Efficient acquisition rules for model-based approximate Bayesian computation

Digital Object Identifier (DOI):
10.1214/18-BA1121

Efficient acquisition rules for model-based approximate Bayesian computation

Marko Järvenpää[1], Michael U. Gutmann[2], Arijus Pleska[1], Aki Vehtari[1], and Pekka Marttinen[1]

[1]Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University
[2]School of Informatics, University of Edinburgh

August 9, 2018

Abstract

Approximate Bayesian computation (ABC) is a method for Bayesian inference when the likelihood is unavailable but simulating from the model is possible. However, many ABC algorithms require a large number of simulations, which can be costly. To reduce the computational cost, Bayesian optimisation (BO) and surrogate models such as Gaussian processes have been proposed. Bayesian optimisation enables one to intelligently decide where to evaluate the model next but common BO strategies are not designed for the goal of estimating the posterior distribution. Our paper addresses this gap in the literature. We propose to compute the uncertainty in the ABC posterior density, which is due to a lack of simulations to estimate this quantity accurately, and define a loss function that measures this uncertainty. We then propose to select the next evaluation location to minimise the expected loss. Experiments show that the proposed method often produces the most accurate approximations as compared to common BO strategies.

1 Introduction

We consider the problem of Bayesian inference of some unknown parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ of a simulation model. Such models are typically not amenable to any analytical treatment but they can be simulated with any parameter $\boldsymbol{\theta} \in \Theta$ to produce data $\mathbf{x}_{\boldsymbol{\theta}} \in \mathcal{X}$. Simulation models are also called simulator-based or implicit models [Diggle and Gratton, 1984]. Our prior knowledge about the unknown parameter $\boldsymbol{\theta}$ is represented by the prior probability density $\pi(\boldsymbol{\theta})$ and the goal of the analysis is to update our knowledge about the parameters $\boldsymbol{\theta}$ after we have observed data $\mathbf{x}_{obs} \in \mathcal{X}$.

If evaluating the likelihood function $\pi(\mathbf{x} \,|\, \boldsymbol{\theta})$ is feasible, the posterior distribution can be computed directly using Bayes' theorem

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

In this article we focus on simulation models that have intractable likelihoods. This means that one can only simulate from the model, that is, draw samples $\mathbf{x}_{\boldsymbol{\theta}} \sim \pi(\cdot \,|\, \boldsymbol{\theta})$, but not evaluate the likelihood function $\pi(\mathbf{x}_{obs} \,|\, \boldsymbol{\theta})$ at all so that the standard Bayesian approach cannot be used. For example, possibly high-dimensional unobservable latent random quantities present in the simulation model can make evaluating the likelihood impossible. Such difficulties occur in many areas of science, and typical application fields include population genetics Beaumont et al. [2002], Numminen et al. [2013], genomics Marttinen et al.

[2015], Järvenpää et al. [2017], ecology [Wood, 2010, Hartig et al., 2011] and psychology Turner and Van Zandt [2012], see e.g. Lintusaari et al. [2017] and references therein for further examples.

Approximate Bayesian computation (ABC) replaces likelihood evaluations with model simulations[1], see e.g. Marin et al. [2012], Turner and Van Zandt [2012], Lintusaari et al. [2017] for an overview. The main idea of the basic ABC algorithm is to draw a parameter value from the prior distribution, simulate a data set with the given parameter value, and accept the value as a draw from the (approximate) posterior if the discrepancy between the simulated and observed data is small enough. This algorithm produces samples from the approximate posterior distribution

$$\pi_\varepsilon^{\mathrm{ABC}}(\boldsymbol{\theta}\,|\,\mathbf{x}_{obs}) \propto \pi(\boldsymbol{\theta}) \int \pi_\varepsilon(\mathbf{x}_{obs}\,|\,\mathbf{x})\pi(\mathbf{x}\,|\,\boldsymbol{\theta})\,\mathrm{d}\mathbf{x}, \tag{2}$$

where $\pi_\varepsilon(\mathbf{x}_{obs}\,|\,\mathbf{x}) \propto \mathbb{1}_{\Delta(\mathbf{x}_{obs},\mathbf{x})\leq\varepsilon}$, although other choices of π_ε are also possible [Wilkinson, 2013]. The function $\Delta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is the discrepancy that tells how different the simulated and observed data sets are, and it is often formed by combining a set of summary statistics even though, occasionally, the output data have a relatively small dimension and the data sets can be compared directly. Sometimes the discrepancy function may be available from previous analyses with similar models or can be constructed based on expert knowledge of the application field. The discrepancy and the summaries affect the approximations and their choice is an active research topic [Blum et al., 2013, Fearnhead and Prangle, 2012, Gutmann et al., 2017]. In this work, we are concerned with another, equally important, research question, namely given a suitable discrepancy function, how to perform the inference in a computationally efficient manner.

The threshold ε controls the trade-off between the accuracy of the approximation and computational cost: a small ε yields accurate approximations but requires more simulations, see e.g. Marin et al. [2012]. Given t samples from the model for some $\boldsymbol{\theta}$, that is, $\mathbf{x}_{\boldsymbol{\theta}}^{(i)} \sim \pi(\cdot\,|\,\boldsymbol{\theta})$ for $i = 1, \ldots, t$, the value of the ABC posterior in Equation (2) can be estimated as

$$\pi_\varepsilon^{\mathrm{ABC}}(\boldsymbol{\theta}\,|\,\mathbf{x}_{obs}) \mathrel{\overset{\sim}{\propto}} \pi(\boldsymbol{\theta}) \sum_{i=1}^{t} \pi_\varepsilon(\mathbf{x}_{obs}\,|\,\mathbf{x}_{\boldsymbol{\theta}}^{(i)}), \tag{3}$$

where "$\overset{\sim}{\propto}$" means that the left-hand side is approximately proportional to the right-hand side and where the extra approximation is due to replacing the integral with the Monte-Carlo sum.

Algorithms based on Markov Chain Monte Carlo and sequential Monte Carlo have been used to improve the efficiency of ABC as compared to the basic rejection sampler [Marjoram et al., 2003, Sisson et al., 2007, Beaumont et al., 2009, Toni et al., 2009, Marin et al., 2012, Lenormand et al., 2013]. Unfortunately, the sampling based methods still require a very large number of simulations. In this paper we focus on the challenging scenario where the number of available simulations is limited, e.g. to fewer than a thousand, rendering these sampling-based ABC methods infeasible. Different modelling approaches have also been proposed to reduce the number of simulations required. For example, in the synthetic likelihood method summary statistics are assumed to follow the Gaussian density [Wood, 2010, Price et al., 2017] and the resulting likelihood approximation can be used together with MCMC but evaluating the synthetic likelihood is typically still very expensive. Wilkinson [2014], Meeds and Welling [2014], Jabot et al. [2014], Kandasamy et al. [2015], Drovandi et al. [2015], Gutmann and Corander [2016], Järvenpää et al. [2017] all use Gaussian processes (GP) to accelerate ABC in various ways. Some other alternative approaches are considered by Fan et al. [2013], Papamakarios and Murray [2016]. Also, Beaumont et al. [2002], Blum [2010], Blum and François [2010] have used modelling as a post-processing step to correct the approximation error of the nonzero threshold.

While probabilistic modelling has been used to accelerate ABC inference, and strategies have been proposed for selecting which parameter to simulate next, little work has focused on trying to quantify the amount of uncertainty in the estimator of the ABC posterior density under the chosen modelling assumptions. This

[1] Such approaches are also called likelihood-free in the literature although this name can be considered a misnomer. Namely, while the user does not need to provide the likelihood of the simulator model, many methods construct some sort of likelihood approximation implicitly or explicitly.

uncertainty is due to a finite computational budget to perform the inference and could be thus also called as "computational uncertainty". Consequently, little has been done to design strategies that directly aim to minimise this uncertainty. To our knowledge, only Kandasamy et al. [2015] have used the uncertainty in the likelihood function to propose new simulation locations in a query-efficient way, but they assumed that the likelihood can be evaluated, although with high a computational cost. Also, Wilkinson [2014] modelled the uncertainty in the likelihood to rule out regions with negligible posterior probability. Rasmussen [2003] used GP regression to accelerate Hybrid Monte Carlo but did not consider the setting of ABC. Osborne et al. [2012] developed an active learning scheme to select evaluations to estimate integrals such as the model evidence under GP modelling assumptions, however, their approach is designed for estimating this particular scalar value. Finally, Gutmann and Corander [2016] proposed Bayesian optimisation to efficiently select new evaluation locations. While the BO strategies they used to illustrate the framework worked reasonably, their approach does not directly address the goal of ABC, that is to learn the posterior accurately.

In this article we propose an acquisition function for selecting the next evaluation location tailored specifically for ABC. The acquisition function measures the expected uncertainty in the estimate of the (unnormalised) ABC posterior density function over a future evaluation of the simulation model, and proposes the next simulation location so that this expected uncertainty is minimised. We also consider some variants of this strategy. More scecifically, in Section 2 we formulate our probabilistic approach on a general level. In Section 3 we propose a particular algorithm, based on modelling the discrepancy with a GP. Section 4 contains experiments. Some additional details of our algorithms are discussed in Section 5 and Section 6 concludes the article. Technical details and additional experiments are presented in the supplementary material.

2 Problem formulation

We start by presenting the main idea of the probabilistic framework for query-efficient ABC inference. Suppose we have training data $D_{1:t} = \{(\mathbf{x}_i, \boldsymbol{\theta}_i)\}_{i=1}^t$ of simulation outputs $\mathbf{x}_i \in \mathcal{X}$ that were generated by simulating the model with parameters $\boldsymbol{\theta}_i$. Suppose also that we have a Bayesian model that describes our uncertainty about the future simulation output $\mathbf{x}^* \in \mathcal{X}$ with parameter $\boldsymbol{\theta}^*$ conditional on the training data $D_{1:t}$. This uncertainty is represented by a probability measure $\Pi(\mathrm{d}\mathbf{x}^* \,|\, \boldsymbol{\theta}^*, D_{1:t})$[2]. Instead of modelling the full data $\mathbf{x}^* \in \mathcal{X}$, we note that in practice it is reasonable to model only some summary statistics $s(\mathbf{x}^*) \in S \subset \mathbb{R}^r$. Alternatively, the discrepancy between the observed data and simulator output can be modelled as is done later in this article. Importantly, our estimate for the ABC posterior probability density function π^{ABC} actually depends on the training data if e.g. Equation (3) is used, and can therefore also be considered a random quantity. Given the training data $D_{1:t}$, we assume that, using our Bayesian model, we can represent the uncertainty in π^{ABC} using a probability measure $\Pi(\mathrm{d}\pi^{\mathrm{ABC}} \,|\, D_{1:t})$ over the space of (suitable smooth) density functions $\pi^{\mathrm{ABC}} : \Theta \to \mathbb{R}_+$, where the probability measure Π now describes the uncertainty in the ABC posterior.

If the amount of available simulations is limited due to a high computational cost, we may have considerable uncertainty of the ABC posterior π^{ABC}. Let $\mathcal{L}_{\pi^{\mathrm{ABC}}}(D_{1:t})$ denote the loss due to our uncertainty about the ABC posterior density. This loss function could, for example, measure overall uncertainty in the probability density π^{ABC} or the uncertainty of a particular point estimate of interest such as the posterior mean. In the latter case, for a scalar θ, we could choose $\mathcal{L}_{\pi^{\mathrm{ABC}}}(D_{1:t}) = \mathbb{V}(\int_\Theta \theta \pi^{\mathrm{ABC}}(\theta) \,\mathrm{d}\theta \,|\, D_{1:t})$, where the variance (assuming it exists) is taken with respect to the probability measure $\Pi(\mathrm{d}\pi^{\mathrm{ABC}} \,|\, D_{1:t})$.

We consider the sequential setting where, at each iteration, we need to decide the next evaluation location. After each iteration, we can compute the uncertainty in the ABC posterior and the corresponding loss function, and fit a model that predicts the next simulation output, given all data available at the time. Our aim is to choose the next evaluation location $\boldsymbol{\theta}^* = \boldsymbol{\theta}_{t+1}$ such that the expected loss, after having simulated

[2] We use $\Pi(\cdot)$ to denote the probability measure of a random quantity that can be interpreted from the argument. Similarly, $\pi(\cdot)$ denotes a probability density function whenever the corresponding random vector is assumed to be absolutely continuous.

the model at this location, is minimised. That is, we want to minimise

$$\mathbb{E}_{\mathbf{x}^* \,|\, \boldsymbol{\theta}^*, D_{1:t}}(\mathcal{L}_{\pi^{\mathrm{ABC}}}(D_{1:t} \cup \{\mathbf{x}^*, \boldsymbol{\theta}^*\})) = \int \mathcal{L}_{\pi^{\mathrm{ABC}}}(D_{1:t} \cup \{\mathbf{x}^*, \boldsymbol{\theta}^*\}) \Pi(\mathrm{d}\mathbf{x}^* \,|\, \boldsymbol{\theta}^*, D_{1:t}) \tag{4}$$

with respect to $\boldsymbol{\theta}^*$, where we need to average over unknown unknown simulator output $\mathbf{x}^* = \mathbf{x}_{t+1}$ at parameter $\boldsymbol{\theta}^*$ using the model for the new simulator output $\Pi(\mathrm{d}\mathbf{x}^* \,|\, \boldsymbol{\theta}^*, D_{1:t})$. If the loss function measures the uncertainty of the ABC posterior density, then this approach is, by construction, a one step-ahead Bayes optimal solution to a decision problem of minimising the expected uncertainty and offers thus a query-efficient approach to determine the next evaluation location.

This approach resembles the entropy search (ES) method [Hennig and Schuler, 2012, Hernández-Lobato et al., 2014]. Other related approaches have been proposed by Wang et al. [2016], Bijl et al. [2016], Wang and Jegelka [2017]. Different from these approaches, our main goal is to select the parameter for a future run of the costly simulation model so that the uncertainty in the approximate posterior is minimised. ES, in contrast, is designed for query-efficient global optimisation and it aims to find a parameter value that maximises the objective function, and to minimise the uncertainty related to this maximiser. We note that the rationale of our approach is essentially the same as in probabilistic numerics literature (see e.g. Hennig et al. [2015]) or in sequential Bayesian experimental design (see Ryan et al. [2016] for a recent survey). However, different from these approaches, our interest is to design the evaluations to minimise the uncertainty in a quantity that itself describes the uncertainty of the parameters of a costly simulation model. The uncertainty in the former is due to a limited budget for model simulations that we can control, while the uncertainty in the latter is caused by noisy observations that have already been provided to us and are considered here as fixed.

The framework outlined above requires some modelling choices and can lead to computational challenges as the selection of the future evaluation location itself can require costly evaluations (as is the case of ES). In the following section we propose an efficient algorithm based on a loss function that measures the uncertainty in the (unnormalised) ABC posterior over the parameter space and a GP surrogate model. We also consider some alternative strategies that are more heuristic but easier to evaluate. While our approach can be extended to a batch setting where multiple acquisitions are computed in parallel, in this article we restrict our discussion to the sequential case. We note that the outlined strategy is “myopic”, meaning that the expected uncertainty after the next evaluation is considered only, and the number of simulations left in a limited budget is not taken into account, see e.g. González et al. [2016] for some discussion in BO context; non-myopic strategies are also beyond the scope of this work.

Details of our approach appear in the next section, but the main idea is illustrated in Figure 1. We model the discrepancy $\Delta_{\boldsymbol{\theta}} = \Delta(\mathbf{x}_{obs}, \mathbf{x}_{\boldsymbol{\theta}})$ with GP regression (Figure 1a). The ABC posterior is proportional to the prior times the probability of obtaining a discrepancy realisation that is below the threshold when the model is simulated. However, because the GP is fitted with limited training data, this probability cannot be estimated exactly, causing uncertainty in the ABC posterior density function (Figure 1b). We propose an acquisition function that selects the next evaluation location to minimise the expected variance of the (unnormalised) ABC posterior density over the parameter space.

3 Nonparametric modelling and parameter acquisition

This section contains the details of our algorithms. Section 3.1 describes the GP model for the discrepancy, which permits closed-form equations for many of the required quantities to estimate the posterior, which are derived in Section 3.2. In Sections 3.3 and 3.4 we formulate the proposed acquisition functions, and handling uncertainty in GP hyperparameters is briefly discussed in Section 3.5.

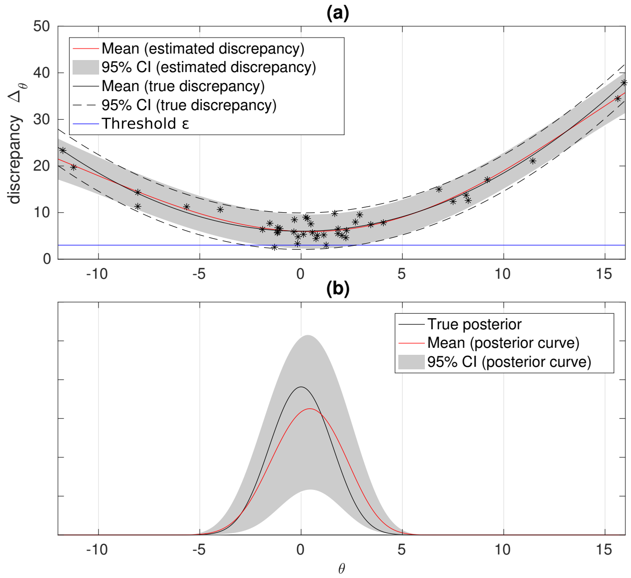

Figure 1: (a) The estimated and true discrepancy distributions are compared. Most of the evaluations are successfully chosen on the modal area of the posterior, leading to a good approximation there. (b) The red curve shows the mean of the unnormalised posterior density function and the grey area its 95% pointwise credible interval.

