Bayesian Inference for Gaussian Process Classifiers with Annealing and Pseudo-Marginal MCMC

Maurizio Filippone
School of Computing Science, University of Glasgow
Email: maurizio.filippone@glasgow.ac.uk

***Abstract*—Kernel methods have revolutionized the fields of pattern recognition and machine learning. Their success, however, critically depends on the choice of kernel parameters. Using Gaussian process (GP) classification as a working example, this paper focuses on Bayesian inference of covariance (kernel) parameters using Markov chain Monte Carlo (MCMC) methods. The motivation is that, compared to standard optimization of kernel parameters, they have been systematically demonstrated to be superior in quantifying uncertainty in predictions. Recently, the Pseudo-Marginal MCMC approach has been proposed as a practical inference tool for GP models. In particular, it amounts in replacing the analytically intractable marginal likelihood by an unbiased estimate obtainable by approximate methods and importance sampling. After discussing the potential drawbacks in employing importance sampling, this paper proposes the application of annealed importance sampling. The results empirically demonstrate that compared to importance sampling, annealed importance sampling can reduce the variance of the estimate of the marginal likelihood exponentially in the number of data at a computational cost that scales only polynomially. The results on real data demonstrate that employing annealed importance sampling in the Pseudo-Marginal MCMC approach represents a step forward in the development of fully automated exact inference engines for GP models.**

I. Introduction

Kernel methods have revolutionized the fields of pattern recognition and machine learning due to their nonlinear and nonparametric modeling capabilities [1]. Their success, however, critically depends on the choice of kernel parameters. In applications where accurate quantification of uncertainty in predictions is of primary interest, it has been argued that optimization of kernel parameters may not be desirable, and that inference using Bayesian techniques represents a much more reliable alternative [2], [3], [4], [5], [6].

This paper focuses in particular on the problem of inferring covariance (kernel) parameters of Gaussian Process classification models using Markov chain Monte Carlo (MCMC) techniques. The choice of using GP classification as a working example is that they are formulated in probabilistic terms and are therefore particularly suitable candidates for carrying out Bayesian inference of their kernel parameters. The choice of employing MCMC based inference techniques is that for general GP models and for general kernels they offer an all-purpose solution to do so up to a given precision [7], as discussed in [4], [8], [9]. The formulation of GP classifiers, and that of GP models in general, makes use of a set of latent variables $\mathbf{f}$ that are assumed to be distributed according to a GP prior with covariance parameterized by a set of parameters $\boldsymbol{\theta}$. The application of MCMC to directly draw samples form the posterior distribution over covariance parameters would require the evaluation of the so called marginal likelihood, namely the likelihood where latent variables are integrated out of the model, which is analytically intractable.

Recently, the Pseudo-Marginal (PM) MCMC approach has been proposed as a practical way to efficiently infer covariance parameters in Gaussian process classifiers exactly [3]. In this approach, computations do not rely on the actual marginal likelihood, but on an unbiased estimate obtained by approximate methods and Importance Sampling (IS). While the sampling of covariance parameters using PM MCMC improves on previous approaches for inferring covariance parameters, a large variance in the estimate of the marginal likelihood can negatively impact the efficiency of the PM MCMC approach, making convergence slow and efficiency low. In [3], IS was based on an importance distribution obtained by Gaussian approximations to the posterior over latent variables [10], [11], [12]. For certain values of the covariance parameters, the posterior over latent variables can be strongly non-Gaussian and the approximation can be poor, thus leading to a large variance in the IS estimate of the marginal likelihood [11]. This effect is exacerbated by the dimensionality of the problem that makes the variance of IS grow exponentially large [13]. In the case of GP classification, estimating the marginal likelihood entails an integration in as many dimensions as the number of data, so this effect might be problematic in the case of large data sets.

This paper presents the application of Annealed Importance Sampling (AIS) [13] to obtain a low-variance unbiased estimate of the marginal likelihood[1]. This paper empirically demonstrate that compared to IS, AIS can reduce the variance of the estimate of the marginal likelihood exponentially in the number of data at a computational cost that scales only polynomially. Finally, two versions of PM MCMC approaches, employing AIS and IS respectively, are compared on five real data sets. The results on these data demonstrate that employing AIS in the PM MCMC approach represents a step forward in the development of fully automated exact Bayesian inference engines for GP classifiers.

The remainder of this paper is organized as follows. Sections II and III review GP models and their fully Bayesian treatment using the PM MCMC approach. Section IV presents AIS to obtain an unbiased estimate of the marginal likelihood in GP models that can be used in the PM MCMC approach. Section V reports results on synthetic and real data, and section VI reports the conclusions.

