Adaptive Multiple Importance Sampling for Gaussian Processes

Xiaoyu Xiong[a*] and Václav Šmídl[b] and Maurizio Filippone[c]

[a] *School of Computing Science, University of Glasgow, UK*
[b] *Institute of Information Theory and Automation, Czech Academy of Sciences, Czech Republic*
[c] *Department of Data Science, EURECOM, France*

Abstract: In applications of Gaussian processes where quantification of uncertainty is a strict requirement, it is necessary to accurately characterize the posterior distribution over Gaussian process covariance parameters. This is normally done by means of standard Markov chain Monte Carlo (MCMC) algorithms, which require repeated expensive calculations involving the marginal likelihood. Motivated by the desire to avoid the inefficiencies of MCMC algorithms rejecting a considerable amount of expensive proposals, this paper develops an alternative inference framework based on Adaptive Multiple Importance Sampling (AMIS). In particular, this paper studies the application of AMIS for Gaussian processes in the case of a Gaussian likelihood, and proposes a novel Pseudo-Marginal-based AMIS algorithm for non-Gaussian likelihoods, where the marginal likelihood is unbiasedly estimated. The results suggest that the proposed framework outperforms MCMC-based inference of covariance parameters in a wide range of scenarios.

Keywords: Gaussian processes; Bayesian inference; Markov chain Monte Carlo; Importance sampling

1. Introduction

Gaussian Processes (GPs) have proved to be a successful class of statistical inference methods for data analysis in several applied domains, such as pattern recognition [1–3], neuroimaging [4], signal processing [5], Bayesian optimization [6], and emulation and calibration of computer codes [7]. GP models are attractive due to their nonparametric formulation that yields the possibility to flexibly model data; in addition, with a suitable parameterization, they offer the possibility to gain some insights into the application under study. These properties hinge on the parameterization of the GP covariance function and on the way GP covariance parameters are optimized or inferred.

It is acknowledged that optimizing covariance parameters can severely affect the ability of GP models to quantify uncertainty in predictions [3, 4, 8, 9]. Therefore, in applications where this is undesirable, it is necessary to accurately characterize the posterior distribution over covariance parameters and propagate this source of uncertainty forward to predictions. This task, which is the focus of this work, is particularly challenging when dealing with GPs. Inference of GP covariance parameters in closed form is generally analytically intractable, and when resorting to standard inference methods a complication arises from the difficulties associated with having to repeatedly compute the so called marginal likelihood (and possibly the gradient of its logarithm). The marginal likelihood is computable in the case of a Gaussian likelihood, but extremely costly due to the need

*Corresponding author. Email: x.xiong.1@research.gla.ac.uk

to carry out a number of operations that is cubic with the number of input vectors. On the other hand, when the likelihood function is not Gaussian, e.g., in classification, in ordinal regression, or in Cox-processes, the marginal likelihood is not even computable analytically.

In response to these challenges, a large body of the literature develops approximate inference methods [1, 10–14] which, although successful in many cases, give no guarantees on the amount of bias they introduce. With regards to quantifying uncertainty without introducing any bias, there have been attempts to employ Markov chain Monte Carlo (MCMC) techniques; we can broadly divide such attempts in works that propose reparameterization techniques [8, 15–17], or methods that carry out inference based on unbiased computations of the marginal likelihood [3, 18, 19]. Although these approaches proved successful in a variety of scenarios, employing MCMC algorithms may lead to inefficiencies; for instance, optimal acceptance rates for popular MCMC algorithms such as the Metropolis-Hastings (MH) algorithm (around 25% [20]) and the Hybrid Monte Carlo (HMC) algorithm (about 65% [21, 22]) indicate that several expensive computations are wasted. Introducing adaptivity into MCMC proposal mechanisms to improve efficiency may lead to convergence issues [23].

In this paper we develop a general framework to carry out Bayesian inference for GPs aimed at overcoming the aforementioned limitations of MCMC methods, where expectations under the posterior distribution over covariance parameters are carried out by means of the Adaptive Multiple Importance Sampling (AMIS) algorithm [24]. The application of this framework to the Gaussian likelihood case, although novel, is relatively straightforward given that the likelihood is computable. In the case of non-Gaussian likelihoods, the inability to compute the likelihood exactly motivates us to propose a novel version of AMIS where the likelihood is unbiasedly estimated. Inspired by the Pseudo-Marginal MCMC approaches [25], we propose the Pseudo-Marginal AMIS (PM-AMIS) algorithm, and provide a theoretical analysis showing under which conditions PM-AMIS yields expectations under the posterior over GP covariance parameters without introducing any bias. The proposed PM-AMIS is an instance of the Importance Sampling squared (IS^2) algorithms [26, 27] that are gaining popularity as practical Bayesian inference methods.

In summary, the main contributions of this work are: (i) the application of AMIS to infer GP covariance parameters with any likelihood; (ii) a theoretical analysis of PM-AMIS; (iii) an extensive comparison of convergence speed with respect to computational complexity of AMIS versus MCMC methods. Table 1 illustrates where our work fits in the literature of Bayesian inference for GP covariance parameters and beyond.

The results demonstrate the value of our proposal. In particular, the results indicate that AMIS is competitive with MCMC algorithms in terms of convergence speed over computational cost. Furthermore, the results suggest that AMIS is a valid alternative to MCMC algorithms even in the case of moderately large dimensional parameter spaces, which is common when employing richly parameterized covariances (e.g., Automatic Relevance Determination (ARD) covariances [28]). Given that importance sampling-based inference methods, unlike MCMC algorithms, are inherently parallel, the results suggest a promising direction to speed up inference of GP covariance parameters.

The paper is organized as follows. In section 2 we provide a brief overview of GP regression and Bayesian inference. Section 3 presents the proposed Adaptive Multiple Importance Sampling for Gaussian Processes. Section 4 reports the experiments and the results, and section 5 reports our conclusions and suggestions for future work.

Table 1. Schematic representation of where the proposed contribution fits within the literature of inferring GP covariance parameters. In this work, we propose AMIS for Gaussian processes and PM-AMIS and study its application to Gaussian processes; the latter can be employed whenever an unbiased estimate of the (marginal) likelihood is available. The list of references is not exhaustive but illustrates some of the key works and reviews in this field.

Inference	Models	(Marginal) Likelihood		Reparameterizations
		Computable	Estimated	
MCMC	Others	[29]	[25]	[30]
	GPs	[8]	[3]	[17]
AMIS	Others	[24]	PM-AMIS	–
	GPs	AMIS for GPs		–

2. Bayesian Gaussian Processes

2.1. *Gaussian Processes*

Consider a supervised learning scenario. Let $\mathbf{X}$ be a set of n input vectors $\mathbf{x}_i \in \mathbb{R}^d (1 \leq i \leq n)$, and let $\mathbf{y}$ be the vector consisting of the corresponding labels y_i. In most GP models, the labels are assumed to be conditionally independent given a set of n latent variables. Such latent variables are modelled as realizations of a function $f(\mathbf{x})$ at the input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, i.e., $\mathbf{f} = \{f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)\}$. Latent variables are used to express the likelihood function, which under the assumption of independence becomes $p(\mathbf{y} \mid \mathbf{f}) = \prod_{i=1}^{n} p(y_i \mid f_i)$, where $p(y_i \mid f_i)$ depends on the data being modelled (e.g., Gaussian for regression, Bernoulli for probit classification with probability $P(y_i = 1) = \Phi(f(\mathbf{x}_i))$ where Φ is defined as the cumulative normal distribution).

What characterizes GP models is the way the latent variables are specified. In particular, we assume that the function $f(\mathbf{x})$ is distributed as a GP, which implies that the latent function values $\mathbf{f}$ are jointly distributed as a Gaussian $\mathbf{f} \mid \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$, where $\mathbf{K}$ is the covariance matrix. The entries of the covariance matrix $\mathbf{K}$ are specified by a covariance (kernel) function with hyperparameters $\boldsymbol{\theta}$ between pairs of input vectors. In this work, two different covariance functions are considered. The first is the RBF (Radial Basis Function) covariance defined as:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma \exp\left\{ -\frac{1}{\tau^2} \parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2 \right\} \tag{1}$$

The parameter τ defines the length-scale of the interaction between the input vectors, while σ represents the marginal variance for each latent variable. The second is the ARD covariance, which takes the form:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma \exp\left\{ -\sum_{r=1}^{d} \frac{1}{{\tau_r}^2} (\mathbf{x}_{i(r)} - \mathbf{x}_{j(r)})^2 \right\} \tag{2}$$

The advantage of the ARD covariance is that it accounts for the influence of each feature on the mapping between inputs and labels, with smaller values of parameters $(\tau_1, ..., \tau_d)$ indicating a higher influence of the corresponding features [5]. For simplicity of notation,

in the remainder of the paper we will denote by $\boldsymbol{\theta}$ the vector of all covariance parameters.

When making predictions, using a point estimate of $\boldsymbol{\theta}$ has been reported to potentially underestimate the uncertainty in predictions or yield inaccurate assessment of the relative influence of different features [2–4]. Therefore, a Bayesian approach is usually adopted to overcome these limitations, which entails characterizing the posterior distribution over covariance parameters. In order to do so, it is necessary to employ methods, such as MCMC, that require computing the marginal likelihood every time $\boldsymbol{\theta}$ is updated. We now discuss the challenges associated with the computation of the marginal likelihood for the special case of a Gaussian likelihood and the more general case of non-Gaussian likelihoods.

2.1.1. Gaussian likelihood

In the GP regression setting, the observations $\mathbf{y}$ are modeled to be Gaussian distributed with mean of $\mathbf{f}$ (latent variables) and covariance $\lambda\mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix, and λ is the variance of the Gaussian noise on the observations. In this setting, the likelihood $p(\mathbf{y} \mid \mathbf{f})$ and the GP priors $p(\mathbf{f} \mid \boldsymbol{\theta})$ form a conjugate pair, so latent variables can be integrated out of the model leading to $\mathbf{y} \mid \boldsymbol{\theta} \sim \mathcal{N}(0, \mathbf{C})$, where $\mathbf{C} = \mathbf{K} + \lambda\mathbf{I}$. This yields the log-marginal likelihood

$$\log[p(\mathbf{y}|\boldsymbol{\theta})] = -\frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}\mathbf{y}^\top\mathbf{C}^{-1}\mathbf{y} + \text{const.}$$

in closed form. Although computable, the log-marginal likelihood requires computing the log determinant of $\mathbf{C}$ and solving a linear system involving $\mathbf{C}$. These calculations are usually carried out by factorizing the matrix $\mathbf{C}$ using the Cholesky decomposition, giving $\mathbf{C} = \mathbf{L}\mathbf{L}^\top$, with $\mathbf{L}$ being lower triangular. The Cholesky algorithm requires $O(n^3)$ operations, but subsequently computing the terms of the marginal likelihood requires at most $O(n^2)$ operations [1].

