Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

Adaptive Importance Sampling for Bayesian Inference in Gaussian Process models ⋆

Dejan Petelin * **Matej Gašperin** * **Václav Šmídl** **

* *Department of Systems and Control, Jožef Stefan Institute, Ljubljana, Slovenia (e-mail: dejan.petelin@ijs.si, matej.gasperin@ijs.si).*
** *Regional Innovation Centre for Electrical Engineering, University of West Bohemia, Pilsen, Czech Republic (e-mail: vsmidl@rice.zcu.cz).*

Abstract: Gaussian process (GP) models are nowadays considered among the standard tools in modern control system engineering. They are routinely used for model-based control, time-series prediction, modelling and estimation in engineering applications. While the underlying theory is completely in line with the principles of Bayesian inference, in practice this property is lost due to approximation steps in the GP inference. In this paper we propose a novel inference algorithm for GP models, which relies on adaptive importance sampling strategy to numerically evaluate the intractable marginalization over the hyperparameters. This is required in the case of broad-peaked or multi-modal posterior distribution of the hyperparameters where the point approximations turn out to be insufficient. The benefits of the algorithm are that is retains the Bayesian nature of the inference, has sufficient convergence properties, relatively low computational load and does not require heavy prior knowledge due to its adaptive nature. All the key advantages are demonstrated in practice using numerical examples.

Keywords: Adaptive importance sampling, Gaussian processes, Bayesian inference, numerical integration methods.

1. INTRODUCTION

Gaussian process (GP) models were received a lot of attention in the control systems community in recent years. They provide flexible tools for various problems, such as time series prediction (Deisenroth et al., 2009), dynamic systems control (Kocijan and Murray-Smith, 2005) or state-space model identification (Deisenroth et al., 2012). Due to their properties, the GP models are especially suitable for modelling when data are unreliable, noisy or missing, which was demonstrated with numerous successful practical applications (Likar and Kocijan, 2007). However, the fully Bayesian inference in GP models is computationally demanding and has lead to prevalence of more or less rough approximation.

Full Bayesian treatment of GP inference requires to perform the integration over the posterior distribution of a moderate number of hyperparameters. Even though most calculations in GP can be analytically solvable, this particular integral is not. The common solution is that the integration over the posterior of the hyperparameters is approximated using only a single point estimate (ML estimate). However, in Bayesian inference, all uncertainty should be taken into consideration. A possible way to perform the numerical integration is by using Markov chain Monte Carlo (MCMC) methods (Williams and Rasmussen, 1996; Neal, 1997) and MCMC using control variables (Titsias et al., 2009). Computational less demanding methods such as slice sampling (Murray and Adams, 2010) and eliptical slice sampling (Murray et al., 2010) were also investigated. For binary classification where likelihood is non-Gaussian deterministic methods were proposed (Kuss and Rasmussen, 2005; Rue et al., 2009).

In this paper, we investigate the use of adaptive importance sampling (AIS) approach (Oh and Berger, 1992; Šmidl and Hofman, 2013). It is based on a parametric form of the proposal density, the parameters of which are estimated from previously drawn particles. The key feature of this approach is its ability to update the parameters recursively for each realization of the particle. The problem that needs to be addressed in the context of GP are initialization of the parameter statistics, which is a case specific problem. Poor choice of the initial statistics may have significant impact on the inference process.

Performance of the proposed AIS-GP algorithm is demonstrated on a typical regression problem, with insufficient training data (Rasmussen and Williams, 2005). In such cases, the hyperparameter posterior distribution is multimodal and point approximation inference algorithms are bound to end up in a single optimum, resulting in possible model over-fitting, insufficient uncertainty representation and poor predictive capabilities. We show, that these issues can be successfully addressed by the AIS, while due to the adaptive nature, the algorithm in insensitive to the

⋆ The work was supported by the Slovenian research agency trough grants P2- 0001, Z2-5477, L2-5475 and by the European Regional Development Fund and Ministry of Education, Youth and Sports of the Czech Republic under project No. CZ.1.05/2.1.00/03.0094: Regional Innovation Centre for Electrical Engineering (RICE).

selection of the proposal density of importance sampling (IS) scheme.

The paper starts with a brief introduction to GP regression in Section 2, followed by a more detailed discussion on the problem of marginalization over the hyperparameters in Section 3. An introduction to the AIS and its inclusion in and the final AIS-GP algorithm are given in Section 4. The performance of the AIS-GP in comparison to the more standard methods is made in Section 5.

