Bayesian optimisation 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

(Dated: 21 May 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 propose 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, our strategy combines Gaussian process regression of the discrepancy to build a surrogate surface with Bayesian optimisation to actively acquire training data. We derive and make use of an acquisition function tailored for the purpose of minimising the expected uncertainty in the approximate posterior density. The resulting algorithm (Bayesian optimisation for likelihood-free inference, BOLFI) is applied to the problems of summarising 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.

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 non-linear dynamical system involving an unrestricted number of latent variables. Likelihood-free inference methods, also known as Approximate Bayesian computation (ABC, see Marin et al., 2012; Lintusaari et al., 2017, 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 (Weyant, Schafer & Wood-Vasey, 2013) and weak lensing peak counts (Lin & Kilbinger, 2015), analysing the galaxy halo-connection (Hahn et al., 2017), inferring the photometric and size evolution of galaxies (Carassou et al., 2017), measuring cosmological redshift distributions (Kacprzak et al., 2018), estimating the ionising background from the Lyman-α and Lyman-β forests (Davies et al., 2018).

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 $\epsilon$. 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 Akeret et al., 2015; Ishida et al., 2015; Jennings & Madigan, 2017, 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 $O(10^4 - 10^6)$ 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 introduce BOLFI (Bayesian optimisation for likelihood-free inference), an algorithm which combines probabilistic modelling of the discrepancy with optimisation to facilitate likelihood-free inference (Gutmann & Corander, 2016). We focus on computable parametric approximations to the true likelihood, rendering the approach completely $\epsilon$-free. We extend the work of Järvenpää et al. (2017) to propose an acquisition function for Bayesian optimisation specifically tailored to perform efficient and accurate ABC. 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 minimised. As a result, we show that high-fidelity posterior inferences can be obtained with orders of magnitude fewer simulations than with likelihood-free...
of the simulator which is minimised in parameter space by our acquisition rule.

The structure of this paper is as follows. In section II, we provide a review of the formalism for the inference of simulator-based statistical models. In section III, we introduce BOLFI and discuss the regression and optimisation strategies. In particular, we provide the appropriate acquisition rule for ABC in the parametric approach to likelihood approximation. Applications are given in section IV. The developed method is discussed in section 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 (figure 2), where $\theta$ are the parameters of interest, and $d$ the simulated data.

The simplest case (figure 2, left) is when the simulator is a deterministic function of its input and does not use any random variable, i.e.

$$P(\theta|d) = \delta(d - \hat{d}(\theta)).$$

(1)

In a more generic scenario (figure 2, right), the simulator is stochastic, in the sense that the data are drawn from an overall (but often unknown analytically) probability distribution function (pdf) $P(d|\theta)$. Equation (1) does not hold in this case. The scatter between different realisations of $d$ given the same $\theta$ 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 $d$ 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 realisation 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

\[
P(\theta | d) = \frac{P(d | \theta) P(\theta)}{P(d)},
\]

for the observed data \(d_o\), i.e.

\[
P(\theta | d)_{d=d_o} = L(\theta) \frac{P(\theta)}{Z_d},
\]

where the exact likelihood for the problem is defined as

\[
L(\theta) \equiv P(d | \theta)_{d=d_o}.
\]

It is generally of unknown analytical form. The normalisation constant is \(Z_d \equiv P(d)_{d=d_o}\).

C. Approximate Bayesian computation

![Diagram of hierarchical representation of approximate Bayesian inference problem for simulator-based statistical models, with a compression of the raw data to a set of summary statistics.]

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

On a physical and statistical level, the approximation consists of compressing the full data \(d_o\) to a set of summary statistics \(\Phi_o\) before performing inference. Similarly, simulated data \(d_s\) are compressed to simulated summary statistics \(\Phi_s\). This can be seen as adding a layer to the Bayesian hierarchical model (figure 3). The purpose of this operation is to filter out the information in \(d\) that is not deemed relevant to the inference of \(\theta\), so as to reduce the dimensionality of the problem. Ideally, \(\Phi\) should be sufficient for parameters \(\theta\), i.e. formally \(P(\theta | \Phi) = P(\theta | d)\) or equivalently \(P(d | \Phi, \theta) = P(d | \Phi)\), which happens when the compression is lossless. However, sufficient summary statistics are generally unknown or even impossible to design; therefore the compression from \(d\) to \(\Phi\) will usually be lossy. The approximate inference problem to be solved is now \(P(\theta | \Phi) = P(\Phi | \theta) \frac{P(\theta)}{P(\Phi)}\) for the observed summary statistics \(\Phi_o\), i.e.

\[
P(\theta | \Phi)_{\Phi=\Phi_o} = L(\theta) \frac{P(\theta)}{Z_{\Phi}}.
\]

In other words, \(L(\theta)\) is replaced by

\[
L(\theta) \equiv P(\Phi | \theta)_{\Phi=\Phi_o},
\]

and \(Z_{\Phi}\) by \(Z_{\Phi} \equiv P(\Phi)_{\Phi=\Phi_o}\). Inference of model 3 gives

\[
P(\theta, d | \Phi) \propto P(d | \Phi, \theta) P(\theta, d)\]

with, after marginalisation over \(d\),

\[
P(\theta | \Phi) \propto \int P(\theta, d | \Phi) \, dd.
\]

Therefore, the approximate likelihood \(L(\theta)\) must satisfy

\[
L(\theta) \propto \int P(\Phi | d) P(\Phi | \theta) P(d) \, dd.
\]

In many cases, the compression from \(d\) to \(\Phi\) is deterministic, i.e.

\[
P(\Phi | d) = \delta_D(\Phi - \Phi(d)),
\]

which simplifies the integral over \(d\) in equations (8) and (9).

On a practical level, \(L(\theta)\) is still of unknown analytical form (which is a property of \(P(\Phi | \theta)\) inherited from \(P(d | \theta)\) in model 3). Therefore, it has to be approximated using the simulator. We denote by \(\tilde{L}^N(\theta)\) an estimate of \(L(\theta)\) computed using \(N\) realisations of the simulator. The limiting approximation, in the case where infinite computer resources were available, is denoted by \(\tilde{L}(\theta)\), such that

\[
\tilde{L}^N(\theta) \xrightarrow{N \rightarrow \infty} \tilde{L}(\theta).
\]

Note that \(\tilde{L}(\theta)\) can be different from \(L(\theta)\), depending on the assumptions made to construct \(\tilde{L}^N(\theta)\). These are discussed in section 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., $\Phi_\theta$ is an entirely deterministic function of $\theta$ (see figure 2, left). Equivalently, all the conditional probabilities appearing in equation (7) reduce to Dirac delta distributions given by equations (1) and (10). In this case, one can directly use the approximate likelihood given by equation (6), complemented by an assumption on the functional shape of $P(\Phi|\theta)$.

2. Parametric approximations and the synthetic likelihood

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

For example, if $\Phi_\theta$ 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. 

$$L(\theta) = \exp \left[ \ell(\theta) \right]$$

$$-2\ell(\theta) \equiv \log |2\pi \Sigma_\theta| + (\Phi_\theta - \mu_\theta)^T \Sigma_\theta^{-1} (\Phi_\theta - \mu_\theta),$$

where the mean and covariance matrix,

$$\mu_\theta \equiv E[\Phi_\theta] \quad \text{and} \quad \Sigma_\theta \equiv E[(\Phi_\theta - \mu_\theta)(\Phi_\theta - \mu_\theta)^T],$$

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

$$\hat{\mu}_\theta \equiv E[N(\Phi_\theta)] \quad \text{and} \quad \hat{\Sigma}_\theta \equiv E[N((\Phi_\theta - \hat{\mu}_\theta)(\Phi_\theta - \hat{\mu}_\theta)^T)].$$

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

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

where $\Delta_\theta$ is a non-negative function of $\Phi_\theta$ and $\Phi_\theta$ (usually with a maximum at zero). A computable approximation of the likelihood is then given by

$$\tilde{L}^N(\theta) \equiv E^N[\kappa(\Delta_\theta)].$$

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

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

where $P^N(\Delta_\theta \leq \varepsilon)$ is the empirical probability that the discrepancy is below the threshold. $\tilde{L}^N(\theta)$ can be straightforwardly evaluated by running simulations, computing $\Delta_\theta$ and using $\Delta_\theta \leq \varepsilon$ as a criterion for acceptance or rejection of proposed samples. Such an approach is usually referred to as Approximate Bayesian computation (ABC) in the astrophysics literature, although a more explicit denomination would be likelihood-free rejection sampling (see e.g. Marin et al., 2012).

