Batch simulations and uncertainty quantification in Gaussian process surrogate approximate Bayesian computation

Marko Järvenpää*, Aki Vehtari, Pekka Marttinen
Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University

Abstract

The computational efficiency of approximate Bayesian computation (ABC) has been improved by using surrogate models such as Gaussian processes (GP). In one such promising framework the discrepancy between the simulated and observed data is modelled with a GP which is further used to form a model-based estimator for the intractable posterior. In this article we improve this approach in several ways. We develop batch-sequential Bayesian experimental design strategies to parallellise the expensive simulations. In earlier work only sequential strategies have been used. Current surrogate-based ABC methods also do not fully account the uncertainty due to the limited budget of simulations as they output only a point estimate of the ABC posterior. We propose a numerical method to fully quantify the uncertainty in, for example, ABC posterior moments. We also provide some new analysis on the GP modelling assumptions in the resulting improved framework called Bayesian ABC and discuss its connection to Bayesian quadrature (BQ) and Bayesian optimisation (BO). Experiments with toy and real-world simulation models demonstrate advantages of the proposed techniques.

1 INTRODUCTION

Approximate Bayesian computation (Beaumont et al., 2002; Marin et al., 2012) is used for Bayesian inference when the likelihood function of a statistical model of interest is intractable, i.e., when the analytical form of the likelihood is either unavailable or too costly to evaluate, but simulating the model is feasible. The main idea of the ABC rejection sampler (Pritchard et al., 1999) is to draw a parameter from the prior, use it to simulate one pseudo-data set and finally accept the parameter as a draw from an approximate posterior if the discrepancy between the simulated and observed data sets is small enough. While the computational efficiency of this basic ABC algorithm has been improved in several ways, many models e.g. in genomics and epidemiology (Numminen et al., 2013; Marttinen et al., 2015), astronomy (Rogers et al., 2019) and climate science (Holden et al., 2018) are expensive-to-simulate which makes the sampling-based ABC inference algorithms infeasible. To increase sample-efficiency of ABC, various methods using surrogate models such as neural networks (Papamakarios and Murray, 2016; Papamakarios et al., 2019; Greenberg et al., 2019) and Gaussian processes (Meeds and Welling, 2014; Wilkinson, 2014; Gutmann and Corander, 2016; Järvenpää et al., 2018, 2019) have been proposed.

In one promising surrogate-based ABC framework the discrepancy between the observed and simulated data, a key quantity in ABC, is modelled with a GP (Gutmann and Corander, 2016; Järvenpää et al., 2018, 2019). The GP model is then used to form an estimator for the (approximate) posterior and to adaptively select new evaluation locations. Sequential *Bayesian experimental design* (also known as *active learning*) methods to select the simulation locations so as to maximise the sample-efficiency were proposed by Järvenpää et al. (2019). However, their methods allow to run only one simulation at a time although in practice one often has access to multiple cores to run some of the simulations in parallel. In this work, we resolve this limitation by developing batch simulation methods which are then shown to considerably decrease the wall-time needed for ABC inference. Our approach (Section 4) is based on a Bayesian decision theoretic framework recently developed by Järvenpää et al. (2020) who, however, assumed that expensive and potentially noisy likelihood evaluations are available (e.g. by syn-

[*] Current address: Department of Biostatistics, University of Oslo, email: `m.j.jarvenpaa@medisin.uio.no`

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

thetic likelihood method (Wood, 2010; Price et al., 2018)). In this work we instead focus on ABC scenario where only less than a thousand model simulations can be obtained.

In practice the posterior distribution is often summarised for further decision making using e.g. the expectation and variance. When the computational resources for ABC inference are limited, it would be important to assess the accuracy of such summaries, but this has not been done in earlier work. As the second main contribution of this article, we devise an approximate numerical method to propagate the uncertainty of the discrepancy, represented by the GP model, to the resulting ABC posterior summaries (Section 5). Such uncertainty estimates are useful for assessing the accuracy of the inference and guiding the termination of the inference algorithm. We call the resulting improved framework as *Bayesian ABC* in analogy with the related problems of *Bayesian* quadrature and *Bayesian* optimisation.

We also provide new insights on the underlying GP modelling assumptions (Appendix A.2) and on the connections between Bayesian ABC, BQ and BO to improve understanding of these conceptually similar techniques (Section 6). Finally, we demonstrate the ABC posterior uncertainty quantification and show that Bayesian ABC framework is well-suited for parallel simulations using several numerical experiments (Section 7).

2 BRIEF BACKGROUND ON ABC

We denote the (continuous) parameters of the statistical model of interest with $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. The posterior distribution, that describes our knowledge of $\boldsymbol{\theta}$ given some observed data $\mathbf{x}_\mathrm{o} \in \mathcal{X}$ and a prior density $\pi(\boldsymbol{\theta})$, can then be computed using Bayes' theorem

$$\pi(\boldsymbol{\theta} \,|\, \mathbf{x}_\mathrm{o}) = \frac{\pi(\boldsymbol{\theta})\pi(\mathbf{x}_\mathrm{o} \,|\, \boldsymbol{\theta})}{\int_\Theta \pi(\boldsymbol{\theta}')\pi(\mathbf{x}_\mathrm{o} \,|\, \boldsymbol{\theta}')\,\mathrm{d}\boldsymbol{\theta}'}. \tag{1}$$

If the likelihood function $\pi(\mathbf{x}_\mathrm{o} \,|\, \boldsymbol{\theta})$ is intractable, evaluating (1) even up-to-normalisation becomes infeasible. Standard ABC algorithms such as the ABC rejection sampler instead target the approximate posterior

$$\pi_{\mathrm{ABC}}(\boldsymbol{\theta}|\mathbf{x}_\mathrm{o}) \triangleq \frac{\pi(\boldsymbol{\theta})\int_{\mathcal{X}}\pi_\varepsilon(\mathbf{x}_\mathrm{o}|\mathbf{x})\pi(\mathbf{x}|\boldsymbol{\theta})\,\mathrm{d}\mathbf{x}}{\int_\Theta \pi(\boldsymbol{\theta}')\int_{\mathcal{X}}\pi_\varepsilon(\mathbf{x}_\mathrm{o}|\mathbf{x}')\pi(\mathbf{x}'|\boldsymbol{\theta}')\,\mathrm{d}\mathbf{x}'\,\mathrm{d}\boldsymbol{\theta}'}, \tag{2}$$

where $\pi_\varepsilon(\mathbf{x}_\mathrm{o} \,|\, \mathbf{x}) = \mathbb{1}_{\Delta(\mathbf{x}_\mathrm{o},\mathbf{x}) \leq \varepsilon}$. Other choices of kernel π_ε are also possible (Wilkinson, 2013). Above, $\Delta : \mathcal{X}^2 \to \mathbb{R}_+$ is the discrepancy function used to compare the similarity of the data sets and ε is a threshold parameter. Small ε produces good approximations but renders sampling-based ABC methods inefficient. A well-constructed discrepancy function is an important ingredient of accurate ABC inference (Marin et al., 2012). In this article we assume a suitable discrepancy function is already available (e.g. constructed based on expert opinion, earlier analyses on other similar models, pilot runs or distance measures between raw data sets (Park et al., 2016; Bernton et al., 2019)) and focus on approximating any given ABC posterior in (2) as well as possible given only a limited budget of simulations.

