterior distribution is also computed using nested sampling for the Laplacian and Student-t cases, whereas that of the uniform likelihood is obtained through an Approximate Bayesian Computation (rejection ABC).

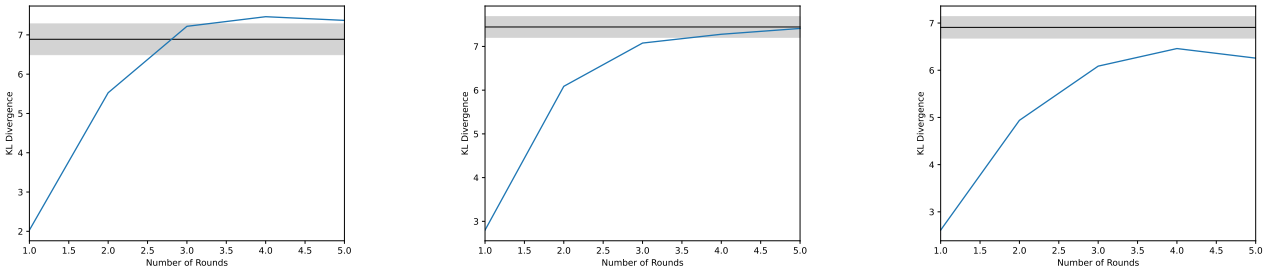

FIG. 4. Kullback-Leibler divergence between the prior and posterior on the parameters as a function of the number of simulations; (left) uniform noise; (center) Student-t noise; (right) asymmetric Laplacian noise. The black line corresponds to the value computed via nested sampling / rejection ABC; the estimated error is also shown as a gray band.

put is the predicted CMB temperature power spectrum

$$\mathcal{C}_\ell = \frac{1}{2\ell+1}\sum_{m=-\ell}^{\ell} |a_{\ell,m}|^2, \qquad 2 \leqslant \ell \leqslant 2058 \tag{22}$$

where $a_{\ell,m}$ are the coefficients of the harmonic expansion of the CMB signal. To this data, we add the standard scaled χ^2 noise

$$\frac{2\ell+1}{C_\ell + N_\ell}\mathcal{C}_\ell \;\sim\; \chi^2(2\ell+1). \tag{23}$$

We apply several rounds of Sequential LSBI, each drawing $k = 10^4$ simulations from the emulator, but keep $N = 100$ since a larger number is computationally unmanageable without parallelization. The parameter samples are drawn from the prior displayed in Eqs. C1 and C2 and, as before, the observed data is generated by running the simulator on a known set of parameters θ^*. The posterior is also obtained by nested sampling with `dynesty`.

The output of this calculation is shown in Figs. 6 and 7. The first figure displays the prior and rounds 1 and 2, while the second shows rounds 3 to 5; in both cases the nested sampling result is shown. It can be noted by eye that the posterior coincides well with the result of nested sampling after four to five rounds of LSBI. This suggests that, although the CMB power spectrum is not well approximated by a linear model at first instance, sequential LSBI succeeds at yielding a narrow sampling distribution about θ^*, thus iteratively approximating the correct posterior.

Figure 5 displays the evolution of the KL divergence up to 8 rounds of LSBI. This result provides further evidence that sequential LSBI converges after $\mathcal{O}(1)$ rounds, thus keeping the total number of simulations within $\mathcal{O}(10^4)$. Nevertheless, we note that the $\mathcal{D}_{\mathrm{KL}}$ is slightly overestimated, suggesting that the LSBI posterior overestimates the true distribution to a small extent. Reproducing this calculation with smaller values of N, we have noticed that the overconfidence increases as N is decreased. Therefore, as discussed in Sections II B and II C,

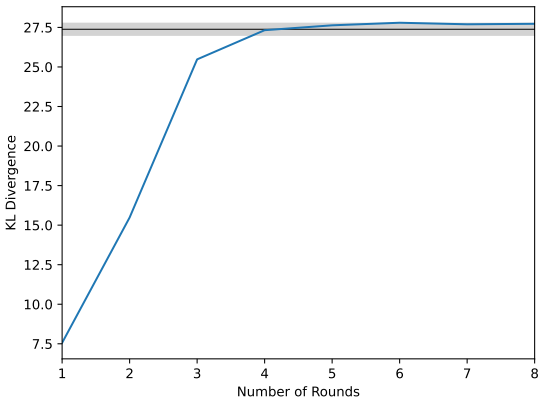

FIG. 5. $\mathcal{D}_{\rm KL}$ between the prior and posterior for each round of Sequential LSBI for the simulated CMB data, displayed in Figures 6 and 7. The black line corresponds to the value computed via nested sampling. The estimated error nested sampling is also shown as a gray band.

the choice of N must be large enough to guarantee the convergence of the integral in Eq. 19. In general we recommend a value at least of order 10^3, but as evidenced by Figure 5, a smaller value may still yield accurate results.

