Bayesian optimization for likelihood-free cosmological inference

Florent Leclercq[*]

Imperial Centre for Inference and Cosmology (ICIC) & Astrophysics Group, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, United Kingdom

(Received 24 May 2018; published 12 September 2018)

Many cosmological models have only a finite number of parameters of interest, but a very expensive data-generating process and an intractable likelihood function. We address the problem of performing likelihood-free Bayesian inference from such black-box simulation-based models, under the constraint of a very limited simulation budget (typically a few thousand). To do so, we adopt an approach based on the likelihood of an alternative parametric model. Conventional approaches to approximate Bayesian computation such as likelihood-free rejection sampling are impractical for the considered problem, due to the lack of knowledge about how the parameters affect the discrepancy between observed and simulated data. As a response, we make use of a strategy previously developed in the machine learning literature (Bayesian optimization for likelihood-free inference, BOLFI), which combines Gaussian process regression of the discrepancy to build a surrogate surface with Bayesian optimization to actively acquire training data. We extend the method by deriving an acquisition function tailored for the purpose of minimizing the expected uncertainty in the approximate posterior density, in the parametric approach. The resulting algorithm is applied to the problems of summarizing Gaussian signals and inferring cosmological parameters from the joint lightcurve analysis supernovae data. We show that the number of required simulations is reduced by several orders of magnitude, and that the proposed acquisition function produces more accurate posterior approximations, as compared to common strategies.

DOI: 10.1103/PhysRevD.98.063511

I. INTRODUCTION

We consider the problem of Bayesian inference from cosmological data, in the common scenario where we can generate synthetic data through forward simulations, but where the exact likelihood function is intractable. The generative process can be extremely general: it may be a noisy nonlinear dynamical system involving an unrestricted number of latent variables. Likelihood-free inference methods, also known as approximate Bayesian computation (ABC) (see [1,2] for reviews), replace likelihood calculations with data model evaluations. In recent years, they have emerged as a viable alternative to likelihood-based techniques, when the simulator is sufficiently cheap. Applications in cosmology include measuring cosmological parameters from type Ia supernovae [3] and weak lensing peak counts [4], analyzing the galaxy halo connection [5], inferring the photometric and size evolution of galaxies [6], measuring cosmological redshift distributions [7], estimating the ionizing background from the Lyman-α and Lyman-β forests [8].

In its simplest form, ABC takes the form of likelihood-free rejection sampling and involves forward simulating data from parameters drawn from the prior, then accepting parameters when the discrepancy (by some measure) between simulated data and observed data is smaller than a user-specified threshold ε. Such an approach tends to be extremely expensive since many simulated data sets get rejected, due to the lack of knowledge about the relation between the model parameters and the corresponding discrepancy. Variants of likelihood-free rejection sampling such as population (or sequential) Monte Carlo ABC [(PMC-ABC) or (SMC-ABC)] (see [9–11] for implementations aimed at astrophysical applications) improve upon this scheme by making the proposal adaptive; however, they do not use a probabilistic model for the relation between parameters and discrepancies (also known as a surrogate surface), so that their practical use usually necessitates $\mathcal{O}(10^4 - 10^6)$ evaluations of the simulator.

In this paper, we address the challenging problem where the number of simulations is extremely limited, e.g., to a few thousand, rendering the use of sampling-based ABC methods impossible. To this end, we use Bayesian optimization for likelihood-free inference (BOLFI) [12], an algorithm which combines probabilistic modeling of the discrepancy with optimization to facilitate likelihood-free inference. Since it was introduced, BOLFI has been applied to various statistical problems in science, including inference of the Ricker model [12], the Lotka-Volterra

[*]florent.leclercq@polytechnique.org; http://www.florent-leclercq.eu/

predator-prey model and population genetic models [13], pathogen spread models [2], atomistic structure models in materials [14], and cognitive models in human-computer interaction [15]. This work aims at introducing BOLFI in cosmological data analysis and at presenting its first cosmological application. We focus on computable parametric approximations to the true likelihood (also known as synthetic likelihoods), rendering the approach completely ϵ-free. Recently, Järvenpää *et al.* [16] introduced an acquisition function for Bayesian optimization (the expected integrated variance), specifically tailored to perform efficient and accurate ABC. We extend their work by deriving the expression of the expected integrated variance in the parametric approach. This acquisition function measures the expected uncertainty in the estimate of the BOLFI posterior density, which is due to the limited number of simulations, over the future evaluation of the simulation model. The next simulation location is proposed so that this expected uncertainty is minimized. As a result, high-fidelity posterior inferences can be obtained with orders of magnitude fewer simulations than with likelihood-free rejection sampling. As examples, we demonstrate the use of BOLFI on the problems of summarizing Gaussian signals and inferring cosmological parameters from the joint lightcurve analysis (JLA) supernovae data set [17].

The structure of this paper is as follows. In Sec. II, we provide a review of the formalism for the inference of simulator-based statistical models. In Sec. III, we describe BOLFI and discuss the regression and optimization strategies. In particular, we provide the optimal acquisition rule for ABC in the parametric approach to likelihood approximation. Applications are given in Sec. IV. The developed method is discussed in Sec. V in the context of cosmological data analysis. Section VI concludes the paper. Mathematical details and descriptions of the case studies are presented in the Appendices.

II. INFERENCE OF SIMULATOR-BASED STATISTICAL MODELS

A. Simulator-based statistical models

Simulator-based statistical models (also known as generative models) can be written in a hierarchical form (Fig. 1), where $\boldsymbol{\theta}$ are the parameters of interest, and $\mathbf{d}$ the simulated data. $\mathcal{P}(\boldsymbol{\theta})$ is the prior probability distribution of $\boldsymbol{\theta}$ and $\mathcal{P}(\mathbf{d}|\boldsymbol{\theta})$ is the sampling distribution of $\mathbf{d}$ given $\boldsymbol{\theta}$.

The simplest case (Fig. 1, left) is when the simulator is a deterministic function of its input and does not use any random variable, i.e.,

$$\mathcal{P}(\mathbf{d}|\boldsymbol{\theta}) = \delta_{\mathrm{D}}(\mathbf{d} - \hat{\mathbf{d}}(\boldsymbol{\theta})), \tag{1}$$

where δ_{D} is a Dirac delta distribution and $\hat{\mathbf{d}}$ a deterministic function of $\boldsymbol{\theta}$.

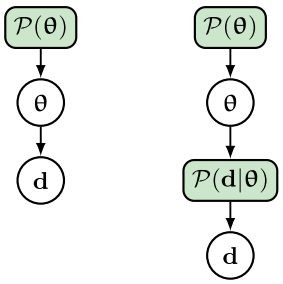

FIG. 1. Hierarchical representation of the exact Bayesian problem for simulator-based statistical models of different complexities: a deterministic simulator (left), and a stochastic simulator (right).

In a more generic scenario (Fig. 1, right), the simulator is stochastic, in the sense that the data are drawn from an overall (but often unknown analytically) probability distribution function (pdf) $\mathcal{P}(\mathbf{d}|\boldsymbol{\theta})$. Equation (1) does not hold in this case. The scatter between different realizations of $\mathbf{d}$ given the same $\boldsymbol{\theta}$ can have various origins. In the simplest case, it only reflects the intrinsic uncertainty, which is of interest. More generically, additional nuisance parameters can be at play to produce the data $\mathbf{d}$ and will contribute to the uncertainty. This "latent space" can often be hundred-to-multi-million dimensional. Simulator-based cosmological models are typically of this kind: although the physical and observational processes simulated are repeatable features about which inferences can be made, the particular realization of Fourier phases of the data is entirely noise driven. Ideally, phase-dependent quantities should not contribute to any measure of match or mismatch between model and data.

B. The exact Bayesian problem

The inference problem is to evaluate the probability of $\boldsymbol{\theta}$ given $\mathbf{d}$,

$$\mathcal{P}(\boldsymbol{\theta}|\mathbf{d}) = \mathcal{P}(\mathbf{d}|\boldsymbol{\theta})\frac{\mathcal{P}(\boldsymbol{\theta})}{\mathcal{P}(\mathbf{d})}, \tag{2}$$

for the observed data $\mathbf{d}_{\mathrm{O}}$, i.e.,

$$\mathcal{P}(\boldsymbol{\theta}|\mathbf{d})_{|\mathbf{d}=\mathbf{d}_{\mathrm{O}}} = \mathcal{L}(\boldsymbol{\theta})\frac{\mathcal{P}(\boldsymbol{\theta})}{Z_{\mathbf{d}}}, \tag{3}$$

where the exact likelihood for the problem is defined as

$$\mathcal{L}(\boldsymbol{\theta}) \equiv \mathcal{P}(\mathbf{d}|\boldsymbol{\theta})_{|\mathbf{d}=\mathbf{d}_{\mathrm{O}}}. \tag{4}$$

It is generally of unknown analytical form. The normalization constant is $Z_{\mathbf{d}} \equiv \mathcal{P}(\mathbf{d})_{|\mathbf{d}=\mathbf{d}_{\mathrm{O}}}$, where $\mathcal{P}(\mathbf{d})$ is the marginal distribution of $\mathbf{d}$.

C. Approximate Bayesian computation

Inference of simulator-based statistical models is usually based on a finite set of simulated data $\mathbf{d}_{\boldsymbol{\theta}}$, generated with parameter value $\boldsymbol{\theta}$, and on a measurement of the discrepancy between simulated data and observed data $\mathbf{d}_{\mathrm{O}}$. This discrepancy is used to define an approximation to the exact likelihood $\mathcal{L}(\boldsymbol{\theta})$. The approximation happens on multiple levels.

On a physical and statistical level, the approximation consists of compressing the full data $\mathbf{d}_{\mathrm{O}}$ to a set of summary statistics $\boldsymbol{\Phi}_{\mathrm{O}}$ before performing inference. Similarly, simulated data $\mathbf{d}_{\boldsymbol{\theta}}$ are compressed to simulated summary statistics $\boldsymbol{\Phi}_{\boldsymbol{\theta}}$. This can be seen as adding a layer to the Bayesian hierarchical model (Fig. 2). The purpose of this operation is to filter out the information in $\mathbf{d}$ that is not deemed relevant to the inference of $\boldsymbol{\theta}$, so as to reduce the dimensionality of the problem. Ideally, $\boldsymbol{\Phi}$ should be *sufficient* for parameters $\boldsymbol{\theta}$, i.e., formally $\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{\Phi}) = \mathcal{P}(\boldsymbol{\theta}|\boldsymbol{\Phi},\mathbf{d})$ or equivalently $\mathcal{P}(\mathbf{d}|\boldsymbol{\Phi},\boldsymbol{\theta}) = \mathcal{P}(\mathbf{d}|\boldsymbol{\Phi})$, which happens when the compression is lossless. However, sufficient summary statistics are generally unknown or even impossible to design; therefore the compression from $\mathbf{d}$ to $\boldsymbol{\Phi}$ will usually be lossy. The approximate inference problem to be solved is now $\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{\Phi}) = \mathcal{P}(\boldsymbol{\Phi}|\boldsymbol{\theta})\frac{\mathcal{P}(\boldsymbol{\theta})}{\mathcal{P}(\boldsymbol{\Phi})}$ for the observed summary statistics $\boldsymbol{\Phi}_{\mathrm{O}}$, i.e.,

$$\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{\Phi})_{|\boldsymbol{\Phi}=\boldsymbol{\Phi}_{\mathrm{O}}} = L(\boldsymbol{\theta})\frac{\mathcal{P}(\boldsymbol{\theta})}{Z_{\boldsymbol{\Phi}}}. \tag{5}$$

In other words, $\mathcal{L}(\boldsymbol{\theta})$ is replaced by

$$L(\boldsymbol{\theta}) \equiv \mathcal{P}(\boldsymbol{\Phi}|\boldsymbol{\theta})_{|\boldsymbol{\Phi}=\boldsymbol{\Phi}_{\mathrm{O}}}, \tag{6}$$

and $Z_{\mathbf{d}}$ by $Z_{\boldsymbol{\Phi}} \equiv \mathcal{P}(\boldsymbol{\Phi})_{|\boldsymbol{\Phi}=\boldsymbol{\Phi}_{\mathrm{O}}}$. Inference of model 2 gives

$$\mathcal{P}(\boldsymbol{\theta},\mathbf{d}|\boldsymbol{\Phi}) \propto \mathcal{P}(\boldsymbol{\Phi}|\mathbf{d})\mathcal{P}(\mathbf{d}|\boldsymbol{\theta})\mathcal{P}(\boldsymbol{\theta}), \tag{7}$$

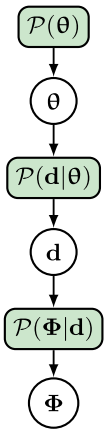

FIG. 2. Hierarchical representation of the approximate Bayesian inference problem for simulator-based statistical models, with a compression of the raw data to a set of summary statistics.

with, after marginalization over $\mathbf{d}$,

$$\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{\Phi}) = \int \mathcal{P}(\boldsymbol{\theta},\mathbf{d}|\boldsymbol{\Phi})\mathrm{d}\mathbf{d}. \tag{8}$$

Therefore, the approximate likelihood $L(\boldsymbol{\theta})$ must satisfy

$$L(\boldsymbol{\theta}) \propto \int \mathcal{P}(\boldsymbol{\Phi}|\mathbf{d})_{|\boldsymbol{\Phi}=\boldsymbol{\Phi}_{\mathrm{O}}}\mathcal{P}(\mathbf{d}|\boldsymbol{\theta})\mathrm{d}\mathbf{d}. \tag{9}$$

In many cases, the compression from $\mathbf{d}$ to $\boldsymbol{\Phi}$ is deterministic, i.e.,

$$\mathcal{P}(\boldsymbol{\Phi}|\mathbf{d}) = \delta_{\mathrm{D}}(\boldsymbol{\Phi} - \hat{\boldsymbol{\Phi}}(\mathbf{d})), \tag{10}$$

which simplifies the integral over $\mathbf{d}$ in Eqs. (8) and (9).

On a practical level, $L(\boldsymbol{\theta})$ is still of unknown analytical form [which is a property of $\mathcal{P}(\boldsymbol{\Phi}|\boldsymbol{\theta})$ inherited from $\mathcal{P}(\mathbf{d}|\boldsymbol{\theta})$ in model 2]. Therefore, it has to be approximated using the simulator. We denote by $\hat{L}^N(\boldsymbol{\theta})$ an estimate of $L(\boldsymbol{\theta})$ computed using N realizations of the simulator. The limiting approximation, in the case where infinite computer resources were available, is denoted by $\tilde{L}(\boldsymbol{\theta})$, such that

$$\hat{L}^N(\boldsymbol{\theta}) \underset{N\to\infty}{\to} \tilde{L}(\boldsymbol{\theta}). \tag{11}$$

Note that $\tilde{L}(\boldsymbol{\theta})$ can be different from $L(\boldsymbol{\theta})$, depending on the assumptions made to construct $\hat{L}^N(\boldsymbol{\theta})$. These are discussed in Sec. II D.

D. Computable approximations of the likelihood

1. Deterministic simulators

The simplest possible case is when the simulator does not use any random variable, i.e., $\boldsymbol{\Phi}_{\boldsymbol{\theta}}$ is an entirely deterministic function of $\boldsymbol{\theta}$ (see Fig. 1, left). Equivalently, all the conditional probabilities appearing in Eq. (7) reduce to Dirac delta distributions given by Eqs. (1) and (10). In this case, one can directly use the approximate likelihood given by Eq. (6), complemented by an assumption on the functional shape of $\mathcal{P}(\boldsymbol{\Phi}|\boldsymbol{\theta})$.

2. Parametric approximations and the synthetic likelihood

When the simulator is not deterministic, the pdf $\mathcal{P}(\boldsymbol{\Phi}|\boldsymbol{\theta})$ is unknown analytically. Nonetheless, in some situations, it may be reasonably assumed to follow specific parametric forms.

For example, if $\boldsymbol{\Phi}_{\boldsymbol{\theta}}$ is obtained through averaging a sufficient number of independent and identically distributed variables contained in $\mathbf{d}$, the central limit theorem suggests that a Gaussian distribution is appropriate, i.e., $\tilde{L}(\boldsymbol{\theta}) = \exp[\tilde{\ell}(\boldsymbol{\theta})]$ with

$$-2\tilde{\ell}(\boldsymbol{\theta}) \equiv \log|2\pi\boldsymbol{\Sigma}_{\boldsymbol{\theta}}| + (\boldsymbol{\Phi}_{\mathrm{O}} - \boldsymbol{\mu}_{\boldsymbol{\theta}})^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\Phi}_{\mathrm{O}} - \boldsymbol{\mu}_{\boldsymbol{\theta}}), \quad (12)$$

where the mean and covariance matrix,

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} \equiv \mathrm{E}[\boldsymbol{\Phi}_{\boldsymbol{\theta}}] \quad \text{and} \quad \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \equiv \mathrm{E}[(\boldsymbol{\Phi}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_{\boldsymbol{\theta}})(\boldsymbol{\Phi}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_{\boldsymbol{\theta}})^{\top}], \quad (13)$$

can depend on $\boldsymbol{\theta}$. This is an approximation of $L(\boldsymbol{\theta})$, unless the summary statistics $\boldsymbol{\Phi}_{\boldsymbol{\theta}}$ are indeed Gaussian distributed. $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ are generally unknown, but can be estimated using the simulator: given a set of N simulations $\{\boldsymbol{\Phi}_{\boldsymbol{\theta}}^{(i)}\}$, drawn independently from $\mathcal{P}(\boldsymbol{\Phi}|\boldsymbol{\theta})$, one can define

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} \equiv \mathrm{E}^{N}[\boldsymbol{\Phi}_{\boldsymbol{\theta}}] \text{ and } \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} \equiv \mathrm{E}^{N}[(\boldsymbol{\Phi}_{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})(\boldsymbol{\Phi}_{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})^{\top}], \quad (14)$$

where E^{N} stands for the empirical average over the set of simulations. A computable approximation of the likelihood is therefore $\hat{L}^{N}(\boldsymbol{\theta}) = \exp[\hat{\ell}^{N}(\boldsymbol{\theta})]$, where

$$-2\hat{\ell}^{N}(\boldsymbol{\theta}) \equiv \log|2\pi\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}| + (\boldsymbol{\Phi}_{\mathrm{O}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})^{\top}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\Phi}_{\mathrm{O}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}). \quad (15)$$

Due to the approximation of the expectation E with an empirical average E^{N}, both $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ become random objects. The approximation of the likelihood $\hat{L}^{N}(\boldsymbol{\theta})$ is therefore a random function with some intrinsic uncertainty itself, and its computation is a stochastic process. This is further discussed using a simple example in Sec. IV A.

The approximation given in Eq. (15), known as the synthetic likelihood [18,19], has already been applied successfully to perform approximate inference in several scientific fields. However, as pointed out by Sellentin and Heavens [20], for inference from Gaussian-distributed summaries $\boldsymbol{\Phi}_{\boldsymbol{\theta}}$ with an estimated covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$, a different parametric form, namely a multivariate t-distribution, should rather be used. The investigation of a synthetic t-likelihood is left to future investigations.

In Sec. IV A and Appendix B, we extend previous work on the Gaussian synthetic likelihood and introduce a Gamma synthetic likelihood for case where the $\boldsymbol{\Phi}_{\boldsymbol{\theta}}$ are (or can be assumed to be) Gamma distributed.