3.1 GP model for the discrepancy

We consider the discrepancy $\Delta_{\boldsymbol{\theta}}$ a stochastic process indexed by $\boldsymbol{\theta}$ i.e. a random function of the parameter $\boldsymbol{\theta}$. We assume that the discrepancy can be modelled by a Gaussian distribution[3], that is $\Delta_{\boldsymbol{\theta}} \sim \mathcal{N}(f(\boldsymbol{\theta}), \sigma_n^2)$ for some unknown suitably smooth function $f : \Theta \to \mathbb{R}$ and variance $\sigma_n^2 \in \mathbb{R}_+$ both of which need to be estimated. We place a Gaussian process prior on f so that $f \sim \mathcal{GP}(\mu(\boldsymbol{\theta}), k(\boldsymbol{\theta}, \boldsymbol{\theta}'))$. While other choices are also possible, in this paper we consider $\mu(\boldsymbol{\theta}) = 0$, and use the squared exponential covariance function $k(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sigma_f^2 \exp(-\sum_{i=1}^p (\theta_i - \theta_i')^2/(2l_i^2))$. There are thus $p + 2$ hyperparameters to infer, denoted by $\boldsymbol{\phi} = (\sigma_f^2, l_1, \ldots, l_p, \sigma_n^2)$.

Conditioned on the obtained training data $D_{1:t} = \{(\Delta_i, \boldsymbol{\theta}_i)\}_{i=1}^t$, which consists of realised discrepancy-parameter pairs, and the GP hyperparameters $\boldsymbol{\phi}$, our knowledge of the function f evaluated at an arbitrary point $\boldsymbol{\theta} \in \Theta$ can be shown to be $f(\boldsymbol{\theta}) \,|\, D_{1:t}, \boldsymbol{\theta}, \boldsymbol{\phi} \sim \mathcal{N}(m_{1:t}(\boldsymbol{\theta}), v_{1:t}^2(\boldsymbol{\theta}))$, where

$$m_{1:t}(\boldsymbol{\theta}) = k(\boldsymbol{\theta}, \boldsymbol{\theta}_{1:t}) K(\boldsymbol{\theta}_{1:t})^{-1} \Delta_{1:t}, \tag{5}$$

$$v_{1:t}^2(\boldsymbol{\theta}) = k(\boldsymbol{\theta}, \boldsymbol{\theta}) - k(\boldsymbol{\theta}, \boldsymbol{\theta}_{1:t}) K^{-1}(\boldsymbol{\theta}_{1:t}) k(\boldsymbol{\theta}_{1:t}, \boldsymbol{\theta}) \tag{6}$$

and $K(\boldsymbol{\theta}_{1:t}) = k(\boldsymbol{\theta}_{1:t}, \boldsymbol{\theta}_{1:t}) + \sigma_n^2 \mathbf{I}$. Above we defined $k(\boldsymbol{\theta}, \boldsymbol{\theta}_{1:t}) = (k(\boldsymbol{\theta}, \boldsymbol{\theta}_1), \ldots, k(\boldsymbol{\theta}, \boldsymbol{\theta}_t))^T$, $k(\boldsymbol{\theta}_{1:t}, \boldsymbol{\theta}_{1:t})_{ij} = k(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ for $i, j = 1, \ldots, t$ and similarly for $k(\boldsymbol{\theta}_{1:t}, \boldsymbol{\theta})$. We have also used $\Delta_{1:t} = (\Delta_1, \ldots, \Delta_t)^T$. A comprehensive presentation of GP regression can be found in Rasmussen and Williams [2006].

3.2 Quantifying the uncertainty of the ABC posterior estimate

As in Gutmann and Corander [2016], one can compute the posterior predictive density for a new discrepancy value at $\boldsymbol{\theta}$ using $\Delta_{\boldsymbol{\theta}} \,|\, D_{1:t}, \boldsymbol{\theta}, \boldsymbol{\phi} \sim \mathcal{N}(m_{1:t}(\boldsymbol{\theta}), v_{1:t}^2(\boldsymbol{\theta}) + \sigma_n^2)$ and obtain a model-based estimate for the

[3] Alternatively, we could model some transformation of the discrepancy such as $\log \Delta_{\boldsymbol{\theta}}$. In that case, the the following analysis goes similarly.

acceptance probability given the modelling assumptions and training data $D_{1:t}$, as

$$\mathbb{P}(\Delta_{\boldsymbol{\theta}} \leq \varepsilon \,|\, D_{1:t}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \Phi\left((\varepsilon - m_{1:t}(\boldsymbol{\theta})) / \sqrt{\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta})} \right), \tag{7}$$

where $\Phi(z) = \int_{-\infty}^{z} \exp(-t^2/2)\,\mathrm{d}t/(2\pi)$ is the cdf of the standard normal distribution. This probability is approximately proportional to the likelihood and yields a useful point estimate of the likelihood function. We here take a different approach and explicitly exploit the fact that part of the probability mass of $\Delta_{\boldsymbol{\theta}}$ is due to our uncertainty in the latent function f and GP hyperparameters $\boldsymbol{\phi}$. For simplicity, we first assume that the GP hyperparameters $\boldsymbol{\phi}$ are known and discuss relaxing this assumption in a later section. Indeed, if we knew f, the (unnormalised) ABC posterior $\tilde{\pi}_{\varepsilon}^{\mathrm{ABC}}(\boldsymbol{\theta})$ and the acceptance probability $p_a(\boldsymbol{\theta})$[4] could be computed as

$$\tilde{\pi}_{\varepsilon}^{\mathrm{ABC}}(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}) p_a(\boldsymbol{\theta}), \quad p_a(\boldsymbol{\theta}) = \Phi\left(\frac{\varepsilon - f(\boldsymbol{\theta})}{\sigma_n} \right). \tag{8}$$

With a limited number of discrepancy–parameter pairs in $D_{1:t}$ there is uncertainty in the values of the function f (and in GP hyperparameters $\boldsymbol{\phi}$) which we propose to quantify and attempt to minimise in order to accurately estimate the ABC posterior. The following result (whose proof is found in the supplementary material) allows us to compute the expectation and the variance of the unnormalised ABC posterior.

Lemma 3.1. *Under the GP model described in Section 3.1, the pointwise expectation and variance of* $\tilde{\pi}_{\varepsilon}^{ABC}$ *with respect to* $\Pi(\mathrm{d}f \,|\, D_{1:t})$ *are*

$$\mathbb{E}(\tilde{\pi}_{\varepsilon}^{ABC}(\boldsymbol{\theta}) \,|\, D_{1:t}) = \pi(\boldsymbol{\theta})\, \Phi\left(\frac{\varepsilon - m_{1:t}(\boldsymbol{\theta})}{\sqrt{\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta})}} \right), \tag{9}$$

$$\begin{aligned} \mathbb{V}(\tilde{\pi}_{\varepsilon}^{ABC}(\boldsymbol{\theta}) \,|\, D_{1:t}) = \pi^2(\boldsymbol{\theta}) \Bigg[&\Phi\left(\frac{\varepsilon - m_{1:t}(\boldsymbol{\theta})}{\sqrt{\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta})}} \right) \Phi\left(\frac{m_{1:t}(\boldsymbol{\theta}) - \varepsilon}{\sqrt{\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta})}} \right) \\ &- 2T\left(\frac{\varepsilon - m_{1:t}(\boldsymbol{\theta})}{\sqrt{\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta})}}, \frac{\sigma_n}{\sqrt{\sigma_n^2 + 2v_{1:t}^2(\boldsymbol{\theta})}} \right) \Bigg], \end{aligned} \tag{10}$$

where $m_{1:t}(\boldsymbol{\theta})$ *and* $v_{1:t}^2(\boldsymbol{\theta})$ *are given by Equations (5) and (6), respectively, and* $T(\cdot,\cdot)$ *is Owen's t-function which satisfies*

$$T(h, a) = \frac{1}{2\pi} \int_0^a \frac{e^{-h^2(1+x^2)/2}}{1+x^2}\, \mathrm{d}x, \tag{11}$$

for $h, a \in \mathbb{R}$.

We note that Equation (9) equals the product of the prior density $\pi(\boldsymbol{\theta})$ and the point estimate of the likelihood shown in Equation 7 which was used in Gutmann and Corander [2016]. The variance of the unnormalised ABC posterior in Equation (10) depends on Owen's t-function that needs to be computed numerically. However, there exists an efficient algorithm to evaluate its values by Patefield and Tandy [2000].

It is of interest to examine when the variance in Equation (10) is large. If a parameter $\boldsymbol{\theta}$ satisfies $m_{1:t}(\boldsymbol{\theta}) = \varepsilon$, then the first term of Equation (10) is maximised, and in this case the second term is maximised for $\boldsymbol{\theta}$ values where the posterior variance $v_{1:t}^2(\boldsymbol{\theta})$ is large. On the other hand, if $m_{1:t}(\boldsymbol{\theta}) \gg \varepsilon$ but $v_{1:t}(\boldsymbol{\theta}) \gg |m_{1:t}(\boldsymbol{\theta}) - \varepsilon|$, the first term in Equation (10) is approximately maximised and the second term is also close to its maximum value, especially if also $v_{1:t}(\boldsymbol{\theta}) \gg \sigma_n$. Because the ABC threshold ε is usually chosen very small, we thus conclude that the variance in Equation (10) tends to be high in regions where the mean of the discrepancy $m_{1:t}(\boldsymbol{\theta})$ is small and/or the variance of the latent function $v_{1:t}^2(\boldsymbol{\theta})$ is large relative to the mean function.

[4] This notation should not be confused with a probability distribution function which is always denoted with $\pi(\cdot)$.

Some further insight to Equation (10) is obtained by using the approximation $\mathbb{V}_{f(\boldsymbol{\theta})\,|\,D_{1:t}}(\tilde{\pi}_{\varepsilon}^{\mathrm{ABC}}(\boldsymbol{\theta})) \approx ((\tilde{\pi}_{\varepsilon}^{\mathrm{ABC}})'(\mathbb{E}_{f(\boldsymbol{\theta})\,|\,D_{1:t}}(f(\boldsymbol{\theta}))))^2\mathbb{V}_{f(\boldsymbol{\theta})\,|\,D_{1:t}}(f(\boldsymbol{\theta}))$, where the formula $(\tilde{\pi}_{\varepsilon}^{\mathrm{ABC}})'(\mathbb{E}_{f(\boldsymbol{\theta})\,|\,D_{1:t}}(f(\boldsymbol{\theta})))$ denotes the derivative of $\tilde{\pi}_{\varepsilon}^{\mathrm{ABC}}$ with respect to $f(\boldsymbol{\theta})$ evaluated at $\mathbb{E}_{f(\boldsymbol{\theta})\,|\,D_{1:t}}(f(\boldsymbol{\theta}))$. This approximation, also known as the delta method, produces

$$-\log\mathbb{V}_{f(\boldsymbol{\theta})\,|\,D_{1:t}}(\tilde{\pi}_{\varepsilon}^{\mathrm{ABC}}(\boldsymbol{\theta})) \approx \frac{(\varepsilon - m_{1:t}(\boldsymbol{\theta}))^2}{\sigma_n^2} - \log(v_{1:t}^2(\boldsymbol{\theta})) - 2\log\pi(\boldsymbol{\theta}) + \log(2\pi\sigma_n^2), \tag{12}$$

where the last term is constant and can be dropped. This equation has some similarity with the lower confidence bound (LCB) criteria used in bandit problems and Bayesian optimisation

$$\mathrm{LCB}(\boldsymbol{\theta}) = m_{1:t}(\boldsymbol{\theta}) - \beta_t\sqrt{v_{1:t}^2(\boldsymbol{\theta})}, \tag{13}$$

where β_t is a tradeoff parameter. Both equations produce small values if the mean of the discrepancy $m_{1:t}(\boldsymbol{\theta})$ is small (assuming also that $\varepsilon \leq m_{1:t}(\boldsymbol{\theta})$) or the variance $v_{1:t}^2(\boldsymbol{\theta})$ is large relative to the mean $m_{1:t}(\boldsymbol{\theta})$. However, the LCB tradeoff parameter β_t typically depends on the iteration t and the dimension of the parameter space (see Srinivas et al. [2010] for theoretical analysis) while in the posterior variance (Equation (12)) this tradeoff is determined automatically in a nonlinear fashion and, unlike for LCB, the variance formula depends also on the prior density $\pi(\boldsymbol{\theta})$.

Other useful facts for $\tilde{\pi}_{\varepsilon}^{\mathrm{ABC}}$ can also be obtained. Some of these formulas are not required to apply our methodology and are thus included as supplementary material. For instance, if $\pi(\boldsymbol{\theta}) > 0$ then the cdf for $\tilde{\pi}_{\varepsilon}^{\mathrm{ABC}}(\boldsymbol{\theta})$ is

$$F_{\tilde{\pi}_{\varepsilon}^{\mathrm{ABC}}(\boldsymbol{\theta})}(z) = \Phi\left(\frac{\sigma_n\Phi^{-1}(z/\pi(\boldsymbol{\theta})) + m_{1:t}(\boldsymbol{\theta}) - \varepsilon}{v_{1:t}(\boldsymbol{\theta})}\right), \tag{14}$$

if $z \in (0, \pi(\boldsymbol{\theta}))$, zero if $z \leq 0$, and 1 if $z \geq \pi(\boldsymbol{\theta})$. This formula enables the computation of quantiles which can be used for assessing the uncertainty via credible intervals. Setting $\alpha = F_{\tilde{\pi}_{\varepsilon}^{\mathrm{ABC}}(\boldsymbol{\theta})}(z)$, where $\alpha \in (0,1)$ and solving for z yields the α-quantile that was already used in Figure 1b,

$$z_{\alpha} = \pi(\boldsymbol{\theta})\Phi\left(\frac{v_{1:t}(\boldsymbol{\theta})\Phi^{-1}(\alpha) - m_{1:t}(\boldsymbol{\theta}) + \varepsilon}{\sigma_n}\right). \tag{15}$$

From the above equation we see, e.g., that the median is given by $\pi(\boldsymbol{\theta})\Phi((\varepsilon - m_{1:t}(\boldsymbol{\theta}))/\sigma_n)$.

Above we assumed that the GP hyperparameters $\boldsymbol{\phi}$ are known but in practice these need to be estimated. One can use the MAP-estimate in the place of the fixed values in the previous formulae. The MAP-estimate is computed by maximising the logarithm of the marginal posterior

$$\boldsymbol{\phi}_{1:t}^{\mathrm{MAP}} = \arg\max_{\boldsymbol{\phi}}\left(\log\pi(\boldsymbol{\phi}) - \frac{1}{2}\Delta_{1:t}^T K^{-1}(\boldsymbol{\theta}_{1:t})\Delta_{1:t} - \frac{1}{2}\log\det(K(\boldsymbol{\theta}_{1:t}))\right), \tag{16}$$

where $\pi(\boldsymbol{\phi})$ is the prior density for GP hyperparameters and where the covariance function in $K(\boldsymbol{\theta}_{1:t}) = k(\boldsymbol{\theta}_{1:t}, \boldsymbol{\theta}_{1:t}) + \sigma_n^2\mathbf{I}$ depends naturally also on $\boldsymbol{\phi}$. For the rest of the paper, we assume that the MAP estimate is used for GP hyperparameters, however, we also briefly discuss how one could integrate over them in Section 3.5.

3.3 Efficient parameter acquisition

We define our loss function $\mathcal{L}_{\pi_{\varepsilon}^{\mathrm{ABC}}}$ for model-based ABC inference as

$$\mathcal{L}_{\pi_{\varepsilon}^{\mathrm{ABC}}}(D_{1:t}) = \int_{\Theta}\mathbb{V}(\tilde{\pi}_{\varepsilon}^{\mathrm{ABC}}(\boldsymbol{\theta})\,|\,D_{1:t})\,\mathrm{d}\boldsymbol{\theta} = \int_{\Theta}\pi^2(\boldsymbol{\theta})\mathbb{V}(p_a(\boldsymbol{\theta})\,|\,D_{1:t})\,\mathrm{d}\boldsymbol{\theta}, \tag{17}$$

where $\tilde{\pi}_{\varepsilon}^{\mathrm{ABC}}(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})p_a(\boldsymbol{\theta})$ is the unnormalised ABC posterior and the variance is taken with respect to the unknown latent function f conditioned on the training data $D_{1:t}$. We call the function in Equation (17) as the integrated variance loss function. It measures the uncertainty in the unnormalised ABC posterior density averaged over the parameter space Θ. The loss function is defined in terms of the unnormalised ABC posterior because we are here interested in minimising the uncertainty in the posterior shape. Also, this choice allows tractable computations unlike some other potential choices such as defining the integrated variance over the normalised ABC posterior density function. However, in principle, other loss functions, suitable for particular problem at hand, could be defined.

We obtain the following formula for computing the expected integrated variance loss function $L_{1:t}(\boldsymbol{\theta}^*)$ (abbreviated as "expintvar") when the new candidate evaluation location is $\boldsymbol{\theta}^*$. The proof is rather technical and can be found in the supplementary.