3 BAYESIAN ABC FRAMEWORK

We describe our Bayesian ABC framework here. The main difference to earlier work (Järvenpää et al., 2019) is that we use a hierarchical GP model and, most importantly, explicitly quantify the uncertainty of the ABC posterior instead of resorting to point estimation. The main idea is to explicitly use another layer of Bayesian inference to estimate the ABC posterior in (2). The previously simulated discrepancy-parameter-pairs are treated as data to learn a surrogate model, which will predict the discrepancy for a given parameter value. The surrogate model is further used to form an estimator for the ABC posterior in (2) and to adaptively acquire new data.

We assume that each discrepancy evaluation, denoted by Δ_i at the corresponding parameter $\boldsymbol{\theta}_i$, is generated as

$$\Delta_i = f(\boldsymbol{\theta}_i) + \nu_i, \quad \nu_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_n^2), \tag{3}$$

where $\sigma_n^2 > 0$ is the variance of the discrepancy[1]. To encode the assumptions of smoothness and e.g. potential quadratic shape of the discrepancy $\Delta_{\boldsymbol{\theta}}$, in this work its unknown mean function f is given a hierarchical GP prior

$$\begin{aligned} f \,|\, \boldsymbol{\gamma} &\sim \mathcal{GP}(m_0(\boldsymbol{\theta}), k_{\boldsymbol{\phi}}(\boldsymbol{\theta}, \boldsymbol{\theta}')), \\ m_0(\boldsymbol{\theta}) &\triangleq \sum_{i=1}^r \gamma_i h_i(\boldsymbol{\theta}), \quad \boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{b}, \mathbf{B}), \end{aligned} \tag{4}$$

where $k_{\boldsymbol{\phi}} : \Theta^2 \to \mathbb{R}$ is a covariance function with hyperparameters $\boldsymbol{\phi}$ and $h_i : \Theta \to \mathbb{R}$ are basis functions (both assumed continuous). We marginalise $\boldsymbol{\gamma}$ in (4), as in O'Hagan and Kingman (1978), and Riihimäki and Vehtari (2014), to obtain the GP prior

$$f \sim \mathcal{GP}(\mathbf{h}(\boldsymbol{\theta})^\top \mathbf{b}, k_{\boldsymbol{\phi}}(\boldsymbol{\theta}, \boldsymbol{\theta}') + \mathbf{h}(\boldsymbol{\theta})^\top \mathbf{B}\mathbf{h}(\boldsymbol{\theta}')), \tag{5}$$

where $\mathbf{h}(\boldsymbol{\theta}) \in \mathbb{R}^r$ is a column vector consisting of the basis functions h_i evaluated at $\boldsymbol{\theta}$. For now, we assume the GP hyperparameters $\boldsymbol{\psi} \triangleq (\sigma_n^2, \boldsymbol{\phi})$ are fixed and omit $\boldsymbol{\psi}$ from our notation for brevity.

Given training data $D_t \triangleq \{(\Delta_i, \boldsymbol{\theta}_i)\}_{i=1}^t$, we obtain $f \,|\, D_t \sim \mathcal{GP}(m_t(\boldsymbol{\theta}), c_t(\boldsymbol{\theta}, \boldsymbol{\theta}'))$,

$$m_t(\boldsymbol{\theta}) \triangleq \mathbf{k}_t(\boldsymbol{\theta})\mathbf{K}_t^{-1}\boldsymbol{\Delta}_t + \mathbf{R}_t^\top(\boldsymbol{\theta})\bar{\boldsymbol{\gamma}}_t, \tag{6}$$

[1] While this modelling assumption may seem strong, it has been used successfully before. We now give a justification for this modelling choice in Appendix A.2.

$$\begin{aligned} c_t(\boldsymbol{\theta},\boldsymbol{\theta}') \triangleq\ & k(\boldsymbol{\theta},\boldsymbol{\theta}') - \mathbf{k}_t(\boldsymbol{\theta})\mathbf{K}_t^{-1}\mathbf{k}_t^\top(\boldsymbol{\theta}') \\ & + \mathbf{R}_t^\top(\boldsymbol{\theta})[\mathbf{B}^{-1} + \mathbf{H}_t\mathbf{K}_t^{-1}\mathbf{H}_t^\top]^{-1}\mathbf{R}_t(\boldsymbol{\theta}'), \end{aligned} \tag{7}$$

where $[\mathbf{K}_t]_{ij} \triangleq k(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j) + \mathbb{1}_{i=j}\sigma_n^2$, $\mathbf{k}_t(\boldsymbol{\theta}) \triangleq (k(\boldsymbol{\theta},\boldsymbol{\theta}_1),\ldots,k(\boldsymbol{\theta},\boldsymbol{\theta}_t))^\top$, $\boldsymbol{\Delta}_t \triangleq (\Delta_1,\ldots,\Delta_t)^\top$ and

$$\bar{\boldsymbol{\gamma}}_t \triangleq [\mathbf{B}^{-1}+\mathbf{H}_t\mathbf{K}_t^{-1}\mathbf{H}_t^\top]^{-1}(\mathbf{H}_t\mathbf{K}_t^{-1}\boldsymbol{\Delta}_t+\mathbf{B}^{-1}\mathbf{b}), \tag{8}$$

$$\mathbf{R}_t(\boldsymbol{\theta}) \triangleq \mathbf{H}(\boldsymbol{\theta}) - \mathbf{H}_t\mathbf{K}_t^{-1}\mathbf{k}_t^\top(\boldsymbol{\theta}). \tag{9}$$

Above $\bar{\boldsymbol{\gamma}}_t$ is the generalised least-squares estimate, $\mathbf{H}_t$ is the $r \times t$ matrix whose columns consist of basis function values evaluated at $\boldsymbol{\theta}_{1:t}$, $\boldsymbol{\theta}_{1:t}$ is a $p \times t$ matrix, and $\mathbf{H}(\boldsymbol{\theta}) \in \mathbb{R}^r$ is the corresponding vector of test point $\boldsymbol{\theta}$. We also define $s_t^2(\boldsymbol{\theta}) \triangleq c_t(\boldsymbol{\theta},\boldsymbol{\theta})$ and $\Pi_{D_t}^f \triangleq \mathcal{GP}(m_t(\boldsymbol{\theta}), c_t(\boldsymbol{\theta},\boldsymbol{\theta}'))$. For further details of GP regression, see e.g. Rasmussen and Williams (2006).

If the true discrepancy mean function f and the variance of the discrepancy σ_n^2 were known, the ABC posterior would be

$$\pi_{\text{ABC}}^f(\boldsymbol{\theta}) \triangleq \frac{\pi(\boldsymbol{\theta})\Phi\left((\varepsilon - f(\boldsymbol{\theta}))/\sigma_n\right)}{\int_\Theta \pi(\boldsymbol{\theta}')\Phi\left((\varepsilon - f(\boldsymbol{\theta}'))/\sigma_n\right) \mathrm{d}\boldsymbol{\theta}'}, \tag{10}$$

where $\Phi(\cdot)$ is the Gaussian cdf. This fact follows from (2) and the Gaussian modelling assumption (3). In practice f is unknown but our knowledge about f is represented by the posterior $f \,|\, D_t \sim \Pi_{D_t}^f$. Since the ABC posterior π_{ABC}^f in (10) depends on f, it is also a random quantity and its posterior can be obtained by propagating the uncertainty in f through the mapping $f \mapsto \pi_{\text{ABC}}^f$.