IV. CONCLUSION

In this paper, we have developed the theoretical framework of Linear Simulation-based Inference (LSBI), an application of likelihood-free inference where the model is approximated by a linear function of its parameters and the noise is assumed to be Gaussian with zero mean. Under these circumstances, it is possible to obtain analytical expressions for the posterior distributions of hyper-parameters of the linear likelihood, given a set of samples $S = (\theta^{(i)}, D^{(i)})$, where $D^{(i)}$ are obtained by running simulations on the $\theta^{(i)}$. These parameter samples can be drawn from a distribution other than the prior, so we can exploit this to sequentially approximate the posterior in the vicinity of the observed data.

The analysis of the toy models in Section III illustrates the extent of the resilience of LSBI to deviations from its assumptions. When the error is non-Gaussian, LSBI can still yield accurate estimates, although not universally so, and sequential LSBI provides a way to effectively treat non-linear models. Furthermore, its application to the pseudo-realistic model for the CMB power spectrum demonstrates that it is possible to obtain significant information gain and generate posteriors that agree with the underlying parameters while maintaining explainability and intellectual oversight. We also find that convergence is achieved after $\mathcal{O}(10^4)$ simulations, competitive with state-of-the-art neural density estimation methods [5–7].

Further efforts should be directed towards testing LSBI on more realistic examples, such as the CMB with foregrounds, Baryon Acoustic Oscillations (BAO), or supernovae surveys. In addition, extending this analysis to Gaussian-mixture models may be helpful in better approximating non-Gaussian and multimodal likelihoods.

V. MATERIALS

The source code for the data and plots shown in this paper can be found at https://github.com/ngm29/astro-lsbi/tree/main.

