Fast Variational Inference for Gaussian Process Models through KL-Correction

Nathaniel J. King and Neil D. Lawrence

Department of Computer Science,
University of Sheffield, Regent Court,
211 Portobello Street, Sheffield,
S1 4DP. United Kingdom
{nat,neil}@dcs.shef.ac.uk

Abstract Variational inference is a flexible approach to solving problems of intractability in Bayesian models. Unfortunately the convergence of variational methods is often slow. We review a recently suggested variational approach for approximate inference in Gaussian process (GP) models and show how convergence may be dramatically improved through the use of a positive correction term to the standard variational bound. We refer to the modified bound as a KL-corrected bound. The KL-corrected bound is a lower bound on the true likelihood, but an upper bound on the original variational bound. Timing comparisons between optimisation of the two bounds show that optimisation of the new bound consistently improves the speed of convergence.

1 Introduction

A key problem with many variational approximations is the slow speed of convergence. In this paper we will show how the speed of convergence for variational approximations can be radically improved by 'KL-correction' of the variational bound. Empirically we find that our approach dramatically improves convergence speed for a range of benchmark data sets.

We consider the variational approximation proposed independently by [1] and [2]. This approximation allows us to consider the process of inference in the Gaussian process independently of the noise model [2]. We follow [2] in referring to this formulation of the variational approach as probabilistic point assimilation (PPA).

The paper is laid out as follows, in Sections 2 and 3, we introduce notation and describe the underlying probabilistic model, as well as the PPA variational approximation and the KL-corrected bound. In Section 4 we demonstrate the performance of the approach on some benchmark data sets, including timing comparisons, and we conclude in Section 5 with a short discussion.

2 Gaussian Processes

Consider a data set consisting of input data, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^{\mathrm{T}}$, and labels, $\mathbf{y} = [y_1, \ldots, y_N]^{\mathrm{T}}$. We will assume that the labels are dependent on an $N \times 1$

vector, $\mathbf{f} = [f_1, \ldots, f_N]^{\mathrm{T}}$ through a 'noise model' $p(y_n|f_n)$. The label y_n relates to $\mathbf{x}_n$ through the latent variable f_n. In the case of our simple classification noise model the relationship to f_n is given by,

$$p(y_n|f_n) = \phi(y_n f_n),$$

where $\phi(z) = \int_{-\infty}^{z} N(t|0,1)\, dt$ is the cumulative Gaussian distribution function and $N(\mathbf{z}|\boldsymbol{\mu}, \Sigma)$ denotes a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance Σ.

The latent variable is normally then related to the input data through a Gaussian process prior [3,4] over $\mathbf{f}$. For the moment we depart from this approach and define an additional spherical distribution over $\mathbf{f}$,

$$p(\mathbf{f}|\bar{\mathbf{f}}, \beta) = \prod_{n=1}^{N} p(f_n|\bar{f}_n, \beta) = \prod_{n=1}^{N} N(f_n|\bar{f}_n, \beta^{-1}),$$

where the β is a precision (inverse variance), and $\bar{\mathbf{f}}$ is a vector of means, the nth element being $\bar{f}_n$. Clearly under this definition $\mathbf{y}$ is independent of $\mathbf{X}$, to rectify this we now introduce a prior distribution over $\bar{\mathbf{f}}$,

$$p(\bar{\mathbf{f}}|\mathbf{X}, \boldsymbol{\theta}) = N(\bar{\mathbf{f}}|\mathbf{0}, \mathbf{K}),$$

which is a Gaussian process prior over $\bar{\mathbf{f}}$ with a mean of zero and a covariance function $\mathbf{K}$. This matrix is a function of $\mathbf{X}$ and its form is controlled by a set of parameters, $\boldsymbol{\theta}$. Note that this prior distribution can be combined with our distribution over $\mathbf{f}$ to obtain

$$p(\mathbf{f}|\mathbf{0}, \mathbf{K}) = \int \prod_{n=1}^{N} p(f_n|\bar{f}_n, \beta)\, N(\bar{\mathbf{f}}|\mathbf{0}, \mathbf{K})\, d\bar{\mathbf{f}} = N(\mathbf{f}|\mathbf{0}, \mathbf{K} + \beta^{-1}\mathbf{I}),$$

which, since a diagonal term is often added to the kernel matrix, does not in practice lead to a richer model. However, as we shall see, augmentation of the basic model with the vector of means $\bar{\mathbf{f}}$ renders the application of variational approaches to the model more convenient.