[1] The code to reproduce all the results in this paper can be found here: www.dcs.gla.ac.uk/∼maurizio/pages/code_ea_mcmc_ais/

II. Bayesian Inference for GP Classification

Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n input data where $\mathbf{x}_i \in R^d$, and let $\mathbf{y} = \{y_1, \ldots, y_n\}$ be a set of associated observed binary responses $y_i \in \{-1, +1\}$. GP classification models are a class of hierarchical models where labels $\mathbf{y}$ are modeled as being independently distributed according to a Bernoulli distribution. The probability of class $+1$ for an input $\mathbf{x}_i$ is based on a latent variable f_i and is defined as $p(y_i = +1|f_i) = \Phi(f_i)$, where Φ is the cumulative distribution function of the standard normal distribution, so that $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} \Phi(y_i f_i)$. Latent variables $\mathbf{f} = \{f_1, \ldots, f_n\}$ are assumed to be distributed according to a GP prior, where a GP is a set of random variables characterized by the fact that any finite subset of them is jointly Gaussian. GPs are specified by a mean function and a covariance function; for the sake of simplicity, in the remainder of this paper we will employ zero mean GPs. The covariance function $k(\mathbf{x}, \mathbf{x}'|\boldsymbol{\theta})$ gives the covariance between latent variables at inputs $\mathbf{x}$ and $\mathbf{x}'$ and it is assumed to be parameterized by a set of parameters $\boldsymbol{\theta}$. This specification results in a multivariate Gaussian prior over the latent variables $p(\mathbf{f}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, K)$ with K defined as an $n \times n$ matrix with entries $k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta})$.

A GP can be viewed as a prior over functions and it is appealing in situations where it is difficult to specify a parametric form for the function mapping X into the probabilities of class labels. The covariance plays the role of the kernel in kernel machines, and in the remainder of this paper it will be assumed to be the Radial Basis Function (RBF) covariance

$$k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta}) = \sigma \exp\left[-\frac{1}{2}\sum_{r=1}^{d} \frac{(x_{ir} - x_{jr})^2}{\tau_r^2}\right]. \qquad (1)$$

There can be one length-scale parameters τ_r for each feature, which is a suitable modelling assumption for Automatic Relevance Determination (ARD) [14], or there can be one global length-scale parameter τ such that $\tau_1 = \ldots = \tau_d = \tau$. The parameter σ represents the variance of the marginal distribution of each latent variable. A complete specification of a fully Bayesian GP classifier requires a prior $p(\boldsymbol{\theta})$ over $\boldsymbol{\theta}$.

When predicting the label y_* for a new input data $\mathbf{x}_*$, it is necessary to estimate or infer all unobserved quantities in the model, namely $\mathbf{f}$ and $\boldsymbol{\theta}$. An appealing way of calculating predictive distributions is as follows:

$$p(y_*|\mathbf{y}) = \int p(y_*|f_*) p(f_*|\mathbf{f}, \boldsymbol{\theta}) p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y}) df_* d\mathbf{f} d\boldsymbol{\theta}. \qquad (2)$$

In the last expression predictions are no longer conditioned on latent variables and covariance parameters, as they are integrated out of the model. Crucially, such an integration accounts for the uncertainty in latent variables and covariance parameters based on their posterior distribution $p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})$.

In order to compute the predictive distribution in eq. 2, a standard way to proceed is to approximate it using a Monte Carlo estimate:

$$p(y_*|\mathbf{y}) \simeq \frac{1}{N}\sum_{i=1}^{N} \int p(y_*|f_*) p(f_*|\mathbf{f}^{(i)}, \boldsymbol{\theta}^{(i)}) df_*, \qquad (3)$$

provided that samples from the posterior $p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})$ are available. Note that in the case of GP classification, the remaining integral has a closed form solution [10].

As it is not possible to directly draw samples from $p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})$, alternative ways to characterize it have been proposed. A popular way to do so employs deterministic approximations to integrate out latent variables [11], [12], but there is no way to quantify the error introduced by these approximation. Also, quadrature is usually employed to integrate out covariance parameters, thus limiting the applicability of GP models to problems with few covariance parameters [5]. Such limitations might not be acceptable in some pattern recognition applications, so we propose Markov chain Monte Carlo (MCMC) based inference as a general framework for tackling inference problems exactly in GP models. The idea underpinning MCMC methods for GP models is to set up a Markov chain with $p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})$ as invariant distribution.

To date, most MCMC approaches applied to GP models alternate updates of latent variables and covariance parameters. All these approaches, however, face the complexity of having to decouple latent variables and covariance parameters, whose posterior dependence makes convergence to the posterior distribution slow. Reparameterization techniques are a popular way to attempt to decouple the two groups of variables [15], [16], [17]. Also, jointly sampling $\mathbf{f}$ and $\boldsymbol{\theta}$ has been attempted in [18], [19], and it is based on approximations to the posterior over latent variables. Despite these efforts, a satisfactory way of sampling the parameters $\boldsymbol{\theta}$ for general GP models is still missing, as demonstrated in a recent comparative study [8].