It is interesting to note that the parametric approximate likelihood approach of section II D 2 can be embedded into the non-parametric approach. Indeed, $\Delta_\theta$ can be defined as

$$\Delta_\theta^{C_\theta} = \log |2\pi C_\theta| + (\Phi_\theta - \Phi_\theta)^T C_\theta^{-1} (\Phi_\theta - \Phi_\theta)$$

for some positive semi-definite matrix $C_\theta$. The second term is the square of the Mahalanobis distance, which has already been applied successfully to perform approximate inference in several scientific fields. However, as pointed out by Sellentin & Heavens (2016), for inference from Gaussian-distributed summaries $\Phi_\theta$ with an estimated covariance matrix $\hat{\Sigma}_\theta$, 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 section IV A and appendix B, we extend previous work on the Gaussian synthetic likelihood and introduce a Gamma synthetic likelihood for cases where the $\Phi_\theta$ are (or can be assumed to be) Gamma-distributed.

3. Non-parametric approximations and likelihood-free rejection sampling
includes the Euclidean distance as a special case, when \( C_\theta \) is the identity matrix. Using an exponential kernel \( \kappa(u) = \exp(-u/2) \) and \( C_\theta = \hat{\Sigma}_\theta \) gives \( \hat{L}(\theta) = \mathbb{E}[\kappa(\Delta_{\theta}^{C_\theta})] \) and \( \hat{L}^N(\theta) = \mathbb{E}^n[\kappa(\Delta_{\theta}^{C_\theta})] \) with
\[
-2\log[\kappa(\Delta_{\theta}^{C_\theta})] = \log|2\pi\hat{\Sigma}_\theta|
+ (\Phi_O - \Phi_\theta)^T\hat{\Sigma}_\theta^{-1}(\Phi_O - \Phi_\theta),
\]
the form of which is similar to equation (15). In fact, Gutmann & Corander (2016, proposition 1) show that the synthetic likelihood satisfies
\[
\hat{\ell}(\theta) = J(\theta) + \text{constant}, \quad \text{and}
\hat{\ell}^N(\theta) = \hat{J}^N(\theta) + \text{constant},
\]
where
\[
J(\theta) \equiv \mathbb{E}[\Delta_{\theta}^{C_\theta}]
\]
and
\[
\hat{J}^N(\theta) \equiv \mathbb{E}^n[\Delta_{\theta}^{C_\theta}]
\]
are respectively the expectation and the empirical average of the discrepancy \( \Delta_{\theta}^{C_\theta} \), for \( C_\theta = \hat{\Sigma}_\theta \).

III. REGRESSION AND OPTIMISATION FOR LIKELIHOOD-FREE INFERENCE

A. Computational difficulties with likelihood-free rejection sampling

We have seen in section II D that computable approximations \( \hat{L}^N(\theta) \) of the likelihood \( L(\theta) \) 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 (section II D 3), the expectations are approximated by empirical probabilities that the discrepancy is below the threshold \( \varepsilon \). 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 \( \varepsilon \) is small, leading to a computationally inefficient algorithm.

2. It does not make assumptions about the shape or smoothness of the target function \( L(\theta) \), hence accepted samples can not “share” information in parameter space.

3. It uses a fixed proposal distribution (typically the prior \( P(\theta) \)) 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 & Corander (2016), who combine regression of the discrepancy (addressing issues 1 and 2) with Bayesian optimisation (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 & Corander (2016) for a treatment of the non-parametric approach.

B. Regression of the discrepancy

The standard approach to obtain a computable approximate likelihood relies on empirical averages (equations (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, \( \hat{\ell}(\theta) \) can be interpreted as a regression function with the model parameters \( \theta \) (the “predictors”) as the independent input variables and the discrepancy \( \Delta_{\theta} \) as the response variable. Therefore, in the present approach, we consider an approximation of the intractable expectation defining \( \hat{J}(\theta) \) in equation (23) based on a regression analysis of \( \Delta_{\theta} \), instead of sample averages. Explicitly, we consider
\[
\hat{J}^{(t)}(\theta) \equiv \mathbb{E}^{(t)}[\Delta_{\theta}^{C_\theta}],
\]
FIG. 4. Illustration of Gaussian process regression in one-dimension, for the target test-function \( f : \theta \mapsto 2 - \exp\left(\frac{-1}{2}\right) - \exp\left(\frac{-1}{10}\right) \) (dashed line). Training data are acquired (red dots); they are subject to a Gaussian observation noise with standard deviation \( \sigma_n = 0.03 \). 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.
where the superscript (t) stands for “training” and the expectation \( E^{(t)} \) is taken under the probabilistic model defined in the following.

Inferring \( J(\theta) \) via regression requires a training data set \( \{ (\theta^{(i)}, \Delta_\theta^{(i)}) \} \) where the discrepancies are computed from the simulated summary statistics \( \Phi^{(i)}_\theta \). 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 section III A). A regression-based approach also allows incorporating a smoothness assumption about \( J(\theta) \). In this way, samples of the training set can “share” the information of the computed \( \Delta_\theta \) in the neighbourhood of \( \theta \) (thus addressing issue 2). This suggests that fewer simulated data are needed to reach a certain level of accuracy when learning the target function \( J(\theta) \).

In this work, we rely on Gaussian process (GP) regression in order to construct a prediction for \( J(\theta) \). 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 complex non-linear, or multi-modal features. Second, GPs allow extrapolating the prediction into regions is useful for actively constructing the training data via rejection sampling (thus addressing issue 1, see section III A). A regression-based approach also allows incorporating a smoothness assumption about \( J(\theta) \). In this way, samples of the training set can “share” the information of the computed \( \Delta_\theta \) in the neighbourhood of \( \theta \) (thus addressing issue 2). This suggests that fewer simulated data are needed to reach a certain level of accuracy when learning the target function \( J(\theta) \).

We now briefly review Gaussian process regression. Suppose that we have a set of \( t \) training points, \( \{(\theta^{(i)}, f^{(i)})\} \), of the function \( f \) that we want to regress. We assume that \( f \) is a Gaussian process with prior mean function \( m(\theta) \) and covariance function \( \kappa(\theta, \theta') \) also known as the kernel (see Rasmussen & Williams, 2006). The joint probability distribution of the training set is therefore \( P(f|\Theta) \propto \exp[\ell(f|\Theta)] \) where the exponent \( \ell(f|\Theta) \) is

\[
\frac{1}{2} \sum_{i,j=1}^{t} \left[ f^{(i)} - m(\theta^{(i)}) \right]^T \kappa(\theta^{(i)}, \theta^{(j)})^{-1} \left[ f^{(j)} - m(\theta^{(j)}) \right].
\]

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

\[
\kappa(\theta, \theta') \equiv \sigma^2 \exp\left[ -\frac{1}{2} \sum_p \left( \frac{\theta_p - \theta'_p}{\lambda_p} \right)^2 \right] + \sigma^2_n \delta(\theta, \theta').
\]

The \( \lambda_p \) and \( \theta'_p \) are the components of \( \theta \) and \( \theta' \), respectively. In the last term, \( \delta(\theta, \theta') \) is one if and only if \( \theta = \theta' \) and zero otherwise. The hyperparameters are \( C \), the \( \lambda_p \) (the length scales controlling the amount of correlation between points, and hence the allowed wiggliness of \( f \)), \( \sigma^2_f \) (the signal variance, i.e. the marginal variance of \( f \) at a point \( \theta \) if the observation noise was zero), and \( \sigma^2_n \) (the observation noise). For the results of this paper, GP hyperparameters were learned from the training set using L-BFGS (Byrd et al., 1995), a popular optimiser for machine learning, and updated every time the training set was augmented with 10 samples.