Proposition 3.2. *Under the GP model described in Section 3.1, the expected integrated variance after running the simulation model with parameter $\boldsymbol{\theta}^*$ is given by*

$$L_{1:t}(\boldsymbol{\theta}^*) = \mathbb{E}_{\Delta^* \,|\, \boldsymbol{\theta}^*, D_{1:t}} \int_\Theta \pi^2(\boldsymbol{\theta}) \mathbb{V}(p_a(\boldsymbol{\theta}) \,|\, \Delta^*, \boldsymbol{\theta}^*, D_{1:t}) \, \mathrm{d}\boldsymbol{\theta} \tag{18}$$

$$= 2 \int_\Theta \pi^2(\boldsymbol{\theta}) \Bigg[T\left(\frac{\varepsilon - m_{1:t}(\boldsymbol{\theta})}{\sqrt{\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta})}}, \sqrt{\frac{\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta}) - \tau_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta}) + \tau_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}} \right) - T\left(\frac{\varepsilon - m_{1:t}(\boldsymbol{\theta})}{\sqrt{\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta})}}, \frac{\sigma_n}{\sqrt{\sigma_n^2 + 2v_{1:t}^2(\boldsymbol{\theta})}} \right) \Bigg] \mathrm{d}\boldsymbol{\theta}, \tag{19}$$

where the variance of $p_a(\boldsymbol{\theta})$ is taken with respect to $\Pi(\mathrm{d}f \,|\, \Delta^*, \boldsymbol{\theta}^*, D_{1:t})$, the function $T(\cdot, \cdot)$ is the Owen's t-function as in Equation (11) and

$$\tau_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \frac{cov_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta}^*)}, \tag{20}$$

where $cov_{1:t}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = k(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - k(\boldsymbol{\theta}, \boldsymbol{\theta}_{1:t}) K^{-1}(\boldsymbol{\theta}_{1:t}) k(\boldsymbol{\theta}_{1:t}, \boldsymbol{\theta}^*)$ is the posterior covariance between the evaluation point $\boldsymbol{\theta}$ and the candidate location for the next evaluation $\boldsymbol{\theta}^*$.

A future evaluation at $\boldsymbol{\theta}^*$ causes a deterministic reduction of the GP variance that is given by the Equation (20). However, the variance of the unnormalised ABC posterior depends on the realisation of the discrepancy Δ^* and we need to average over $\pi(\Delta^* \,|\, \boldsymbol{\theta}^*, D_{1:t})$. It is easy to see that if $\tau_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \to v_{1:t}^2(\boldsymbol{\theta})$, then the integrand at the corresponding parameter $\boldsymbol{\theta}$ approaches zero. It can also be shown (using Owen [1980, Eq. 2.3]) that if $\tau_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = 0$, then the integrand in Equation (19) equals the current variance given by Equation (10).

While some of the derivations could be done analytically, computing the expected integrated variance requires integration over the parameter space Θ. This can be done with Monte Carlo or quasi-Monte Carlo methods; here we use importance sampling (IS) to approximate the integral when $p > 2$. Using the IS estimator [Robert and Casella, 2004, Eq. 3.10, p. 95], we obtain

$$L_{1:t}(\boldsymbol{\theta}^*) = 2 \int_\Theta \pi^2(\boldsymbol{\theta}) g_{1:t+1}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \, \mathrm{d}\boldsymbol{\theta} \approx 2 \sum_{i=1}^{s} \omega^{(i)} \pi^2(\boldsymbol{\theta}^{(i)}) g_{1:t+1}(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^*), \tag{21}$$

where $g_{1:t+1}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is the term inside the square brackets in Equation (19) and the importance weights are given by

$$\omega^{(i)} = \frac{1}{\pi^2(\boldsymbol{\theta}^{(i)}) \mathbb{V}(p_a(\boldsymbol{\theta}^{(i)}) \,|\, D_{1:t})} \Bigg/ \sum_{j=1}^{s} \frac{1}{\pi^2(\boldsymbol{\theta}^{(j)}) \mathbb{V}(p_a(\boldsymbol{\theta}^{(j)}) \,|\, D_{1:t})}, \tag{22}$$

and $\boldsymbol{\theta}^{(i)} \sim \pi_q(\cdot)$ for $i = 1, \ldots, s$. The importance distribution $\pi_q(\boldsymbol{\theta})$ is proportional to the prior squared times the current variance of the unnormalised ABC posterior i.e. $\pi_q(\boldsymbol{\theta}) \propto \pi^2(\boldsymbol{\theta}) \mathbb{V}(p_a(\boldsymbol{\theta}) \,|\, D_{1:t})$. This importance distribution is a reasonable choice, because one evaluation is unlikely to change the variance surface much and the expected variance thus has similar shape as the current variance surface. It is easy to see that if the prior is bounded and proper i.e. $\pi(\boldsymbol{\theta}) < \infty$ and $\int_\Theta \pi(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta} = 1$, then π_q defines a valid probability density function (up to normalisation). Because the normalising constant of π_q is unavailable, we need to normalise the weights in Equation (22). Generating samples from the importance distribution π_q is not straightforward but can be done using (e.g. adaptive) Metropolis algorithm or using sequential Monte Carlo methods.

As outlined in Section 2, the new evaluation location is chosen to minimise the expected loss, that is

$$\boldsymbol{\theta}_{t+1} \in \{\boldsymbol{\theta} \in \Theta : \boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}^* \in \Theta} L_{1:t}(\boldsymbol{\theta}^*)\}, \tag{23}$$

where the right hand side is a set of parameters because the minimiser may not be unique. We call this new strategy 'an acquisition rule' according to the nomenclature in the Bayesian optimisation literature. Unlike in BO, however, our aim is not to optimise the discrepancy but to minimise our uncertainty in the ABC posterior approximation. The second term in Equation (19) does not depend on $\boldsymbol{\theta}^*$ and its value can be computed just once (or omitted completely) and the normalisation of the prior density $\pi(\boldsymbol{\theta})$ does not affect the solution of (23) as it only scales the objective function. Gradient-based optimisation with multiple starting points can be used for solving (23) and the gradient is derived in the supplementary material. The resulting algorithm for estimating the ABC posterior is outlined as Algorithm 1.

Algorithm 1 GP-based ABC inference using the expected integrated variance acquisition function.

```
Generate initial training locations $\boldsymbol{\theta}_{1:t_0} \sim \pi(\cdot)$
for $t = 1 : t_0$ do
    Simulate $\mathbf{x}_t \sim \pi(\cdot \mid \boldsymbol{\theta}_t)$
    Compute $\Delta_t \leftarrow \Delta(\mathbf{x}_{obs}, \mathbf{x}_t)$
end for
for $t = t_0 : t_{\max} - 1$ do
    Estimate GP hyperparameters $\boldsymbol{\phi}_{1:t}^{\mathrm{MAP}}$ using $D_{1:t}$ and Equation (16)
    Precompute Cholesky factorisation for the GP prediction
    Simulate evaluation points $\boldsymbol{\theta}^{(i)}$ and weights $\omega^{(i)}$ for $i = 1, \ldots, s$ by sampling from $\pi_q(\cdot)$
    Precompute the second term in Equation (19)
    Obtain $\boldsymbol{\theta}_{t+1}$ by solving the optimisation problem in Equation (23)
    Simulate $\mathbf{x}_{t+1} \sim \pi(\cdot \mid \boldsymbol{\theta}_{t+1})$
    Compute $\Delta_{t+1} \leftarrow \Delta(\mathbf{x}_{obs}, \mathbf{x}_{t+1})$
    Update the training data $D_{1:t+1} \leftarrow D_{1:t} \cup \{(\Delta_{t+1}, \boldsymbol{\theta}_{t+1})\}$
end for
Estimate GP hyperparameters $\boldsymbol{\phi}_{1:t_{\max}}^{\mathrm{MAP}}$ using $D_{1:t_{\max}}$ and Equation (16)
Simulate samples $\boldsymbol{\vartheta}^{(1:n)}$ from the density defined by Equation (9)
return $\boldsymbol{\vartheta}^{(1:n)}$ as a sample from the approximate posterior density
```

3.4 Alternative acquisition rules

We briefly discuss some alternative acquisition rules for ABC inference. Their derivations follow directly from our previous analysis and we include these strategies in our experiments in Section 4. One such alternative to the expected integrated variance strategy is to evaluate where the current uncertainty of the unnormalised ABC posterior is highest. This approach is similar to Kandasamy et al. [2015]. This strategy is a reasonable heuristic in the sense that the next evaluation location is where improvement in estimation accuracy is needed most, although it does not account for how large an improvement can be expected at the location, or overall. This approach requires solving the optimisation problem

$$\boldsymbol{\theta}_{t+1} \in \{\boldsymbol{\theta} \in \Theta : \boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}^* \in \Theta} \pi^2(\boldsymbol{\theta}^*)\mathbb{V}(p_a(\boldsymbol{\theta}^*) \,|\, D_{1:t})\} \tag{24}$$

$$= \left\{\boldsymbol{\theta} \in \Theta : \boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}^* \in \Theta} \left(\log \pi(\boldsymbol{\theta}^*) + \log \sqrt{\mathbb{V}(p_a(\boldsymbol{\theta}^*) \,|\, D_{1:t})}\right)\right\}, \tag{25}$$

where the current variance $\mathbb{V}(p_a(\boldsymbol{\theta}) \,|\, D_{1:t}))$ is given by Equation (10) and is taken with respect to $\Pi(\mathrm{d}f \,|\, D_{1:t})$. We call this method the "maxvar" acquisition rule. The gradient of this acquisition function is derived in the supplementary material.

To encourage further exploration, similarly to Gutmann and Corander [2016], we also consider a stochastic variant of the maxvar acquisition rule in Equation (25). Specifically, we generate the evaluation point randomly according to the variance surface $\pi_q(\boldsymbol{\theta}) \propto \pi^2(\boldsymbol{\theta})\mathbb{V}(p_a(\boldsymbol{\theta}) \,|\, D_{1:t})$ which we also use as an importance distribution for the expected integrated variance acquisition function as discussed earlier. That is, instead of finding the maximiser, we generate $\boldsymbol{\theta}_{t+1} \sim \pi_q(\boldsymbol{\theta})$. This strategy requires generating random samples from $\pi_q(\boldsymbol{\theta})$ but sampling (and optimising) the variance function can be done fast compared to the time required to run the simulation model. We call this method "rand_maxvar".

The stochastic acquisition rule is reminiscent of Thompson sampling, but it is actually quite different. In our method, acquisitions are drawn at random from the probability distribution which is proportional to the (point-wise) variance of the approximate posterior density. In Thompson sampling, instead, one generates a posterior density realisation from the model, and chooses the next point as the maximiser of this realisation.

The maxvar and rand_maxvar strategies avoid the integration over the parameter space that is necessary for the expintvar method. However, one could replace the integration in expintvar by only a single evaluation at the candidate point. In other words, this method chooses a location with the highest expected reduction in the uncertainty of the unnormalised ABC posterior in that particular location. We call this variant the "expdiffvar" from now on.

A comparison of the acquisition functions in a one-dimensional toy problem is shown in Figure 2. A simulation model has been run eight times and the acquisition functions for selecting the ninth evaluation location (shown with triangles) are plotted for comparison. The current variance surface (maxvar) and the expected integrated variance (expintvar) function are plotted with three values of the threshold ε. Unlike the current variance surface, the expected integrated variance appears insensitive to the value of the threshold. Figure 2b shows also that using the MAP-estimate for the GP hyperparameters causes underestimation of the variance of the unnormalised ABC posterior. In the next section we show how the uncertainty in GP hyperparameters is (approximately) taken into account.

3.5 Uncertainty in hyperparameters

Above we assumed that either the GP hyperparameters $\boldsymbol{\phi}$ are known or MAP estimates are used. Here we briefly discuss how the uncertainty in the GP hyperparameters could also be taken into account. Integrating over the uncertainty in the GP hyperparameters requires Monte Carlo sampling as in Murray and Adams [2010]. An alternative approach is to use central composite design [Rue et al., 2009, Vanhatalo et al., 2010]. Briefly, in central composite design (CCD) certain design points $\boldsymbol{\phi}^i$ are chosen and each of them is given a weight $\omega^i \propto \pi(\boldsymbol{\phi}^i \,|\, D_{1:t})\gamma^i \propto \pi(D_{1:t} \,|\, \boldsymbol{\phi}^i)\pi(\boldsymbol{\phi}^i)\gamma^i$, where γ^i is a design weight. This approach has the advantage that the amount of design points grows only moderately with increased dimension and has been shown to yield good accuracy in practice. Further details on choosing the design points and their weights are given in Vanhatalo et al. [2010].

Integrating over the uncertainty in the GP hyperparameters $\boldsymbol{\phi}$ in Lemma 3.1 leads to the following calculations. Using the law of total expectation yields

$$\mathbb{E}(p_a(\boldsymbol{\theta})) = \mathbb{E}_{\boldsymbol{\phi}}\mathbb{E}_{f\,|\,\boldsymbol{\phi}}(p_a(\boldsymbol{\theta})) \approx \sum_i \omega^i\, \mathbb{E}_{f\,|\,\boldsymbol{\phi}=\boldsymbol{\phi}^i}(p_a(\boldsymbol{\theta})) = \sum_i \omega^i\, \Phi\left(a(\boldsymbol{\theta}, \boldsymbol{\phi}^i)\right), \tag{26}$$

where the grid points and the corresponding weights are $\boldsymbol{\phi}^i$ and ω^i, respectively, and where $a(\boldsymbol{\theta}, \boldsymbol{\phi}^i) = (\varepsilon - m_{1:t}(\boldsymbol{\theta} \,|\, \boldsymbol{\phi}^i))/\sqrt{(\sigma_n^2)^i + v_{1:t}^2(\boldsymbol{\theta} \,|\, \boldsymbol{\phi}^i)}$. If Monte Carlo sampling is used, then $\omega^i = 1/s$ for all $i = 1, \ldots, s$,

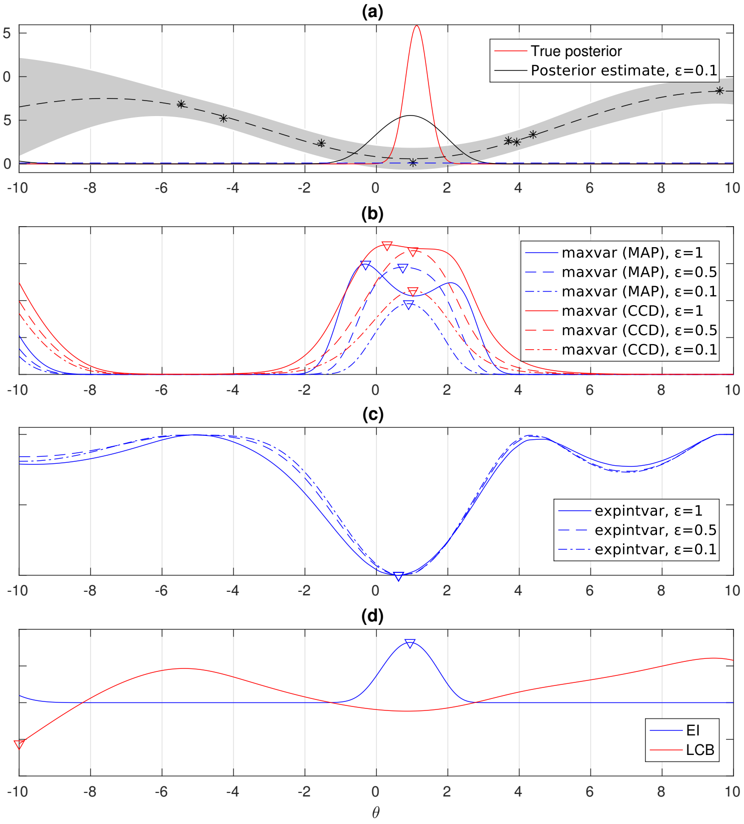

Figure 2: (a) The discrepancy observations (black stars) and the estimate of the ABC posterior density based on eight training data points (with $\varepsilon = 0.1$) as compared to the true posterior. (b) The variance of the unnormalised ABC posterior is computed using the MAP estimate (maxvar (MAP)) or CCD integration (maxvar (CCD)) for GP hyperparameters and for three values of the threshold ε. Details of the CCD integration are in Section 3.5. (c) Expected integrated variance (expintvar) acquisition function. (d) Expected improvement (EI) and lower confidence bound (LCB) criteria (scaled to fit the same figure). Note that the scales of the variance function (b) and acquisition functions computed with different thresholds, (c) and (d), are not comparable.

where s is the number of samples. Similarly, for the variance we obtain

$$\begin{aligned}\mathbb{V}(p_a(\boldsymbol{\theta})) &= \mathbb{E}(p_a(\boldsymbol{\theta})^2) - [\mathbb{E}(p_a(\boldsymbol{\theta}))]^2 \\ &= \mathbb{E}_{\boldsymbol{\phi}}\mathbb{E}_{f\,|\,\boldsymbol{\phi}}(p_a(\boldsymbol{\theta})^2) - [\mathbb{E}_{\boldsymbol{\phi}}\mathbb{E}_{f\,|\,\boldsymbol{\phi}}(p_a(\boldsymbol{\theta}))]^2 \\ &\approx \sum_i \omega^i \left[\Phi(a(\boldsymbol{\theta},\boldsymbol{\phi}^i)) - 2T(a(\boldsymbol{\theta},\boldsymbol{\phi}^i), b(\boldsymbol{\theta},\boldsymbol{\phi}^i))\right] - \left[\sum_i \omega^i\, \Phi(a(\boldsymbol{\theta},\boldsymbol{\phi}^i))\right]^2, \end{aligned} \tag{27}$$

where $b(\boldsymbol{\theta},\boldsymbol{\phi}^i) = (\sigma_n)^i/\sqrt{(\sigma_n^2)^i + 2v_{1:t}^2(\boldsymbol{\theta}\,|\,\boldsymbol{\phi}^i)}$. This formula with CCD integration was already used in Figure 2b.

One can also take into account the uncertainty in GP hyperparameters in the expected integrated variance acquisition function. The posterior predictive distribution for a future simulation is then approximated by a Gaussian mixture and one can make the simplification by (incorrectly) assuming that the future evaluation will not affect the GP hyperparameters but only the latent function f. Evaluating the resulting acquisition function requires a large number of calls to Owen's t-function and GP formulas and is computationally more costly and thus possibly impractical. Alternatively, one can define an integrated[5] acquisition function as in Snoek et al. [2012], Hernández-Lobato et al. [2014], Wang and Jegelka [2017] which only requires computing Equation (19) for each sampled GP hyperparameter $\boldsymbol{\phi}^i$. The integrated acquisition function is then averaged over these values. Alternatively, one could simply use the posterior mean of the hyperparameters in the place of the MAP estimate. However, we leave a detailed analysis for future work.