3. Nonparametric approximations and likelihood-free rejection sampling

An alternative to assuming a parametric form for $L(\boldsymbol{\theta})$ is to replace it by a kernel density estimate of the distribution of a discrepancy between simulated and observed summary statistics, i.e.,

$$\tilde{L}(\boldsymbol{\theta}) \equiv \mathrm{E}[\kappa(\Delta_{\boldsymbol{\theta}})], \quad (16)$$

where $\Delta_{\boldsymbol{\theta}}$ is a non-negative function of $\boldsymbol{\Phi}_{\mathrm{O}}$ and $\boldsymbol{\Phi}_{\boldsymbol{\theta}}$ (usually of $\boldsymbol{\Phi}_{\mathrm{O}} - \boldsymbol{\Phi}_{\boldsymbol{\theta}}$) which can also possibly depend on $\boldsymbol{\theta}$ and any variable used internally by the simulator, and the kernel κ is a non-negative, univariate function independent of $\boldsymbol{\theta}$ (usually with a maximum at zero). A computable approximation of the likelihood is then given by

$$\hat{L}^{N}(\boldsymbol{\theta}) \equiv \mathrm{E}^{N}[\kappa(\Delta_{\boldsymbol{\theta}})]. \quad (17)$$

For likelihood-free inference, κ is often chosen as the uniform kernel on the interval $[0, \varepsilon)$, i.e., $\kappa(u) \propto \chi_{[0,\varepsilon)}(u)$, where ε is called the threshold and the indicator function $\chi_{[0,\varepsilon)}$ equals one if $u \in [0, \varepsilon)$ and zero otherwise. This yields

$$\tilde{L}(\boldsymbol{\theta}) \propto \mathcal{P}(\Delta_{\boldsymbol{\theta}} \leq \varepsilon) \quad \text{and} \quad \hat{L}^{N}(\boldsymbol{\theta}) \propto \mathcal{P}^{N}(\Delta_{\boldsymbol{\theta}} \leq \varepsilon), \quad (18)$$

where $\mathcal{P}^{N}(\Delta_{\boldsymbol{\theta}} \leq \varepsilon)$ is the empirical probability that the discrepancy is below the threshold. $\hat{L}^{N}(\boldsymbol{\theta})$ can be straightforwardly evaluated by running simulations, computing $\Delta_{\boldsymbol{\theta}}$ and using $\Delta_{\boldsymbol{\theta}} \leq \varepsilon$ as a criterion for acceptance or rejection of proposed samples. Such an approach is often simply (or mistakenly) referred to as approximate Bayesian computation (ABC) in the astrophysics literature, although the more appropriate and explicit denomination is likelihood-free rejection sampling (see e.g., [1]).

It is interesting to note that the parametric approximate likelihood approach of Sec. II D 2 can be embedded into the nonparametric approach. Indeed, $\Delta_{\boldsymbol{\theta}}$ can be defined as

$$\Delta_{\boldsymbol{\theta}}^{\mathbf{C}_{\boldsymbol{\theta}}} \equiv \log|2\pi\mathbf{C}_{\boldsymbol{\theta}}| + (\boldsymbol{\Phi}_{\mathrm{O}} - \boldsymbol{\Phi}_{\boldsymbol{\theta}})^{\top}\mathbf{C}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\Phi}_{\mathrm{O}} - \boldsymbol{\Phi}_{\boldsymbol{\theta}}) \quad (19)$$

for some positive semidefinite matrix $\mathbf{C}_{\boldsymbol{\theta}}$. The second term is the square of the Mahalanobis distance, which includes the Euclidean distance as a special case, when $\mathbf{C}_{\boldsymbol{\theta}}$ is the identity matrix. Using an exponential kernel $\kappa(u) = \exp(-u/2)$ and $\mathbf{C}_{\boldsymbol{\theta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ gives $\tilde{L}(\boldsymbol{\theta}) = \mathrm{E}[\kappa(\Delta_{\boldsymbol{\theta}}^{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}})]$ and $\hat{L}^{N}(\boldsymbol{\theta}) = \mathrm{E}^{N}[\kappa(\Delta_{\boldsymbol{\theta}}^{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}})]$ with

$$-2\log[\kappa(\Delta_{\boldsymbol{\theta}}^{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}})] = \log|2\pi\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}| + (\boldsymbol{\Phi}_{\mathrm{O}} - \boldsymbol{\Phi}_{\boldsymbol{\theta}})^{\top}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\Phi}_{\mathrm{O}} - \boldsymbol{\Phi}_{\boldsymbol{\theta}}), \quad (20)$$

the form of which is similar to Eq. (15). In fact, Gutmann and Corander [12] (proposition 1) show that the synthetic likelihood satisfies

$$-2\tilde{\ell}(\boldsymbol{\theta}) = J(\boldsymbol{\theta}) + \text{constant}, \quad \text{and} \quad (21)$$

$$-2\hat{\ell}^{N}(\boldsymbol{\theta}) = \hat{J}^{N}(\boldsymbol{\theta}) + \text{constant}, \quad (22)$$

where

$$J(\boldsymbol{\theta}) \equiv \mathrm{E}[\Delta_{\boldsymbol{\theta}}^{\mathbf{C}_{\boldsymbol{\theta}}}] \quad (23)$$

and

$$\hat{J}^{N}(\boldsymbol{\theta}) \equiv \mathrm{E}^{N}[\Delta_{\boldsymbol{\theta}}^{\mathbf{C}_{\boldsymbol{\theta}}}] \quad (24)$$

are respectively the expectation and the empirical average of the discrepancy $\Delta_{\boldsymbol{\theta}}^{\mathbf{C}_{\boldsymbol{\theta}}}$, for $\mathbf{C}_{\boldsymbol{\theta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$.

III. REGRESSION AND OPTIMIZATION FOR LIKELIHOOD-FREE INFERENCE

A. Computational difficulties with likelihood-free rejection sampling

We have seen in Sec. II D that computable approximations $\hat{L}^N(\boldsymbol{\theta})$ of the likelihood $L(\boldsymbol{\theta})$ are stochastic processes, due to the use of simulations to approximate intractable expectations. In the most popular ABC approach, i.e., likelihood-free rejection sampling (see Sec. II D 3), the expectations are approximated by empirical probabilities that the discrepancy is below the threshold ε. While this approach allows inference of simulator-based statistical models with minimal assumptions, it suffers from several limitations that can make its use impossible in practice.

(1) It rejects most of the proposed samples when ε is small, leading to a computationally inefficient algorithm.
(2) It does not make assumptions about the shape or smoothness of the target function $L(\boldsymbol{\theta})$, hence accepted samples cannot "share" information in parameter space.
(3) It uses a fixed proposal distribution [typically the prior $\mathcal{P}(\boldsymbol{\theta})$] and does not make use of already accepted samples to update the proposal of new points.
(4) It aims at equal accuracy for all regions in parameter space, regardless of the values of the likelihood.

To overcome these issues, the proposed approach follows closely Gutmann and Corander [12], who combine regression of the discrepancy (addressing issues 1 and 2) with Bayesian optimization (addressing issues 3 and 4) in order to improve the computational efficiency of inference of simulator-based models. In this work, we focus on parametric approximations of the likelihood; we refer to Gutmann and Corander [12] for a treatment of the nonparametric approach.

B. Regression of the discrepancy

The standard approach to obtain a computable approximate likelihood relies on empirical averages [Eqs. (14) and (24)]. However, such sample averages are not the only way to approximate intractable expectations. Equations (21) and (23) show that, up to constants and the sign, $\tilde{\mathscr{E}}(\boldsymbol{\theta})$ can be interpreted as a regression function with the model parameters $\boldsymbol{\theta}$ (the "predictors") as the independent input variables and the discrepancy $\boldsymbol{\Delta}_{\boldsymbol{\theta}}$ as the response variable. Therefore, in the present approach, we consider an approximation of the intractable expectation defining $J(\boldsymbol{\theta})$ in Eq. (23) based on a regression analysis of $\boldsymbol{\Delta}_{\boldsymbol{\theta}}$, instead of sample averages. Explicitly, we consider

$$\hat{J}^{(\mathrm{t})}(\boldsymbol{\theta}) \equiv \mathrm{E}^{(\mathrm{t})}[\boldsymbol{\Delta}_{\boldsymbol{\theta}}^{\mathbf{C}_{\boldsymbol{\theta}}}], \tag{25}$$

where the superscript (t) stands for "training" and the expectation $\mathrm{E}^{(\mathrm{t})}$ is taken under the probabilistic model defined in the following.

Inferring $J(\boldsymbol{\theta})$ via regression requires a training data set $\{(\boldsymbol{\theta}^{(i)}, \boldsymbol{\Delta}_{\boldsymbol{\theta}}^{(i)})\}$ where the discrepancies are computed from the simulated summary statistics $\boldsymbol{\Phi}_{\boldsymbol{\theta}}^{(i)}$. Building this training set requires to run simulations, but does not involve an accept/reject criterion as does likelihood-free rejection sampling (thus addressing issue 1, see Sec. III A). A regression-based approach also allows incorporating a smoothness assumption about $J(\boldsymbol{\theta})$. In this way, samples of the training set can share the information of the computed $\boldsymbol{\Delta}_{\boldsymbol{\theta}}$ in the neighborhood of $\boldsymbol{\theta}$ (thus addressing issue 2). This suggests that fewer simulated data are needed to reach a certain level of accuracy when learning the target function $J(\boldsymbol{\theta})$.

In this work, we rely on Gaussian process (GP) regression in order to construct a prediction for $J(\boldsymbol{\theta})$. There are several reasons why this choice is advantageous for likelihood-free inference. First, GPs are a general-purpose regressor, able to deal with a large variety of functional shapes for $J(\boldsymbol{\theta})$, including potentially complex nonlinear, or multimodal features. Second, GPs provide not only a prediction (the mean of the regressed function), but also the uncertainty of the regression. This is useful for actively constructing the training data via Bayesian optimization, as we show in Sec. III E. Finally, GPs allow extrapolating the prediction into regions of the parameter space where no training points are available. These three properties are shown in Fig. 3 for a multimodal test function subject to observation noise.

We now briefly review Gaussian process regression. Suppose that we have a set of t training points,

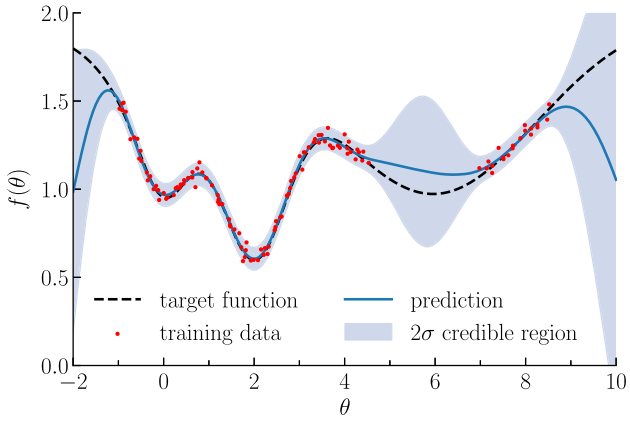

FIG. 3. Illustration of Gaussian process regression in one dimension, for the target test function $f:\theta \mapsto 2 - \exp[-(\theta-2)^2] - \exp[-(\theta-6)^2/10] - 1/(\theta^2+1)$ (dashed line). Training data are acquired (red dots); they are subject to a Gaussian observation noise with standard deviation $\sigma_{\mathrm{n}} = 0.03$. The blue line shows the mean prediction $\mu(\theta)$ of the Gaussian process regression, and the shaded region the corresponding $2\sigma(\theta)$ uncertainty. Gaussian processes allow interpolating and extrapolating predictions in regions of parameter space where training data are absent.

$(\boldsymbol{\Theta}, \mathbf{f}) \equiv \{(\boldsymbol{\theta}^{(i)}, f^{(i)} = f(\boldsymbol{\theta}^{(i)})\}$, of the function f that we want to regress. We assume that f is a Gaussian process with prior mean function $m(\boldsymbol{\theta})$ and covariance function $\kappa(\boldsymbol{\theta}, \boldsymbol{\theta}')$ also known as the kernel (see [21]). The joint probability distribution of the training set is therefore $\mathcal{P}(\mathbf{f}|\boldsymbol{\Theta}) \propto \exp[\ell(\mathbf{f}|\boldsymbol{\Theta})]$, where the exponent $\ell(\mathbf{f}|\boldsymbol{\Theta})$ is

$$-\frac{1}{2}\sum_{i,j=1}^{t}[f^{(i)} - m(\boldsymbol{\theta}^{(i)})]^{\top}\kappa(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(j)})^{-1}[f^{(j)} - m(\boldsymbol{\theta}^{(j)})]. \quad (26)$$

The mean function $m(\boldsymbol{\theta})$ and the kernel $\kappa(\boldsymbol{\theta}, \boldsymbol{\theta}')$ define the functional shape and smoothness allowed for the prediction. Standard choices are respectively a constant and a squared exponential (the radial basis function), subject to additive Gaussian observation noise with variance σ_{n}^2. Explicitly, $m(\boldsymbol{\theta}) \equiv C$ and

$$\kappa(\boldsymbol{\theta}, \boldsymbol{\theta}') \equiv \sigma_f^2 \exp\left[-\frac{1}{2}\sum_p \left(\frac{\theta_p - \theta_p'}{\lambda_p}\right)^2\right] + \sigma_{\mathrm{n}}^2 \delta_{\mathrm{K}}(\boldsymbol{\theta}, \boldsymbol{\theta}'). \quad (27)$$

The θ_p and θ_p' are the components of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, respectively. In the last term, $\delta_{\mathrm{K}}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is one if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}'$ and zero otherwise. The hyperparameters are C, the λ_p (the length scales controlling the amount of correlation between points, and hence the allowed wiggliness of f), σ_f^2 (the signal variance, i.e., the marginal variance of f at a point $\boldsymbol{\theta}$ if the observation noise was zero), and σ_{n}^2 (the observation noise). For the results of this paper, GP hyperparameters were learned from the training set using L-BFGS [22], a popular optimizer for machine learning, and updated every time the training set was augmented with ten samples.

The predicted value $f_\star$ at a new point $\boldsymbol{\theta}_\star$ can be obtained from the fact that $(\{\boldsymbol{\Theta}, \boldsymbol{\theta}_\star\}, \{\mathbf{f}, f_\star\})$ form jointly a random realization of the Gaussian process f. Thus, the target pdf $\mathcal{P}(f_\star|\mathbf{f}, \boldsymbol{\Theta}, \boldsymbol{\theta}_\star)$ can be obtained from conditioning the joint pdf $\mathcal{P}(\mathbf{f}, f_\star|\boldsymbol{\Theta}, \boldsymbol{\theta}_\star)$ to the values of the training set $\mathbf{f}$. The result is (see [21], Sec. II.7)

$$\mathcal{P}(f_\star|\mathbf{f}, \boldsymbol{\Theta}, \boldsymbol{\theta}_\star) \propto \exp\left[-\frac{1}{2}\left(\frac{f_\star - \mu(\boldsymbol{\theta}_\star)}{\sigma(\boldsymbol{\theta}_\star)}\right)^2\right], \quad (28)$$

$$\mu(\boldsymbol{\theta}_\star) \equiv m(\boldsymbol{\theta}_\star) + \underline{\mathbf{K}}_\star^{\intercal}\underline{\underline{\mathbf{K}}}^{-1}(\mathbf{f} - \mathbf{m}), \quad (29)$$

$$\sigma^2(\boldsymbol{\theta}_\star) \equiv K_{\star\star} - \underline{\mathbf{K}}_\star^{\intercal}\underline{\underline{\mathbf{K}}}^{-1}\underline{\mathbf{K}}_\star, \quad (30)$$

where we use the definitions

$$K_{\star\star} \equiv \kappa(\boldsymbol{\theta}_\star, \boldsymbol{\theta}_\star), \quad (31)$$

$$\mathbf{m} \equiv (m(\boldsymbol{\theta}^{(i)}))^{\intercal} \quad \text{for } \boldsymbol{\theta}^{(i)} \in \boldsymbol{\Theta}, \quad (32)$$

$$\underline{\mathbf{K}}_\star \equiv (\kappa(\boldsymbol{\theta}_\star, \boldsymbol{\theta}^{(i)}))^{\intercal} \quad \text{for } \boldsymbol{\theta}^{(i)} \in \boldsymbol{\Theta}, \quad (33)$$

$$(\underline{\underline{\mathbf{K}}})_{ij} \equiv \kappa(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(j)}) \quad \text{for } \{\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(j)}\} \in \boldsymbol{\Theta}^2. \quad (34)$$

C. Bayesian optimization

The second major ingredient of the proposed approach is Bayesian optimization, which allows the inference of the regression function $J(\boldsymbol{\theta})$ while avoiding unnecessary computations. It allows active construction of the training data set $\{(\boldsymbol{\theta}^{(i)}, \Delta_{\boldsymbol{\theta}}^{(i)})\}$, updating the proposal of new points using the regressed $\hat{J}^{(\mathrm{t})}(\boldsymbol{\theta})$ (thus addressing issue 3 with likelihood-free rejection sampling, see Sec. III A). Further, since we are mostly interested in the regions of the parameter space where the variance of the approximate posterior is large (due to its stochasticity), the acquisition rules can prioritize these regions, so as to obtain a better approximation of $J(\boldsymbol{\theta})$ there (thus addressing issue 4).

Bayesian optimization is a decision-making framework under uncertainty, for the automatic learning of unknown functions. It aims at gathering training data in such a manner as to evaluate the regression model the least number of times while revealing as much information as possible about the target function and, in particular, the location of the optimum or optima. The method proceeds by iteratively picking predictors to be probed (i.e., simulations to be run) in a manner that trades off *exploration* (parameters for which the outcome is most uncertain) and *exploitation* (parameters which are expected to have a good outcome for the targeted application). In many contexts, Bayesian optimization has been shown to obtain better results with fewer simulations than grid search or random search, due to its ability to reason about the interest of simulations before they are run (see [23] for a review). Figure 4 illustrates Bayesian optimization in combination with Gaussian process regression, applied to finding the minimum of the test function of Fig. 3.

In the following, we give a brief overview of the elements of Bayesian optimization used in this paper. In order to add a new point to the training data set $(\boldsymbol{\Theta}, \mathbf{f}) \equiv \{(\boldsymbol{\theta}^{(i)}, f^{(i)} = f(\boldsymbol{\theta}^{(i)})\}$, Bayesian optimization uses an acquisition function $\mathcal{A}(\boldsymbol{\theta})$ that estimates how useful the evaluation of the simulator at $\boldsymbol{\theta}$ will be in order to learn the target function. The acquisition function is constructed from the posterior predictive distribution of f given the training set $(\boldsymbol{\Theta}, \mathbf{f})$, i.e., from the mean prediction $\mu(\boldsymbol{\theta})$ and the uncertainty $\sigma(\boldsymbol{\theta})$ of the regression analysis [Eqs. (29) and (30)]. The optimum of the acquisition function in parameter space determines the next point $\boldsymbol{\theta}_\star \equiv \mathrm{argopt}_{\boldsymbol{\theta}}\mathcal{A}(\boldsymbol{\theta})$ to be evaluated by the simulator (argopt = argmax or argmin depending on how the acquisition function is defined), so that the training set can be augmented with $(\boldsymbol{\theta}_\star, f(\boldsymbol{\theta}_\star))$. The acquisition function is a scalar function whose evaluation should be reasonably expensive, so that its optimum can be found by simple search methods such as gradient descent.