At this point it is useful to notice that samples from the posterior distribution of latent variables and covariance parameters can be obtained by alternating the sampling from $p(\mathbf{f}|\boldsymbol{\theta}, \mathbf{y})$ and $p(\boldsymbol{\theta}|\mathbf{y})$. Obtaining samples from $p(\boldsymbol{\theta}|\mathbf{y})$ is obviously difficult, as it requires the marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta})$; except for the case of a Gaussian likelihood, evaluating the marginal likelihood entails an integration which cannot be computed analytically [10]. In the next section we will focus on the PM MCMC approach as a practical way of dealing with this problem.

Obtaining samples from $p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})$, instead, can be done efficiently using Elliptical Slice Sampling (Ell-SS) [20]. Ell-SS defines a transition operator $T(\mathbf{f}'|\mathbf{f})$, and is a variant of Slice Sampling [21] adapted to the sampling of latent variables in GP models. Ell-SS begins by randomly choosing a threshold η for $\log[p(\mathbf{y}|\mathbf{f})]$

$$u \sim U[0, 1] \qquad \eta = \log[p(\mathbf{y}|\mathbf{f})] + \log[u] \qquad (4)$$

and by drawing a set of latent variables $\mathbf{z}$ from the prior $\mathcal{N}(\mathbf{0}, K)$. Then, a combination of $\mathbf{f}$ and $\mathbf{z}$ is sought, such that the log-likelihood of the resulting combination is larger than the threshold η. Such a combination is defined by means of sine and cosine of an auxiliary variable α, which makes the resulting combination spanning a domain of points that is an ellipse in the latent variable space. The search procedure is based on slice sampling on α starting from the interval $[0, 2\pi]$. Due to the fact that Ell-SS does not require any tuning and it has been shown to be very efficient for several GP models [8], it is the operator that will be used in the remainder of this paper to sample latent variables. However, note that latent variables can be also efficiently sampled by means of a variant of Hybrid Monte Carlo [8].

III. Pseudo-Marginal Inference for GP models

For the sake of simplicity, this work will focus on the Metropolis-Hastings (MH) algorithm [9], [22] to obtain samples from the posterior distribution over covariance parameters. The MH algorithm is based on the iteration of the following two steps: (i) proposing a new set of parameters $\boldsymbol{\theta}'$ drawing from a user defined proposal distribution $\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})$ and (ii) evaluating the Hastings ratio

$$\tilde{z} = \frac{p(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} \frac{\pi(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})} \tag{5}$$

to accept or reject $\boldsymbol{\theta}'$. As previously discussed, the marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})d\mathbf{f}$ cannot be computed analytically, except for the case of a Gaussian likelihood.

The PM approach in [3] builds upon a remarkable theoretical result [23], [24] stating that it is possible to plug an unbiased estimate of the marginal likelihood $\tilde{p}(\mathbf{y}|\boldsymbol{\theta})$ in the Hastings ratio

$$\tilde{z} = \frac{\tilde{p}(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{\tilde{p}(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} \frac{\pi(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})} \tag{6}$$

and still obtain an MCMC algorithm sampling from the correct posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$. In [3] an unbiased estimate of the marginal likelihood was obtained as follows. First, an approximation of the posterior over latent variables $p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})$, say $q(\mathbf{f}|\boldsymbol{\theta}, \mathbf{y})$, was obtained by means of approximate methods, such as for example the Laplace Approximation (LA) or Expectation Propagation. Second, based on $q(\mathbf{f}|\boldsymbol{\theta}, \mathbf{y})$, it was proposed to get an unbiased estimate of the marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ using IS. In particular, this was achieved by drawing N_{imp} samples $\mathbf{f}^{(i)}$ from the approximating distribution $q(\mathbf{f}|\boldsymbol{\theta}, \mathbf{y})$. Defining

$$w_{\text{IS}}^{(i)} = \frac{p(\mathbf{y}|\mathbf{f}^{(i)})p(\mathbf{f}^{(i)}|\boldsymbol{\theta})}{q(\mathbf{f}^{(i)}|\boldsymbol{\theta}, \mathbf{y})}, \tag{7}$$

the marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ was approximated by

$$\tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \simeq \frac{1}{N_{\text{imp}}} \sum_{i=1}^{N_{\text{imp}}} w_{\text{IS}}^{(i)}. \tag{8}$$

Such an estimate is unbiased and the closer $q(\mathbf{f}|\boldsymbol{\theta}, \mathbf{y})$ is to $p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})$ the lower the variance of the estimate [13].

In the experiments shown in [3] this estimate was adequate for the problems that were analyzed, especially when accurate approximations based on Expectation Propagation were used. However, the variance of the IS estimate grows exponentially with the dimensionality of the integral [13], and this might represent a limitation when applying PM MCMC to large data sets. In particular, a large variance in the estimate of $p(\mathbf{y}|\boldsymbol{\theta})$ can eventually lead to the acceptance of a $\boldsymbol{\theta}$ because the corresponding marginal likelihood is overestimated. If the overestimation is severe, it is unlikely that any new proposal will be accepted, resulting in slow convergence and low efficiency. The aim of this paper is to present a methodology based on AIS [13] which is capable of mitigating this effect.