2. GAUSSIAN PROCESS MODELS

The GP models are probabilistic, non-parametric models based on the principles of Bayesian probability. They differ from most of the other black-box identification approaches in that they search for relationships among the measured data rather than try to approximate the modelled system by fitting the parameters of the selected basis functions. The output of the GP models is a normal distribution, expressed in terms of the mean and the variance. Their modelling properties are reviewed in (Rasmussen and Williams, 2005).

GP models can be easily utilized for regression, where the task is to infer a mapping from a set of N D-dimensional input vectors $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T$ to a vector of output data $\mathbf{y} = [y_1, y_2, \ldots, y_N]$. These observed inputs and outputs compose the training dataset D. The outputs are often assumed to be noisy realizations of an underlying function $f(\mathbf{x}_i)$. GP provides the probability distribution over estimates of $f(\mathbf{X})$.

$$f(\mathbf{x}_t) \sim \mathcal{GP}\left(m_{\boldsymbol{\theta}}(\mathbf{x}^*), k_{\boldsymbol{\theta}}(\mathbf{x}^*, \mathbf{x}^{*^T})\right), \tag{1}$$

where a GP is fully specified by its mean $\mathbf{m}_{\boldsymbol{\theta}}$ and covariance function $k_{\boldsymbol{\theta}}(\mathbf{x}^*, \mathbf{x}^{*^T})$. The marginalization property for Gaussian distributions results in the fact that any finite subset of function values $\mathbf{f} = (f(\mathbf{x}_1)f(\mathbf{x}_2), \ldots, f(\mathbf{x}_n))'$ has a joint Gaussian distribution, for which

$$\mathbf{f} \sim \mathcal{N}(0, \boldsymbol{\Sigma}), \tag{2}$$

where $\Sigma_{ij} = k_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j)$.

A common aim in regression is to predict the output y^* in an unobserved test location $\mathbf{x}^*$ given the training data D and a known covariance function $k_{\boldsymbol{\theta}}$. The posterior predictive distribution can be obtained fist constructing the joint posterior distribution $p(y^*, \mathbf{f}|\mathbf{X}, \mathbf{x}^*, \mathbf{y})$ using the Bayes' rule. The posterior predictive distribution is obtained by marginalizing over the function f

$$p(y^*|\mathbf{X}, \mathbf{x}^*, \mathbf{y}) = \int p(y^*|\mathbf{f}, \mathbf{x}^*) p(\mathbf{f}|\mathbf{X}, \mathbf{y}) d\mathbf{f}, \tag{3}$$

where the likelihood $p(\mathbf{f}|\mathbf{X}, \mathbf{y}) \sim \mathcal{N}(\mathbf{f}, \sigma^2\mathbf{I})$. Assuming zero mean GP prior leads to a Gaussian predictive distribution (Rasmussen and Williams, 2005).

Inference in GP firstly involves finding the form of the covariance function $k_{\boldsymbol{\theta}}(x_i, x_j)$. The value of covariance function $k(\mathbf{x}_i, \mathbf{x}_j)$ expresses the correlation between the individual outputs $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ with respect to inputs $\mathbf{x}_i$ and $\mathbf{x}_j$. It should be noted that the covariance function can be any function that generates a positive semi-definite covariance matrix. Typically, our prior over the function value $p(\mathbf{f})$ is too weak to quantify aspects of the covariance function. We use a hierarchical model using hyperparameters. Assuming stationary data contaminated with white noise most commonly used covariance function is composition of the square exponential covariance function and constant covariance function:

$$k_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j) = v_1 \exp\left[-\frac{1}{2}\sum_{d=1}^{D} \omega_d (x_{dp} - x_{dq})^2\right] + \delta_{pq} v_0, \tag{4}$$

where w_d are the automatic relevance determination (ARD) hyperparameters, v_1 and v_0 are hyperparameters of the covariance function, D is the input dimension, and $\delta_{pq} = 1$ if $p = q$ and 0 otherwise. Hyperparameters can be written as a vector $\boldsymbol{\theta} = [w_1, \ldots, \omega_D, v_1, v_0]^T$. The ω_d indicate the importance of individual inputs. If w_d is zero or near zero, it means the inputs in dimension d contain little information and could possibly be discarded.

