Gaussian Processes for Big Data through Stochastic Variational Inference

James Hensman
Department of Computer Science
The University of Sheffield
james.hensman@shef.ac.uk

Neil Lawrence
Department of Computer Science
The University of Sheffield
n.lawrence@dcs.shef.ac.uk

1 Introduction

Gaussian processes [GP 10] are perhaps the dominant approach for inference on functions. They underpin a range of algorithms for regression, classification and unsupervised learning. Unfortunately, exact inference in a GP has complexity $\mathcal{O}(n^3)$ with storage demands of $\mathcal{O}(n^2)$ and this hinders application of these models for 'big data'. Various approximate techniques have been suggested [see e.g. 1, 11, 9, 12] which lead to a computational complexity of $\mathcal{O}(nm^2)$ and storage demands of $\mathcal{O}(nm)$ where m is a user selected parameter governing a number of "inducing variables". However, even the reduced storage requirements which are linear in the data set size are prohibitive for big data, where n can be many millions. For parametric models, stochastic gradient descent is often applied to resolve this storage issue, but in the GP domain, it hasn't been clear how this should be performed. In this paper we show how recent advances in variational inference [5, 6] can be combined with the idea of inducing variables to develop a practical algorithm for fitting GPs based around stochastic variational inference (SVI).

2 Sparse GPs Revisited

We start with a succint rederivation of the variational approach to inducing variables of Titsias [12]. This allows us to introduce notation and derive expressions which allow for the formulation of a SVI algorithm.

Consider a data vector[1] $\mathbf{y}$, where each entry y_i is a noisy observation of the function $f(\mathbf{x}_i)$, for all the points $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$. We consider the noise to be independent Gaussian with precision β. Introducing a Gaussian process prior over $f(\cdot)$, let the vector $\mathbf{f}$ contain values of the function at the points $\mathbf{X}$. We shall also introduce a set of *inducing variables*: let the vector $\mathbf{u}$ contain values of the function f at the points $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^m$ which live in the same space as $\mathbf{X}$. Using standard Gaussian process methodologies, we can write

$$p(\mathbf{y} \mid \mathbf{f}) = \mathcal{N}\left(\mathbf{y}|\mathbf{f}, \beta^{-1}\mathbf{I}\right), \quad p(\mathbf{f} \mid \mathbf{u}) = \mathcal{N}\left(\mathbf{f}|\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{u}, \widetilde{\mathbf{K}}\right), \quad p(\mathbf{u}) = \mathcal{N}\left(\mathbf{u}|\mathbf{0}, \mathbf{K}_{mm}\right), \quad (1)$$

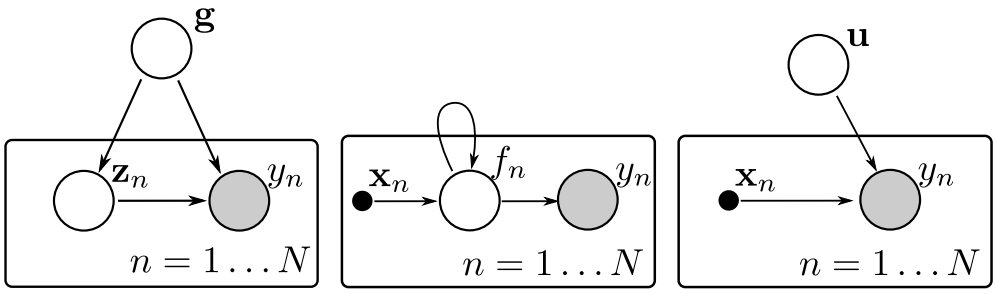
where $\mathbf{K}_{mm}$ is the covariance function evaluated between all the inducing points and $\mathbf{K}_{nm}$ is the covariance function between all inducing points and training points and we have defined with $\widetilde{\mathbf{K}} = \mathbf{K}_{nn} - \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}$.

We first apply Jensen's inequality on the conditional probability $p(\mathbf{y} \mid \mathbf{u})$:

$$\log p(\mathbf{y} \mid \mathbf{u}) = \log \left\langle p(\mathbf{y} \mid \mathbf{f}) \right\rangle_{p(\mathbf{f} \mid \mathbf{u})} \geq \left\langle \log p(\mathbf{y} \mid \mathbf{f}) \right\rangle_{p(\mathbf{f} \mid \mathbf{u})} \triangleq \mathcal{L}_1. \quad (2)$$

where $\left\langle g(x) \right\rangle_{p(x)}$ denotes the expectation of $g(x)$ under $p(x)$. For Gaussian noise the integral on the left is tractable, but it results in an expression containing $\mathbf{K}_{nn}^{-1}$, which has a computational

[1] Our derivation trivially extends to multiple independent output dimensions, but we omit them here for clarity.