The algorithm needs to be initialized with an initial training set. In numerical experiments, we found that building this initial set by drawing from the prior (as would typically be done in likelihood-free rejection sampling) can result in difficulties with the first iterations of Gaussian

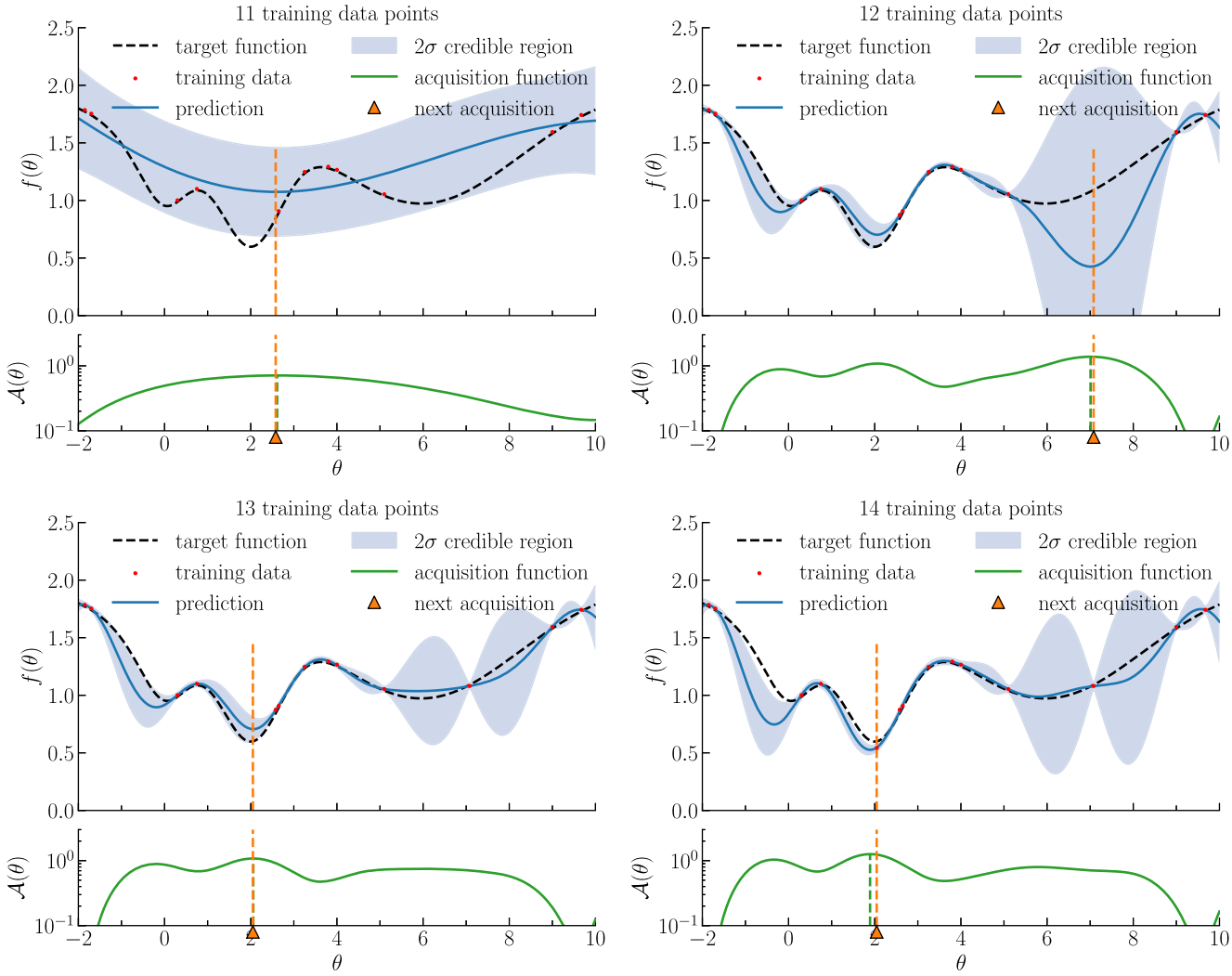

FIG. 4. Illustration of four consecutive steps of Bayesian optimization to learn the test function of Fig. 3. For each step, the top panel shows the training data points (red dots) and the regression (blue line and shaded region). The bottom panel shows the acquisition function (the expected improvement, solid green line) with its maximiser (dashed green line). The next acquisition point, i.e., where to run a simulation to be added to the training set, is shown in orange; it differs from the maximiser of the acquisition function by a small random number. The acquisition function used is the expected improvement, aiming at finding the minimum of f. Hyperparameters of the regression kernel are optimised after each acquisition. As can observed, Bayesian optimization implements a trade-off between exploration (evaluation of the target function where the variance is large, e.g., after 12 points) and exploitation (evaluation of the target function close to the predicted minimum, e.g., after 11, 13, and 14 points).

process regression. Uniformly distributed points within the boundaries of the GP are also a poor choice, as they will result in an uneven initial sampling of the parameter space. To circumvent this issue, we build the initial training set using a low-discrepancy quasirandom Sobol sequence [24], which covers the parameter space more evenly.

D. Expressions for the approximate posterior

As discussed in Sec. III B, using $\Delta_{\boldsymbol{\theta}}^{\mathbf{C}_{\boldsymbol{\theta}}}$ as the regressed quantity directly gives an estimate of $J(\boldsymbol{\theta})$ in Eq. (23). The response variable is thus $f(\boldsymbol{\theta}) \equiv \Delta_{\boldsymbol{\theta}}^{\mathbf{C}_{\boldsymbol{\theta}}}$ and the regression then gives

$$\hat{J}^{(t)}(\boldsymbol{\theta}) = \mu(\boldsymbol{\theta}). \tag{35}$$

In the parametric approach to likelihood approximation, this is equivalent to an approximation of $-2\tilde{\ell}(\boldsymbol{\theta}) = -2\log\tilde{L}(\boldsymbol{\theta})$ [see Eq. (21)]. The expectation of the (unnormalized) approximate posterior is therefore directly given as [see Eq. (5)]

$$\mathrm{E}^{(t)}[\mathcal{P}_{\mathrm{BOLFI}}(\boldsymbol{\theta}|\boldsymbol{\Phi}_{\mathrm{O}}, \mathbf{f}, \boldsymbol{\Theta})] \equiv \mathcal{P}(\boldsymbol{\theta}) \exp\left(-\frac{1}{2}\mu(\boldsymbol{\theta})\right), \tag{36}$$

where $\mathcal{P}_{\mathrm{BOLFI}}(\boldsymbol{\theta}|\boldsymbol{\Phi}_{\mathrm{O}}, \mathbf{f}, \boldsymbol{\Theta}) \approx Z_{\boldsymbol{\Phi}} \times \mathcal{P}(\boldsymbol{\theta}|\boldsymbol{\Phi})_{|\boldsymbol{\Phi}=\boldsymbol{\Phi}_{\mathrm{O}}}$.

The estimate of the variance of $f(\boldsymbol{\theta})$ can also be propagated to the approximate posterior, giving

$$\mathrm{V}^{(\mathrm{t})}[\mathcal{P}_{\mathrm{BOLFI}}(\boldsymbol{\theta}|\boldsymbol{\Phi}_{\mathrm{O}},\mathbf{f},\boldsymbol{\Theta})] \equiv \frac{\mathcal{P}(\boldsymbol{\theta})^2}{4}\exp\left[-\mu(\boldsymbol{\theta})\right]\sigma^2(\boldsymbol{\theta}). \quad (37)$$

Details of the computations can be found in Appendix A 1.

Expressions for the BOLFI posterior in the nonparametric approach with the uniform kernel can also be derived (see [16], lemma 3.1). As this paper focuses on the parametric approach, we refer to the literature for the former case.

E. Acquisition rules

1. Expected improvement

Standard Bayesian optimization uses acquisition functions that estimate how useful the next evaluation of the simulator will be in order to find the minimum or minima of the target function. While several other choices are possible (see e.g., [23]), in this work we discuss the acquisition function known as *expected improvement* (EI). The *improvement* is defined by $I(\boldsymbol{\theta}_\star) = \max\left[\min(\mathbf{f}) - f(\boldsymbol{\theta}_\star), 0\right]$, and the expected improvement is $\mathrm{EI}(\boldsymbol{\theta}_\star) \equiv \mathrm{E}^{(\mathrm{t})}[I(\boldsymbol{\theta}_\star)]$, where the expectation is taken with respect to the random observation assuming decision $\boldsymbol{\theta}_\star$. For a Gaussian process regressor, this evaluates to (see [23], Sec. 2.3)

$$\mathrm{EI}(\boldsymbol{\theta}_\star) \equiv \sigma(\boldsymbol{\theta}_\star)[z\Phi(z) + \phi(z)], \quad \text{with} \quad z \equiv \frac{\min(\mathbf{f}) - \mu(\boldsymbol{\theta}_\star)}{\sigma(\boldsymbol{\theta}_\star)}, \quad (38)$$

or $\mathrm{EI}(\boldsymbol{\theta}_\star) \equiv 0$ if $\sigma(\boldsymbol{\theta}_\star) = 0$, where ϕ and Φ denote respectively the pdf and the cumulative distribution function (cdf) of the unit-variance zero-mean Gaussian. The decision rule is to select the location $\boldsymbol{\theta}_\star$ that maximizes $\mathrm{EI}(\boldsymbol{\theta}_\star)$.

The EI criterion can be interpreted as follows: since the goal is to find the minimum of f, a reward equal to the improvement $\min(\mathbf{f}) - f(\boldsymbol{\theta}_\star)$ is received if $f(\boldsymbol{\theta}_\star)$ is smaller than all the values observed so far, otherwise no reward is received. The first term appearing in Eq. (38) is maximized when evaluating at points with high uncertainty (exploration); and, at fixed variance, the second term is maximized by evaluating at points with low mean (exploitation). The expected improvement therefore automatically captures the exploration-exploitation trade-off as a result of the Bayesian decision-theoretic treatment.

2. Expected integrated variance

As pointed out by Jäarvenpää *et al.* [16], in Bayesian optimization for approximate Bayesian computation, the goal should not be to find the minimum of $J(\boldsymbol{\theta})$, but rather to minimize the expected uncertainty in the estimate of the approximate posterior over the future evaluation of the simulator at $\boldsymbol{\theta}_\star$. Consequently, they propose an acquisition function, known as the *expected integrated variance* (ExpIntVar or EIV in the following) that selects the next evaluation location to minimize the expected variance of the future posterior density $\mathcal{P}_{\mathrm{BOLFI}}(\boldsymbol{\theta}|\boldsymbol{\Phi}_{\mathrm{O}},\mathbf{f},\boldsymbol{\Theta},\boldsymbol{\theta}_\star)$ over the parameter space. The framework used is Bayesian decision theory. Formally, the loss due to our uncertain knowledge of the approximate posterior density can be defined as

$$\mathcal{L}[\mathcal{P}_{\mathrm{BOLFI}}(\boldsymbol{\theta}|\boldsymbol{\Phi}_{\mathrm{O}},\mathbf{f},\boldsymbol{\Theta})] = \int \mathrm{V}^{(\mathrm{t})}[\mathcal{P}_{\mathrm{BOLFI}}(\boldsymbol{\theta}|\boldsymbol{\Phi}_{\mathrm{O}},\mathbf{f},\boldsymbol{\Theta})]\mathrm{d}\boldsymbol{\theta}, \quad (39)$$

and the acquisition rule is to select the location $\boldsymbol{\theta}_\star$ that minimizes

$$\begin{aligned}\mathrm{EIV}(\boldsymbol{\theta}_\star) &\equiv \mathrm{E}^{(\mathrm{t})}[\mathcal{L}[\mathcal{P}_{\mathrm{BOLFI}}(\boldsymbol{\theta}|\boldsymbol{\Phi}_{\mathrm{O}},\mathbf{f},\boldsymbol{\Theta},f_\star,\boldsymbol{\theta}_\star)]] \\ &= \int \mathcal{L}[\mathcal{P}_{\mathrm{BOLFI}}(\boldsymbol{\theta}|\boldsymbol{\Phi}_{\mathrm{O}},\mathbf{f},\boldsymbol{\Theta},f_\star,\boldsymbol{\theta}_\star)] \\ &\quad \times \mathcal{P}(f_\star|\mathbf{f},\boldsymbol{\Theta},\boldsymbol{\theta}_\star)\mathrm{d}f_\star \end{aligned} \quad (40)$$

with respect to $\boldsymbol{\theta}_\star$, where we have to marginalize over the unknown simulator output $f_\star$ using the probabilistic model $\mathcal{P}(f_\star|\mathbf{f},\boldsymbol{\Theta},\boldsymbol{\theta}_\star)$ [Eqs. (28)–(30)].

Jäarvenpää *et al.* [16] (proposition 3.2) derive the expressions for the expected integrated variance for a GP model in the nonparametric approach. In Appendix A, we extend this work and derive the ExpIntVar acquisition function and its gradient in the parametric approach. The result is the following: under the GP model, the expected integrated variance after running the simulation model with parameter $\boldsymbol{\theta}_\star$ is given by

$$\mathrm{EIV}(\boldsymbol{\theta}_\star) = \int \frac{\mathcal{P}(\boldsymbol{\theta})^2}{4}\exp\left[-\mu(\boldsymbol{\theta})\right][\sigma^2(\boldsymbol{\theta}) - \tau^2(\boldsymbol{\theta},\boldsymbol{\theta}_\star)]\mathrm{d}\boldsymbol{\theta}, \quad (41)$$

with

$$\tau^2(\boldsymbol{\theta},\boldsymbol{\theta}_\star) \equiv \frac{\mathrm{cov}^2(\boldsymbol{\theta},\boldsymbol{\theta}_\star)}{\sigma^2(\boldsymbol{\theta}_\star)}, \quad (42)$$

where $\mathrm{cov}(\boldsymbol{\theta},\boldsymbol{\theta}_\star) \equiv \kappa(\boldsymbol{\theta},\boldsymbol{\theta}_\star) - \underline{\mathbf{K}}^{\intercal}\underline{\underline{\mathbf{K}}}^{-1}\underline{\mathbf{K}}_\star$ is the GP posterior predicted covariance between the evaluation point $\boldsymbol{\theta}$ in the integral and the candidate location for the next evaluation $\boldsymbol{\theta}_\star$. Note that in addition to the notations given by Eqs. (31)–(34), we have introduced the vector

$$\underline{\mathbf{K}} \equiv (\kappa(\boldsymbol{\theta},\boldsymbol{\theta}^{(i)}))^{\intercal} \quad \text{for } \boldsymbol{\theta}^{(i)} \in \boldsymbol{\Theta}. \quad (43)$$

It is of interest to examine when the integrand in Eq. (41) is small. As for the EI [Eq. (38)], optimal values are found when the mean of the discrepancy $\mu(\boldsymbol{\theta})$ is small or the variance $\sigma^2(\boldsymbol{\theta})$ is large. This effect is what yields the trade-off between exploitation and exploration for the ExpIntVar acquisition rule. However, unlike in standard Bayesian optimization strategies such as the EI, the trade-off is a

nonlocal process (due to the integration over the parameter space), and also depends on the prior, so as to minimize the uncertainty in the posterior (and not likelihood) approximation.

Computing the expected integrated variance requires integration over the parameter space. In this work, the integration is performed on a regular grid of 50 points per dimension within the GP boundaries. In high dimension, the integral can become prohibitively expensive to compute on a grid. As discussed by Jäarvenpää *et al.* [16], it can then be evaluated with Monte Carlo or quasi–Monte Carlo methods such as importance sampling.

In numerical experiments, we have found that the ExpIntVar criterion (as any acquisition function for Bayesian optimization) has some sensitivity to the initial training set. In particular, the initial set (built from a Sobol sequence or otherwise) shall sample sufficiently well the GP domain, which shall encompass the prior. This ensures that the prior volume is never wider than the training data. Under this condition, as Jäarvenpää *et al.* [16], we have found that ExpIntVar is stable, in the sense that it produces consistent BOLFI posteriors over different realizations of the initial training data set and simulator outputs.

3. Stochastic versus deterministic acquisition rules

The above rules do not guarantee that the selected $\boldsymbol{\theta}_\star$ is different from a previously acquired $\boldsymbol{\theta}^{(i)}$. Gutmann and Corander [12] (see in particular Appendix C) found that this can result in a poor exploration of the parameter space, and propose to add a stochastic element to the decision rule in order to avoid getting stuck at one point. In some experiments, we followed this prescription by adding an "acquisition noise" of strength $\sigma_{\rm a}^p$ to each component of the optimizer of the acquisition function. More precisely, $\boldsymbol{\theta}_\star$ is sampled from the Gaussian distribution $\mathcal{G}(\boldsymbol{\theta}_{\rm opt}, \mathbf{D})$, where $\boldsymbol{\theta}_{\rm opt} \equiv \mathrm{argopt}_{\boldsymbol{\theta}} \mathcal{A}(\boldsymbol{\theta})$ and $\mathbf{D}$ is the diagonal covariance matrix of components $(\sigma_{\rm a}^p)^2$. The $\sigma_{\rm a}^p$ are chosen to be of order $\lambda_p/10$.

For a more extensive discussion and comparison of various stochastic and deterministic acquisition rules, the reader is referred to Järvenpää *et al.* [16].

IV. APPLICATIONS

In this section, we show the application of BOLFI to several application studies. In particular, we discuss the simulator and the computable approximation of the likelihood to be used, and compare BOLFI to likelihood-free rejection sampling in terms of computational efficiency. In all cases, we show that BOLFI reduces the amount of required simulations by several orders of magnitude.

In Sec. IV A, we discuss the toy problem of summarizing Gaussian signals (i.e., inferring the unknown mean and/or variance of Gaussian-distributed data). In Sec. IV B, we show the first application of BOLFI to a real cosmological problem using actual observational data: the inference of cosmological parameters from supernovae data. For each test case, we refer to the corresponding section in the Appendices for the details of the data model and inference assumptions.

A. Summarizing Gaussian signals

A simple toy model can be constructed from the general problem of summarizing Gaussian signals with unknown mean, or with unknown mean and variance. This example allows for the comparison of BOLFI and likelihood-free rejection sampling to the true posterior conditional on the full data, which is known analytically. All the details of this model are given in Appendix B.