The predicted value \( f_\ast \) at a new point \( \theta_\ast \) can be obtained from the fact that \( (\{\Theta, \theta_\ast\}, \{f, f_\ast\}) \) form jointly a random realisation of the Gaussian process \( f \). Thus, the target pdf \( P(f_\ast|f, \Theta, \theta_\ast) \) can be obtained from conditioning the joint pdf \( P(f, f_\ast|\Theta, \theta_\ast) \) to the values of the training set \( f \). The result is (see Rasmussen & Williams, 2006, section 2.7)

\[
P(f_\ast|f, \Theta, \theta_\ast) \propto \exp\left[ -\frac{1}{2} \left( \frac{f_\ast - \mu(\theta_\ast)}{\sigma(\theta_\ast)} \right)^2 \right],
\]

\[
\mu(\theta_\ast) \equiv m(\theta_\ast) + \kappa^\top \kappa^{-1}(f - m),
\]

\[
\sigma^2(\theta_\ast) \equiv K_{\ast\ast} - \kappa^\top \kappa^{-1} \kappa_{\ast},
\]

where we use the definitions

\[
K_{\ast\ast} \equiv \kappa(\theta_\ast, \theta_\ast),
\]

\[
m \equiv \left( m(\theta^{(i)}) \right)^\top \text{ for } \theta^{(i)} \in \Theta,
\]

\[
\kappa \equiv \left( \kappa(\theta_\ast, \theta^{(i)}) \right)^\top \text{ for } \theta^{(i)} \in \Theta,
\]

\[
(K)_{ij} \equiv \kappa(\theta^{(i)}, \theta^{(j)}) \text{ for } \{\theta^{(i)}, \theta^{(j)}\} \in \Theta^2.
\]

### C. Bayesian optimisation

The second major ingredient of the proposed approach is Bayesian optimisation, which allows the inference of the regression function \( J(\theta) \) while avoiding unnecessary computations. It allows active construction of the training data set \( \{ (\theta^{(i)}, \Delta_\theta^{(i)}) \} \), updating the proposal of new points using the regressed \( \hat{J}^{(i)}(\theta) \) (thus addressing issue 3 with likelihood-free rejection sampling, see section III A). Further, since we are mostly interested in the regions of the parameter space where no training points are available. These three properties are shown in figure 4 for a multi-modal test function subject to observation noise.
In the following, we give a brief overview of the elements of Bayesian optimisation used in this paper. In order to add a new point to the training data set \((\Theta, f) \equiv \{(\theta^{(i)}, f^{(i)}) = f(\theta^{(i)})\}\), Bayesian optimisation uses an acquisition function \(A(\theta)\) that estimates how useful the evaluation of the simulator at \(\theta\) 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 \((\Theta, f)\), i.e. from the mean prediction \(\mu(\theta)\) and the uncertainty \(\sigma(\theta)\) of the regression analysis (equations (29) and (30)). The optimum of

(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 optimisation 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 Brochu, Cora & de Freitas, 2010, for a review). Figure 5 illustrates Bayesian optimisation in combination with Gaussian process regression, applied to finding the minimum of the test function of figure 4.
the acquisition function in parameter space determines the next point \( \theta_* \equiv \text{argopt}_g A(\theta) \) 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 \( (\theta_*, f(\theta_*)) \). 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 initialised 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 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 quasi-random Sobol sequence (Sobol, 1967), which covers the parameter space more evenly.

### D. Expressions for the approximate posterior

As discussed in section III B, using \( \Delta \theta_{\text{c.o}} \) as the regressed quantity directly gives an estimate of \( J(\theta) \) in equation (23). The response variable is thus \( f(\theta) \equiv \Delta \theta_{\text{c.o}} \) and the regression then gives,

\[
\hat{J}(\theta) = \mu(\theta).
\]

In the parametric approach to likelihood approximation, this is equivalent to an approximation of \( -2 \ell(\theta) = -2 \log L(\theta) \) (see equation (21)). The expectation of the (unnormalised) approximate posterior is therefore directly given as (see equation (5))

\[
E(\theta | \Phi_\Theta, f, \Theta) \equiv P(\theta) \exp \left( -\frac{1}{2} \mu(\theta) \right),
\]

where \( P_B(\theta|\Phi_\Theta,f,\Theta) \approx Z_\theta \times P(\theta|\Phi_\Theta)P_\Phi=\Phi_\Theta \).

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

\[
V(\theta | \Phi_\Theta, f, \Theta) \equiv \frac{P(\theta)^2}{4} \exp \left( -\mu(\theta) \right) \sigma^2(\theta).
\]

These expressions were already used in figure 1. Details of the computations can be found in appendix A 1.

Expressions for the BOLFI posterior in the non-parametric approach with the uniform kernel can also be derived (Järvenpää et al., 2017, 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 optimisation 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. Brochu, Cora & de Freitas, 2010), in this work we discuss the acquisition function known as expected improvement (EI). The improvement is defined by \( I(\theta_*) = \max \{ \min(f) - f(\theta_*), 0 \} \), and the expected improvement is \( \text{EI}(\theta_*) \equiv E(I(\theta_*)) \), where the expectation is taken with respect to the random observation assuming decision \( \theta_* \). For a Gaussian process regressor, this evaluates to (see Brochu, Cora & de Freitas, 2010, section 2.3)

\[
\text{EI}(\theta_*) \equiv \sigma(\theta_*) \left[ z \Phi(z) + \phi(z) \right], \\
\text{with } z \equiv \frac{\min(f) - \mu(\theta_*)}{\sigma(\theta_*)},
\]

or \( \text{EI}(\theta_*) \equiv 0 \) if \( \sigma(\theta_*) = 0 \), where \( \phi \) and \( \Phi \) 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 \( \theta_* \) that maximises \( \text{EI}(\theta_*) \).

The EI criterion can be interpreted as follows: since the goal is to find the minimum of \( f \), a reward equal to the improvement \( \min(f) - f(\theta_*) \) is received if \( f(\theta_*) \) is smaller than all the values observed so far, otherwise no reward is received. The first term appearing in equation (38) is maximised when evaluating at points with high uncertainty (exploration); and, at fixed variance, the second term is maximised 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 Järvenpää et al. (2017), in Bayesian optimisation for Approximate Bayesian computation, the goal should not be to find the minimum of \( J(\theta) \), but rather to minimise the expected uncertainty in the estimate of the approximate posterior over the future evaluation of the simulator at \( \theta_* \). 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 minimise the expected variance of the future posterior density \( P_B(\theta_*|\Phi_\Theta,f,\Theta_*) \) over the parameter space. The framework used is Bayesian decision theory. Formally, the loss due to our uncertain knowledge of the approxi-
mate posterior density can be defined as

$$\mathcal{L}[P_{\text{BOLFI}}(\theta|\Phi_O, f, \Theta)] = \int V^{(1)}[P_{\text{BOLFI}}(\theta|\Phi_O, f, \Theta)] \, d\theta,$$

and the acquisition rule is to select the location $\theta_*$ that minimises

$$\text{EIV}(\theta_*) = E^{(1)}[\mathcal{L}[P_{\text{BOLFI}}(\theta|\Phi_O, f, \Theta, f_*, \theta_*)]]$$

$$= \int \mathcal{L}[P_{\text{BOLFI}}(\theta|\Phi_O, f, \Theta, f_*, \theta_*)] \, P(f_*|f, \Theta, \theta) \, df_*$$

(39)

(40)

with respect to $\theta_*$, where we have to marginalise over the unknown simulator output $f_*$ using the probabilistic model $P(f_*|f, \Theta, \theta_*)$ (equations (28)–(30)).

Järvenpää et al. (2017, proposition 3.2) derive the expressions for the expected integrated variance for a GP model in the non-parametric 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 $\theta_*$ is given by

$$\text{EIV}(\theta_*) = \int \frac{P(\theta)^2}{4} \, \exp[-\mu(\theta)] \, [\sigma^2(\theta) - \tau^2(\theta, \theta_*))] \, d\theta,$$

(41)

where

$$\tau^2(\theta, \theta_*) \equiv \frac{\text{cov}^2(\theta, \theta_*)}{\sigma^2(\theta_*)},$$

(42)

$$\text{cov}(\theta, \theta_*) \equiv \kappa(\theta, \theta_*) - \mathbf{K}^\top \mathbf{K}^{-1} \mathbf{K},$$

is the GP posterior predicted covariance between the evaluation point $\theta$ in the integral and the candidate location for the next evaluation $\theta_*$. Note that in addition to the notations given by equations (31)–(34), we have introduced the vector

$$\mathbf{K} \equiv \left(\kappa(\theta, \theta^{(i)})\right)^T \text{ for } \theta^{(i)} \in \Theta.$$

(43)

It is of interest to examine when the integrand in equation (41) is small. As for the EI (equation (38)), optimal values are found when the mean of the discrepancy $\mu(\theta)$ is small or the variance $\sigma^2(\theta)$ is large. This effect is what yields the trade-off between exploitation and exploration for the ExpIntVar acquisition rule. However, unlike in standard Bayesian optimisation strategies such as the EI, the trade-off is a non-local process (due to the integration over the parameter space), and also depends on the prior, so as to minimise 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 Järvenpää et al. (2017), it can then be evaluated with Monte Carlo or quasi-Monte Carlo methods such as importance sampling.