4 Experiments

We compare the proposed expected integrated variance acquisition rule (expintvar) to commonly used BO strategies: expected improvement (EI) and lower confidence bound (LCB) criterion, see e.g. Shahriari et al. [2015]. We use the same trade-off parameter for LCB as Gutmann and Corander [2016], but unlike them, we consider the deterministic LCB rule. As a simple baseline, we also draw points sequentially from the uniform distribution, abbreviated as "unif". We also included the probability of improvement (PI) strategy in preliminary experiments, but it resulted in poor estimates and was therefore excluded from the comparisons.

In addition to expintvar, we also include the maxvar, rand_maxvar and expdiffvar strategies, which were briefly described in Section 3.4, to our list of methods to be compared. The MAP estimate for the GP hyperparameters $\boldsymbol{\phi}$ is used in all the experiments. We use MATLAB and GPstuff 4.6 [Vanhatalo et al., 2013] for GP fitting. For fast and accurate computation of Owen's t-function, we use a C-implementation of the algorithm by Patefield and Tandy [2000]. The algorithms in this article are also made available in the ELFI (engine for likelihood-free inference) Python software package by Lintusaari et al. [2018].

The total variation (TV) distance is used for assessing the accuracy of the posterior approximation. It is defined as $\text{TV} = 1/2 \int_{\Theta} |\widehat{\pi}_{\varepsilon}^{\text{ABC}}(\boldsymbol{\theta}) - \pi^{\text{true}}(\boldsymbol{\theta})| \, \mathrm{d}\boldsymbol{\theta}$, where $\widehat{\pi}_{\varepsilon}^{\text{ABC}}$ is the estimated ABC posterior and π^{true} is the reference distribution. As the reference we use the exact ABC posterior with the same threshold as used for the approximations but in some scenarios, however, the reference distribution is the exact posterior for computational convenience and to see the overall approximation quality. The point estimate of the ABC posterior density function for the comparisons is always computed using Equation (9). We demonstrate our approach with multiple toy models as well as two realistic models. While the likelihood is actually available for the toy models, we restrict our comparison to the model-based ABC methods, the focus of this work, at the same time acknowledging that in practice with likelihood available the standard methods, such as MCMC, are expected to outperform the likelihood-free alternatives. An overview of the results is given in Table 1 and discussed in detail in the following sections.

4.1 Synthetic 2D simulation models

To compare the different acquisition strategies first without the need to actually handle different simulation models, we construct "synthetic" discrepancies by adding Gaussian noise to certain parametric curves, and use these to simulate the discrepancy realisations directly. The exact ABC posterior that is used as a reference distribution here is computed using the posterior density given by Equation (8) with a small predefined threshold ε. As test cases we consider 1) a unimodal density with two correlated variables, 2) a bimodal density, 3) a density where the first parameter is (almost) unidentifiable, and 4) a banana shaped density. For all cases, a uniform prior was assumed. The resulting exact ABC posterior densities are illustrated in Figure 3. (See supplementary material for additional details). The integration and sampling steps required by expintvar and rand_maxvar strategies are performed in a 2D grid of 50^2 evaluation locations. The initial training set size is $t_0 = 10$ and the initial training sets for the repeated experiments are generated randomly from the uniform prior as is done in the other test cases as well.

The threshold is fixed so that differences in approximation quality between the acquisition methods are solely caused by the selection of the evaluation locations. However, because selecting a reasonable threshold can be challenging in practice, we also examine how updating this value adaptively during the acquisitions

[5] Note that the term "integrated" here refers to integrating over GP hyperparameters $\boldsymbol{\phi}$ and not for integrating over the parameter space Θ as in Equation (19).

	expintvar	expdiffvar	maxvar	rand_maxvar	LCB	EI	unif
unimodal*	1.00	1.39	1.52	1.23	1.27	2.54	**0.97**
bimodal*	**1.00**	1.23	1.24	1.03	1.04	1.51	1.13
unidentifiable*	**1.00**	1.11	1.21	1.12	1.01	1.58	1.49
banana*	**1.00**	1.12	1.23	1.09	1.08	1.67	1.47
Gaussian (strong prior)	**1.00**	1.13	1.18	1.24	1.68	2.68	1.02
Gaussian (weak prior)	**1.00**	1.20	1.16	1.07	1.10	1.59	1.83
Gaussian 3d	1.00	1.26	1.16	**0.94**	1.26	1.67	2.24
Gaussian 6d	1.00	1.06	1.08	**0.98**	1.14	1.42	1.94
Gaussian 10d	**1.00**	1.08	1.08	1.17	1.21	1.51	1.45
Lotka-Volterra	**1.00**	1.20	1.37	1.10	1.15	1.85	1.62

Table 1: Results for the test problems. The numbers in the table represent the median of the area under the TV curve (TV values as a function of iteration) scaled so that the proposed expintvar method obtains value one. Smaller values mean better average performance. In the first four test problems (marked with *), the reference distribution is the exact ABC posterior obtained using the same threshold as the model-based estimate. In the other cases, TV distance is computed with respect to the 'true' posterior. For the Gaussian 3d-10d examples, the TV represents the average TV of marginal densities.

affects the results. In the supplementary we show results when the threshold is constantly updated so that it matches either the 0.01th or the 0.05th quantile of the realised discrepancies.

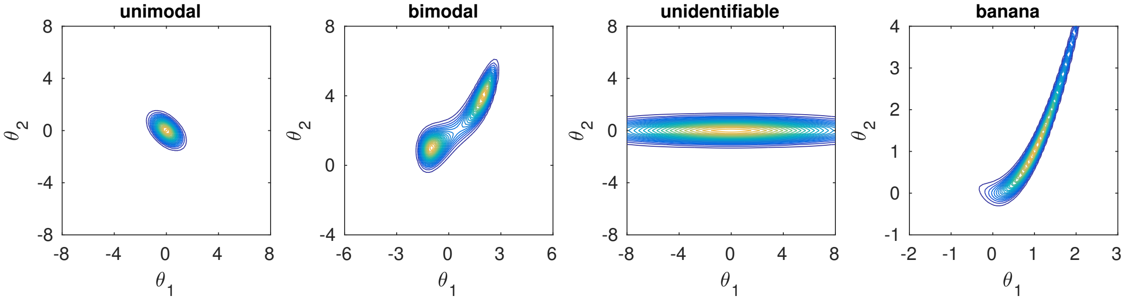

Figure 3: Exact ABC posterior densities for the synthetic 2d test problems.

The results in Figure 4 indicate that the expintvar is the best method overall but also rand_maxvar produces good results. Of the common alternatives, LCB is clearly the best and produces results with similar accuracy as rand_maxvar. The performance of the EI strategy is poor because it tends to focus evaluations greedily around the mode and samples insufficiently in the tail areas which often results in poor estimates to the ABC posterior density and high variability between different experiments. Interestingly, the uniform strategy produces the best estimates in the case of the unimodal example. Most of the acquisitions are not focused on the modal region but because the modelling assumptions hold everywhere and the parameter space is rather small, the extrapolation seems to work well in this case.

4.2 Gaussian simulation model

A simple Gaussian simulation model is used to study the effect of prior strength and the dimension of the parameter space. Data points are generated independently from $\mathbf{x}_i \sim \mathcal{N}(\cdot \,|\, \boldsymbol{\theta}, \boldsymbol{\Sigma}), i = 1, \dots, n$, where $\boldsymbol{\theta} \in \Theta = [0, 8]^p$ needs to be estimated and the covariance matrix $\boldsymbol{\Sigma}$ is known. If $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{a}, \mathbf{B})$ truncated to Θ,

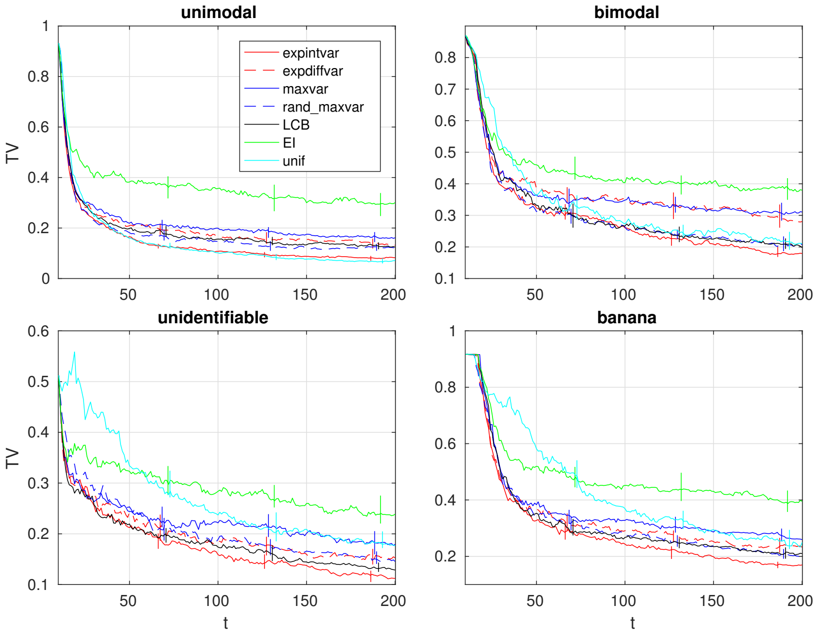

Figure 4: Median of the TV distance between the estimated ABC posterior and the corresponding exact ABC posterior over 100 experiments. Vertical lines show the 95% confidence interval of the median computed using the bootstrap.

the true posterior is $\mathcal{N}(\boldsymbol{\theta}\,|\,\mathbf{a}^\star, \mathbf{B}^\star)$ truncated to Θ, where $\mathbf{a}^\star = \mathbf{B}^\star(\mathbf{B}^{-1}\mathbf{a}+n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_{obs})$, $\mathbf{B}^\star = (\mathbf{B}^{-1}+n\boldsymbol{\Sigma}^{-1})^{-1}$ and $\bar{\mathbf{x}}_{obs} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i$ is the sample mean. As discrepancy, we use the Mahalanobis distance $\Delta_{\boldsymbol{\theta}} = ((\bar{\mathbf{x}}_{obs}-\bar{\mathbf{x}}_{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_{obs}-\bar{\mathbf{x}}_{\boldsymbol{\theta}}))^{1/2}$. The true posterior is used for comparisons in all the following experiments with the Gaussian model.

4.2.1 Strength of the prior

In the first experiment we set $p = 2$, $n = 5$, $\boldsymbol{\Sigma}_{ii} = 1$, and $\boldsymbol{\Sigma}_{ij} = 0.5$ for $i \neq j$. The initial training set size is $t_0 = 10$ and the threshold is fixed to $\varepsilon = 0.1$. Integration and sampling for expintvar and rand_maxvar are done as in Section 4.1. The true data mean of $\boldsymbol{\theta}$ is $[2, 2]^T$. The mean of the (truncated) Gaussian prior is $\mathbf{a} = [5, 5]^T$ and the covariance matrix is $\mathbf{B} = b^2\mathbf{I}$. We vary b, allowing us to study the impact of prior strength relative to the likelihood. Figure 5 shows the results, and we see that the proposed acquisition rules perform consistently well regardless the strength of prior, and focus the evaluations on the posterior modal region. On the other hand, LCB samples where the discrepancies are small, i.e. in areas of high likelihood, leading to sub-optimal posterior estimation whenever the prior is also informative. Comparing Figures 5a and 5b shows that using the expintvar strategy also avoids unnecessary evaluations on the boundary, which is often undesired also in the Bayesian optimisation methods, see Siivola et al. [2017] for a discussion. Curiously, the uniform sampling (unif rule) works well when prior information is strong.

4.2.2 High-dimensional test cases

Next we investigate the effect of the dimension p of the parameter space. The settings are as before, except that now we use uniform priors supported on $\Theta = [0, 8]^p$ and the threshold is set adaptively to the 0.01th quantile as described in Section 4.1. Further, $n = 15$, and the initial training set sizes are $t_0 = 20$ (3d) and $t_0 = 30$ (6d and 10d). Adaptive MCMC (with multiple chains) is used to sample from the model-based ABC posterior estimates required in the line 17 of Algorithm 1 and, in the case of expintvar and rand_maxvar, from the probability density $\pi_q(\boldsymbol{\theta})$. For expintvar we use $s = 500$ importance samples in 3d and $s = 200$ in

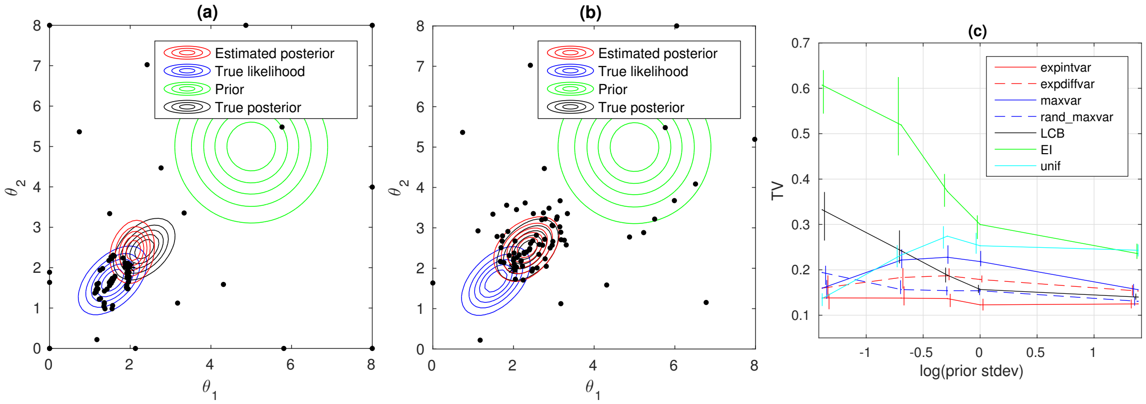

Figure 5: Acquired training data locations (black dots) for (a) LCB, (b) expintvar after 70 acquisitions. As discussed in Gutmann and Corander [2016], the LCB strategy ignores prior information which here leads to suboptimal selection of evaluation locations. (c) Median TV between the estimated ABC posterior and the corresponding true posterior as a function of the standard deviation (stdev) of the Gaussian prior over 100 experiments and after 200 evaluations (small stdev corresponds to strong prior information).

6d and 10d. Unlike in the other test problems, the TV distance measures here the average of TVs between all the marginal densities.

Figure 6 shows the results. With $p \leq 6$, the rand_maxvar is the most accurate and slightly better than expintvar strategy. However, in 10d it suffers from instability in MCMC convergence. Detailed examination shows that the method often produces multimodal posterior estimates which makes the sampling difficult. Such densities are likely a result of the random acquisitions. Namely, even if the uncertainty is high in some region, it can happen that no evaluations occur there during the available iterations, due to the randomness and the curse of dimensionality. EI also tends to produce multimodal difficult-to-sample posterior estimates but similar issues were only rarely observed with other strategies. The results suggest that in high dimensions the strategies that select the acquisition locations deterministically should be preferred over the stochastic ones.

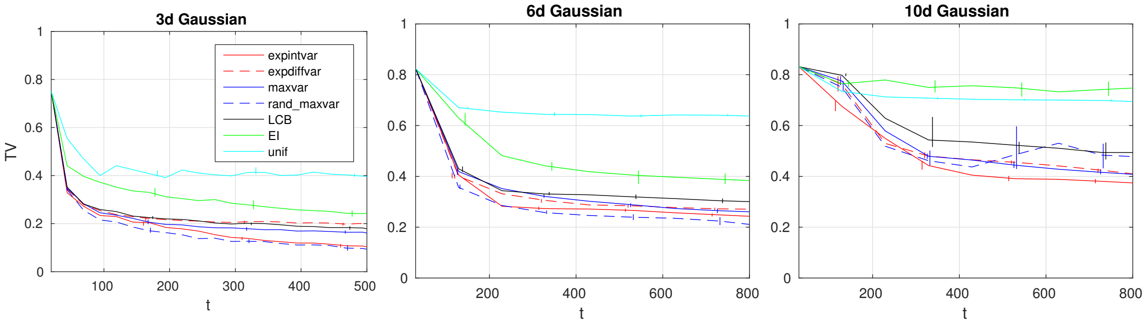

Figure 6: Median of the average marginal TVs between the estimated ABC posterior and the corresponding true posterior over 100 experiments in the 3d, 6d and 10d Gaussian toy simulation model.

4.3 Realistic simulation models

We consider the Lotka-Volterra model and a model of bacterial infections in day care centers to illustrate the proposed acquisition methods in practical modelling situations.

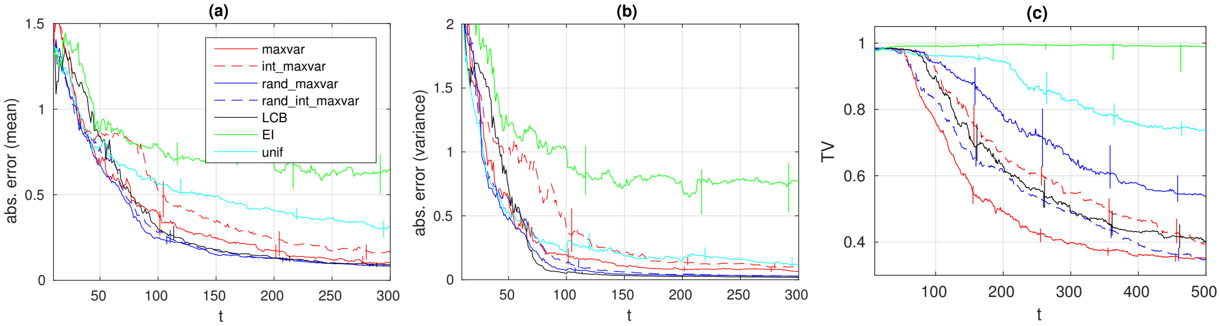

Figure 7: Median of the mean absolute error in the (a) ABC posterior mean, and (b) ABC posterior variance as compared to the true posterior over 100 experiments in the Lotka-Volterra model. Panel (c) shows the TV distance between the estimated ABC posterior and the true posterior.