1. Unknown mean, known variance

We first consider the problem, already discussed by Gutmann and Corander [12], where the data $\mathbf{d}$ are a vector of n components drawn from a Gaussian with unknown mean μ and known variance $\sigma^2_{\rm true}$. The empirical mean Φ^1 is a sufficient summary statistic for the problem of inferring μ. The distribution of simulated Φ^1_μ takes a simple form, $\Phi^1_\mu \sim \mathcal{G}(\mu, \sigma^2_{\rm true}/n)$. Using here the true variance, the discrepancy and synthetic likelihood are

$$\Delta^1_\mu = -2\hat{\ell}^N_1(\mu) = \log\left(\frac{2\pi\sigma^2_{\rm true}}{n}\right) + n\frac{(\Phi^1_{\rm O} - \hat{\mu}^1_\mu)^2}{\sigma^2_{\rm true}}, \tag{44}$$

where $\hat{\mu}^1_\mu$ is an average of N realizations of Φ^1_μ. In Fig. 5 (lower panel), the black dots show simulations of Δ^1_μ for different values of μ. We have $\hat{\mu}^1_\mu \sim \mathcal{G}(\mu, \sigma^2_{\rm true}/(Nn))$, therefore the stochastic process defining the discrepancy can be written

$$\Delta^1_\mu = \log\left(\frac{2\pi\sigma^2_{\rm true}}{n}\right) + n\frac{(\Phi^1_{\rm O} - \mu - g)^2}{\sigma^2_{\rm true}}, \qquad g \sim \mathcal{G}(0, \sigma^2_g), \tag{45}$$

where $\sigma^2_g \equiv \sigma^2_{\rm true}/(Nn)$. Each realization of g gives a different mapping $\mu \mapsto \Delta^1_\mu$. In Fig. 5, we show one such realization in the lower panel, and the corresponding approximate posterior in the upper panel. Using the percent point function (inverse of the cdf) of the Gaussian $\mathcal{G}(0, \sigma^2_g)$, we also show in red the mean and 2σ credible interval of the true stochastic process.

The GP regression using the simulations shown as the training set is represented in blue in the lower panel of Fig. 5. The corresponding BOLFI posterior and its variance, defined by Eqs. (36) and (37), are shown in purple in the upper panel. The uncertainty in the estimate of the posterior (shaded purple region) is due to the limited number of available simulations (and not to the noisiness of individual training points). It is the expectation of this uncertainty under the next

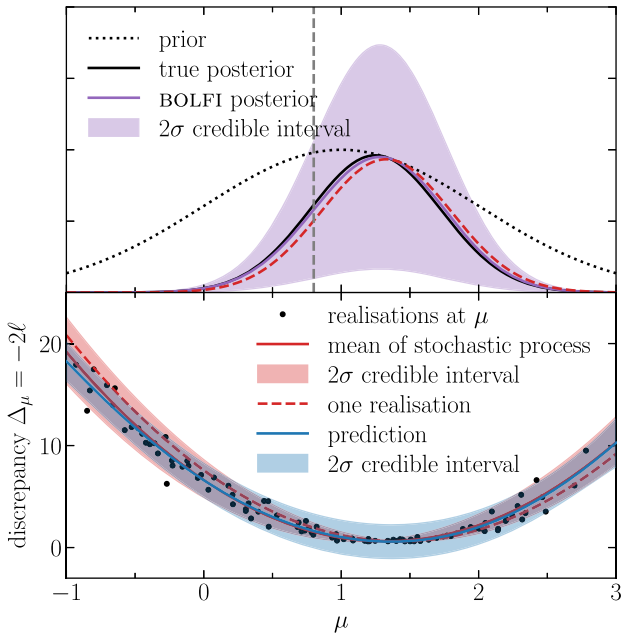

FIG. 5. Illustration of BOLFI for a one-dimensional problem, the inference of the unknown mean μ of a Gaussian. Lower panel: The discrepancy Δ_μ (i.e., twice the negative loglikelihood) is a stochastic process due to the limited computational resources. Its mean and the 2σ credible interval are shown in red. The dashed red line shows one realization of the stochastic process as a function of μ. Simulations at different μ are shown as black dots. BOLFI builds a probabilistic model for the discrepancy, the mean and 2σ credible interval of which are shown in blue. Upper panel: The expectation of the (rescaled) BOLFI posterior and its 2σ credible interval are shown in comparison to the exact posterior for the problem. The dashed red line shows the posterior obtained from the corresponding realization of the stochastic process of the lower panel.

evaluation of the simulator which is minimized in parameter space by the ExpIntVar acquisition rule.

2. Unknown mean and variance

We now consider the problem where the full data set $\mathbf{d}$ is a vector of n components drawn from a Gaussian with unknown mean μ and unknown variance σ^2. The aim is the two-dimensional inference of $\boldsymbol{\theta} \equiv (\mu, \sigma^2)$. Evidently, the true likelihood $\mathcal{L}(\mu, \sigma^2)$ for this problem is the Gaussian characterized by (μ, σ^2). The Gaussian-inverse-Gamma distribution is the conjugate prior for this likelihood. It is described by four parameters. Adopting a Gaussian-inverse-Gamma prior characterized by $(\alpha, \beta, \eta, \lambda)$ yields a Gaussian-inverse-Gamma posterior characterized by $(\alpha', \beta', \eta', \lambda')$ given by Eqs. (B8)–(B11). This is the analytic solution to which we compare our approximate results.

For the numerical approach, we forward model the problem using a simulator that draws from the prior, simulates $N = 10$ realizations of the Gaussian signal, and compresses them to two summary statistics, the empirical mean and variance, respectively Φ^1 and Φ^2. The graphical probabilistic model is given in Fig. 8. It is a noise-free simulator without latent variables (of the type given by Fig. 1, right) completed by a deterministic compression of the full data. Note that the vector $\boldsymbol{\Phi} \equiv (\Phi^1, \Phi^2)$ is a sufficient statistic for the inference of (μ, σ^2). To perform likelihood-free inference, we also need a computable approximation $\hat{L}^N(\mu, \sigma^2)$ of the true likelihood. We derive such an approximation in Sec. B 3 using a parametric approach, under the assumptions (exactly verified in this example) that Φ^1 is Gaussian distributed and Φ^2 is Gamma distributed. We name it the Gaussian-Gamma synthetic likelihood.

The posterior obtained from likelihood-free rejection sampling is shown in green in Fig. 6 (left) in comparison to the prior (in blue) and the analytic posterior (in orange). It was obtained from 5,000 accepted samples using a threshold of $\varepsilon = 4$ on $-2\hat{\ell}^N$. The entire run required $\sim 350,000$ forward simulations in total, the vast majority of which have been rejected. The rejection-sampling posterior is a fair approximation to the true posterior, unbiased but broader, as expected from a rejection-sampling method.

For comparison, the posterior obtained via BOLFI is shown in red in Fig. 6 (right). BOLFI was initialized using a Sobol sequence of 20 members to compute the original surrogate surface, and Bayesian optimization with the ExpIntVar acquisition function and acquisition noise was run to acquire 230 more samples. As can be observed, BOLFI allows very precise likelihood-free inference; in particular, the 1σ, 2σ and 3σ contours (the latter corresponding to the 0.27% least likely events) of the analytic posterior are reconstructed almost perfectly. The overall cost to get these results is only 2,500 simulations with BOLFI versus $\sim 350,000$ with rejection sampling (for a poorer approximation of the analytic posterior), which corresponds to a reduction by 2 orders of magnitude.

B. Supernova cosmology

In this section, we present the first application of BOLFI to a cosmological inference problem. Specifically, we perform an analysis of the joint lightcurve analysis (JLA) data set, consisting of the B-band peak apparent magnitudes m_B of 740 type Ia supernovae (SN Ia) with redshift z between 0.01 and 1.3 [17]: $\mathbf{d}_\mathrm{O} \equiv (m^k_\mathrm{B,O})$ for $k \in [\![1, 740]\!]$. The details of the data model and inference assumptions are given in Appendix C. For the purpose of validating BOLFI, we assume a Gaussian synthetic likelihood (see Sec. C 4), allowing us to demonstrate the fidelity of the BOLFI posterior against the exact likelihood-based solution obtained via Markov chain Monte Carlo (MCMC). This analysis can also be compared to the proof of concept for another likelihood-free method, density estimation for likelihood-free inference (DELFI) [25,26], as the assumptions are very similar.

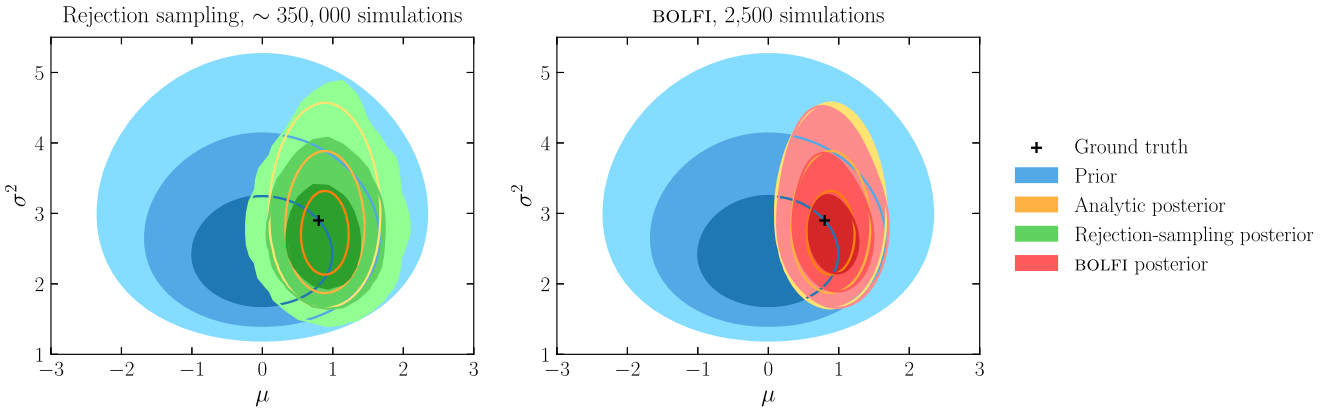

FIG. 6. Prior and posterior for the joint inference of the mean and variance of Gaussian signals. The prior and exact posterior (from the analytic solution) are Gaussian-inverse-Gamma distributed and shown in blue and orange, respectively. In the left panel, the approximate rejection-sampling posterior, based on 5,000 samples accepted out of ∼350, 000 simulations, is shown in green. It loosely encloses the exact posterior. In the right panel, the approximate BOLFI posterior, based on 2,500 simulations only, is shown in red. It is a much finer approximation of the exact posterior. For all distributions, the 1σ, 2σ and 3σ contours are shown.

As described in Appendix C, the full problem is six dimensional; however, in this work, we focus on the inference of the two physically relevant quantities, namely Ω_m (the matter density of the Universe) and w (the equation of state of dark energy, assumed constant), and marginalize over the other four (nuisance) parameters (α, β, M_B, δM). We assume a Gaussian prior,

$$\begin{pmatrix} \Omega_m \\ w \end{pmatrix} \sim \mathcal{G}\left[\begin{pmatrix} 0.3 \\ -0.75 \end{pmatrix}, \begin{pmatrix} 0.4^2 & -0.24 \\ -0.24 & 0.75^2 \end{pmatrix}\right], \tag{46}$$

which is roughly aligned with the direction of the well-known $\Omega_m - w$ degeneracy. We generated 10^6 samples (out of $\sim 6 \times 10^6$ data model evaluations) of the posterior for the exact six-dimensional Bayesian problem via MCMC (performed using the EMCEE code, [27]), ensuring sufficient convergence to characterize the 3σ contours of the distribution.[1] The prior and the exact posterior are shown in blue and orange, respectively, in Fig. 7.

For likelihood-free inference, the simulator takes as input Ω_m and w and simulates N realizations of the magnitudes m_B of the 740 supernovae at their redshifts. Consistently with the Gaussian likelihood used in the MCMC analysis, we assume a Gaussian synthetic likelihood with a fixed covariance matrix $\mathbf{C}$. The observed data $\mathbf{d}_O$ and the covariance matrix $\mathbf{C}$ are shown in Fig. 10.

The approximate posterior obtained from likelihood-free rejection sampling is shown in green in Fig. 7. It was obtained from 5,000 accepted samples using a (conservative) threshold of $\varepsilon = 650$ on $\Delta_{(\Omega_m, w)}$, chosen so that the acceptance ratio was not below 0.01. The entire run required ∼450, 000 simulations in total. The approximate posterior obtained via BOLFI is shown in red in Fig. 7. BOLFI was initialized with a Sobol sequence of 20 samples, and 100 acquisitions were performed according to the ExpIntVar criterion, without acquisition noise. The BOLFI posterior is a much finer approximation to the true posterior than the one obtained from likelihood-free rejection sampling. It is remarkable that only 100 acquisitions are enough to learn the nontrivial banana shape of the posterior. Only the 3σ contour, which is usually not shown in cosmology papers (e.g., [17]), notably deviates from the MCMC posterior. This is due to the fact that we used one realization of the stochastic process defining $\Delta_{(\Omega_m, w)}$ and only $N = 50$ realizations per (Ω_m, w); the marginalization over the four nuisance parameters is therefore partial, yielding slightly smaller credible contours. However, a better approximation could be obtained straightfowardly, if desired, by investing more computational resources (increasing N), without requiring more acquisitions.

As we used $N = 50$, the total cost for BOLFI is 6,000 simulations. This is a reduction by ∼2 orders of magnitude with respect to likelihood-free rejection sampling (∼450, 000 simulations) and 3 orders of magnitude with respect to MCMC sampling of the exact posterior (6×10^6 simulations). It is also interesting to note that our BOLFI analysis required a factor of ∼3 fewer simulations than the recently introduced DELFI procedure [26], which used 20,000 simulations drawn from the prior for the analysis of the JLA.[2]

[1] The final Gelman-Rubin statistic [28] was $R - 1 \leq 5 \times 10^{-4}$ for each of the six parameters.

[2] A notable difference is that DELFI allowed the authors to perform the joint inference of the six parameters of the problem, whereas we only get the distribution of Ω_m and w. However, since these are the only two physically interesting parameters, inference of the nuisance parameters is not deemed crucial for this example.

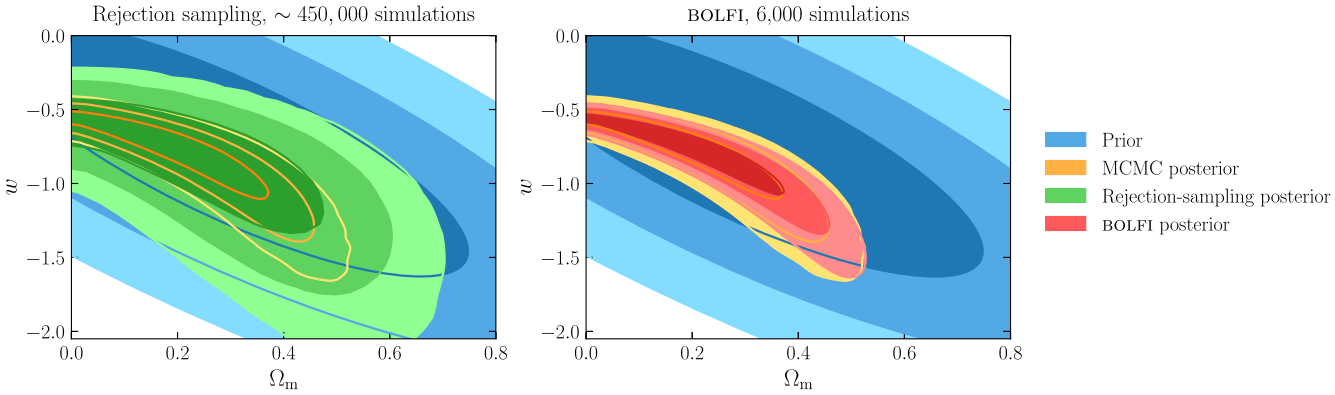

FIG. 7. Prior and posterior distributions for the joint inference of the matter density of the Universe, Ω_m, and the dark energy equation of state, w, from the JLA supernovae data set. The prior and exact posterior distribution (obtained from a long MCMC run requiring $\sim 6 \times 10^6$ data model evaluations) are shown in blue and orange, respectively. In the left panel, the approximate rejection-sampling posterior, based on 5,000 samples accepted out of $\sim 450,000$ simulations, is shown in green. In the right panel, the approximate BOLFI posterior, based on 6,000 simulations only, is shown in red. For all distributions, the 1σ, 2σ and 3σ contours are shown.

V. DISCUSSION

A. Benefits and limitations of the proposed approach for cosmological inferences

As noted in the Introduction, likelihood-free rejection sampling, when at all viable, is extremely costly in terms of the number of required simulations. In contrast, the BOLFI approach relies on a GP probabilistic model for the discrepancy, and therefore allows the incorporation of a smoothness assumption about the approximate likelihood $L(\boldsymbol{\theta})$. The smoothness assumption allows simulations in the training set to share information about their value of $\Delta_{\boldsymbol{\theta}}$ in the neighborhood of $\boldsymbol{\theta}$, which suggests that fewer simulations are needed to reach a certain level of accuracy. Indeed, the number of simulations required is typically reduced by 2 to 3 orders of magnitude, for a better final approximation of the posterior, as demonstrated by our tests in Sec. IV and in the statistical literature (see [12]).

A second benefit of BOLFI is that it actively acquires training data through Bayesian optimization. The trade-off between computational cost and statistical performance is still present, but in a modified form: the trade-off parameter is the size of the training set used in the regression. Within the training set, the user is free to choose which areas of the parameter space should be prioritized, so as to approximate the regression function more accurately there. In contrast, in ABC strategies that rely on drawing from a fixed proposal distribution (often the prior), or variants such as PMC-ABC, a fixed computational cost needs to be paid per value of $\boldsymbol{\theta}$ regardless of the value of $\Delta_{\boldsymbol{\theta}}$.

Finally, by focusing on parametric approximations to the exact likelihood, the approach proposed in this work is totally "ϵ-free," meaning that no threshold (which is often regarded as an unappealing *ad hoc* element) is required. As likelihood-based techniques, the parametric version of BOLFI has the drawback that assuming a wrong form for the synthetic likelihood or miscalculating values of its parameters (such as the covariance matrix) can potentially bias the approximate posterior and/or lead to an underestimation of credible regions. Nevertheless, massive data compression procedures can make the assumptions going into the choice of a Gaussian synthetic likelihood (almost) true by construction (see Sec. V B 4).

Of course, regressing the discrepancy and optimizing the acquisition function are not free of computational cost. However, the run-time for realistic cosmological simulation models can be hours or days. In comparison, the computational overhead introduced by BOLFI is negligible.

Likelihood-free inference should also be compared to existing likelihood-based techniques for cosmology such as Gibbs sampling or Hamiltonian Monte Carlo (e.g., [29,30] for the cosmic microwave background; [31–33] for galaxy clustering; [34] for weak lensing). The principal difference between these techniques and BOLFI lies in its likelihood-free nature. Likelihood-free inference has particular appeal for cosmological data analysis, since encoding complex physical phenomena and realistic observational effects into forward simulations is much easier than designing an approximate likelihood which incorporates these effects and solving the inverse problem. While the numerical complexity of likelihood-based techniques typically requires to approximate complex data models in order to access required products (conditionals or gradients of the pdfs) and to allow for sufficiently fast execution speeds, BOLFI performs inference from full-scale black-box data models. In the future, such an approach is expected to allow previously infeasible analyses, relying on a much more precise modeling of cosmological data, including in particular the complicated systematics they experience. However, while the physics and instruments will be more accurately modeled, the statistical approximation introduced with respect to likelihood-based techniques should be kept in mind.

Other key aspects of BOLFI for cosmological data analysis are the arbitrary choice of the statistical summaries and the easy joint treatment of different data sets. Indeed, as the data compression from $\mathbf{d}$ to $\mathbf{\Phi}$ is included in the simulator (see Sec. II C), summary statistics do not need to be quantities that can be physically modeled (such as the power spectrum) and can be chosen robustly to model misspecification. For example, for the microwave sky, the summaries could be the cross spectra between different frequency maps; and for imaging surveys, the cross-correlation between different bands. Furthermore, joint analyses of correlated data sets, which is usually challenging in likelihood-based approaches (as they require a good model for the joint likelihood) can be performed straightforwardly in a likelihood-free approach.

