

Adaptive Gaussian process approximation for Bayesian inference with expensive likelihood functions

Hongqiao Wang¹ and Jinglai Li²

¹Institute of Natural Sciences and School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai 200240, China.

²Institute of Natural Sciences, School of Mathematical Sciences, and MOE Key Laboratory of Scientific and Engineering Computing, Shanghai Jiao Tong University, Shanghai 200240, China. (Corresponding author).

Keywords: Active learning, Bayesian inference, entropy, Gaussian process, inverse problems

Abstract

We consider Bayesian inference problems with computationally intensive likelihood functions. We propose a Gaussian process (GP) based method to approximate the joint distribution of the unknown parameters and the data, built upon a recent work [10]. In particular, we write the joint density approximately as a product of an approximate posterior density and an exponentiated GP surrogate. We then provide an adaptive algorithm to construct such an approximation, where an active learning method is used to choose the design points. With numerical examples, we illustrate that the proposed method has competitive performance against existing approaches for Bayesian computation.

1 Introduction

The Bayesian inference is a popular method to estimate unknown parameters from data, and a major advantage of the method is its ability to quantify uncertainty in the inference results [6]. In this work we consider Bayesian inference problems where the likelihood functions are highly expensive to evaluate. A typical example of this type of problems is the Bayesian inverse problems [33], where the parameters of interest can not be observed directly and need to be estimated from indirect data. Such problems arise from many real-world applications, ranging from carbon capture [13] to chemical kinetics [7]. In Bayesian inverse problems, the mappings from the parameter of interest to the observable quantities, often known as the forward models, are often computationally intensive, e.g., involving simulating large scale computer models.

Due to the high computational cost, common numerical implementations of Bayesian inferences, such as the Markov chain Monte Carlo (MCMC) [1] methods can be prohibitively expensive. A simple idea to accelerate the computation of the posterior is to construct a computationally inexpensive surrogate or an approximation of the posterior distribution with a limited number of likelihood function evaluations. To this end, a particular convenient choice for surrogate function is the Gaussian Process (GP) model [35]. The idea of using the GP model to approximate the posterior or the likelihood function dates back to the so-called Bayesian quadrature (or Bayesian Monte Carlo) approaches [23, 28, 27, 11], which were designed to perform numerical integrations in a Bayesian fashion (for example, to compute the evidence in Bayesian inference problems [26]). Unlike the Bayesian quadrature methods, the goal of this work is to construct an approximation of the posterior distribution. To this end, a recent work [10] approximates the joint distribution of the unknown parameter and the data (which can also be viewed as the un-normalized posterior distribution) with an exponentiated GP model, where the design points, i.e., the points where the likelihood function is evaluated, are chosen with an active learning strategy. In particular, they determine the design points by sequentially maximizing the variance in the posterior approximation. Other ideas of using the GP approximation to accelerate the Bayesian computation can be found in [4, 3], and so on. The method presented in this work also intends to approximate the un-normalized posterior distribution. The main contribution of the work is the following. We write the unnormalized posterior distribution as a product of an approximate posterior density and an exponentiated GP surrogate. The intuition behind this formulation is that, the GP model can be more effectively constructed if we factor out a good approximation of the posterior (see Section 2.3 for a detailed explanation). As we may not know a good approximate posterior density in advance, we develop an algorithm to adaptively construct the product-form approximation of the un-normalized posterior distribution. Another difference between our method can that in [10] is the learning strategy for selecting the design points. Namely, we use the entropy rather than the variance as the selection criterion, which can better represent the uncertainty in the approximation. Numerical examples illustrate that the proposed method can substantially improve the performance of the GP approximation.

We note that, other surrogate models, notably the generalized polynomial chaos (gPC) expansion [15, 17, 18, 19, 21], have also been used to accelerate the Bayesian computation. Detailed comparison of the two type of the surrogates is not discussed in this work and those who are interested in this matter may consult [24].

The rest of the paper is organized as the following. In Section 2 we present the adaptive GP algorithm to construct the posterior approximation and the active leaning method to determine the design points. In section 3, we give two examples to illustrate the performance of the proposed method. Finally section 4 provides some concluding remarks.

2 The adaptive GP method

2.1 Problem Setup

A Bayesian inference problem aims to estimate an unknown parameter \mathbf{x} from data \mathbf{d} , and specifically it computes the posterior distribution of \mathbf{x} using the Bayes' formula:

$$\pi(\mathbf{x}|\mathbf{d}) \propto \pi(\mathbf{x}, \mathbf{d}) = l(\mathbf{d}|\mathbf{x})\pi(\mathbf{x}), \quad (1)$$

where $l(\mathbf{d}|\mathbf{x})$ is the likelihood function and $\pi(\mathbf{x})$ is the prior distribution of \mathbf{x} . When the Bayesian method is applied to inverse problems, the data and the forward model enter the formulation through the likelihood function. Namely, suppose that there is a function (termed as the forward function or the forward model) that maps the parameter of interest \mathbf{x} to the observable quantity \mathbf{y} :

$$\mathbf{y} = \mathbf{G}(\mathbf{x}) + \mathbf{z},$$

where \mathbf{z} is the observation error. Now we further assume that the distribution density of the observation noise \mathbf{z} , $p_z(\mathbf{z})$, is available, and it follows directly that the likelihood function is given by

$$l(\mathbf{d}|\mathbf{x}) = p_z(\mathbf{d} - \mathbf{G}(\mathbf{x})).$$

In what follows we shall omit the argument \mathbf{d} in the likelihood function and denote it as $l(\mathbf{x})$ for simplicity. It is easy to see that each evaluation of the likelihood function $l(\mathbf{x})$ requires to evaluate the forward function $\mathbf{G}(\mathbf{x})$. In practice, the forward function $\mathbf{G}(\mathbf{x})$ often represents a large-scale computer model, and thus the evaluation of $l(\mathbf{x})$ can be highly computational demanding. Due to the high computational cost, the brute-force Monte Carlo simulation can not be used for such problems, and we resort to an alternative method to compute the posterior distributions, using the GP surrogate model. A brief description of the GP method is provided in next section.

2.2 The GP model

Given a real-valued function $g(\mathbf{x})$, the GP or the Kriging method constructs a surrogate model of $g(\mathbf{x})$ in a nonparameteric Bayesian regression framework [35, 22, 25]. Specifically the target function $g(\mathbf{x})$ is cast as a Gaussian random process whose mean is $\mu(\mathbf{x})$ and covariance is specified by a kernel function $k(\mathbf{x}, \mathbf{x}')$, namely,

$$\text{COV}[g(\mathbf{x}), g(\mathbf{x}')] = k(\mathbf{x}, \mathbf{x}').$$

The kernel $k(\mathbf{x}, \mathbf{x}')$ is positive semidefinite and bounded. Now let us assume that m evaluations of the function $g(\mathbf{x})$ are performed at parameter values $\mathbf{X}^* := [\mathbf{x}_1^*, \dots, \mathbf{x}_m^*]$, yielding function evaluations $\mathbf{y}^* := [y_1^*, \dots, y_m^*]$, where