4.3.1 Lotka-Volterra model

The Lotka-Volterra (LV) model [Toni et al., 2009] is described by differential equations $x_1'(t) = \theta_1 x_1(t) - x_1(t)x_2(t)$ and $x_2'(t) = \theta_2 x_1(t)x_2(t) - x_2(t)$, where $x_1(t)$ and $x_2(t)$ describe the evolution of prey and predator populations as a function of time t, respectively, and $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is the unknown parameter to be estimated. We use a similar experiment design as in Toni et al. [2009] but with discrepancy $\Delta_{\boldsymbol{\theta}} = \log \sum_{ij}(x_j^{\text{obs}}(t_i) - x_j^{\text{mod}}(t_i, \boldsymbol{\theta}))^2$, where $x_j^{\text{obs}}(t_i)$ for $j \in \{1, 2\}$, $i \in \{1, \ldots, 8\}$ denote the noisy observations at times t_i, and $x_j^{\text{mod}}(t_i, \boldsymbol{\theta})$ are the corresponding predictions. Up to the log transformation, this is the same as used by Toni et al. [2009]. We also experimented with another discrepancy, where the squared differences were replaced by absolute differences; however, the results were similar. In comparisons we use the uniform prior with support on $[0, 5]^2$, and the reference is the exact posterior distribution that can be computed analytically. We set $t_0 = 10$. The threshold is set to match the smallest observed discrepancy realisations and the integration and sampling required by expintvar and rand_maxvar are done as in Section 4.1.

The results are presented in Figure 7. We see that the expintvar strategy produces the best posterior mean estimates (Figure 7a), while the best posterior variance estimates are obtained by LCB and rand_maxvar (Figure 7b). However, the expintvar strategy clearly produces the most accurate posterior approximations in terms of TV distance followed by rand_maxvar and the LCB strategy (Figure 7c).

4.3.2 Bacterial infections model

Finally, we show the promise of our method using a simulation model that describes transmission dynamics of bacterial infections in day care centers. The model has three parameters: an internal infection parameter $\beta \in [0, 11]$, an external infection parameter $\Lambda \in [0, 2]$ and a co-infection parameter $\theta \in [0, 1]$. Full details of the model and data are described in Numminen et al. [2013]. The true posterior is not available and thus an ABC posterior computed using PMC-ABC algorithm, which required over two million simulations, is used as the reference distribution [Numminen et al., 2013]. We use the same experimental setup and discrepancy as Gutmann and Corander [2016], who used the model to illustrate their approach. Specifically, the initial training data size is $t_0 = 20$ and the uniform prior is used. Adaptive MCMC is again used to sample from the model-based posterior estimates and from the probability density $\pi_q(\boldsymbol{\theta})$. For expintvar we use $s = 500$ importance samples.

Figure 8 shows the results. Unlike in the other test cases, expintvar and rand_maxvar tend to produce slightly wider credible intervals for the marginal ABC posterior distributions than the other methods. Similarly, Gutmann and Corander [2016] obtained conservative estimates of these credible intervals with their stochastic variant of the LCB acquisition rule. To explain this, we investigated the GP modelling assumptions in more detail. Running a high number of additional bacterial model simulations indicates that the discrepancy is well approximated with a Gaussian in the modal area. On the other hand, the variance of

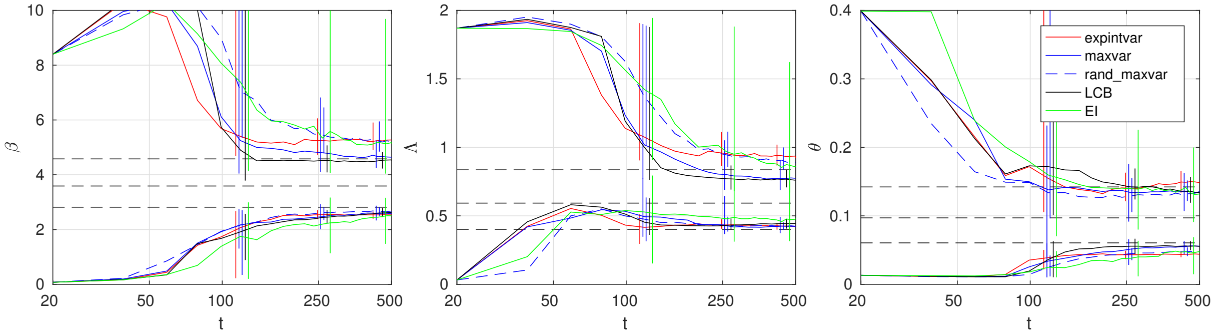

Figure 8: Comparison of the 95% credible interval estimates in the bacterial model. The black dashed lines show the corresponding credible interval estimates (computed using over 2 million model simulations with ABC-PMC algorithm) by Numminen et al. [2013] and the vertical lines show the 75% interval of the realisations over 100 experiments. The x-axis shows the iterations t on the log-scale.

the discrepancy, represented by the noise parameter σ_n^2 in the GP model, is not exactly constant as assumed in the GP but grows towards the tail areas. This explains why the expintvar and rand_maxvar strategies, that tend to include more evaluations in the tails than the other methods, have a tendency to over-estimate the value of σ_n^2. This further causes slightly overestimated credible intervals for the marginal ABC posterior distribution, as illustrated by Järvenpää et al. [2017]. We can also see that the expintvar strategy is more stable than the other methods in the sense that its results are more consistent over the 100 different realisations of the initial training data sets and simulator outputs than those of the other methods.

The maxvar and (deterministic) LCB produce credible interval estimates that are overall closest to the ABC posterior computed in Numminen et al. [2013]. However, the credible region for the parameter Λ is often underestimated possibly due to excess exploitation producing too small variance estimates σ_n^2. Using input-dependent GP model as in Järvenpää et al. [2017] would likely improve the approximation quality but we do not investigate this possibility here. While the EI strategy appears to work well on average in this example, it actually has high variability and occasionally produces too narrow posterior estimates and thus performs poorly overall. Comparing our posterior approximations using both the proposed acquisition methods as well as the (deterministic) LCB strategy to those in Gutmann and Corander [2016] shows that our estimates, both expintvar and (deterministic) LCB, are more accurate than the experiment reported in their paper.

In summary, despite some violations of the GP model assumptions, we were able to obtain posterior estimates that were very similar to those presented in Numminen et al. [2013] with only a fraction of simulations (500 vs 2,000,000), and without a need to use a computer cluster. We also showed that the proposed methods, especially expintvar, work consistently over different simulation model realisations, which is important with any realistic model where extensive running times may prohibit proper assessment of stochastic variability.

5 Discussion

In this section we offer guidelines for potential users and discuss some additional details of our algorithms. While the developed methods worked well, it may not be clear for an end user which method to use in practice. First of all, if the likelihood can be evaluated, there is usually no reason to consider ABC. Furthermore, if this is not the case but the simulation is fast, e.g. less than a second, standard ABC techniques may suffice. If the simulation is slow then the techniques in this paper become useful. Our technique with the expintvar strategy has a sound Bayesian decision theoretic basis and it performed the best overall, producing consistent approximation quality in different scenarios. We thus recommend this strategy. However, expintvar required up to 1 minute of computation time for selecting the next evaluation location in our 6d Gaussian test problem while this optimisation step took up to 4s for UCB and EI, and at most 8s for maxvar. The corresponding time for rand_maxvar was 30s. In our 2d Gaussian case, the expintvar strategy required at most 10s and all the other methods less than a second. These values are, however, descriptive since the computational time depends on various settings and the amount of training data. All these times are in any case negligible compared to the run time of many realistic simulation models which can be hours or days. In the supplementary material we provide further analysis of computation time using Big O notation and we further consider an alternative approach where a non-uniform acceptance threshold is used which allows for slightly faster computations. However, in high dimensions, when $p \gtrsim 10$, we recommend maxvar because we expect the estimated ABC posterior uncertainty to be inflated then anyway.

While not designed for ABC, the LCB criterion still worked surprisingly well overall and offers another reasonable choice in practice. However, standard LCB is not suitable if the prior is informative and its accuracy deteriorated in the high-dimensional experiment. EI (and PI) performed poorly and we see little reason to use them unless the goal is to learn only the maximiser of the discrepancy. Furthermore, unlike the standard BO strategies such as LCB [with the exception of Shahriari et al., 2016], the developed acquisition rules do not necessarily require a box-constrained domain. Namely, if the prior is a bounded and proper density (to ensure that the acquisition functions are bounded and the ABC posterior defines a valid pdf), the requirement of the bounded support can be relaxed.

In addition to the acquisition strategy, the posterior approximation quality also depends on the GP model and some other choices. For example, the proposed acquisition strategies and the final posterior estimate depend on the threshold. We either assumed this value to be known or used a heuristic approach and set the threshold to the 0.01th quantile of the realised discrepancies. We also considered other choices but this approach worked well. In principle, the strategy for selecting the threshold could also vary during the iterations. While some ABC methods bypass selecting the threshold, they may not be applicable when the budget for simulations is very small. Our framework is also applicable for model-based ABC methods that do not require the threshold.

We used the zero mean GP model in our experiments. While Wilkinson [2014], Drovandi et al. [2015], Gutmann and Corander [2016] considered certain parametric mean functions which might help focusing the simulations on the modal area, our choice is a safe option. Namely, if there is a large region containing no simulations, the discrepancy tends to zero there. Thus, the uncertainty will be high in the region, attracting future simulations. Futhermore, even though in some studies [e.g. Snoek et al., 2012], the Matern kernel has been empirically shown to perform slightly better than the squared exponential, we expect the discrepancy in many ABC modelling applications to be smooth and, consequently, used the squared exponential kernel.

To demonstrate the framework, we chose to model the discrepancy with a GP. However, this approach may not be optimal if the Gaussianity assumption is violated [Gutmann and Corander, 2016, Järvenpää et al., 2017]. Non-Gaussian measurement models can be used but the acquisition criteria in Equation (10) or (19) may become costly to evaluate. One could also model the log-likelihood directly with a GP and select the evaluations at the maximiser of the variance of the likelihood function [Kandasamy et al., 2015], which is similar to our maxvar criterion. One could also model the individual summaries with independent GPs as in Jabot et al. [2014], Meeds and Welling [2014]. In both of these cases the evaluation locations could be chosen based on the ideas in Section 2.

An alternative to the proposed stochastic acquisition rule is to sample new evaluation locations from the current ABC posterior estimate. This approach seems to work well in some scenarios but no systematic comparison was done. However, the posterior estimate could get stuck to a poor region due to an "unlucky" discrepancy realisation, after which new evaluations would be focused on this seemingly good region and the method has little chance to escape from the local optimum.

While our approach is designed for fitting costly simulation models, we note that it can be useful even when the simulation model is relatively cheap to run. For example, for a developer of a simulation model, it may be useful to first obtain rough estimates for the model parameters before using costly computations for final and accurate results. Our derivations are also applicable for estimating the tail probabilities of Gaussian processes over some parameter domain. An approach similar to ours has also been applied to the problem of estimating an excursion set by Chevalier et al. [2014]. However, the objective of their work is to identify the set of points that are below a fixed threshold instead of learning the corresponding tail probability under GP surrogate model assumptions.

6 Conclusions

We considered the challenging problem of performing Bayesian inference when the likelihood function cannot be evaluated and simulating data from the statistical model is costly. We proposed to use another instance of Bayesian inference to quantify the uncertainty in the approximate posterior due to the limited budget of simulations and to design the simulations to minimise the expected uncertainty in the posterior approximation. Such computations can be costly themselves but we chose a loss function that measures such uncertainty and allowed developing a tractable and practical algorithm for selecting the next evaluation location to run the simulation model. Notably, compared to many realistic simulation models, the run time of which can be hours or days, the computational overhead introduced by our approach is negligible. Experiments demonstrated that the proposed method performs better than or similarly to the commonly used Bayesian optimisation strategies and other, more heuristic approaches obtained as a by-product of our derivations. Our approach also takes prior density into account, does not require box-constrained parameter spaces and has a sound decision-theoretic basis that extends to other ABC surrogate modelling scenarios beyond those considered in this article.

As future work, other surrogate models and principled approaches for selecting the threshold could be investigated. We here focused on single acquisitions but our approach in principle extends to batch acquisitions as well. This enables parallelised inference, which is particularly useful when computationally very costly simulation models need to be fitted.

Acknowledgements

This work was funded by the Academy of Finland (grants no. 286607 and 294015 to PM). We acknowledge the computational resources provided by Aalto Science-IT project.

References

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002. 1, 2

M. A. Beaumont, J.-M. Cornuet, J.-M. Marin, and C. P. Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990, 2009. 2

H. Bijl, T. B. Schön, J. van Wingerden, and M. Verhaegen. A sequential Monte Carlo approach to Thompson sampling for Bayesian optimization. Available at `https://arxiv.org/abs/1604.00169`, 2016. 4

M. G. B. Blum. Approximate Bayesian Computation: a nonparametric perspective. *Journal of American Statistical Association*, 105(491):1178–1187, 2010. 2

M. G. B. Blum and O. François. Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, 20(1):63–73, 2010. 2

M. G. B. Blum, M. A. Nunes, D. Prangle, and S. A. Sisson. A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28(2):189–208, 2013. 2

C. Chevalier, J. Bect, D. Ginsbourger, E. Vazquez, V. Picheny, and Y. Richet. Fast Parallel Kriging-Based Stepwise Uncertainty Reduction With Application to the Identification of an Excursion Set. *Technometrics*, 56(4):455–465, 2014. 18

P. J. Diggle and R. J. Gratton. Monte Carlo Methods of Inference for Implicit Statistical Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):193–227, 1984. 1

C. C. Drovandi, M. T. Moores, and R. J. Boys. Accelerating pseudo-marginal MCMC using Gaussian processes. Available at `http://eprints.qut.edu.au/90973/`. Accessed 11-9-2017, 2015. 2, 18

Y. Fan, D. J. Nott, and S. A. Sisson. Approximate Bayesian computation via regression density estimation. *Stat*, 2(1):34–48, 2013. 2

P. Fearnhead and D. Prangle. Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 74(3):419–474, 2012. 2

J. González, M. Osborne, and N. D. Lawrence. GLASSES: Relieving The Myopia Of Bayesian Optimisation. In *Proceedings of the Nineteenth International Workshop on Artificial Intelligence and Statistics*, 2016. 4

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. 2, 3, 5, 6, 10, 12, 15, 16, 17, 18

M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander. Likelihood-free inference via classification. *Statistics and Computing*, March 2017. 2

F. Hartig, J. M. Calabrese, B. Reineking, T. Wiegand, and A. Huth. Statistical inference for stochastic simulation models–theory and application. *Ecology Letters*, 14(8):816–27, 2011. 2

P. Hennig and C. J. Schuler. Entropy Search for Information-Efficient Global Optimization. *Journal of Machine Learning Research*, 13(1999):1809–1837, 2012. 4

P. Hennig, M. A. Osborne, and M. Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 471 (2179):20150142, 2015. 4

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. 4, 12

F. Jabot, G. Lagarrigues, B. Courbaud, and N. Dumoulin. A comparison of emulation methods for Approximate Bayesian Computation. Available at `http://arxiv.org/abs/1412.7560`, 2014. 2, 18

M. Järvenpää, M. Gutmann, A. Vehtari, and P. Marttinen. Gaussian process modeling in approximate Bayesian computation to estimate horizontal gene transfer in bacteria. Available at `https://arxiv.org/abs/1610.06462`, 2017. 2, 17, 18

K. Kandasamy, J. Schneider, and B. Póczos. Bayesian active learning for posterior estimation. In *International Joint Conference on Artificial Intelligence*, pages 3605–3611, 2015. 2, 3, 9, 18

M. Lenormand, F. Jabot, and G. Deffuant. Adaptive approximate Bayesian computation for complex models. *Computational Statistics*, 28(6):2777–2796, 2013. 2

J. Lintusaari, M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander. Fundamentals and Recent Developments in Approximate Bayesian Computation. *Systematic biology*, 66(1):e66–e82, 2017. 2

J. Lintusaari, H. Vuollekoski, A. Kangasrääsiö, K. Skytén, M. Järvenpää, M. U. Gutmann, A. Vehtari, J. Corander, and S. Kaski. ELFI: Engine for Likelihood Free Inference. *Journal of Machine Learning Research*, 2018. Accepted for publication. 12

J. M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012. 2

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. 2

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. 1

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. 2, 18

I. Murray and R. P. Adams. Slice sampling covariance hyperparameters of latent Gaussian models. *Advances in Neural Information Processing Systems*, 2(1):9, 2010. 10

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. 1, 16, 17

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. 3

D. B. Owen. Tables for computing bivariate normal probabilities. *The Annals of Mathematical Statistics*, 27(4):1075–1090, 12 1956. 23

D. B. Owen. A table of normal integrals. *Communications in Statistics - Simulation and Computation*, 9 (4):389–419, 1980. 8

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. 2

M. Patefield and D. Tandy. Fast and accurate Calculation of Owen's T-Function. *Journal of Statistical Software*, 5(5):1–25, 2000. 6, 12

L. F. Price, C. C. Drovandi, A. Lee, and D. J. Nott. Bayesian synthetic likelihood. Journal of Computational and Graphical Statistics. (In Press), 2017. 2

C. E. Rasmussen. Gaussian Processes to Speed up Hybrid Monte Carlo for Expensive Bayesian Integrals. *Bayesian Statistics 7*, pages 651–659, 2003. 3

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006. 5, 22, 30

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, second edition, 2004. 8

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 71(2):319–392, 2009. 10

E. G. Ryan, C. C. Drovandi, J. M. Mcgree, and A. N. Pettitt. A Review of Modern Computational Algorithms for Bayesian Optimal Design. *International Statistical Review*, 84(1):128–154, 2016. 4

B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1), 2015. 12

B. Shahriari, A. Bouchard-Côté, and N. de Freitas. Unbounded Bayesian Optimization via Regularization. In *International Conference on Artificial Intelligence and Statistics*, 2016. 18

E. Siivola, A. Vehtari, J. Vanhatalo, and J. González. Bayesian optimization with virtual derivative sign observations. Available at https://arxiv.org/abs/1704.00963, 2017. 14

S. A. Sisson, Y. Fan, and M. M. Tanaka. Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 104(6):1760–5, 2007. 2

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. 12, 18

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 International Conference on Machine Learning*, pages 1015–1022, 2010. 7

T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society, Interface*, 6(31):187–202, 2009. 2, 16

B. M. Turner and T. Van Zandt. A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, 56(2):69–85, 2012. 2

J. Vanhatalo, V. Pietiläinen, and A. Vehtari. Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in Medicine*, 29(15):1580–1607, 2010. 10

J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, and A. Vehtari. GPstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research*, 14:1175–1179, 2013. 12

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. 4, 12

Z. Wang, B. Zhou, and S. Jegelka. Optimization as Estimation with Gaussian Processes in Bandit Settings. In *In proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2016. 4

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. 2, 29

R. D. Wilkinson. Accelerating ABC methods using Gaussian processes. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, 2014. 2, 3, 18

S. N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466:1102–1104, 2010. 2

A Proofs

We provide the derivations of Lemma 3.1 and Proposition 3.2 below.