Computing the distribution of π_{ABC}^f is difficult due to its nonlinear dependence on f and because f is infinite-dimensional. However, the pointwise mean, variance and quantiles of the unnormalised ABC posterior

$$\tilde{\pi}_{\text{ABC}}^f(\boldsymbol{\theta}) \triangleq \pi(\boldsymbol{\theta})\Phi((\varepsilon - f(\boldsymbol{\theta}))/\sigma_n), \tag{11}$$

i.e. the numerator of (10), can be computed analytically as shown by Järvenpää et al. (2019) in the case of a zero mean GP prior. It is easy to see that their formulas also hold for our more general GP model in (4). For example,

$$\mathbb{E}_{f \,|\, D_t}(\tilde{\pi}_{\text{ABC}}^f(\boldsymbol{\theta})) = \pi(\boldsymbol{\theta})\Phi(a_t(\boldsymbol{\theta})), \tag{12}$$

$$a_t(\boldsymbol{\theta}) \triangleq (\varepsilon - m_t(\boldsymbol{\theta}))/\sqrt{\sigma_n^2 + s_t^2(\boldsymbol{\theta})}, \tag{13}$$

$$\text{med}_{f \,|\, D_t}(\tilde{\pi}_{\text{ABC}}^f(\boldsymbol{\theta})) = \pi(\boldsymbol{\theta})\Phi\left((\varepsilon - m_t(\boldsymbol{\theta}))/\sigma_n\right), \tag{14}$$

where med is the marginal (i.e. elementwise) median. While these formulas are useful, they do not allow one to assess the uncertainty of e.g. posterior mean $\int_\Theta \boldsymbol{\theta}\, \pi_{\text{ABC}}^f(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$. We resolve this limitation in Section 5.

4 PARALLEL SIMULATIONS

We aim to find the most informative simulation locations for obtaining the best possible estimate of the ABC posterior π_{ABC}^f given the postulated GP model. Principled sequential designs, where one simulation is run at a time, were developed by Järvenpää et al. (2019). In practice, to decrease the wall-time needed for the inference task, one could run some of the simulations in parallel. In the following, we apply Bayesian experimental design theory (Chaloner and Verdinelli, 1995; Järvenpää et al., 2020) for the (synchronous) batch setting where b simulations are simultaneously selected to be computed in parallel.

4.1 DECISION-THEORETIC APPROACH

Consider a loss function $l : \mathscr{D}^2 \to \mathbb{R}_+$ so that $l(\pi_{\text{ABC}}, d)$ quantifies the penalty of reporting $d \in \mathscr{D}$ as our ABC posterior when the true one is $\pi_{\text{ABC}} \in \mathscr{D}$. Given D_t, the one-batch-ahead Bayes-optimal selection of the next batch of b evaluations $\boldsymbol{\theta}^{\text{opt}} = [\boldsymbol{\theta}_1^{\text{opt}},\ldots,\boldsymbol{\theta}_b^{\text{opt}}]$ is then

$$\boldsymbol{\theta}^{\text{opt}} = \operatorname*{arg\,min}_{\boldsymbol{\theta}^* \in \Theta^b} L_t(\boldsymbol{\theta}^*), \quad \text{where} \tag{15}$$

$$L_t(\boldsymbol{\theta}^*) = \mathbb{E}_{\boldsymbol{\Delta}^* \,|\, \boldsymbol{\theta}^*, D_t}\Big(\underbrace{\min_{d \in \mathscr{D}} \mathbb{E}_{f \,|\, D_t \cup D^*} l(\pi_{\text{ABC}}^f, d)}_{\triangleq \mathcal{L}(\Pi_{D_t \cup D^*}^f)}\Big). \tag{16}$$

In (16), we calculate an expectation over future discrepancy evaluations $\boldsymbol{\Delta}^* = (\Delta_1^*,\ldots,\Delta_b^*)^\top$ at locations $\boldsymbol{\theta}^*$, assuming they follow our current GP model. The expectation is taken of the *Bayes risk* $\mathcal{L}(\Pi_{D_t \cup D^*}^f)$ resulting from the nested decision problem of choosing the estimator d, assuming $\boldsymbol{\Delta}^*$ are known and merged with current data D_t via $D^* \triangleq \{(\Delta_i^*, \boldsymbol{\theta}_i^*)\}_{i=1}^b$. While the main quantity of interest in the Bayesian ABC framework is the ABC posterior π_{ABC}^f in (10), in practice it is desirable to use a loss function $\tilde{l}$ based on the unnormalised distribution $\tilde{\pi}_{\text{ABC}}^f$. Such a simplification, also used by Kandasamy et al. (2017); Sinsbeck and Nowak (2017); Järvenpää et al. (2019, 2020), allows efficient computations. Furthermore, evaluations that are optimal for estimating $\tilde{\pi}_{\text{ABC}}^f$ will be informative about the related quantity π_{ABC}^f.

Consider L^2 loss function $\tilde{l}_2 \triangleq \int_\Theta (\tilde{\pi}_{\text{ABC}}^f(\boldsymbol{\theta}) - \tilde{d}(\boldsymbol{\theta}))^2\, \mathrm{d}\boldsymbol{\theta}$ between the unnormalised ABC posterior $\tilde{\pi}_{\text{ABC}}^f$ and its estimator $\tilde{d}$ (both supposed to be square-integrable in Θ i.e. $\tilde{\pi}_{\text{ABC}}^f, \tilde{d} \in L^2(\Theta)$). Then the optimal estimator is the mean in (12) (Sinsbeck and Nowak, 2017; Järvenpää et al., 2020). If we instead consider L^1 loss $\tilde{l}_1 \triangleq \int_\Theta |\tilde{\pi}_{\text{ABC}}^f(\boldsymbol{\theta}) - \tilde{d}(\boldsymbol{\theta})|\, \mathrm{d}\boldsymbol{\theta}$ (supposing $\tilde{\pi}_{\text{ABC}}^f, \tilde{d} \in L^1(\Theta)$), then the marginal median in (14) is the optimal estimator. Corresponding Bayes risks, denoted $\mathcal{L}^{\text{v}}$ and $\mathcal{L}^{\text{m}}$, respectively, can be computed as follows:

Lemma 4.1. *Consider the GP model in Section 3. The Bayes risks for the L^2 and L^1 losses are given by*

$$\begin{aligned} \mathcal{L}^{\text{v}}(\Pi_{D_t}^f) = \int_\Theta \pi^2(\boldsymbol{\theta}) \Big[& \Phi(a_t(\boldsymbol{\theta}))\Phi(-a_t(\boldsymbol{\theta})) \\ & - 2T\Big(a_t(\boldsymbol{\theta}), \sigma_n/\sqrt{\sigma_n^2 + 2s_t^2(\boldsymbol{\theta})}\Big)\Big] \mathrm{d}\boldsymbol{\theta}, \end{aligned} \tag{17}$$

$$\mathcal{L}^{\mathrm{m}}(\Pi^f_{D_t}) = 2\int_\Theta \pi(\boldsymbol{\theta}) T(a_t(\boldsymbol{\theta}), s_t(\boldsymbol{\theta})/\sigma_n)\,\mathrm{d}\boldsymbol{\theta}, \tag{18}$$

respectively, where $a_t(\boldsymbol{\theta})$ is given by (13) and $T(\cdot,\cdot)$ denotes Owen's T function (Owen, 1956).