The marginal likelihood of a data set can be obtained through marginalisation of the latent variables $\mathbf{f}$ and $\bar{\mathbf{f}}$,

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \beta) = \int N(\bar{\mathbf{f}}|\mathbf{0}, \mathbf{K}) \prod_{n=1}^{N} \int p(y_n|f_n)\, p(f_n|\bar{f}_n, \beta)\, df_n d\bar{\mathbf{f}}. \tag{1}$$

In practise we will find that for non-Gaussian noise models this marginal likelihood will not be tractable, forcing us to turn to approximate methods.

2.1 Variational Inference

Variational inference is a popular choice for approximate inference in Bayesian models. In [2] we showed how to implement variational inference in Gaussian processes in a generic manner, we refered to this approach as probabilistic point assimilation (PPA). The same approach was also independently suggested by [1] in the context of multi-class classification in Gaussian processes.

The first step in PPA is to introduce an approximating distribution, $q\left(\bar{\mathbf{f}}\right)$, for the mean parameters giving

$$\begin{aligned}\log p\left(\mathbf{y}|\mathbf{X},\boldsymbol{\theta},\beta\right) \geq \sum_{n=1}^{N} \left\langle \log p\left(y_n|\bar{f}_n,\beta\right)\right\rangle_{q\left(\bar{\mathbf{f}}\right)} + \left\langle \log p\left(\bar{\mathbf{f}}|\mathbf{X},\boldsymbol{\theta}\right)\right\rangle_{q\left(\bar{\mathbf{f}}\right)} \\ - \left\langle \log q\left(\bar{\mathbf{f}}\right)\right\rangle_{q\left(\bar{\mathbf{f}}\right)}. \qquad (2)\end{aligned}$$

This is in effect the standard variational formalism for Gaussian processes. Ideally we would now seek to maximise the bound through free-form optimisation with respect to $q\left(\bar{\mathbf{f}}\right)$ [5]. Unfortunately, for most noise models, such a free form optimisation of the bound is not possible. The next step is, therefore, to assume a form for $q\left(\bar{\mathbf{f}}\right)$ which renders the bound tractable. Seeger [6] made the natural assumption that $q\left(\bar{\mathbf{f}}\right)$ is a Gaussian process and sought its mean and covariance by maximising the resulting bound. Unfortunately this approach greatly complicates the process of inference as it demands gradient based optimisation of the variational bound, which for practicality often requires further constraints on the posterior covariance matrix. In PPA we depart from the standard approach through introduction of a further approximating distribution, $q\left(\mathbf{f}\right)$, to lower bound the first term of (2),

$$\begin{aligned}\log p\left(\mathbf{y}|\mathbf{X},\boldsymbol{\theta},\beta\right) \geq \sum_{n=1}^{N} \left\langle \log p\left(y_n|f_n\right)\right\rangle_{q(f_n)} + \sum_{n=1}^{N} \left\langle \log p\left(f_n|\bar{f}_n,\beta\right)\right\rangle_{q\left(\bar{f}_n\right)q(f_n)} \\ + \left\langle \log p\left(\bar{\mathbf{f}}|\mathbf{X},\boldsymbol{\theta}\right)\right\rangle_{q\left(\bar{\mathbf{f}}\right)} - \sum_{n=1}^{N} \left\langle \log q\left(f_n\right)\right\rangle_{q(f_n)} \\ - \left\langle \log q\left(\bar{\mathbf{f}}\right)\right\rangle_{q\left(\bar{\mathbf{f}}\right)} = \mathcal{L}. \qquad (3)\end{aligned}$$