2.1.2. Non-Gaussian likelihoods

In the case of non-Gaussian likelihoods, the likelihood $p(\mathbf{y} \mid \mathbf{f})$ and the GP prior $p(\mathbf{f} \mid \boldsymbol{\theta})$ are no longer conjugate. As a consequence, it is not possible to solve the integral needed to integrate out latent variables

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})d\mathbf{f}$$

and this requires some form of approximation. A notable example is GP probit classification, which is what we explore in detail in this paper. In this case, the observations $\mathbf{y}$ are assumed to be Bernoulli distributed with success probability given by [1]:

$$p(y_i \mid f_i) = \Phi(y_i f_i) \tag{3}$$

For GPs with non-Gaussian likelihoods, there have been several proposals on how to carry out approximation to integrate out the latent variables, or to avoid approximations altogether. The focus of this paper is on methods that do not introduce any bias in the calculation of expectation under the posterior over covariance parameters, and we will discuss these approaches in detail in the next sections.

2.2. *Bayesian inference of covariance parameters*

For simplicity of notation, we denote the posterior distribution over covariance parameters by

$$\pi(\boldsymbol{\theta}) := p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{4}$$

where $p(\boldsymbol{\theta})$ encodes any prior knowledge on the parameters $\boldsymbol{\theta}$. Within the Bayesian framework, we are usually interested in calculating expectations of functions of $\boldsymbol{\theta}$ with respect to the posterior distribution, i.e., $E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})]$. For instance, setting $h(\boldsymbol{\theta}) = p(y_\star \mid \boldsymbol{\theta}, \mathbf{x}_\star, \mathbf{y}, \mathbf{X})$, we obtain the predictive distribution for the label $y_\star$ associated with a new input vector $\mathbf{x}_\star$.

The denominator needed to normalize the posterior distribution $\pi(\boldsymbol{\theta})$ is intractable, so it is not possible to characterize the posterior distribution analytically. Despite this complication, it is possible to resort to a Monte Carlo approximation to compute expectations under the posterior distribution of $\boldsymbol{\theta}$

$$E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})] \simeq \frac{1}{N} \sum_{i=1}^{N} h(\boldsymbol{\theta}^{(i)}) \tag{5}$$

where $\boldsymbol{\theta}^{(i)}$ denotes the ith of N samples from $\pi(\boldsymbol{\theta})$. However, as it is generally not feasible to draw samples from $\pi(\boldsymbol{\theta})$ directly, it is necessary to resort to MCMC methods to generate samples from the posterior $\pi(\boldsymbol{\theta})$.

An alternative way to compute expectations is by means of importance sampling, which takes the following form:

$$E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})] = \int h(\boldsymbol{\theta}) \frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{6}$$

where $q(\boldsymbol{\theta})$ is the importance distribution. The corresponding Monte Carlo approximation is of the form:

$$E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})] \simeq \frac{1}{N} \sum_{i=1}^{N} h(\boldsymbol{\theta}^{(i)}) \frac{\pi(\boldsymbol{\theta}^{(i)})}{q(\boldsymbol{\theta}^{(i)})} \tag{7}$$

where the samples $\boldsymbol{\theta}^{(i)}$ are now drawn from the importance sampling distribution $q(\boldsymbol{\theta})$. The key to make this Monte Carlo estimator accurate is to choose $q(\boldsymbol{\theta})$ to be similar to the function that needs to be integrated, that is $h(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. It is easy to verify that when this is the case, the variance of the importance sampling estimator is zero. Therefore, the success of importance sampling relies on constructing a tractable importance distribution $q(\boldsymbol{\theta})$ that well approximates $h(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. In the remainder of this paper we study and evaluate methods that adaptively construct $q(\boldsymbol{\theta})$.

Both Monte Carlo approximations in equations (5) and (7) converge to the desired expectation, and in practice, they can estimate the desired integral to a given level of precision [31, 32]. The experimental part of this work is devoted to the study of the convergence properties of the expectation $E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})]$ with respect to the computational effort needed to carry out the Monte Carlo approximations in Equations (5) and (7).

2.3. *Pseudo-Marginal MCMC for inference of covariance parameters*

Standard MCMC algorithms to draw from the posterior $\pi(\boldsymbol{\theta})$ require calculating the marginal likelihood and the gradient of its logarithm exactly. When the likelihood is not Gaussian, computing the expectation in equation (5) becomes unfeasible because of the inability to calculate the marginal likelihood exactly. In cases where the marginal likelihood can be unbiasedly estimated, it is possible to resort to so called Pseudo-Marginal MCMC approaches. Taking the Metropolis-Hastings algorithm as an example, it is possible to replace the exact calculation of the Hastings ratio

$$\frac{p(\mathbf{y} \mid \boldsymbol{\theta}')p(\boldsymbol{\theta}')}{p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}$$

with an approximation where the marginal likelihood is unbiasedly estimated

$$\frac{\tilde{p}(\mathbf{y} \mid \boldsymbol{\theta}')p(\boldsymbol{\theta}')}{\tilde{p}(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}$$

where $\tilde{p}(\mathbf{y} \mid \boldsymbol{\theta})$ denotes such an approximation. Interestingly, the introduction of this approximation does not affect the properties of the MCMC approach that still yields samples from the correct posterior $\pi(\boldsymbol{\theta})$. The effect of introducing this approximation, however, is that the efficiency of the corresponding MCMC approach is reduced; this is due to the possibility that the algorithm accepts a proposal with a largely overestimated value of the marginal likelihood, making it difficult for any new proposals to be accepted.

By inspecting the GP marginal likelihood

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \int p(\mathbf{y} \mid \mathbf{f})p(\mathbf{f} \mid \boldsymbol{\theta})d\mathbf{f} \tag{8}$$

we observe that we can attempt to unbiasedly estimate this integral using importance sampling:

$$\tilde{p}(\mathbf{y} \mid \boldsymbol{\theta}) \simeq \frac{1}{N_{\text{imp}}} \sum_{i=1}^{N_{\text{imp}}} \frac{p(\mathbf{y} \mid \mathbf{f}_i)p(\mathbf{f}_i \mid \boldsymbol{\theta})}{q(\mathbf{f}_i \mid \mathbf{y}, \boldsymbol{\theta})} \tag{9}$$

Here N_{imp} is the number of samples $\mathbf{f}_i$ drawn from the importance density $q(\mathbf{f} \mid \mathbf{y}, \boldsymbol{\theta})$. The motivation for attempting this approximation is to leverage the various successful attempts that construct accurate approximations to the posterior distribution over the latent variables $p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta}) \propto p(\mathbf{y} \mid \mathbf{f})p(\mathbf{f} \mid \boldsymbol{\theta})$. The accuracy of the approximations to the posterior over latent variables directly affects the accuracy of the importance sampling estimates of the marginal likelihood. Despite introducing some noise in the calculation of the Hastings ratio, the resulting MCMC approach has been shown to yield state-of-the-art performance in sampling from the posterior over GP covariance parameters [3]. In this paper, we investigate approximations $q(\mathbf{f} \mid \mathbf{y}, \boldsymbol{\theta})$ to the posterior obtained by the Laplace Approximation (LA) and Expectation Propagation (EP) algorithms [1, 12].

3. Adaptive Multiple Importance Sampling for Gaussian Processes

Inefficiencies arising from the use of MCMC methods to sample from the posterior distribution over covariance parameters are due to the fact that several proposals are rejected. To mitigate this issue, some adaptation mechanisms of the proposals can be used based on previous MCMC samples, but the chain resulting from the adaptivity is no longer Markovian. As a result, elaborate ergodicity results are needed to establish convergence to the true posterior distribution [23, 33, 34].

In response to this, Cappe et al. [35] proposed a universal adaptive sampling scheme called Population Monte Carlo (PMC), where the difference from Sequential Monte Carlo (SMC) [36] is that the target distribution becomes static. This method is reported to have better adaptivity than MCMC since the use of importance sampling removes the issue of ergodicity. At each iteration of PMC, the Sampling Importance Resampling (SIR) [37] particle filter is used to generate samples that are assumed to be marginally distributed from the target distribution and hence, the approach is unbiased and can be stopped at any time. Moreover, the importance distribution can be adapted using part (generated at each iteration) or all of the importance sample sequence. Douc et al. [38, 39] also introduced updating mechanisms for the weights of the mixture in the so called D-kernel PMC, which lead to a reduction either in Kullback divergence between the mixture and the target distribution or in the asymptotic variance for a function of interest. An earlier adaptive importance sampling strategy was proposed in [40].

Cournet et al. [24] proposed a new perspective of adaptive importance sampling (AIS), called Adaptive Multiple Importance Sampling, which differs from the aforementioned PMC methods because the importance weights of all simulations, produced previously as well as currently, are re-evaluated at each iteration. This method follows the 'deterministic multiple mixture' sampling scheme of Owen and Zhou [41]. The corresponding importance weight takes the form

$$w_i^t = f(\boldsymbol{\theta}_i^t) / \frac{1}{\sum_{t=0}^{T-1} N_t} \sum_{t=0}^{T-1} N_t q_t(\boldsymbol{\theta}_i^t; \widehat{\boldsymbol{\gamma}_t}) \tag{10}$$

where T is the total number of iterations, $f(\cdot)$ denotes the target distribution $\pi(\cdot)$ up to a constant, i.e., $\pi(\cdot) \propto f(\cdot)$, $q_t(\cdot)$ denotes the importance density at iteration t with sequentially updated parameters $\widehat{\boldsymbol{\gamma}_t}$ and $\boldsymbol{\theta}_i^t$ are samples drawn from $q_t(\cdot)$ with $0 \leq t \leq T-1$, $1 \leq i \leq N_t$.

The fixed denominator in Equation (10) gives the name 'deterministic multiple mixture'. The motivation is that this construction can achieve an upper bound on the asymptotic variance of the estimator without rejecting any simulations [41]. In AMIS, the parameters $\boldsymbol{\gamma}$ of a parametric importance function $q_t(\boldsymbol{\theta}; \boldsymbol{\gamma})$ are sequentially updated using the entire sequence of weighted importance samples, based on efficiency criteria such as moment matching, minimum Kullback divergence with respect to the target, or minimum variance of the weights (see, e.g. [42] for stochastic gradient-based optimization of these efficiency criteria). This leads to a sequence of importance distributions, $q_1(\boldsymbol{\theta}; \widehat{\boldsymbol{\gamma}_1}), ..., q_T(\boldsymbol{\theta}; \widehat{\boldsymbol{\gamma}_T})$ that progressively improves on the approximation to the posterior over $\boldsymbol{\theta}$. Algorithm 1 gives the pseudo code of the generic AMIS algorithm.