Up to this point, fixed values have been assumed for the hyperparameters that determine the shape of the covariance function. However, in order to perform a full Bayesian inference the effect of unknown hyperparameters $\boldsymbol{\theta}$ has to be taken into account. This results in the following form of the predictive distribution

$$\begin{aligned} p(y^*|\mathbf{x}^*, \mathbf{y}) &= \iint p(y^*, \mathbf{f}, \boldsymbol{\theta}|\mathbf{y}, \mathbf{x}^*) d\mathbf{f} d\boldsymbol{\theta} \\ &= \iint (y^*|\mathbf{f}, \mathbf{x}^*)\, p(y^*, \mathbf{f}|\boldsymbol{\theta}, \mathbf{x}^*)\, p(\boldsymbol{\theta})\, d\mathbf{f} d\boldsymbol{\theta}, \end{aligned} \tag{5}$$

where we avoid notational clutter by omitting the conditioning on the training data $\mathbf{X}$. The computation of such integrals can be difficult due to the intractable nature of the non-linear functions. A solution to the problem of intractable integrals is to adopt either approximations or numerical integration methods.

3. INFERENCE OVER THE HYPERPARAMETERS

The implementation of Bayesian inference according to equations (5) involves the evaluation of several integrals. If these integrations are analytically intractable, solving them requires some approximation method. In practice, especially the marginalization over the hyperparameters $\boldsymbol{\theta}$ may be difficult (Rasmussen and Williams, 2005). A computationally attractive approach is to select only a point estimate for hyperparameters, which can be obtained by maximizing the marginal likelihood w.r.t. to the hyperparameters (type II maximum likelihood) (Rasmussen and Williams, 2005).

Recall, that the predictive distribution of y^* is calculated by integrating over $\mathbf{y}$ and hyperparameters $\boldsymbol{\theta}$ out of the joint distribution, repeated here for convenience

$$\begin{aligned} p(y^*|\mathbf{x}^*, \mathbf{y}) &= \iint (y^*|\mathbf{f}, \mathbf{x}^*)\, p(y^*, \mathbf{f}|\boldsymbol{\theta}, \mathbf{x}^*)\, p(\boldsymbol{\theta})\, d\mathbf{f} d\boldsymbol{\theta} \\ &= \int p(y^*|\mathbf{x}^*, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \end{aligned} \tag{6}$$

where only the marginalization over the training outputs $\mathbf{y}$ is tractable. As discussed before, the functional form of the likelihood term determines whether the integral over latent variables is analytically tractable. It is usually tractable in the case of regression when likelihood is Gaussian. However, the integral over $\boldsymbol{\theta}$ is usually analytically intractable. This papers deals with different numerical

approximations for the integration over the distribution of the hyperparameters.

3.1 Type II Maximum-likelihood approximation

A computationally attractive approach is that instead of approximating the entire posterior distribution over the hyperparameters, we only approximate it with one point estimate of the hyperparameters $\hat{\boldsymbol{\theta}}$. If this estimate maximizes the likelihood function of the hyperparameters, it is referred to as evidence maximization or *type II maximum likelihood (ML-II)* estimate (Rasmussen and Williams, 2005), which maximizes the likelihood function of the hyperparameters. However, using this approximation we are not coherent with Bayesian inference, which opens up the possibility of over-fitting.

In practice, the following negative log-likelihood function is minimized:

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2}\log(\mid \mathbf{K} \mid) - \frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y} - \frac{N}{2}\log(2\pi). \quad (7)$$

Since the covariance matrix $\mathbf{K}$ in (7) depends on $\boldsymbol{\theta}$, the likelihood function is non-linear and multi-modal. In fact if the posterior distribution of the hyperparameters is narrow ML-II estimate can lead to equally good results compared to those of full Bayesian inference. In the case when the posterior distribution is multi-modal, gradient based method is guaranteed only to find a local minimum. Therefore, the hyperparameters should either be optimized several times with different initializations, or other global optimization method should be used, in order to find the global maximum (Petelin et al., 2011).

3.2 Monte-Carlo approximation of the posterior predictive distribution

The intractable integral over the posterior distribution of the hyperparameters can be approximated with several methods, e.g. grid integration (Šimandl et al., 2006), Gaussian approximations and linearization (Rue et al., 2009), sigma-point approximation (Särkkä, 2011) and Monte-Carlo based approximations (Doucet et al., 2001). The computational efficiency and convergence properties of Monte-Carlo methods based on IS depends on the choice of the proposal density. We first start with a brief introduction to the IS algorithm, followed by the AIS approach.