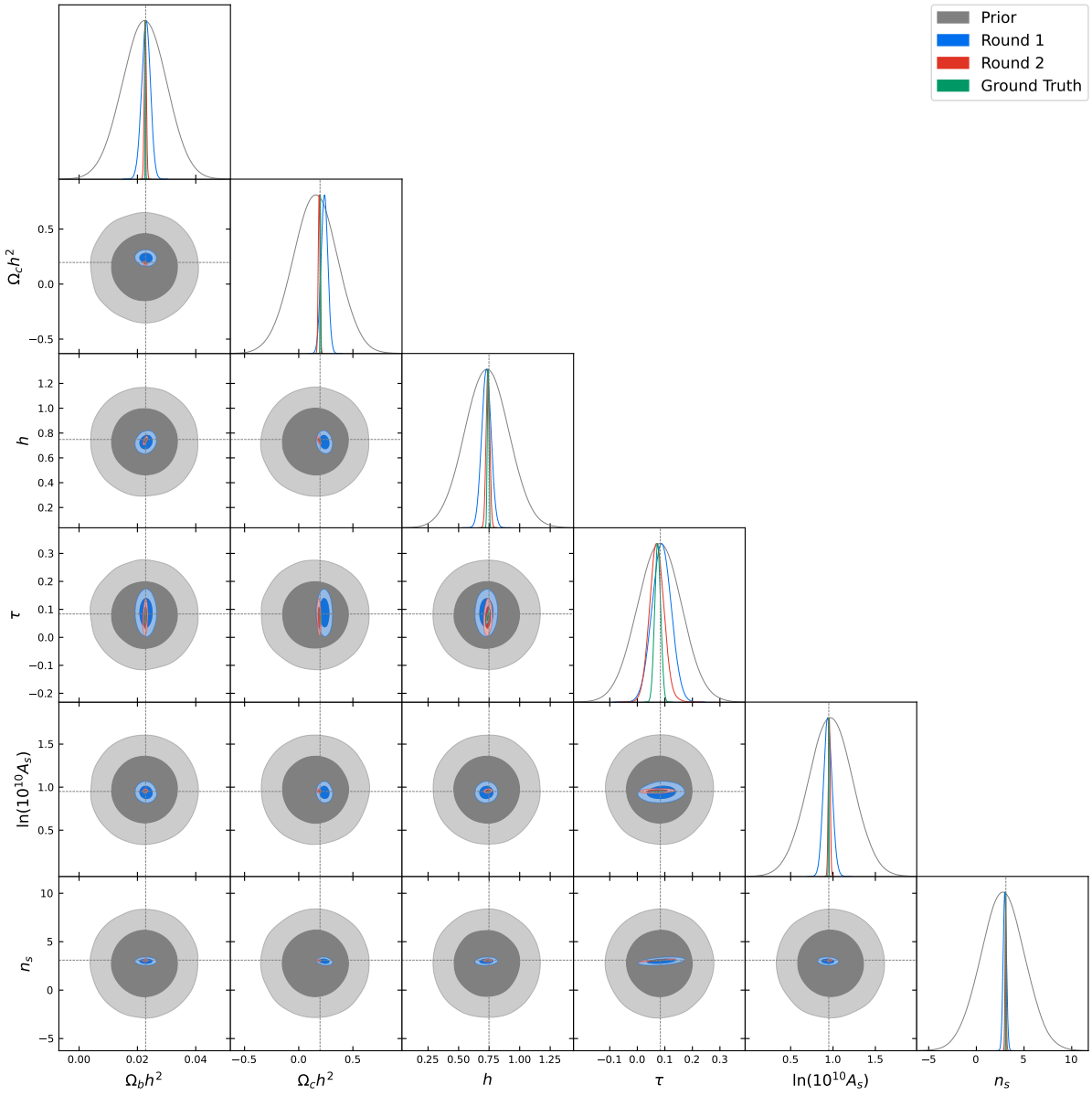

FIG. 6. The plot displays the two-dimensional posterior distributions given by the first two rounds of sequential LSBI, where each round corresponds to the output of LSBI after $k = 10^4$ simulations. The prior distribution and the result of nested sampling on the dataset (labeled "Ground Truth") are also shown. The dashed lines indicate the values of the 'real' parameters θ^*.