(a) Requirements for SVI (b) Gaussian Process regression (c) Variational GP regression

Figure 1: Graphical models showing (a) the reqired form for a probabilistic model for SVI (reproduced from [6]), with *global* variables $\mathbf{g}$ and latent variables $\mathbf{z}$. (b) The graphical model corresponding to Gaussian process regression, where connectivity between the values of the function f_i is denoted by a loop around the plate. (c) The graphical model corresponding to the sparse GP model, with inducing variables $\mathbf{u}$ working as global variables, and the term $\mathcal{L}_1$ acting as $\log p(y_i \,|\, \mathbf{u}, \mathbf{x}_i)$. Marginalisation of $\mathbf{u}$ leads to the variational DTC formulation, introducing dependencies between the observations.

complexity of $\mathcal{O}(n^3)$, whilst computation of $\mathcal{L}_1$ has complexity $\mathcal{O}(m^3)$. In fact this lower bound can be show to be separable across $\mathbf{y}$ allowing us to write,

$$\exp(\mathcal{L}_1) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|\mu_i, \beta^{-1}\right) \exp(-\frac{1}{2}\beta\tilde{k}_{i,i}) \tag{3}$$

where $\boldsymbol{\mu} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{u}$ and $\tilde{k}_{i,i}$ is the ith diagonal element of $\widetilde{\mathbf{K}}$. It is clear that when the inducing variables are placed at the training data locations (i.e. $\mathbf{u} = \mathbf{f}$, $\mathbf{K}_{mm}$=$\mathbf{K}_{nm}$=$\mathbf{K}_{nn}$ so $\widetilde{\mathbf{K}}$=$\mathbf{0}$) we recover $\exp(\mathcal{L}_1) = p(\mathbf{y}|\mathbf{f})$ and the bound becomes equality because $p(\mathbf{f}|\mathbf{u})$ is degenerate. However, since $m = n$ there is no gain in computational efficiency in this case. When $m < n$ the quality of the bound is improved by ensuring that $\mathbf{Z}$ (which are variational parameters) are distributed amongst the training data $\mathbf{X}$ such that all $\tilde{k}_{i,i}$ are small: this ensures that the expectations in (2) are only taken across a narrow domain ($\tilde{k}_{i,i}$ is the marginal variance of $p(f_i|\mathbf{u})$). This keeps Jensen's bound tight.

To recover the bound of Titsias [12] we now marginalize the inducing variables

$$\log p(\mathbf{y} \,|\, \mathbf{X}) = \log \int p(\mathbf{y} \,|\, \mathbf{u}) p(\mathbf{u}) \, \mathrm{d}\mathbf{u} \geq \log \int \exp\left\{\mathcal{L}_1\right\} p(\mathbf{u}) \, \mathrm{d}\mathbf{u} \triangleq \mathcal{L}_2, \tag{4}$$

which, with some linear algebraic manipulation, leads to

$$\mathcal{L}_2 = \log \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn} + \beta^{-1}\mathbf{I}\right) - \frac{1}{2}\beta \mathrm{tr}\left(\widetilde{\mathbf{K}}\right), \tag{5}$$

matching the result of Titsias, with the implicit approximating distribution $q(\mathbf{u})$ having precision $\boldsymbol{\Lambda} = \beta\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1} + \mathbf{K}_{mm}^{-1}$ and mean $\hat{\mathbf{u}} = \beta\boldsymbol{\Lambda}^{-1}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}\mathbf{y}$.

3 SVI for GPs

The novel approach in this paper is to retain the inducing variables explicitly in our representation. Stochastic variational inference (SVI) allows variational inference for very large data sets, but it can only be applied to probabilistic models which have a set of *global* variables, and which factorise in the observations and latent variables as Figure 1(a). Gaussian Processes do not have global variables and exhibit no such factorisation (Figure 1(b)). By introducing inducing variables $\mathbf{u}$, we have an appropriate model for SVI (Figure 1(c)). Unfortunately, marginalising $\mathbf{u}$ re-introduces dependencies between the observations, and eliminates the global parameters. In the following, we derive a lower bound on $\mathcal{L}_2$ which includes an explicit variational distribution $q(\mathbf{u})$, enabling SVI. We then derive the required natural gradients and discuss how latent variables might be used.