Importance sampling Assuming one can efficiently sample from the distribution $p(\boldsymbol{\theta}|y)$, the integration over posterior distribution $p(\boldsymbol{\theta}|y)$ can be approximated by

$$p\left(y^*|\mathbf{x}^*,\mathbf{y}\right) \approx \frac{1}{N}\sum_{i=1}^{N} p\left(y^*|\boldsymbol{\theta}^{(i)},\mathbf{x}^*,\mathbf{y}\right). \quad (8)$$

Therefore, (6) is evaluated by drawing samples from the proposal distribution $q\left(\boldsymbol{\theta}\right)$ such that

$$p\left(\boldsymbol{\theta}\right) = \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta}) \dot{\propto} \sum_{i=1}^{n} \frac{p(\boldsymbol{\theta}^{(i)})}{q(\boldsymbol{\theta}^{(i)})} \delta\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)}\right) \quad (9)$$

$$= \sum_{i=1}^{n} \tilde{w}^{(i)} \delta\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)}\right) \quad (10)$$

where $w^{(i)} = \frac{p(\boldsymbol{\theta}^{(i)})}{q(\boldsymbol{\theta}^{(i)})}$ is the non-normalized weight, and $\tilde{w}^{(i)} = w^{(i)} / \sum_{i=1}^{n} w^{(i)}$ is the normalized weight. The main advantage of this approach is that under mild conditions it converges to the unknown function $p(\boldsymbol{\theta})$ with probability one.

The approximation of the predictive distribution (6) is

$$p\left(y^*|\mathbf{x}^*,\mathbf{y}\right) \approx \sum_{i=1}^{n} \tilde{w}^{(i)} p\left(y^*|\mathbf{x},\boldsymbol{\theta}^{(i)}\right) \delta\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)}\right), \quad (11)$$

which is a Gaussian mixture. For easier use, we project (11) to a single Gaussian using moment matching.

$$p\left(y^*|\mathbf{x}^*,\mathbf{y}\right) \approx \mathcal{N}(\hat{y}^*, \mathbf{P}), \quad (12)$$

$$\hat{y}^* = \sum_{i=1}^{n} \tilde{w}^{(i)} \hat{y_i}^*, \quad (13)$$

$$\mathbf{P} = \sum_{j=1}^{n} \tilde{w}^{(i)} (\mathbf{P}_i + \hat{y_i}^* \hat{y_i}^{*T}) - \hat{y}^* \hat{y}^{*T}, \quad (14)$$

where $\hat{y}_i$ and $\mathbf{P}_i$ are the mean value and covariance matrix of the i-th mixture component. This projection ensures that the prediction output follows the GP framework and the GP-AIS can thus be used in recursive algorithms, e.g. modelling of dynamical systems.

Laplace approximation The initial approximation for the proposal distribution is the Laplace approximation, where the mean value is the ML estimate of the parameters and covariance is the inverse of the Hessian matrix. This approximation will be good if the posterior for $\boldsymbol{\theta}$ is fairly well peaked.

$$p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}) \approx \mathcal{N}\left(\boldsymbol{\theta}_{ML}, \mathbf{H}_{ML}^{-1}\right), \quad (15)$$

where

$$\mathbf{H}_{ML} = \left[-\frac{d^2 L\left(\boldsymbol{\theta}\right)}{d\boldsymbol{\theta}^2}|_{\boldsymbol{\theta}_{ML}}\right] \quad (16)$$

is the Hessian matrix evaluated at $\boldsymbol{\theta}_{ML}$ and $L\left(\boldsymbol{\theta}\right)$ is the log-likelihood function. This may be a reasonable approximation in the case of uni-modal posterior distributions.

3.3 Adaptive importance sampling

The rate of convergence for IS heavily depends on the chosen proposal, which is difficult to choose for a general problem. An elegant solution is to choose the proposal function from a parametric family, $q\left(\boldsymbol{\theta}|\mu,\mathbf{P}\right)$, and iteratively estimate the vector of parameters from the sampled particles (Oh and Berger, 1992). The idea of the AIS is to draw batches of samples of size n_k from the parametric proposal and after each batch compute the sufficient statistics of the proposal distribution parameters. The statistics are used to update the estimates which are used for generation of the next batch. The key attribute of the AIS is the possibility to shorten the batch to $n_k = 1$ and run the algorithm in completely recursive manner. Therefore, we can use the exponential forgetting of the previous data to preserve recursiveness by introducing additional parameter $0 < \lambda \leq 1$. This way, the influence of the poor initial conditions can be suppressed by exponential weighting of the older contributions. The update of statistics in AIS is:

$$\nu_{k+1} = \lambda\nu_k + w^{(i)}, \quad (17)$$

$$\mathbf{m}_{k+1} = \lambda\mathbf{m}_k + w^{(i)}\boldsymbol{\theta}^{(i)}, \quad (18)$$

$$\mathbf{S}_{k+1} = \lambda\mathbf{S}_k + w^{(i)}\boldsymbol{\theta}^{(i)}\boldsymbol{\theta}^{(i)T}, \quad (19)$$

with estimated values of the parameters $\boldsymbol{\mu}_{k+1} = \mathbf{m}_{k+1}\nu_{k+1}^{-1}$ and $\mathbf{P}_{k+1} = \mathbf{S}_{k+1}\nu_{k+1}^{-1} - \boldsymbol{\mu}_{k+1}\boldsymbol{\mu}_{k+1}^T$. The full algorithm is adapted from (Oh and Berger, 1992) and is given in Algorithm 1.

Algorithm 1: AIS with forgetting

(1) Set batch counter $k = 0$ and initial statistics ω_0, μ_0, $\mathbf{P}_0$ and forgetting λ
(2) Draw n_k samples $\boldsymbol{\theta}^{(i)}$ from the proposal $q\left(\boldsymbol{\theta}|\mu_{\mathbf{k}},\mathbf{P}_k\right)$ and compute their weights

$$w^{(i)} = \frac{p(\boldsymbol{\theta}^{(i)})}{q\left(\boldsymbol{\theta}|\mu_k,\mathbf{P}_k\right)}, i = 1\ldots n \tag{20}$$

(3) Evaluate statistics μ_{k+1} and $\mathbf{P}_{k+1}$ using (18), (19) and (17).

Example 1: Consider a non-Gaussian bi-modal likelihood function as shown in Figure 1 to be the target density. The AIS algorithm uses a Gaussian proposal distribution, initialized at $\mathbf{x}_0 = [-3 \quad 1]^T$ and $\boldsymbol{\Sigma}_0 = \mathrm{diag}(0.3 \quad 0.3)$. The updated proposals according to Algorithm 1 are shown in Figure 1. It can be seen that even in the relatively demanding case, the proposal converges to a distribution that adequately covers both peaks of the target density, which ensures optimal utilization of the samples.

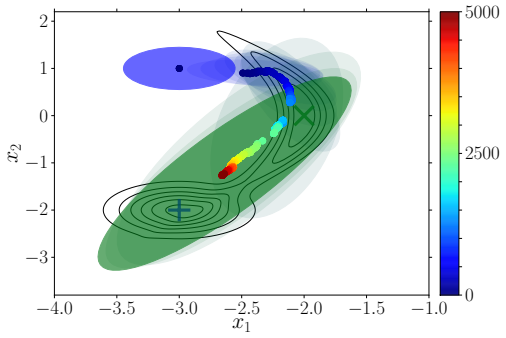


Fig. 1. Adaptation of the proposal distribution.

3.4 GP regression with AIS

The AIS scheme is implemented in GP inference and we will refer to the resulting algorithm as AIS-GP. The proposed implementation is designed with the aim to have minimum parameterization. Therefore, we include the optimum step, where the initial guess for the hyperparameter posterior distribution is obtained by Laplace approximation.

In the AIS-GP implementation, the proposal statistics were updated with every sample (effectively batch size $n_k = 1$). It would make sense to use $n_k > 1$ is the algorithm is parallelized as for each sample, GP inference is independent from other samples.

4. EXPERIMENTS AND VALIDATION

The inference with AIS-GP algorithm is evaluated and compared to other methods (Maximum likelihood, Grid, IS) on an example with two local optima. Example consists

Algorithm 2: GP-AIS

(1) **Initialize** Set initial hyperparameter estimate θ_0, initial covariance $\boldsymbol{\Sigma}_\theta$, number of particles used for AIS, N
(2) (Optional) Find optimal hyperparameter values using ML-II optimization and construct the Laplace approximation and set $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_{ML}$ and $\boldsymbol{\Sigma}_\theta = \mathbf{H}$
(3) **for** $i = 1 : N$
Sample from the distribution $\boldsymbol{\theta}^{(i)} \sim \mathcal{N}(\boldsymbol{\theta}^{(i)}, \Sigma_i)$ and compute