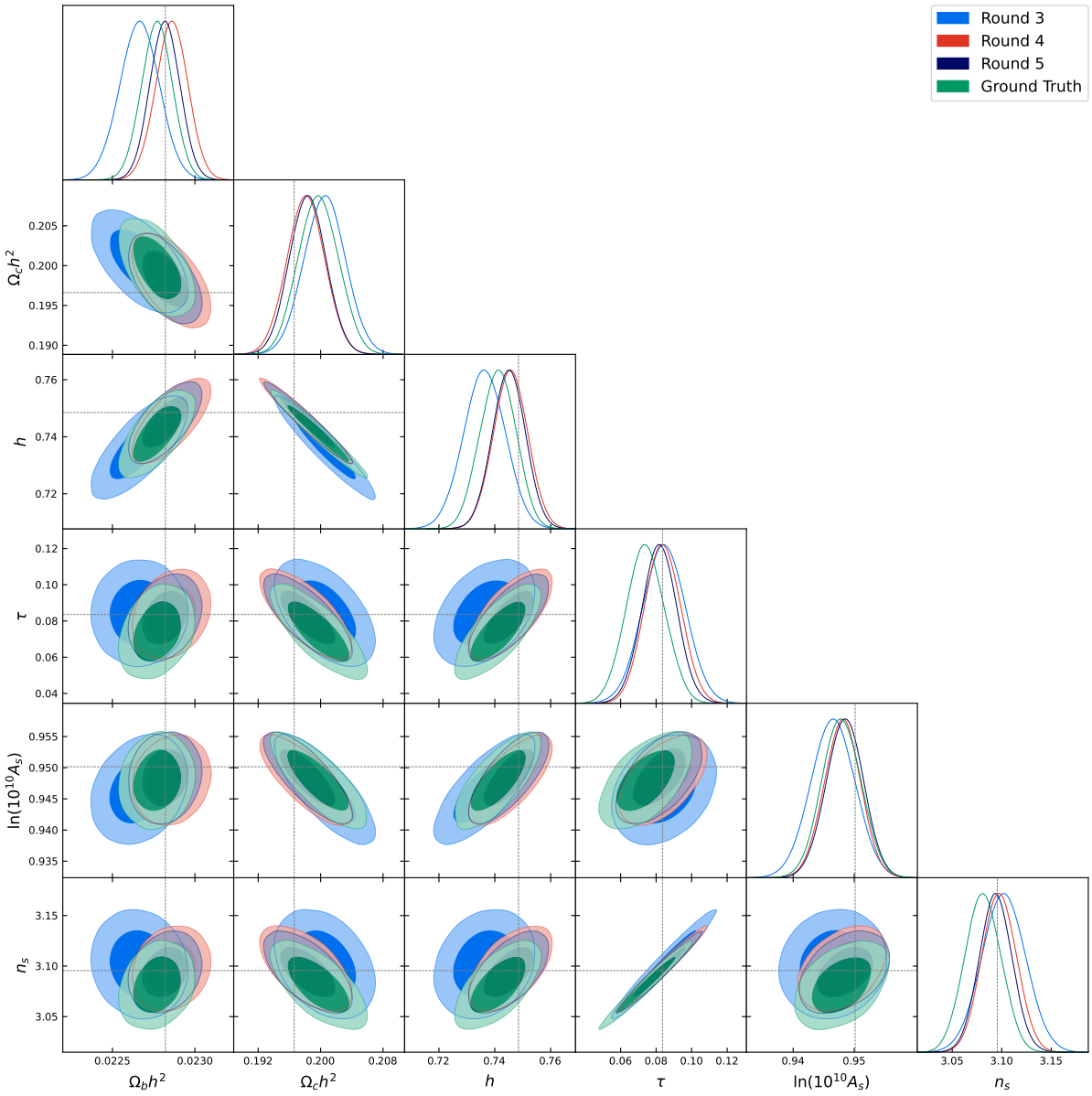

FIG. 7. The plot displays the two-dimensional posterior distributions given by rounds three through five of sequential LSBI, together with the result of nested sampling on the dataset (labeled "Ground Truth"). The dashed lines indicate the values of the 'real' parameters θ^*.

Appendix A: LSBI with Conjugate Priors

Consider a simulator $\mathcal{S}_\mathcal{M}$ which emulates a model $\mathcal{M}$ that can be approximated linearly. We ignore the values of the hyper-parameters m, M, and C, but we may infer them by performing k *independent* simulations $S = \{(D_i, \theta_i)\}$, where the $\{\theta_i\}$ may be drawn from an arbitrary distribution. Defining the sample covariances and means for the data and parameters,

$$\Theta = \tfrac{1}{k}\sum_i (\theta_i - \overline{\theta})(\theta_i - \overline{\theta})^{\mathrm{T}}, \qquad \Delta = \tfrac{1}{k}\sum_i (D_i - \overline{D})(D_i - \overline{D})^{\mathrm{T}}, \tag{A1}$$

$$\Psi = \tfrac{1}{k}\sum_i (D_i - \overline{D})(\theta_i - \overline{\theta})^{\mathrm{T}}, \qquad \overline{D} = \tfrac{1}{k}\sum_i D_i, \qquad \overline{\theta} = \tfrac{1}{k}\sum_i \theta_i, \tag{A2}$$

then, after some algebra, the joint likelihood for the simulations can be written as:

$$\begin{aligned}
\log p(\{D_i\}|\{\theta_i\}, m, M, C) &\equiv \textstyle\sum_i \log p(D_i|\theta_i, m, M, C) \\
&= -\tfrac{k}{2}\log|2\pi C| - \tfrac{1}{2}\textstyle\sum_i (D_i - m - M\theta_i)^{\mathrm{T}} C^{-1} (D_i - m - M\theta_i) \\
&= -\tfrac{k}{2}\log|2\pi C| - \tfrac{1}{2}\left[m - (\overline{D} - M\overline{\theta})\right]^{\mathrm{T}} (C/k)^{-1} \left[m - (\overline{D} - M\overline{\theta})\right] \\
&\quad - \tfrac{k}{2}\operatorname{tr}\left(\Theta(M - \Psi\Theta^{-1})^{\mathrm{T}} C^{-1} (M - \Psi\Theta^{-1})\right) \\
&\quad - \tfrac{k}{2}\operatorname{tr}\left((\Delta - \Psi\Theta^{-1}\Psi^{\mathrm{T}}) C^{-1}\right).
\end{aligned}$$

Using the previous result, we can infer via Bayes' theorem

$$p(m|M, C, S) = \frac{p\left(\{D_i\}|\{\theta_i\}, m, M, C\right)}{p\left(\{D_i\}|\{\theta_i\}, M, C\right)} \cdot p\left(m|M, C, \{\theta^{(i)}\}\right).$$