IV. Marginal Likelihood estimation with Annealed Importance Sampling

AIS is an extension of IS where the weights in eq. 7 are computed based on a sequence of distributions going from one that is easy to sample from to the posterior distribution of interest. Following the derivation in [13], define $g_s(\mathbf{f})$ as the unnormalized density of a distribution which is easy to sample from; in the next section we will study two of such distributions. Also, define

$$g_0(\mathbf{f}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta}) \propto p(\mathbf{f}|\boldsymbol{\theta}, \mathbf{y}). \tag{9}$$

AIS defines a sequence of intermediate unnormalized distributions

$$g_j(\mathbf{f}) = g_0(\mathbf{f})^{\beta_j} g_s(\mathbf{f})^{1-\beta_j} \tag{10}$$

with $1 = \beta_0 > \ldots > \beta_s = 0$. The AIS sampling procedure begins by drawing one sample $\mathbf{f}_{s-1}$ from $g_s(\mathbf{f})$. After that, for $i = s-1, \ldots, 1$, a new $\mathbf{f}_{i-1}$ is obtained from $\mathbf{f}_i$ by iterating a transition operator $T_i(\mathbf{f}'|\mathbf{f})$ that leaves the normalized version of $g_i(\mathbf{f})$ invariant. Finally, computing the average of the following weights

$$w_{\text{AIS}}^{(i)} = \frac{g_{s-1}(\mathbf{f}_{s-1})}{g_s(\mathbf{f}_{s-1})} \frac{g_{s-2}(\mathbf{f}_{s-2})}{g_{s-1}(\mathbf{f}_{s-2})} \cdots \frac{g_1(\mathbf{f}_1)}{g_2(\mathbf{f}_1)} \frac{g_0(\mathbf{f}_0)}{g_1(\mathbf{f}_0)} \tag{11}$$

yields an unbiased estimate of the ratio of the normalizing constants of $g_0(\mathbf{f})$ and $g_s(\mathbf{f})$, which immediately yields an unbiased estimate of $p(\mathbf{y}|\boldsymbol{\theta})$. For numerical reasons, it is safe to implement the calculations using logarithm transformations. Also, note that although the annealing strategy is inherently serial, the computations with respect to multiple importance samples can be parallelized. We now analyze two ways of implementing AIS for GP models, which are visually illustrated in fig. 1.

A. Annealing from the prior

When annealing from the prior, the intermediate distributions are between $g_s(\mathbf{f}) = \mathcal{N}(\mathbf{f}|0, K)$ and $g_0(\mathbf{f}) = \mathcal{N}(\mathbf{f}|0, K)p(\mathbf{y}|\mathbf{f})$, namely

$$g_j(\mathbf{f}) = \mathcal{N}(\mathbf{f}|0, K) \left[p(\mathbf{y}|\mathbf{f})\right]^{\beta_j}. \tag{12}$$

Employing Ell-SS as a transition operator for $\mathbf{f}$ for the intermediate unnormalized distributions $g_j(\mathbf{f})$ is straightforward, as the log-likelihood is simply scaled by β_j. Annealing from the prior was proposed in [11] where it was reported that a sequence of 8000 annealed distributions was employed. This is because the prior and the posterior look very much different (see fig. 1) and the only way to ensure a smooth transition from the prior to the posterior is by using several intermediate distributions. This is problematic from a computational perspective, as the calculation of the marginal likelihood has to be done at each iteration of the PM approach to sample from the posterior distribution over $\boldsymbol{\theta}$. We therefore propose an alternative starting distribution $g_s(\mathbf{f})$ that leads to a reduction in the number of intermediate distributions while obtaining estimates of the marginal likelihood that are accurate enough to ensure good sampling efficiency when used in the PM MCMC approach.