Each of the two lower bounds we have made use of can independently be made to be equalities if their variational distributions are optimised, however when combined they will only reach equality if the true posterior distribution factorises. For later convenience we shall refer to this bound (3) as the standard variational approach. The key advantage associated with introduction of the second lower bound is that we can now perform free-form optimisation of the posterior approximations [5] in the manner of standard variational inference. Under free-form optimisation it turns out that the approximating distribution over $\mathbf{f}$ factorises, $q\left(\mathbf{f}\right) = \prod_{n=1}^{N} q\left(f_n\right)$, with each factor being given by

$$q\left(f_n\right) \propto p\left(y_n|f_n\right) \exp \left\langle \log p\left(f_n|\bar{f}_n,\beta\right)\right\rangle_{q\left(\bar{f}_n\right)}.$$

Recalling that $p\left(f_n|\bar{f}_n,\beta\right)$ is a Gaussian distribution, we can re-write this formula as

$$q\left(f_n\right) = \frac{1}{Z_n} p\left(y_n|f_n\right) N\left(f_n|\left\langle\bar{f}_n\right\rangle,\beta^{-1}\right), \qquad (4)$$

where the normalisation constant is given by Z_n. The tractability of the normalisation constant is dependent on the form of the noise model. However, even when Z_n is analytically intractable, the integral can be solved numerically through quadrature.

Different Noise Models A key advantage of the PPA approach is that we can make use of many different noise models in (4) without significantly changing our algorithm. This is achieved in the following manner. It is well known (see *e.g.* [7]) that expectations under distributions of the form given in (4) can be computed through differentiation of $\log Z_n$. The mean of (4) can be shown to be

$$\langle f_n \rangle = \langle \bar{f}_n \rangle + \beta^{-1} g_n$$

where $g_n = \nabla_{\langle \bar{f}_n \rangle} \log Z_n$ and the second moment can be shown to be

$$\langle f_n^2 \rangle = 2\beta^{-2} \Gamma_n + \beta^{-1} + 2 \langle \bar{f}_n \rangle \langle f_n \rangle - \langle \bar{f}_n \rangle^2$$

with $\Gamma_n = \nabla_{\beta_n^{-1}} \log Z_n$. For a given noise model of interest, it is therefore only necessary to compute $\log Z_n = \log \int p\left(y_n | f_n\right) N\left(f_n | \langle \bar{f}_n \rangle, \beta^{-1}\right) df_n$ for it to be used in the inference process. This was our main motivation in describing this model within [2].

Approximating Distribution for $\bar{\mathbf{f}}$ The moments under $q\left(f_n\right)$ can be used to find the form of the approximating component associated with the mean vector $\bar{\mathbf{f}}$. Free-form optimisation of the variational bound with respect to $q\left(\bar{\mathbf{f}}\right)$ recovers

$$q\left(\bar{\mathbf{f}}\right) \propto p\left(\bar{\mathbf{f}} | \mathbf{X}, \boldsymbol{\theta}\right) \prod_{n=1}^{N} \exp \left\langle \log p\left(f_n | \bar{f}_n, \beta_n\right) \right\rangle . \qquad (5)$$

This implies that $q\left(\bar{\mathbf{f}}\right)$ has the form of a Gaussian process,

$$q\left(\bar{\mathbf{f}}\right) = N\left(\bar{\mathbf{f}} | \mu, \mathbf{C}\right)$$

whose posterior covariance function is given by $\mathbf{C} = \left(\mathbf{K}^{-1} + \beta \mathbf{I}\right)^{-1}$, while the posterior mean function is given by $\mu = \beta \mathbf{C} \langle \mathbf{f} \rangle$. Computation of the required moments under this process posterior is straightforward, the first moment is given by $\langle \bar{\mathbf{f}} \rangle = \mu$ and the second moment by $\langle \bar{\mathbf{f}} \bar{\mathbf{f}}^{\mathrm{T}} \rangle = \mathbf{C} + \mu \mu^{T}$. Note that the first and second moment of our posterior approximation can be computed by inspection; contrast this with the situation in [6] where these moments must be found through gradient based methods.