$$y_i^* = g(\mathbf{x}_i^*) \quad \text{for } i = 1, \dots, m.$$

Suppose that we want to predict the function values at points $\mathbf{D} := [\mathbf{x}_1, \dots, \mathbf{x}_{m'}]$, i.e., $\mathbf{y} = [y_1, \dots, y_{m'}]$ where $y_i = g(\mathbf{x}_i)$. The sets \mathbf{X}^* and \mathbf{D} are often known as the training and the test points respectively. The joint prior distribution of $(\mathbf{y}^*, \mathbf{y})$ is,

$$\begin{bmatrix} \mathbf{y}^* \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mu(\mathbf{X}^*) \\ \mu(\mathbf{D}) \end{bmatrix}, \begin{bmatrix} K(\mathbf{X}^*, \mathbf{X}^*) & K(\mathbf{X}^*, \mathbf{D}) \\ K(\mathbf{D}, \mathbf{X}^*) & K(\mathbf{D}, \mathbf{D}) \end{bmatrix} \right), \quad (2)$$

where we use the notation $K(\mathbf{A}, \mathbf{B})$ to denote the matrix of the covariance evaluated at all pairs of points in set \mathbf{A} and in set \mathbf{B} . The posterior distribution of \mathbf{y} is also Gaussian:

$$\mathbf{y} \mid \mathbf{D}, \mathbf{X}^*, \mathbf{y}^* \sim \mathcal{N}(\mathbf{u}, \Sigma), \quad (3a)$$

where the posterior mean is

$$\mathbf{u} = \mu(\mathbf{D}) + K(\mathbf{D}, \mathbf{X}^*)K(\mathbf{X}^*, \mathbf{X}^*)^{-1}(\mathbf{y} - \mu(\mathbf{D})), \quad (3b)$$

and the posterior covariance matrix is

$$\Sigma = K(\mathbf{D}, \mathbf{D}) - K(\mathbf{D}, \mathbf{X}^*)K(\mathbf{X}^*, \mathbf{X}^*)^{-1}K(\mathbf{X}^*, \mathbf{D}). \quad (3c)$$

Here we only provide a brief introduction to the GP method tailored for our own purposes, and readers who are interested in further details may consult the aforementioned references.

2.3 The adaptive GP algorithm

Now we discuss how to use the GP method to compute the posterior distribution in our problem. A straightforward idea is to construct the surrogate model directly for the log-likelihood function $\log l(\mathbf{x})$, and such a method has been used in the aforementioned works [26, 10]. A difficulty in this approach is that the target function $\log l(\mathbf{x})$ can be highly nonlinear and fast varying, and thus are not well described by a GP model. We here present an adaptive scheme to alleviate the difficulty.

We first write the unnormalized posterior, i.e., the joint distribution $\pi(\mathbf{x}, \mathbf{d})$, as

$$f(\mathbf{x}) = l(\mathbf{x})\pi(\mathbf{x}) = \exp(g(\mathbf{x}))p(\mathbf{x}),$$

where $p(\mathbf{x})$ is a probability distribution that we are free to choose and

$$g(\mathbf{x}) = \log(f(\mathbf{x})/p(\mathbf{x})). \quad (4)$$

We work on the log posterior distribution since the log smoothes out a function and is more conducive for the GP modeling. Also, by doing this we ensure the non-negativity of the obtained approximate posterior. We then sample the function $g(\mathbf{x})$ at certain locations and construct the GP surrogate of $g(\mathbf{x})$. It should be noted that, the distribution $p(\mathbf{x})$ plays an important role in the surrogate construction as a good choice of $p(\mathbf{x})$ can significantly improve the accuracy of the GP surrogate models. In particular, if we take $p(\mathbf{x})$ to be exactly the posterior $\pi(\mathbf{x}|\mathbf{d})$, it follows immediately that $g(\mathbf{x})$ in Eq (4) is a constant. This then gives us the intuition that, if $p(\mathbf{x})$ is a good approximation to the posterior distribution $\pi(\mathbf{x}|\mathbf{d})$, $g(\mathbf{x})$ is a mildly varying function which is easy to approximate. In other word, we can improve the performance of the GP surrogate by factoring out a good approximation of the posterior. Certainly, this can not be done in one step, as the posterior is not known in advance. We present here an adaptive framework to construct a sequence of pairs $\{p_i(\mathbf{x}), \exp(\hat{g}_i(\mathbf{x}))\}$, the product of which evolves to a good approximation of the unnormalized posterior $f(\mathbf{x})$. Roughly speaking the algorithm performs the following iterations: in the n -th cycle, given the current

guess of the posterior distribution $p_n(\mathbf{x})$, we construct a GP surrogate $\hat{g}_n(\mathbf{x})$ of $g_n(\mathbf{x})$ which is given by

$$g_n(\mathbf{x}) = \log(f(\mathbf{x})/p_n(\mathbf{x})),$$

and we then compute a new (and possibly better) posterior approximation $p_{n+1}(\mathbf{x})$ using

$$p_{n+1}(\mathbf{x}) \propto \exp(\hat{g}_n(\mathbf{x}))p_n(\mathbf{x}).$$

Finally we want to specify stopping criteria for the iteration, and the iteration terminates if either of the following two conditions are satisfied. The first is that the maximum number of iterations is reached. Our second stopping condition is based on the Kullback-Leiber (KL) divergence between p_{n-1} and p_n , which reads,

$$D_{KL}(p_{n-1}, p_n) = \int \log \frac{p_{n-1}(\mathbf{x})}{p_n(\mathbf{x})} p_{n-1}(\mathbf{x}) d\mathbf{x}. \quad (5)$$

Specifically the second stopping condition is that $D_{KL}(p_{n-1}, p_n)$ is smaller than a prescribed value D_{\max} in K consecutive iterations. That is, if the computed posterior approximation does not change much in a certain number of consecutive iterations, the algorithm terminates. The complete scheme is described in Algorithm 1.

Algorithm 1 The adaptive GP algorithm

Require: $m_0, n_{\max}, M, D_{\max}, k_{\max}$

- 1: let $\hat{p}_0(x) = \pi(x)$; let $n = 0$; let $k = 0$;
- 2: choose m_0 initial design points: $\{x_1, \dots, x_{m_0}\}$, and compute $y_i = f(x_i)$ for $i = 1 \dots m_0$;
- 3: let $S_0 = \{(x_1, y_1), \dots, (x_{m_0}, y_{m_0})\}$;
- 4: **for** $n=0$ **do** n_{\max}
- 5: let $g_n(x) = \log(f(x)/\hat{p}_n(x))$;
- 6: construct a GP surrogate model $\hat{g}_n(x)$ for the function $g_n(x)$ with data set S_n ;
- 7: draw a set of M samples from the approximate posterior

$$p_{n+1}(x) \propto \exp(\hat{g}_n(x))\hat{p}_n(x)$$

with MCMC, denoted as A_n ;

- 8: obtain an estimated PDF from samples A_n , denoted as \hat{p}_{n+1} ;
 - 9: compute $D_{KL}(\hat{p}_{n-1}, \hat{p}_n)$;
 - 10: **if** $D_{KL}(\hat{p}_{n-1}, \hat{p}_n) < D_{\max}$ **then** $k = k + 1$;
 - 11: **else** $k = 0$;
 - 12: **end if**
 - 13: **if** $k = K$ **then** break the FOR loop;
 - 14: **else**
 - 15: select m design points: $\{x_1, \dots, x_m\}$, evaluate $f(x_i)$ for $i = 1 \dots m$, and let $S_{n+1} = S_n \cup \{(x_1, y_1), \dots, (x_m, y_m)\}$;
 - 16: **end if**
 - 17: **end for**
-