$$w^{(i)} = \frac{p(\boldsymbol{\theta}^{(i)})}{q(\boldsymbol{\theta}^{(i)})},$$
$$p(y_i^*|\mathbf{x}^*, \boldsymbol{\theta}^{(i)}) = \mathcal{GP}\left(m_{\boldsymbol{\theta}^{(i)}}(\mathbf{x}^*), k_{\boldsymbol{\theta}^{(i)}}(\mathbf{x}^*, \mathbf{x}^{*^T})\right)$$

where $p(\boldsymbol{\theta}^{(i)})$ is the likelihood of the sample $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}^{(i)})$.
end for
(4) Normalize importance weights and project the predictions to a single Gaussian using moment matching (14)

of 20 data points for training and 80 data points for validation. The data points represent a typical regression problem and are by a GP with (4) where $[v_1, \omega_1, v_0] = [1, 1, 0.1]$. Additionally, the data points are contaminated by white noise. 90% of the data points by white noise with standard deviation 0.1 and the rest of the data points with standard deviation 1. Both, training data points and validation data points are depicted in Figures 2(b,c) and 3c, denoted as "+" and "·" respectively.

The data is modelled by using a covariance function (4). As the input data is one dimensional, there are three hyperparameters of the covariance function. With the aim to make the inference steps easier to visualize, the hyperparameter v_1 is fixed to $v_1 = \log(1) = 0$. Despite only two variable hyperparameters there still exist two local optima. The contour of the GP model's marginal likelihood for hyperparameter values $w_1 \in [-2.5, 1.5]$ and $w_1 \in [-3.5, 0.5]$ is depicted in Figure 2a. The optimum L_f^1 corresponds to a relatively complicated model with low noise, whereas the optimum L_f^2 corresponds to a much simpler model with more noise.

By running multiple optimization procedures with random initial values we determined that 64.2% optimizations finish in optima L_f^1, where $\mathcal{L}(\boldsymbol{\theta}) = 23.8545$ and 35.5% in optima L_f^2, where $\mathcal{L}(\boldsymbol{\theta}) = 23.3807$. A small portion of optimization procedures finish in 3 other local optima with much lower log marginal likelihood and their effect is negligible. However, as both local optima have very similar log marginal likelihood values it is not possible for the model to confidently reject either of the two possibilities.

The algorithms are validated and compared in terms of their predictive capability by computing the Standardized Mean Squared Error (SMSE). Additionally, the quality of the prediction variance of all the approaches is compared with the Mean Standardized Log Loss (MSLL) (Rasmussen and Williams, 2005). The MSLL is obtained

by averaging log predictive density over the validation set and subtracting the same score for a trivial model.While SMSE calculates only the error of the prediction mean value, the MSLL also accounts for the likelihood of the prediction.

The AIS-GP algorithm is compared to Type II Maximum-likelihood approximation, standard importance sampling (IS) and uniform point-mass approximation. The proposal density in IS algorithm is set to Laplace approximation in either of the two optima. The point-mass approximation serves as a reference and the grid is uniformly spaced over the sufficient partition of the support. All the numerical approximation methods use the same amount of samples. The inference in all algorithms was carried out using 10.000 samples.

The results of the IS sampling scheme with proposals from both optima and GP inference are depicted in Figure 2a. It can be seen that in both cases the other optima is not covered good enough and consequently the obtained model is still over-fitted to optima used for calculating the proposal. The results obtained by the AIS-GP are depicted in Figure 3. It can be seen from Figure 3b that the samples from adaptive proposal also covers the region of optima L_f^2, even though the initial proposal is positioned in optima L_f^1. The evolution of the proposal can be seen in Figure 3a. However, same result is also achieved when the initial estimate is a Laplace transfomration at the second optimum.

Error measures of all conducted methods are given in Table 1. It can be seen that comparing sampling methods the AIS results the lowest error measure values. The

Table 1. SMSE and MSLL error measure values for different algorithms.

Method	SMSE	MSLL
$\mathbf{ML_1}$	0.677435	1.164701
$\mathbf{ML_2}$	0.337965	−0.548378
$\mathbf{ML_{avg}}$	0.554891	0.553064
$\mathbf{IS_1}$	0.659408	−0.073132
$\mathbf{IS_2}$	0.506627	−0.419936
AIS	0.321127	−0.556568
Point-mass	0.504723	−0.373281

results show the advantage of the approximate integration methods, as the achieved errors are lower than in ML approximation. Furthermore, among the numerical approximation methods, using the same amount of samples, the AIS sampling strategy systematically outperforms the IS and the point-mass algorithms. The advantage over the point-mass integration arises from the fact, that many samples in point-mass integration lie in the regions with extremely low likelihood values and contribute very little to the result.