In this paper, we use a Gaussian importance density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, that is, $\boldsymbol{\gamma}_t = (\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$. We also choose moment matching as the efficiency criterion to estimate

Algorithm 1 Generic AMIS as analysed by Cornuet et al. [24]

- At iteration $t = 0$,
  (1) Generate N_0 independent samples $\boldsymbol{\theta}_i^0 (1 \leq i \leq N_0)$ from the initial importance density $q_0(\boldsymbol{\theta}; \widehat{\boldsymbol{\gamma}_0})$
  (2) For $1 \leq i \leq N_0$, compute $\delta_i^0 = N_0 q_0(\boldsymbol{\theta}_i^0; \widehat{\boldsymbol{\gamma}_0}), w_i^0 = f(\boldsymbol{\theta}_i^0) \Big/ q_0(\boldsymbol{\theta}_i^0; \widehat{\boldsymbol{\gamma}_0})$
  (3) Estimate $\widehat{\boldsymbol{\gamma}_1}$ of $q_1(\boldsymbol{\theta}; \widehat{\boldsymbol{\gamma}_1})$ using the weighted samples $(\{\boldsymbol{\theta}_1^0, w_1^0\}$, ..., $\{\boldsymbol{\theta}_{N_0}^0, w_{N_0}^0\})$ and a well-chosen efficiency criterion for estimation.
- At iteration $t = 1$, ..., $T-1$,
  (1) Generate N_t independent samples $\boldsymbol{\theta}_i^t (1 \leq i \leq N_t)$ from $q_t(\boldsymbol{\theta}; \widehat{\boldsymbol{\gamma}_t})$
  (2) For $1 \leq i \leq N_t$, compute the multiple mixture at $\boldsymbol{\theta}_i^t$

$$\delta_i^t = N_0 q_0(\boldsymbol{\theta}_i^t; \widehat{\boldsymbol{\gamma}_0}) + \sum_{l=1}^{t} N_l q_l(\boldsymbol{\theta}_i^t; \widehat{\boldsymbol{\gamma}_t})$$

  and derive the importance weights of $\boldsymbol{\theta}_i^t$

$$w_i^t = f(\boldsymbol{\theta}_i^t) \Big/ \Big[\delta_i^t \Big/ \sum_{j=0}^{t} N_j \Big]$$

  (3) For $1 \leq l \leq t-1$ and $1 \leq i \leq N_l$, update the past importance weights as

$$\delta_i^l \leftarrow \delta_i^l + N_t q_t(\boldsymbol{\theta}_i^l; \widehat{\boldsymbol{\gamma}_t}) \qquad \text{and} \qquad w_i^l \leftarrow f(\boldsymbol{\theta}_i^l) \Big/ \Big[\delta_i^l \Big/ \sum_{j=0}^{t} N_j \Big]$$

  (4) Estimate $\widehat{\boldsymbol{\gamma}_{t+1}}$ using all the weighted samples

$$(\{\boldsymbol{\theta}_1^0, w_1^0\}, ..., \{\boldsymbol{\theta}_{N_0}^0, w_{N_0}^0\}, ..., \{\boldsymbol{\theta}_1^t, w_1^t\}, ..., \{\boldsymbol{\theta}_{N_t}^t, w_{N_t}^t\})$$

  and the same efficiency criterion for estimation.

$\widehat{\boldsymbol{\gamma}_t} = (\hat{\boldsymbol{\mu}}^t, \hat{\boldsymbol{\Sigma}}^t)$ using the self-normalized AMIS estimator:

$$\hat{\boldsymbol{\mu}}^t = \frac{\sum_{l=0}^{t} \sum_{i=1}^{N_l} w_i^l \boldsymbol{\theta}_i^l}{\sum_{l=0}^{t} \sum_{i=1}^{N_l} w_i^l} \qquad \text{and} \qquad \hat{\boldsymbol{\Sigma}}^t = \frac{\sum_{l=0}^{t} \sum_{i=1}^{N_l} w_i^l (\boldsymbol{\theta}_i^l - \hat{\boldsymbol{\mu}}^t)(\boldsymbol{\theta}_i^l - \hat{\boldsymbol{\mu}}^t)^T}{\sum_{l=0}^{t} \sum_{i=1}^{N_l} w_i^l}$$

Despite the efficiency brought by AMIS compared with other AIS techniques, proving convergence of this algorithm is not straightforward. The work in [43] proposed a modified version of AMIS (named as MAMIS in this paper), aimed at obtaining a variant of AMIS where convergence can be more easily established. In MAMIS, the updated parameters $\widehat{\boldsymbol{\gamma}_t}$ are estimated based on samples produced at iteration t only, i.e., $\boldsymbol{\theta}_1^t, ..., \boldsymbol{\theta}_{N_t}^t$, with classical weights $f(\boldsymbol{\theta}_i^t)/q(\boldsymbol{\theta}_i^t; \widehat{\boldsymbol{\gamma}_t})$. The weights of all simulations are then updated according to

Equation (10) to give the final output, and it is recommended to increase the sample size N_t so as to improve the accuracy of $\widehat{\boldsymbol{\gamma}_t}$. MAMIS effectively solves any convergence issues of AMIS, but using less samples to update the importance distribution may lead to slower convergence, as we report in the results.

3.1. *Pseudo-marginal AMIS*

The above AMIS/MAMIS estimators are designed for the general analytically computable marginal likelihood, such as in the case of GP regression. In this paper, we propose AMIS to sample from the posterior over model parameters where the likelihood is analytically intractable but can be unbiasedly estimated. In practice, we modify AMIS by replacing the exact calculation of the marginal likelihood with an unbiased estimate, giving an unbiased estimate of the posterior up to a normalizing constant:

$$\tilde{f}(\boldsymbol{\theta}) = \tilde{p}(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{11}$$

We refer to this as Pseudo-Marginal AMIS (PM-AMIS), inspired by the name pseudo-marginal MCMC that was given to the class of MCMC algorithms replacing exact calculations of the likelihood with unbiased estimates [25]. The pseudo-code of PM-AMIS is similar to that of AMIS described in Algorithm 1, except that the target distribution up to a constant $f(\boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$ is replaced by the above unbiased estimate $\tilde{f}(\boldsymbol{\theta})$.

It is known that despite the fact that calculations are approximate, Pseudo-Marginal MCMC methods yields samples from the correct posterior distribution over covariance parameters, so a natural question is whether the same argument holds for our proposal. In the remainder of this section, we provide an analysis of the properties of Pseudo-Marginal AMIS, discussing the conditions under which it yields unbiased expectations with respect to the posterior distribution over covariance parameters. As in [26, 27], we introduce a random variable z whose distribution (denoted by $p(z \mid \boldsymbol{\theta})$ herein) is determined by the randomness occurring when carrying out the unbiased estimation of the likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$. Define:

$$z = \log \tilde{p}(\mathbf{y} \mid \boldsymbol{\theta}) - \log p(\mathbf{y} \mid \boldsymbol{\theta}) \tag{12}$$

that is,

$$\tilde{p}(\mathbf{y} \mid \boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta}) e^z \tag{13}$$

Due to the unbiased property ($E[\tilde{p}(\mathbf{y} \mid \boldsymbol{\theta})] = p(\mathbf{y} \mid \boldsymbol{\theta})$), we readily verify that $E[e^z] = 1$. For the sake of clarity, it is useful to define the unnormalized joint density of z and $\boldsymbol{\theta}$ as:

$$f(z, \boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta}) e^z p(z \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{14}$$

with a corresponding normalized version

$$\pi(z, \boldsymbol{\theta}) = \frac{f(z, \boldsymbol{\theta})}{Z} \tag{15}$$

Marginalizing this joint density with respect to z

$$\int \pi(z, \boldsymbol{\theta}) dz = \int \frac{f(z, \boldsymbol{\theta})}{Z} dz = \frac{p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{Z} E[e^z] = \frac{f(\boldsymbol{\theta})}{Z} \tag{16}$$

yields the target posterior $\pi(\boldsymbol{\theta})$ of interest defined in Equation (4).

Recall that our objective is analyzing expectations under the posterior over the parameters $\pi(\boldsymbol{\theta})$ of some function $h(\boldsymbol{\theta})$

$$E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})] = \int h(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int h(\boldsymbol{\theta}) \frac{f(\boldsymbol{\theta})}{Z} d\boldsymbol{\theta} \tag{17}$$

We begin our analysis by substituting Equation (16) into Equation (17), obtaining

$$E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})] = \int h(\boldsymbol{\theta}) \frac{f(z, \boldsymbol{\theta})}{Z} d\boldsymbol{\theta} dz \tag{18}$$

In PM-AMIS, let N_t denote the number of samples generated at each iteration t, $q_t(\boldsymbol{\theta})$ denote the importance density at each iteration for $\pi(\boldsymbol{\theta})$. We also define

$$q_t(z, \boldsymbol{\theta}) = p(z \mid \boldsymbol{\theta}) q_t(\boldsymbol{\theta}) \tag{19}$$

as the joint importance density at each iteration for $\pi(z, \boldsymbol{\theta})$, $(z_i^t, \boldsymbol{\theta}_i^t)$ as samples drawn from $q_t(z, \boldsymbol{\theta})$ with $0 \leq t \leq T, 1 \leq i \leq N_t$.

Since in a practical setting $f(z, \boldsymbol{\theta})$ is the only function that we can evaluate, the expectation defined in Equation (18) is estimated by the self-normalized PM-AMIS estimator:

$$E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})] \approx \frac{1}{\sum_{t=0}^{T} \sum_{i=1}^{N_t} {w_i}^t} \sum_{t=0}^{T} \sum_{i=1}^{N_t} {w_i}^t h(\boldsymbol{\theta}_i^t) \tag{20}$$

where the weights of this estimator are computed as

$${w_i}^t = \frac{f(z_i^t, \boldsymbol{\theta}_i^t)}{\frac{1}{\sum_{j=0}^{T} N_j} \sum_{l=0}^{T} N_l q_l(z_i^t, \boldsymbol{\theta}_i^t)} \tag{21}$$

Expanding the terms in the computations of the weights, namely substituting Equations (14) and (19) into Equation (21), we have

$$\begin{aligned} {w_i}^t &= \frac{p(\mathbf{y} \mid \boldsymbol{\theta}_i^t) e^{z_i^t} p(z_i^t \mid \boldsymbol{\theta}_i^t) p(\boldsymbol{\theta}_i^t)}{\frac{1}{\sum_{j=0}^{T} N_j} \sum_{l=0}^{T} N_l p(z_i^t \mid \boldsymbol{\theta}_i^t) q_l(\boldsymbol{\theta}_i^t)} \\ &= \frac{p(\mathbf{y} \mid \boldsymbol{\theta}_i^t) e^{z_i^t} p(\boldsymbol{\theta}_i^t)}{\frac{1}{\sum_{j=0}^{T} N_j} \sum_{l=0}^{T} N_l q_l(\boldsymbol{\theta}_i^t)} \end{aligned} \tag{22}$$

which can be rewritten in terms of the unbiased estimate of the marginal likelihood as

$$w_i{}^t = \frac{\tilde{p}(\mathbf{y} \mid \boldsymbol{\theta}_i^t)p(\boldsymbol{\theta}_i^t)}{\frac{1}{\sum_{j=0}^{T} N_j}\sum_{l=0}^{T} N_l q_l(\boldsymbol{\theta}_i^t)} = \frac{\tilde{f}(\boldsymbol{\theta}_i^t)}{\frac{1}{\sum_{j=0}^{T} N_j}\sum_{l=0}^{T} N_l q_l(\boldsymbol{\theta}_i^t)} \tag{23}$$

Equation (23) shows how the importance weights can be computed by the unbiased estimator of the marginal likelihood.