Some remarks on the implementation of Algorithm 1 are listed in order:

- In Line 6, we construct the GP model for $g_n(\mathbf{x})$ using the procedure described in Section 2.2. The hyperparameters of the GP model are determined by maximizing the marginal likelihood function [35].
- In Line 7, we resort to the MCMC method to draw a rather large number of samples from the approximate posterior distribution; this procedure, however, does not require to evaluate the true likelihood function and is not computationally expensive.
- In Line 8, we need to compute the density function of a distribution p_{n+1} from the samples X_n , and here we use the Gaussian mixture method [20] to estimate the density. Certainly there will be estimation errors in this procedure and so we denote the estimated density as \hat{p}_{n+1} to distinguish it from the true density p_{n+1} .
- In Line 9, we find that it is rather costly to compute the KLD between p_{n-1} and p_n . We instead use the KLD between \hat{p}_{n-1} and \hat{p}_n , which is much easier to compute as the distributions are available as Gaussian mixtures.
- In Line 15, we need to determine the design points, i.e., the locations where we evaluate the true function. The choice of design points is critical to the performance of the proposed adaptive GP algorithm, and we use an active learning method to determine the points, which is presented in Section 2.4.

2.4 Active learning for the design points

In the GP literature, the determination of the design points is often cast as an experimental design problem, i.e., to find the experimental parameters that can provide us the most information. The problem has received considerable attention and a number of methods and criteria have been proposed to select the points, such as, the Mutual Information criterion [14], the Integrated Mean Square Error (IMSE) [30], the Integrated Posterior Variance (IVAR) [8], and the active learning MacKay (ALM) criterion [16], just to name a few. Here we choose to use an active learning strategy, that adds one design point a time, primarily for that it is easy to implement.

A common active learning strategy is to choose the point that has the largest uncertainty, and to this end we need a function that can measure or quantify the uncertainty in the approximation reconstructed. In the usual GP problems, the variance of the GP model $\hat{g}(\mathbf{x})$ is a natural choice for such a measure of uncertainty (which yields the ALM method), because the distribution of $\hat{g}(\mathbf{x})$ is Gaussian. In our problems, however, the function of interest is the posterior approximation $\hat{f}(\mathbf{x}) = \exp(\hat{g}(\mathbf{x}))p(\mathbf{x})$ rather than the GP model $\hat{g}(\mathbf{x})$ itself, and thus we should measure the uncertainty in $f(\mathbf{x})$. In [10], the variance of the posterior approximation \hat{f} is used as the measure function. However, since the distribution of $\hat{f}(\mathbf{x})$ is not Gaussian, the variance may not provide a good estimate of the uncertainty. On the other hand, the entropy is a commonly used measure to quantify the uncertainty in a random variable [29, 32], and here we use it as our design criterion.

Specifically, suppose that, at point \mathbf{x} , the distribution of $\hat{f}(\mathbf{x})$ is $\pi_f(\hat{f})$, and the entropy of $\hat{f}(\mathbf{x})$ is defined as

$$\mathbb{H}(\hat{f}(\mathbf{x})) = - \int \log(\pi_f(\hat{f})) \pi_f(\hat{f}) d\hat{f}. \quad (6)$$

Thus we choose a new design point by

$$\max_{\mathbf{x} \in \Omega} \mathbb{H}(\hat{f}(\mathbf{x})),$$

where Ω is a bounded subspace of the state space of \mathbf{x} . In the present problem, the distribution of $\hat{g}(\mathbf{x})$ is Gaussian and let us assume its mean and variance are μ and σ^2 respectively. It follows that the distribution of $\hat{f}(\mathbf{x})$ is log-normal and the entropy of it can be computed analytically:

$$\mathbb{H}(\hat{f}) = \mu + \frac{1}{\ln} (2\pi e \sigma^2). \quad (7)$$

We want to emphasize that the entropy based active learning method is different from the usual maximum entropy method for experimental design, e.g., [31]. The purpose of the maximum entropy method in [31], is the find the design points that maximize the information gain of an inference problem, while in our problem, we use the entropy as a measure of uncertainty.

Now suppose that we have a set of existing data points, and we want to choose m new design points. We use the following scheme to sequentially choose the new points:

1. Construct a GP model $\hat{g}(\mathbf{x})$ for $g(\mathbf{x})$ using data set S ;
2. Compute $\mathbf{x}^* = \arg \max_{\mathbf{x} \in \Omega} \mathbb{H}(\hat{f}(\mathbf{x}))$;
3. Evaluate $y^* = g(\mathbf{x}^*)$ and let $S = S \cup \{(\mathbf{x}^*, y^*)\}$;

Note that the key in the adaptive scheme is Step 2, where we seek the point \mathbf{x} that maximizes the entropy $\mathbb{H}(\hat{f}(\mathbf{x}))$ in Ω . This is a quite challenging problem from an optimization perspective, because the gradient of the objective function can not be easily obtained and the problem may have multiple local maxima. However, in the numerical tests, we have found that, our algorithm does not strictly require the optimality of the solution and it performs well as long as a good design point can be found in each step. Thus here we use a stochastic search method, the simulated annealing algorithm [12], to find the design point. We have also tested other meta-heuristic optimization algorithms, and the performances do not vary significantly.

3 Numerical examples

3.1 The Rosenbrock function

We first test our method on a two-dimensional mathematical example. The likelihood function is

$$l(\mathbf{x}) = \exp \left(-\frac{1}{100} (x_1 - 1)^2 - (x_1^2 - x_2)^2 \right), \quad (8)$$

which is the well-known Rosenbrock function, and the prior $\pi(\mathbf{x})$ is a uniform distribution defined on $[-5, 5] \times [-5, 5]$. The resulting unnormalized posterior is shown in Fig. 1 (left). The function has a “banana shape”, and is often used as a test problem for Bayesian computation methods.

We now apply the proposed adaptive GP method to compute the posterior for this problem. In this example, we let $m_0 = 20$ and the samples in S_0 were randomly drawn according to the prior distribution. We also choose $m = 10$: namely, 10 new design points are computed in each iteration. In the algorithm, we need to sample from the approximate posterior distribution in each iteration, and here we draw $M = 2 \times 10^4$ samples with the delayed rejection adaptive Metropolis algorithm (DRAM) [9]. We restate that the 2×10^4 MCMC samples are generated from the approximate posterior distribution and thus it does not require to evaluate the true likelihood function. We also set the parameters that specify the termination conditions to be $n_{\max} = 100$, $D_{\max} = 0.01$ and $K = 5$. The algorithm terminates with 13 iterations and totally 140 evaluations of the true likelihood function are used. In Figs. 1 (right), we plot the KL difference in two consecutive iterations, which is used as one of our stopping criteria, against the number of iterations. To illustrate the performance of our method, we use the KL distance and the Hellinger distance which is defined as,

$$D_H(p_1, p_2) = \frac{1}{2} \int (\sqrt{p_1(\mathbf{x})} - \sqrt{p_2(\mathbf{x})})^2 d\mathbf{x},$$