We also see that the form of $q\left(\bar{\mathbf{f}}\right)$ is not directly dependent on the form of the noise model. This dependence occurs through the latent variables $\mathbf{f}$.

3 Updating Parameters

One of the advantages of the Gaussian process framework is that we can seek to optimise kernel parameters through optimisation of the model's log-likelihood. In approximate variational inference direct optimisation of the marginal likelihood is not possible; instead we seek to maximise the variational lower bound. For our model the relevant terms of the bound are

$$L\left(\beta,\boldsymbol{\theta}\right)=\left\langle \log p\left(\bar{\mathbf{f}}|\mathbf{X},\boldsymbol{\theta}\right)\right\rangle_{q\left(\bar{\mathbf{f}}\right)}+\sum_{n=1}^{N}\left\langle \log p\left(f_n|\bar{f}_n,\beta\right)\right\rangle_{q\left(\bar{f}_n\right)q\left(f_n\right)}=\mathcal{L}\left(\theta\right) \quad (6)$$

The bound is normally optimised with respect to $\boldsymbol{\theta}$ by gradient based methods.

3.1 KL-Corrected Inference

A common problem with variational methods is slow convergence to a maximum. This can occur if the quality of the bound as a function of the parameters, $\mathcal{L}\left(\boldsymbol{\theta}\right)$, falls away rapidly as $\boldsymbol{\theta}$ changes. In other words convergence will be slow if the quality of the bound is very sensitive to changes in the parameters. The effect is shown in Figure 1(a). The motivation behind this paper was to discover whether we could obtain an upper bound, $\mathcal{L}'\left(\boldsymbol{\theta}\right)$, on (3) which is also a lower bound on the true likelihood, then we are also likely to achieve faster convergence. The intuition behind this idea is shown schematically in Figure 1(b). If $\mathcal{L}'\left(\boldsymbol{\theta}\right)$ is an upper bound on $\mathcal{L}\left(\boldsymbol{\theta}\right)$ and a lower bound on the true likelihood $L\left(\boldsymbol{\theta}\right)$ then its maxima is likely to be closer to the maxima of $L\left(\boldsymbol{\theta}\right)$ than the maxima of $\mathcal{L}\left(\boldsymbol{\theta}\right)$ is.

An Improved Bound Ideally we would like to optimise the marginal likelihood,

$$L\left(\boldsymbol{\theta}\right)=\log p\left(\mathbf{y}|\mathbf{X},\boldsymbol{\theta},\beta\right)=\log\int\prod_{n=1}^{N}p\left(y_n|\bar{f}_n,\beta\right)p\left(\bar{\mathbf{f}}|\mathbf{X},\boldsymbol{\theta}\right)d\bar{\mathbf{f}}, \quad (7)$$

with respect to $\boldsymbol{\theta}$; unfortunately the integral is, in general, intractable. We previously discussed the fact that the log of the noise model can be lower bounded variationally. This lower bound is maintained when taking the exponential of both sides (as the exponential is a monotonic function). Thus, the noise model is lower bounded by

$$\begin{aligned}p\left(y_n|\bar{f}_n,\beta_n\right)\geq\exp\Big(&\left\langle \log p\left(y_n|f_n\right)\right\rangle_{q\left(f_n\right)}+\left\langle \log p\left(f_n|\bar{f}_n,\beta\right)\right\rangle_{q\left(f_n\right)}\\&-\left\langle \log q\left(f_n\right)\right\rangle_{q\left(f_n\right)}\Big). \quad (8)\end{aligned}$$