Importantly, as a general inference technique, BOLFI can be embedded into larger probabilistic schemes such as Gibbs or Hamiltonian-within-Gibbs samplers. Indeed, as posterior predictive distributions for conditionals and gradients of GPs are analytically tractable, it is easy to obtain samples of the BOLFI approximate posterior for use in larger models. BOLFI can therefore allow parts of a larger Bayesian hierarchical model to be treated as black boxes, without compromising the tractability of the entire model.

B. Possible extensions

1. High-dimensional inference

In this proof-of-concept paper, we focused on two-dimensional problems. Likelihood-free inference is in general very difficult when the dimensionality of the parameter space is large, due to the curse of dimensionality, which makes the volume exponentially larger with $\dim \boldsymbol{\theta}$. In BOLFI, this difficulty manifests itself in the form of a hard regression problem which needs to be solved. The areas in the parameter space where the discrepancy is small tend to be narrow in high dimension, therefore discovering these areas becomes more challenging as the dimension increases. The optimization of GP kernel parameters, which control the shapes of allowed features, also becomes more difficult. Furthermore, finding the global optimum of the acquisition function becomes more demanding (especially with the ones designed for ABC such as ExpIntVar, which have a high degree of structure—see Fig. 12, bottom right panel).

Nevertheless, Järvenpää *et al.* [16] showed on a toy simulation model (a Gaussian) that up to ten-dimensional inference is possible with BOLFI. As usual cosmological models do not include more than ten free physical parameters, we do not expect this limitation to be a hindrance. Any additional nuisance parameter or latent variable used internally by the simulator (such as α, β, M_{B}, δM in supernova cosmology, see Sec. IV B) can be automatically marginalized over, by using N realizations per $\boldsymbol{\theta}$. Recent advances in high-dimensional implementation of the synthetic likelihood [35] and high-dimensional Bayesian optimization (e.g., [36,37]), could also be exploited. In future work, we will address the problem of high-dimensional likelihood-free inference in a cosmological context.

2. Scalability with the number of acquisitions and probabilistic model for the discrepancy

In addition to the fundamental issues with high-dimensional likelihood-free inference described in the previous section, practical difficulties can be met.

Gaussian process regression requires the inversion of a matrix $\underline{\underline{\mathbf{K}}}$ of size $t \times t$, where t is the size of the training set. The complexity is $\mathcal{O}(t^3)$, which limits the size of the training set to a few thousand. Improving GPs with respect to this inversion is still subject to research (see [21], Chap. 8). For example, "sparse" Gaussian process regression reduces the complexity by introducing auxiliary "inducing variables." Techniques inspired by the solution to the Wiener filtering problem in cosmology, such as preconditioned conjugate gradient or messenger field algorithms could also be used [38–40]. Another strategy would be to divide the regression problem spatially into several patches with a lower number of training points [41]. Such approaches are possible extensions of the presented method.

In the GP probabilistic model employed to model the discrepancy, the variance depends only on the training locations, not on the obtained values [see Eq. (30)]. Furthermore, a stationary kernel is assumed. However, depending on the simulator, the discrepancy can show heteroscedasticity (i.e., its variance can depend on $\boldsymbol{\theta}$—see e.g., Fig. 5, bottom panel). Such cases could be handled by nonstationary GP kernels or different probabilistic models for the discrepancy, allowing a heteroscedastic regression.

3. Acquisition rules

As shown in our examples, attention should be given to the selection of an efficient acquisition rule. Although standard Bayesian optimization strategies such as the EI are reasonably effective, they are usually too greedy, focusing nearly all the sampling effort near the estimated minimum of the discrepancy and gathering too little information about other regions in the domain (see Fig. 12, bottom left panel). This implies that, unless the acquisition noise is high, the tails of the posterior will not be as well approximated as the modal areas. In contrast, the ExpIntVar acquisition rule, derived in this work for the parametric approach, addresses the inefficient use of resources in likelihood-free rejection sampling by directly targeting the regions of the parameter space where improvement in the estimation accuracy of the approximate posterior is needed most. In our experiments, ExpIntVar seems to correct—at least partially—for the well-known effect in Bayesian optimization of overexploration of the

domain boundaries, which becomes more problematic in high dimension.

Acquisition strategies examined so far in the literature (see [16] for a comparative study) have focused on single acquisitions and are all "myopic," in the sense that they reason only about the expected utility of the next acquisition, and the number of simulations left in a limited budget is not taken into account. Improvement of acquisition rules enabling batch acquisitions and nonmyopic reasoning are left to future extensions of BOLFI.

4. Data compression

In addition to the problem of the curse of dimensionality in parameter space, discussed in Sec. V B 1, likelihood-free inference usually suffers from difficulties in the measuring the (mis)match between simulations and observations if the data space also has high dimension. As discussed in Sec. II C, simulator-based models include a data compression step. The comparison in data space can be made more easily if dim $\mathbf{\Phi}$ is reduced. In future work, we will therefore aim at combining BOLFI with massive and (close to) optimal data compression strategies. These include MOPED [42], the score function [43], or information-maximizing neural networks [44]. Using such efficient data compression techniques, the number of simulations required for inference with BOLFI will be reduced even more, and the number of parameters treated could be increased.

Parametric approximations to the exact likelihood depend on quantities that have to be estimated using the simulator (typically for the Gaussian synthetic likelihood, the inverse covariance matrix of the summaries). Unlike supernova cosmology where the covariance matrix is easily obtained, in many cases it is prohibitively expensive to run enough simulations to estimate the required quantities, especially when they vary with the model parameters. In this context, massive data compression offers a way forward, reducing enormously the number of required simulations and making the analysis feasible when otherwise it might be essentially impossible [45,46].

An additional advantage of several data compression strategies is that they support the choice of a Gaussian synthetic likelihood. Indeed, the central limit theorem (for MOPED) or the form of the network's reward function (for information-maximizing neural networks) assist in giving the compressed data a near-Gaussian distribution. Furthermore, testing the Gaussian assumption for the synthetic likelihood will be far easier in a smaller number of dimensions than in the original high-dimensional data space.

C. Parallelization and computational efficiency

While MCMC sampling has to be done sequentially, BOLFI lends itself to more parallelization. In an efficient strategy, a master process performs the regression and decides on acquisition locations, then dispatches simulations to be run by different workers. In this way, many simulations can be run simultaneously in parallel, or even on different machines. This allows fast application of the method and makes it particularly suitable for grid computing. Extensions of the probabilistic model and of the acquisition rules, discussed in Secs. V B 2 and V B 3, would open the possibility of doing asynchronous acquisitions. Different workers would then work completely independently and decide on their acquisitions locally, while just sharing a pool of simulations to update their beliefs given all the evidence available.

While the construction of the training set depends on the observed data $\mathbf{\Phi}_\mathrm{O}$ (through the acquisition function), simulations can nevertheless be reused as long as summaries $\mathbf{\Phi}_{\boldsymbol{\theta}}$ are saved. This means that if one acquires new data $\mathbf{\Phi}'_\mathrm{O}$, the existing $\mathbf{\Phi}_{\boldsymbol{\theta}}$ (or a subset of them) can be used to compute the new discrepancy $\Delta_{\boldsymbol{\theta}}(\mathbf{\Phi}_{\boldsymbol{\theta}}, \mathbf{\Phi}'_\mathrm{O})$. Building an initial training set in this fashion can massively speed up the inference of $\mathcal{P}(\boldsymbol{\theta}|\mathbf{\Phi})_{\mathbf{\Phi}=\mathbf{\Phi}'_\mathrm{O}}$, whereas likelihood-based techniques would require a new MCMC.

D. Comparison to previous work

As discussed in the Introduction, likelihood-free rejection sampling is not a viable strategy for various problems that BOLFI can tackle. In recent work, another algorithm for scalable likelihood-free inference in cosmology, DELFI, was introduced [25,26]. The approach relies on estimating the joint probability $\mathcal{P}(\boldsymbol{\theta}, \mathbf{\Phi})$ via density estimation. This idea also relates to the work of Hahn *et al.* [47], who fit the sampling distribution of summaries $\mathcal{P}(\mathbf{\Phi}|\boldsymbol{\theta})$ using Gaussian mixture density estimation or independent component analysis, before using it for parameter estimation. This section discusses the principal similarities and differences.

The main difference between BOLFI and DELFI is the data acquisition. Training data are actively acquired in BOLFI, contrary to DELFI which, in the simplest scheme, draws from the prior. The reduction in the number of simulations for the inference of cosmological parameters (see Sec. IV B) can be interpreted as the effect of the Bayesian optimization procedure in combination with the ExpIntVar acquisition function. Using a purposefully constructed surrogate surface instead of a fixed proposal distribution, BOLFI focuses the simulation effort to reveal as much information as possible about the target posterior. In particular, its ability to reason about the quality of simulations before they are run is an essential element. Acquisition via Bayesian optimization almost certainly remains more efficient than even the PMC version of DELFI, which learns a better proposal distribution but still chooses parameters randomly. In future cosmological applications with simulators that are expensive and/or have a large latent space, an active data acquisition procedure could be crucial in order to provide a good model for the noisy approximate likelihood in the interesting regions of

parameter space, and to reduce the computational cost. This comes at the expense of a reduction of the parallelization potential: with a fixed proposal distribution (like in DELFI and unlike in BOLFI), the entire set of simulations can be run at the same time.

The second comment is related to the dimensionality of problems which can be addressed. Like DELFI, BOLFI relies on a probabilistic model to make ABC more efficient. However, the quantities employed differ, since in DELFI the relation between the parameters $\boldsymbol{\theta}$ and the summary statistics $\boldsymbol{\Phi}$ is modeled (via density estimation), while BOLFI focuses on the relation between the parameters $\boldsymbol{\theta}$ and the discrepancy $\Delta_{\boldsymbol{\theta}}$ (via regression). Summary statistics are multidimensional while the discrepancy is a univariate scalar quantity. Thus, DELFI requires to solve a density estimation problem in $\dim\boldsymbol{\theta} + \dim\boldsymbol{\Phi}$ (which equals $2 \times \dim\boldsymbol{\theta}$ if the compression from [43] is used), while BOLFI requires to solve a regression problem in $\dim\boldsymbol{\theta}$. Both tasks are expected to become more difficult as $\dim\boldsymbol{\theta}$ increases (a symptom of the curse of dimensionality, see Sec. V B 1), but the upper limits on $\dim\boldsymbol{\theta}$ for practical applications may differ. Further investigations are required to compare the respective maximal dimensions of problems that can be addressed by BOLFI and DELFI.

Finally, as argued by Alsing *et al.* [26], DELFI readily provides an estimate of the approximate evidence. In contrast, as in likelihood-based techniques, integration over parameter space is required with BOLFI to get

$$Z_{\boldsymbol{\Phi}} = \left(\int \mathcal{P}(\boldsymbol{\Phi}|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \right)_{\boldsymbol{\Phi}=\boldsymbol{\Phi}_{\mathrm{O}}} . \tag{47}$$

However, due to the GP model, the integral can be more easily computed, using the same strategies as for the integral appearing in ExpIntVar (see Sec. III E 2): only the GP predicted values are required at discrete locations on a grid (in low dimension) or at the positions of importance samples. A potential caveat is that DELFI has only been demonstrated to work in combination with the score function [43], which is necessary to reduce the dimensionality of $\boldsymbol{\Phi}$ before estimating the density.[3] The score function produces summaries that are only sufficient up to linear order in the loglikelihood. However, in ABC, care is required to perform model selection if the summary statistics are insufficient. Indeed, Robert *et al.* [48] [Eq. (1)] show that, in such a case, the approximate Bayes factor can be arbitrarily biased and that the approximation error is unrelated to the computational effort invested in running the ABC algorithm. Moreover, sufficiency for models $\mathcal{M}_1$ and $\mathcal{M}_2$ alone, or even for both of them—even if approximately realized via Alsing and Wandelt's procedure—does not guarantee sufficiency to compare the two different models $\mathcal{M}_1$ and $\mathcal{M}_2$ [49]. As the assumptions behind BOLFI do not necessarily necessitate to reduce $\dim\boldsymbol{\Phi}$ ($\Delta_{\boldsymbol{\theta}}$ is always a univariate scalar quantity, see above), these difficulties could be alleviated with BOLFI by carefully designing sufficient summary statistics for model comparison within the black-box simulator, if they exist.

[3]In contrast, Sec. IV B showed, for the same supernovae problem, that BOLFI can still operate if the comparison is done in the full 740-dimensional data space.

VI. CONCLUSION

Likelihood-free inference methods allow Bayesian inference of the parameters of simulator-based statistical models with no reference to the likelihood function. This is of particular interest for data analysis in cosmology, where complex physical and observational processes can usually be simulated forward but not handled in the inverse problem.

In this paper, we considered the demanding problem of performing Bayesian inference when simulating data from the model is extremely costly. We have seen that likelihood-free rejection sampling suffers from a vanishingly small acceptance rate when the threshold ϵ goes to zero, leading to the need for a prohibitively large number of simulations. This high cost is largely due to the lack of knowledge about the functional relation between the model parameters and the discrepancy. As a response, we have described a new approach to likelihood-free inference, BOLFI, that uses regression to infer this relation, and optimization to actively build the training data set. A crucial ingredient is the acquisition function derived in this work, with which training data are acquired such that the expected uncertainty in the final estimate of the posterior is minimized.

In case studies, we have shown that BOLFI is able to precisely recover the true posterior, even far in its tails, with as few as 6,000 simulations, in contrast to likelihood-free rejection sampling or likelihood-based MCMC techniques which require orders of magnitude more simulations. The reduction in the number of required simulations accelerated the inference massively.

This study opens up a wide range of possible extensions, discussed in Sec. V B. It also allows for novel analyses of cosmological data from fully nonlinear simulator-based models, as required e.g., for the cosmic web (see the discussions in [50–52]). Other applications may include the cosmic microwave background, weak gravitational lensing or intensity mapping experiments. We therefore anticipate that BOLFI will be a major ingredient in principled, simulator-based inference for the coming era of massive cosmological data.

ACKNOWLEDGMENTS

The author thanks Jens Jasche and Wolfgang Enzi for the collaboration that triggered this project, and Alan Heavens for useful discussions and a careful reading of the manuscript. This work has made use of a modified version of the

engine for likelihood-free inference (ELFI) [53] code. The author acknowledges funding from the Imperial College London Research Fellowship Scheme.

APPENDIX A: DERIVATIONS OF THE MATHEMATICAL RESULTS

1. Expressions for the approximate posterior

If we knew the target function f, the BOLFI posterior would be given as

$$\mathcal{P}_{\mathrm{BOLFI}}(\boldsymbol{\theta}|\boldsymbol{\Phi}_{\mathrm{O}}) \equiv \mathcal{P}(\boldsymbol{\theta}) \exp\left(-\frac{1}{2} f(\boldsymbol{\theta})\right) \propto \mathcal{P}(\boldsymbol{\theta}) \exp\left(\tilde{\ell}(\boldsymbol{\theta})\right). \tag{A1}$$

However, due to the limited computational resources we only have a finite training set $(\boldsymbol{\Theta}, \mathbf{f})$, which implies that there is uncertainty in the values of $f(\boldsymbol{\theta})$, and therefore that the approximate posterior is itself a stochastic process. To get its expectation under the model, the loglikelihood $\tilde{\ell}(\boldsymbol{\theta})$ is replaced by its expectation under the model, i.e., $-\frac{1}{2}\mu(\boldsymbol{\theta})$ [up to constants, see Eqs. (21) and (35)], giving Eq. (36).

Similarly, if the function f was known, the variance of the approximate posterior could be computed by standard propagation of uncertainties,

$$\begin{aligned} \mathrm{V}[\mathcal{P}_{\mathrm{BOLFI}}(\boldsymbol{\theta}|\boldsymbol{\Phi}_{\mathrm{O}}, \mathbf{f}, \boldsymbol{\Theta})] &= \left|\frac{\partial}{\partial f} \mathcal{P}(\boldsymbol{\theta}) \exp\left(-\frac{1}{2} f\right)\right|^2 \mathrm{V}[f] \\ &= \frac{\mathcal{P}(\boldsymbol{\theta})^2}{4} \exp(-f)\ \mathrm{V}[f]. \end{aligned} \tag{A2}$$

The argument of the exponential is $-f(\boldsymbol{\theta}) = 2\tilde{\ell}(\boldsymbol{\theta})$; it should be replaced by its expectation under the model, $-\mu(\boldsymbol{\theta})$. The variance of f under the model is, by definition, $\mathrm{V}^{(\mathrm{t})}[f] = \sigma^2(\boldsymbol{\theta})$. The result for $\mathrm{V}^{(\mathrm{t})}[\mathcal{P}_{\mathrm{BOLFI}}(\boldsymbol{\theta}|\boldsymbol{\Phi}_{\mathrm{O}}, \mathbf{f}, \boldsymbol{\Theta})]$ is therefore given by Eq. (37).

2. The ExpIntVar acquisition function in the parametric approach

We start by deriving the probability distributions for the GP mean and variance after one future observation $(\boldsymbol{\theta}_\star, f_\star)$ is added to the training set $(\boldsymbol{\Theta}, \mathbf{f})$. We denote them by $\mu_\star$ and $\sigma^2_\star$ respectively. These quantities are random functions of $\boldsymbol{\theta}$ since the new value $f_\star$ is unknown. Assuming that the GP mean is $m(\boldsymbol{\theta}) = 0$ for simplicity, and using Eq. (29) with the full training set $\{(\boldsymbol{\Theta}, \mathbf{f}), (\boldsymbol{\theta}_\star, f_\star)\}$, we get

$$\mu_\star(\boldsymbol{\theta}) = \begin{pmatrix} \underline{\underline{\mathbf{K}}} \\ \kappa(\boldsymbol{\theta}, \boldsymbol{\theta}_\star) \end{pmatrix}^{\top} \begin{pmatrix} \underline{\underline{\mathbf{K}}} & \underline{\mathbf{K}}_\star \\ \underline{\mathbf{K}}_\star^{\top} & K_{\star\star} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{f} \\ f_\star \end{pmatrix}, \tag{A3}$$

using the notations of Eqs. (31)–(34) and (43). By means of a standard formula for block matrix inversion, we get

$$\begin{aligned} \mu_\star(\boldsymbol{\theta}) &= \underline{\mathbf{K}}^{\intercal} \underline{\underline{\mathbf{K}}}^{-1} \mathbf{f} + [\kappa(\boldsymbol{\theta}, \boldsymbol{\theta}_\star) - \underline{\mathbf{K}}^{\intercal} \underline{\underline{\mathbf{K}}}^{-1} \underline{\mathbf{K}}_\star] \\ &\quad \times [K_{\star\star} - \underline{\mathbf{K}}_\star^{\intercal} \underline{\underline{\mathbf{K}}}^{-1} \underline{\mathbf{K}}_\star]^{-1} [f_\star - \underline{\mathbf{K}}_\star^{\intercal} \underline{\underline{\mathbf{K}}}^{-1} \mathbf{f}] \\ &= \mu(\boldsymbol{\theta}) + \mathrm{cov}(\boldsymbol{\theta}, \boldsymbol{\theta}_\star) \times [\sigma^2(\boldsymbol{\theta}_\star)]^{-1} [f_\star - \mu(\boldsymbol{\theta}_\star)]. \end{aligned} \tag{A4}$$

According to the GP model trained with $\{(\boldsymbol{\Theta}, \mathbf{f})\}$, the unknown future observation $f_\star$ is Gaussian distributed, i.e., $\mathcal{P}(f_\star|\mathbf{f}, \boldsymbol{\Theta}, \boldsymbol{\theta}_\star) = \mathcal{G}(\mu(\boldsymbol{\theta}_\star), \sigma^2(\boldsymbol{\theta}_\star))$. Thus, $[\sigma^2(\boldsymbol{\theta}_\star)]^{-1}[f_\star - \mu(\boldsymbol{\theta}_\star)]$ is Gaussian distributed with mean zero and variance $[\sigma^2(\boldsymbol{\theta}_\star)]^{-1}$, and $\mu_\star(\boldsymbol{\theta})$ is Gaussian distributed with mean $\mu(\boldsymbol{\theta})$ and variance $\tau^2(\boldsymbol{\theta}, \boldsymbol{\theta}_\star)$,

$$\mathcal{P}(\mu_\star(\boldsymbol{\theta})|\mathbf{f}, \boldsymbol{\Theta}, \boldsymbol{\theta}_\star) = \mathcal{G}(\mu(\boldsymbol{\theta}), \tau^2(\boldsymbol{\theta}, \boldsymbol{\theta}_\star)), \tag{A5}$$

using the notation introduced in Eq. (42).

