Computational Statistics and Data Analysis 103 (2016) 367–383

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Gaussian process hyper-parameter estimation using Parallel Asymptotically Independent Markov Sampling

A. Garbuno-Inigo [a,*], F.A. DiazDelaO [a], K.M. Zuev [b]

[a] Institute for Risk and Uncertainty, School of Engineering, University of Liverpool, Brownlow Hill, Liverpool L69 3GH, United Kingdom
[b] Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125, USA

ARTICLE INFO

Article history:
Received 25 June 2015
Received in revised form 3 December 2015
Accepted 29 May 2016
Available online 6 June 2016

Keywords:
Gaussian process
Hyper-parameter
Marginalisation
Optimisation
MCMC
Simulated annealing

ABSTRACT

Gaussian process emulators of computationally expensive computer codes provide fast statistical approximations to model physical processes. The training of these surrogates depends on the set of design points chosen to run the simulator. Due to computational cost, such training set is bound to be limited and quantifying the resulting uncertainty in the hyper-parameters of the emulator by uni-modal distributions is likely to induce bias. In order to quantify this uncertainty, this paper proposes a computationally efficient sampler based on an extension of Asymptotically Independent Markov Sampling, a recently developed algorithm for Bayesian inference. Structural uncertainty of the emulator is obtained as a by-product of the Bayesian treatment of the hyper-parameters. Additionally, the user can choose to perform stochastic optimisation to sample from a neighbourhood of the Maximum a Posteriori estimate, even in the presence of multimodality. Model uncertainty is also acknowledged through numerical stabilisation measures by including a nugget term in the formulation of the probability model. The efficiency of the proposed sampler is illustrated in examples where multi-modal distributions are encountered. For the purpose of reproducibility, further development, and use in other applications the code used to generate the examples is freely available for download at https://github.com/agarbuno/paims_codes.

1. Introduction

Computationally expensive computer codes are frequently needed to implement mathematical models which are assumed to be reliable approximations to physical processes. Such simulators often require intensive use of computational resources that makes them inefficient if further exploitation of the code is needed for optimisation, uncertainty propagation and sensitivity analysis (Forrester et al., 2008; Kennedy and O'Hagan, 2001a). For this reason, surrogate models are needed to perform fast approximations to the output of demanding simulators and enable efficient exploration and exploitation of the input space. In this context, Gaussian processes are a common choice to build statistical surrogates – also known as *emulators* – which allow to take into account the uncertainty derived from the inability to evaluate the original model in the whole input space. Gaussian processes have become popular in recent years due to their ability to fit complex mappings between outputs and inputs by means of a non-parametric hierarchical structure. Such applications are found, amongst many other areas, in Machine Learning (Rasmussen and Williams, 2006), Spatial Statistics (Cressie, 1993, with the name of Kriging), likelihood-free Bayesian Inference (Wilkinson, 2014) and Genetics (Kalaitzis and Lawrence, 2011).

* Corresponding author.
E-mail address: agarbuno@liv.ac.uk (A. Garbuno-Inigo).

http://dx.doi.org/10.1016/j.csda.2016.05.019
0167-9473/

To build an emulator, a number of runs from the simulator is needed, but due to computing limitations only a small number of evaluations can be performed. With a small amount of data, it is possible that the uncertainty in the parameters of the model cannot be described by a uni-modal distribution. In such scenarios, Model Uncertainty Analysis (Draper, 1995) is capable of setting a proper framework that acknowledges all uncertainties related to the idealisations made through the modelling assumptions and the available, albeit limited information. To this end, *hierarchical modelling* should be taken into account. This corresponds to adding a layer of structural uncertainty to the assumed emulator either in a continuous or discrete manner (see Draper, 1995, Section 4). In the case of Gaussian processes, continuous structural uncertainty can be accounted for as a natural by-product of Bayesian inference. Hence, this is pursued in this work by focusing on samplers capable of exploring multi-modal distributions.

In order for the Gaussian process to be able to replicate the relationship between inputs and outputs and make predictions, a training phase is necessary. Such training involves the estimation of the parameters of the Gaussian process from the data collected by running the simulator. These parameters are referred to as *hyper-parameters*. The selection of the hyper-parameters is usually done by using Maximum Likelihood estimates (MLE), or their Bayesian counterpart Maximum a Posteriori estimates (MAP) (Oakley, 1999; Rasmussen and Williams, 2006), or by sampling from the posterior distribution (Williams and Rasmussen, 1996) in a fully Bayesian manner.