to quantify the difference between the computed approximation and the true posterior. We plot the KL (left) and the Hellinger (right) distances between the approximate posterior and the true posterior distribution in Figs. 2. It can be seen from the figures that, the computed approximation converges very well to the true posterior in terms of both distance measures, as the iteration proceeds. We then plot the approximate posterior obtained in the 7th, 9th, 11th and 13th iterations in Figs. 3, in which we can visualize how the quality of the approximation increases as the iterations proceed. In each of the plots, we also show the design points (red dots) that have been used up to the given iteration. As a comparison, we also compute the GP approximation of the posterior with the aforementioned Bayesian active posterior estimation (BAPE) method developed in [10]. In particular, we implement the BAPE method using totally 140 design points and this way it matches the number of design points of our method. The results are shown in Figs 4. The figure on the left shows the posterior distribution computed with all the 140 design points (corresponding to the 13th iteration in our method), and as one can see, the BAPE method can also obtain a good approximation of the posterior distribution. To compare the performance of the two methods, we compute the KL divergence between the true posterior and the approximation obtained with different numbers of design points by the BAPE and our adaptive GP (AGP) methods. We plot the KL distance against the number of design points in Fig. 4 (left). One can see from the figure that, with the same number of design points, the approximate posterior obtained by the proposed AGP method is significantly closer to the true posterior than the results of the BAPE method.

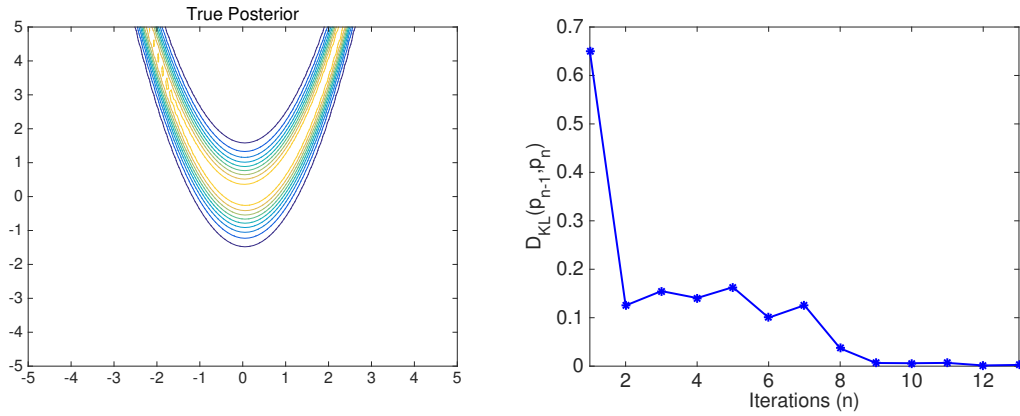


Figure 1: Left: the true posterior distribution. Right: the KL distance between p_{n-1} and p_n , plotted against the number of iterations.

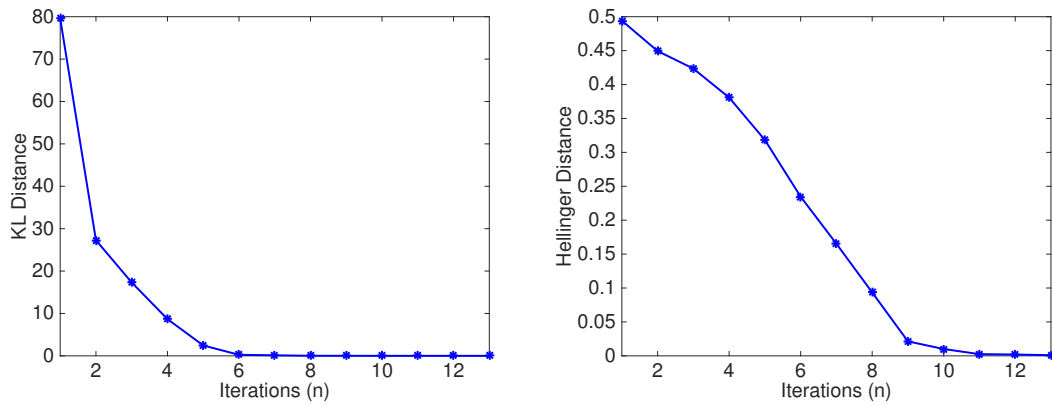


Figure 2: The KL (left) and the Hellinger (right) distances between the obtained approximation and the true posterior, plotted against the number of iterations.

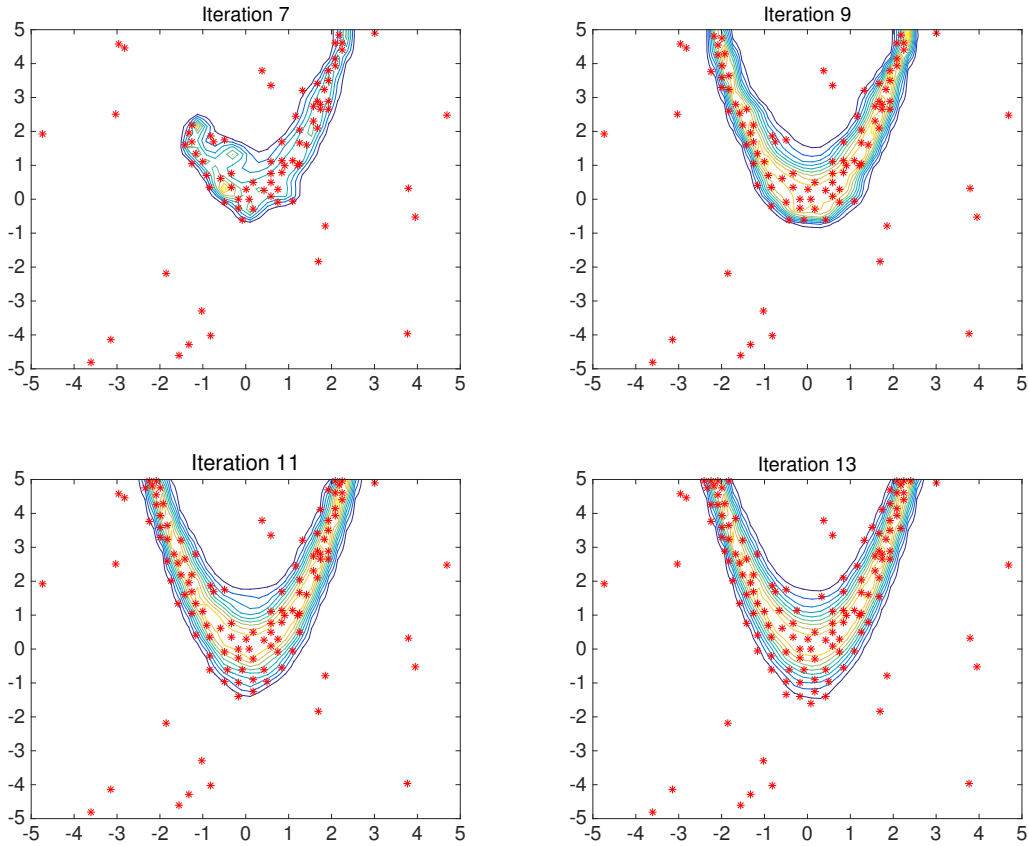


Figure 3: The approximate posterior distribution obtained at the 7th, 9th, 11th and 13th iterations respectively. The red dots are the design points that have been used.