Similar calculations for the variance show that

$$\sigma^2_\star(\boldsymbol{\theta}) = \sigma^2(\boldsymbol{\theta}) - \tau^2(\boldsymbol{\theta}, \boldsymbol{\theta}_\star), \tag{A6}$$

and therefore

$$\mathcal{P}(\sigma^2_\star(\boldsymbol{\theta})|\mathbf{f}, \boldsymbol{\Theta}, \boldsymbol{\theta}_\star) = \delta_{\mathrm{D}}(\sigma^2_\star(\boldsymbol{\theta}) - \sigma^2(\boldsymbol{\theta}) + \tau^2(\boldsymbol{\theta}, \boldsymbol{\theta}_\star)). \tag{A7}$$

This formula means that the reduction in the GP variance is deterministic and depends only on the new location $\boldsymbol{\theta}_\star$, independently of the future observation $f_\star$.

We now derive the expression for the expected integrated variance in the parametric approach:

$$\begin{aligned} &\mathrm{EIV}(\boldsymbol{\theta}_\star) \\ &\quad \equiv \mathrm{E}^{(\mathrm{t})}[\mathcal{L}[\mathcal{P}_{\mathrm{BOLFI}}(\boldsymbol{\theta}|\boldsymbol{\Phi}_{\mathrm{O}}, \mathbf{f}, \boldsymbol{\Theta}, f_\star, \boldsymbol{\theta}_\star)]] \\ &\quad = \int \mathcal{L}[\mathcal{P}_{\mathrm{BOLFI}}(\boldsymbol{\theta}|\boldsymbol{\Phi}_{\mathrm{O}}, \mathbf{f}, \boldsymbol{\Theta}, f_\star, \boldsymbol{\theta}_\star)] \mathcal{P}(f_\star|\mathbf{f}, \boldsymbol{\Theta}, \boldsymbol{\theta}_\star) \mathrm{d}f_\star \\ &\quad = \int\int \mathrm{V}[\mathcal{P}_{\mathrm{BOLFI}}(\boldsymbol{\theta}|\boldsymbol{\Phi}_{\mathrm{O}}, \mathbf{f}, \boldsymbol{\Theta})] \mathrm{d}\boldsymbol{\theta} \mathcal{P}(f_\star|\mathbf{f}, \boldsymbol{\Theta}, \boldsymbol{\theta}_\star) \mathrm{d}f_\star \\ &\quad = \int \mathcal{P}(\boldsymbol{\theta})^2 w^2(\boldsymbol{\theta}, \boldsymbol{\theta}_\star) \mathrm{d}\boldsymbol{\theta}, \end{aligned} \tag{A8}$$

where in the last line we have interchanged the order of integration, used Eq. (37), and introduced

$$\begin{aligned} w^2(\boldsymbol{\theta}, \boldsymbol{\theta}_\star) &\equiv \int \frac{1}{4} \exp[-\mu_\star(\boldsymbol{\theta})] \sigma^2_\star(\boldsymbol{\theta}) \mathcal{P}(f_\star|\mathbf{f}, \boldsymbol{\Theta}, \boldsymbol{\theta}_\star) \mathrm{d}f_\star \\ &= \mathrm{E}^{(\mathrm{t})}\left[\frac{1}{4} \exp[-\mu_\star(\boldsymbol{\theta})] \sigma^2_\star(\boldsymbol{\theta})\right], \end{aligned} \tag{A9}$$

that is to say the expectation of $\frac{1}{4}\exp[-\mu_\star(\boldsymbol{\theta})]\sigma^2_\star(\boldsymbol{\theta})$ under the GP model trained with $\{(\boldsymbol{\Theta}, \mathbf{f}), (\boldsymbol{\theta}_\star, f_\star)\}$. This expectation can be treated using Eqs. (A5) and (A6), assuming that mean and variance are independent: $\sigma^2_\star(\boldsymbol{\theta})$ becomes deterministically $\sigma^2(\boldsymbol{\theta}) - \tau^2(\boldsymbol{\theta}, \boldsymbol{\theta}_\star)$ under the model. As in Sec. A 1, the argument of the exponential, $\mu_\star(\boldsymbol{\theta})$, is replaced by its mean $\mu(\boldsymbol{\theta})$. The final result is

$$w^2(\boldsymbol{\theta},\boldsymbol{\theta}_\star)=\frac{1}{4}\exp\left[-\mu(\boldsymbol{\theta})\right]\left[\sigma^2(\boldsymbol{\theta})-\tau^2(\boldsymbol{\theta},\boldsymbol{\theta}_\star)\right]. \quad \text{(A10)}$$

3. Gradient of the ExpIntVar acquisition function in the parametric approach

In this section we derive the gradient of the expected integrated variance in the parametric approach, which can be used to find its minimum in parameter space. Inverting the differentiation and the integration, we have

$$\begin{aligned}\frac{\mathrm{dEIV}(\boldsymbol{\theta}_\star)}{\mathrm{d}\boldsymbol{\theta}_\star}&=\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}_\star}\int\mathcal{P}(\boldsymbol{\theta})^2 w^2(\boldsymbol{\theta},\boldsymbol{\theta}_\star)\mathrm{d}\boldsymbol{\theta}\\&=\int\mathcal{P}(\boldsymbol{\theta})^2\frac{\partial w^2(\boldsymbol{\theta},\boldsymbol{\theta}_\star)}{\partial\boldsymbol{\theta}_\star}\mathrm{d}\boldsymbol{\theta},\end{aligned} \quad \text{(A11)}$$

where

$$\begin{aligned}\frac{\partial w^2(\boldsymbol{\theta},\boldsymbol{\theta}_\star)}{\partial\boldsymbol{\theta}_\star}&=\frac{\partial}{\partial\boldsymbol{\theta}_\star}\left\{\frac{1}{4}\exp\left[-\mu(\boldsymbol{\theta})\right]\left[\sigma^2(\boldsymbol{\theta})-\tau^2(\boldsymbol{\theta},\boldsymbol{\theta}_\star)\right]\right\}\\&=-\frac{1}{4}\exp\left[-\mu(\boldsymbol{\theta})\right]\frac{\partial\tau^2(\boldsymbol{\theta},\boldsymbol{\theta}_\star)}{\partial\boldsymbol{\theta}_\star},\end{aligned} \quad \text{(A12)}$$

with

$$\begin{aligned}\frac{\partial\tau^2(\boldsymbol{\theta},\boldsymbol{\theta}_\star)}{\partial\boldsymbol{\theta}_\star}&=2\frac{\mathrm{cov}(\boldsymbol{\theta},\boldsymbol{\theta}_\star)}{\sigma^2(\boldsymbol{\theta}_\star)}\frac{\partial\mathrm{cov}(\boldsymbol{\theta},\boldsymbol{\theta}_\star)}{\partial\boldsymbol{\theta}_\star}\\&\quad-\frac{\mathrm{cov}(\boldsymbol{\theta},\boldsymbol{\theta}_\star)}{\sigma^4(\boldsymbol{\theta}_\star)}\frac{\partial\sigma^2(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_\star},\end{aligned} \quad \text{(A13)}$$

$$\frac{\partial\mathrm{cov}(\boldsymbol{\theta},\boldsymbol{\theta}_\star)}{\partial\boldsymbol{\theta}_\star}=\frac{\partial\kappa(\boldsymbol{\theta},\boldsymbol{\theta}_\star)}{\partial\boldsymbol{\theta}_\star}-\underline{\mathbf{K}}^{\intercal}\underline{\underline{\mathbf{K}}}^{-1}\frac{\partial\underline{\mathbf{K}}_\star}{\partial\boldsymbol{\theta}_\star}. \quad \text{(A14)}$$

The integral in Eq. (A11) can be evaluated similarly as discussed in Sec. III E 2.

APPENDIX B: SUMMARIZING GAUSSIAN SIGNALS

This Appendix gives the details of the problem of summarizing Gaussian signals discussed in Sec. IV A.

1. Forward modeling

The problem considered is the joint inference of the mean μ and of the variance σ^2 of a Gaussian $\mathcal{G}$, from which we have n samples that constitute the observed data $\mathbf{d}_\mathrm{O}$. The true likelihood for this problem is therefore

$$\mathcal{L}(\mu,\sigma^2)\equiv\mathcal{P}(\mathbf{d}|\mu,\sigma^2)_{|\mathbf{d}=\mathbf{d}_\mathrm{O}}=\mathcal{G}(\mathbf{d}|\mu,\sigma^2)_{|\mathbf{d}=\mathbf{d}_\mathrm{O}}. \quad \text{(B1)}$$

The Gaussian-inverse-Gamma is the natural prior for this problem, as it is conjugate for the Gaussian distribution with unknown mean and variance. It is a two-dimensional distribution characterized by four hyperparameters $(\alpha,\beta,\eta,\lambda)$. Samples of this prior can be straightforwardly generated by first sampling σ from the inverse-Gamma distribution Γ^{-1} with shape parameter α and scale parameter β, then by drawing μ from the Gaussian distribution $\mathcal{G}$ with mean η and variance σ^2/λ.

A noise-free simulator can be designed for this inference problem by taking the operations successively,

$$\sigma^2\curvearrowleft\mathcal{P}(\sigma^2|\alpha,\beta)=\Gamma^{-1}(\sigma^2|\alpha,\beta), \quad \text{(B2)}$$

$$\mu\curvearrowleft\mathcal{P}(\mu|\sigma^2,\eta,\lambda)=\mathcal{G}(\mu|\eta,\sigma^2/\lambda), \quad \text{(B3)}$$

$$\mathbf{d}\curvearrowleft\mathcal{P}(\mathbf{d}|\mu,\sigma^2)=\mathcal{G}(\mathbf{d}|\mu,\sigma^2). \quad \text{(B4)}$$

After the full data $\mathbf{d}$ are generated, they can be compressed to summary statistics. A simple choice is the empirical estimator for the mean and (unbiased) variance, defined by

$$\Phi^1(\mathbf{d})=\frac{1}{n}\sum_{k=1}^{n}d_k, \quad \text{(B5)}$$

$$\Phi^2(\mathbf{d})=\frac{1}{n-1}\sum_{k=1}^{n}(d_k-\Phi^1(\mathbf{d}))^2. \quad \text{(B6)}$$

$\boldsymbol{\Phi}=(\Phi^1,\Phi^2)$ is a sufficient summary statistic for the inference of (μ,σ^2). For this model, no information is lost in the reduction from $\mathbf{d}$ to $\boldsymbol{\Phi}$, which ensures $L(\boldsymbol{\theta})\propto\mathcal{L}(\boldsymbol{\theta})$. Furthermore, the distribution of the summary statistics $\boldsymbol{\Phi}_{(\mu,\sigma^2)}$ are here known:

$$\Phi^1_{(\mu,\sigma^2)}\sim\mathcal{G}\left(\mu,\frac{\sigma^2}{n}\right)\ \text{and}\ \Phi^2_{(\mu,\sigma^2)}\sim\Gamma\left(\frac{n-1}{2},\frac{2\sigma^2}{n-1}\right), \quad \text{(B7)}$$

where Γ is the Gamma distribution parametrized by its shape and scale.

The hierarchical graphical representation of the simulator is shown in Fig. 8.

2. Analytic solution

The exact solution of the problem described in the previous section is known analytically: the posterior is Gaussian-inverse-Gamma distributed, with parameters $(\alpha',\beta',\eta',\lambda')$ given by

$$\alpha'=\alpha+\frac{n}{2}, \quad \text{(B8)}$$

$$\beta'=\beta+\frac{n\lambda}{\lambda+n}\frac{(\Phi^1_\mathrm{O}-\eta)^2}{2}+\frac{n-1}{2}\Phi^2_\mathrm{O}, \quad \text{(B9)}$$

$$\eta'=\frac{\lambda\eta+n\Phi^1_\mathrm{O}}{\lambda+n}, \quad \text{(B10)}$$

$$\lambda'=\lambda+n, \quad \text{(B11)}$$

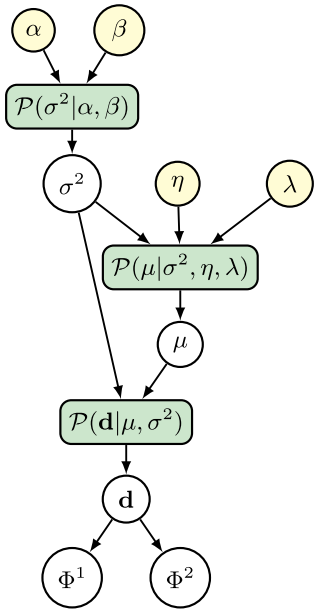

FIG. 8. Hierarchical forward model for the problem of summarizing simulated Gaussian signals. The upper part corresponds to the generation of random variables from the two-dimensional Gaussian-inverse-Gamma prior parametrized by $(\alpha,\beta,\eta,\lambda)$: first σ^2 is drawn from $\mathcal{P}(\sigma^2|\alpha,\beta)$ (an inverse-Gamma distribution with shape parameter α and scale parameter β), then μ is drawn from $\mathcal{P}(\mu|\sigma^2,\eta,\lambda)$ (a Gaussian distribution with mean η and variance σ^2/λ). A Gaussian likelihood $\mathcal{P}(\mathbf{d}|\mu,\sigma^2)$ with mean μ and variance σ^2 gives the data $\mathbf{d}$. Finally, the simulator produces two summary statistics: the estimated mean and variance, Φ^1 and Φ^2 respectively.

where $\Phi^1_{\rm O}$ and $\Phi^2_{\rm O}$ are the summary statistics of the observed data, defined by applying Eqs. (B5) and (B6) to $\mathbf{d}_{\rm O}$.

For the experiment described in Sec. IV A 1, we have used $n=10$ and $N=20$. The data have been generated from ground truth parameters $\mu_{\rm true}=0.8$ and $\sigma^2_{\rm true}=2.9$. We have measured $\Phi^1_{\rm O}=1.3212$, and have chosen a Gaussian prior on μ with mean unity and variance unity. The exact posterior is therefore a Gaussian with mean 1.2490 and variance 0.2248.

For the experiment described in Sec. IV A 2, we have used $n=50$ and $N=10$. The data have been generated from ground truth parameters $\mu_{\rm true}=0.8$ and $\sigma^2_{\rm true}=2.9$ (shown as the plus in Fig. 6). We have measured $\Phi^1_{\rm O}=0.9925$ and $\Phi^2_{\rm O}=2.8499$. We have chosen a prior with parameters $(\alpha,\beta,\eta,\lambda)=(22,54,0,6)$. The exact posterior has therefore parameters $(\alpha',\beta',\eta',\lambda')=(47,127.8885,0.8862,56)$.

3. Derivation of the Gaussian-Gamma synthetic likelihood for likelihood-free inference

For likelihood-free inference, a computable approximation $\hat{L}^N(\mu,\sigma^2)$ to the true likelihood given by Eq. (B1) is required. In this section, we design a parametric form for $\hat{L}^N(\mu,\sigma^2)$ which we call the Gaussian-Gamma synthetic likelihood.

As the approach is likelihood free, $\hat{L}^N(\mu,\sigma^2)$ should be based only on realizations of the summary statistics. Using the simulator described in Sec. B 1, we can generate N realizations of Φ^1 and Φ^2 for each pair of input parameters (μ,σ^2). Assuming exchangeability, we can use the Ansatz $L(\mu,\sigma^2)\equiv L_1(\mu,\sigma^2)L_2(\mu,\sigma^2)$ and $\hat{L}^N(\mu,\sigma^2)\equiv \hat{L}^N_1(\mu,\sigma^2)\hat{L}^N_2(\mu,\sigma^2)$, or using the loglikelihood,

$$\hat{\ell}^N(\mu,\sigma^2)\equiv\hat{\ell}^N_1(\mu,\sigma^2)+\hat{\ell}^N_2(\mu,\sigma^2), \tag{B12}$$

where the first term depends only on Φ^1 and the second on Φ^2. They are discussed successively in the following.

Φ^1 is the empirical mean of the independent and identically distributed components of $\mathbf{d}$, obtained through averaging. As discussed in Sec. II D 2, the Gaussian parametric approximation also known as the synthetic likelihood is appropriate in this case. We therefore define

$$-2\hat{\ell}^N_1(\mu,\sigma^2)\equiv\log|2\pi\hat{v}^1_{(\mu,\sigma^2)}|+\frac{(\Phi^1_{\rm O}-\hat{\mu}^1_{(\mu,\sigma^2)})^2}{\hat{v}^1_{(\mu,\sigma^2)}}, \tag{B13}$$

where $\hat{\mu}^1_{(\mu,\sigma^2)}$ and $\hat{v}^1_{(\mu,\sigma^2)}$ are respectively the empirical mean and variance of the simulated Φ^1, i.e.,

$$\hat{\mu}^1_{(\mu,\sigma^2)}\equiv\mathrm{E}^N[\Phi^1_{(\mu,\sigma^2)}], \tag{B14}$$

$$\hat{v}^1_{(\mu,\sigma^2)}\equiv\mathrm{E}^N[(\Phi^1_{(\mu,\sigma^2)}-\hat{\mu}^1_{(\mu,\sigma^2)})^2]. \tag{B15}$$

As $\mathcal{P}(\Phi^1|\mu,\sigma^2)$ is actually a Gaussian distribution, the equality $\tilde{L}_1(\mu,\sigma^2)=L_1(\mu,\sigma^2)$ holds without approximation, in the limit of infinite computer resources. From Eq. (B7), we also have

$$\hat{\mu}^1_{(\mu,\sigma^2)}\sim\mathcal{G}\left(\mu,\frac{\sigma^2}{Nn}\right)\quad\text{and}$$
$$\hat{v}^1_{(\mu,\sigma^2)}\sim\Gamma\left(\frac{N-1}{2},\frac{2\sigma^2}{n(N-1)}\right), \tag{B16}$$

which allows a closed-form definition of the stochastic process defining $\hat{L}^N_1(\mu,\sigma^2)$.