Proof of Lemma 3.1. Using the law of the unconscious statistician, we can write

$$\mathbb{E}(p_a(\boldsymbol{\theta})) = \int_{-\infty}^{\infty} \Phi\left(\frac{\varepsilon - f}{\sigma_n}\right) \mathcal{N}(f \,|\, m_{1:t}(\boldsymbol{\theta}), v_{1:t}^2(\boldsymbol{\theta}))\,\mathrm{d}f. \tag{28}$$

Now, using the fact $\Phi(x) = 1 - \Phi(-x)$, a standard result for Gaussian moments derived in Rasmussen and Williams [2006, p. 74] and $\mathbb{E}(\pi(\boldsymbol{\theta})p_a(\boldsymbol{\theta})) = \pi(\boldsymbol{\theta})\mathbb{E}(p_a(\boldsymbol{\theta}))$, (which holds because the prior density at $\boldsymbol{\theta}$ is a scalar) one obtains Equation (9).

A formula for the variance of $\tilde{\pi}_{\varepsilon}^{\mathrm{ABC}}(\boldsymbol{\theta})$ can be obtained similarly. First we see that

$$\begin{aligned} \mathbb{V}(p_a(\boldsymbol{\theta})) &= \mathbb{E}(p_a(\boldsymbol{\theta})^2) - [\mathbb{E}(p_a(\boldsymbol{\theta}))]^2 & (29) \\ &= \int_{-\infty}^{\infty} \Phi^2\left(\frac{\varepsilon - f}{\sigma_n}\right) \mathcal{N}(f \,|\, m_{1:t}(\boldsymbol{\theta}), v_{1:t}^2(\boldsymbol{\theta}))\,\mathrm{d}f - [\mathbb{E}(p_a(\boldsymbol{\theta}))]^2. & (30) \end{aligned}$$

The first term of Equation (30) can be further written as

$$\begin{aligned}&\int_{-\infty}^{\infty}\int_{-\infty}^{\varepsilon}\mathcal{N}(z_1\,|\,f,\sigma_n^2)\,\mathrm{d}z_1\int_{-\infty}^{\varepsilon}\mathcal{N}(z_2\,|\,f,\sigma_n^2)\,\mathrm{d}z_2\,\mathcal{N}(f\,|\,m_{1:t}(\boldsymbol{\theta}),v_{1:t}^2(\boldsymbol{\theta}))\,\mathrm{d}f\\&=\int_{-\infty}^{\varepsilon}\int_{-\infty}^{\varepsilon}\int_{-\infty}^{\infty}\mathcal{N}(f\,|\,z_1,\sigma_n^2)\mathcal{N}(f\,|\,z_2,\sigma_n^2)\mathcal{N}(f\,|\,m_{1:t}(\boldsymbol{\theta}),v_{1:t}^2(\boldsymbol{\theta}))\,\mathrm{d}f\,\mathrm{d}z_1\,\mathrm{d}z_2\end{aligned}\tag{31}$$

where we have used Fubini-Tonelli theorem to change the order of integration. The integrand can be now written as an unnormalised Gaussian pdf for f and, after integrating over f and some further calculations, the resulting formula can be recognised as

$$\Phi_2(\varepsilon\mathbb{1}\,|\,m_{1:t}(\boldsymbol{\theta})\mathbb{1},V_{1:t}(\boldsymbol{\theta})),\tag{32}$$

where $\mathbb{1}=[1,1]^T$ and the function Φ_2 denotes the bivariate Gaussian cdf with mean $m_{1:t}(\boldsymbol{\theta})\mathbb{1}$ and covariance matrix

$$V_{1:t}(\boldsymbol{\theta})=\begin{bmatrix}\sigma_n^2+v_{1:t}^2(\boldsymbol{\theta}) & v_{1:t}^2(\boldsymbol{\theta})\\ v_{1:t}^2(\boldsymbol{\theta}) & \sigma_n^2+v_{1:t}^2(\boldsymbol{\theta})\end{bmatrix},\tag{33}$$

which is clearly symmetric and positive definite since $\sigma_n^2>0$.

Denoting the correlation coefficient $\rho(\boldsymbol{\theta})=v_{1:t}^2(\boldsymbol{\theta})/(\sigma_n^2+v_{1:t}^2(\boldsymbol{\theta}))$, we see that Equation (32) can be further written as

$$\begin{aligned}&\Phi_2\left(\frac{\varepsilon-m_{1:t}(\boldsymbol{\theta})}{\sqrt{\sigma_n^2+v_{1:t}^2(\boldsymbol{\theta})}}\mathbb{1}\,\middle|\,\mathbf{0},\begin{bmatrix}1 & \rho(\boldsymbol{\theta})\\ \rho(\boldsymbol{\theta}) & 1\end{bmatrix}\right)\\&=\Phi\left(\frac{\varepsilon-m_{1:t}(\boldsymbol{\theta})}{\sqrt{\sigma_n^2+v_{1:t}^2(\boldsymbol{\theta})}}\right)-2T\left(\frac{\varepsilon-m_{1:t}(\boldsymbol{\theta})}{\sqrt{\sigma_n^2+v_{1:t}^2(\boldsymbol{\theta})}},\frac{1-\rho(\boldsymbol{\theta})}{\sqrt{1-\rho^2(\boldsymbol{\theta})}}\right)\end{aligned}\tag{34}$$

where $\mathbf{0}=[0,0]^T$. The equality follows from the connection between bivariate Gaussian cdf and Owen's t-function, see Owen [1956] for this fact and its proof. Also, simple calculations show that

$$\frac{1-\rho(\boldsymbol{\theta})}{\sqrt{1-\rho^2(\boldsymbol{\theta})}}=\frac{\sigma_n}{\sqrt{\sigma_n^2+2v_{1:t}^2(\boldsymbol{\theta})}}.\tag{35}$$

The rest of the result now follows from the derivations above, Equation (9) and the fact $\mathbb{V}(\pi(\boldsymbol{\theta})p_a(\boldsymbol{\theta}))=\pi^2(\boldsymbol{\theta})\mathbb{V}(p_a(\boldsymbol{\theta}))$. □

Proof of Proposition 3.2. We first derive the probability densities for the GP mean and covariance function after a future observation is obtained. These quantities are random variables because the new discrepancy realisation Δ^* is unknown. Given the training data $D_{1:t}=\{(\Delta_i,\boldsymbol{\theta}_i)\}_{i=1}^t$, the mean function $m_{1:t+1}(\boldsymbol{\theta})$ can be written as

$$m_{1:t+1}(\boldsymbol{\theta})=[k_{\boldsymbol{\theta},\boldsymbol{\theta}_{1:t}},k_{\boldsymbol{\theta},\boldsymbol{\theta}^*}]\begin{bmatrix}K(\boldsymbol{\theta}_{1:t}) & k(\boldsymbol{\theta}_{1:t},\boldsymbol{\theta}^*)\\ k(\boldsymbol{\theta}^*,\boldsymbol{\theta}_{1:t}) & K(\boldsymbol{\theta}^*)\end{bmatrix}^{-1}\begin{bmatrix}\Delta_{1:t}\\ \Delta^*\end{bmatrix}\tag{36}$$

$$\begin{aligned}&=k(\boldsymbol{\theta},\boldsymbol{\theta}_{1:t})K^{-1}(\boldsymbol{\theta}_{1:t})\Delta_{1:t}+[k(\boldsymbol{\theta},\boldsymbol{\theta}^*)-k(\boldsymbol{\theta},\boldsymbol{\theta}_{1:t})K^{-1}(\boldsymbol{\theta}_{1:t})k(\boldsymbol{\theta}_{1:t},\boldsymbol{\theta}^*)]\\&\quad\cdot[K(\boldsymbol{\theta}^*)-k(\boldsymbol{\theta}^*,\boldsymbol{\theta}_{1:t})K^{-1}(\boldsymbol{\theta}_{1:t})k(\boldsymbol{\theta}_{1:t},\boldsymbol{\theta}^*)]^{-1}[\Delta^*-k(\boldsymbol{\theta}^*,\boldsymbol{\theta}_{1:t})K^{-1}(\boldsymbol{\theta}_{1:t})\Delta_{1:t}]\\&=m_{1:t}(\boldsymbol{\theta})+\mathrm{cov}_{1:t}(\boldsymbol{\theta},\boldsymbol{\theta}^*)(v_{1:t}^2(\boldsymbol{\theta}^*)+\sigma_n^2)^{-1}(\Delta^*-m_{1:t}(\boldsymbol{\theta}^*)),\end{aligned}\tag{37}$$

where we have used a well-known formula for blockwise inversion. According to the current GP model, the unknown future discrepancy Δ^* follows a Gaussian density i.e. $\Delta^*\,|\,\boldsymbol{\theta}^*,D_{1:t}\sim\mathcal{N}(m_{1:t}(\boldsymbol{\theta}^*),v_{1:t}^2(\boldsymbol{\theta}^*)+\sigma_n^2)$ so that

$$m_{1:t+1}(\boldsymbol{\theta})\sim\mathcal{N}(m_{1:t}(\boldsymbol{\theta}),\tau_{1:t}^2(\boldsymbol{\theta},\boldsymbol{\theta}^*)),\tag{38}$$

where we have denoted $\tau^2_{1:t}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \mathrm{cov}^2_{1:t}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)/(\sigma_n^2 + v^2_{1:t}(\boldsymbol{\theta}^*))$.

Similar computations as for the mean function show that

$$v^2_{1:t+1}(\boldsymbol{\theta}) = v^2_{1:t}(\boldsymbol{\theta}) - \tau^2_{1:t}(\boldsymbol{\theta}, \boldsymbol{\theta}^*). \tag{39}$$

This formula further shows that the change in the GP variance is deterministic and depends only on the chosen evaluation location $\boldsymbol{\theta}^*$.

Changing the order of integration using Tonelli theorem we obtain

$$\begin{aligned}
L_{1:t}(\boldsymbol{\theta}^*) &= \mathbb{E}_{\Delta^* \,|\, \boldsymbol{\theta}^*, D_{1:t}} \int_\Theta \pi^2(\boldsymbol{\theta}) \mathbb{V}(p_a(\boldsymbol{\theta}) \,|\, \Delta^*, \boldsymbol{\theta}^*, D_{1:t}) \,\mathrm{d}\boldsymbol{\theta} && (40)\\
&= \int_{-\infty}^{\infty} \int_\Theta \pi^2(\boldsymbol{\theta}) \mathbb{V}(p_a(\boldsymbol{\theta}) \,|\, \Delta^*, \boldsymbol{\theta}^*, D_{1:t}) \,\mathrm{d}\boldsymbol{\theta} \,\mathcal{N}(\Delta^* \,|\, m_{1:t}(\boldsymbol{\theta}^*), \sigma_n^2 + v^2_{1:t}(\boldsymbol{\theta}^*)) \,\mathrm{d}\Delta^* && (41)\\
&= \int_\Theta \pi^2(\boldsymbol{\theta}) \underbrace{\int_{-\infty}^{\infty} \mathbb{V}(p_a(\boldsymbol{\theta}) \,|\, \Delta^*, \boldsymbol{\theta}^*, D_{1:t}) \mathcal{N}(\Delta^* \,|\, m_{1:t}(\boldsymbol{\theta}^*), \sigma_n^2 + v^2_{1:t}(\boldsymbol{\theta}^*)) \,\mathrm{d}\Delta^*}_{=: g^2_{1:t+1}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)} \,\mathrm{d}\boldsymbol{\theta} && (42)\\
&= \int_\Theta \pi^2(\boldsymbol{\theta}) g^2_{1:t+1}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \,\mathrm{d}\boldsymbol{\theta}, && (43)
\end{aligned}$$

To simplify the notation in the following formulas, we define $m_0 := m_{1:t}(\boldsymbol{\theta})$, $m_1 := m_{1:t+1}(\boldsymbol{\theta})$ and similarly for the variance terms v_0^2 and v_1^2. We also define $\tau_0^2(\boldsymbol{\theta}^*) := \tau^2_{1:t}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$. Using these conventions and Equations (38) and (39) we see that

$$\begin{aligned}
&g^2_{1:t+1}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\\
&= \int_0^\infty \int_{-\infty}^{\infty} \mathbb{V}(p_a(\boldsymbol{\theta}, m_1, v_1^2)) \mathcal{N}(m_1 \,|\, m_0, \tau_0^2(\boldsymbol{\theta}^*)) \delta(v_1^2 - v_0^2 + \tau_0^2(\boldsymbol{\theta}^*)) \,\mathrm{d}m_1 \,\mathrm{d}v_1^2 && (44)\\
&= \int_{-\infty}^{\infty} \Phi\left(\frac{\varepsilon - m_1}{\sqrt{\sigma_n^2 + v_1^2 - \tau_0^2(\boldsymbol{\theta}^*)}}\right) - \Phi^2\left(\frac{\varepsilon - m_1}{\sqrt{\sigma_n^2 + v_1^2 - \tau_0^2(\boldsymbol{\theta}^*)}}\right) \mathcal{N}(m_1 \,|\, m_0, \tau_0^2(\boldsymbol{\theta}^*)) \,\mathrm{d}m_1 && (45)\\
&\quad - \frac{1}{\pi} \int_{-\infty}^{\infty} \int_0^{\frac{\sigma_n}{\sqrt{\sigma_n^2 + 2v_1^2 - 2\tau_0^2(\boldsymbol{\theta}^*)}}} \frac{e^{-\frac{1}{2}\left(\frac{\varepsilon - m_1}{\sqrt{\sigma_n^2 + v_1^2 - \tau_0^2(\boldsymbol{\theta}^*)}}\right)^2 (1+x^2)}}{1 + x^2} \,\mathrm{d}x \,\mathcal{N}(m_1 \,|\, m_0, \tau_0^2(\boldsymbol{\theta}^*)) \,\mathrm{d}m_1 && (46)
\end{aligned}$$

The integral of Equation (45) can be computed using the equations in the proof of Lemma 3.1 and after some straightforward computations one obtains

$$2T\left(\frac{\varepsilon - m_0}{\sqrt{\sigma_n^2 + v_0^2}}, \sqrt{\frac{\sigma_n^2 + v_0^2(\boldsymbol{\theta}) - \tau_0^2(\boldsymbol{\theta}^*)}{\sigma_n^2 + v_0^2(\boldsymbol{\theta}) + \tau_0^2(\boldsymbol{\theta}^*)}}\right). \tag{47}$$

To simplify Equation (46), we use again the Fubini-Tonelli theorem to change the order of integration and some straightforward (but tedious) manipulations to obtain

$$\begin{aligned}
&- \frac{1}{\pi} \int_0^{\frac{\sigma_n}{\sqrt{\sigma_n^2 + 2v_1^2 - 2\tau_0^2(\boldsymbol{\theta}^*)}}} \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}\frac{(\varepsilon - m_1)^2(1+x^2)}{\sigma_n^2 + v_1^2 - \tau_0^2(\boldsymbol{\theta}^*)}}}{1 + x^2} \mathcal{N}(m_1 \,|\, m_0, \tau_0^2(\boldsymbol{\theta}^*)) \,\mathrm{d}m_1 \,\mathrm{d}x && (48)\\
&= -\frac{1}{\pi} \int_0^{\frac{\sigma_n}{\sqrt{\sigma_n^2 + 2v_1^2 - 2\tau_0^2(\boldsymbol{\theta}^*)}}} \int_{-\infty}^{\infty} \frac{\sqrt{2\pi} c(x)}{1 + x^2} \mathcal{N}(m_1 \,|\, \varepsilon, c^2(x)) \mathcal{N}(m_1 \,|\, m_0, \tau_0^2(\boldsymbol{\theta}^*)) \,\mathrm{d}m_1 \,\mathrm{d}x && (49)\\
&= -\frac{1}{\pi} \int_0^{\frac{\sigma_n}{\sqrt{\sigma_n^2 + 2v_1^2 - 2\tau_0^2(\boldsymbol{\theta}^*)}}} \frac{1}{1 + x^2} \sqrt{\frac{\sigma_n^2 + v_0^2 - \tau_0^2(\boldsymbol{\theta}^*)}{\sigma_n^2 + v_0^2 + x^2\tau_0^2(\boldsymbol{\theta}^*)}} e^{-\frac{(\varepsilon - m_1)^2(1+x^2)}{\sigma_n^2 + v_0^2 + x^2\tau_0^2(\boldsymbol{\theta}^*)}} \,\mathrm{d}x && (50)
\end{aligned}$$

where we have defined $c(x) := (\sigma_n^2 + v_1^2 - \tau_0^2(\boldsymbol{\theta}^*))/(1+x^2)$ and used again the well-known product rule for Gaussian pdfs.