We now appeal to Lemma 1 in [24], which gives the conditions under which the self-normalized estimator of AMIS will converge to Equation (17). Following the conditions in Lemma 1 in [24], when T and N_0, ..., N_{T-1} are fixed, and when N_T goes to infinity, $w_i{}^t$ (Equation (21)) becomes:

$$w_i{}^t \simeq \frac{f(z_i^t, \boldsymbol{\theta}_i^t)}{q_T(z_i^t, \boldsymbol{\theta}_i^t)} \tag{24}$$

Then we have

$$\begin{aligned}
E_{q_t(z,\boldsymbol{\theta})}\left[\frac{1}{\sum_{t=0}^{T}\sum_{i=1}^{N_t} w_i{}^t}\sum_{t=0}^{T}\sum_{i=1}^{N_t} w_i{}^t h(\boldsymbol{\theta}_i^t)\right] &= \frac{1}{Z\sum_{t=0}^{T} N_t}\sum_{t=0}^{T} N_t \int h(\boldsymbol{\theta})\frac{f(z,\boldsymbol{\theta})}{q_T(z,\boldsymbol{\theta})} q_T(z,\boldsymbol{\theta}) d\boldsymbol{\theta} dz \\
&= \frac{1}{\sum_{t=0}^{T} N_t}\sum_{t=0}^{T} N_t \int h(\boldsymbol{\theta})\frac{f(z,\boldsymbol{\theta})}{Z} d\boldsymbol{\theta} dz \\
&= \frac{1}{\sum_{t=0}^{T} N_t}\sum_{t=0}^{T} N_t \int h(\boldsymbol{\theta})\frac{f(\boldsymbol{\theta})}{Z} d\boldsymbol{\theta} \\
&= \int h(\boldsymbol{\theta})\frac{f(\boldsymbol{\theta})}{Z} d\boldsymbol{\theta} = E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})]
\end{aligned}$$

where the normalizing constant Z is estimated by $\frac{\sum_{t=0}^{T}\sum_{i=1}^{N_t} w_i{}^t}{\sum_{t=0}^{T} N_t}$.

Therefore, under the conditions that T and N_0, ..., N_{T-1} are fixed and that N_T goes to infinity, which are the same conditions mentioned in Lemma 1 in [24], the estimator of Equation (20) proves to be an unbiased estimator of $E_{\pi(\boldsymbol{\theta})}[h(\boldsymbol{\theta})]$. As noted in [24], we remark that these conditions might prove restrictive in practice; however, these conditions provide some solid grounds onto which convergence can be established for AMIS. Furthermore, we note that in a practical setting, when in doubt as to whether convergence might be an issue, it is always possible to switch to the modified version of AMIS [43] during execution.

4. Experiments

4.1. *Competing sampling methods*

In this section, we present the state-of-the-art MCMC and AIS sampling methods considered in this work. The aim is to evaluate whether AIS (AMIS/MAMIS) can improve speed of convergence with respect to computational complexity compared to MCMC approaches. The competing sampling algorithms considered in this work are given in Table 2.

Table 2. Competing sampling algorithms considered in this work

Sampler	**Tuning parameters**
Metropolis-Hastings [44, 45] (MH)	Covariance matrix $\boldsymbol{\Sigma}$
Hybrid Monte Carlo [29, 46] (HMC)	Mass matrix $\boldsymbol{\Sigma}$, Leapfrog stepsize ϵ, Number of leapfrog steps L
No-U-Turn Sampler [47] (NUTS)	Mass matrix $\boldsymbol{\Sigma}$, Leapfrog stepsize ϵ
NUTS with Dual Averaging [47] (NUTSDA)	Mass matrix $\boldsymbol{\Sigma}$
Slice Sampling [48] (SS)	Width of the initial bracket

4.2. *Data sets*

The sampling methods considered in this work are tested on six benchmark datasets from the University of California, Irvine (UCI) repository [49]. The Concrete, Housing and Parkinsons datasets are for GP regression, whereas the Glass, Thyroid and Breast datasets are for GP classification. The number of data points and features for each dataset are given in Table 3. For the original Parkinsons dataset we randomly sampled 4 records for each of the 42 patients, resulting in 168 data points in total.

Table 3. Data sets

	Data sets for regression			Data sets for classification		
	Concrete	Housing	Parkinsons	Glass	Thyroid	Breast
n	1030	506	168	214	215	682
d	8	13	20	9	5	9

Notes: n denotes the number of data points and d denotes the number of features.

4.3. *Experimental setup*

4.3.1. Settings for GP regression

We compare three different covariances for the proposals of the MH algorithm. The first is proportional to the identity matrix. The second and third covariances are proportional to the inverse of the negative Hessian of the log-posterior (denoted by $\mathbf{H}$) evaluated at the mode (denoted by $\mathbf{m}$); one uses the full Hessian matrix, whereas the other uses its diagonal only, namely $\text{diag}((-\mathbf{H})^{-1})$. The mode $\mathbf{m}$ is found by the maximum likelihood optimization using the 'BFGS' method.

Thus the proposals that we compare in this work take the form of $\mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}, \alpha\mathbf{I})$, $\mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}, \alpha(-\mathbf{H})^{-1})$, and $\mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}, \alpha\ \text{diag}((-\mathbf{H})^{-1}))$, where α is a tuning parameter. We tune α in pilot runs until we get the desired acceptance rate (around 25%), as suggested by Roberts et al. [20].

The approximate distribution $\mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}, (-\mathbf{H})^{-1})$ is used as the initial importance density for AMIS/MAMIS. This approximation is also used to initialize several other samplers considered in this work (listed in Table 2). In this way, valid summary inference from multiple independent sequences can be obtained [31]. For AMIS/MAMIS, we explored two different strategies to update the covariance of the importance density. One updates the full covariance, whereas the other updates the diagonal of the covariance only. The

first two rows of Table 4 show the experimental settings for AMIS/MAMIS.

Motivated by the fact that knowledge of the scales and the correlations of the position variables can improve the performance of HMC [22], we also chose three types of mass matrices for HMC, namely the identity matrix, the inverse of the approximate covariance, and the inverse of the diagonal of the approximate covariance. We set the maximum leapfrog steps to be 10. We then tune the stepsize ϵ until a suggested acceptance rate (around 65%) is reached [21, 22]. The three forms of mass matrix apply to NUTS, NUTSDA as well; a full description of the pseudo codes of these algorithms can be found in Algorithms 3 and 6 in [47], respectively. NUTS requires the tuning of a stepsize ϵ. After a few trials, we set the stepsize of NUTS to 0.1. Although tuning leapfrog steps and stepsize is not an issue in NUTSDA, the parameters (γ, t_0, κ) for the dual averaging scheme therein have to be tuned by hand to produce reasonable results. After trying a few settings for each parameter, we decided to proceed with the values $\gamma = 0.05$, $t_0 = 30$, and $\kappa = 0.75$ in both the RBF and ARD covariance cases.

The slice sampling algorithm adopted in this paper makes component-wise updates of the parameters, where a new sample is drawn according to the 'stepping out' and 'shrinkage' procedures as described in [48]. In our implementation, we set the estimate of the typical size of a slice w to 1.5.

Table 4. Settings for AMIS/MAMIS/PM-AMIS

	RBF covariance		ARD covariance	
	T	N_t	T	N_t
AMIS	1120	25	280	100
MAMIS	46	$26t$	5	$3000 + 1000t$
PM-AMIS	60	400	60	400

Notes: T is the total number of iterations and N_t is the sample size at each iteration t.

4.3.2. Settings for GP classification

As a representative example of GP models with non-Gaussian likelihoods, we consider probit classification. Since the likelihood is analytically intractable and thus unbiasedly estimated, the critical property of reversibility and preservation of volume of HMC, NUTS, and NUTSDA is no longer satisfied. Also, slice sampling with the noisy estimate $\tilde{f}(\boldsymbol{\theta})$ is still valid, but naively running standard SS with the noisy estimate $\tilde{f}(\boldsymbol{\theta})$ worked very poorly as reported in [19]. As a result, we only compare PM-AMIS and Pseudo-Marginal MH (PM-MH) to infer covariance parameters in GP classification.

Both the EP and LA approximations are used to obtain importance densities to unbiasedly estimate the marginal likelihood. The last row of Table 4 shows the settings of PM-AMIS in both the RBF and ARD cases except for the Breast dataset in the ARD case using LA approximation, where the total number of iterations T is set to 240 for the sake of presentation. The initial importance density is obtained by the same optimization method as described in Section 4.3.1 except that the gradient required to perform the optimization cannot be computed analytically but is estimated from the EP or LA approximations. We update the full covariance of the importance density during the adaptation process. The proposal of PM-MH also takes the form of $\mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}, \alpha(-\mathbf{H})^{-1})$ where H is the Hessian matrix obtained again from the EP or LA approximate marginal likelihood. The collection of samples follows an initial tuning of α to reach the recommended acceptance rate

of around 25%.

4.4. *Convergence analysis*

Since the classic $\hat{R}$ score is for MCMC convergence analysis and not suitable for importance sampling, convergence analysis here is performed based on the IQR (interquartile range) of the expectation of norm of parameters ($E_{p(\boldsymbol{\theta}|\mathbf{y},\mathbf{X})}[\|\boldsymbol{\theta}\|]$) over several repetitions against the number of $O(n^3)$ operations. This is reported to be a more reliable measure of complexity than running time, as many other factors, which do not relate directly to the actual computing complexity of the algorithm, can affect the running time [17]. In GP regression the IQR is computed over 100 repetitions, whereas in GP classification it is based on 20 repetitions.

For AMIS/MAMIS/SS/MH, the computational complexity lies in the computation of the function of $f(\boldsymbol{\theta})$, where one $O(n^3)$ operation is required to perform the Cholesky decomposition of the covariance matrix $\mathbf{C}$. For HMC/NUTS/NUTSDA where computing the gradient is necessary, two extra $O(n^3)$ operations are needed for the computation of the inverse of the covariance matrix $\mathbf{C}$.