We call $L_t(\boldsymbol{\theta}^*)$ as an *acquisition function*. Expected integrated variance (EIV) and expected integrated MAD[2] (EIMAD) acquisition functions, denoted $L^{\mathrm{v}}_t(\boldsymbol{\theta}^*)$ and $L^{\mathrm{m}}_t(\boldsymbol{\theta}^*)$, respectively, can be computed as follows:

Proposition 4.2. *Consider the GP model in Section 3. The* EIV *and* EIMAD *acquisition functions are*

$$L^{\mathrm{v}}_t(\boldsymbol{\theta}^*) = 2\int_\Theta \pi^2(\boldsymbol{\theta}) \Bigg[T\left(a_t(\boldsymbol{\theta}), \frac{\sqrt{\sigma_n^2 + s_t^2(\boldsymbol{\theta}) - \tau_t^2(\boldsymbol{\theta};\boldsymbol{\theta}^*)}}{\sqrt{\sigma_n^2 + s_t^2(\boldsymbol{\theta}) + \tau_t^2(\boldsymbol{\theta};\boldsymbol{\theta}^*)}}\right) - T\left(a_t(\boldsymbol{\theta}), \frac{\sigma_n}{\sqrt{\sigma_n^2 + 2s_t^2(\boldsymbol{\theta})}}\right)\Bigg]\mathrm{d}\boldsymbol{\theta}, \tag{19}$$

$$L^{\mathrm{m}}_t(\boldsymbol{\theta}^*) = 2\int_\Theta \pi(\boldsymbol{\theta}) T\left(a_t(\boldsymbol{\theta}), \frac{\sqrt{s_t^2(\boldsymbol{\theta}) - \tau_t^2(\boldsymbol{\theta};\boldsymbol{\theta}^*)}}{\sqrt{\sigma_n^2 + \tau_t^2(\boldsymbol{\theta};\boldsymbol{\theta}^*)}}\right)\mathrm{d}\boldsymbol{\theta}, \tag{20}$$

respectively, where $a_t(\boldsymbol{\theta})$ is given by (13) and

$$\tau_t^2(\boldsymbol{\theta};\boldsymbol{\theta}^*) = c_t(\boldsymbol{\theta},\boldsymbol{\theta}^*)[c_t(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) + \sigma_n^2\mathbf{I}]^{-1} c_t(\boldsymbol{\theta}^*,\boldsymbol{\theta}). \tag{21}$$

This result generalizes EIV in Järvenpää et al. (2019) to the batch setting. The proofs are given in Appendix A.1.

4.2 DETAILS ON COMPUTATION

Finding the one-batch-ahead optimal design $\boldsymbol{\theta}^{\mathrm{opt}}$ requires global optimisation over Θ^b for both EIV and EIMAD. As this is infeasible with large batch size b and/or the dimension p of $\boldsymbol{\theta}$, we use greedy optimisation: For $r = 1, \ldots, b$, the rth point $\boldsymbol{\theta}^{\mathrm{opt}}_r$ in the batch is chosen by optimising $L_t([\boldsymbol{\theta}^*_1, \ldots, \boldsymbol{\theta}^*_r])$ with respect to $\boldsymbol{\theta}^*_r$ when the earlier points $\boldsymbol{\theta}^*_1, \ldots, \boldsymbol{\theta}^*_{r-1}$ are kept fixed to their already determined values. This simplifies the pb-dimensional optimisation problem to a sequence of easier p-dimensional problems. Similar techniques have been used in batch BO, see Ginsbourger et al. (2010); Snoek et al. (2012); Wilson et al. (2018). Theory of submodular optimisation has been used to study greedy batch designs (Bach, 2013; Wilson et al., 2018; Järvenpää et al., 2020). Unfortunately, such analysis hardly extends to our case because the acquisition functions in Proposition 4.2 depend on $\boldsymbol{\theta}^*$ in a rather complex way. Using the facts that $T(h,a)$ is non-decreasing for $a \geq 0$ and $\tau_t^2(\boldsymbol{\theta};\boldsymbol{\theta}^*)$ cannot decrease as more points are included to $\boldsymbol{\theta}^*$, we nevertheless see that both EIV and EIMAD are non-increasing as set functions of $\boldsymbol{\theta}^*$. We can thus expect the greedy optimisation to be useful in practice as is seen empirically in Section 7.

Another potential computational difficulty is the integration over Θ in (19) and (20). Many state-of-the-art BO methods, such as Hennig and Schuler (2012); Hernández-Lobato et al. (2014); Wu and Frazier (2016), also require similar computations. We approximate the integral using numerical integration for $p \leq 2$ and self-normalised importance sampling (IS), where the current loss function interpreted as an unnormalised density is the instrumental distribution, for $p > 2$. Full details and the pseudocode of our algorithm can be found in Appendix B.

4.3 HEURISTIC BASELINE BATCH METHODS

We consider also heuristic acquisition functions which evaluate where the pointwise uncertainty of $\tilde{\pi}^f_{\mathrm{ABC}}(\boldsymbol{\theta})$ is highest. Such intuitive strategies are also known as *uncertainty sampling* and used e.g. by Gunter et al. (2014); Järvenpää et al. (2019); Chai and Garnett (2019). When the variance is used as the measure of the uncertainty of $\tilde{\pi}^f_{\mathrm{ABC}}(\boldsymbol{\theta})$, we call the method as MAXV. When MAD is used, we obtain an alternative strategy called analogously MAXMAD. The resulting acquisition functions can be computed using the integrands of (17) and (18).

Finally, we propose a heuristic approach, also used for batch BO (Snoek et al., 2012), to parallellise MAXV and MAXMAD strategies: The first point in the batch is chosen as in the sequential case. The other points are iteratively selected as the locations where the expected pointwise variance (or MAD) of $\tilde{\pi}^f_{\mathrm{ABC}}(\boldsymbol{\theta})$, taken with respect to the discrepancy values of the pending points (i.e. points that have been already chosen to the current batch) is highest. The resulting acquisition functions are immediately obtained as the integrands of (19) and (20).

5 UNCERTAINTY QUANTIFICATION OF THE ABC POSTERIOR

Pointwise marginal uncertainty of the unnormalised ABC posterior $\tilde{\pi}^f_{\mathrm{ABC}}$ was used in previous section for selecting the simulation locations adaptively. However, knowing the value of $\tilde{\pi}^f_{\mathrm{ABC}}$ and its marginal uncertainty in some individual $\boldsymbol{\theta}$-values is not very helpful for summarising and understanding the accuracy of the final estimate of the ABC posterior. Computing the distribution of the moments and marginals of the normalised ABC posterior π^f_{ABC} in (10) is clearly more intuitive. See Fig. 1 for a 1D demonstration of this approach.

To access the posterior of π^f_{ABC}, one could fix a sample path $f^{(i)} \sim \Pi^f_{D_t}$, then use it to fix a realisation of the ABC posterior $\pi^{f^{(i)}}_{\mathrm{ABC}}$ using (10) and finally use e.g. MCMC to sample from $\pi^{f^{(i)}}_{\mathrm{ABC}}$. This would be repeated s times and the resulting set of samples $\{\{\boldsymbol{\theta}^{(i,j)}\}_{j=1}^n\}_{i=1}^s$

[2] Mean absolute deviation (around median).

