Fast Inference in Non-Conjugate Gaussian Process Models via Data Augmentation

Florian Wenzel[*]
TU Kaiserslautern, Germany

Théo Galy-Fajou[*]
TU Berlin, Germany

Christian Donner
TU Berlin, Germany

Marius Kloft
TU Kaiserslautern, Germany
University of Southern California, USA

Manfred Opper
TU Berlin, Germany

Abstract

We present `AugmentedGaussianProcesses.jl`, a software package for augmented stochastic variational inference (ASVI) for Gaussian process models with non-conjugate likelihood functions. The idea of ASVI is to find an augmentation of the original GP model which renders the model conditionally conjugate and perform inference in the augmented model. We evaluate our method for three GP models (GP classification, robust GP regression and a Bayesian SVM model) and demonstrate that it is up to two orders of magnitude faster than the state-of-the-art.

1 Introduction

Gaussian processes (GPs) Rasmussen and Williams (2005) provide a popular Bayesian non-parametric non-linear approach on inference on functions. A wide range of interesting Bayesian models can be obtained by connecting a GP prior with a suitable likelihood function. The simplest case is as Gaussian likelihood leading to a GP regression model. This model is conjugate and computing the posterior is easy and given in closed-form. For other likelihood functions, as e.g. the probit and logistic likelihood (used for classification), the posterior is intractable and one has to resort to sampling or approximate inference methods (e.g. Dezfouli and Bonilla, 2015; Buchholz et al., 2018).

Recently, another approach based on data augmentations has been proposed (Wenzel et al., 2017, 2018). The idea is to aim on augmenting the non-conjugate GP model by a set of auxiliary variables such it becomes conditionally conjugate. In a conditionally conjugate model, applying stochastic variational inference (Hoffman et al., 2013) becomes easy and the updates are given in closed-form.

In this paper we present `AugmentedGaussianProcesses.jl`, a software package for inference in non-conjugate GP models. The package is coded in the programming language Julia (Bezanson et al., 2017) and implements a set of data augmentation based variational inference algorithms. We experiment with three different GP models and find that AugmentedGaussianProcesses.jl is up to two orders of magnitude faster than GPFlow, another state-of-the-art package for non-conjugate GP inference.

Moreover, we aim on providing a useful tool for research on GPs. The software package provides an easy to use framework for implementing new augmentation stochastic variational inference (ASVI) algorithms for other GP models. The code can be found at `https://github.com/theogf/AugmentedGaussianProcesses.jl`.

[*] equal contributions, contact: wenzelfl@hu-berlin.de

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

2 Augmented variational inference for Gaussian processes

Former work on augmenting GP models focuses on single specific GP models, e.g. GP classification (Wenzel et al., 2018). Here, we present a more general view on the problem and provide some ideas on how to find suitable augmentations.

2.1 The Augmentation

We are interested in models which consist of a GP prior on a latent function $f \sim \mathrm{GP}(0, k)$, where k is the kernel function and the data y is connected to f via a non-conjugate likelihood $p(y|f)$. We now aim on finding an augmented representation of the model which renders the model conditionally conjugate. Let ω be potential augmentation, then the augmented joint distribution is

$$p(y, f, \omega) = p(y|f, \omega)p(\omega)p(f). \tag{1}$$

The original model can be restored by marginalizing ω, i.e. $p(y, f) = \int p(y, f, \omega)d\omega$.

The goal is to find an augmentation ω, such that the augmented likelihood $p(y|f, \omega)$ becomes conjugate to the prior distributions $p(f)$ and $p(\omega)$ and the expectations of the log complete conditional distributions $\log p(f|\omega, y)$ and $\log p(\omega|f, y)$ can be computed in closed-form. These expectation are the basis for the variational inference updates in section 2.2.

How to find a suitable augmentation? Many popular likelihood functions can be expressed as a *scale mixture of Gaussians*

$$p(y|f) = \int \mathcal{N}(y; Bf, \mathrm{diag}(\omega^{-1}))p(\omega)d\omega,$$

where B is a matrix (Palmer et al., 2006). This representation directly leads to the augmented likelihood $p(y|\omega, f) = \mathcal{N}(y; Bf, \mathrm{diag}(\omega^{-1}))$ which is conjugate in f, i.e. the posterior is Gaussian again. This is a good starting point for a successful augmentation and the only remaining requirement that needs to be met is conjugacy in ω.