3.1 Global variables

To apply SVI to a GP model, we must have a set of global variables. The variables $\mathbf{u}$ will perform this role, and we introduce a variational distribution $q(\mathbf{u})$, and use it to lower bound the quantity $p(\mathbf{y} \mid \mathbf{X})$.

$$\log p(\mathbf{y} \mid \mathbf{X}) \geq \langle \mathcal{L}_1 + \log p(\mathbf{u}) - \log q(\mathbf{u}) \rangle_{q(\mathbf{u})} \triangleq \mathcal{L}_3. \tag{6}$$

From the above we know that the optimal distribution is Gaussian, and we parametrise it as $q(\mathbf{u}) = \mathcal{N}(\mathbf{u} \mid \mathbf{m}, \mathbf{S})$. Thus the bound $\mathcal{L}_3$ becomes

$$\mathcal{L}_3 = \sum_{i=1}^{n} \left\{ \log \mathcal{N}(y_i \mid \mathbf{k}_{mn}^\top \mathbf{K}_{mm}^{-1} \mathbf{m}, \beta^{-1}) - \frac{1}{2}\beta \widetilde{k}_{i,i} - \frac{1}{2}\mathrm{tr}\left(\mathbf{S}\mathbf{\Lambda}_i\right) \right\} - \mathrm{KL}\left[q(\mathbf{u}) || p(\mathbf{u})\right] \tag{7}$$

with $\mathbf{k}_{mn}$ being a vector of the i^{th} column of $\mathbf{K}_{mn}$ and $\mathbf{\Lambda}_i = \beta \mathbf{K}_{mm}^{-1} \mathbf{k}_{mn} \mathbf{k}_{mn}^\top \mathbf{K}_{mm}^{-1}$. The gradients of $\mathcal{L}_3$ with respect to the parameters of $q(\mathbf{u})$ are

$$\frac{\partial \mathcal{L}_3}{\partial \mathbf{m}} = \beta \mathbf{K}_{mm}^{-1} \mathbf{K}_{mn} \mathbf{y} - \mathbf{\Lambda} \mathbf{m}, \quad \frac{\partial \mathcal{L}_3}{\partial \mathbf{S}} = \frac{1}{2}\mathbf{S}^{-1} - \frac{1}{2}\mathbf{\Lambda}. \tag{8}$$

It is immediately apparent that setting the derivatives to zero recovers the optimal solution found in the previous section, namely $\mathbf{S}{=}\mathbf{\Lambda}^{-1}$, $\mathbf{m}{=}\hat{\mathbf{u}}$. It follows that $\mathcal{L}_2 \geq \mathcal{L}_3$, with equality at this unique maximum.

The key propery of $\mathcal{L}_3$ is that is can be written as a sum of n terms, each corresponding to one input-output pair $\{\mathbf{x}_i, y_i\}$: we have induced the necessary factorisation to perform stochastic gradient methods on the distribution $q(\mathbf{u})$.

3.2 Natural gradients

Stochastic variational inference works by taking steps in the direction of the approximate *natural* gradient $\widetilde{\mathbf{g}}(\boldsymbol{\theta})$, which is given by the usual gradient re-scaled by the inverse Fisher information: $\widetilde{\mathbf{g}}(\boldsymbol{\theta}){=}G(\boldsymbol{\theta})^{-1}\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}$. To work with the natural gradients of the distribution $q(\mathbf{u})$, we first recall the canonical and expectaion parameters $\boldsymbol{\theta}_1{=}\mathbf{S}^{-1}\mathbf{m}$, $\boldsymbol{\theta}_2{=}-\frac{1}{2}\mathbf{S}^{-1}$ and $\boldsymbol{\eta}_1{=}\mathbf{m}$, $\boldsymbol{\eta}_2{=}\mathbf{m}\mathbf{m}^\top{+}\mathbf{S}$. In the exponential family, properties of the Fisher information reveal the following simplification of the natural gradient [5]:

$$\widetilde{\mathbf{g}}(\boldsymbol{\theta}) = G(\boldsymbol{\theta})^{-1}\frac{\partial \mathcal{L}_3}{\partial \boldsymbol{\theta}} = \frac{\partial \mathcal{L}_3}{\partial \boldsymbol{\eta}}. \tag{9}$$