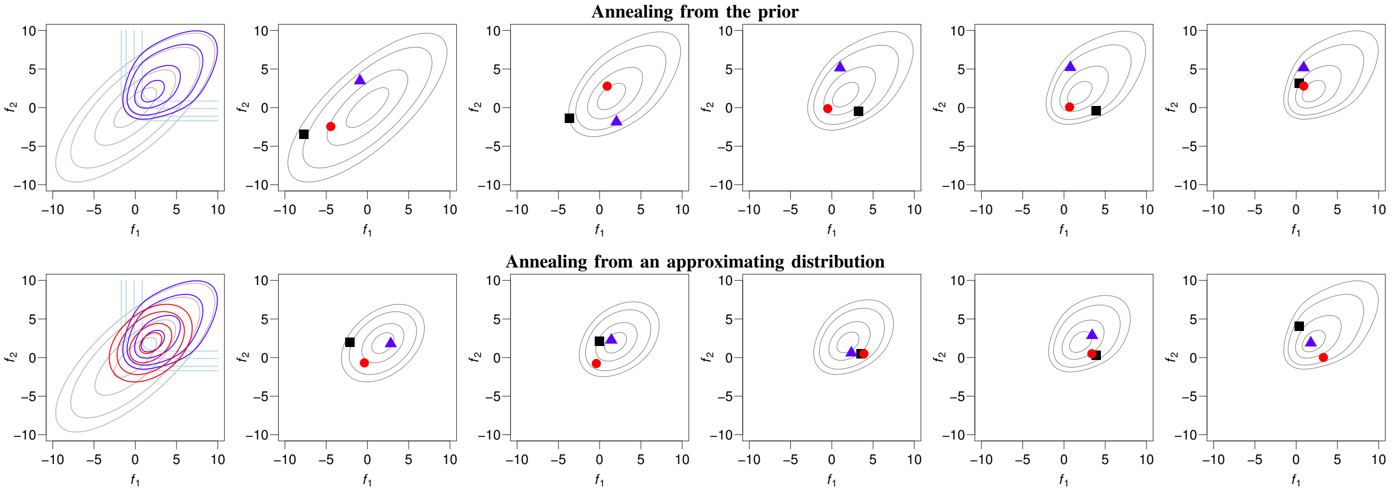

Fig. 1. Illustration of the annealing strategies studied in this work. The figure was generated as follows. The input data X comprises two data points in two dimensions with $\mathbf{x}_1 = (-1, -1)$ and $\mathbf{x}_2 = (1, 1)$, and corresponding labels $\mathbf{y} = (1, 1)$. The covariance is the one in eq. 1 with $\sigma = 15$ and $\tau = \exp(-1)$. The leftmost plots show the multiplication of the GP prior (grey) and the likelihood (light blue) resulting in the posterior distribution over the two latent variables (blue). The first row of the figure shows the annealing procedure from the GP prior to the posterior. The leftmost plot in the second row shows prior, likelihood and posterior as before, along with the Gaussian approximation given by the LA algorithm (red). The remaining plots in the second row show the annealing procedure from the approximating Gaussian distribution to the posterior. In both cases, we defined $\beta_j = \exp(-j/2)$, thus assuming a geometric spacing for the β's. Three samples drawn from $g_s(\mathbf{f})$ and propagated using operators $T_i(\mathbf{f}'|\mathbf{f})$ (one iteration of Ell-SS) have also been added to the plots.

B. *Annealing from an approximating distribution*

Several Gaussian-based approximation schemes to integrate out latent variables have been proposed for GP models [25], [26]. When an approximation to the posterior over latent variables is available, it might be reasonable to construct the sequence of intermediate distributions in AIS starting from it rather than the prior. When annealing from an approximating Gaussian distribution, the intermediate distributions are between $g_s(\mathbf{f}) = q(\mathbf{f}|\boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \Sigma)$ and $g_0(\mathbf{f}) = \mathcal{N}(\mathbf{f}|0, K)p(\mathbf{y}|\mathbf{f})$. In order to employ Ell-SS as a transition operator $T_i(\mathbf{f}'|\mathbf{f})$, it is useful to write the unnormalized intermediate distributions as

$$g_j(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \Sigma) \left[\frac{\mathcal{N}(\mathbf{f}|0, K)p(\mathbf{y}|\mathbf{f})}{\mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \Sigma)} \right]^{\beta_j} . \tag{13}$$

In this way, the model can be interpreted as having a prior $\mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \Sigma)$ and a likelihood given by the term in square brackets; applying Ell-SS to this formulation is straightforward.

V. Experimental results

The first part of this section, compares the behavior of IS and AIS in the case of synthetic data. The second part of this section, reports an analysis of IS and AIS when employed in the PM MCMC approach applied to real data. In all experiments, the approximation was based on the Laplace Approximation (LA) algorithm. Also, we imposed Gamma priors on the parameters $\mathrm{Ga}(\sigma|a = 1.1, b = 0.1)$ and $\mathrm{Ga}(\tau_i|a = 1, b = 1)$ for the ARD covariance and $\mathrm{Ga}(\tau|a = 1, b = 1/\sqrt{d})$ for the isotropic covariance, where a and b are shape and rate parameters respectively. Following the recommendations in [13], [27], $s = \sqrt{n}$ intermediate distributions were defined based on a geometric spacing of the β's. In particular, this was implemented by setting $s/2 - 1$ uniformly spaced values of $\log[\beta]$ between $\log[1]$ and $\log[0.2]$, $s/2$ uniformly spaced values between $\log[0.2]$ and $\log\left[10^{-6}\right]$, and finally $\beta_s = 0$. In AIS, the transitions $T_i(\mathbf{f}'|\mathbf{f})$ involved one iteration of Ell-SS.