We choose the conjugate prior $m|M, C, \{\theta^{(i)}\} \sim \mathcal{N}(0, C)$, giving the posterior

$$p(m|M, C, S) = \frac{1}{\sqrt{(2\pi)^d |\frac{C}{k+1}|}} \exp\left(-\frac{1}{2}\left[m - \tfrac{k}{k+1}(\overline{D} - M\overline{\theta})\right]^{\mathrm{T}} \left[C/(k+1)\right]^{-1} \left[m - \tfrac{k}{k+1}(\overline{D} - M\overline{\theta})\right]\right). \tag{A3}$$

Hence, we find a new distribution, with m integrated out (the running evidence),

$$\begin{aligned}
\log p(\{D_i\}|\{\theta_i\}, M, C) &= -\tfrac{k}{2}\log|2\pi C| - \tfrac{d}{2}\log(k+1) \\
&\quad - \tfrac{1}{2}\tfrac{k}{k+1}(\overline{D} - M\overline{\theta})^{\mathrm{T}} C^{-1} (\overline{D} - M\overline{\theta}) \\
&\quad - \tfrac{k}{2}\operatorname{tr}\left(\Theta(M - \Psi\Theta^{-1})^{\mathrm{T}} C^{-1} (M - \Psi\Theta^{-1})\right) \\
&\quad - \tfrac{k}{2}\operatorname{tr}\left((\Delta - \Psi\Theta^{-1}\Psi^{\mathrm{T}}) C^{-1}\right) \\
&= -\tfrac{k}{2}\log|2\pi C| - \tfrac{d}{2}\log(k+1) \\
&\quad - \tfrac{k}{2}\operatorname{tr}\left(\Theta_*(M - \widetilde{\Psi}\Theta_*^{-1})^{\mathrm{T}} C^{-1} (M - \widetilde{\Psi}\Theta_*^{-1})\right) \\
&\quad - \tfrac{k}{2}\operatorname{tr}\left((\widetilde{\Delta} - \widetilde{\Psi}\Theta_*^{-1}\widetilde{\Psi}^{\mathrm{T}}) C^{-1}\right),
\end{aligned}$$

where

$$\widetilde{\Theta} \equiv \Theta + \tfrac{1}{k+1}\overline{\theta}\,\overline{\theta}^{\mathrm{T}}, \qquad \widetilde{\Psi} \equiv \Psi + \tfrac{1}{k+1}\overline{D}\,\overline{\theta}^{\mathrm{T}}, \qquad \widetilde{\Delta} \equiv \Delta + \tfrac{1}{k+1}\overline{D}\,\overline{D}^{\mathrm{T}}. \tag{A4}$$

A second application of Bayes' theorem gives

$$p(M|C, S) = \frac{p\left(\{D_i\}|\{\theta_i\}, M, C\right)}{p\left(\{D_i\}|\{\theta_i\}, C\right)} \cdot p\left(M|C, \{\theta^{(i)}\}\right).$$

We choose the conjugate prior $M|C, \{\theta^{(i)}\} \sim \mathcal{N}(0, C, \Theta^{-1})$, giving the posterior

$$p(M|C, S) = \frac{1}{\sqrt{(2\pi)^{nd} |\frac{C}{k}|^n |\Theta_*^{-1}|^d}} \exp\left(-\frac{1}{2}\operatorname{tr}\left\{\Theta_*\left[M - \widetilde{\Psi}\Theta_*^{-1}\right]^{\mathrm{T}} \left(\frac{C}{k}\right)^{-1} \left[M - \widetilde{\Psi}\Theta_*^{-1}\right]\right\}\right), \tag{A5}$$

where

$$\Theta_* \equiv \widetilde{\Theta} + \tfrac{1}{k}\Theta = \tfrac{k+1}{k}\Theta + \tfrac{1}{k+1}\overline{\theta}\,\overline{\theta}^{\mathrm{T}}. \tag{A6}$$

The running evidence after marginalization over M is then,

$$\begin{aligned}\log p(\{D_i\}|\{\theta_i\}, C) = &-\tfrac{k}{2}\log|2\pi C| - \tfrac{d}{2}\log(k+1) - \tfrac{dn}{2}\log k + \tfrac{d}{2}\log|\Theta_*^{-1}\Theta| \\ &- \tfrac{k}{2}\,\mathrm{tr}\left\{\left(\widetilde{\Delta} - \widetilde{\Psi}\left[\Theta_*^{-1} + \widetilde{\Theta}^{-1} - \Theta^{-1}\right]\widetilde{\Psi}^{\mathrm{T}}\right)C^{-1}\right\}.\end{aligned}$$