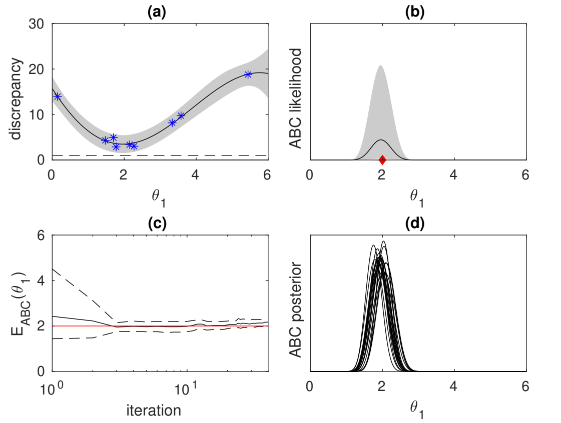

Figure 1: Demonstration of ABC posterior uncertainty quantification using Lorenz model from Section 7.2 with parameter θ_2 fixed. (a) GP model for Δ_{θ_1} (blue dashed line ε, blue stars 9 discrepancy evaluations), (b) uncertainty of unnormalised ABC posterior $\tilde{\pi}^f_{\text{ABC}}$, (c) evolution of model-based ABC posterior expectation (black line) and its 95% CI (dashed black) for 40 iterations, (d) uncertainty of ABC posterior π^f_{ABC} corresponding (b).

(where n is the length of the MCMC chain for each posterior realisation $i = 1, \ldots, s$) approximately describes the posterior of π^f_{ABC} given D_t (see Fig. 1d). The uncertainty of the GP hyperparameters $\boldsymbol{\psi}$ could also be taken into account by drawing $\boldsymbol{\psi}^{(i)} \sim \pi(\boldsymbol{\psi} \,|\, D_t)$ as the very first step but we here consider $\boldsymbol{\psi}$ as known for simplicity although this can cause underestimation of the uncertainty of π^f_{ABC}. The outlined approach involves a major computational challenge as evaluating the s sample paths at n distinct sets of test points scales[3] as $\mathcal{O}(s(nt^2 + tn^2) + sn^3)$.

We propose the following computationally cheaper approach: In small dimensions, when $p \leq 2$, we evaluate each sample path $f^{(i)}, i = 1, \ldots, s$ at $\bar{n}^p$ fixed grid points and compute the required integrations numerically. This approach scales as $\mathcal{O}(\bar{n}^p t^2 + \bar{n}^{2p}(t+s) + \bar{n}^{3p})$. If $p > 2$, then self-normalised importance sampling is used. We draw n samples from an instrumental density, defined so that its unnormalised pdf at $\boldsymbol{\theta}$ equals the α-quantile of $\tilde{\pi}^f_{\text{ABC}}(\boldsymbol{\theta})$. This is computed using (A.23) of Appendix and we use $\alpha = 0.95$. The samples are thinned and the resulting $\tilde{n} \ll n$ representative samples $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{\tilde{n}}$ are used to compute the normalised importance weights $\omega^{(i,j)}$ for each sampled posterior $i = 1, \ldots, s$. The output is a set of weighted sample sets $\{\{(\omega^{(i,j)}, \boldsymbol{\theta}^{(j)})\}_{j=1}^{\tilde{n}}\}_{i=1}^{s}$ from which moments and marginal densities can be computed using standard Monte Carlo estimators for each $i = 1, \ldots, s$. This approach requires only one MCMC sampling from the instrumental density which scales as $\mathcal{O}(nt^2)$, i.e. only linearly with respect to n, so that n can be large. Total cost is $\mathcal{O}((n + \tilde{n})t^2 + \tilde{n}^2(t + s) + \tilde{n}^3)$.

This approach has nevertheless some limitations: The computations are only approximate because $\tilde{n}$ and s are finite. Also, if the uncertainty of π^f_{ABC} is substantial, choosing a good instrumental density can be difficult. This is because some of the sampled posteriors are then necessarily quite different from any single instrumental density producing possibly poor approximation. In our experiments this however happened only with early iterations and can be detected e.g. by monitoring the distribution of effective sample sizes for $i = 1, \ldots, s$. In Section 7 we demonstrate that the uncertainty quantification is still feasible and beneficial for low-dimensional cases. The proposed approach also works with other GP modelling situations such as Järvenpää et al. (2020).

6 ON RELATED GP-BASED METHODS

In this section we briefly discuss the relation between Bayesian ABC, BQ and BO to facilitate better understanding of these conceptually similar inference methods.

6.1 RELATION TO BAYESIAN QUADRATURE

In Bayesian quadrature one aims to compute integral $I_f \triangleq \int_{\mathbb{R}^p} f(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$, where $f : \mathbb{R}^p \to \mathbb{R}$ is an expensive black-box function and $\pi(\boldsymbol{\theta})$ is a known density, e.g. Gaussian. If a GP prior is placed on f, given some evaluations $\{(f_i, \boldsymbol{\theta}_i)\}_{i=1}^t$ where $f_i = f(\boldsymbol{\theta}_i)$, the posterior of I_f, describing one's knowledge of the value of this integral, is Gaussian whose mean and variance can be computed analytically for some choices of $k(\boldsymbol{\theta}, \boldsymbol{\theta}')$ and $\pi(\boldsymbol{\theta})$ (for details, see O'Hagan (1991); Briol et al. (2019)). Also, BQ methods for computing integrals of the form $I^g_f \triangleq \int_\Theta g(f(\boldsymbol{\theta}))\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$ with some known (non-negative) function $g : \mathbb{R} \to \mathbb{R}_+$, such as marginal likelihoods, have been developed by Osborne et al. (2012); Gunter et al. (2014); Chai and Garnett (2019).

Our approach in Section 5 is instead developed for quantifying the uncertainty in either the whole function $\pi^f_{\text{ABC}} : \Theta \to \mathbb{R}_+$, which we here write as

$$\pi^f_{\text{ABC}}(\boldsymbol{\theta}) = \frac{g(f(\boldsymbol{\theta}))\pi(\boldsymbol{\theta})}{\int_\Theta g(f(\boldsymbol{\theta}'))\pi(\boldsymbol{\theta}')\,\mathrm{d}\boldsymbol{\theta}'}, \tag{22}$$

or some corresponding moments such as the expectation $\int_\Theta \boldsymbol{\theta}\, \pi^f_{\text{ABC}}(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \in \mathbb{R}^p$. To our knowledge, computation

[3] Approximations such as random Fourier features (RFF) (Rahimi and Recht, 2008) and those by Pleiss et al. (2018) can be used to reduce this cost, e.g. Hernández-Lobato et al. (2014); Wang and Jegelka (2017) used RFF to approximately optimise GP sample paths. However, this produces tradeoff between exact GP but small n vs. inexact GP but large n which we do not analyse in this work.

of these quantities probabilistically has not been considered before. In particular, we used the "0-1 kernel" $\mathbb{1}_{\Delta_{\boldsymbol{\theta}} \leq \varepsilon}$ in (2) corresponding to $g(f(\boldsymbol{\theta})) = \Phi((\varepsilon - f(\boldsymbol{\theta}))/\sigma_n)$ in (22). Osborne et al. (2012); Gutmann and Corander (2016); Acerbi (2018); Järvenpää et al. (2020) instead modelled the log-likelihood with GP to reckon the non-negativity of the likelihood and the high dynamic range of the log-likelihood. This would correspond to $g(f(\boldsymbol{\theta})) = \exp(f(\boldsymbol{\theta}))$ in (22).

Osborne et al. (2012); Gunter et al. (2014); Chai and Garnett (2019) used linearisation approximations in their algorithms for estimating integrals of the form I_f^g. Similarly, if both $\Phi(\cdot)$-terms in our case in (10) were linear for f, then the numerator and denominator in (10) would have joint Gaussian density leading to tractable computations. However, we observed that the resulting densities can be highly non-Gaussian so that any linearisation approach can result poor quality approximations. For this reason we considered simulation-based approach in Section 5.