For PM-AMIS/PM-MH, the computational complexity largely comes from the EP or LA approximation of the posterior of the latent variables in order to compute the unbiased estimate $\tilde{f}(\boldsymbol{\theta})$. Both EP and LA approximations require two Cholesky decomposition ($O(n^3)$ operations); one is for the decomposition of the covariance matrix $\mathbf{K}$ of the GP prior, while the other is for the decomposition of the covariance of the approximating Gaussian. Each iteration of EP and LA requires five $O(n^3)$ operations and one $O(n^3)$ operation, respectively. In the LA approximation, two extra $O(n^3)$ operations are needed to compute the covariance of the Gaussian approximation.

4.5. *Results*

4.5.1. Convergence of samplers for GP regression

In this section, we present the comparison of convergence of samplers for GP regression considered in this paper (Table 5). Details of convergence results of AMIS family (AMIS/MAMIS), MH family (MH-I/MH-D/MH-H) and HMC family (standard HMC, NUTS, NUTSDA) can be found in Appendices A and B. Figure 1 shows the result of AMIS compared to the various competitors, where for the sake of brevity, we only report the results of their best configurations. The results are shown for the three regression datasets for both the RBF and ARD covariances. It is interesting to see that AMIS/MAMIS performs best among all methods in terms of convergence speed in the RBF covariance case. In the ARD covariance case, AMIS also converges much faster than the other approaches. However, our experiments show that in this case, although MAMIS converges faster than the other approaches in the Concrete dataset, it converges slowly in the Housing and Parkinsons datasets, which is probably due to the higher dimensionality compared to the previous cases.

In cases where MAMIS converges slowly, we can exploit the fact that AMIS converges faster to construct hybrid sampling schemes where MAMIS is initialized from a run of AMIS. In this way, we can leverage the fast adaptation of AMIS, while ensuring that the overall scheme does not introduce any bias. In the experiments, we tested this AMIS-MAMIS combination in cases where MAMIS converges slowly. These results are reported in Figure B2(f), B3(f) where EOT (end of tuning) indicates the point where we switched

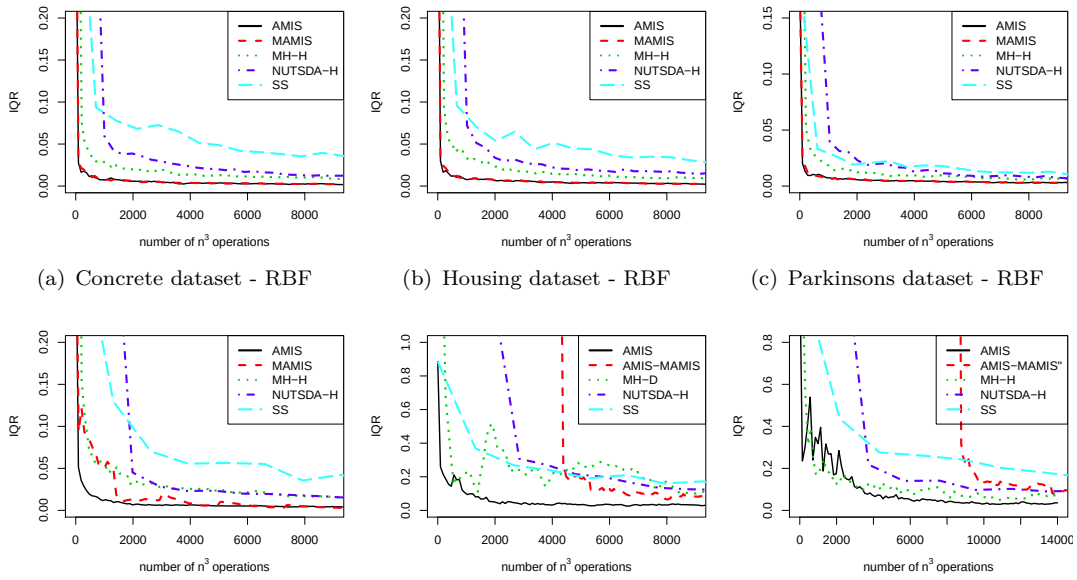
Table 5. Notation for the samplers used in the experiments

AMIS/MAMIS	AMIS/MAMIS for GP regression where the full covariance matrix of the proposal distribution is updated at each iteration
AMIS-D/MAMIS-D	AMIS/MAMIS for GP regression where only the diagonal of the covariance matrix of the proposal distribution is updated at each iteration
MH-I	MH for GP regression where the covariance of the starting proposal distribution for tuning is the identity matrix
MH-D	MH for GP regression where the covariance of the starting proposal distribution for tuning is the diagonal of the approximate covariance from the optimization
MH-H	MH for GP regression where the covariance of the starting proposal distribution for tuning is the approximate covariance from the optimization
HMC-I/NUTS-I/NUTSDA-I	HMC family for GP regression where the mass matrix is the identity matrix
HMC-D/NUTS-D/NUTSDA-D	HMC family for GP regression where the mass matrix is the inverse of the diagonal of the approximate covariance from the optimization
HMC-H/NUTS-H/NUTSDA-H	HMC family for GP regression where the mass matrix is the inverse of the approximate covariance from the optimization
PM-AMIS	AMIS for GP classification where the full covariance matrix of the proposal distribution is updated at each iteration
PM-MH	MH for GP classification where the covariance of the starting proposal distribution for tuning is the approximate covariance from the optimization

to MAMIS. Three settings (Table 6) of AMIS-MAMIS were tested for the Parkinsons dataset.

For the Housing dataset, we tested only AMIS-MAMIS in Table 6. The results for the Housing and Parkinsons datasets in the ARD covariance case prove the convergence of AMIS-MAMIS. In particular, AMIS-MAMIS and AMIS-MAMIS" seem to compete well with the other MCMC approaches in terms of convergence for the Housing dataset and the Parkinsons dataset respectively. As shown in Figure B3(f), the best performance of AMIS-MAMIS" for the Parkinsons dataset suggests that for higher dimensional problems, a more accurate initialization and a larger sample size at each iteration for MAMIS are necessary to achieve faster convergence.

Another attempt made in this paper to improve convergence speed of the adaptive importance sampling schemes is to regularize the estimation of the parameters of the importance distribution as illustrated in [50]. The regularization stems from the use of an informative prior on $\boldsymbol{\gamma}$ of the importance distribution $q_t(\boldsymbol{\gamma})$ of MAMIS and treat the update of these parameters in a Bayesian fashion [51]. This construction makes it possible to avoid situations where the importance distribution degenerates to low rank due to



(a) Concrete dataset - RBF (b) Housing dataset - RBF (c) Parkinsons dataset - RBF

(d) Concrete dataset - ARD (e) Housing dataset - ARD (f) Parkinsons dataset - ARD

Figure 1. Convergence of AMIS, Best of MAMIS, Best of MH family, Best of HMC family, SS for GP regression.

Table 6. Settings for AMIS-MAMIS

	N_t for MAMIS	$\star$ number of tuning samples for MAMIS	$\star\star$ corresponding tuning cost
AMIS-MAMIS	$1000t$	13000	4333
AMIS-MAMIS'	$5000t$	13000	4333
AMIS-MAMIS"	$5000t$	26000	8667

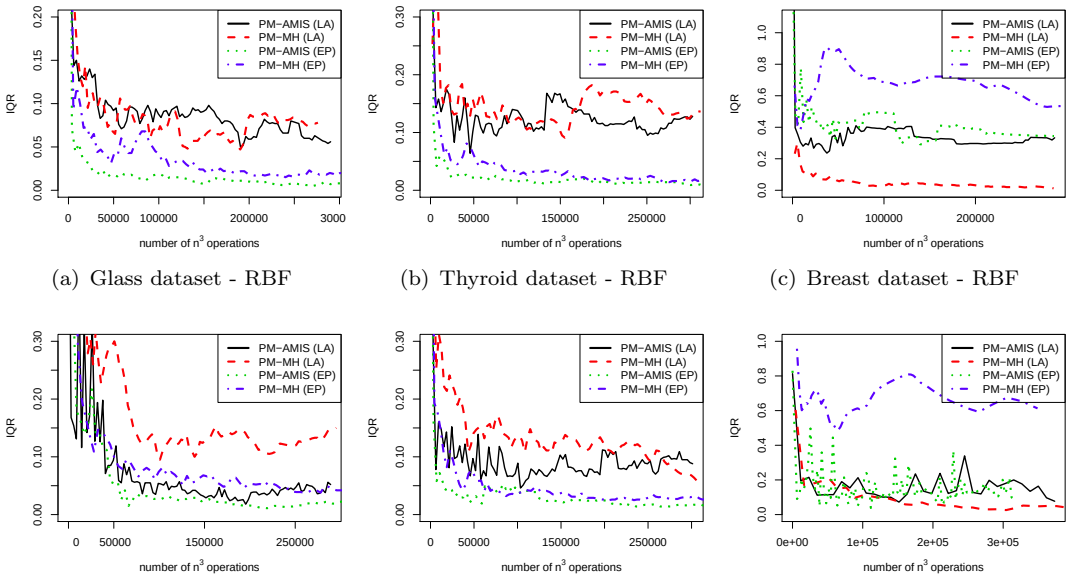
Notes: N_t is the sample size at each iteration t. $\star$ This refers to the number of samples generated from AMIS for tuning the initial importance density of MAMIS. $\star\star$ Unit of the tuning cost: number of n^3 operations.

few importance weights dominating all the others. In this work, we use an informative prior based on a Gaussian approximation to the posterior over covariance parameters. We denote this method by MAMIS-P and in the ARD covariance case it was tested only in the Housing dataset. The result indicates that even though MAMIS-P improves on MAMIS, its convergence is slower than AMIS-MAMIS (Figure B2(f)).

4.5.2. Convergence of samplers for GP classification

The comparison of convergence of samplers for GP classification (PM-AMIS and PM-MH) is presented in this section.

Figure 2 shows the results of PM-AMIS and PM-MH using EP and LA approximation (in order to obtain a Gaussian approximation to the posterior of the latent variables $\mathbf{f}$) with $N_{\text{imp}} = 64$, where N_{imp} denotes the number of importance samples of latent variables $\mathbf{f}$ to estimate the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$. The results indicate that PM-AMIS



(a) Glass dataset - RBF (b) Thyroid dataset - RBF (c) Breast dataset - RBF

(d) Glass dataset - ARD (e) Thyroid dataset - ARD (f) Breast dataset - ARD

Figure 2. Convergence of PM-AMIS, PM-MH using EP and LA approximation for GP classification.

is competitive with PM-MH in terms of convergence speed in all the EP approximation cases and in most of the LA approximation cases. The results also seem to suggest that PM-AMIS/PM-MH converge faster with the EP approximation than with the LA approximation in most cases, which we attribute to the fact that EP yields a more accurate approximation to the posterior over covariance parameters than LA [12, 13]. We also tested the performance of PM-AMIS and PM-MH with $N_{\text{imp}} = 1$, the results of which are shown in Appendix C. As expected, both PM-AMIS and PM-MH algorithms with higher number of importance samples converge much faster than those with lower number of importance samples.