A third and final application of Bayes' theorem

$$p(C|S) = \frac{p(\{D_i\}|\{\theta_i\}, C)}{p(\{D_i\}|\{\theta_i\})} \cdot p\left(C, \{\theta^{(i)}\}\right)$$

with conjugate prior $C|\{\theta^{(i)}\} \sim \mathcal{W}^{-1}(C_0, \nu_0)$, gives the posterior

$$p(C|S) = \frac{|C|^{-(\nu+d+1)/2}}{N(S)}\exp\left(-\frac{1}{2}\,\mathrm{tr}\left\{\left[k\left(\widetilde{\Delta} - \widetilde{\Psi}\left[\Theta_*^{-1} + \widetilde{\Theta}^{-1} - \Theta^{-1}\right]\widetilde{\Psi}^{\mathrm{T}}\right) + C_0\right]C^{-1}\right\}\right) \tag{A7}$$

with $\nu = \nu_0 + k$ and $\nu_0 > d$, and

$$N(S) = 2^{\nu d/2}\Gamma_d[\tfrac{\nu}{2}] \times \left|k\left(\widetilde{\Delta} - \widetilde{\Psi}\left[\Theta_*^{-1} + \widetilde{\Theta}^{-1} - \Theta^{-1}\right]\widetilde{\Psi}^{\mathrm{T}}\right) + C_0\right|^{-\nu/2}$$

. The running evidence is then

$$\begin{aligned}\log p(\{D_i\}|\{\theta_i\}) &\equiv \frac{p(\{D_i\}|\{\theta_i\}, C)}{p(C|S)} \cdot p(C) \\ &= -\tfrac{kd}{2}\log\pi - \tfrac{d}{2}\log(k+1) - \tfrac{dn}{2}\log k + \log\big(\Gamma_d[\tfrac{\nu}{2}]/\Gamma_d[\tfrac{\nu_0}{2}]\big) \\ &\quad + \tfrac{d}{2}\log|\widetilde{\Theta}^{-1}\Theta| - \tfrac{1}{2}\log\left\{\left|k\left(\Delta - \Psi\left[\Theta_*^{-1} + \widetilde{\Theta}^{-1} - \Theta^{-1}\right]\Psi^{\mathrm{T}}\right) + C_0\right|^{\nu}/|C_0|^{\nu_0}\right\}.\end{aligned}$$

Finally, we can compute the total evidence for the simulations, where we assume that the parameter samples have been drawn from a Gaussian, $\theta^{(i)} \sim \mathcal{N}(\overline{\theta}, \Theta)$

$$\begin{aligned}\log p(S) &= \log p(\{D_i\}|\{\theta_i\}) + \log p(\{\theta_i\}) \\ &= \log p(\{D_i\}|\{\theta_i\}) - \tfrac{k}{2}\log|2\pi\Theta| - \tfrac{1}{2}nk \\ &= -\tfrac{kd}{2}\log\pi - \tfrac{d}{2}\log(k+1) - \tfrac{dn}{2}\log k + \log\big(\Gamma_d[\tfrac{\nu}{2}]/\Gamma_d[\tfrac{\nu_0}{2}]\big) - \tfrac{1}{2}nk \\ &\quad + \tfrac{d}{2}\log|\widetilde{\Theta}^{-1}\Theta| - \tfrac{k}{2}\log|2\pi\Theta| - \tfrac{1}{2}\log\left\{\left|k\left(\Delta - \Psi\left[\Theta_*^{-1} + \widetilde{\Theta}^{-1} - \Theta^{-1}\right]\Psi^{\mathrm{T}}\right) + C_0\right|^{\nu}/|C_0|^{\nu_0}\right\}.\end{aligned}$$

Appendix B: Model Comparison and Data Compression

If the implicit likelihood is inferred for a different model $\mathcal{M}'$, the Bayes' ratio between the two models may be calculated,

$$\mathcal{B} = \frac{\langle p(D_{\rm obs}|\mathcal{M})\rangle_{m,M,C}}{\langle p(D_{\rm obs}|\mathcal{M}')\rangle_{m',M',C'}} \tag{B1}$$