In this paper we assume a scenario where the task of generating new runs from the simulator is prohibitive. Such limited information is not enough to completely identify either a candidate or a region of appropriate candidates for the hyper-parameters. In this scenario, traditional optimisation routines (Nocedal and Wright, 2004) are not able to guarantee global optima when looking for the MLE or MAP, and a Bayesian treatment is the only option to account for all the uncertainties in the modelling. In the literature, however, it is common to see that MLE or MAP alternatives are preferred (Kennedy and O'Hagan, 2001a; Gibbs, 1998) due to the numerical burden of maximising the likelihood function or because it is assumed that Bayesian integration will not produce results worth the effort. Though it is a strong argument in favour of estimating isolated candidates, in high-dimensional applications it is difficult to assess if the number of runs of the simulator is sufficient to produce robust hyper-parameters. Robustness is usually measured with a prediction-oriented metric such as root-mean-square error (RMSE) (Kennedy and O'Hagan, 2001b), ignoring uncertainty and risk assessment of choosing a single candidate of the hyper-parameters by an inference process with limited data. In order to account for such uncertainty in the hyper-parameters when making predictions, numerical integration should be performed. However, methods as quadrature approximation become infeasible as the number of dimensions increases (Kennedy and O'Hagan, 2001a). Therefore, an appropriate approach is to perform Monte Carlo integration (MacKay, 1998). This allows to approximate any integral by means of a weighted sum, given a sample from the *correct* distribution.

In Gaussian processes, as in many other applications of statistics, the target distribution of the hyper-parameters cannot be sampled directly and one should resort to Markov Chain Monte Carlo (MCMC) methods (Robert and Casella, 2004). MCMC methods are powerful statistical tools but have a number of drawbacks if not tuned properly, particularly if one wishes to sample from multi-modal distributions (Neal, 2001; Hankin, 2005). One of such limitations is the tuning of the proposal distribution, which allows the generation of a candidate in the chain. This proposal function has to be tuned with parameters that define its ability to move through the sample space. If an excessively wide spread is selected, this will produce samples with space-filling properties but which are likely to be rejected. On the other hand, having a narrower spread will cause an inefficient exploration of the sample space by taking short updates of the states of the chain, known in the literature as *Random Walk* behaviour (Neal, 1993). In practice it is desirable to use a proposal distribution which is capable of balancing both extremes. To find an appropriate tuning in high-dimensional spaces with sets of highly correlated variables can be an overwhelming task and often MCMC samplers can become expensive due to the long time needed to reach stationarity (Ching and Chen, 2007). Neal (1998) and Williams and Rasmussen (1996) favour the Hybrid Monte Carlo (HMC) method to generate a sample from the posterior distribution, preventing the random walk behaviour of traditional MCMC methods. If tuned correctly, HMC should be able to explore most of the input space (Liu, 2008). Such tuning process is problem-dependent and there is no guarantee that the method will sample from all existing modes, thus failing to adapt well to multi-modal distributions (Neal, 2011).

This paper proposes a sampler for the hyper-parameters of a Gaussian process based on recently developed methods for Bayesian inference problems. Additionally, it uses the Transitional Markov Chain Monte Carlo (TMCMC) method of Ching and Chen (2007) to set a framework for the parallelisation of Asymptotically Independent Markov Sampling in both the context of hyper-parameter sampling (AIMS) (Beck and Zuev, 2013) and in stochastic optimisation (AIMS-OPT) (Zuev and Beck, 2013) reminiscent of Stochastic Subset Optimisation (Taflanidis and Beck, 2008a,b). Such an extension is built using concepts of Particle Filtering methods (Andrieu et al., 2010; Gramacy and Polson, 2009), Adaptive Sequential Monte Carlo (Del Moral et al., 2006, 2012) and Delayed Rejection Samplers (Zuev and Katafygiotis, 2011; Mira, 2001). AIMS is chosen since it provides a framework for Sequential Monte Carlo sampling (Neal, 1996, 2001; Del Moral et al., 2006) which automatically chooses the sequence of transitions. Moreover, it uses most of the information generated in the previous step in the sequence as opposed to traditional sequential methods, thus building a robust sampler when applied to multi-modal distributions. Finally, by using the AIMS-OPT algorithm a solution is built by means of a nested sequence of subsets, which converges to the optimal solution set. The algorithm can be terminated prematurely given a previously chosen accuracy threshold, thus providing a set of nearly optimal solutions. Whether it is composed by a single element, or a set of elements whose objective function differs by a negligible quantity, a full characterisation of the optimal solution is achieved. This contrasts with the