Φ^2 is the empirical variance of the components of $\mathbf{d}$. As noted in Eq. (B7), $\mathcal{P}(\Phi^2|\mu,\sigma^2)$ is a Gamma distribution. Consequently, we introduce for $\Phi^2_{\rm O}$ a Gamma synthetic likelihood, namely

$$-2\hat{\ell}^N_2(\mu,\sigma^2)\equiv-2(\hat{k}_{(\mu,\sigma^2)}-1)\log\Phi^2_{\rm O}+\frac{2\Phi^2_{\rm O}}{\hat{\theta}_{(\mu,\sigma^2)}}$$
$$+2\hat{k}_{(\mu,\sigma^2)}\log\hat{\theta}_{(\mu,\sigma^2)}+2\log\Gamma(\hat{k}_{(\mu,\sigma^2)}). \tag{B17}$$

The question is now to use the simulator in order to learn the shape and scale parameters $\hat{k}_{(\mu,\sigma^2)}$ and $\hat{\theta}_{(\mu,\sigma^2)}$. To do so, the simplest possibility is the methods of moments: using a Gaussian approximation to the first two moments of the Gamma distribution, we have

$$\hat{\mu}^2_{(\mu,\sigma^2)} \approx \hat{k}_{(\mu,\sigma^2)}\hat{\theta}_{(\mu,\sigma^2)}, \quad \text{and} \tag{B18}$$

$$\hat{v}^2_{(\mu,\sigma^2)} \approx \hat{k}_{(\mu,\sigma^2)}(\hat{\theta}_{(\mu,\sigma^2)})^2, \tag{B19}$$

where $\hat{\mu}^2_{(\mu,\sigma^2)}$ and $\hat{v}^2_{(\mu,\sigma^2)}$ are the empirical mean and variance of Φ^2, respectively, defined as in Eqs. (B14) and (B15). Solving this system for $\hat{k}_{(\mu,\sigma^2)}$ and $\hat{\theta}_{(\mu,\sigma^2)}$, we obtain the parameters of $\hat{\ell}^N_2$,

$$\hat{k}_{(\mu,\sigma^2)} \approx \frac{(\hat{\mu}^2_{(\mu,\sigma^2)})^2}{\hat{v}^2_{(\mu,\sigma^2)}}, \quad \text{and} \tag{B20}$$

$$\hat{\theta}_{(\mu,\sigma^2)} \approx \frac{\hat{v}^2_{(\mu,\sigma^2)}}{\hat{\mu}^2_{(\mu,\sigma^2)}}. \tag{B21}$$

As $\mathcal{P}(\Phi^2|\mu,\sigma^2)$ is known to be a Gamma distribution, we have, as for the first term, $\tilde{L}^2(\mu,\sigma^2) = L^2(\mu,\sigma^2)$ in the limit of infinite computer resources. $\hat{\mu}^2_{(\mu,\sigma^2)}$ is the sum of N independent random variables, identically distributed according to a Gamma distribution with the same scale parameter. Therefore, it obeys

$$\hat{\mu}^2_{(\mu,\sigma^2)} \sim \Gamma\left(\frac{N(n-1)}{2}, \frac{2\sigma^2}{N(n-1)}\right). \tag{B22}$$

Unlike $\hat{\mu}^2_{(\mu,\sigma^2)}$, there is no closed-form expression for $\hat{v}^2_{(\mu,\sigma^2)}$, $\hat{k}_{(\mu,\sigma^2)}$ and $\hat{\theta}_{(\mu,\sigma^2)}$ with standard probability distributions. However, these quantities, as well as $\hat{L}^N_2(\mu,\sigma^2)$, can be easily simulated using their defining equations.

The resulting approximate likelihood $\hat{L}^N(\mu,\sigma^2)$ is the product of a Gaussian synthetic likelihood for Φ^1 and a Gamma synthetic likelihood for Φ^2. It is shown in Fig. 9. There, the different panels show that realizations become smoother as N increases, i.e., with more computational resources.

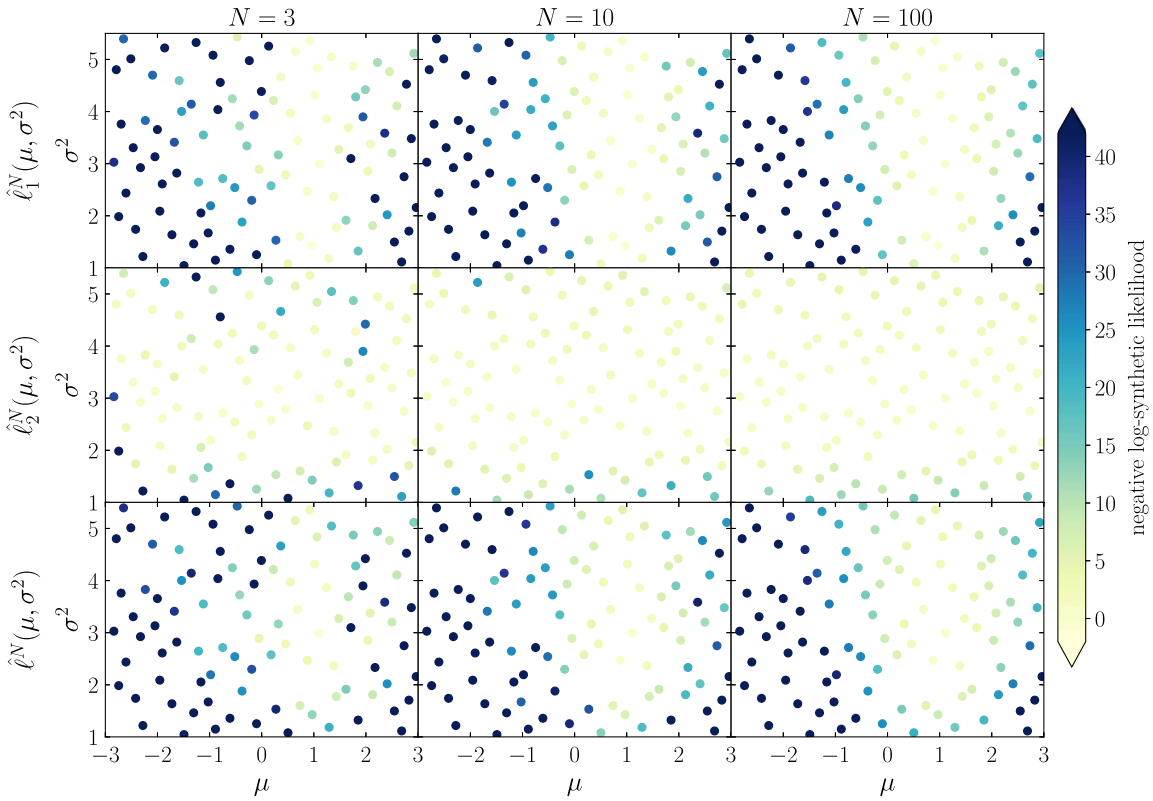

FIG. 9. Illustration of the Gaussian-Gamma synthetic likelihood as a stochastic process. The observed data have been generated using $\mu_{\text{true}} = 0.8$ and $\sigma^2_{\text{true}} = 2.9$. The 100 sampling points form a low-discrepancy quasirandom Sobol sequence in parameter space. The three rows show respectively the first term $\hat{\ell}^N_1(\mu,\sigma^2)$ (a Gaussian synthetic likelihood for Φ^1), the second term $\hat{\ell}^N_2(\mu,\sigma^2)$ (a Gamma synthetic likelihood for Φ^2), and their sum $\hat{\ell}^N(\mu,\sigma^2)$. The three columns show a varying number of simulations per value of (μ,σ^2): $N = 3$, $N = 10$, $N = 100$. The use of simulations makes the synthetic likelihood a stochastic process. Its noisiness decreases as N increases, i.e., as more computational resources are invested.

APPENDIX C: SUPERNOVA COSMOLOGY

This Appendix gives the details of the data model and the modeling assumptions for the problem of inferring cosmological parameters from the JLA catalog, presented in Sec. IV B.

1. Data samples

Type Ia supernovae (SNe Ia) are "standard candles," i.e., astrophysical objects that precisely map the distance-redshift relation in the nearby Universe. As such, they are one of the most sensitive probes of the late-time expansion history of the Universe. The joint lightcurve analysis (JLA) [17] is a compiled catalog of 740 SNe Ia. 374 objects in the redshift range $0.03 \leq z \leq 0.41$ have been identified by the Sloan Digital Sky Survey phase II (SDSS-II) supernova survey [54] and confirmed as SNe Ia by spectroscopic follow-up observations. The remaining objects come from the earlier C11 compilation [55]: 118 are low-z ($z \leq 0.08$) SNe Ia from the third release [56] of photometric data acquired at the Whipple Observatory of the Harvard-Smithsonian Center for Astrophysics (CfA3). 239 SNe Ia in the redshift range $0.12 \leq z \leq 1.07$ have been observed by the Supernova Legacy Survey (SNLS) [57,58]. Finally, nine objects are high-redshift SNe Ia ($0.8 \leq z \leq 1.4$) observed by the Hubble Space Telescope (HST) [59].

For each supernova, the JLA catalog provides a rich variety of information. The full data set comprises lightcurves in different bands and spectroscopic or photometric observations of each SN Ia. These products are then used to estimate the redshift z, the apparent magnitude m, the color at maximum brightness C and a time-stretching parameter for the lightcurve, X_1. In particular, the catalog includes several estimations of the redshift z. In this work, we use $z = z_{\mathrm{CMB}}$, the cosmological redshift of the object in the frame of the cosmic microwave background (CMB), including peculiar velocity corrections. For our data vector $\mathbf{d}_{\mathrm{O}}$, we use the estimated B-band peak magnitudes in the rest frame, denoted $(m^k_{\mathrm{B,O}})$ for $k \in [\![1, 740]\!]$ (as in the body of the paper, the subscript O stands for "observed"). The magnitudes are plotted as a function of redshift in the Hubble diagram shown in Fig. 10 (left). The JLA catalog also provides some properties of the SNe host galaxies, in particular the stellar mass M_{stellar}. We denote by $\mathbf{z}_{\mathrm{O}} \equiv (z^k_{\mathrm{O}})$, $\mathbf{X}_{1,\mathrm{O}} \equiv (X^k_{1,\mathrm{O}})$, $\mathbf{C}_{\mathrm{O}} \equiv (C^k_{\mathrm{O}})$, $\mathbf{M}_{\mathrm{stellar,O}} \equiv (M^k_{\mathrm{stellar,O}})$ for $k \in [\![1, 740]\!]$, and $\mathbf{m}_{\mathrm{O}} \equiv (\mathbf{z}_{\mathrm{O}}, \mathbf{X}_{1,\mathrm{O}}, \mathbf{C}_{\mathrm{O}}, \mathbf{M}_{\mathrm{stellar,O}})$ the metadata used in the analysis.

2. Supernova data model and distance estimates

Distance estimation with SNe Ia is based on the assumption that they are standardizable objects, which is quantified by a linear model for the apparent magnitude:

$$m_{\mathrm{B}} = 5\log_{10}\left[\frac{D_{\mathrm{L}}(z)}{10\ \mathrm{pc}}\right] + \widetilde{M}_{\mathrm{B}}(M_{\mathrm{stellar}}, M_{\mathrm{B}}, \delta M) - \alpha X_1 + \beta C. \tag{C1}$$

The absolute magnitude $\widetilde{M}_{\mathrm{B}}$ depends on the stellar mass of the host galaxy, M_{stellar}. This dependence is assumed to be captured by the relation [55]

$$\widetilde{M}_{\mathrm{B}}(M_{\mathrm{stellar}}, M_{\mathrm{B}}, \delta M) = M_{\mathrm{B}} + \delta M \Theta(M_{\mathrm{stellar}} - 10^{10}\mathrm{M}_{\odot}), \tag{C2}$$

where Θ is the Heaviside function and $\mathrm{M}_{\odot}$ the mass of the Sun. The lightcurve calibration model therefore comprises

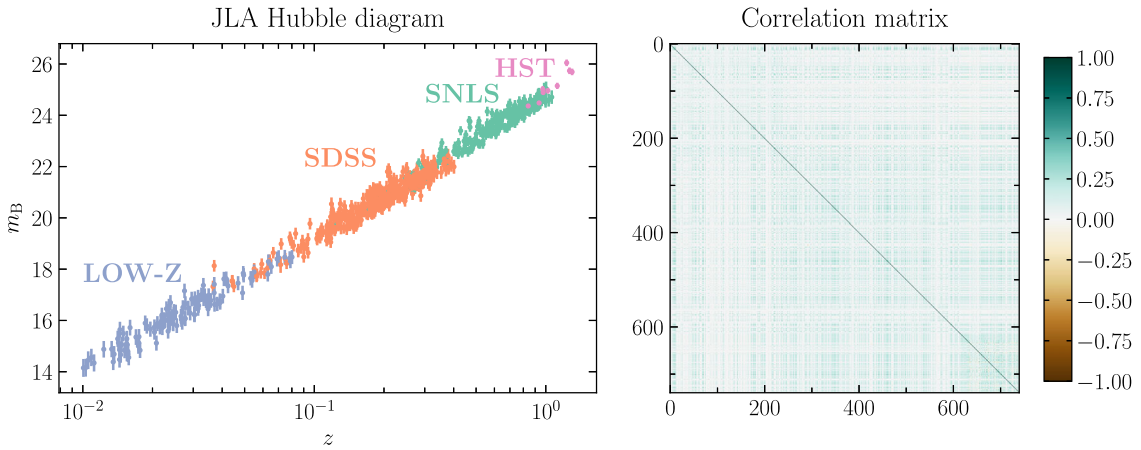

FIG. 10. Left panel: JLA Hubble diagram, representing the observed apparent magnitudes m_{B} of 740 type Ia supernovae as a function of their redshift. The error bars represented correspond to $2\Delta m_{\mathrm{B}}$, where Δm_{B} is included in the JLA catalog but not used in this work. The different colors correspond to the different observational programs used in the compilation. Right panel: Correlation matrix of the observed apparent magnitudes, taking into account statistical and various systematic uncertainties (see [17], Sec. V. 5, for details on the construction of the covariance matrix).

four nuisance parameters $(\alpha, \beta, M_B, \delta M)$. They are assumed to be independent of host galaxy properties.

The cosmological model enters in the analysis through the distance-redshift relation. We assume a flat Universe containing cold dark matter and a dark energy component (wCDM hereafter). A wCDM Universe is characterized by two physical parameters Ω_m (the matter density) and w (the equation of state of dark energy, assumed constant in time). The luminosity distance appearing in Eq. (C1) is given by (e.g., [60], Sec. VII)

$$D_L(z) = \frac{(1+z)\,c}{H_0}\int_0^z \frac{dz'}{E(z')},$$
$$E(z) \equiv \sqrt{\Omega_m(1+z)^3 + (1-\Omega_m)(1+z)^{3(w+1)}}, \quad \text{(C3)}$$

where c is the speed of light in vacuum and $H_0 \equiv 100h\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$.

3. Forward modeling

The data model described in the previous section can be simulated forward by taking the following operations successively:

$$(\Omega_m, w) \curvearrowleft \mathcal{P}(\Omega_m, w|\boldsymbol{\omega}, \mathbf{S}), \quad \text{(C4)}$$

$$(\alpha, \beta, M_B, \delta M) \curvearrowleft \mathcal{P}(\alpha, \beta, M_B, \delta M|\mathbf{M}), \quad \text{(C5)}$$

$$D_L(\mathbf{z}_O) \curvearrowleft \mathcal{P}(D_L(\mathbf{z}_O)|\Omega_m, w), \quad \text{(C6)}$$

$$\mathbf{d} \curvearrowleft \mathcal{P}(\mathbf{d}|D_L(\mathbf{z}_O), \alpha, \beta, M_B, \delta M, \mathbf{m}_O). \quad \text{(C7)}$$

The last two steps are deterministic: in Eq. (C6), the luminosity distance at the observed redshifts is computed via Eq. (C3), and in Eq. (C7), the predicted data $\mathbf{d}_{(\Omega_m,w)} \equiv (m^k_{B,(\Omega_m,w)})$ come from Eqs. (C1) and (C2). We can therefore write

$$\begin{aligned}
&\mathcal{P}(D_L(\mathbf{z}_O)|\Omega_m, w)\\
&\quad = \delta_D(D_L(\mathbf{z}_O) - \hat{D}_L(\mathbf{z}_O, \Omega_m, w)),\\
&\mathcal{P}(\mathbf{d}|D_L(\mathbf{z}_O), \alpha, \beta, M_B, \delta M, \mathbf{m}_O)\\
&\quad = \delta_D(\mathbf{d} - \hat{\mathbf{d}}(D_L(\mathbf{z}_O), \alpha, \beta, M_B, \delta M, \mathbf{m}_O)),\\
&\mathcal{P}(\mathbf{d}|\Omega_m, w, \mathbf{M}, \mathbf{m}_O)\\
&\quad = \delta_D(\mathbf{d} - \hat{\mathbf{d}}(D_L(\mathbf{z}_O), \alpha, \beta, M_B, \delta M, \mathbf{m}_O))\\
&\qquad \times \delta_D(D_L(\mathbf{z}_O) - \hat{D}_L(\mathbf{z}_O, \Omega_m, w))\\
&\qquad \times \mathcal{P}(\alpha, \beta, M_B, \delta M|\mathbf{M}). \quad \text{(C8)}
\end{aligned}$$

The probability $\mathcal{P}(\Omega_m, w|\boldsymbol{\omega}, \mathbf{S})$ appearing in Eq. (C4) is the Gaussian prior given by Eq. (46), i.e., $\mathcal{P}(\Omega_m, w|\boldsymbol{\omega}, \mathbf{S}) \equiv \mathcal{G}(\boldsymbol{\omega}, \mathbf{S})$ with

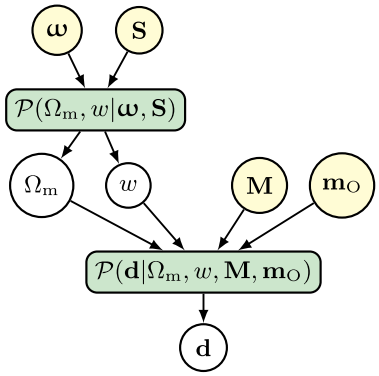

FIG. 11. Hierarchical forward model for the analysis of the JLA type Ia supernovae catalog. The prior on the physical parameters Ω_m and w is a Gaussian with mean $\boldsymbol{\omega}$ and covariance matrix $\mathbf{S}$. The data generating process uses four nuisance parameters, the distribution of which is characterized by the hyperparameters $\mathbf{M}$ and the supernovae metadata $\mathbf{m}_O$.

$$\boldsymbol{\omega} \equiv \begin{pmatrix} 0.3 \\ 0.75 \end{pmatrix} \quad \text{and} \quad \mathbf{S} \equiv \begin{pmatrix} 0.4^2 & -0.24 \\ -0.24 & 0.75^2 \end{pmatrix}. \quad \text{(C9)}$$

Finally, $\mathcal{P}(\alpha, \beta, M_B, \delta M|\mathbf{M})$ is the sampling distribution of nuisance parameters, characterized by hyperparameters $\mathbf{M}$. Following previous studies, we choose broad, independent Gaussian priors on each of the four parameters. Specifically, we assume

$$\begin{pmatrix} \alpha \\ \beta \\ M_B \\ \delta M \end{pmatrix} \sim \mathcal{G}\left[\begin{pmatrix} 0.125 \\ 2.6 \\ -19.05 \\ -0.05 \end{pmatrix}, \begin{pmatrix} 0.025^2 & 0 & 0 & 0 \\ 0 & 0.25^2 & 0 & 0 \\ 0 & 0 & 0.1^2 & 0 \\ 0 & 0 & 0 & 0.03^2 \end{pmatrix}\right]. \quad \text{(C10)}$$

The hierarchical graphical representation of the simulator is shown in Fig. 11.