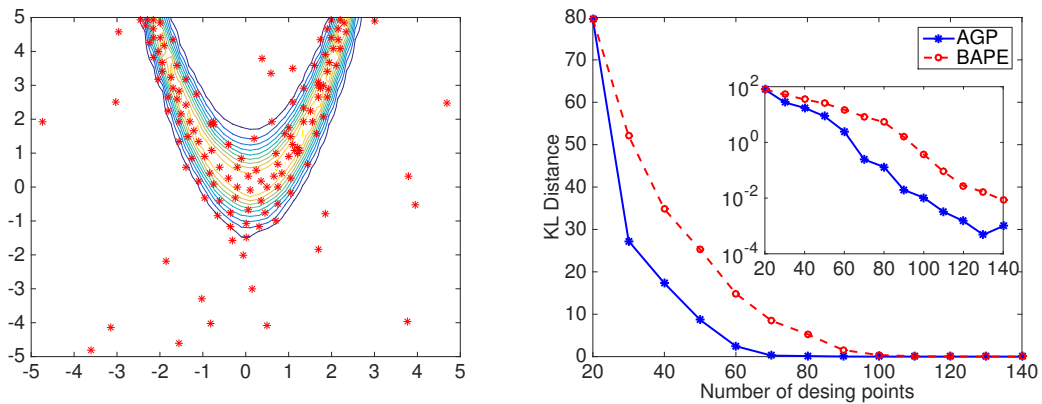


Figure 4: Left: the GP approximation of the posterior distribution obtained with the BAPE method using 140 design points (red dots). Right: the KL distance between the true posterior and the approximation computed with the AGP (solid line) and the BAPE (dashed line) methods, plotted again the number of design points used; the inset is the same plot on a logarithmic scale.

3.2 Genetic toggle switch

We now apply the proposed method to a real-world inference problem. Namely, we consider the kinetics of a genetic toggle switch, which was first studied in [5] and later numerically investigated in [17]. The toggle switch consists of two repressible promoters arranged in a mutually inhibitory network: promoter 1 and promoter 2. Either promoter transcribes a repressor for the other one, and moreover, either repressor may be induced by an external chemical or thermal signal. Genetic circuits of this form can be modeled by the following differential-algebraic equation system [5]:

$$\frac{du}{dt} = \frac{\alpha_1}{1 + v^\beta} - u, \quad (9a)$$

$$\frac{dv}{dt} = \frac{\alpha_2}{1 + w^\gamma} - v, \quad (9b)$$

$$w = \frac{u}{1 + ([\text{IPTG}]/K)^\eta}. \quad (9c)$$

In the equations above, u and v are respectively the concentration of repressors 1 and 2; α_1 and α_2 are the effective rates of synthesis of the repressors; γ and β represent cooperativity of repression of the two promoters; and $[\text{IPTG}]$ is the concentration of IPTG, the chemical compound that induces the switch. Parameters K and η describe binding of IPTG with the first repressor. For more details of the model, we refer to [5].

	α_1	α_2	γ	β	η	K
n	[120, 200]	[15.0, 16.0]	[2.1, 2.9]	[0.85, 1.15]	[1.3, 2.7]	$[2.3, 3.7] \times 10^{-5}$

Table 1: The prior domains of the parameters.

The experiments are performed with several selected values of $[\text{IPTG}]$: 1×10^{-6} , 5×10^{-4} , 7×10^{-4} , 1×10^{-3} , 3×10^{-3} , 5×10^{-3} respectively, and for each experiment, the measurement of v is taken at $t = 10$. The goal is to infer the six parameters

$$\mathbf{x} = [\alpha_1, \alpha_2, \gamma, \beta, \eta, K],$$

from the measurements of v . We use synthetic data in this problem, and specifically we assume that the true values of the parameters are

$$\mathbf{x}_{\text{true}} = [143, 15.95, 2.70, 0.96, 2.34, 2.70 \times 10^{-5}].$$

The data is simulated using the model described by Eqs. (9) with the true parameter values and measurement noise is then added to the simulated data. The measurement noise here is assumed Gaussian and zero-mean, with a variance σ^2 . In the numerical experiments, we consider a large noise case where $\sigma^2 = 5 \times 10^{-4}$ and a small noise case where $\sigma^2 = 1.25 \times 10^{-4}$. We assume that the priors of the six parameters are all uniform and independent of each other, where the domains of the uniform priors are given in Table 1.

We want to use this example to make a detailed comparison of the proposed AGP algorithm with some other popular methods. Thus, we employ four different methods to

compute the posterior distribution in this example: the direct MCMC algorithm with the true likelihood function, the proposed AGP algorithm, the BAPE method [10], and the spectral likelihood expansion (SLE) method [21] which constructs the gPC surrogate for the likelihood function using non-intrusive approaches.

We first consider the large noise case. We draw 3×10^5 samples from the true posterior distribution with a DRAM algorithm and use the results as the reference posterior distribution. We then apply the AGP method to approximate the posterior distribution, where we use $m_0 = 50$ initial design points randomly drawn from the prior and $m = 50$ design points in each iteration. We also choose the termination parameters to be $n_{\max} = 100$, $D_{\max} = 0.05$ and $K = 5$. The algorithm terminates in 18 iterations, resulting in totally 950 evaluations of the true likelihood function. We note that each evaluation of the likelihood function involves a full simulation of the underlying model described by Eqs. (9). After obtaining the approximate posterior distribution, we draw 3×10^5 samples from it using a DRAM MCMC simulation. We then compute the approximate posterior with the BAPE method using also 950 likelihood evaluations, and draw 3×10^5 samples from it using a DRAM MCMC simulation. Finally we approximate the posterior with the SLE-gPC method where the gPC expansion coefficients are computed using the least square method with design points determined by the Sobol sequences based quasi Monte Carlo (QMC) method. The gPC degree is automatically by the algorithm using the leave-one-out (LOO) cross validation, and in the QMC scheme, we set the number of design points to be 950.

We now compare the results of these methods. First we estimate the posterior distributions of the six parameters by all the four methods and show the results in Fig. 5. One can see from the figure that, the distributions computed by both the BAPE and the AGP methods are rather close to those of direct MCMC (which are regarded as the true posteriors), while the results of SLE-gPC deviate evidently from the MCMC results, especially for the two parameters α_2 and β . As for the comparison of BAPE and AGP, the figures show that both methods can produce reasonably good approximations of the posterior distributions in this case. So for a quantitative evaluation for the performance of the methods, we compute the KLD from the approximate posterior distributions to the true posterior densities, and show the results in Table 2. We can see from the table that, the posterior distributions computed by the AGP method are closer (in terms of KLD) to the true posteriors than the other two methods for all the 6 parameters.