3. Stochastic versus deterministic acquisition rules

The above rules do not guarantee that the selected $\theta_*$ is different from a previously-acquired $\theta^{(i)}$. Gutmann & Corander (2016, 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 $\sigma^2_\lambda$ to each component of the optimiser of the acquisition function. More precisely, $\theta_*$ is sampled from the Gaussian distribution $G(\theta_\text{opt}, D)$, where $\theta_\text{opt} = \arg\text{opt}_A(\theta)$ and $D$ is the diagonal covariance matrix of components $(\sigma^2_\lambda)^2$. The $\sigma^2_\lambda$ are chosen to be of order $\lambda_\text{p}/10$.

For a more extensive discussion and comparison of various stochastic and deterministic acquisition rules, the reader is referred to Järvenpää et al. (2017).

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 section IV A, we discuss the toy problem of summarising Gaussian signals (i.e. inferring the unknown mean and/or variance of Gaussian-distributed data). In section 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. Summarising Gaussian signals

A simple toy model can be constructed from the general problem of summarising 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 & Corander (2016), where the data $d$ are a vector of $n$ components drawn from a Gaussian with unknown mean $\mu$ and known variance $\sigma^2_\text{true}$. The empirical
mean $\Phi^1$ is a sufficient summary statistic for the problem of inferring $\mu$. The distribution of simulated $\Phi^1_{\hat{\mu}}$ takes a simple form, $\Phi^1_{\hat{\mu}} \sim \mathcal{G}(\mu, \sigma^2_{\text{true}}/n)$. Using here the true variance, the discrepancy and synthetic likelihood are

$$\Delta^1_\mu = -2\ell^N_1(\mu) = \log \left( \frac{2\pi \sigma^2_{\text{true}}}{n} \right) + n \frac{(\Phi_0^1 - \hat{\mu}_{\text{true}})^2}{\sigma^2_{\text{true}}}, \quad (44)$$

where $\hat{\mu}_{\text{true}}$ is an average of $N$ realisations of $\Phi^1_{\hat{\mu}}$. In figure 1 (lower panel), the black dots show simulations of $\Delta^1_\mu$ for different values of $\mu$. We have $\hat{\mu}_{\text{true}} \sim \mathcal{G}(\mu, \sigma^2_{\text{true}}/(Nn))$, therefore the stochastic process defining the discrepancy can be written

$$\Delta^1_\mu = \log \left( \frac{2\pi \sigma^2_{\text{true}}}{n} \right) + n \frac{(\Phi_0^1 - \mu - g)^2}{\sigma^2_{\text{true}}}, \quad g \sim \mathcal{G}(0, \sigma^2_g), \quad (45)$$

where $\sigma^2_g \equiv \sigma^2_{\text{true}}/(Nn)$. Each realisation of $g$ gives a different mapping $\mu \mapsto \Delta^1_\mu$. In figure 1, we show one such realisation 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 figure 1. The corresponding BOLFI posterior and its variance, defined by equations (36) and (37), is shown in purple in the upper panel.

2. Unknown mean and variance

We now consider the problem where the full data set $d$ is a vector of $n$ components drawn from a Gaussian with unknown mean $\mu$ and unknown variance $\sigma^2$. The aim is the two-dimensional inference of $\Theta \equiv (\mu, \sigma^2)$. Evidently, the true likelihood $\mathcal{L}(\mu, \sigma^2)$ for this problem is the Gaussian characterised by $(\mu, \sigma^2)$. The Gaussian-inverse-Gamma distribution is the conjugate prior for this likelihood. It is described by four parameters. Adopting a Gaussian-inverse-Gamma prior characterised by $(\alpha, \beta, \eta, \lambda)$ yields an Gaussian-inverse-Gamma posterior characterised by $(\alpha', \beta', \eta', \lambda')$ given by equations (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 $N = 10$ realisations of the Gaussian signal, and compresses them to two summary statistics, the empirical mean and variance, respectively $\Phi^1$ and $\Phi^2$. The graphical probabilistic model is given in figure B.1. It is a noise-free simulator without latent variables (of the type given by figure 2, right) completed by a deterministic compression of the full data. Note that the vector $\Phi \equiv (\Phi^1, \Phi^2)$ is a sufficient statistic for the inference of $(\mu, \sigma^2)$. 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 section B.3 using a parametric approach, under the assumptions (exactly verified in this example) that $\Phi^1$ is Gaussian-distributed and $\Phi^2$ is Gamma-distributed. We name it the Gaussian-Gamma synthetic likelihood.

The posterior obtained from likelihood-free rejection sampling is shown in green in figure 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 $-2\ell^N$. 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 figure 6 (right). BOLFI was initialised using a Sobol sequence of 20 members to compute the original surrogate surface, and Bayesian optimisation 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 (Betoule et al., 2014): $d_O = (m_B^{i/O})$ for $i \in [1, 740]$. 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 section 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, DELFI (Density Estimation for Likelihood-Free Inference, Alsing, Wandelt & Feeney, 2018), as the assumptions are very similar.

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 $\Omega_m$ (the matter density of the Universe) and $\omega$ (the equation of state of dark energy, assumed constant), and marginalise over the other four (nuisance) parameters ($\alpha$, $\beta$, $\eta$, $\lambda$).
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.

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 ~ 6 × 10^8 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 ~ 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.

β, M_B, δM). We assume a Gaussian prior,

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

which is roughly aligned with the direction of the well-known Ω_m − w degeneracy. We generated 10^6 samples (out of ~ 6 × 10^9 data model evaluations) of the posterior for the exact six-dimensional Bayesian problem via MCMC (performed using the EMCEE code, Foreman-Mackey et al., 2013), ensuring sufficient convergence to characterise the 3σ contours of the distribution. The prior and the exact posterior are shown in blue and orange, respectively, in figure 7.

For likelihood-free inference, the simulator takes as input Ω_m and w and simulates N realisations 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 C. The observed data d_O and the covariance matrix C are shown in figure C.1.

The approximate posterior obtained from likelihood-free rejection sampling is shown in green in figure 7. It was obtained from 5,000 accepted samples using a (conservative) threshold of ε = 650 on Δ(Ω_m,w), chosen so that the acceptance ratio was not below 0.01. The entire run required ~ 450,000 simulations in total. The ap-

---

1 The final Gelman-Rubin statistic (Gelman & Rubin, 1992) was R − 1 ≤ 5 × 10^{-4} for each of the six parameters.
proximate posterior obtained via BOLFI is shown in red in figure 7. BOLFI was initialised 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 non-trivial banana shape of the posterior. Only the $3\sigma$ contour (which is usually not shown in cosmology papers, e.g. Betoule et al., 2014) notably deviates from the MCMC posterior. This is due to the fact that we used one realisation of the stochastic process defining $\Delta(\Omega_m, w)$ and only $N = 50$ realisations per $(\Omega_m, w)$; the marginalisation over the four nuisance parameters is therefore partial, yielding slightly smaller credible contours. However, a better approximation could be obtained straightforwardly, 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 $\sim 2$ orders of magnitude with respect to likelihood-free rejection sampling ($\sim 450,000$ simulations) and 3 orders of magnitude with respect to MCMC sampling of the exact posterior ($6 \times 10^6$ simulations). It is also interesting to note that BOLFI requires a factor of $\sim 3$ fewer simulations than the recently-introduced DELFI procedure (Alsing, Wandelt & Feeney, 2018), which used 20,000 simulations drawn from the prior for the analysis of the JLA.²

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 $L(\theta)$. The smoothness assumption allows simulations in the training set to “share” information about their value of $\Delta_\theta$ in the neighbourhood of $\theta$, 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 section IV and in the statistical literature (see Gutmann & Corander, 2016).

A second benefit of BOLFI is that it actively acquires training data through Bayesian optimisation. 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 prioritised, 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 $\theta$ regardless of the value of $\Delta_\theta$.

Finally, by focusing on parametric approximations to the exact likelihood, the approach proposed in this work is totally “$\varepsilon$-free”, meaning that no threshold (which is often regarded as an unappealing ad hoc element) is required.

Of course, regressing the discrepancy and optimising the acquisition function are not free of computational cost. However, the runtime 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. Wandelt, Larson & Lakshminarayanan, 2004; Eriksen et al., 2004 for the cosmic microwave background: Jasche et al., 2010; Jasche & Lavaux, 2015; Jasche, Leclercq & Wandelt, 2015 for galaxy clustering; Alsing et al., 2016 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 modelling of cosmological data, including in particular the complicated systematics they experience. However, while the physics and instruments will be more accurately modelled, 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 $d$ to $\Phi$ is included in the simulator (see section II C), summary statistics do not need to be quantities that can be physically mo-

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 $\Omega_m$ and $w$. However, since these are the only two physically-interesting parameters, inference of the nuisance parameters is not deemed crucial for this example.
modelled (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 a 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 dimension space is large, due to the curse of dimensionality, general very difficult when the dimensionality of the parameter space is large, due to the curse of dimensionality, and hence becomes more challenging as the dimension increases. The optimisation 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 figure C.3, bottom right panel).