4. Discrepancy

Following Betoule *et al.* [17] (formula 15), we define the discrepancy between observed and simulated data as

$$\Delta_{(\Omega_m,w)} \equiv (\mathbf{d}_O - \hat{\boldsymbol{\mu}}_{(\Omega_m,w)})^\top \mathbf{C}^{-1} (\mathbf{d}_O - \hat{\boldsymbol{\mu}}_{(\Omega_m,w)}), \quad \text{(C11)}$$

where $\hat{\boldsymbol{\mu}}_{(\Omega_m,w)}$ is the average of N simulated realizations of $\mathbf{d}_{(\Omega_m,w)} \equiv (m^k_{B,(\Omega_m,w)})$ for $k \in [\![1, 740]\!]$. This is equivalent to assuming a Gaussian synthetic likelihood (see Sec. II D 2) in approximate Bayesian computation, and to using a Gaussian likelihood for the exact Bayesian problem, solved by MCMC sampling for reference. Betoule *et al.* [17] (Sec. 5. 5), constructed a covariance matrix $\mathbf{C}_{(\alpha,\beta)}$ which accounts for the uncertainty in the color, stretch and redshift of each supernova, depending on the nuisance parameters α and β, but dropped the term $\log|2\pi\mathbf{C}_{(\alpha,\beta)}|$ from the

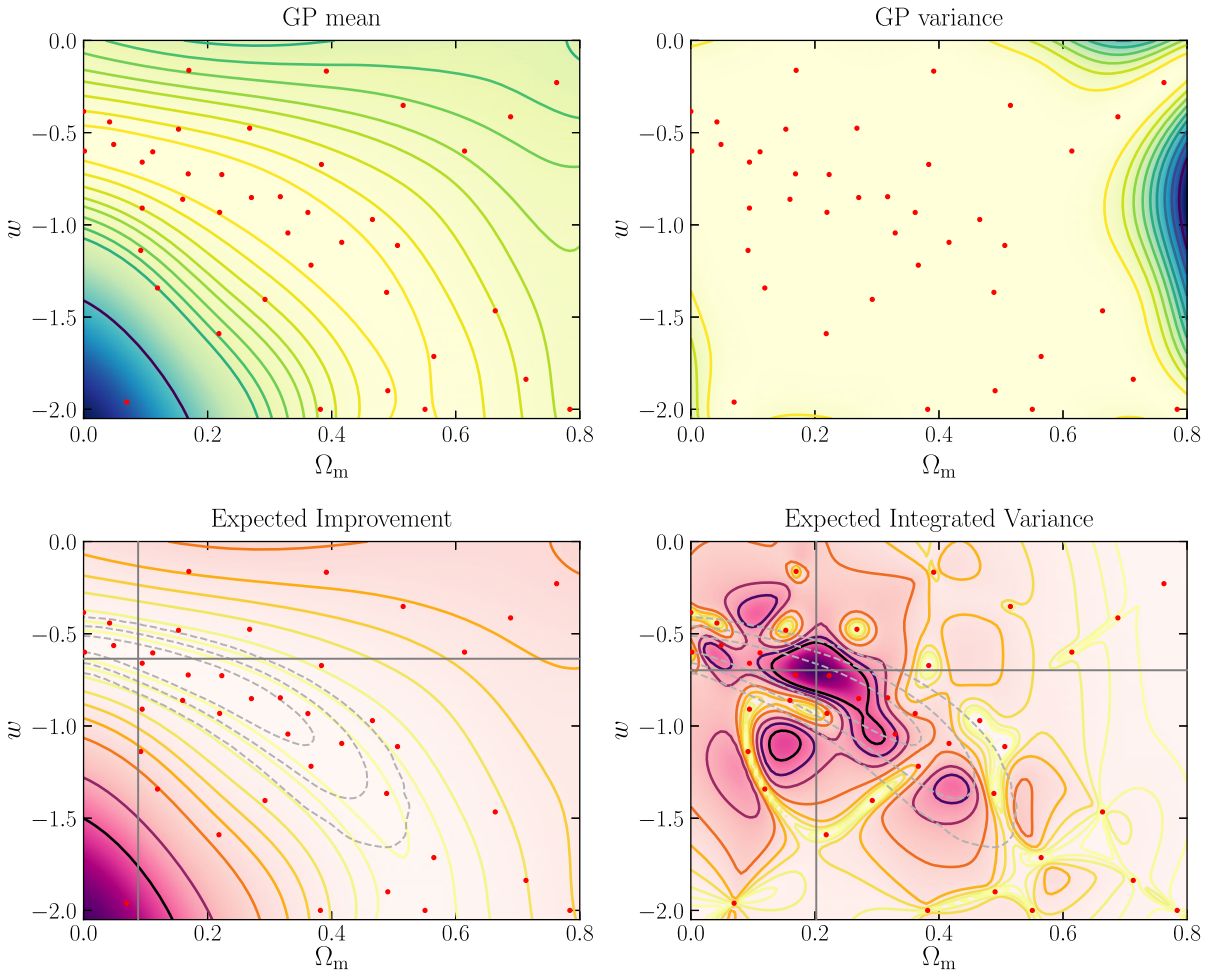

FIG. 12. BOLFI at work after 20 acquisitions for the supernovae cosmology problem. Top panels: Isocontours of the Gaussian process model for the discrepancy $\Delta_{(\Omega_m,w)}$. The mean (left) and variance (right) are shown in arbitrary units. The red dots mark the location of the training parameters (Ω_m, w). Bottom panels: Isocontours of the acquisition surfaces built from the Gaussian process, using two different acquisition rules: the expected improvement (which is maximized, left), and the expected integrated variance (which is minimized, right). Units are arbitrary. The location of the next acquisition (i.e., the optimizer) is marked by the cross, and the contours of the exact posterior are plotted as dashed gray lines for reference. The initial training set is composed of 20 samples, and the expected integrated variance has been used for the 20 acquisitions shown.

definition of the discrepancy. Since α and β are very well constrained by the data, the dependence of $\mathbf{C}_{(\alpha,\beta)}$ has a weak effect on the final inference results. Therefore, in this work (and as in [26]), we assume a fixed covariance matrix $\mathbf{C}$ where the parameters α and β are taken at their maximum *a posteriori* value ($\alpha = 0.1256$, $\beta = 2.6342$). This also justifies dropping the constant term $\log|2\pi\mathbf{C}|$ from the definition of the discrepancy.

We used the data (version 6) and the python script provided along with the JLA[4] to generate the 740×740 covariance matrix $\mathbf{C}$. The associated correlation matrix is shown in Fig. 10 (right).

[4]These products are available at http://supernovae.in2p3.fr/sdss_snls_jla/ReadMe.html.

5. Acquisition

For the analysis described in Sec. IV B, we used $N = 50$ simulations per point (Ω_m, w), and the ExpIntVar rule without acquisition noise. Figure 12 shows the state of BOLFI after 20 acquisitions, for a training set of 40 samples. As can be observed in the lower panels, the different acquisition functions implement a different trade-off between exploration and exploitation. In particular, the ExpIntVar surface has a much more complex structure. Simulations surrounding the 3σ contour of the posterior have already been run (exploration). The proposed acquisition is in a region of high estimated density (exploitation), but not yet fully sampled. On the contrary, the next acquisition suggested by the EI criterion stays in the "valley" (the innermost contour line) where lies the estimated optimum, meaning that the tails of the posterior will hardly be sufficiently sampled.

[1] J.-M. Marin, P. Pudlo, C. P. Robert, and R. Ryder, Approximate Bayesian computational methods, Stat. Comput. **22**, 1167 (2012).

[2] J. Lintusaari, M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander, Fundamentals and recent developments in approximate Bayesian computation, Syst. Biol. **66**, e66 (2017).

[3] A. Weyant, C. Schafer, and W. M. Wood-Vasey, Likelihood-free cosmological inference with type Ia supernovae: Approximate Bayesian computation for a complete treatment of uncertainty, Astrophys. J. **764**, 116 (2013).

[4] C.-A. Lin and M. Kilbinger, A new model to predict weak-lensing peak counts. II. Parameter constraint strategies, Astron. Astrophys. **583**, A70 (2015).

[5] C. Hahn, M. Vakili, K. Walsh, A. P. Hearin, D. W. Hogg, and D. Campbell, Approximate Bayesian computation in large-scale structure: Constraining the galaxy-halo connection, Mon. Not. R. Astron. Soc. **469**, 2791 (2017).

[6] S. Carassou, V. de Lapparent, E. Bertin, and D. Le Borgne, Inferring the photometric and size evolution of galaxies from image simulations. I. Method, Astron. Astrophys. **605**, A9 (2017).

[7] T. Kacprzak, J. Herbel, A. Amara, and A. Réfrégier, Accelerating approximate Bayesian computation with quantile regression: Application to cosmological redshift distributions, J. Cosmol. Astropart. Phys. 02 (2018) 042.

[8] F. B. Davies, J. F. Hennawi, A.-C. Eilers, and Z. Lukić, A new method to measure the post-reionization ionizing background from the joint distribution of Lyα and Lyβ forest transmission, Astrophys. J. **855**, 106 (2018).

[9] J. Akeret, A. Refregier, A. Amara, S. Seehars, and C. Hasner, Approximate Bayesian computation for forward modeling in cosmology, J. Cosmol. Astropart. Phys. 08 (2015) 043.

[10] E. E. O. Ishida, S. D. P. Vitenti, M. Penna-Lima, J. Cisewski, R. S. de Souza, A. M. M. Trindade, E. Cameron, and V. C. Busti (COIN Collaboration), COSMOABC: Likelihood-free inference via population Monte Carlo approximate Bayesian computation, Astron. Computing **13**, 1 (2015).

[11] E. Jennings and M. Madigan, astroABC : An approximate Bayesian computation sequential Monte Carlo sampler for cosmological parameter estimation, Astron. Computing **19**, 16 (2017).

[12] M. U. Gutmann and J. Corander, Bayesian optimization for likelihood-free inference of simulator-based statistical models, J. Machine Learning Res. **17**, 1 (2016).

[13] M. Järvenpää, M. Gutmann, A. Vehtari, and P. Marttinen, Gaussian process modeling in approximate Bayesian computation to estimate horizontal gene transfer in bacteria, arXiv:1610.06462.

[14] M. Todorović, M. U. Gutmann, J. Corander, and P. Rinke, Efficient Bayesian inference of atomistic structure in complex functional materials, arXiv:1708.09274.

[15] A. Kangasrääsiö, K. Athukorala, A. Howes, J. Corander, S. Kaski, and A. Oulasvirta, CHI '17 Proceedings of the 2017 CHI conference on human factors in computing systems, Inferring cognitive models from data using approximate Bayesian computation, arXiv:1612.00653.

[16] M. Järvenpää, M. U. Gutmann, A. Pleska, A. Vehtari, and P. Marttinen, Efficient acquisition rules for model-based approximate Bayesian computation, arXiv:1704.00520.

[17] M. Betoule *et al.*, Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples, Astron. Astrophys. **568**, A22 (2014).

[18] S. N. Wood, Statistical inference for noisy nonlinear ecological dynamic systems, Nature (London) **466**, 1102 (2010).

[19] L. F. Price, C. C. Drovandi, A. Lee, and D. J. Nott, Bayesian synthetic likelihood, J. Comput. Graph. Stat. **27**, 1 (2018).

[20] E. Sellentin and A. F. Heavens, Parameter inference with estimated covariance matrices, Mon. Not. R. Astron. Soc. **456**, L132 (2016).

[21] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, Adaptative Computation and Machine Learning Series (University Press Group Limited, 2006).

[22] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, A limited memory algorithm for bound constrained optimization, SIAM J. Sci. Comput. **16**, 1190 (1995).

[23] E. Brochu, V. M. Cora, and N. de Freitas, A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning, arXiv:1012.2599.

[24] I. M. Sobol, On the distribution of points in a cube and the approximate evaluation of integrals, USSR computational mathematics and mathematical physics **7**, 86 (1967).

[25] G. Papamakarios and I. Murray, Fast ε-free inference of simulation models with Bayesian conditional density estimation, arXiv:1605.06376.

[26] J. Alsing, B. Wandelt, and S. Feeney, Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology, Mon. Not. R. Astron. Soc. **477**, 2874 (2018).

[27] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, emcee: The MCMC Hammer, Publ. Astron. Soc. Pac. **125**, 306 (2013).

[28] A. Gelman and D. B. Rubin, Inference from iterative simulation using multiple sequences, Stat. Sci. **7**, 457 (1992).

[29] B. D. Wandelt, D. L. Larson, and A. Lakshminarayanan, Global, exact cosmic microwave background data analysis using Gibbs sampling, Phys. Rev. D **70**, 083511 (2004).

[30] H. K. Eriksen, I. J. O'Dwyer, J. B. Jewell, B. D. Wandelt, D. L. Larson, K. M. Górski, S. Levin, A. J. Banday, and P. B. Lilje, Power spectrum estimation from high-resolution maps by Gibbs sampling, Astrophys. J. Suppl. Ser. **155**, 227 (2004).

[31] J. Jasche, F. S. Kitaura, B. D. Wandelt, and T. A. Enßlin, Bayesian power-spectrum inference for large-scale structure data, Mon. Not. R. Astron. Soc. **406**, 60 (2010).

[32] J. Jasche and G. Lavaux, Matrix-free large-scale Bayesian inference in cosmology, Mon. Not. R. Astron. Soc. **447**, 1204 (2015).

[33] J. Jasche, F. Leclercq, and B. D. Wandelt, Past and present cosmic structure in the SDSS DR7 main sample, J. Cosmol. Astropart. Phys. 01 (2015) 036.

[34] J. Alsing, A. Heavens, A. H. Jaffe, A. Kiessling, B. Wandelt, and T. Hoffmann, Hierarchical cosmic shear power spectrum inference, Mon. Not. R. Astron. Soc. **455**, 4452 (2016).

[35] V. M. H. Ong, D. J. Nott, M.-N. Tran, S. A. Sisson, and C. C. Drovandi, Likelihood-free inference in high dimensions with synthetic likelihood (2017).
[36] Z. Wang, F. Hutter, M. Zoghi, D. Matheson, and N. De Freitas, Bayesian optimization in a billion dimensions via random embeddings, arXiv:1301.1942.
[37] K. Kandasamy, J. Schneider, and B. Póczos, High dimensional Bayesian optimisation and bandits via additive models, arXiv:1503.01673.
[38] F. Elsner and B. D. Wandelt, Efficient Wiener filtering without preconditioning, Astron. Astrophys. **549**, A111 (2013).
[39] D. Kodi Ramanah, G. Lavaux, and B. D. Wandelt, Wiener filter reloaded: Fast signal reconstruction without preconditioning, Mon. Not. R. Astron. Soc. **468**, 1782 (2017).
[40] J. Papez, L. Grigori, and R. Stompor, Solving linear equations with messenger-field and conjugate gradients techniques - an application to CMB data analysis, arXiv:1803.03462.
[41] C. Park and D. Apley, Patchwork kriging for large-scale gaussian process regression, arXiv:1701.06655.
[42] A. F. Heavens, R. Jimenez, and O. Lahav, Massive lossless data compression and multiple parameter estimation from galaxy spectra, Mon. Not. R. Astron. Soc. **317**, 965 (2000).
[43] J. Alsing and B. Wandelt, Generalized massive optimal data compression, Mon. Not. R. Astron. Soc. **476**, L60 (2018).
[44] T. Charnock, G. Lavaux, and B. D. Wandelt, Automatic physical inference with information maximizing neural networks, Phys. Rev. D **97**, 083004 (2018).
[45] A. F. Heavens, E. Sellentin, D. de Mijolla, and A. Vianello, Massive data compression for parameter-dependent covariance matrices, Mon. Not. R. Astron. Soc. **472**, 4244 (2017).
[46] D. Gualdi, H. Gil-Marín, R. L. Schuhmann, M. Manera, B. Joachimi, and O. Lahav, Enhancing BOSS bispectrum cosmological constraints with maximal compression, arXiv:1806.02853.
[47] C. Hahn, F. Beutler, M. Sinha, A. Berlind, S. Ho, and D. W. Hogg, Likelihood non-Gaussianity in large-scale structure analyses, arXiv:1803.06348.
[48] C. P. Robert, J.-M. Cornuet, J.-M. Marin, and N. Pillai, Lack of confidence in approximate Bayesian computation model choice, Proc. Natl. Acad. Sci. U.S.A. **108**, 15112 (2011).
[49] X. Didelot, R. G. Everitt, A. M. Johansen, and D. J. Lawson, Likelihood-free estimation of model evidence, Bayesian Anal. **6**, 49 (2011).
[50] F. Leclercq, J. Jasche, and B. Wandelt, Bayesian analysis of the dynamic cosmic web in the SDSS galaxy survey, J. Cosmol. Astropart. Phys. 06 (**2015**) 015.
[51] F. Leclercq, G. Lavaux, J. Jasche, and B. Wandelt, Comparing cosmic web classifiers using information theory, J. Cosmol. Astropart. Phys. 08 (**2016**) 027.
[52] F. Leclercq, J. Jasche, G. Lavaux, B. Wandelt, and W. Percival, The phase-space structure of nearby dark matter as constrained by the SDSS, J. Cosmol. Astropart. Phys. 06 (2017) 049.
[53] J. Lintusaari, H. Vuollekoski, A. Kangasrääsiö, K. Skytén, M. Järvenpää, M. Gutmann, A. Vehtari, J. Corander, and S. Kaski, ELFI: Engine for likelihood free inference, arXiv:1708.00707.
[54] M. Sako *et al.*, The data release of the Sloan Digital Sky Survey-II Supernova survey, Publ. Astron. Soc. Pac. **130**, 064002 (2018).
[55] A. Conley *et al.*, Supernova constraints and systematic uncertainties from the first three years of the Supernova legacy survey, Astrophys. J. Suppl. Ser. **192**, 1 (2011).
[56] M. Hicken *et al.*, CfA3: 185 Type Ia Supernova light curves from the CfA, Astrophys. J. **700**, 331 (2009).
[57] P. Astier *et al.*, The Supernova legacy survey: Measurement of Ω_M, Ω_Λ and w from the first year data set, Astron. Astrophys. **447**, 31 (2006).
[58] M. Sullivan *et al.*, SNLS3: Constraints on dark energy combining the supernova legacy survey three-year data with other probes, Astrophys. J. **737**, 102 (2011).
[59] A. G. Riess *et al.*, New Hubble space telescope discoveries of type Ia supernovae at $z > 1$: Narrowing constraints on the early behavior of dark energy, Astrophys. J. **659**, 98 (2007).
[60] D. W. Hogg, Distance measures in cosmology, arXiv:astro-ph/9905116.