However, this calculation requires a $d \times d$ matrix to be inverted (see Eq. 6), which scales as $\mathcal{O}(d^\alpha \times N)$, where $2 < \alpha \leqslant 3$ depending on the algorithm used. To increase the computational efficiency, Alsing et. al. [29, 30] and Heavens et. al. [21] remark that for a homoscedastic Gaussian likelihood, the data may be mapped into a set of n summary statistics via the linear compression

$$D \;\mapsto\; M^{\mathrm{T}}C^{-1}(D - m) \tag{B2}$$

without information loss. In our case, the Gaussian likelihood is an approximation, so we may only claim lossless compression to the extent that the approximation is good. The advantage of this method is that C^{-1} may be drawn directly from the distribution

$$C^{-1}|S \sim \mathcal{W}\left([(k-1)(\Delta - \Psi\Theta^{-1}\Psi^{\mathrm{T}})]^{-1}, \nu\right) \tag{B3}$$

and thus, save for one-time inversion of the scale matrix, the complexity of the compression is no more than $\mathcal{O}(n \times d^2 \times N)$.

We propose a slightly different data compression scheme,

$$x = \Gamma M^{\mathrm{T}} C^{-1}(D - m), \quad \Gamma \equiv (M^{\mathrm{T}} C^{-1} M)^{-1}, \tag{B4}$$

where Γ can be computed in $\mathcal{O}(n^2 \times d^2 \times N)$ time. We substitute into Eq. 6 to find

$$x|\mathcal{M} \sim \mathcal{N}(\mu, \Sigma + \Gamma).$$

The matrix $\Sigma + \Gamma$ is only $n \times n$, so if we assume that $n \ll d$, this represents a potential reduction in computational complexity.

Appendix C: Priors

Prior for the ΛCDM parameters, $\theta_{\mathrm{CMB}} = (\Omega_{\mathrm{b}}h^2, \Omega_{\mathrm{c}}h^2, \tau, \ln(10^{10}A_{\mathrm{s}}), n_{\mathrm{s}}, H_0)$

$$\mu_{\mathrm{CMB}} = (2.22 \times 10^{-2}, 0.120, 6.66 \times 10^{-2}, 3.05, 0.964 \times 10^{-1}, 67.3), \tag{C1}$$

$$\Sigma_{\mathrm{CMB}} = \mathrm{diag}(1.05 \times 10^{-3}, 8.28 \times 10^{-3}, 3.47 \times 10^{-2}, 1.47 \times 10^{-1}, 2.64 \times 10^{-2}, 3.38). \tag{C2}$$

[1] K. Cranmer, J. Brehmer, and G. Louppe, The frontier of simulation-based inference, Proceedings of the National Academy of Sciences **117**, 30055 (2020).

[2] D. B. Rubin, Bayesianly justifiable and relevant frequency calculations for the applied statistician, The Annals of Statistics , 1151 (1984).

[3] P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré, Markov chain monte carlo without likelihoods, Proceedings of the National Academy of Sciences **100**, 15324 (2003).

[4] S. A. Sisson, Y. Fan, and M. M. Tanaka, Sequential monte carlo without likelihoods, Proceedings of the National Academy of Sciences **104**, 1760 (2007).

[5] G. Papamakarios and I. Murray, Fast ε-free inference of simulation models with bayesian conditional density estimation, Advances in neural information processing systems **29** (2016).

[6] J. Alsing, T. Charnock, S. Feeney, and B. Wandelt, Fast likelihood-free cosmology with neural density estimators and active learning, Monthly Notices of the Royal Astronomical Society **488**, 4440 (2019).

[7] A. Cole, B. K. Miller, S. J. Witte, M. X. Cai, M. W. Grootes, F. Nattino, and C. Weniger, Fast and credible likelihood-free cosmology with truncated marginal neural ratio estimation, Journal of Cosmology and Astroparticle Physics **2022** (09), 004.

[8] P. Lemos, M. Cranmer, M. Abidi, C. Hahn, M. Eickenberg, E. Massara, D. Yallup, and S. Ho, Robust simulation-based inference in cosmology with bayesian neural networks, Machine Learning: Science and Technology **4**, 01LT01 (2023).

[9] G. Papamakarios, Neural density estimation and likelihood-free inference, arXiv preprint arXiv:1910.13233 (2019).

[10] S. Dupourqué, N. Clerc, E. Pointecouteau, D. Eckert, S. Ettori, and F. Vazza, Investigating the turbulent hot gas in x-cop galaxy clusters, Astronomy & Astrophysics **673**, A91 (2023).