6.2 RELATION TO BAYESIAN OPTIMISATION

Suppose now $f : \Theta \subset \mathbb{R}^p \to \mathbb{R}$ is an expensive, black-box function to be minimised. In BO, a GP prior is placed on f and the future locations for obtaining (possibly noisy) evaluations of f are chosen adaptively by optimising an acquisition function that, in some sense, measures the potential improvement in the knowledge of the minimum point $\boldsymbol{\theta}^\star \triangleq \arg\min_{\boldsymbol{\theta}\in\Theta} f(\boldsymbol{\theta})$ or the corresponding function value $f^\star \triangleq \min_{\boldsymbol{\theta}\in\Theta} f(\boldsymbol{\theta})$ brought by the extra evaluation. For example, (predictive) entropy search (Hennig and Schuler, 2012; Hernández-Lobato et al., 2014) use an acquisition function that measures the expected reduction in the differential entropy of the posterior of $\boldsymbol{\theta}^\star$. Wang and Jegelka (2017) similarly considered the posterior of $f^\star$. The important difference between these methods (or BO in general) and Bayesian ABC is that the quantity of interest in Bayesian ABC is not the minimiser of f but the full ABC posterior density π^f_{ABC} (or $\tilde{\pi}^f_{\text{ABC}}$). Also, BO is rarely introduced this way in literature, simple acquisition functions such as the expected improvement and lower confidence bound (LCB) are often used and the posterior of $\boldsymbol{\theta}^\star$ or $f^\star$ is rarely considered.

In the *BOLFI* framework (Gutmann and Corander, 2016), the function f was however taken to be the ABC discrepancy $\Delta_{\boldsymbol{\theta}}$, and LCB acquisition function $\text{LCB}(\boldsymbol{\theta}) = m_t(\boldsymbol{\theta}) - \beta_t s_t(\boldsymbol{\theta})$ (Srinivas et al., 2010) was used for illustrating their approach of learning the ABC posterior. This is reasonable because to learn the ABC posterior one needs to evaluate in the regions with small discrepancy. We have the following new result that relates LCB to the Bayesian ABC framework:

Proposition 6.1. *If the prior is uniform over Θ (and may be improper), i.e. if $\pi(\boldsymbol{\theta}) \propto \mathbb{1}_{\boldsymbol{\theta}\in\Theta}$, then the point chosen by the LCB acquisition function with parameter β_t is the same as the point maximising the $\Phi(\beta_t)$-quantile of the unnormalised ABC posterior $\tilde{\pi}^f_{\text{ABC}}(\boldsymbol{\theta})$ for any ε.*

This result gives an interpretation for the LCB tradeoff parameter β_t in the ABC setting. However, instead of using LCB for Bayesian ABC, it is clearly more reasonable to evaluate where the variance (or some other measure of uncertainty) is large as already discussed e.g. by Kandasamy et al. (2017); Järvenpää et al. (2019). Järvenpää et al. (2019) showed empirically that EIV consistently works better than LCB in their sequential scenario when the goal is to learn the ABC posterior. For this reason, we do not use (batch) BO methods in this article.

7 EXPERIMENTS

We first consider four 2D toy problems to see how the proposed method performs with a well-specified GP model. We then focus on more typical scenarios where the GP modelling assumptions do not hold exactly using three real-world simulation models. We compare the performance of the sequential and synchronous batch versions of the acquisition methods of Section 4. As a simple baseline, we consider random points drawn from the prior (abbreviated as RAND). We also briefly demonstrate the uncertainty quantification of the ABC posterior. We do not consider synthetic likelihood method (as e.g. in Järvenpää et al. (2020)) because it requires hundreds of evaluations for each proposed parameter and is thus not applicable here. For similar reason, we do not consider sampling-based ABC methods.

Locations for fitting the initial GP model are sampled from the uniform prior in all cases. We take 10 initial points for 2D and 20 for 3D and 4D cases. We use $\mathbf{b} = \mathbf{0}$, $B_{ij} = 10^2 \mathbb{1}_{i=j}$ and include basis functions of the form $1, \theta_i, \theta_i^2$. The discrepancy $\Delta_{\boldsymbol{\theta}}$ is assumed smooth and we use the squared exponential covariance function $k(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sigma_f^2 \exp(-\frac{1}{2}\sum_{i=1}^p (\theta_i - \theta_i')^2/l_i^2)$. GP hyperparameters $\boldsymbol{\psi} = (\sigma_n^2, l_1, \ldots, l_p, \sigma_f^2)$ are given weakly informative priors and their values are obtained using MAP estimation at each iteration.

ABC-MCMC (Marjoram et al., 2003) with extensive simulations is used to compute the ground truth ABC posterior for the real-world models. For simplicity and to ensure meaningful comparisons to ground-truth, we fix ε to certain small predefined values although, in practice, its value is set adaptively (Järvenpää et al., 2019) or based on pilot runs. We compute the estimate of the unnormalised ABC posterior using (12) for MAXV, EIV, RAND and (14) for MAXMAD, EIMAD. Adaptive MCMC is used

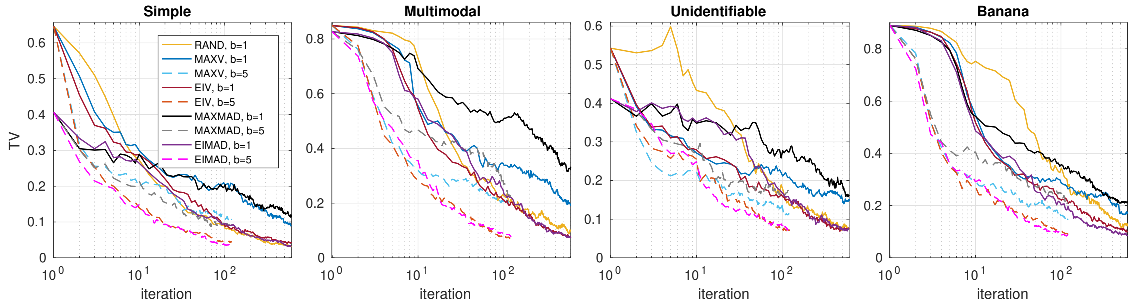

Figure 2: Results for the 2D toy simulation models over 600 iterations and two batch sizes b.

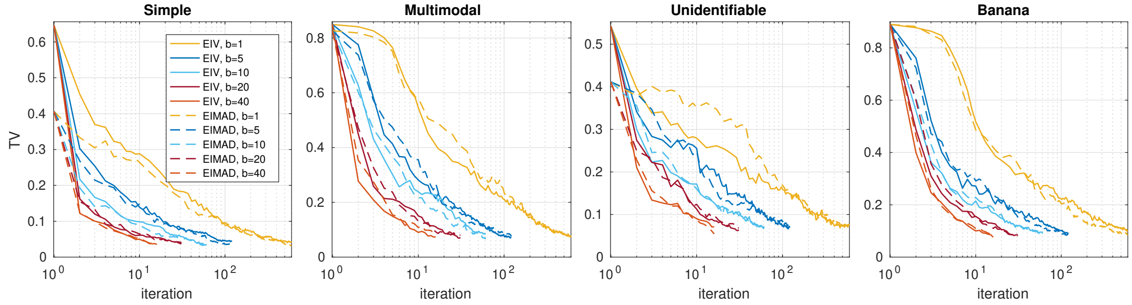

Figure 3: Results for 2D toy simulation models with two acquisition functions and various batch sizes.

to sample from the resulting ABC posterior estimates and from the instrumental densities needed for the IS approximations. TV denotes the median total variation distance between the estimated ABC posterior and the true one (2D) or the average TV between their marginal TV values (3D, 4D) computed numerically over 50 repeated runs. Iteration (i.e. number of batches chosen) serves as a proxy to wall-time. The number of simulations i.e. the maximum value of t is fixed in all experiments and the batch methods thus finish earlier.