We then consider the small noise case where we use the same implementation configurations as the large noise case. In this case, the AGP algorithm uses 1200 true likelihood evaluations, and as before, we also compute the posterior using the SLE and the BAPE methods with the same number of true likelihood evaluations. We then compute the posterior directly with 3×10^5 MCMC samples and use the results as the true posterior. In Figs. 6 we compare the marginal posterior distributions compared with the four methods. Similar to the large noise case, the figures show that the results of the SLE-gPC are of very low accuracy, while both the AGP and the BAPE methods yield rather good results. Once again, we show in Table 2 the KLD from the marginal posterior distributions computed with the three approximate methods to the true posterior (those computed by the direct MCMC). These quantitative comparison results indicate that, the AGP method yields better results than the other two methods in terms of the KLD. Thus, we can conclude that our AGP method has the best performance in both

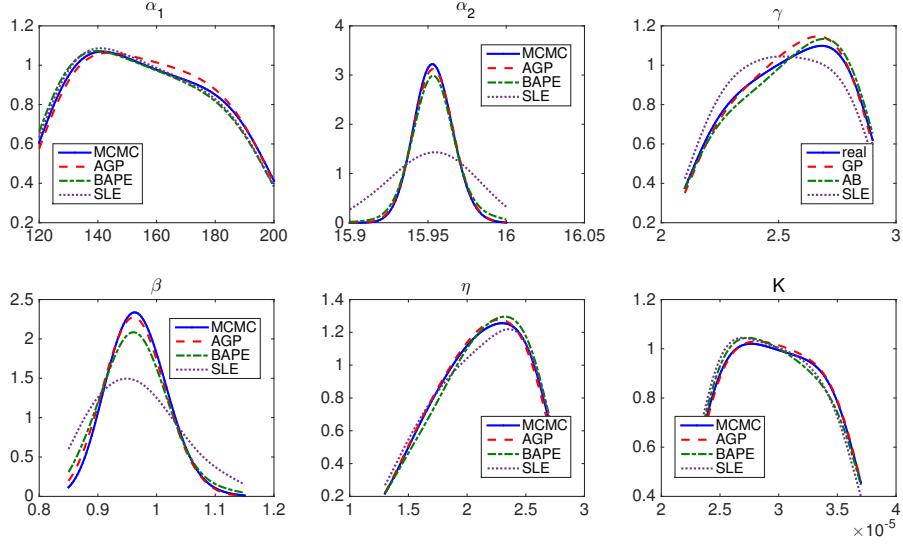


Figure 5: The genetic toggle example: large noise case. The marginal distributions of the six parameters, computed with the four different methods.

the large and the small noise cases.

	method	α_1	α_2	γ	β	η	K
KLD	AGP	1.5×10^{-4}	0.0069	0.0020	0.014	0.0041	0.0075
	BAPE	4.5×10^{-3}	0.043	0.0022	0.039	0.0095	0.013
	SLE	1.8×10^{-3}	0.44	0.0012	0.18	0.0078	0.0013

Table 2: The large noise case: the KLD between the marginal posterior distributions computed with the three approximate methods and those computed with standard MCMC.

	method	α_1	α_2	γ	β	η	K
KLD	AGP	0.0032	0.015	0.0039	0.012	0.0096	0.014
	BAPE	0.015	0.036	0.028	0.057	0.0075	0.074
	SLE	0.0035	1.1	0.021	0.46	0.050	0.010

Table 3: The small noise case: the KLD between the marginal posterior distributions computed with the three approximate methods and those computed with standard MCMC.

3.3 The human body sway problem

Finally, we apply the proposed method to a human body sway problem. This problem has received considerable attention as the body sway may provide information about the physiological status of a person [34]. Several mathematical models have been proposed

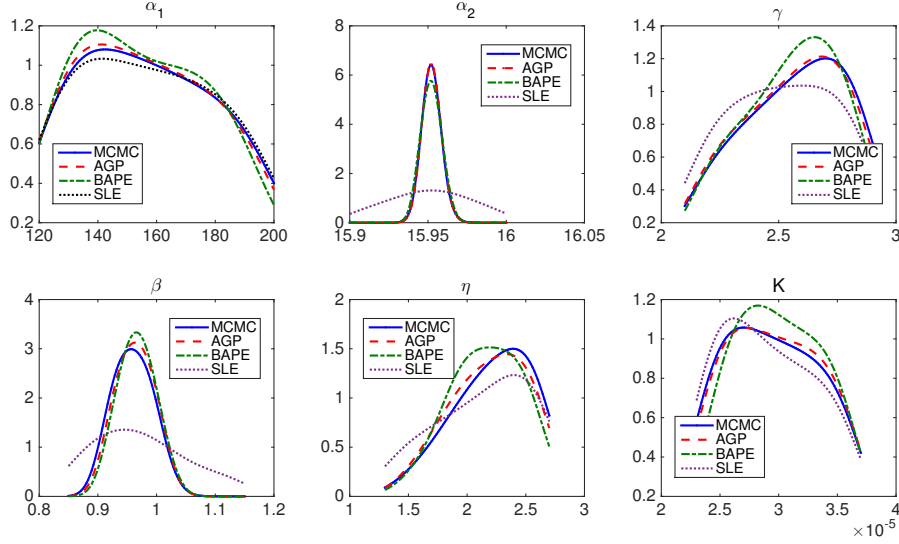


Figure 6: The genetic toggle example: small noise case. The marginal distributions of the six parameters, computed with the four different methods.

to describe the sway motion, and here we consider the single-link inverted pendulum (SLIP) model proposed in [2], which assumes that the body is maintained in an upright position by an active and a passive proportional-derivative controller.

Specifically, the SLIP model is given by the following stochastic delay differential equation (SDDE) [2]:

$$I\ddot{\theta}(t) = mgh\theta(t) - [K\theta(t) + B\dot{\theta}(t) + f_P(\theta(t - \Delta)) + f_D(\dot{\theta}(t - \Delta))] + \xi(t). \quad (10)$$

In this equation, I is the moment of inertia of the body, θ is the tilt angle ($\dot{\theta}$ and $\ddot{\theta}$ are its first and second derivatives respectively), m denotes the body mass, g is the gravitational acceleration, h is the distance between 3D center-of-mass (COM) and the ankle joint, and ξ is a zero-mean Gaussian noise with variance σ^2 . K and P are the passive stiffness and passive damping parameters, and $f_P(\theta(t - \Delta))$ and $f_D(\dot{\theta}(t - \Delta))$ are active stiffness and active damping terms where Δ is the time delay. We now specify the active stiffness $f_P(\theta(t - \Delta))$ and the active damping $f_D(\dot{\theta}(t - \Delta))$. We first define two functions $c_1(\theta(t - \Delta)) = \theta(\dot{\theta}(t - \Delta) - a_s\theta(t - \Delta))$ and $c_2(\theta(t - \Delta)) = \theta(t - \Delta)^2 + (\dot{\theta}(t - \Delta))^2$. We then have,

$$f_P(\theta(t - \Delta)) = \begin{cases} P\theta(t - \Delta), & \text{if } c_1(\theta(t - \Delta)) > 0 \text{ and } c_2(\theta(t - \Delta)) > r^2; \\ 0, & \text{otherwise;} \end{cases} \quad (11)$$

and

$$f_D(\dot{\theta}(t - \Delta)) = \begin{cases} D\dot{\theta}(t - \Delta) & \text{if } c_1(\theta(t)) > 0 \text{ and } c_2(\theta(t)) > r^2; \\ 0, & \text{otherwise;} \end{cases} \quad (12)$$

where r is the radius of the “quiet zone” (active control is off). The slope a_s depends on the level of control, C_{ON} , as $a_s = -\tan(\pi(C_{ON} - 0.5))$. In this model, five key parameters (P , D , Δ , σ , and C_{ON}) can not be measured directly and need to be inferred