In section 4 we present three models where this approach is successful. We obtain conditionally conjugate augmentations for a GP classification model based on the logistic likelihood, a robust GP regression model based on the Student-t likelihood and a Bayesian SVM model.

2.2 Inference in the augmented model

We now assume that the augmentation, discussed in the previous section, was successful and we obtained an augmented model $p(y, f, \omega) = p(y|f, \omega)p(f)p(\omega)$ that is conditionally conjugate. In a conditionally conjugate model variational inference is easy and block coordinate ascent updates can be computed in closed-form. We follow as structured mean-field approach (Wainwright and Jordan, 2008) and assume a decoupling between the latent GP f and the auxiliary variable ω in the variational distribution $q(f, \omega) = q(f)q(\omega)$. We alternate between updating $q(f)$ and $q(\omega)$ by using the typical coordinate ascent (CAVI) updates building on expectations of the log complete conditionals (Blei et al., 2017).

The hyperparameter of the latent GP (e.g. length scale) are learned by optimizing the variational lower bound as function of the hyper parameters. We alternate between updating the variational parameters and the hyperparameters (Wenzel et al., 2018).

Scaling to big datasets. Direct inference for GPs has a cubic computational complexity $\mathcal{O}(N^3)$. To scale our model to big datasets we approximate the latent GP by a *sparse GP* building on *inducing points* (Hensman et al., 2013). This reduces the complexity to $\mathcal{O}(M^3)$, where M is the number of inducing points. Using inducing points allows us to employ stochastic variational inference (SVI) (Hoffman et al., 2013) that computes the updates based on mini-batches of the data.

2.3 Why use the augmentation stochastic variational inference (ASVI) approach?

The goal of our augmentation is to frame the model to be conditionally conjugate. Conditionally conjugate models allow for using the standard SVI algorithm (Hoffman et al., 2013) which has the following advantages.

1. The updates are computed in **closed-form**. This is opposed to most non-conjugate variational inference approaches which use sampling or numerical quadrature.
2. The (global) variational updates of the GP parameters ($q(f)$) can be interpreted as **natural gradient** updates. Natural gradients have the advantage that they provide effective second-order optimization (Hoffman et al., 2013; Jähnichen et al., 2018). In most other approaches the updates are based on ordinary Euclidean gradients.
3. The (local) variational updates of the auxiliary variables ($q(\omega)$) are optimized using *co-ordinate ascent*, i.e. in each iteration each parameter is set to its optimal value given the remaining parameters.

Other approaches which apply the methodology of variational inference directly to the non-conjugate model have the disadvantages that in most cases the updates are not computed in closed-form and inference is slower due to inefficient Euclidean gradient updates. In the experiments we show that our approach achieves speed-ups up to two orders of magnitude.

3 `AugmentedGaussianProcesses.jl` : A Julia package for fast inference in augmented GP models

We implement a framework for augmented stochastic variational inference (ASVI) for non-conjugate GP models in the programming language Julia (Bezanson et al., 2017). Our implementation focuses on:

1. Fast and stable inference for applying non-conjugate GP models to real-world applications.
2. Flexible framework for research which supports easy implementation of new GP models.
3. Implementation in the dynamic programming language Julia for fast code execution.

We decided to implement the package in Julia because there is a growing interest in Julia in the scientific community. Julia has the advantage that it has the syntax simplicity of Python while having the efficiency of C. It can be easily bridged with other languages (R, Matlab, Python,. . .) and parallelization and GPU support are straight-forward.

However, there are only two GP packages available in Julia. First, GaussianProcesses.jl[2] which offers a large variety of likelihoods, but relies on MCMC sampling and gradient descent for optimization, it is therefore quite slow and scale badly to big datasets. Second, Stheno.jl[3] which focuces on multi-output regression and is not scalable. Our package tries to close that gap aiming on efficient inference for non-conjugate GP models.