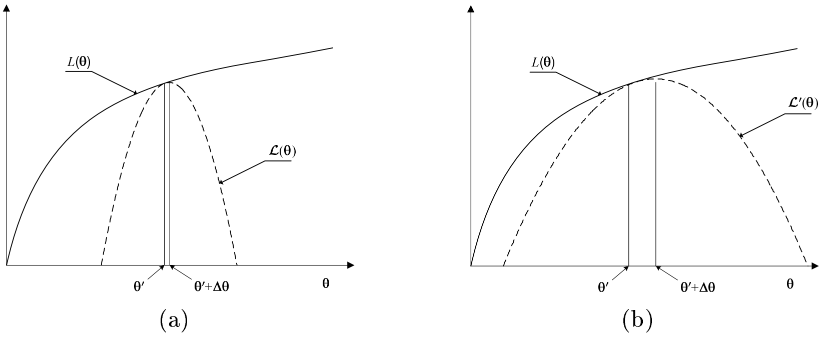


Figure 1. Variational optimisation. (a) The schematic shows the log likelihood, $L(\boldsymbol{\theta})$ as a function of the parameters and the variational lower bound, $\mathcal{L}(\boldsymbol{\theta})$. The lower bound is shown as being quadratic in the parameters. The bound shown has been maximised with respect to the q-distributions for a set of parameters $\boldsymbol{\theta}'$, however the bound falls away sharply for quite small changes in $\boldsymbol{\theta}$. As a result optimisation of the lower bound with respect to $\boldsymbol{\theta}$ leads to only a small change $\Delta\boldsymbol{\theta}$. Many iterations are required for convergence. (b) Here we show a schematic of the effect of KL-correction of the bound. The bound is less sensitive to the variational distributions and it falls away from the likelihood less quickly, as a result larger steps are taken when $\boldsymbol{\theta}$ is optimised.

Substituting this expression into the marginal log likelihood (7) gives the following lower bound

$$\mathcal{L}'(\boldsymbol{\theta}) = \log \int \prod_{n=1}^{N} \exp \left\langle \log p\left(f_n|\bar{f}_n, \beta\right)\right\rangle_{q(f_n)} p\left(\bar{\mathbf{f}}|\mathbf{X}, \boldsymbol{\theta}\right) d\bar{\mathbf{f}} - \sum_{n=1}^{N} \left\langle \log q\left(f_n\right)\right\rangle_{q(f_n)}$$
$$+ \sum_{n=1}^{N} \left\langle \log p\left(y_n|f_n\right)\right\rangle_{q(f_n)} \leq L(\boldsymbol{\theta}) . \tag{9}$$

Note that the only term in this bound which is now dependent on $\boldsymbol{\theta}$ is the first term. The integral in this term can be computed analytically. To see this we first rewrite it as a Gaussian integral,

$$\begin{aligned}\mathcal{L}'(\boldsymbol{\theta}) &= \log \int \prod_{n=1}^{N} \exp \left\langle \log p\left(f_n|\bar{f}_n, \beta\right)\right\rangle_{q(f_n)} p\left(\bar{\mathbf{f}}|\mathbf{X}, \boldsymbol{\theta}\right) d\bar{\mathbf{f}} + \text{const} \\ &= \log \int \prod_{n=1}^{N} N\left(\left\langle f_n\right\rangle |\bar{f}_n, \beta^{-1}\right) p\left(\bar{\mathbf{f}}|\mathbf{X}, \boldsymbol{\theta}\right) d\bar{\mathbf{f}} + \text{const},\end{aligned} \tag{10}$$

leading to a tractable objective function for $\boldsymbol{\theta}$ that does not directly depend on $q\left(\bar{\mathbf{f}}\right)$. The result is a new bound that is actually an upper bound on the original variational lower bound. It thus has the characteristics suggested in Section 3.1 which are conducive to faster convergence.