7.1 TOY SIMULATION MODELS

Fig. 2 shows the results with sequential methods ($b = 1$) and the corresponding batch methods with $b = 5$ for four synthetically constructed toy models. These were taken from Järvenpää et al. (2019) and are illustrated in the Appendix B. In Fig. 3 the effect of batch size b is studied for the two best performing methods.

7.2 REAL-WORLD SIMULATION MODELS

Lorenz model. This modified version of the well-known Lorenz weather prediction model describes the dynamics of slow weather variables and their dependence on unobserved fast weather variables over a certain period of time. The model is represented by a coupled stochastic differential equation which can only be solved numerically resulting in an intractable likelihood function. The model has two parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)$ which we estimate from timeseries data generated using $\boldsymbol{\theta} = (2, 0.1)$. See Thomas et al. (2018) for full details of the model and the experimental set-up that we also use here, with the exception that we use wider uniform prior $\boldsymbol{\theta} \sim \mathcal{U}([0, 5] \times [0, 0.5])$. The discrepancy is formed as a Mahalanobis distance from the six summary statistics by Hakkarainen et al. (2012). The results are shown in Fig. 4(a). Furthermore, Fig. 4(b-c) demonstrates the uncertainty quantification of the model-based ABC posterior expectation. See Appendix B.1 for the details of the numerical computations used. The effect of batch size is shown in Fig. 5(c).

Bacterial infections model. This model describes transmission dynamics of bacterial infections in day care centres and features intractable likelihood. The model has been developed by Numminen et al. (2013) and used previously by Gutmann and Corander (2016); Järvenpää et al. (2019) as an ABC benchmark problem. We estimate the internal, external and co-infection parameters $\beta \in [0, 11], \Lambda \in [0, 2]$ and $\theta \in [0, 1]$, respectively, using true data (Numminen et al., 2013) and uniform priors. The discrepancy is formed as in Gutmann and Corander (2016), see Appendix B.3 for details. The results with all methods are shown in Fig. 5(a) and Fig. 5(b) shows the effect of batch size for the two best performing methods.

Additional details, e.g., on the optimisation of the acquisition function, MCMC methods used, computational costs, and additional experimental results can be found in the

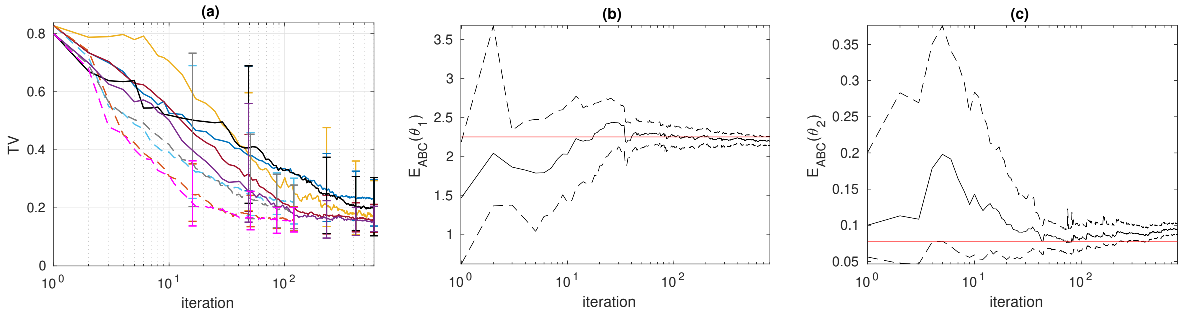

Figure 4: (a) Lorenz model. The intervals show the 90% variability. See Fig. 2 for the legend. (b-c) Black line is the mean and dashed black the 95% CI of the ABC posterior expectations. Red line shows the true value.

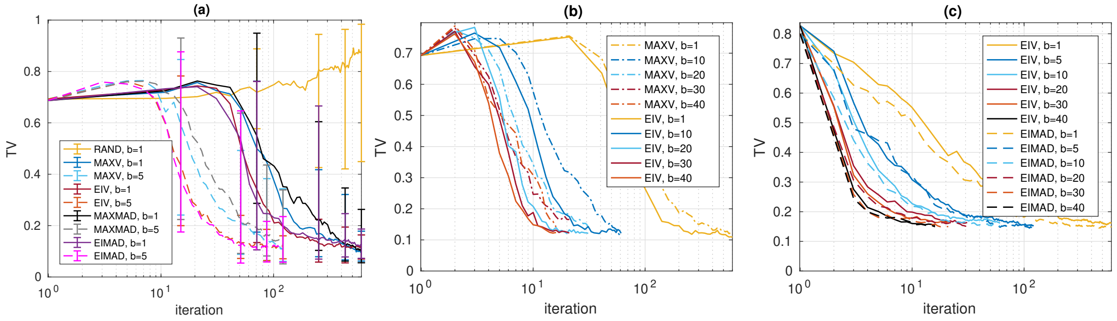

Figure 5: (a) Bacterial infections model. The intervals show the 90% variability. (b) Bacterial infections model with different batch sizes and two chosen acquisition methods. (c) Additional experiments with Lorenz model.

Appendix B and C. The results for our third, additional ABC benchmark scenario, **g-and-k model**, are shown in the Appendix D.

7.3 DISCUSSION ON THE RESULTS

In general, we obtain reasonable posterior approximations considering the very limited budget of simulations. EIV and EIMAD tend to produce more stable, accurate but also more conservative estimates than MAXV and MAXMAD. Difference in approximation quality between EIV and EIMAD, both based on the same Bayesian decision theoretic framework but different loss functions, was small. While RAND worked well in 2D cases and is fully parallellisable, it unsurprisingly produced poor posterior approximations in higher dimensions. In all cases, our batch strategies produced similar evaluation locations as the corresponding sequential methods leading to substantial improvements in wall-time when the simulations are costly. Unlike in the related problem of BO, batch points need not always be diverse because the simulations are stochastic and simulating multiple times at nearby points can be useful. On the other hand, already a single simulation can be enough to effectively rule out large tail regions. The proposed methods automatically balance between these two situations.

Fig. 4(b-c) shows the evolution of the uncertainty in the ABC posterior expectation of the Lorenz model over 800 iterations in the case of sequential EIV. The convergence is approximately towards the true ABC posterior expectation due to a slight GP misspecification. Similarly, the ABC posterior marginals of the bacterial infection model in Appendix C contain some uncertainty after 600 iterations which our approach allows to rigorously quantify. Due to the approximations involved and because this approach is not designed to account for the error due to approximating the intractable ground-truth posterior with the ABC posterior in the first place, we however suggest to interpret the uncertainty estimates with care. Developing more effective (analytical) methods for computing these uncertainty estimates is an interesting avenue for future work. The connection to BQ methods outlined in Section 6.1 can be helpful for achieving this goal.