A. *Synthetic data*

The aim of this section is to highlight the potential inefficiency in employing IS to obtain an unbiased estimate of the marginal likelihood and to demonstrate the effectiveness of AIS in dealing with this problem. In particular, this can be problematic in large dimensions, namely when analyzing large amounts of data. In order to show this effect, we generated data sets with an increasing number of data $n = 10, 50, 100, 500, 1000$ in two dimensions with a balanced class distribution. Data were generated drawing input vectors uniformly in the unit square and a latent function from a GP with covariance in eq. 1 with $\sigma = 20$ and a global $\tau = 0.255$. This combination of covariance parameters leads to a strongly non-Gaussian posterior distribution over the latent variables making IS perform poorly when n is large.

In order to obtain a measure of variability of the IS and AIS estimators of the marginal likelihood, we analyze the standard deviation of the estimator of $\log[p(\mathbf{y}|\boldsymbol{\theta})]$

$$r = \mathrm{st\ dev} \left\{ \log_{10} \left[\tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \right] \right\} . \tag{14}$$

In the experiments, r was estimated based on 50 repetitions; fig. 2 shows the distribution of r based on 50 draws of $\boldsymbol{\theta}$ from the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ obtained from a preliminary run of an MCMC algorithm. Ideally, a perfect estimator of the marginal likelihood would yield a degenerate distribution of r over posterior samples of $\boldsymbol{\theta}$ at zero. In practice, the distribution of r indicates the variability (across posterior samples of $\boldsymbol{\theta}$) around an average value of the standard deviation of the estimator of the logarithm of the marginal likelihood. The representation in $\log_{10}$ is helpful to get an idea of the order of magnitude of such a variability. For instance, a distribution of r across posterior samples of $\boldsymbol{\theta}$ concentrated around 2 would mean

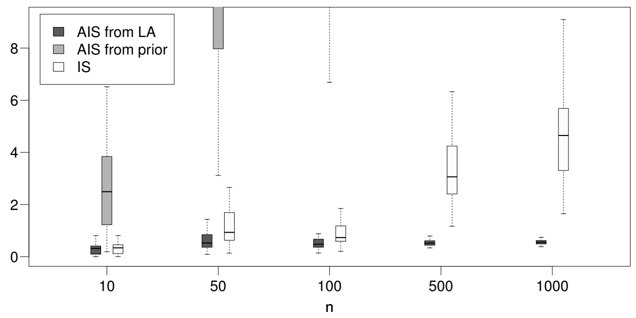

Fig. 2. This figure shows a measure of the quality of the IS and AIS (annealing from the prior and from an approximating distribution obtained by the LA algorithm) estimators of the marginal likelihood. The boxplot summarizes the distribution of r in eq. 14 for 50 values of $\boldsymbol{\theta}$ drawn from $p(\boldsymbol{\theta}|\mathbf{y})$.

that, on average, the estimates of the marginal likelihood span roughly two orders of magnitude.

Fig. 2 shows the distribution of r for AIS when annealing from the prior and from an approximating distribution, along with the distribution of r for IS as in [3]. In all methods we set $N_{\mathrm{imp}} = 4$. The results confirm that annealing from the prior offers much poorer estimates of the marginal likelihood compared to annealing from an approximating distribution and will not be considered further. The analysis of the results in fig. 2 reveal that when annealing from an approximating distribution, the reduction in variance of the estimate of the marginal likelihood compared to IS is exponential in n. When comparing the computational cost of running IS and AIS, instead, we notice that AIS increases it by a factor which scales only polynomially with n. This is because, after approximating the posterior over $\mathbf{f}$ (that typically costs $O(n^3)$ operations), in AIS drawing the initial importance samples, iterating Ell-SS, and computing the weights w_{AIS} costs $O(n^2)$ operations; this needs to be done as many times as the number of intermediate distributions s, which in our case means $O(\sqrt{n})$ times. In IS, drawing the importance samples and computing the weights w_{IS} requires $O(n^2)$ operations.

B. Real data

This section reports an analysis of the PM MCMC approach applied to five UCI data sets [28] when the marginal likelihood is estimated using AIS and IS. The Glass data set is multi-class, and we turned it into a two class data set by considering the data labelled as “window glass” as one class and data labelled as “non-window glass” as the other class. In all data sets, features were normalized to have zero mean and unit standard deviation. All experiments were repeated varying the number of importance samples $N_{\mathrm{imp}} = 1, 10$, and employing isotropic and ARD RBF covariance functions as in eq. 1.

In order to tune the MH proposal, we ran a preliminary MCMC algorithm for 2000 iterations. This was initialized from the prior and the marginal likelihood in the Hastings ratio was obtained by the LA algorithm. The proposal was then adapted to obtain an acceptance rate between 20% and 30%. This set up was useful in order to avoid problems in tuning the proposal mechanism when a noisy version of the marginal likelihood is used, which may lead to a poor acceptance rate independently of the proposal mechanism. Tab. I reports the average acceptance rate when switching to an unbiased version of the marginal likelihood obtained by IS or AIS for different values of N_{imp} after the adaptive phase. The average acceptance rate was computed based on 1500 iterations, collected after discarding 500 iterations, and over 5 parallel chains.