5. Conclusions

In this paper we proposed the use of adaptive importance sampling techniques to compute expectations under the posterior distribution of covariance parameters in Gaussian processes. The motivation for our proposal is based on a number of observations related to the complexity of dealing with the calculation of the marginal likelihood. In GPs with a Gaussian likelihood, calculating the marginal likelihood and the gradient of its logarithm with respect to covariance parameters is expensive and the rejection of proposals of standard MCMC algorithms leads to a waste of computations. In GPs with non-Gaussian likelihoods, pseudo marginal MCMC approaches bypass the need to compute the marginal likelihood exactly, but may suffer from inefficiencies due to the fact that when a proposal is accepted and the marginal likelihood is largely overestimated, it becomes difficult for the chain to accept any other proposal. A further motivation behind our work is that importance sampling-based algorithms are generally easy to implement and tune, and can be massively parallelized.

The extensive set of results reported in this paper supports our intuition that importance

sampling-based inference of covariance parameters is competitive with MCMC algorithms. In particular, the results indicate that it is possible to achieve convergence of expectations under the posterior distribution of covariance parameters faster than employing MCMC methods in a wide range of scenarios. Even in the case of around twenty parameters, where importance sampling based methods start to degrade in performance, our proposal is still competitive with MCMC approaches.

Acknowledgments

VS acknowledges support from the Norwegian Financial Mechanism 2009-2014 under Project Contract no. MSMT-28477/2014. MF gratefully acknowledges support from the AXA Research Fund.

References

[1] Rasmussen CE, Williams C. Gaussian Processes for Machine Learning. Cambridge, Massachusetts: MIT Press; 2006.

[2] Bishop CM. Pattern recognition and machine learning (information science and statistics). Secaucus, NJ, USA: Springer-Verlag New York, Inc.; 2006.

[3] Filippone M, Girolami M. Pseudo-marginal Bayesian inference for Gaussian processes. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014;36(11):2214–2226.

[4] Filippone M, Marquand AF, Blain CRV, Williams SCR, Mourão-Miranda J, Girolami M. Probabilistic Prediction of Neurological Disorders with a Statistical Assessment of Neuroimaging Data Modalities. Annals of Applied Statistics. 2012;6(4):1883–1905.

[5] Kim S, Valente F, Filippone M, Vinciarelli A. Predictiong Continuous Confilict Perception with Bayesian Gaussian Processes. IEEE Transactions on Affective Computing. 2014; 5(2):187–200.

[6] Jones DR, Schonlau M, Welch WJ. Efficient Global Optimization of Expensive Black-Box Functions. Journal of Global Optimization. 1998;13(4):455–492.

[7] Kennedy MC, O'Hagan A. Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2001;63(3):425–464.

[8] Neal RM. Regression and classification using Gaussian process priors (with discussion). Bayesian Statistics. 1999;6:475–501.

[9] Taylor MB, Diggle JP. INLA or MCMC? A Tutorial and Comparative Evaluation for Spatial Prediction in log-Gaussian Cox Processes. 2012; arXiv:1202.1738.

[10] Williams CKI, Barber D. Bayesian classification with Gaussian processes. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998;20:1342–1351.

[11] Opper M, Winther O. Gaussian processes for classification: Mean-field algorithms. Neural Computation. 2000;12(11):2655–2684.

[12] Kuss M, Rasmussen CE. Assessing Approximate Inference for Binary Gaussian Process Classification. Journal of Machine Learning Research. 2005;6:1679–1704.

[13] Nickisch H, Rasmussen CE. Approximations for Binary Gaussian Process Classification. Journal of Machine Learning Research. 2008 Oct;9:2035–2078.

[14] Hensman J, Alexander, Filippone M, Ghahramani Z. MCMC for variationally sparse Gaussian processes; 2015. Report no.:; arXiv:1506.04000.

[15] Murray I, Adams RP. Slice sampling covariance hyperparameters of latent Gaussian models. In: Lafferty JD, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A, editors. Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada. Curran Associates, Inc.; 2010. p. 1732–1740.

[16] Vanhatalo J, Vehtari A. Sparse Log Gaussian Processes via MCMC for Spatial Epidemiology. Journal of Machine Learning Research - Proceedings Track. 2007;1:73–89.
[17] Filippone M, Zhong M, Girolami M. A comparative evaluation of stochastic-based inference methods for Gaussian process models. Machine Learning. 2013;93(1):93–114.
[18] Filippone M. Bayesian inference for Gaussian process classifiers with annealing and pseudo-marginal MCMC. In: 22nd International Conference on Pattern Recognition, ICPR 2014, Stockholm, Sweden, August 24-28, 2014. IEEE; 2014. p. 614–619.
[19] Murray I, Graham MM. Pseudo-Marginal Slice Sampling. eprint arXiv:151002958v1. 2015;.
[20] Roberts GO, Gelman A, Gilks WR. Weak convergence and optimal scaling of random walk Metropolis algorithms. Annals of Applied Probability. 1997;7:110–120.
[21] Beskos A, Pillai N, Roberts GO, Sanz-Serna JM, Stuart AM. Optimal tuning of hybrid Monte Carlo algorithm. Bernoulli. 2013;19:1501–1534.
[22] Neal RM. Handbook of Markov chain Monte Carlo, chapter 5: MCMC using hamitonian dynamics. Boca Raton, London: CRC Press; 2011.
[23] Andrieu C, Robert CP. Controlled MCMC for optimal sampling. Bernoulli. 2001;9:395–422.
[24] Cornuet JM, Marin JM, Mira A, Robert CP. Adaptive multiple importance sampling. Scandinavian Journal of Statistics. 2012;39:798–812.
[25] Andrieu C, Roberts GO. The pseudo-marginal approach for efficient Monte Carlo computations. The Annals of Statistics. 2009 Apr;37(2):697–725.
[26] Pitt M, Silva R, Giordani P, Kohn R. On some properties of markov chain monte carlo simulation methods based on the particle filter. Journal of Econometrics. 2012;171:134–151.
[27] Tran MN, Scharth M, Pitt M, Kohn R. Importance sampling squared for Bayesian inference in latent variable models. eprint arXiv:13093339. 2014;.
[28] MacKay DJ. Bayesian non-linear modelling for the prediction competition. In: In ASHRAE Transactions, V.100, Pt.2. ASHRAE; 1994. p. 1053–1062.
[29] Neal RM. Probabilistic inference using Markov chain Monte Carlo methods. Dept. of Computer Science, University of Toronto; 1993. Report No.: CRG-TR-93-1.
[30] Papaspiliopoulos O, Roberts GO, Sköld M. A general framework for the parametrization of hierarchical models. Statist Sci. 2007 02;22(1):59–73; Available from: http://dx.doi.org/10.1214/088342307000000014.
[31] Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statistical Science. 1992;7(4):457–472.
[32] Flegal JM, Haran M, Jones GL. Markov Chain Monte Carlo: Can We Trust the Third Significant Figure? Statistical Science. 2007 Mar;23(2):250–260.
[33] Haario H, Saksman E, Tamminen J. Adaptive proposal distribution for random walk Metropolis algorithm. Computational Statistics. 1999;14:375–395.
[34] Haario H, Saksman E, Tamminen J. An adaptive Metropolis algorithm. Bernoulli. 2001; 7:223–242.
[35] Cappe O, Guillin A, Marin JM, Robert CP. Population Monte Carlo. Journal of Computational and Graphical Statistics. 2004;13:907–929.
[36] Doucet A, deFreitas N, Gordon N. Sequential MCMC in practice. New York : Springer-Verlag; 2001.
[37] Rubin D. Using the SIR algorithm to simulate posterior distributions. in Bayesian Statistics 3 (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 395-402. Oxford: Oxford Univ. Press; 1988.
[38] Douc R, Guillin A, Marin JM, Robert C. Convergence of adaptive mixtures of importance sampling schemes. Ann Statist. 2007a;35:420–448.
[39] Douc R, Guillin A, Marin JM, Robert C. Minimum variance importance sampling via population monte carlo. ESAIM: Probab Stat. 2007b;11:427–447.
[40] Oh MS, Berger JO. Adaptive Importance Sampling in Monte Carlo Integration. Journal of Statistical Computing and Simulation. 1992;41:143–168.
[41] Owen A, Zhou Y. Safe and effective importance sampling. J Amer Statist Assoc. 2000; 95:135–143.
[42] Ortiz L, Kaelbling L. Adaptive importance sampling for estimation in structured domains.

In: Proceedings of the Sixteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-2000). Morgan Kaufmann Publishers, San Francisco, CA.; 2000. p. 446–454.
[43] Marin JM, Pudlo P, Sedki M. Consistency of the Adaptive Multiple Importance Sampling. eprint arXiv:12112548v2. 2014;.
[44] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. The Journal of Chemical Physics. 1953 Jun;21(6):1087–1092.
[45] Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970 Apr;57(1):97–109.
[46] Duane S, Kennedy AD, Pendleton BJ, Roweth D. Hybrid Monte Carlo. Physics Letters B. 1987;195(2):216–222.
[47] Hoffman MD, Gelman A. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. Journal of Machine Learning Research. 2014;15(1):1593–1623.
[48] Neal RM. Slice Sampling. Annals of Statistics. 2003;31:705–767.
[49] Asuncion A, Newman DJ. UCI machine learning repository. 2007; Available from: https://archive.ics.uci.edu/ml/datasets.html.
[50] Šmídl V, Hofman R. Efficient Sequential Monte Carlo Sampling for Continuous Monitoring of a Radiation Situation. Technometrics. 2014;56(4):514–528.
[51] Kulhavý R. Recursive nonlinear estimation: A geometric approach. Berlin: Springer-Verlag GmbH; 1996.