[11] M. Gatti, N. Jeffrey, L. Whiteway, J. Williamson, B. Jain, V. Ajani, D. Anbajagane, G. Giannini, C. Zhou, A. Porredon, *et al.*, Dark energy survey year 3 results: Simulation-based cosmological inference with wavelet harmonics, scattering transforms, and moments of weak lensing mass maps. validation on simulations, Physical Review D **109**, 063534 (2024).

[12] M. Crisostomi, K. Dey, E. Barausse, and R. Trotta, Neural posterior estimation with guaranteed exact coverage: The ringdown of gw150914, Physical Review D **108**, 044029 (2023).

[13] K. Christy, E. J. Baxter, and J. Kumar, Applying simulation-based inference to spectral and spatial information from the galactic center gamma-ray excess, arXiv preprint arXiv:2402.04549 (2024).

[14] J. Harnois-Deraps, S. Heydenreich, B. Giblin, N. Martinet, T. Troester, M. Asgari, P. Burger, T. Castro, K. Dolag, C. Heymans, *et al.*, Kids-1000 and des-y1 combined: Cosmology from peak count statistics, arXiv preprint arXiv:2405.10312 (2024).

[15] B. Moser, T. Kacprzak, S. Fischbacher, A. Refregier, D. Grimm, and L. Tortorelli, Simulation-based inference of deep fields: galaxy population model and redshift distributions, Journal of Cosmology and Astroparticle Physics **2024** (05), 049.

[16] C. P. Novaes, L. Thiele, J. Armijo, S. Cheng, J. A. Cowell, G. A. Marques, E. G. Ferreira, M. Shirasaki, K. Osato, and J. Liu, Cosmology from hsc y1 weak lensing with combined higher-order statistics and simulation-based inference, arXiv preprint arXiv:2409.01301 (2024).

[17] S. Fischbacher, B. Moser, T. Kacprzak, J. Herbel, L. Tortorelli, U. Schmitt, A. Refregier, and A. Amara, `galsbi` : A python package for the galsbi galaxy population model, arXiv preprint arXiv:2412.08722 (2024).

[18] D. Castelvecchi, Can we open the black box of ai?, Nature News **538**, 20 (2016).

[19] J. Hermans, A. Delaunoy, F. Rozet, A. Wehenkel, V. Begy, and G. Louppe, A trust crisis in simulation-based inference? your posterior approximations can be unfaithful, arXiv preprint arXiv:2110.06581 (2021).

[20] F. Leclercq, W. Enzi, J. Jasche, and A. Heavens, Primordial power spectrum and cosmology from black-box galaxy surveys, Monthly Notices of the Royal Astronomical Society **490**, 4237 (2019).

[21] A. F. Heavens, R. Jimenez, and O. Lahav, Massive lossless data compression and multiple parameter estimation from galaxy spectra, Monthly Notices of the Royal Astronomical Society **317**, 965 (2000).

[22] A. K. Gupta and D. K. Nagar, *Matrix variate distributions* (Chapman and Hall/CRC, 2018).

[23] D. Piras and A. S. Mancini, Cosmopower-jax: high-dimensional bayesian inference with differentiable cosmological emulators, arXiv preprint arXiv:2305.06347 (2023).

[24] A. Spurio Mancini, D. Piras, J. Alsing, B. Joachimi, and M. P. Hobson, Cosmopower: emulating cosmological power spectra for accelerated bayesian inference from next-generation surveys, Monthly Notices of the Royal Astronomical Society **511**, 1771 (2022).

[25] J. S. Speagle, dynesty: a dynamic nested sampling package for estimating bayesian posteriors and evidences, Monthly Notices of the Royal Astronomical Society **493**, 3132 (2020).

[26] S. Koposov, J. Speagle, K. Barbary, G. Ashton, E. Bennett, J. Buchner, C. Scheffler, B. Cook, C. Talbot, J. Guillochon, *et al.*, joshspeagle/dynesty: v2. 0.0, Zenodo (2022).

[27] E. Higson, W. Handley, M. Hobson, and A. Lasenby, Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation, Statistics and Computing **29**, 891 (2019).

[28] A. Lewis, Getdist: a python package for analysing monte carlo samples, arXiv preprint arXiv:1910.13970 (2019).

[29] J. Alsing and B. Wandelt, Generalized massive optimal data compression, Monthly Notices of the Royal Astronomical Society: Letters **476**, L60 (2018).

[30] J. Alsing, B. Wandelt, and S. Feeney, Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology, Monthly Notices of the Royal Astronomical Society **477**, 2874 (2018).