5. CONCLUSION

The aim of this paper is to discuss the inference in GP models at one particular level, namely the inference over the vector of hyperparameters. An established approach for this step is to search for a point estimate (e.g. Maximum-Likelihood) and use it to approximate the hyperparameter posterior. The use of this approach is well justified in the case where enough training data is available and the actual posterior is well-peaked. However, it is often overlooked, that in general, this approximation results in a non-Bayesian inference procedure, which can lead to undesired effects, such as over-fitting. This can be mitigated by application of the Laplace approximation at the point of estimate.

In this paper we propose to use the adaptive importance sampling (AIS) approach to extend the Laplace approximation to the full Bayesian inference. We have demonstrated that the AIS is able overcome local extremes and cover larger areas of the parameter space. The simulated results confirmed that the achieved prediction errors with the fixed amount of computations, are indeed lower with the AIS-GP algorithm than with comparable methods. The effect of numerical approximations are expected to bring an even greater advantage in problems with higher dimensions.

The paper only discussed GP-AIS in the context of regression problems, but the ideas can be extended to GP modeling of dynamical systems. Implementation of GP-AIS to inference to GP model of state-space systems is one of the possible topics for future work, as the lack of sufficient training data and overfitting can be a major issue.

REFERENCES

Deisenroth, M.P., Huber, M.F., and Hanebeck, U.D. (2009). Analytic moment-based Gaussian process filtering. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, 225–232. ACM, New York, NY, USA.

Deisenroth, M., Turner, R., Huber, M., Hanebeck, U., and Rasmussen, C. (2012). Robust filtering and smoothing with Gaussian processes. *IEEE Transactions on Automatic Control*, 57(7), 1865–1871.

Doucet, A., De Freitas, N., and Gordon, N. (eds.) (2001). *Sequential Monte Carlo methods in practice*. Springer New York.

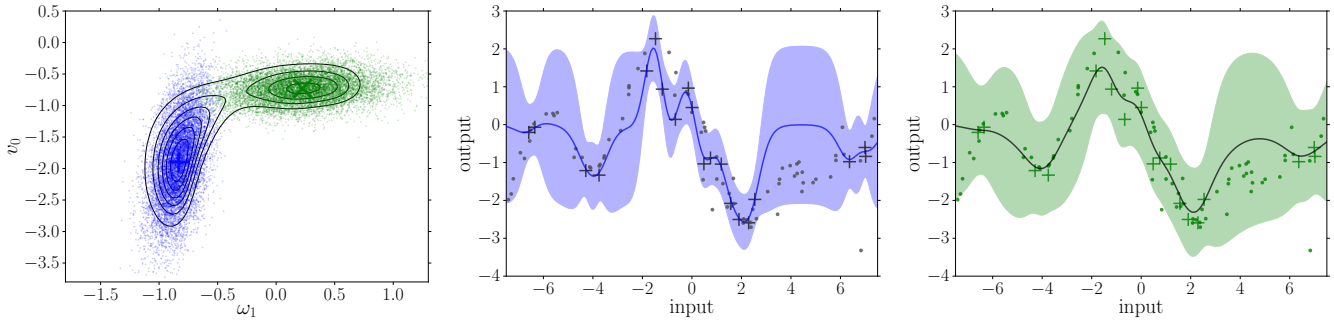
Kocijan, J. and Murray-Smith, R. (2005). Nonlinear predictive control with a Gaussian process model. In R. Murray-Smith and R. Shorten (eds.), *Switching and Learning in Feedback Systems*, volume 3355 of *Lecture Notes in Computer Science*, 185–200. Springer Berlin Heidelberg.

Kuss, M. and Rasmussen, C.E. (2005). Assessing approximate inference for binary Gaussian process classification. *The Journal of Machine Learning Research*, 6, 1679–1704.

Likar, B. and Kocijan, J. (2007). Predictive control of a gas-liquid separation plant based on a Gaussian process model. *Computers & Chemical Engineering*, 31(3), 142–152.

Murray, I. and Adams, R. (2010). Slice sampling covariance hyperparameters of latent Gaussian models. In J. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems 23*, 1732–1740.