Nevertheless, Järvenpää et al. (2017) showed on a toy simulation model (a Gaussian) that up to 10-dimensional inference is possible with BOLFI. As usual cosmological models do not include more than 10 free physical parameters, and additional nuisance parameters can be automatically marginalised over, we do not expect this limitation to be a hindrance. Recent advances in high-dimensional Bayesian optimisation (e.g. Wang et al., 2013; Kandasamy, Schneider & Póczos, 2015) could also be exploited. In future work, we will address the problem of high-dimensional likelihood-free inference.

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 \( \mathbf{K} \) of size \( t \times t \), where \( t \) is the size of the training set. The complexity is \( O(t^3) \), 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 Rasmussen & Williams, 2006, chapter 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 (Elsner & Wandelt, 2013; Kodi Ramanah, Lavaux & Wandelt, 2017; Papez, Grigori & Stompor, 2018).

Another strategy would be to divide the regression problem spatially into several patches with a lower number of training points (Park & Apley, 2017). 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 equation (30)). Furthermore, a stationary kernel is assumed. However, depending on the simulator, the discrepancy can show heteroscedasticity (i.e. its variance can depend on \( \theta \) – see for example figure 1). Such cases could be handled by non-stationary GP kernels or different probabilistic models for the discrepancy, allowing an 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 optimisation 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 figure C.3, 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 optimisation of overexploration of the domain boundaries, which becomes more problematic in high dimension.

Acquisition strategies examined so far in the literature
(see Järvenpää et al., 2017, 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 non-myopic 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 section 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 section II C, simulator-based models include a data compression step. The comparison in data space can be made more easily if \( \dim \Phi \) is reduced. In future work, we will therefore aim at combining BOLFI with massive and (close to) optimal data compression strategies. These include MOPED (Heavens, Jimenez & Lavaux, 2000), the score function (Alsing & Wandelt, 2018), or information-maximising neural networks (Charnock, Lavaux & Wandelt, 2018).

C. Parallelisation and computational efficiency

While MCMC sampling has to be done sequentially, BOLFI lends itself to more parallelisation. 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 section 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 \( \Phi_0 \) (through the acquisition function), simulations can nevertheless be re-used as long as summaries \( \Phi_0' \) are saved. This means that if one acquires new data \( \Phi_0'' \), the existing \( \Phi_0'' \) (or a subset of them) can be used to compute the new discrepancy \( \Delta_0(\Phi_0, \Phi_0') \). Building an initial training set in this fashion can massively speed-up the inference of \( P(\theta|\Phi)_{\Phi=\Phi_0'} \), 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, an algorithm for scalable likelihood-free inference in cosmology (Delfi, Alsing, Wandelt & Feeney, 2018) was introduced. 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 by a factor of \( \sim 3 \) for the inference of cosmological parameters (see section IV B) can be interpreted as the effect of the Bayesian optimisation 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. 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 parallelisation 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 \( \theta \) and the summary statistics \( \Phi \) is modelled (via density estimation), while BOLFI focuses on the relation between the parameters \( \theta \) and the discrepancy \( \Delta_0 \) (via regression). Summary statistics are multi-dimensional while the discrepancy is a univariate scalar quantity. Thus, Delfi requires to solve a density estimation problem in \( \dim \theta + \dim \Phi \) (which equals \( 2 \times \dim \theta \) if the compression from Alsing & Wandelt, 2018 is used), while BOLFI requires to solve a regression problem in \( \dim \theta \). Both tasks are expected to become more difficult as \( \dim \theta \) increases (a symptom of the curse of dimensionality, see section V B 1), but the upper limits on \( \dim \theta \) 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, Wandelt & Feeney (2018), 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

\[ Z_{\Phi} = \left( \int P(\Phi|\theta) d\theta \right)_{\Phi=\Phi_O}. \] (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 section 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 already been demonstrated to work in combination with the score function (Alsing & Wandelt, 2018), which is necessary to reduce the dimensionality of \( \Phi \) before estimating the density.\(^3\) The score function produces summaries that are only sufficient up to linear order in the log-likelihood. However, in ABC, care is required to perform model selection if the summary statistics are insufficient. Indeed, Robert et al. (2011, equation 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 \( M_1 \) and \( M_2 \) alone, or even for both of them – even if approximately realised via Alsing & Wandelt’s procedure – does not guarantee sufficiency to compare the two different models \( M_1 \) and \( M_2 \) (Didelot et al., 2011). As the assumptions behind BOLFI do not necessarily necessitate to reduce \( \dim \Phi \) (\( \Delta_{\Phi} \) 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.

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 \( \varepsilon \) 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 introduced a new approach to likelihood-free inference, BOLFI, that uses regression to infer this relation, and optimisation 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 minimised.

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 section V.B. It also allows for novel analyses of cosmological data from fully non-linear simulator-based models, as required for example for the cosmic web (see the discussions in Leclercq, Jasche & Wandelt, 2015; Leclercq et al., 2016; Leclercq et al., 2017). 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.

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

\[ P_{\text{BOLFI}}(\theta|\Phi_O) \equiv P(\theta) \exp \left( \frac{-1}{2} f(\theta) \right) \propto P(\theta) \exp \left( \tilde{\ell}(\theta) \right). \] (A1)

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

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

\[ V[P_{\text{BOLFI}}(\theta|\Phi_O, f, \Theta)] = \left| \frac{\partial}{\partial f} P(\theta) \exp \left( \frac{-1}{2} f \right) \right|^2 V[f] = \frac{P(\theta)^2}{4} \exp(-f) V[f]. \] (A2)

The argument of the exponential is \( -f(\theta) = 2\tilde{\ell}(\theta) \); it should be replaced by its expectation under the model, \( -\mu(\theta) \). The variance of \( f \) under the model
is, by definition, \( V^{(1)}[f] = \sigma^2(\theta) \). The result for \( V^{(1)}[P_{\text{BOFB}}(\theta|\Phi_O, f, \Theta)] \) is therefore given by equation (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 \((\theta_*, f_*)\) is added to the training set \((\Theta, f)\). We denote them by \(\mu_*\) and \(\sigma^2_*\) respectively. These quantities are random functions of \(\theta\) since the new value \(f_*\) is unknown. Assuming that the GP mean is \(m(\theta) = 0\) for simplicity, and using equation (29) with the full training set \(\{(\Theta, f), (\theta_*, f_*)\}\), we get

\[
\mu_*(\theta) = \left( \frac{K}{\kappa(\theta, \theta_*)} \right) \left( \frac{K}{K^T K_*} \right)^{-1} f,
\]

(A3)

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

\[
\mu_*(\theta) = \frac{K^T K_* - f_1}{\kappa(\theta, \theta_*) - K^T K_*} \times
\]

\[
\left[ K_* - \frac{K^T K_*}{\kappa(\theta, \theta_*)} \right]^{-1} f
\]

(A4)

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

\[
\mathcal{P}(\mu_*(\theta)|f, \Theta, \theta_*) \sim \mathcal{G}(\mu(\theta), \tau^2(\theta, \theta_*)),
\]

(A5)

using the notation introduced in equation (42).

Similar calculations for the variance show that

\[
\sigma^2_*(\theta) = \sigma^2(\theta) - \tau^2(\theta, \theta_*),
\]

(A6)

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

\[
w^2(\theta, \theta_*) = \int \frac{1}{4} \exp\left[ -\mu_*(\theta) \right] \sigma^2_*(\theta) \mathcal{P}(f_*|f, \Theta, \theta_*) \, df_* = \mathbb{E}^{(1)} \left[ \frac{1}{4} \exp\left[ -\mu_*(\theta) \right] \sigma^2_*(\theta) \right],
\]

(A9)

that is to say the expectation of \(\frac{1}{4} \exp\left[ -\mu_*(\theta) \right] \sigma^2_*(\theta)\) under the GP model trained with \(\{(\Theta, f), (\theta_*, f_*)\}\). This expectation can be treated using equations (A5) and (A6), assuming that mean and variance are independent: \(\sigma^2_*(\theta)\) becomes deterministically \(\sigma^2(\theta) - \tau^2(\theta, \theta_*)\) under the model. As in section A1, the argument of the exponential, \(\mu_*(\theta)\), is replaced by its mean \(\mu(\theta)\). The final result is

\[
w^2(\theta, \theta_*) = \frac{1}{4} \exp\left[ -\mu(\theta) \right] \left[ \sigma^2(\theta) - \tau^2(\theta, \theta_*) \right].
\]