8 CONCLUSIONS

We considered ABC inference with a limited number of simulations ($t \lesssim 1000$). We outlined a GP surrogate modelling framework called Bayesian ABC where the uncertainty of the ABC posterior distribution due to the limited computational resources is approximately quantified. We also developed batch-sequential Bayesian experimental design strategies to efficiently parallellise the expensive simulations. Experiments suggest that substantial gains in wall-time over previous related work can be obtained.

Acknowledgements

We thank Academy of Finland (grants no. 286607 and 294015 to PM) and Finnish Center for Artificial Intelligence for support of this research. We acknowledge the computational resources provided by Aalto Science-IT project.

References

L. Acerbi. Variational Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems 31*, pages 8223–8233. 2018.

F. Bach. *Learning with Submodular Functions: A Convex Optimization Perspective*. 2013.

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):235–269, 2019.

F.-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic. Probabilistic integration: A role in statistical computation? *Statistical Science*, 34(1): 1–22, 2019.

H. R. Chai and R. Garnett. Improving quadrature for constrained integrands. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2751–2759, 2019.

K. Chaloner and I. Verdinelli. Bayesian Experimental Design: A Review. *Statistical Science*, 10(3):273–304, 1995.

D. Ginsbourger, R. Le Riche, and L. Carraro. *Kriging Is Well-Suited to Parallelize Optimization*, pages 131–162. Springer Berlin Heidelberg, 2010.

D. Greenberg, M. Nonnenmacher, and J. Macke. Automatic posterior transformation for likelihood-free inference. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

T. Gunter, M. A. Osborne, R. Garnett, P. Hennig, and S. J. Roberts. Sampling for inference in probabilistic models with fast Bayesian quadrature. In *Advances in Neural Information Processing Systems 27*. 2014.

M. U. Gutmann and J. Corander. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, 17 (125):1–47, 2016.

H. Haario, M. Laine, A. Mira, and E. Saksman. DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16(4):339–354, 2006.

J. Hakkarainen, A. Ilin, A. Solonen, M. Laine, H. Haario, J. Tamminen, E. Oja, and H. Järvinen. On closure parameter estimation in chaotic systems. *Nonlinear Processes in Geophysis*, 19:127–143, 2012.

P. Hennig and C. J. Schuler. Entropy Search for Information-Efficient Global Optimization. *Journal of Machine Learning Research*, 13(1999):1809–1837, 2012.

J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani. Predictive Entropy Search for Efficient Global Optimization of Black-box Functions. *Advances in Neural Information Processing Systems 28*, pages 1–9, 2014.

P. Holden, N. Edwards, and R. Wilkinson. *ABC for climate: dealing with expensive simulators*, pages 569–95. 2018. In The Handbook of ABC.

M. Järvenpää, M. U. Gutmann, A. Vehtari, and P. Marttinen. Gaussian process modelling in approximate Bayesian computation to estimate horizontal gene transfer in bacteria. *The Annals of Applied Statistics*, 12(4): 2228–2251, 2018.

M. Järvenpää, M. U. Gutmann, A. Pleska, A. Vehtari, and P. Marttinen. Efficient acquisition rules for model-based approximate Bayesian computation. *Bayesian Analysis*, 14(2):595–622, 2019.

M. Järvenpää, M. U. Gutmann, A. Vehtari, and P. Marttinen. Parallel Gaussian process surrogate Bayesian inference with noisy likelihood evaluations. *Bayesian analysis, to appear*, 2020.

K. Kandasamy, J. Schneider, and B. Póczos. Query efficient posterior estimation in scientific experiments via Bayesian active learning. *Artificial Intelligence*, 243: 45–56, 2017.

J. M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

P. Marjoram, J. Molitor, V. Plagnol, and S. Tavare. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15324–8, 2003.

P. Marttinen, M. U. Gutmann, N. J. Croucher, W. P. Hanage, and J. Corander. Recombination produces coherent bacterial species clusters in both core and accessory genomes. *Microbial Genomics*, 1(5), 2015.

E. Meeds and M. Welling. GPS-ABC: Gaussian Process Surrogate Approximate Bayesian Computation. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, 2014.

E. Numminen, L. Cheng, M. Gyllenberg, and J. Corander. Estimating the transmission dynamics of streptococcus pneumoniae from strain prevalence data. *Biometrics*, 69(3):748–757, 2013.

A. O'Hagan. Bayes-Hermite quadrature. *Journal of Statistical Planning and Inference*, 1991.

A. O'Hagan and J. F. C. Kingman. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(1): 1–42, 1978.

M. A. Osborne, D. Duvenaud, R. Garnett, C. E. Rasmussen, S. J. Roberts, and Z. Ghahramani. Active Learning of Model Evidence Using Bayesian Quadrature. *Advances in Neural Information Processing Systems 26*, pages 1–9, 2012.

D. B. Owen. Tables for computing bivariate normal probabilities. *The Annals of Mathematical Statistics*, 27(4): 1075–1090, 1956.

G. Papamakarios and I. Murray. Fast e-free inference of simulation models with Bayesian conditional density estimation. In *Advances in Neural Information Processing Systems 29*, 2016.

G. Papamakarios, D. Sterratt, and I. Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 837–848, 2019.

M. Park, W. Jitkrittum, and D. Sejdinovic. K2-ABC: Approximate Bayesian computation with kernel embeddings. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 398–407, 2016.

M. Patefield and D. Tandy. Fast and accurate calculation of Owen's T function. *Journal of Statistical Software*, 5(5):1–25, 2000.

G. Pleiss, J. Gardner, K. Weinberger, and A. G. Wilson. Constant-time predictive distributions for Gaussian processes. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4114–4123, 2018.

D. Prangle. Adapting the ABC distance function. *Bayesian Analysis*, 12(1):289–309, 2017.

L. F. Price, C. C. Drovandi, A. Lee, and D. J. Nott. Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, 27(1):1–11, 2018.

J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798, 1999.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184. 2008.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

J. Riihimäki and A. Vehtari. Laplace approximation for logistic Gaussian process density estimation and regression. *Bayesian Analysis*, 9(2):425–448, 2014.

K. K. Rogers, H. V. Peiris, A. Pontzen, S. Bird, L. Verde, and A. Font-Ribera. Bayesian emulator optimisation for cosmology: application to the Lyman-alpha forest. *Journal of Cosmology and Astroparticle Physics*, 2019.

M. Sinsbeck and W. Nowak. Sequential design of computer experiments for the solution of Bayesian inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):640–664, 2017.

J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems 25*, pages 1–9, 2012.

N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1015–1022, 2010.

O. Thomas, R. Dutta, J. Corander, S. Kaski, and M. U. Gutmann. Likelihood-free inference by ratio estimation. `arXiv1611.10242`, 2018.

Z. Wang and S. Jegelka. Max-value entropy search for efficient Bayesian optimization. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3627–3635, 2017.

R. D. Wilkinson. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology*, 12(2):129–141, 2013.

R. D. Wilkinson. Accelerating ABC methods using Gaussian processes. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, 2014.

J. Wilson, F. Hutter, and M. Deisenroth. Maximizing acquisition functions for Bayesian optimization. In *Advances in Neural Information Processing Systems 31*, pages 9906–9917. 2018.

S. N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466:1102–1104, 2010.

J. Wu and P. Frazier. The parallel knowledge gradient method for batch Bayesian optimization. In *Advances in Neural Information Processing Systems 29*, pages 3126–3134. 2016.