4 Experiments: `AugmentedGaussianProcesses.jl` in practice

We discuss three different non-conjugate GP models and present conditionally conjugate augmentations. We compare our method against GPFlow (Matthews et al., 2017) which is a state-of-the-art GP inference package implemented in TensorFlow and against the approach by Henao et al. (2014) in the last experiment. For every experiment we use 50 inducing points, squared exponential kernel with automatic relevance determination (one length-scale parameter per dimension) and optimize the hyperparameters using Adam (Kingma and Ba, 2014).

4.1 GP classification with a logistic likelihood

The logistic likelihood is a prominent alternative to the probit link (Mandt et al., 2017) for obtaining a GP classification model. In former work, Wenzel et al. (2018) develop a scalable variational inference algorithm building on a conditionally conjugate augmentation of the logistic function. The augmentation is

$$\sigma(z) = \frac{1}{1+e^{-z}} = \frac{1}{2}\int_0^\infty \frac{1}{2}\exp\left(\frac{z}{2} - \frac{z^2}{2}\omega\right)\mathrm{PG}(\omega|1,0)d\omega,$$

[2] `https://github.com/STOR-i/GaussianProcesses.jl`

[3] `https://github.com/willtebbutt/Stheno.jl`

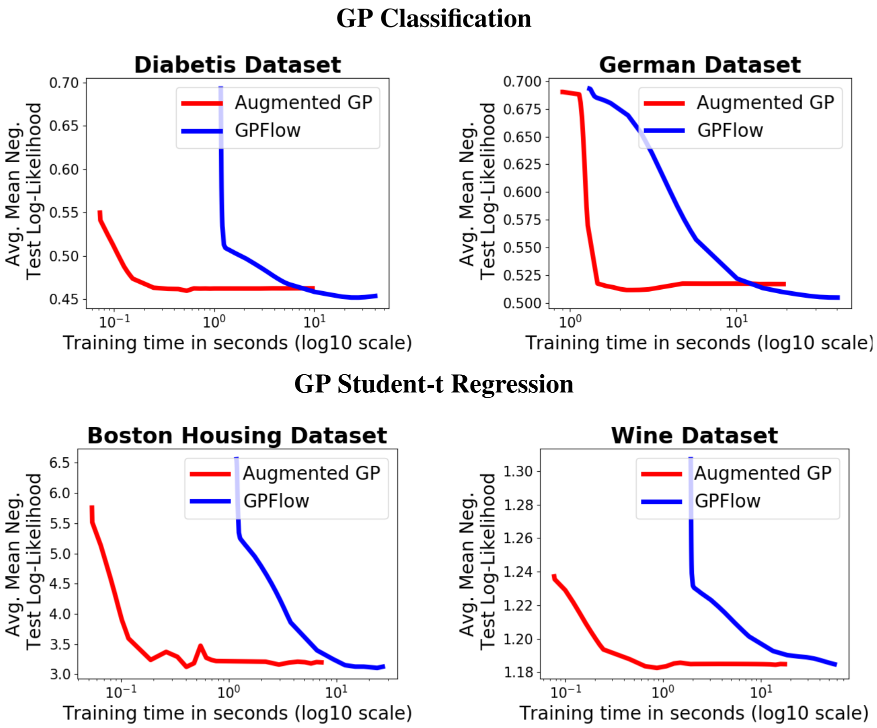

Figure 1: TOP: GP Classification model (logistic likelihood). BOTTOM: Robust GP regression model (Student-t likelihood). For both models, we plot the negative test log-likelihood as a function of time (log-scale) for augmented GP (ours) and GPFlow. We experiment on the UCI datasets Diabetes, German, Boston Housing and Wine.

where PG denotes the Pólya-Gamma distribution (Polson et al., 2013) and $z = yf$. We compare the augmented GP method to the approach by Salimbeni et al. (2018) implemented in GPflow and display the results in figure 1 (top).