Next we define the transformation $\psi(z) := z\sqrt{\sigma_n^2 + v_0^2}/\sqrt{\sigma_n^2 + v_0^2 - \tau_0^2(\boldsymbol{\theta}^*)(1+z^2)}$. Some analysis shows that $\psi'(z) = \sqrt{\sigma_n^2 + v_0^2}\sqrt{\sigma_n^2 + v_0^2 - \tau_0^2(\boldsymbol{\theta}^*)}/(\sigma_n^2 + v_0^2 - \tau_0^2(\boldsymbol{\theta}^*)(1+z^2))^{3/2} > 0$ so that ψ is strictly increasing function, and that it maps the interval $[0, \sigma_n/\sqrt{\sigma_n^2 + 2v_0^2}]$ to $[0, \sigma_n/\sqrt{\sigma_n^2 + 2v_1^2 - 2\tau_0^2(\boldsymbol{\theta}^*)}]$. Substituting $x = \psi(z)$ to the integral in Equation (50) and some straightforward computations show that Equation (50) simplifies to

$$-\frac{1}{\pi}\int_0^{\frac{\sigma_n}{\sqrt{\sigma_n^2+2v_0^2}}} \frac{1}{1+z^2} e^{-\frac{(\varepsilon - m_0)^2(1+z^2)}{2(\sigma_n^2+v_0^2)}}\,\mathrm{d}z = -2T\left(\frac{\varepsilon - m_0}{\sqrt{\sigma_n^2+v_0^2}}, \frac{\sigma_n}{\sqrt{\sigma_n^2+2v_0^2}}\right). \tag{51}$$

The final result (19) now follows from the equations above. □

B Additional derivations and gradients

B.1 Additional derivations

We start by deriving the cdf for the random variable $p_a(\boldsymbol{\theta})$ when the uncertainty in the GP hyperparameters $\boldsymbol{\phi}$ is taken into account. The corresponding results for $\tilde{\pi}_\varepsilon^{\mathrm{ABC}}(\boldsymbol{\theta})$ follow by suitable scaling with the prior pdf $\pi(\boldsymbol{\theta})$. The cdf of $p_a(\boldsymbol{\theta})$ evaluated at $z \in (0,1)$ is

$$\begin{aligned} F_{p_a(\boldsymbol{\theta})}(z) &= \mathbb{P}(p_a(\boldsymbol{\theta}) \leq z) = \int \mathbb{P}(p_a(\boldsymbol{\theta}) \leq z \,|\, \boldsymbol{\phi})\pi(\boldsymbol{\phi})\,\mathrm{d}\boldsymbol{\phi} = \int \mathbb{P}\left(\Phi\left(\frac{\varepsilon - f(\boldsymbol{\theta})}{\sigma_n}\right) \leq z \,\middle|\, \boldsymbol{\phi}\right)\pi(\boldsymbol{\phi})\,\mathrm{d}\boldsymbol{\phi} \\ &= \int \mathbb{P}\left(f(\boldsymbol{\theta})) \geq \varepsilon - \sigma_n\Phi^{-1}(z) \,\middle|\, \boldsymbol{\phi}\right)\pi(\boldsymbol{\phi})\,\mathrm{d}\boldsymbol{\phi} = \int \Phi\left(\frac{\sigma_n\Phi^{-1}(z) + m_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi}) - \varepsilon}{v_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})}\right)\pi(\boldsymbol{\phi})\,\mathrm{d}\boldsymbol{\phi}, \end{aligned} \tag{52}$$

it is zero if $z \leq 0$, and one if $z \geq 1$. In above, the density $\pi(\boldsymbol{\phi})$ describes our knowledge about the GP hyperparameters $\boldsymbol{\phi}$ given the training data $D_{1:t}$ (conditioning on data is ignored to simplify notation). The integral in the equations above is taken over the domain of the GP hyperparameters $\boldsymbol{\phi}$. The integral can be approximated using e.g. CCD as discussed in the main text. If the hyperparameters are fixed and $\pi_{\boldsymbol{\phi}}(\boldsymbol{\phi})$ is replaced with a point mass, one obtains Equation (14).

A formula for the pdf can be obtained by differentiating the cdf. We first realise that $z = \Phi(\Phi^{-1}(z))$ which implies $1 = \frac{\mathrm{d}z}{\mathrm{d}z} = \frac{\mathrm{d}}{\mathrm{d}z}\Phi(\Phi^{-1}(z)) = \Phi'(\Phi^{-1}(z))(\Phi^{-1})'(z)$ for $z \in (0,1)$. This fact is used to further show that

$$(\Phi^{-1})'(z) = \frac{1}{\Phi'(\Phi^{-1}(z))} = \frac{1}{\mathcal{N}(\Phi^{-1}(z)\,|\,0,1)} = \sqrt{2\pi}e^{(\Phi^{-1}(z))^2/2}. \tag{53}$$

Using the Equation (53) allows to compute

$$\pi_{p_a(\boldsymbol{\theta})\,|\,\boldsymbol{\phi}}(z) = \frac{\partial}{\partial z}F_{p_a(\boldsymbol{\theta})\,|\,\boldsymbol{\phi}}(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{(\sigma_n\Phi^{-1}(z)+m_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})-\varepsilon)^2}{2v_{1:t}^2(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})}}\frac{\partial}{\partial z}\frac{\sigma_n\Phi^{-1}(z) + m_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi}) - \varepsilon}{v_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})} \tag{54}$$

$$= \frac{\sigma_n}{v_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})}e^{\frac{(\Phi^{-1}(z))^2}{2} - \frac{(\sigma_n\Phi^{-1}(z)+m_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})-\varepsilon)^2}{2v_{1:t}^2(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})}} \tag{55}$$

$$= \begin{cases} \frac{\sigma_n}{v_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})}e^{\frac{(\varepsilon - m_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi}))^2}{2(\sigma_n^2 - v_{1:t}^2(\boldsymbol{\theta}\,|\,\boldsymbol{\phi}))}}e^{-\frac{\sigma_n^2 - v_{1:t}^2(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})}{2v_{1:t}^2(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})}\left(\Phi^{-1}(z) - \frac{(\varepsilon - m_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi}))\sigma_n}{\sigma_n^2 - v_{1:t}^2(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})}\right)^2}, & \text{if } \sigma_n \neq v_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi}), \\ e^{-\frac{(\varepsilon - m_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi}))^2}{2v_{1:t}^2(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})}}e^{\frac{\varepsilon - m_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})}{v_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})}\Phi^{-1}(z)}, & \text{if } \sigma_n = v_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi}), \end{cases} \tag{56}$$

for $z \in (0,1)$ and it is zero elsewhere. Finally, the pdf is obtained by marginalising the GP hyperparameters, that is

$$\pi_{p_a(\boldsymbol{\theta})}(z) = \int \pi_{p_a(\boldsymbol{\theta})\,|\,\boldsymbol{\phi}}(z)\pi_{\boldsymbol{\phi}}(\boldsymbol{\phi})\,\mathrm{d}\boldsymbol{\phi}. \tag{57}$$

The mean and variance for $\tilde{\pi}_{\varepsilon}^{\mathrm{ABC}}(\boldsymbol{\theta})$ were presented in Section 3.2 and the corresponding results for $p_a(\boldsymbol{\theta})$ follow by setting $\pi(\boldsymbol{\theta}) = 1$. The quantiles can be computed as in Equation (15). If the uncertainty in the GP hyperparameters is taken into account, then numerical root finding such as bisection search is required for inverting the cdf.

Inspecting the pdf given by Equation (56) shows that if $\sigma_n > v_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})$, then the mode of $p_a(\boldsymbol{\theta})\,|\,\boldsymbol{\phi}$ is at $z = \Phi((\varepsilon - m_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi}))/(\sigma_n - v_{1:t}^2(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})/\sigma_n))$. Unsurprisingly, if $m_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})$ is large enough, then there is a mode near $z = 0$. However, if $\sigma_n = v_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})$ and $m_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi}) > \varepsilon$, then the pdf goes to infinity as $z \to 0^+$. Interestingly, if $\sigma_n < v_{1:t}(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})$, then the pdf goes to infinity both as $z \to 0^+$ and $z \to 1^-$.

B.2 Gradient of the expected integrated variance acquisition function

We outline the derivation for the gradient of the expected integrated variance acquisition function (Equation (19)) with respect to the candidate evaluation location $\boldsymbol{\theta}^*$. We consider only the case where either a point estimate or a fixed value is used for the GP hyperparameters $\boldsymbol{\phi}$. First we define

$$c(\boldsymbol{\theta}, \boldsymbol{\theta}^*) := \sqrt{\frac{\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta}) - \tau_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta}) + \tau_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}}. \tag{58}$$

Because the second term in Equation (19) is constant with respect to $\boldsymbol{\theta}^*$, we obtain

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}^*} L_{1:t}(\boldsymbol{\theta}^*) &= 2\,\frac{\partial}{\partial \boldsymbol{\theta}^*} \int_{\Theta} \pi^2(\boldsymbol{\theta}) T\left(\frac{\varepsilon - m_{1:t}(\boldsymbol{\theta})}{\sqrt{\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta})}}, c(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\right) \mathrm{d}\boldsymbol{\theta} && (59)\\
&= \frac{1}{\pi} \int_{\Theta} \pi^2(\boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}^*} \int_0^{c(\boldsymbol{\theta}, \boldsymbol{\theta}^*)} \frac{e^{-\frac{(\varepsilon - m_{1:t}(\boldsymbol{\theta}))^2}{2(\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta}))}(1+x^2)}}{1 + x^2}\,\mathrm{d}x\,\mathrm{d}\boldsymbol{\theta} && (60)\\
&= \frac{1}{\pi} \int_{\Theta} \pi^2(\boldsymbol{\theta}) \frac{e^{-\frac{(\varepsilon - m_{1:t}(\boldsymbol{\theta}))^2}{2(\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta}))}(1+c^2(\boldsymbol{\theta}, \boldsymbol{\theta}^*))}}{1 + c^2(\boldsymbol{\theta}, \boldsymbol{\theta}^*)} \frac{\partial}{\partial \boldsymbol{\theta}^*} c(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\,\mathrm{d}\boldsymbol{\theta}, && (61)
\end{aligned}$$

where we have used the Leibnitz integration rule twice and where the integrations are applied elementwise. Differentiating $c(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ and some further computations produce

$$\frac{\partial}{\partial \boldsymbol{\theta}^*} L_{1:t}(\boldsymbol{\theta}^*) = -\frac{1}{2\pi} \int_{\Theta} \frac{\pi^2(\boldsymbol{\theta}) e^{-\frac{(\varepsilon - m_{1:t}(\boldsymbol{\theta}))^2}{\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta}) + \tau_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}}}{\sqrt{\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta}) + \tau_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}\sqrt{\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta}) - \tau_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}} \frac{\partial}{\partial \boldsymbol{\theta}^*} \tau_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\,\mathrm{d}\boldsymbol{\theta}, \tag{62}$$

where

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}^*} \tau_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &= \frac{2\mathrm{cov}_{1:t}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta}^*)} \frac{\partial}{\partial \boldsymbol{\theta}^*} \mathrm{cov}_{1:t}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \frac{\mathrm{cov}_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}{(\sigma_n^2 + v_{1:t}^2(\boldsymbol{\theta}^*))^2} \frac{\partial}{\partial \boldsymbol{\theta}^*} v_{1:t}^2(\boldsymbol{\theta}^*), && (63)\\
\frac{\partial}{\partial \boldsymbol{\theta}^*} \mathrm{cov}_{1:t}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &= \frac{\partial}{\partial \boldsymbol{\theta}^*} k(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - k(\boldsymbol{\theta}, \boldsymbol{\theta}_{1:t}) K^{-1}(\boldsymbol{\theta}_{1:t}, \boldsymbol{\theta}_{1:t}) \frac{\partial}{\partial \boldsymbol{\theta}^*} k(\boldsymbol{\theta}_{1:t}, \boldsymbol{\theta}^*). && (64)
\end{aligned}$$

The integral in Equation (62) can be approximated similarly as discussed in Section 3.3.

B.3 Gradient of the maxvar acquisition function

We compute the gradient of the maxvar acquisition function with respect to the parameter vector $\boldsymbol{\theta}$. We take into account the uncertainty in the GP hyperparameters but if a point estimate is used instead, the formulae can be simplified by ignoring the summations, setting $\omega^i = \omega^1 = 1$ and replacing $\boldsymbol{\phi}^i$ with the point estimate. First we denote

$$I_1(\boldsymbol{\theta}) = \sum_i \omega^i \Phi(a(\boldsymbol{\theta}, \boldsymbol{\phi}^i)) - \left[\sum_i \omega^i \, \Phi(a(\boldsymbol{\theta}, \boldsymbol{\phi}^i)) \right]^2, \tag{65}$$

$$I_2(\boldsymbol{\theta}) = \frac{1}{\pi} \sum_i \omega^i \int_0^{b(\boldsymbol{\theta}, \boldsymbol{\phi}^i)} \frac{e^{-\frac{1}{2}a^2(\boldsymbol{\theta}, \boldsymbol{\phi}^i)(1+x^2)}}{1+x^2} \, \mathrm{d}x, \tag{66}$$

where

$$a(\boldsymbol{\theta}, \boldsymbol{\phi}^i) = \frac{\varepsilon - m_{1:t}(\boldsymbol{\theta} \,|\, \boldsymbol{\phi}^i)}{\sqrt{(\sigma_n^i)^2 + v_{1:t}^2(\boldsymbol{\theta} \,|\, \boldsymbol{\phi}^i)}}, \quad b(\boldsymbol{\theta}, \boldsymbol{\phi}^i) = \frac{\sigma_n^i}{\sqrt{(\sigma_n^i)^2 + 2v_{1:t}^2(\boldsymbol{\theta} \,|\, \boldsymbol{\phi}^i)}}, \tag{67}$$

so that

$$\pi^2(\boldsymbol{\theta}) \mathbb{V}(p_a(\boldsymbol{\theta})) \approx \pi^2(\boldsymbol{\theta})(I_1(\boldsymbol{\theta}) - I_2(\boldsymbol{\theta})). \tag{68}$$

Differentiating Equation (68) with respect to the parameter vector $\boldsymbol{\theta}$ yields

$$\frac{\partial}{\partial \boldsymbol{\theta}} \pi^2(\boldsymbol{\theta}) \mathbb{V}(p_a(\boldsymbol{\theta})) \approx 2\pi(\boldsymbol{\theta})(I_1(\boldsymbol{\theta}) - I_2(\boldsymbol{\theta})) \frac{\partial \pi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \pi^2(\boldsymbol{\theta}) \left(\frac{\partial I_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\partial I_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right). \tag{69}$$

Computing the derivatives of I_1 produces

$$\begin{aligned} \frac{\partial I_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \sum_i \omega^i \frac{\partial}{\partial \boldsymbol{\theta}} \Phi(a(\boldsymbol{\theta}, \boldsymbol{\phi}^i)) - 2 \left(\sum_i \omega^i \Phi(a(\boldsymbol{\theta}, \boldsymbol{\phi}^i)) \right) \left(\sum_i \omega^i \frac{\partial}{\partial \boldsymbol{\theta}} \Phi(a(\boldsymbol{\theta}, \boldsymbol{\phi}^i)) \right) \\ &= \left(1 - 2 \sum_i \omega^i \Phi(a(\boldsymbol{\theta}, \boldsymbol{\phi}^i)) \right) \sum_i \omega^i \frac{\partial}{\partial \boldsymbol{\theta}} \Phi(a(\boldsymbol{\theta}, \boldsymbol{\phi}^i)) \\ &= \left(1 - 2 \sum_i \omega^i \Phi(a(\boldsymbol{\theta}, \boldsymbol{\phi}^i)) \right) \sum_i \frac{\omega^i e^{-\frac{1}{2}a^2(\boldsymbol{\theta}, \boldsymbol{\phi}^i)}}{\sqrt{2\pi}} \frac{\partial a(\boldsymbol{\theta}, \boldsymbol{\phi}^i)}{\partial \boldsymbol{\theta}}. \end{aligned} \tag{70}$$