(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

\[
\frac{d\text{EIV}(\theta_*)}{d\theta_*} = \frac{d}{d\theta_*} \int \mathcal{P}(\theta) \mathcal{P}(\theta_*) \, d\theta
\]

(A11)

where

\[
\frac{\partial w^2(\theta, \theta_*)}{\partial \theta_*} = \frac{\partial}{\partial \theta_*} \left\{ \frac{1}{4} \exp\left[ -\mu(\theta) \right] \left[ \sigma^2(\theta) - \tau^2(\theta, \theta_*) \right] \right\}
\]

\[= -\frac{1}{4} \exp\left[ -\mu(\theta) \right] \frac{\partial \tau^2(\theta, \theta_*)}{\partial \theta_*},
\]

(A12)

with

\[
\frac{\partial \tau^2(\theta, \theta_*)}{\partial \theta_*} = 2 \frac{\partial \text{cov}(\theta, \theta_*)}{\partial \theta_*} \frac{\partial \text{cov}(\theta, \theta_*)}{\partial \theta_*} - \frac{\partial \text{cov}(\theta, \theta_*)}{\partial \theta_*} \frac{\partial \text{cov}(\theta, \theta_*)}{\partial \theta_*},
\]

(A13)

\[
\frac{\partial \text{cov}(\theta, \theta_*)}{\partial \theta_*} = \frac{\partial \text{cov}(\theta, \theta_*)}{\partial \theta_*} - \frac{K^T K_*}{\kappa(\theta, \theta_*)} \frac{\partial K}{\partial \theta_*},
\]

(A14)

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

Appendix B: Summarising Gaussian signals

This appendix gives the details of the problem of summarising Gaussian signals discussed in section IV A.
After the full data $d$ are generated, they can be compressed to summary statistics. A simple choice is the empirical estimator for the mean and (unbiased) variance, defined by

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

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

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

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

where $\Gamma$ is the Gamma distribution parametrised by its shape and scale.

The hierarchical graphical representation of the simulator is shown in figure B.1.

### 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

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

$$
\beta' = \beta + \frac{n\lambda}{\lambda + n} \left(\Phi^1_O - \eta\right)^2 + \frac{n - 1}{2} \Phi^2_O, \quad (B9)
$$

$$
\eta' = \frac{\lambda \eta + n \Phi^1_O}{\lambda + n}, \quad (B10)
$$

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

where $\Phi^1_O$ and $\Phi^2_O$ are the summary statistics of the observed data, defined by applying equations (B5) and (B6) to $d_O$.

For the experiment described in section IV A 1, we have used $n = 10$ and $N = 20$. The data have been generated from ground truth parameters $\mu_{\text{true}} = 0.8$ and $\sigma^2_{\text{true}} = 2.9$. We have measured $\Phi^1_O = 1.3212$, and have chosen a Gaussian prior on $\mu$ 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 section IV A 2, we have used $n = 50$ and $N = 10$. The data have been generated from ground truth parameters $\mu_{\text{true}} = 0.8$ and $\sigma^2_{\text{true}} = 2.9$ (shown as the plus in figure 6). We have measured $\Phi^1_O = 0.9925$ and $\Phi^2_O = 2.8499$. We have chosen a prior with parameters $(\alpha, \beta, \eta, \lambda) = (22, 54, 0, 6)$. The exact posterior has therefore parameters $(\alpha', \beta', \eta', \lambda') = (47, 127.8885, 0.8862, 56)$. 

---

**FIG. B.1.** Hierarchical forward model for the problem of summarising simulated Gaussian signals. The upper part corresponds to the generation of random variables from the two-dimensional Gaussian-inverse-Gamma prior parametrised by $(\alpha, \beta, \eta, \lambda)$: first $\sigma^2$ is drawn from $\mathcal{P}(\sigma^2|\alpha, \beta)$ (an inverse-Gamma distribution with shape parameter $\alpha$ and scale parameter $\beta$), then $\mu$ is drawn from $\mathcal{P}(\mu|\sigma^2, \eta, \lambda)$ (a Gaussian distribution with mean $\eta$ and variance $\sigma^2/\lambda$). A Gaussian-inverse-Gamma distribution with unknown mean and variance. It is a straightforwardly generated by first sampling $\sigma^2$ from $\mathcal{P}(\sigma^2|\alpha, \beta)$ and then by drawing $\mu$ from the Gaussian distribution $\mathcal{G}$ with mean $\eta$ and variance $\sigma^2/\lambda$. A noise-free simulator can be designed for this inference problem by taking the operations successively

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

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

$$
d \sim \mathcal{P}(d|\mu, \sigma^2) = \mathcal{G}(d|\mu, \sigma^2). \quad (B4)
$$
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 equation (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 realisations of the summary statistics. Using the simulator described in section C3, we can generate \( N \) realisations of \( \Phi^1 \) and \( \Phi^2 \) for each pair of input parameters \((\mu, \sigma^2)\). Assuming exchangeability, we can use the Ansatz \( L(\mu, \sigma^2) \equiv L_1(\mu, \sigma^2) L_2(\mu, \sigma^2) \) and \( \hat{L}^N(\mu, \sigma^2) \equiv \hat{L}_1^N(\mu, \sigma^2) \hat{L}_2^N(\mu, \sigma^2) \), or using the log-likelihood,

\[
\hat{\ell}^N(\mu, \sigma^2) \equiv \hat{\ell}_1^N(\mu, \sigma^2) + \hat{\ell}_2^N(\mu, \sigma^2),
\]

where the first term depends only on \( \Phi^1 \) and the second on \( \Phi^2 \). They are discussed successively in the following.

\( \Phi^1 \) is the empirical mean of the independent and identically-distributed components of \( d \), obtained through averaging. As discussed in section IID2, the Gaussian parametric approximation also known as the synthetic likelihood is appropriate in this case. We therefore define

\[
-2\hat{\ell}_1^N(\mu, \sigma^2) \equiv \log |2\pi \hat{v}^1_{(\mu, \sigma^2)}|^2 + \frac{(\Phi^1 - \hat{\mu}^1_{(\mu, \sigma^2)})^2}{\hat{v}^1_{(\mu, \sigma^2)}},
\]

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

\[
\hat{\mu}^1_{(\mu, \sigma^2)} \equiv \mathbb{E}^N \left[ \Phi^1_{(\mu, \sigma^2)} \right],
\]

\[
\hat{v}^1_{(\mu, \sigma^2)} \equiv \mathbb{E}^N \left[ \left( \Phi^1_{(\mu, \sigma^2)} - \hat{\mu}^1_{(\mu, \sigma^2)} \right)^2 \right].
\]
As $P(\Phi^1|\mu, \sigma^2)$ is actually a Gaussian distribution, the equality $\hat{L}_1^1(\mu, \sigma^2) = L_1(\mu, \sigma^2)$ holds without approximation, in the limit of infinite computer resources. From equation (B7), we also have

$$\hat{\mu}_1^{(\mu, \sigma^2)} \sim \mathcal{G} \left( \frac{\sigma^2}{N_n} \right) \text{ and}$$

$$\hat{v}_1^{(\mu, \sigma^2)} \sim \Gamma \left( \frac{N - 1}{2}, \frac{2\sigma^2}{n(N - 1)} \right),$$

(B16)

which allows a closed-form definition of the stochastic process defining $\hat{L}_1^N(\mu, \sigma^2)$.