Positive Correction Term The KL-corrected bound is still a lower bound on the log-likelihood, however it is typically a tighter bound than the standard variational bound: it contains a correction term which is always positive or zero. The KL-corrected bound (9) can be rewritten using (3) as

$$L(\boldsymbol{\theta}) \geq \mathcal{L}(\boldsymbol{\theta}) + \mathrm{KL}\left(q\left(\bar{\mathbf{f}}\right) || p\left(\bar{\mathbf{f}} | \langle \mathbf{f} \rangle, \mathbf{X}, \boldsymbol{\theta}\right)\right)$$

where $\mathrm{KL}\left(q\left(\bar{\mathbf{f}}\right) || p\left(\bar{\mathbf{f}} | \langle \mathbf{f} \rangle \mathbf{X}, \boldsymbol{\theta}\right)\right)$ is the Kullback-Leibler divergence[1] between the distribution $q\left(\bar{\mathbf{f}}\right)$ and

$$p\left(\bar{\mathbf{f}} | \langle \mathbf{f} \rangle, \mathbf{X}, \boldsymbol{\theta}\right) \propto \prod_{n=1}^{N} N\left(\langle f_n \rangle | \bar{f}_n, \beta^{-1}\right) p\left(\bar{\mathbf{f}} | \mathbf{X}, \boldsymbol{\theta}\right).$$

This implies that the difference between the KL-corrected bound and the traditional variational bound is the Kullback-Leibler divergence between $q\left(\bar{\mathbf{f}}\right)$ and $p\left(\bar{\mathbf{f}} | \langle \mathbf{f} \rangle, \mathbf{X}, \boldsymbol{\theta}\right)$. Inspection of (5) shows that this divergence is zero after updates of $q\left(\bar{\mathbf{f}}\right)$. However, as $\boldsymbol{\theta}$ changes the divergence will become non-zero and provide a positive correction to the standard variational bound in the manner depicted in Figure 1(b). The KL-corrected objective is therefore a lower bound on the marginal likelihood and an upper bound on the traditional variational objective. Optimisations of the KL-corrected objective are therefore guaranteed to converge. In Section 4 we will show that empirically this convergence is much faster than that of the standard variational optimisation. Before that we will consider the KL-corrected bound in more detail.

3.2 Inference on the Corrected Bound

So far we have discussed optimising the KL-corrected bound primarily with respect to the parameters of the kernel function. Optimisation with respect to β is also straightforward. However, we have implicitly assumed that we will find $q(f_n)$ by optimising the original variational bound (which also entails optimisation of $q\left(\bar{\mathbf{f}}\right)$). In this section we consider the possibility of updating $q(f_n)$ through optimisation of the KL-corrected bound. To do this we first consider the dependence of the KL-corrected bound (9) on $q(f_n)$. First we make use of the fact that $p\left(f_n | \bar{f}_n, \beta\right) = N\left(f_n | \bar{f}_n, \beta^{-1}\right)$ to rewrite

$$\left\langle \log p\left(f_n | \bar{f}_n, \beta\right)\right\rangle = \log N\left(\langle f_n \rangle | \bar{f}_n, \beta^{-1}\right) - c_n$$

where c_n has the form $c_n = \frac{\beta}{2}\left(\left\langle f_n^2 \right\rangle - \langle f_n \rangle^2\right)$. The integral in the first term of (9) can now be computed analytically by making use of the fact that $p\left(\bar{\mathbf{f}} | \mathbf{X}, \theta\right) = N\left(\bar{\mathbf{f}} | \mathbf{0}, \mathbf{K}\right)$ and

$$\log \int \prod_{n=1}^{N} N\left(\langle f_n \rangle | \bar{f}_n, \beta^{-1}\right) N\left(\bar{\mathbf{f}} | \mathbf{0}, \mathbf{K}\right) d\bar{\mathbf{f}} = \log N\left(\langle \mathbf{f} \rangle | \mathbf{0}, \left(\mathbf{K} + \beta^{-1}\mathbf{I}\right)\right).$$

[1] The Kullback-Leibler divergence between two distributions is defined as $\mathrm{KL}\left(q(x) || p(x)\right) = \int q(x) \log \frac{q(x)}{p(x)} dx$.