Appendices

Appendices A and B show the convergence results of the samplers for GP regression with the RBF covariance (RBF covariance case) and ARD covariance (ARD covariance case) respectively. The top-left of figures in A and B demonstrate the result of AMIS/MAMIS. It can be seen that AMIS/MAMIS that exploits the full covariance structure of the proposal distribution performs better than the one that only updates the diagonal of the covariance matrix of the proposal density. For the MH family (MH-I/MH-D/MH-H) and HMC family (standard HMC, NUTS, NUTSDA), figures in A and B show that, the methods that make use of the scales and correlation of the parameters, perform better than the one that does not in most cases. Also, NUTS/NUTSDA turns out to converge much faster than the standard HMC due to the fact that standard HMC has to be tuned costly in pilot runs. For MH and standard HMC, the computational cost of tuning is counted when comparing the convergence, as is shown in top-center and top-right of figures in Appendices A and B where the end of tuning (EOT) is indicated by three vertical dotted lines, corresponding to the three variants respectively from left to right. For NUTSDA, the computational cost of tuning the parameters of the dual averaging scheme is also counted when determining the convergence, as is displayed in bottom-right of figures in Appendix A, Figure B1 and bottom-center of Figure B2, B3 with EOT indicated by three vertical dotted lines, relating to the three variants respectively from left to right. Table 7 shows the corresponding computational cost of tuning:

Table 7. Computational cost of tuning for HMC/NUTSDA

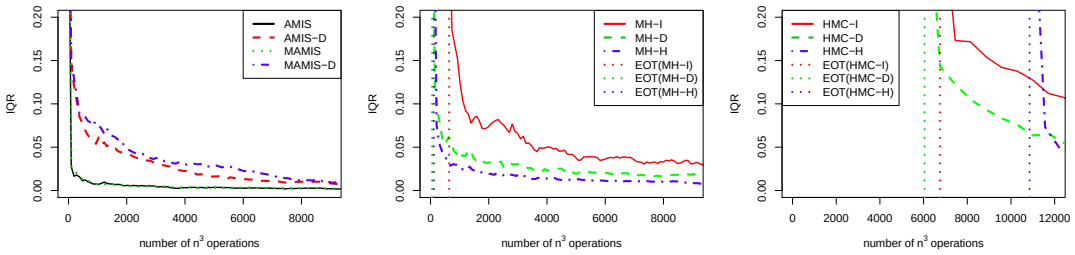
	Concrete		Housing		Parkinsons	
	RBF	ARD	RBF	ARD	RBF	ARD
HMC-I	6747	5910	4779	3924	1561	1340
HMC-D	6042	7316	7281	7726	8883	8469
HMC-H	10851	9451	10987	8860	10871	8736
NUTDA-I	1402	3528	1193	7433	1338	6488
NUTDA-D	1357	1582	1124	2424	975	1951
NUTDA-H	682	1023	670	1866	728	1794

Notes: Unit of the tuning cost: number of n^3 operations.

Appendix C shows the convergence results of PM-AMIS/PM-MH for the RBF (Figure C1) and ARD (Figure C2) cases, respectively. In the figures, LA represents the case where the Gaussian approximation to the posterior of latent variables $\mathbf{f}$ is obtained by LA approximation, whereas EP denotes the case where the Gaussian approximation is obtained by EP approximation. Nimp denotes the number of importance samples of latent variables $\mathbf{f}$ to estimate the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$. As can be seen from the figures, both PM-AMIS and PM-MH algorithms with higher number of importance samples (Nimp=64) converge much faster than those with lower number of importance samples (Nimp=1) in both EP and LA approximation cases as expected. The results also indicate that PM-AMIS is competitive with PM-MH in terms of convergence speed in most of the EP and LA approximation cases. Moreover, PM-AMIS/PM-MH seem to converge faster with EP approximation than with LA approximation in most cases which is probably because EP yields a more accurate approximation than LA as reported in [12, 13].

Appendix D presents all acronyms used in this paper.

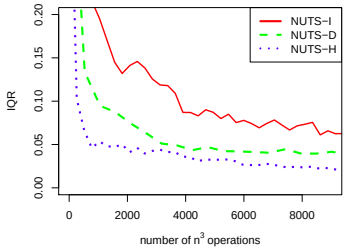
Appendix A. Convergence of samplers for GP regression with the RBF covariance



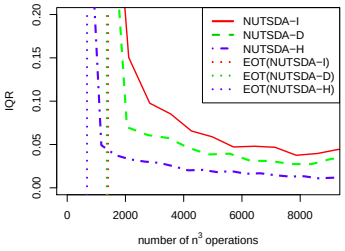
(a) AMIS/MAMIS

(b) MH

(c) HMC

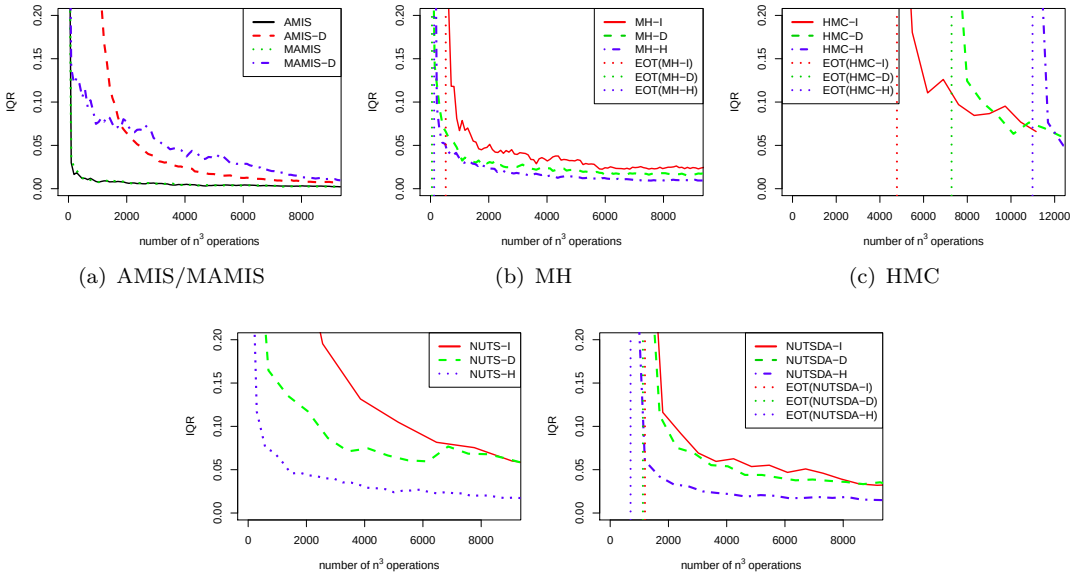


(d) NUTS



(e) NUTSDA

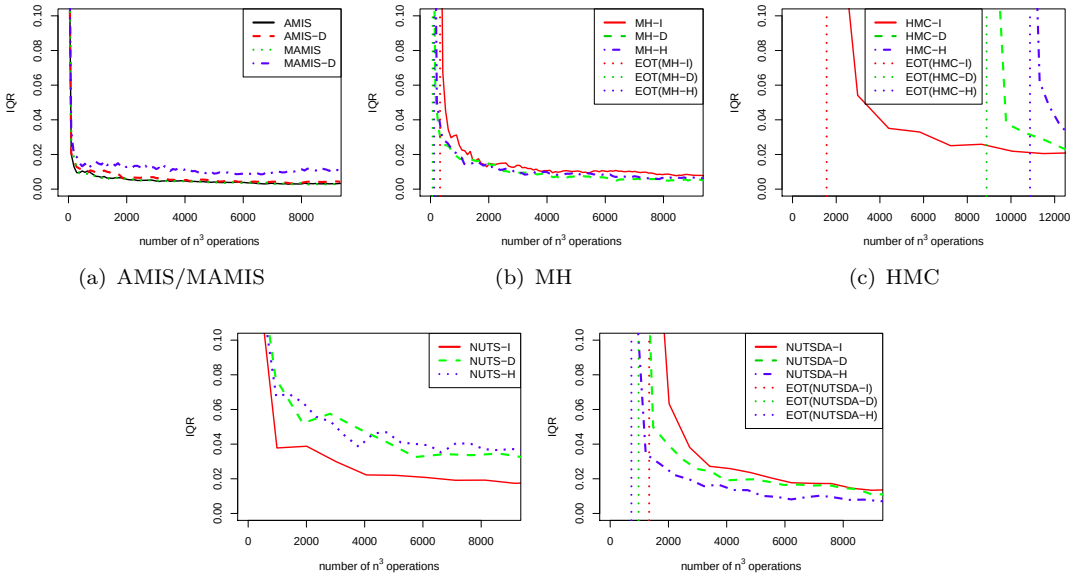
Figure A1. Convergence of AMIS/MAMIS, MH, HMC, NUTS, NUTSDA for the Concrete dataset (RBF covariance case). EOT stands for "end of tuning".



(a) AMIS/MAMIS (b) MH (c) HMC

(d) NUTS (e) NUTSDA

Figure A2. Convergence of AMIS/MAMIS, MH, HMC, NUTS, NUTSDA for the Housing dataset (RBF covariance case). EOT stands for "end of tuning".

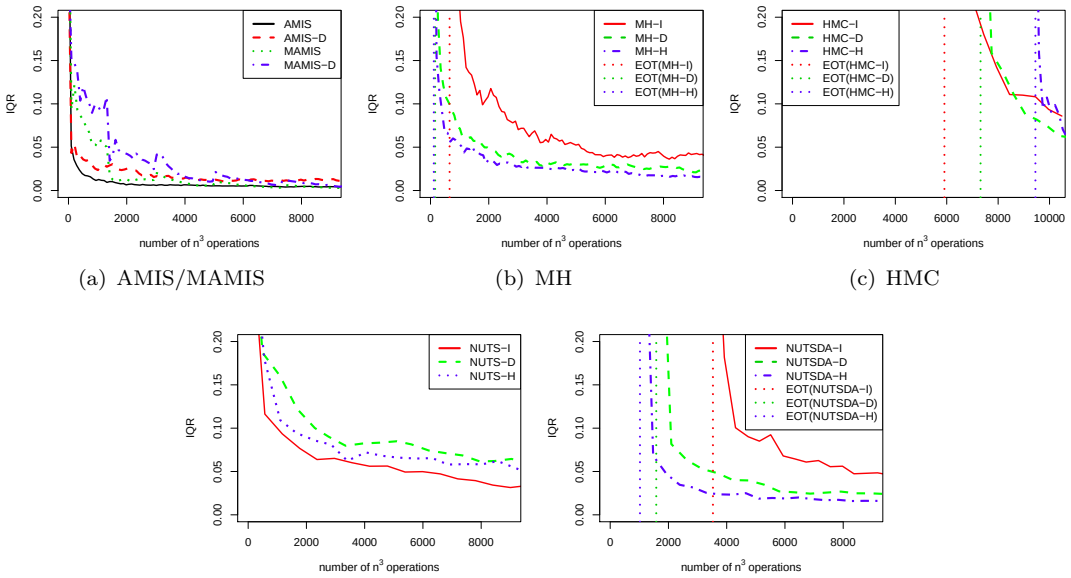


(a) AMIS/MAMIS (b) MH (c) HMC

(d) NUTS (e) NUTSDA

Figure A3. Convergence of AMIS/MAMIS, MH, HMC, NUTS, NUTSDA for the Parkinsons dataset (RBF covariance case). EOT stands for "end of tuning".