$\Phi^2$ is the empirical variance of the components of $d$. As noted in equation (B7), $P(\Phi^2|\mu, \sigma^2)$ is a Gamma distribution. Consequently, we introduce for $\Phi^2_0$ a Gamma synthetic likelihood, namely

$$-2\ell_2^N(\mu, \sigma^2) \equiv -2(\hat{k}_{(\mu, \sigma^2)} - 1) \log \Phi^2_0 + \frac{2\Phi^2_0}{\hat{\theta}_{(\mu, \sigma^2)}} + 2\hat{k}_{(\mu, \sigma^2)} \log \hat{\theta}_{(\mu, \sigma^2)} + 2 \log \Gamma(\hat{k}_{(\mu, \sigma^2)}).$$

(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

$$\hat{\mu}_2^{(\mu, \sigma^2)} \approx \hat{k}_{(\mu, \sigma^2)} \hat{\theta}_{(\mu, \sigma^2)}, \text{ and}$$

$$\hat{v}_2^{(\mu, \sigma^2)} \approx \hat{k}_{(\mu, \sigma^2)} \left( \hat{\theta}_{(\mu, \sigma^2)} \right)^2,$$

(B18)

(B19)

where $\hat{\mu}_2^{(\mu, \sigma^2)}$ and $\hat{v}_2^{(\mu, \sigma^2)}$ are the empirical mean and variance of $\Phi^2$, respectively, defined as in equations (B14) and (B15). Solving this system for $\hat{k}_{(\mu, \sigma^2)}$ and $\hat{\theta}_{(\mu, \sigma^2)}$, we obtain the parameters of $\ell_2^N$,

$$\hat{k}_{(\mu, \sigma^2)} \approx \frac{\left( \hat{\mu}_2^{(\mu, \sigma^2)} \right)^2}{\hat{v}_2^{(\mu, \sigma^2)}}, \text{ and}$$

$$\hat{\theta}_{(\mu, \sigma^2)} \approx \frac{\hat{v}_2^{(\mu, \sigma^2)}}{\hat{\mu}_2^{(\mu, \sigma^2)}}.$$

(B20)

(B21)

As $P(\Phi^2|\mu, \sigma^2)$ is known to be a Gamma distribution, we have, as for the first term, $\hat{L}_2^2(\mu, \sigma^2) = L_2(\mu, \sigma^2)$ 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

$$\hat{\mu}_2^{(\mu, \sigma^2)} \sim \Gamma \left( \frac{N(n - 1)}{2}, \frac{2\sigma^2}{N(n - 1)} \right).$$

(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 $\hat{L}_2^N(\mu, \sigma^2)$, can be easily simulated using their defining equations.

The resulting approximate likelihood $\hat{L}_2^N(\mu, \sigma^2)$ is the product of a Gaussian synthetic likelihood for $\Phi^1$ and a Gamma synthetic likelihood for $\Phi^2$. It is shown in figure B2. There, the different panels show that realisations become smoother as $N$ increases, i.e. with more computational resources.

### Appendix C: Supernova cosmology

This appendix gives the details of the data model and the modelling assumptions for the problem of inferring cosmological parameters from the JLA catalog, presented in section 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, Betoule et al., 2014) is a compiled catalogue 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 (Sako et al., 2014) and confirmed as SNe Ia by spectroscopic follow-up observations. The remaining objects come from the earlier C11 compilation (Conley et al., 2011): $118$ are low-$z$ ($z \leq 0.08$) SNe Ia from the third release (Hicken et al., 2009) of photometric data acquired at the Wipple 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, Astier et al., 2006; Sullivan et al., 2011). Finally, $9$ objects are high-redshift SNe Ia ($0.8 \leq z \leq 1.4$) observed by the Hubble Space Telescope (HST; Riess et al., 2007).

For each supernova, the JLA catalogue 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 colour at maximum brightness $C$ and a time-stretching parameter for the lightcurve, $X_1$. In particular, the catalogue includes several estimations of the redshift $z$. In this work, we use $z = z_{CMB}$, the cosmological redshift of the object in the frame of the cosmic microwave background (CMB), including peculiar velocity corrections. For our data vector $d_{Oj}$, we use the estimated B-band peak magnitudes in the restframe, denoted $(m_{B,0}^{k})$ 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 figure C.1 (left). The JLA ca-
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:

\[ m_B = 5 \log_{10} \left[ \frac{D_L(z)}{10 \text{ pc}} \right] + \tilde{M}_B(M_{\text{stellar}}, M_B, \delta M) - \alpha X_1 + \beta C. \quad (C1) \]

The absolute magnitude \( \tilde{M}_B \) depends on the stellar mass of the host galaxy, \( M_{\text{stellar}} \). This dependence is assumed to be captured by the relation (Conley et al., 2011)

\[ \tilde{M}_B(M_{\text{stellar}}, M_B, \delta M) = M_B + \delta M \Theta (M_{\text{stellar}} - 10^{10} M_\odot), \quad (C2) \]

where \( \Theta \) is the Heaviside function and \( M_\odot \) the mass of the Sun. The lightcurve calibration model therefore comprises 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 (\( \Lambda \)CDM hereafter). A \( \Lambda \)CDM Universe is characterised by two physical parameters \( \Omega_m \) (the matter density) and \( w \) (the equation of state of dark energy, assumed constant in time). The luminosity distance appearing in equation (C1) is given by (e.g. Hogg, 1999, section 7)

\[ D_L(z) = \frac{(1+z)c}{H_0} \int_0^z \frac{dz'}{E(z')}, \quad (C3) \]

\[ E(z) = \sqrt{\Omega_m (1+z)^3 + (1-\Omega_m) (1+z)^3(w+1)}, \quad (C3) \]
where $c$ is the speed of light in vacuum and $H_0 \equiv 100 \, h \, \text{km s}^{-1} \, \text{Mpc}^{-1}$.

3. Forward-modelling

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

\[(\Omega_m, w) \sim P(\Omega_m, w|\omega, S),\]  
\[(\alpha, \beta, M_B, \delta M) \sim P(\alpha, \beta, M_B, \delta M|M),\]  
\[D_L(z_O) \sim P(D_L(z_O)|\Omega_m, w),\]  
\[d \sim P(d|D_L(z_O), \alpha, \beta, M_B, \delta M, m_O).\]

The last two steps are deterministic: in equation (C6), the luminosity distance at the observed redshifts is computed via equation (C3), and in equation (C7), the predicted data $d(\Omega_m, w) \equiv \left(m_B^*|\Omega_m, w\right)$ come from equations...
We can therefore write
\[
\mathcal{P}(D_L(z_O) | \Omega_m, w) = \mathcal{P}(d | \hat{D}_L(z_O) - \hat{D}_L(z_O, \Omega_m, w)),
\]
\[
\mathcal{P}(d | \hat{D}_L(z_O), \alpha, \beta, M_B, \delta M, m_O) = \mathcal{P}(d | \hat{D}_L(z_O), \alpha, \beta, M_B, \delta M, m_O),
\]
\[
\mathcal{P}(d | \Omega_m, w, M, m_O) = \mathcal{P}(|d - \hat{D}_L(z_O), \alpha, \beta, M_B, \delta M, m_O|) \times \mathcal{P}(D_L(z_O) - \hat{D}_L(z_O, \Omega_m, w)) \times \mathcal{P}(\alpha, \beta, M_B, \delta M | M) \quad \text{(C8)}
\]

The probability \(\mathcal{P}(\Omega_m, w | \omega, S)\) appearing in equation (C4) is the Gaussian prior given by equation (46), i.e.
\[
\mathcal{P}(\Omega_m, w | \omega, S) \equiv \mathcal{G}(\omega, S) \text{ with }
\]
\[
\omega \equiv \begin{pmatrix} 0.3 \\ 0.75 \end{pmatrix} \quad \text{and} \quad S \equiv \begin{pmatrix} 0.4 & -0.24 \\ -0.24 & 0.75 \end{pmatrix} \quad \text{(C9)}
\]

Finally, \(\mathcal{P}(\alpha, \beta, M_B, \delta M | M)\) is the sampling distribution of nuisance parameters, characterised by hyper-parameters \(M\). 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 figure C.2.

4. Discrepancy

Following Betoule et al. (2014, formula 15), we define the discrepancy between observed and simulated data as
\[
\Delta_{(\Omega_m, w)} \equiv (d_O - \hat{\mu}_{(\Omega_m, w)})^T \mathbf{C}^{-1} (d_O - \hat{\mu}_{(\Omega_m, w)}), \quad \text{(C11)}
\]
where \(\hat{\mu}_{(\Omega_m, w)}\) is the average of \(N\) simulated realisations of \(d_{(\Omega_m, w)} \equiv (m_{B1i, (\Omega_m, w)}^k)\) for \(k \in [1, 740]\). This is equivalent to assuming a Gaussian synthetic likelihood (see section 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. (2014, section 5.5) constructed a covariance matrix \(\mathbf{C}_{(\alpha, \beta)}\) which accounts for the uncertainty in the colour, stretch and redshift of each supernova, depending on the nuisance parameters \(\alpha\) and \(\beta\), but dropped the term \(\log [2\pi \mathbf{C}_{(\alpha, \beta)}]\) from the definition of the discrepancy. Since \(\alpha\) and \(\beta\) 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 Alsing, Wandelt & Feeney, 2018), we assume a fixed covariance matrix \(\mathbf{C}\) where the parameters \(\alpha\) and \(\beta\) 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 figure C.1 (right).

5. Acquisition

For the analysis described in section IV B, we used \(N = 50\) simulations per point (\(\Omega_m, w\)), and the ExpIntVar rule without acquisition noise. Figure C.3 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.

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 ELFI (Engine for Likelihood-Free Inference, Lintusaari et al., 2017) code. The author acknowledges funding from the Imperial College Research Fellowship programme.