4.2 Robust GP regression with a Student-t likelihood

The Student-t likelihood is a more robust approach to the regression problem than the Gaussian likelihood,. We build on the augmentation presented by Jylänki et al. (2011),

$$t_\nu(yf) = \int_0^\infty \mathcal{N}\left(yf|0,\omega\right)\mathcal{IG}\left(\omega|\frac{\nu}{2},\frac{\nu}{2}\right) d\omega,$$

where t_ν denotes the Student-t distribution with ν is the degree of freedom and $\mathcal{IG}$ is the inverse gamma distribution. We compare our approach to GPflow and present the results in figure 1 (bottom).

4.3 Bayesian SVM

This last example is somewhat a bit peculiar as the likelihood comes from a Bayesian interpretation of the support vector machine (SVM) algorithm introduced by Polson et al. (2011). Wenzel et al. (2017) present a scalable variational inference algorithm based on the augmentation of the hinge loss

$$\max(1 - yf, 0) = -\frac{1}{2}\log\int_0^\infty \frac{1}{\sqrt{2\pi\omega}}\exp\left(-\frac{(\omega + 1 - yf)^2}{2\omega}\right) d\omega$$

introduced by (Henao et al., 2014). Up to our knowledge the only competing state-of-the-art inference method for the Bayesian SVM is an expectation conditional maximization (ECM) approach proposed by Henao et al. (2014). On the UCI dataset Waveform, we obtain the following prediction errors, Augmented GP: 0.09 ± 0.02, ECM: 0.10 ± 0.92 and the the following run times, Augmented GP: 12.5 sec, ECM: 264 sec. More results can be found in Wenzel et al. (2017).

Acknowledgments

We like to thank Simon Danisch for helpful comments on the implementation in Julia. This work was partly funded by the German Research Foundation (DFG) award KL 2698/2-1, by the Federal Ministry of Science and Education (BMBF) awards 031B0187B, 01IS18051A and a GENES doctoral research scholarship.

References

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*.

Buchholz, A., Wenzel, F., and Mandt, S. (2018). Quasi-monte carlo variational inference. *ICML*.

Dezfouli, A. and Bonilla, E. V. (2015). Scalable inference for gaussian process models with black-box likelihoods. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1414–1422.

Henao, R., Yuan, X., and Carin, L. (2014). Bayesian Nonlinear Support Vector Machines and Discriminative Factor Modeling. *NIPS*.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. In *Conference on Uncertainty in Artificial Intellegence*.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic Variational Inference. *JMLR*.

Jähnichen, P., Wenzel, F., Kloft, M., and Mandt, S. (2018). Scalable generalized dynamic topic models. In *AISTATS*.

Jylänki, P., Vanhatalo, J., and Vehtari, A. (2011). Robust gaussian process regression with a student-t likelihood. *Journal of Machine Learning Research*, 12(Nov):3227–3257.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Mandt, S., Wenzel, F., Nakajima, S., Cunningham, J. P., Lippert, C., and Kloft, M. (2017). Sparse Probit Linear Mixed Model. *Machine Learning Journal*.

Matthews, A. G. d. G., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., Ghahramani, Z., and Hensman, J. (2017). GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, 18(40):1–6.

Palmer, J., Kreutz-Delgado, K., Rao, B. D., and Wipf, D. P. (2006). Variational em algorithms for non-gaussian latent variable models. In Weiss, Y., Schölkopf, B., and Platt, J. C., editors, *Advances in Neural Information Processing Systems 18*, pages 1059–1066. MIT Press.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349.

Polson, N. G., Scott, S. L., et al. (2011). Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–23.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Salimbeni, H., Eleftheriadis, S., and Hensman, J. (2018). Natural gradients in practice: Non-conjugate variational inference in gaussian process models. *AISTATS*.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, (1-2):1–305.

Wenzel, F., Deutsch, M., Galy-Fajou, T., and Kloft, M. (2017). Bayesian nonlinear support vector machines for big data. *ECML PKDD*.

Wenzel, F., Galy-Fajou, T., Donner, C., Kloft, M., and Opper, M. (2018). Efficient gaussian process classification using polya-gamma data augmentation. *arXiv*.