A step of length ℓ in the natural gradient direction, using $\boldsymbol{\theta}_{(t+1)} = \boldsymbol{\theta}_{(t)} + \ell\frac{\mathrm{d}\mathcal{L}_3}{\mathrm{d}\boldsymbol{\eta}}$, yields

$$\begin{aligned} \boldsymbol{\theta}_{2(t+1)} &= -\frac{1}{2}\mathbf{S}_{(t+1)}^{-1} = -\frac{1}{2}\mathbf{S}_{(t)}^{-1} + \ell(-\frac{1}{2}\mathbf{\Lambda} + \frac{1}{2}\mathbf{S}_{(t)}^{-1}), \\ \boldsymbol{\theta}_{1(t+1)} &= \mathbf{S}_{(t+1)}^{-1}\mathbf{m}_{(t+1)} = \mathbf{S}_{(t)}^{-1}\mathbf{m}_{(t)} + \ell(\beta\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}\mathbf{y} - \mathbf{S}_{(t)}^{-1}\mathbf{m}_{(t)}), \end{aligned} \tag{10}$$

and taking a step of unit length then recovers the same solution as above by either (4) or (8). It can also be shown that taking this unit step is the same as performing a VB update [5, 6]. We can now obtain stochastic approximations to the natural gradient by considering the data either individually or in mini-batches. We follow a similar procedure for the kernel hyper-parameters and the noise precision β. An illustration is presented in Figure 2.

3.3 Latent variables

The above derivations enable online learning for Gaussian process *regression* using SVI. Several GP based models involve inference of $\mathbf{X}$, such as the GP latent variable model [8, 13] and its extensions [e.g. 2, 3].

To perform stochastic variational inference with latent variables, we require a factorisation as illustrated by Figure 1(a): this factorisation is provided by (7). To get a model like the Bayesian GPLVM, we need a lower bound on $\log p(\mathbf{y})$. In Titsias and Lawrence [13] this was achieved through approximate marginalisation of $\mathcal{L}_2$, w.r.t. $\mathbf{X}$, which leads to an expression depending only

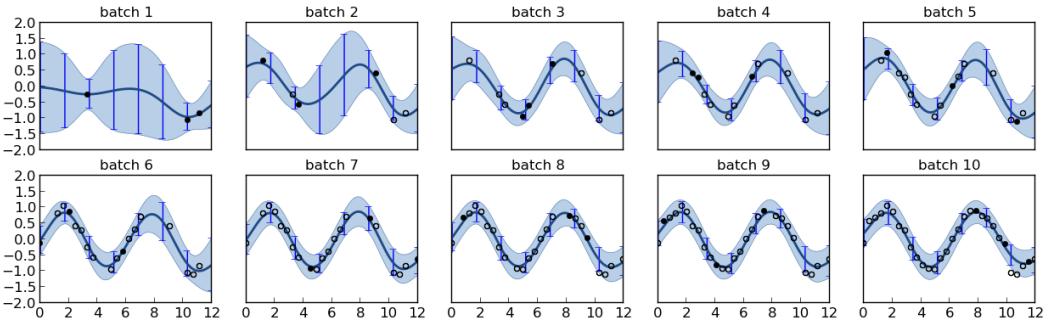


Figure 2: Stochastic variational inference on a trivial GP regression problem. Each pane shows the posterior of the GP after a batch of data, marked as solid points. Previoulsy seen (and discarded) data are marked as empty points, the distribution $q(\mathbf{u})$ is represented by vertical errorbars.

on the parameters of $q(\mathbf{X})$. However this formulation scales poorly, and the variables of the optimisation are closely connected due to the marginalisation of $\mathbf{u}$. The above enables a lower bound to which SVI is immediately applicable:

$$\log p(\mathbf{y}) = \log \int p(\mathbf{y} \mid \mathbf{X}) p(\mathbf{X}) \, \mathrm{d}\mathbf{X} \geq \int q(\mathbf{X}) \left\{ \mathcal{L}_3 + \log p(\mathbf{X}) - \log q(\mathbf{X}) \right\} \mathrm{d}\mathbf{X}. \tag{11}$$

It is straightforward to introduce D output dimensions for the data $\mathbf{Y}$, and following Titsias and Lawrence [13], we use a factorising normal distribution $q(\mathbf{X}) = \prod_{i=1}^{n} q(\mathbf{x}_i)$. The relevant expectations of $\mathcal{L}_3$ are tractable for various choices of covariance function.