The results are variable across data sets and the type of covariance, but the general trend is that employing AIS in the PM MCMC approach improves on the acceptance rate compared to IS. In a few cases, it is striking to see how replacing an approximate marginal likelihood with an unbiased estimate in the Hastings ratio does not affect the acceptance rate, thus confirming the merits of the PM MCMC approach. In general, however, PM MCMC is affected by the use of an estimate of the marginal likelihood. In cases where this happens, AIS consistently offers a way to reduce the variance of the estimate of the marginal likelihood compared to IS, and this improves on the acceptance rate.

VI. Conclusions

This paper presented the application of annealed importance sampling to obtain an unbiased estimate of the marginal likelihood in GP classifiers. Annealed importance sampling for GP classifiers was previously proposed in [11] where the sequence of distributions was constructed from the prior to the posterior over latent variables. Given the difference between these two distributions, the annealing strategy requires the use of several intermediate distributions, thus making this methodology impractical. This paper studied the possibility to construct a sequence of distributions from an approximating distribution rather than the prior, and empirically demonstrated that, compared to importance sampling, this reduces the variance of the estimator of the marginal likelihood exponentially in the number of data. Crucially, this reduction comes at a cost that is only polynomial in the number of data. Also, annealed importance sampling can be easily parallelized.

The motivation for studying this problem was to plug the unbiased estimate of the marginal likelihood in the Hastings ratio in order to obtain an MCMC approach sampling from the correct posterior distribution over covariance parameters. The results on real data show that employing importance sampling within the pseudo-marginal MCMC approach can be satisfactory in many cases. However, in general, annealed importance sampling leads to a lower variance estimator of the marginal likelihood, and the resulting pseudo-marginal MCMC approach significantly improves on the average acceptance rate. These results suggest a promising direction of research towards the development of MCMC methods where the likelihood is estimated in an unbiased fashion, but the acceptance rate is as if the likelihood were known exactly. Given that the computational overhead scales with less than the third power of the number of data, the results indicate that this can be achieved with an acceptable computational cost.

This paper considered GP classification as a working example, and the Laplace approximation algorithm to obtain the importance distribution. A matter of current investigation

TABLE I. Comparison between the average acceptance rate (in %) obtained by the PM MCMC approach using IS and AIS. The number in parentheses represents the standard deviation of the average acceptance rate across five parallel chains.

Isotropic covariance

	Glass		Thyroid		Breast		Pima		Banknote	
	$n = 214$, $d = 9$		$n = 215$, $d = 5$		$n = 682$, $d = 9$		$n = 768$, $d = 8$		$n = 1372$, $d = 4$	
N_{imp}	IS	AIS	IS	AIS	IS	AIS	IS	AIS	IS	AIS
1	$2.8(1.6)$	$5.2(1.9)$	$1.1(1.0)$	$3.2(2.3)$	$17.9(2.4)$	$28.0(2.7)$	$24.8(1.4)$	$29.3(2.6)$	$1.1(0.6)$	$3.2(3.9)$
10	$10.4(3.1)$	$11.4(5.3)$	$4.1(3.8)$	$6.4(3.9)$	$30.5(4.1)$	$36.4(3.5)$	$30.8(2.6)$	$30.8(1.7)$	$4.7(1.0)$	$9.2(5.6)$

ARD covariance

	Glass		Thyroid		Breast		Pima		Banknote	
	$n = 214$, $d = 9$		$n = 215$, $d = 5$		$n = 682$, $d = 9$		$n = 768$, $d = 8$		$n = 1372$, $d = 4$	
N_{imp}	IS	AIS	IS	AIS	IS	AIS	IS	AIS	IS	AIS
1	$1.3(1.3)$	$3.6(2.3)$	$0.4(0.3)$	$2.9(1.8)$	$1.8(1.7)$	$5.0(2.5)$	$17.1(2.7)$	$22.5(3.3)$	$1.3(1.4)$	$4.7(2.1)$
10	$2.5(1.6)$	$4.9(3.2)$	$6.9(2.4)$	$6.4(2.0)$	$7.7(2.6)$	$4.5(1.8)$	$22.8(4.0)$	$24.1(3.9)$	$5.8(3.3)$	$9.2(3.1)$

is the application of the proposed methodology to other GP models and other approximation schemes. Furthermore, this paper focused on the case of full covariance matrices. These results can be extended to deal with sparse inverse covariance matrices, which are popular when modeling spatio-temporal data, thus leading to the possibility to process massive amounts of data due to the use of sparse algebra routines. Finally, this paper did not attempt to optimize the annealing scheme, but it would be sensible to do so in order to minimize the variance of the annealed importance sampling estimator of the marginal likelihood [29].

References

[1] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. New York, NY, USA: Cambridge University Press, 2004.