Murray, I., Adams, R.P., and MacKay, D.J.C. (2010). Elliptical slice sampling. In I. Guyon, G.C. Cawley, G. Dror, V. Lemaire, and A.R. Statnikov (eds.), *Pro-*

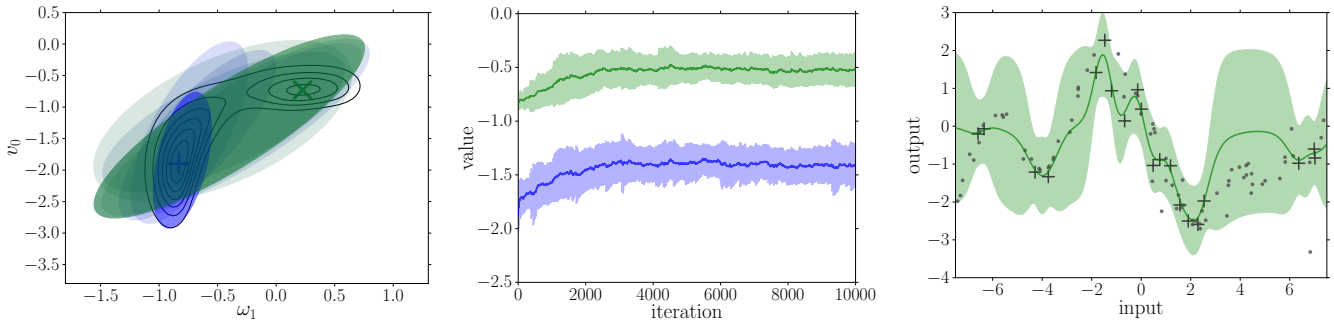


(a) Contour of the marginal likelihood and samples gathered by IS with proposals from both optima L_f^1 and L_f^2.

(b) Predictions based on hyperparameter values (samples) obtained by IS with proposal from L_f^1.

(c) Predictions based on hyperparameter values (samples) obtained by IS with proposal from L_f^2.

Fig. 2. Results of the IS. The marginal likelihood and samples gathered by IS with proposals from both optima, L_f^1 and L_f^1, are depicted in (a). Predictions based on hyperparameter values (samples) obtained by IS with proposals from both optima are depicted in (b) and (c) respectively.



(a) Contour of the marginal likelihood and ellipses presenting individual proposals during AIS.

(b) Parameters (μ) of the proposal distribution as a function of iteration number over 100 MC realizations.

(c) Predictions based on hyperparameter values (samples) obtained by AIS.

Fig. 3. Results of the AIS. Proposals during AIS are depicted in (a). The blue ellipse presents initial proposal and the green one presents the final proposal, while light coloured ellipses present interim proposals. The marginal likelihood and samples gathered by IS with initial proposal from optima L_f^1 are depicted in (b). Predictions based on hyperparameter values (samples) obtained by AIS are depicted in (c).

ceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics.

Neal, R.M. (1997). Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical Report 9702, Department of Statistics, University of Toronto.

Oh, M.S. and Berger, J.O. (1992). Adaptive importance sampling in Monte Carlo integration. *Journal of Statistical Computation and Simulation*, 41 (3), 143–168.

Petelin, D., Filipič, B., and Kocijan, J. (2011). Optimization of Gaussian process models with evolutionary algorithms. In A. Dobnikar, U. Lotrič, and B. Šter (eds.), *Adaptive and Natural Computing Algorithms*, volume 6593 of *Lecture Notes in Computer Science*, 420–429. Springer Berlin Heidelberg.

Rasmussen, C.E. and Williams, C.K.I. (2005). *Gaussian processes for machine learning*. The MIT Press.

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

Särkkä, S. (2011). Learning curves for Gaussian processes via numerical cubature integration. In T. Honkela, W. Duch, M. Girolami, and S. Kaski (eds.), *Artificial Neural Networks and Machine Learning ICANN 2011*, Lecture Notes in Computer Science, 201–208. Springer Berlin Heidelberg.

Titsias, M., Lawrence, N.D., and Rattray, M. (2009). Efficient sampling for Gaussian process inference using control variables. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (eds.), *Advances in Neural Information Processing Systems 21*, 1681–1688.

Šimandl, M., Kralovec, J., and Soderstrom, T. (2006). Advanced point-mass method for nonlinear state estimation. *Automatica*, 42(7), 1133–1145.

Šmidl, V. and Hofman, R. (2013). Adaptive importance sampling in particle filtering. In *Information Fusion (FUSION), 2013 16th International Conference on*, 9–16.

Williams, C.K.I. and Rasmussen, C.E. (1996). Gaussian rocess for regression. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo (eds.), *Advances in Neural Information Processing Systems 8*, 514–520.