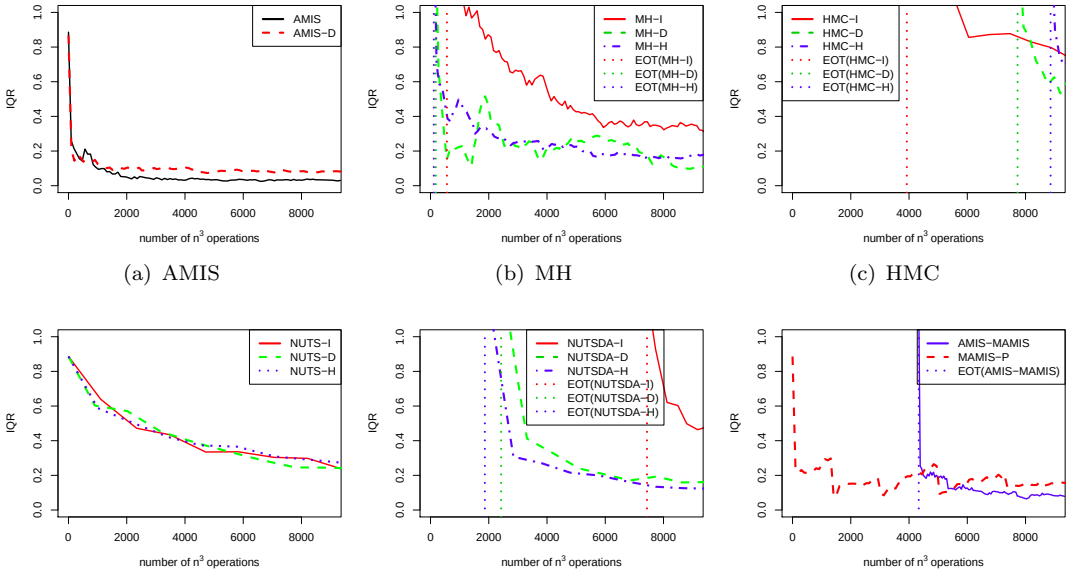
Appendix B. Convergence of samplers for GP regression with the ARD covariance



(a) AMIS/MAMIS (b) MH (c) HMC

(d) NUTS (e) NUTSDA

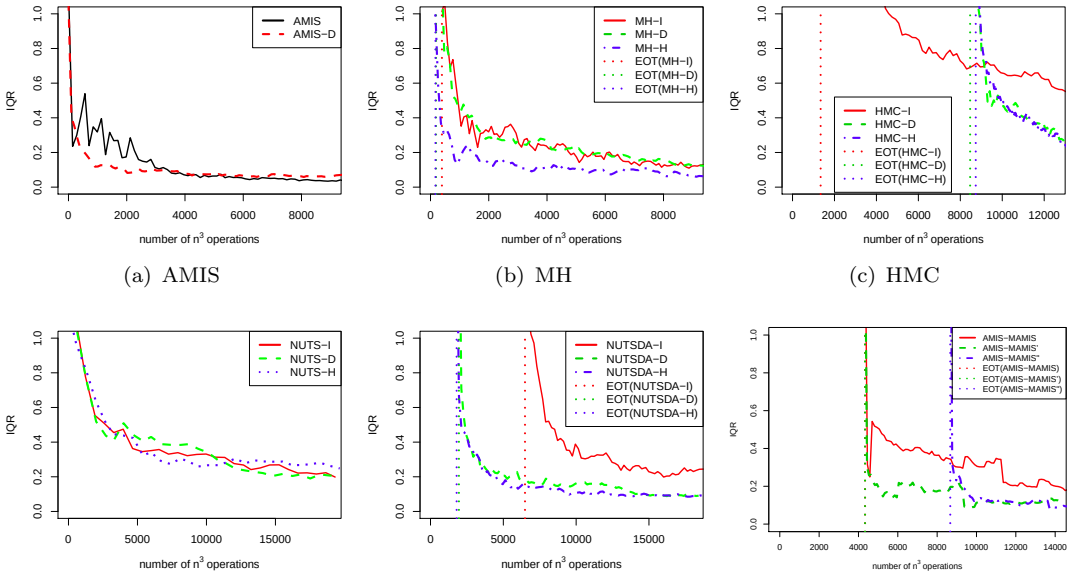
Figure B1. Convergence of AMIS/MAMIS, MH, HMC, NUTS, NUTSDA for the Concrete dataset (ARD covariance case). EOT stands for "end of tuning".



(a) AMIS (b) MH (c) HMC

(d) NUTS (e) NUTSDA (f) MAMIS

Figure B2. Convergence of AMIS/MAMIS, MH, HMC, NUTS, NUTSDA for the Housing dataset (ARD covariance case). EOT stands for "end of tuning".

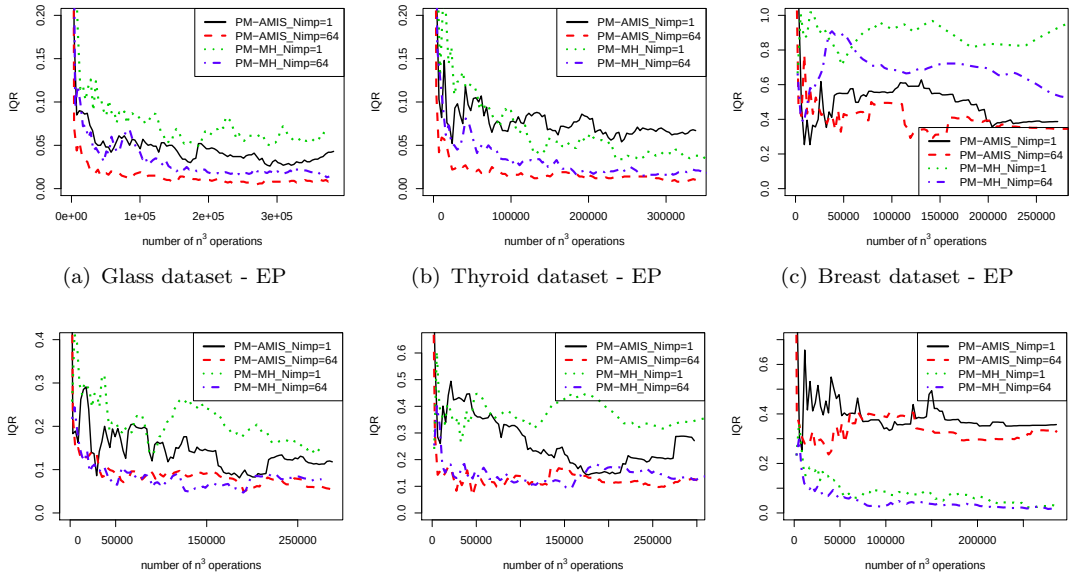


(a) AMIS (b) MH (c) HMC

(d) NUTS (e) NUTSDA (f) MAMIS

Figure B3. Convergence of AMIS/MAMIS, MH, HMC, NUTS, NUTSDA for the Parkinsons dataset (ARD covariance case). EOT stands for "end of tuning".

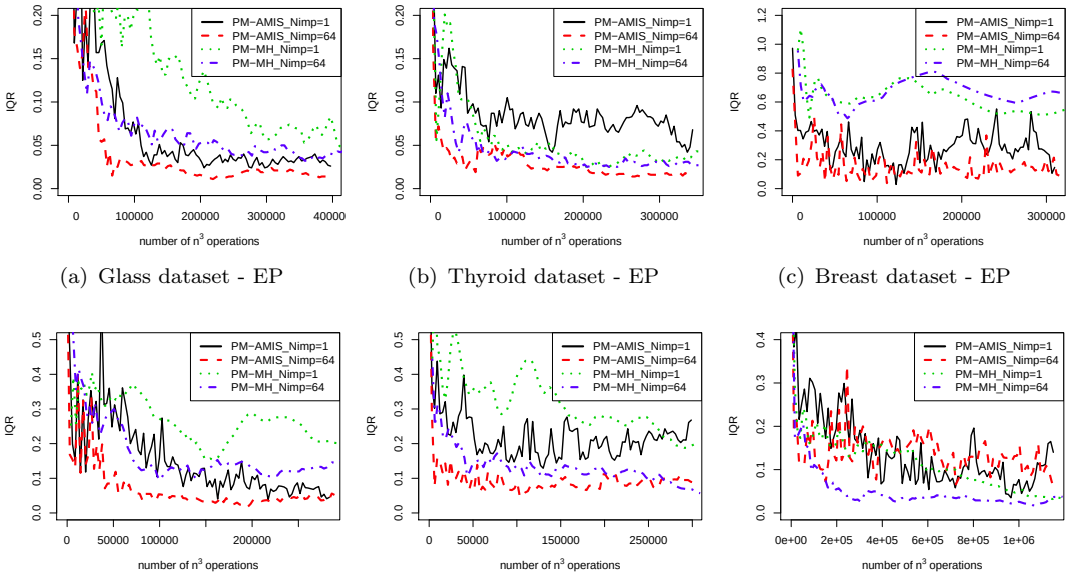
Appendix C. Convergence of samplers for GP classification



(a) Glass dataset - EP (b) Thyroid dataset - EP (c) Breast dataset - EP

(d) Glass dataset - LA (e) Thyroid dataset - LA (f) Breast dataset - LA

Figure C1. Convergence of PM-AMIS, PM-MH using EP and LA approximation for the RBF case. Nimp denotes the number of importance samples of latent variables to estimate the marginal likelihood.



(a) Glass dataset - EP (b) Thyroid dataset - EP (c) Breast dataset - EP

(d) Glass dataset - LA (e) Thyroid dataset - LA (f) Breast dataset - LA

Figure C2. Convergence of PM-AMIS, PM-MH using EP and LA approximation for the ARD case. Nimp denotes the number of importance samples of latent variables to estimate the marginal likelihood.

Appendix D. Acronyms

AIS adaptive importance sampling. 7, 8, 11
AMIS Adaptive Multiple Importance Sampling. 1, 2, 7–9, 11–16, 21–25
ARD Automatic Relevance Determination. 2, 3, 13–17, 21, 24, 25, 27

EP Expectation Propagation. 6, 13, 14, 16, 17, 21, 26, 27

GP Gaussian Process. 1–4, 6, 9, 13–17, 21
GPs Gaussian Processes. 1, 2, 4, 17

HMC Hybrid Monte Carlo. 2, 12–16, 21–25

LA Laplace Approximation. 6, 13, 14, 16, 17, 21, 26, 27

MAMIS modified version of AMIS. 8, 9, 11–16, 21–25
MCMC Markov chain Monte Carlo. 1, 2, 4–7, 9, 11, 14, 15, 17, 18
MH Metropolis-Hastings. 2, 12–16, 21–25

NUTS No-U-Turn Sampler. 12–15, 21–25
NUTSDA NUTS with Dual Averaging. 12–15, 21–25

PM-AMIS Pseudo-Marginal AMIS. 2, 9, 10, 13–17, 21, 26, 27
PM-MH Pseudo-Marginal MH. 13–17, 21, 26, 27
PMC Population Monte Carlo. 7

RBF Radial Basis Function. 3, 13, 14, 16, 17, 21–23, 26

SIR Sampling Importance Resampling. 7
SMC Sequential Monte Carlo. 7
SS Slice Sampling. 12–14, 16