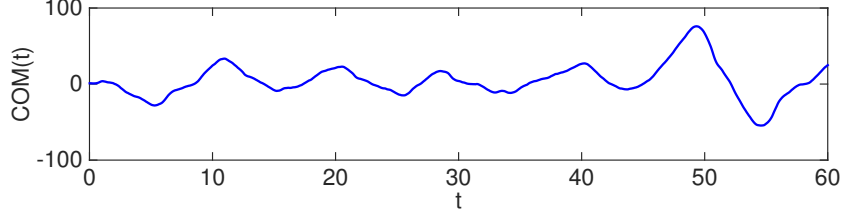


Figure 7: The simulated COM signal.

from the body sway measurements, while the other parameters can either be measured or specified in advance [34]. The COM signal,

$$\text{COM}(t) = h \sin(\theta(t)), \quad (13)$$

is measured and used to infer the five unknown model parameters. The absolute value of the COM amplitude, velocity, acceleration and the power spectral density (PSD) are extracted from the signal. The mean, variance, skewness and kurtosis of each physical quantity are calculated as the data \mathbf{y} (a 16-dimensional vector) to infer the model parameters: $\mathbf{x} = (P, D, \Delta, \sigma, C_{ON})$. Computing the posterior in this problem is rather challenging as the likelihood function $p(\mathbf{y}|\mathbf{x})$ is not available, which make the standard Bayesian inference computation methods such as the MCMC algorithms infeasible. In [34], the parameters were inferred with the approximate Bayesian computation (ABC) method which does not use the likelihood function. Here we compute the posterior with the proposed method. In particular, we compute the likelihood function using the following procedure. For a given parameter value \mathbf{x} , we perform a Monte Carlo simulation for the SDDE model (10) with a given number of samples. For each y_i for $i = 1 \dots 16$, we estimate resulting conditional density function $p_i(y_i|\mathbf{x})$ with the kernel density estimation method, and then we take the likelihood function to be

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{16} p_i(y_i|\mathbf{x}).$$

We note that a single evaluation of the likelihood function requires to repeatedly simulate Eq. (10) a large number of times, which renders the evaluation highly intensive.

In the numerical experiments, we use simulated data, and in particular the true parameters values are set to be $P = 145 \text{ Nm/rad}$, $D = 10 \text{ Nms/rad}$, $\Delta = 0.2 \text{ s}$, $\sigma = 0.45 \text{ Nm}$ and $C_{ON} = 0.75$, and the other parameter values are $g = 9.81 \text{ m/s}^2$, $m = 68 \text{ kg}$, $h = 0.87 \text{ m}$, $I = mh^2$, $K = mgh \times 0.8 \text{ Nm/rad}$, $B = 4 \text{ Nms/rad}$ and $r = 0.004 \text{ rad-rad/s}$. The COM signal generated from the model Eq. (10) with these parameter values is shown in Fig. 7. In the inference, we impose a uniform distribution on each of the five parameters on the following intervals: $P \in [80, 160]$, $D \in [0.05, 30]$, $\Delta \in [0.05, 0.5]$, $\sigma \in [0.05, 0.6]$, and $C_{ON} \in [0.05, 0.85]$. Moreover, for each evaluation of the likelihood function $\pi(\mathbf{y}|\mathbf{x})$, we use 10,000 simulations of Eq. (10), and as a result a direct MCMC simulation of the posterior distribution is computationally infeasible. We apply our AGP method to compute the posterior distribution, and the algorithm parameters are the same as those in the second example. The algorithm terminates in 14 iterations and so total number of true likelihood evaluations is 750. As

a comparison, we also perform the BAPE method with the same number of true likelihood function evaluations. To compare the performance of the two methods, we plot the posterior marginals of the model parameters computed by BAPE in Fig. 8, and those computed by AGP in Fig. 9. We also show the true parameter values as well as the 60% confidence interval in the figures. Here one can see that, for all the posteriors computed with the AGP method, the true parameter values fall in the 60% confidence intervals, while for the results of the BAPE method, the true values of D and Δ fall outside of the 60% confidence intervals, which suggest that the posteriors computed by the AGP method may be more accurate and reliable than those by BAPE.

4 Conclusions

In summary, we have proposed an algorithm to construct GP based approximation for the un-normalized posterior distribution. The method expresses the un-normalized posterior as a product of an approximate posterior density and an exponentiated GP model, and an adaptive scheme is presented to construct such an approximation. We also provide an active learning method that uses maximum entropy as the selection criterion to determine the sampling points. With numerical examples, we show that the method can obtain a rather good approximation of the posterior with a limited number of evaluations of the likelihood functions. We believe the proposed method can be useful in a wide range of practical Bayesian inference problems where the likelihood function are difficult or expensive to evaluate.

Several issues of the proposed algorithm deserve further studies. First, while our numerical experiments illustrate that the algorithm may converge in these examples, a rigorous convergence analysis of the algorithm is still lacking. Secondly, for a posterior distribution with unbounded domain, the resulting approximation may become improper, and thus certain modifications of the algorithm may be needed to address the issue. Finally we note that selecting a good Kernel function for the GP model is a very important issue for GP-based methods, and to this end, a very interesting question is how to choose kernel functions that are specifically suitable for the log-posteriors. We plan to study these issues in future works.

Acknowledgments

The work was partially supported by the National Natural Science Foundation of China under grant number 111771289.

References

- [1] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50(1-2):5–43, 2003.
- [2] Yoshiyuki Asai, Yuichi Tasaka, Kunihiko Nomura, Taishin Nomura, Maura Casadio, and Pietro Morasso. A model of postural control in quiet standing: Robust

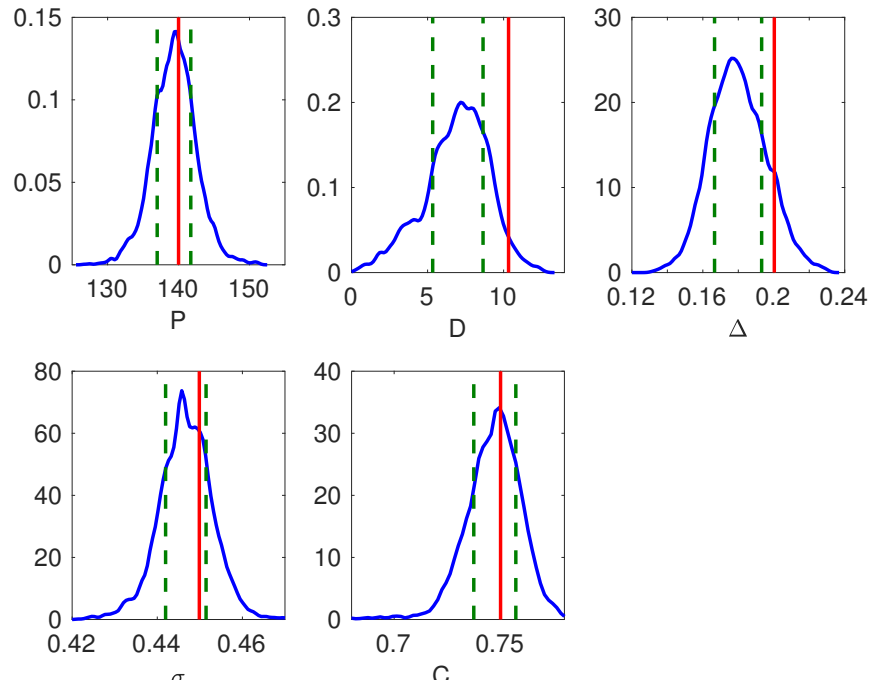


Figure 8: The posteriors of the parameters in the SLIP model, computed by the BAPE method. Also shown in the figures are the 60% confidence interval (dashed vertical lines) and the true parameter values (solid vertical lines).