REFERENCES

(Akeret et al., 2015) J. Akeret, A. Refregier, A. Amara, S. See- hars, C. Hasner, Approximate Bayesian computation for forward modeling in cosmology, Journal of Cosmology and Astroparticle Physics 8, 043 (2015), arXiv:1504.07245.

(Alsing & Wandelt, 2018) J. Alsing, B. Wandelt, Generalized massive optimal data compression, Mon. Not. R. Astron. Soc. 476, L60 (2018), arXiv:1712.00012.

(Alsing, Wandelt & Feeney, 2018) J. Alsing, B. Wandelt, S. Feeney, Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology, Mon. Not. R. Astron. Soc. 477, 2514 (2018), arXiv:1801.01497.

\(^4\) Available at http://supernovae.in2p3.fr/sdss_snls_jla/ReadMe.html.
on and Computing 19, 16 (2017), arXiv:1608.07606 [astro-ph.IM]

(Kacprzak et al., 2018) T. Kacprzak, J. Herbel, A. Amara, A. Réfrégier, Accelerating Approximate Bayesian Computation with Quantile Regression: application to cosmological redshift distributions, Journal of Cosmology and Astroparticle Physics 2, 042 (2018), arXiv:1707.07498.

(Kandasamy, Schneider & Póczos, 2015) K. Kandasamy, J. Schneider, B. Póczos, High Dimensional Bayesian Optimisation and Bandits via Additive Models, in Proceedings of the 32nd International Conference on Machine Learning, Proceedings of Machine Learning Research, Vol. 37, edited by F. Bach, D. Blei (PMLR, Lille, France, 2015) pp. 295–304.

(Kodi Ramanah, Lavaux & Wandelt, 2017) D. Kodi Ramanah, G. Lavaux, B. D. Wandelt, Wiener filter reloaded: fast signal reconstruction without preconditioning, Mon. Not. R. Astron. Soc. 468, 1782 (2017), arXiv:1702.08852.

(Leclercq, Jasche & Wandelt, 2015) F. Leclercq, J. Jasche, B. Wandelt, Bayesian analysis of the dynamic cosmic web in the SDSS galaxy survey, Journal of Cosmology and Astroparticle Physics 6, 015 (2015), arXiv:1502.02690 [astro-ph.CO].

(Leclercq et al., 2016) F. Leclercq, G. Lavaux, J. Jasche, B. Wandelt, Comparing cosmic web classifiers using information theory, Journal of Cosmology and Astroparticle Physics 8, 027 (2016), arXiv:1606.06758 [astro-ph.CO].

(Leclercq et al., 2017) F. Leclercq, J. Jasche, G. Lavaux, B. Wandelt, W. Percival, The phase-space structure of nearby dark matter as constrained by the SDSS, Journal of Cosmology and Astroparticle Physics 6, 049 (2017), arXiv:1601.00093 [astro-ph.CO].

(Lin & Kilbinger, 2015) C.-A. Lin, M. Kilbinger, A new model to predict weak-lensing peak counts. II. Parameter constraint strategies, Astron. & Astrophys. 583, A70 (2015), arXiv:1506.01076.

(Lintusäari et al., 2017) J. Lintusäari, H. Vuollekoski, A. Kangasrääsiö, K. Skytén, M. Järvenpää, M. Gutmann, A. Vehtari, J. Corander, S. Kaski, ELFI: Engine for Likelihood Free Inference, ArXiv e-prints (2017), arXiv:1708.00707 [stat.ML].

(Marin et al., 2012) J.-M. Marin, P. Pudlo, C. P. Robert, R. Ryder, Approximate Bayesian Computational methods, Statistics and Computing 22, 1167 (2012), arXiv:1101.0955 [stat.CO].

(Papez, Grigori & Stompor, 2018) J. Papez, L. Grigori, R. Stompor, Solving linear equations with messenger-field and conjugate gradients techniques - an application to CMB data analysis, ArXiv e-prints (2018), arXiv:1803.03462.

(Park & Apley, 2017) C. Park, D. Apley, Patchwork Kriging for Large-scale Gaussian Process Regression, ArXiv e-prints (2017), arXiv:1701.06655 [cs.LG].

(Price et al., 2017) L. F. Price, C. C. Drovandi, A. Lee, D. J. Nott, Bayesian Synthetic Likelihood, Journal of Computational and Graphical Statistics, In press (2017).

(Rasmussen & Williams, 2006) C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, Adaptive computation and machine learning series (University Press Group Limited, 2006).

(Riess et al., 2007) A. G. Riess, L.-G. Strolger, S. Casertano, H. C. Ferguson, B. Mobasher, B. Gold, P. J. Challis, A. V. Filippenko, S. Jha, W. Li, J. Tonry, R. Foley, R. P. Kirshner, M. Dickinson, E. MacDonald, D. Eisenstein, M. Livio, J. Younger, C. Xu, T. Dahlén, D. Stern, 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), astro-ph/0611572.

(Robert et al., 2011) C. P. Robert, J.-M. Cornuet, J.-M. Marin, N. Pillai, Lack of confidence in approximate Bayesian computation model choice, Proceedings of the National Academy of Sciences of the United States of America 108, 15112 (2011), arXiv:1102.4432 [stat.ME].

(Sako et al., 2014) M. Sako, B. Bassett, A. C. Becker, P. J. Brown, H. Campbell, R. Cane, D. Cinabro, C. B. D’Andrea, K. S. Dawson, F. DeJongh, D. L. Depoy, B. Dilday, M. Doi, A. V. Filippenko, J. A. Fischer, R. J. Foley, J. A. Frieman, L. Galbany, P. M. Garnavich, A. Goobar, R. R. Gupta, G. J. Hill, B. T. Hayden, R. Hlozek, J. A. Holtzman, U. Hopp, S. W. Jha, R. Kessler, W. Kollatschny, G. Leloudas, J. Marriner, J. L. Marshall, R. Miquel, T. Morokuma, J. Mosher, R. C. Nichol, J. Nordin, M. D. Olmstead, L. Ostman, J. L. Prieto, M. Richmond, R. W. Romani, J. Sollerman, M. Stritzinger, D. P. Schneider, M. Smith, J. C. Wheeler, N. Yasuda, C. Zheng, The Data Release of the Sloan Digital Sky Survey-II Supernova Survey, ArXiv e-prints (2014), arXiv:1401.3317 [astro-ph.CO].

(Sellentin & Heavens, 2016) E. Sellentin, A. F. Heavens, Parameter inference with estimated covariance matrices, Mon. Not. R. Astron. Soc. 456, L132 (2016), arXiv:1511.05969 [astro-ph.CO].

(Sobol, 1967) 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).

(Sullivan et al., 2011) M. Sullivan, J. Guy, A. Conley, N. Regnault, P. Astier, C. Ballard, S. Basa, R. G. Carlberg, D. Fouche, D. Hardin, I. M. Hook, D. A. Howell, R. Pain, N. Palanque-Delabrouille, K. M. Perrett, C. J. Pritchet, J. Rich, V. Ruhlmann-Kleider, D. Balam, S. Baumont, R. S. Ellis, S. Fabbro, H. K. Fakhouri, N. Fourmanoit, S. González-Gaitán, M. L. Graham, M. J. Hudson, E. Hsiao, T. Kronborg, C. Lidman, A. M. Mourao, J. D.Neill, S. Perlmutter, P. Ripoche, N. Suzuki, E. S. Walker, SNLS3: Constraints on Dark Energy Combining the Supernova Legacy Survey Three-year Data with Other Probes, Astrophys. J. 737, 102 (2011), arXiv:1104.1444 [astro-ph.CO].

(Wandelt, Larson & Lashkinarayanan, 2004) B. D. Wandelt, D. L. Larson, A. Lashkinarayanan, Global, exact cosmic microwave background data analysis using Gibbs sampling, Phys. Rev. D 70, 083511 (2004), astro-ph/0310080.

(Wang et al., 2013) Z. Wang, M. Zoghi, F. Hutter, M. Matheson, N. De Freitas, Bayesian Optimization in High Dimensions via Random Embeddings, in Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, IJCAI ’13 (AAAI Press, 2013) pp. 1778–1784.

(Weyant, Schafer & Wood-Vasey, 2013) A. Weyant, C. Schafer, W. M. Wood-Vasey, Likelihood-free Cosmological Inference with a Complete Treatment of Uncertainty, Astrophys. J. 764, 116 (2013), arXiv:1206.2563 [astro-ph.CO].

(Wood, 2010) S. N. Wood, Statistical inference for noisy nonlinear ecological dynamic systems, Nature 466, 1102 (2010).