[2] M. Filippone, A. F. Marquand, C. R. V. Blain, S. C. R. Williams, J. Mourão-Miranda, and M. Girolami, “Probabilistic Prediction of Neurological Disorders with a Statistical Assessment of Neuroimaging Data Modalities,” *Annals of Applied Statistics*, vol. 6, no. 4, pp. 1883–1905, 2012.

[3] M. Filippone and M. Girolami, “Pseudo-marginal Bayesian inference for Gaussian processes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[4] R. M. Neal, “Regression and classification using Gaussian process priors (with discussion),” *Bayesian Statistics*, vol. 6, pp. 475–501, 1999.

[5] H. Rue, S. Martino, and N. Chopin, “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 2, pp. 319–392, 2009.

[6] M. B. Taylor and J. P. Diggle, “INLA or MCMC? A Tutorial and Comparative Evaluation for Spatial Prediction in log-Gaussian Cox Processes,” 2012, arXiv:1202.1738.

[7] J. M. Flegal, M. Haran, and G. L. Jones, “Markov Chain Monte Carlo: Can We Trust the Third Significant Figure?” *Statistical Science*, vol. 23, no. 2, pp. 250–260, 2007.

[8] M. Filippone, M. Zhong, and M. Girolami, “A comparative evaluation of stochastic-based inference methods for Gaussian process models,” *Machine Learning*, vol. 93, no. 1, pp. 93–114, 2013.

[9] R. M. Neal, “Probabilistic inference using Markov chain Monte Carlo methods,” Dept. of Computer Science, University of Toronto, Tech. Rep. CRG-TR-93-1, 1993.

[10] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[11] M. Kuss and C. E. Rasmussen, “Assessing Approximate Inference for Binary Gaussian Process Classification,” *Journal of Machine Learning Research*, vol. 6, pp. 1679–1704, 2005.

[12] H. Nickisch and C. E. Rasmussen, “Approximations for Binary Gaussian Process Classification,” *Journal of Machine Learning Research*, vol. 9, pp. 2035–2078, 2008.

[13] R. M. Neal, “Annealed importance sampling,” *Statistics and Computing*, vol. 11, no. 2, pp. 125–139, 2001.

[14] D. J. C. Mackay, “Bayesian methods for backpropagation networks,” in *Models of Neural Networks III*, E. Domany, J. L. van Hemmen, and K. Schulten, Eds. Springer, 1994, ch. 6, pp. 211–254.

[15] I. Murray and R. P. Adams, “Slice sampling covariance hyperparameters of latent Gaussian models,” in *NIPS*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, 2010, pp. 1732–1740.

[16] O. Papaspiliopoulos, G. O. Roberts, and M. Sköld, “A general framework for the parametrization of hierarchical models,” *Statistical Science*, vol. 22, no. 1, pp. 59–73, 2007.

[17] Y. Yu and X.-L. Meng, “To Center or Not to Center: That Is Not the Question–An Ancillarity-Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency,” *Journal of Computational and Graphical Statistics*, vol. 20, no. 3, pp. 531–570, 2011.

[18] L. Knorr-Held and H. Rue, “On Block Updating in Markov Random Field Models for Disease Mapping,” *Scandinavian Journal of Statistics*, vol. 29, no. 4, pp. 597–614, 2002.

[19] H. Rue, I. Steinsland, and S. Erland, “Approximating hidden Gaussian Markov random fields,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 66, no. 4, pp. 877–892, 2004.

[20] I. Murray, R. P. Adams, and D. J. C. MacKay, “Elliptical slice sampling,” *Journal of Machine Learning Research - Proceedings Track*, vol. 9, pp. 541–548, 2010.

[21] R. M. Neal, “Slice Sampling,” *Annals of Statistics*, vol. 31, pp. 705–767, 2003.

[22] W. K. Hastings, “Monte Carlo sampling methods using Markov chains and their applications,” *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[23] M. A. Beaumont, “Estimation of Population Growth or Decline in Genetically Monitored Populations,” *Genetics*, vol. 164, no. 3, pp. 1139–1160, 2003.

[24] C. Andrieu and G. O. Roberts, “The pseudo-marginal approach for efficient Monte Carlo computations,” *The Annals of Statistics*, vol. 37, no. 2, pp. 697–725, 2009.

[25] T. P. Minka, “Expectation Propagation for approximate Bayesian inference,” in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, ser. UAI ’01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.

[26] M. Opper and O. Winther, “Gaussian processes for classification: Mean-field algorithms,” *Neural Computation*, vol. 12, no. 11, pp. 2655–2684, 2000.

[27] R. M. Neal, “Sampling from multimodal distributions using tempered transitions,” *Statistics and Computing*, vol. 6, pp. 353–366, 1996.

[28] A. Asuncion and D. J. Newman, “UCI machine learning repository,” 2007.

[29] G. Behrens, N. Friel, and M. Hurn, “Tuning Tempered Transitions,” *Statistics and Computing*, vol. 22, no. 1, pp. 65–78, 2010.