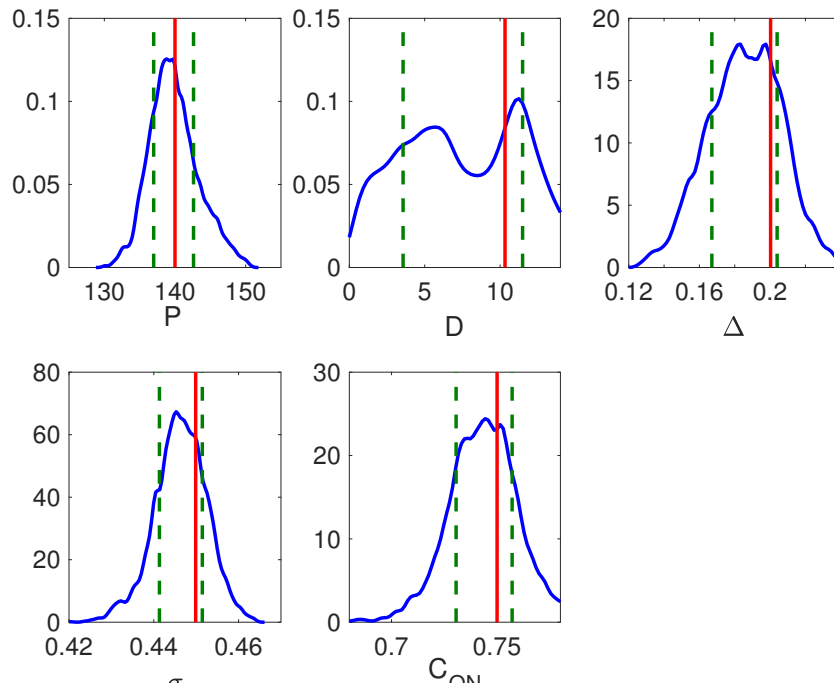


Figure 9: The posteriors of the parameters in the SLIP model, computed by our AGP method. Also shown in the figures are the 60% confidence interval (dashed vertical lines) and the true parameter values (solid vertical lines).

- compensation of delay-induced instability using intermittent activation of feedback control. *Plos One*, 4(7):e6169, 2009.
- [3] I Billionis and N Zabararas. Solution of inverse problems with limited forward solver evaluations: a bayesian perspective. *Inverse Problems*, 30(1):015004, 2013.
- [4] Patrick R Conrad, Youssef M Marzouk, Natesh S Pillai, and Aaron Smith. Accelerating asymptotically exact mcmc for computationally intensive models via local approximations. *Journal of the American Statistical Association*, 111(516):1591–1607, 2016.
- [5] Timothy S Gardner, Charles R Cantor, and James J Collins. Construction of a genetic toggle switch in escherichia coli. *Nature*, 403(6767):339–342, 2000.
- [6] Andrew Gelman, John B Carlin, Hal S Stern, David Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis (3rd edition)*. Chapman & Hall/CRC, 2013.
- [7] Andrew Golightly and Darren J Wilkinson. Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface focus*, 1(6):807–820, 2011.
- [8] Alex Gorodetsky and Youssef Marzouk. Mercer kernels and integrated variance experimental design: connections between gaussian process regression and polynomial approximation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):796–828, 2016.
- [9] Heikki Haario, Marko Laine, Antonietta Mira, and Eero Saksman. DRAM: efficient adaptive MCMC. *Statistics and Computing*, 16(4):339–354, 2006.
- [10] Kirthevasan Kandasamy, Jeff Schneider, and Barnabás Póczos. Bayesian active learning for posterior estimation. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 3605–3611. AAAI Press, 2015.
- [11] Marc Kennedy. Bayesian quadrature with non-normal approximating functions. *Statistics and Computing*, 8(4):365–375, 1998.
- [12] Scott Kirkpatrick, C Daniel Gelatt, Mario P Vecchi, et al. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.
- [13] Bledar A Konomi, Georgios Karagiannis, Kevin Lai, and Guang Lin. Bayesian treed calibration: an application to carbon capture with ax sorbent. *Journal of the American Statistical Association*, in press.
- [14] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(Feb):235–284, 2008.
- [15] Jinglai Li and Youssef M Marzouk. Adaptive construction of surrogates for the bayesian solution of inverse problems. *SIAM Journal on Scientific Computing*, 36(3):A1163–A1186, 2014.

- [16] David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.
- [17] Youssef Marzouk and Dongbin Xiu. A stochastic collocation approach to bayesian inference in inverse problems. *Communications in Computational Physics*, 6(4):826–847, 2009.
- [18] Youssef M Marzouk and Habib N Najm. Dimensionality reduction and polynomial chaos acceleration of bayesian inference in inverse problems. *Journal of Computational Physics*, 228(6):1862–1902, 2009.
- [19] Youssef M Marzouk, Habib N Najm, and Larry A Rahn. Stochastic spectral methods for efficient bayesian solution of inverse problems. *Journal of Computational Physics*, 224(2):560–586, 2007.
- [20] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.
- [21] Joseph B Nagel and Bruno Sudret. Spectral likelihood expansions for bayesian inference. *Journal of Computational Physics*, 309:267–294, 2016.
- [22] Jeremy Oakley and Anthony O’Hagan. Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika*, 89(4):769–784, 2002.
- [23] Anthony O’Hagan. Bayes–Hermite quadrature. *Journal of statistical planning and inference*, 29(3):245–260, 1991.
- [24] Anthony O’Hagan. Polynomial chaos: A tutorial and critique from a statistician’s perspective. *SIAM/ASA Journal on Uncertainty Quantification Uncertainty Quantification*, 20:1–20, 2013.
- [25] Anthony O’Hagan and JFC Kingman. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–42, 1978.
- [26] Michael Osborne, Roman Garnett, Zoubin Ghahramani, David K Duvenaud, Stephen J Roberts, and Carl E Rasmussen. Active learning of model evidence using Bayesian quadrature. In *Advances in neural information processing systems*, pages 46–54, 2012.
- [27] Carl Edward Rasmussen, JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West. Gaussian processes to speed up hybrid monte carlo for expensive bayesian integrals. In *Bayesian Statistics 7*, pages 651–659, 2003.
- [28] Carl Edward Rasmussen and Zoubin Ghahramani. Bayesian Monte Carlo. *Advances in neural information processing systems*, pages 505–512, 2003.

- [29] Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.
- [30] Jerome Sacks, William J Welch, Toby J Mitchell, and Henry P Wynn. Design and analysis of computer experiments. *Statistical science*, pages 409–423, 1989.
- [31] Paola Sebastiani and Henry P Wynn. Maximum entropy sampling and optimal Bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):145–157, 2000.
- [32] Claude E Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.
- [33] Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.
- [34] A. Tietäväinen, M. U. Gutmann, E. Keskinvakkuri, J. Corander, and E. Häggström. Bayesian inference of physiologically meaningful parameters from body sway measurements. *Scientific Reports*, 7(1):3771, 2017.
- [35] Christopher KI Williams and Carl Edward Rasmussen. Gaussian processes for machine learning. *the MIT Press*, 2(3):4, 2006.