To perform SVI in this model, we now alternate between selecting a minibatch of data, and optimisting the relevant variables of $q(\mathbf{X})$ with $q(\mathbf{u})$ fixed, and updating $q(\mathbf{u})$ using the approximate natural gradient. We note that the form of (7) means that each of the latent variable distributions may be updated individually, enabling parallisation across the minibatch.

3.4 Non-Gaussian likelihoods

Another advantage of the factorisation of (7) is that it enables a simple routine for inference with non-Gaussian likelihoods. The usual procedure for fitting GPs with non-Gaussian likelihoods is to approximate the likelihood using either a local variational lower bound [4], or by expectation propagation [7]. These approximations to the likelihood are required because of the connections between the variables $\mathbf{f}$.

In $\mathcal{L}_3$, the bound factorises in such a way that some non-Gaussian likelihoods may be marginalised *exactly*, given the existing approximations. To see this, consider that we are presented not with the vector $\mathbf{y}$, but by a binary vector $\mathbf{t}$ with $t_i \in \{0, 1\}$, and the likelihood $p(\mathbf{t} \mid \mathbf{y}) = \prod_{i=1}^{n} \sigma(y_i)^{t_i} (1 - \sigma(y_i))^{(1-t_i)}$, as in the case of classification. We can bound the marginal likelihood using $p(\mathbf{t} \mid \mathbf{X}) \geq \int p(\mathbf{t} \mid \mathbf{y}) \exp\{\mathcal{L}_3\} \, \mathrm{d}\mathbf{y}$ which involves n independent one dimensional integrals due to the factorising nature of $\mathcal{L}_3$. For the probit likelihood each of these integrals is tractable.

This kind of approximation, where the likelihood is integrated exactly is amenable to SVI in the same manner as the regression case above through computation of the natural gradient.

4 Discussion

We have presented a method for inference in Gaussian process models using stochastic variational inference. These expression potentially allow for the transfer of a multitude of Gaussian process techniques to big data.

References

[1] Lehel Csató and Manfred Opper. Sparse on-line Gaussian processes. *Neural Computation*, 14 (3):641–668, 2002.

[2] Andreas Damianou, Michalis K. Titsias, and Neil D. Lawrence. Variational gaussian process dynamical systems. In Peter Bartlett, Fernando Peirrera, Chris Williams, and John Lafferty, editors, *Advances in Neural Information Processing Systems*, volume 24, Cambridge, MA, 2011. MIT Press.

[3] Andreas Damianou, Carl Henrik Ek, Michalis K. Titsias, and Neil D. Lawrence. Manifold relevance determination. In John Langford and Joelle Pineau, editors, *Proceedings of the International Conference in Machine Learning*, volume 29, San Francisco, CA, 2012. Morgan Kauffman. To appear.

[4] Mark N. Gibbs and David J. C. MacKay. Variational Gaussian process classifiers. *IEEE Transactions on Neural Networks*, 11(6):1458–1464, 2000.

[5] James Hensman, Magnus Rattray, and Neil D. Lawrence. Fast variational inference in the exponential family. *To appear at NIPS 2012*, 2012.

[6] Matthew Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *arXiv preprint arXiv:1206.7051*, 2012.

[7] Malte Kuss and Carl Edward Rasmussen. Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research*, 6:1679–1704, 2005.

[8] Neil D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 11 2005.

[9] Joaquin Quiñonero Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

[10] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. ISBN 0-262-18253-X.

[11] Matthias Seeger, Christopher K. I. Williams, and Neil D. Lawrence. Fast forward selection to speed up sparse Gaussian process regression. In Christopher M. Bishop and Brendan J. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, Key West, FL, 3–6 Jan 2003.

[12] Michalis K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In David van Dyk and Max Welling, editors, *Proceedings of the Twelfth International Workshop on Artificial Intelligence and Statistics*, volume 5, pages 567–574, Clearwater Beach, FL, 16-18 April 2009. JMLR W&CP 5.

[13] Michalis K. Titsias and Neil D. Lawrence. Bayesian Gaussian process latent variable model. In Yee Whye Teh and D. Michael Titterington, editors, *Proceedings of the Thirteenth International Workshop on Artificial Intelligence and Statistics*, volume 9, pages 844–851, Chia Laguna Resort, Sardinia, Italy, 13-16 May 2010. JMLR W&CP 9.