Using the Leibniz integration rule, the gradient of I_2 can be written as

$$\frac{\partial I_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{\pi} \sum_i \omega^i \left(\frac{e^{-\frac{1}{2}a^2(\boldsymbol{\theta}, \boldsymbol{\phi}^i)(1+b^2(\boldsymbol{\theta}, \boldsymbol{\phi}^i))}}{1+b^2(\boldsymbol{\theta}, \boldsymbol{\phi}^i)} \frac{\partial b(\boldsymbol{\theta}, \boldsymbol{\phi}^i)}{\partial \boldsymbol{\theta}} + \int_0^{b(\boldsymbol{\theta}, \boldsymbol{\phi}^i)} \frac{1}{1+x^2} \frac{\partial}{\partial \boldsymbol{\theta}} e^{-\frac{1}{2}a^2(\boldsymbol{\theta}, \boldsymbol{\phi}^i)(1+x^2)} \, \mathrm{d}x \right), \tag{71}$$

where the integration is applied elementwise. The second term in Equation (71) can be further simplified as

$$\begin{aligned} &\int_0^{b(\boldsymbol{\theta}, \boldsymbol{\phi}^i)} \frac{1}{1+x^2} \frac{\partial}{\partial \boldsymbol{\theta}} e^{-\frac{1}{2}a^2(\boldsymbol{\theta}, \boldsymbol{\phi}^i)(1+x^2)} \, \mathrm{d}x \\ &= -\int_0^{b(\boldsymbol{\theta}, \boldsymbol{\phi}^i)} a(\boldsymbol{\theta}, \boldsymbol{\phi}^i) \frac{\partial a(\boldsymbol{\theta}, \boldsymbol{\phi}^i)}{\partial \boldsymbol{\theta}} e^{-\frac{1}{2}a^2(\boldsymbol{\theta}, \boldsymbol{\phi}^i)(1+x^2)} \, \mathrm{d}x \\ &= -a(\boldsymbol{\theta}, \boldsymbol{\phi}^i) \frac{\partial a(\boldsymbol{\theta}, \boldsymbol{\phi}^i)}{\partial \boldsymbol{\theta}} e^{-\frac{1}{2}a^2(\boldsymbol{\theta}, \boldsymbol{\phi}^i)} \int_0^{b(\boldsymbol{\theta}, \boldsymbol{\phi}^i)} e^{-\frac{1}{2}a^2(\boldsymbol{\theta}, \boldsymbol{\phi}^i)x^2} \, \mathrm{d}x \\ &= -\sqrt{2\pi} \frac{\partial a(\boldsymbol{\theta}, \boldsymbol{\phi}^i)}{\partial \boldsymbol{\theta}} e^{-\frac{1}{2}a^2(\boldsymbol{\theta}, \boldsymbol{\phi}^i)} \left(\Phi(a(\boldsymbol{\theta}, \boldsymbol{\phi}^i) b(\boldsymbol{\theta}, \boldsymbol{\phi}^i)) - \Phi(0) \right) \\ &= \sqrt{\pi/2} \, e^{-\frac{1}{2}a^2(\boldsymbol{\theta}, \boldsymbol{\phi}^i)} \left(1 - 2\Phi(a(\boldsymbol{\theta}, \boldsymbol{\phi}^i) b(\boldsymbol{\theta}, \boldsymbol{\phi}^i)) \right) \frac{\partial a(\boldsymbol{\theta}, \boldsymbol{\phi}^i)}{\partial \boldsymbol{\theta}}, \end{aligned} \tag{72}$$

where on the third line we have recognised the integrand as an unnormalised Gaussian pdf. Finally, straightforward calculations show that

$$\frac{\partial a(\boldsymbol{\theta}, \boldsymbol{\phi}^i)}{\partial \boldsymbol{\theta}} = -\frac{1}{\sqrt{(\sigma_n^i)^2 + v_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\phi}^i)}} \frac{\partial m_{1:t}(\boldsymbol{\theta}, \boldsymbol{\phi}^i)}{\partial \boldsymbol{\theta}} - \frac{\varepsilon - m_{1:t}(\boldsymbol{\theta}, \boldsymbol{\phi}^i)}{2((\sigma_n^i)^2 + v_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\phi}^i))^{3/2}} \frac{\partial v_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\phi}^i)}{\partial \boldsymbol{\theta}}, \tag{73}$$

$$\frac{\partial b(\boldsymbol{\theta}, \boldsymbol{\phi}^i)}{\partial \boldsymbol{\theta}} = -\frac{\sigma_n^i}{((\sigma_n^i)^2 + 2v_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\phi}^i))^{3/2}} \frac{\partial v_{1:t}^2(\boldsymbol{\theta}, \boldsymbol{\phi}^i)}{\partial \boldsymbol{\theta}}. \tag{74}$$

Gradients of the GP mean and variance functions m and v^2 depend on the chosen covariance function and are not shown here.

B.4 Gradient of the ABC posterior approximation

The gradient of the posterior approximation with respect to parameter $\boldsymbol{\theta}$ is useful if e.g. Hamiltonian Monte Carlo algorithm is used to sample from this density. The unnormalised approximate posterior is

$$\tilde{\pi}_{\varepsilon}^{\mathrm{ABC}}(\boldsymbol{\theta} \,|\, x_{obs}) = \pi(\boldsymbol{\theta})\, \Phi(a(\boldsymbol{\theta})) \tag{75}$$

where, as earlier, $\pi(\boldsymbol{\theta})$ denotes the prior density and a is defined as in Equation (67). Differentiating the logarithm of Equation (75) yields

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log \tilde{\pi}_{\varepsilon}^{\mathrm{ABC}}(\boldsymbol{\theta} \,|\, x_{obs}) = \frac{\partial \log \pi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \frac{\partial}{\partial \boldsymbol{\theta}} \log \Phi(a(\boldsymbol{\theta})) = \frac{\partial \log \pi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \frac{e^{-\frac{1}{2}a^2(\boldsymbol{\theta})}}{\sqrt{2\pi}\, \Phi(a(\boldsymbol{\theta}))} \frac{\partial a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \tag{76}$$

where the gradient of the term a can be computed as in Equation (73).

C Additional details and experiments

In this section we first briefly discuss the computational cost of our algorithm using Big O notation and then we provide some further details and results of our experiments.

We consider the cost of evaluating the expintvar acquisition function at b arbitrary points $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_b^*$ at the iteration t of Algorithm 1. Computing the Cholesky factorisation on line 8 of Algorithm 1 requires $\mathcal{O}(t^3)$. Generating s samples from the proposal π_q on line 9 requires computing GP mean and variance functions for each proposed point by the MCMC algorithm and the total cost is $\mathcal{O}(st^2)$ when the precomputed Cholesky is used. We can reuse the GP mean and variance function values and line 10 thus requires $\mathcal{O}(\tilde{s})$ where $\tilde{s}$ is the amount of thinned samples that are used for approximating the integral in Equation (19). We obviously have $\tilde{s} \leq s$. (In fact, the second term of Equation (19) is constant with respect to $\boldsymbol{\theta}^*$ and we may not need to evaluate it. However, from the previous discussion we can see that the resulting saving would be small i.e. only $\mathcal{O}(\tilde{s})$.) Computing the value of Equation (20) at the $\tilde{s}$ sampled locations and for all b values of $\boldsymbol{\theta}^*$ can be seen to be $\mathcal{O}(b\tilde{s}t + bt^2)$ when we use the already computed values of $k(\boldsymbol{\theta}, \boldsymbol{\theta}_{1:t})K^{-1}(\boldsymbol{\theta}_{1:t})$ in the formula of $\mathrm{cov}_{1:t}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$. Computing the integral over the first term then requires $\mathcal{O}(\tilde{s})$. The total cost is thus $\mathcal{O}(t^3 + st^2 + b(t^2 + \tilde{s}t))$. If we use grid approximation instead of importance sampling to approximate the integral in Equation (19), we obtain the total cost by setting $s = \tilde{s}$ where s is now the number of grid points.

We obtain $\mathcal{O}(t^3 + bt^2)$ bound for the cost of LCB, maxvar and expdiffvar and we see that the $st^2 + b\tilde{s}t$ term is missing as compared to the corresponding bound of the expintvar acquisition function. There is no acquisition function to evaluate (or optimise) for rand_maxvar rule since we choose the next point directly sampling from a density. This can be done in $\mathcal{O}(t^3 + st^2)$. Finally, we want to emphasise that this analysis is asymptotic and in practice the constant costs of expintvar and its variants due to the need to compute e.g. the Owens t-function and cdf of the standard Gaussian are slightly higher as compared to e.g. LCB rule. However, in practice all these GP computation times are negligible when the simulation time dominates the computation.

In the rest of this section we show further details and results of the experiments. The synthetic test problems in Section 4.1 are designed in the following way. In the "unimodal" example, the mean of the discrepancy is $m(\boldsymbol{\theta}) = 3\sigma + \boldsymbol{\theta}^T\mathbf{S}\boldsymbol{\theta}$, where σ is the standard deviation of the additive Gaussian noise, $\mathbf{S}_{11} = \mathbf{S}_{22} = 1$ and $\mathbf{S}_{12} = \mathbf{S}_{21} = 0.5$. In the "bimodal" example we use $m(\boldsymbol{\theta}) = 3\sigma + 0.2(\theta_2 - \theta_1^2)^2 + 0.75(\theta_2 - \theta_1 - 2)^2$. In the unidentifiable example, the mean of the discrepancy is obtained as $m(\boldsymbol{\theta}) = 3\sigma + 0.01\theta_1^2 + \theta_2^2$. The "banana" example is produced using $m(\boldsymbol{\theta}) = 3\sigma + (1-\theta_1)^2 + 10(\theta_2 - \theta_1^2)^2$. The discrepancy is assumed to follow the Gaussian density, that is, $\Delta_{\boldsymbol{\theta}} \sim \mathcal{N}(m(\boldsymbol{\theta}), \sigma^2)$. The resulting probability densities with $\sigma = 2$ are also illustrated in Figure 3.

We present additional results for the 2d experiments in Section 4.1. The settings are the same except that the threshold is not predefined but is set to the 0.01th quantile of the realised discrepancies, and updated constantly during the acquisitions. Consequently, the selection of the evaluation locations also affects how the threshold is chosen. These results are shown in Figure 9. Figure 10 shows the corresponding results when the threshold is determined using the 0.05th quantile.

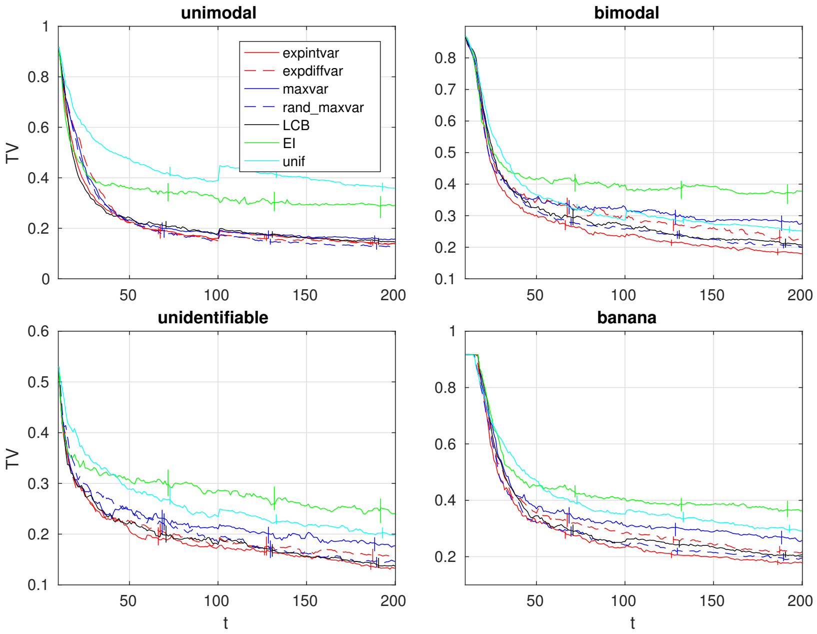

Figure 9: Median of the TV distance between the estimated and the true ABC posterior over 100 experiments. The 0.01th quantile is used for updating the threshold. The results are similar as in Figure 4.

D Non-uniform acceptance threshold

Instead of using the uniform (i.e. "0-1") threshold $\pi_\varepsilon(\mathbf{x}_{\text{obs}} \,|\, \mathbf{x}) \propto \mathbb{1}_{\Delta(\mathbf{x}_{obs},\mathbf{x})\leq\varepsilon}$, other choices are possible. For instance, one can use "Gaussian threshold" $\pi_\varepsilon(\mathbf{x}_{\text{obs}} \,|\, \mathbf{x}) \propto \mathcal{N}(\Delta(\mathbf{x}_{\text{obs}}, \mathbf{x}) \,|\, m_\varepsilon, \sigma_\varepsilon^2)$, where the threshold ε is replaced by two new parameters m_ε and σ_ε^2 that control the quality of the ABC approximation. The unnormalised ABC posterior approximation at $\boldsymbol{\theta}$ is then given by

$$\begin{aligned} \tilde{\pi}_{\text{N}}^{\text{ABC}}(\boldsymbol{\theta}) &= \pi(\boldsymbol{\theta}) \int_{-\infty}^{\infty} \mathcal{N}(\Delta \,|\, m_\varepsilon, \sigma_\varepsilon^2)\, \mathcal{N}(\Delta \,|\, f(\boldsymbol{\theta}), \sigma_n^2)\, \mathrm{d}\Delta & (77) \\ &= \pi(\boldsymbol{\theta})\, \mathcal{N}(f(\boldsymbol{\theta}) \,|\, m_\varepsilon, \sigma_\varepsilon^2 + \sigma_n^2). & (78) \end{aligned}$$

This approach can be seen as an approximation to the uniform threshold but it could be interpreted also as additional Gaussian measurement (or modelling) error as described by Wilkinson [2013].

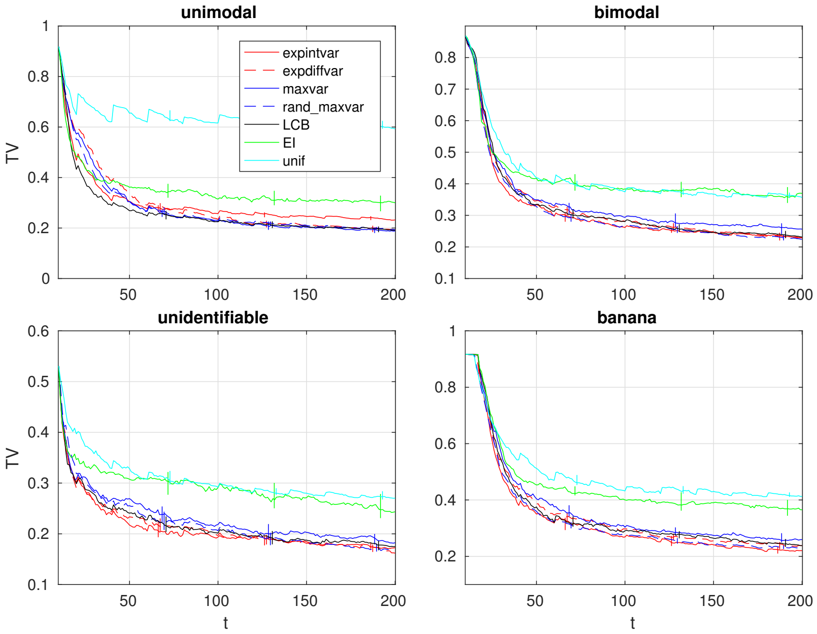

Figure 10: Median of the TV distance between the estimated and the true ABC posterior over 100 experiments. These experiments are as in Figure 9 except that the 0.05th quantile is used. Larger threshold than in Figure 9 generally produces slightly worse posterior estimates.

Proceeding similarly as in the proof of Lemma 3.1 and using Gaussian identities in the appendix of Rasmussen and Williams [2006] to compute the required integrals, the expectation and variance of $\tilde{\pi}^{\mathrm{ABC}}_{\mathrm{N}}$ can be shown to be

$$\mathbb{E}(\tilde{\pi}^{\mathrm{ABC}}_{\mathrm{N}}(\boldsymbol{\theta}) \,|\, D_{1:t}) = \pi(\boldsymbol{\theta}) \mathcal{N}(m_\varepsilon \,|\, m_{1:t}(\boldsymbol{\theta}), \sigma^2 + v^2_{1:t}(\boldsymbol{\theta})), \tag{79}$$

$$\mathbb{V}(\tilde{\pi}^{\mathrm{ABC}}_{\mathrm{N}}(\boldsymbol{\theta}) \,|\, D_{1:t}) = \frac{\pi^2(\boldsymbol{\theta})}{2\sqrt{\pi\sigma^2}} \mathcal{N}\left(m_\varepsilon \,\middle|\, m_{1:t}(\boldsymbol{\theta}), \frac{\sigma^2}{2} + v^2_{1:t}(\boldsymbol{\theta}) \right) - \pi^2(\boldsymbol{\theta}) \left[\mathcal{N}(m_\varepsilon \,|\, m_{1:t}(\boldsymbol{\theta}), \sigma^2 + v^2_{1:t}(\boldsymbol{\theta})) \right]^2, \tag{80}$$

where $\sigma^2 = \sigma^2_\varepsilon + \sigma^2_n$. In addition, the expected integrated variance acquisition function can now be written

$$L_{1:t}(\boldsymbol{\theta}^*) = \int_\Theta \pi^2(\boldsymbol{\theta}) \left[\frac{\mathcal{N}\left(m_\varepsilon \,\middle|\, m_{1:t}(\boldsymbol{\theta}), \frac{\sigma^2}{2} + v^2_{1:t}(\boldsymbol{\theta}) \right)}{2\sqrt{\pi\sigma^2}} - \frac{\mathcal{N}\left(m_\varepsilon \,\middle|\, m_{1:t}(\boldsymbol{\theta}), \frac{\sigma^2 + v^2_{1:t}(\boldsymbol{\theta}) - \tau^2_{1:t}(\boldsymbol{\theta},\boldsymbol{\theta}^*)}{2} \right)}{2\sqrt{\pi}\sqrt{\sigma^2 + v^2_{1:t}(\boldsymbol{\theta}) + \tau^2_{1:t}(\boldsymbol{\theta},\boldsymbol{\theta}^*)}} \right] \mathrm{d}\boldsymbol{\theta}, \tag{81}$$

where the computations follow similarly as in the proof of Proposition 3.2 and by applying Gaussian identities to compute the integrals. (Details are left to the reader.)

The advantage of this approach is that one avoids Owen's t-function evaluations. However, we found determining the two threshold values m_ε and σ^2_ε more challenging than setting the threshold ε for the uniform threshold and thus focused on the latter approach. On the other hand, while running the simulation model typically dominates the total computational cost, these formulae may be useful in high-dimensions where the global optimisation of the acquisition function can also be costly.