We are interested in the dependence of this term on a particular $q\left(f_n\right)$. This can be obtained by factorising the distibution,

$$p\left(\langle\mathbf{f}\rangle\right) = p\left(\langle f_n\rangle \mid \langle\mathbf{f}_{\backslash n}\rangle\right) p\left(\langle\mathbf{f}_{\backslash n}\rangle\right),$$

where only the first term of this factorisation is dependent on $q\left(f_n\right)$. This conditional distribution has the form of a Gaussian,

$$p\left(\langle f_n\rangle \mid \langle\mathbf{f}_{\backslash n}\rangle\right) = N\left(\langle f_n\rangle \mid \mu_n, \sigma_n^2\right),$$

with mean $\mu_n = \mathbf{k}_{\backslash n}^{\mathrm{T}}\left(\mathbf{K}_{\backslash n} + \beta^{-1}\mathbf{I}\right)^{-1}\langle\mathbf{f}_{\backslash n}\rangle$ and variance

$$\sigma_n^2 = \beta^{-1} + k_{nn} - \mathbf{k}_{\backslash n}^{\mathrm{T}}\left(\mathbf{K}_{\backslash n} + \beta^{-1}\mathbf{I}\right)^{-1}\mathbf{k}_{\backslash n},$$

where $\mathbf{k}_{\backslash n}$ is the nth column of the covariance matrix with the nth element removed, $\mathbf{K}_{\backslash n}$ is the covariance matrix with the nth row and column removed and $\mathbf{f}_{\backslash n}$ is the vector $\mathbf{f}$ with the nth element removed. The terms of (9) which are dependent on $q\left(f_n\right)$ are then given by

$$\begin{aligned}\mathcal{L}_n'\left(\boldsymbol{\theta}\right) = &-\left\langle \log N\left(f_n|\mu_n, \sigma_n^2\right)\right\rangle + \left\langle \log p\left(y_n|f_n\right)\right\rangle + \left\langle \log q\left(f_n\right)\right\rangle_{q(f_n)} \\ &+\delta_n - \frac{1}{2}\log 2\pi\sigma_n^2\end{aligned}$$

where

$$\delta_n = \frac{1}{2}\left(\beta - \frac{1}{\sigma_n^2}\right)\left(\left\langle f_n^2\right\rangle - \left\langle f_n\right\rangle^2\right). \tag{11}$$

Now if we assume that δ_n is relatively insensitive to changes in $q\left(f_n\right)$ then we can optimise the KL-corrected bound with respect to $q\left(f_n\right)$ to obtain

$$q\left(f_n\right) \propto p\left(y_n|f_n\right) N\left(f_n|\mu_n, \sigma_n^2\right) \tag{12}$$

where we recall that from our definitions $N\left(f_n|\mu_n, \sigma_n^2\right)$ is the prediction at the nth point having removed the nth point from our data set. In the statistical physics literature this is known as a *cavity* distribution. Such cavity distributions are reminiscent of TAP approximations and the expectation propagation algorithm [8,7]. Of course, there is in general no guarantee that δ_n will be insensitive to changes in $q\left(f_n\right)$, however it is still possible to make use of this update in place of the variational updates (of $q\left(\overline{\mathbf{f}}\right)$ and $q\left(f_n\right)$) but it may be prudent to check that the KL-corrected bound is higher than that generated by the standard variational updates after updating $q\left(f_n\right)$. In the experiments in Section 4 we chose to always make use of the standard update so that we knew that any resulting increase in convergence speed was entirely due to optimisation of (9) with respect to the parameters $\boldsymbol{\theta}$.

3.3 Monitoring Convergence

Convergence of the algorithm can be monitored through evaluation of (9). However, this bound contains an expectation of $\log p\left(y_n|f_n\right)$ under the noise model which will typically require quadrature to compute. However, if we only compute the bound after updating $q(f_n)$ then we find (9) may be replaced by

$$\mathcal{L}'_c(\boldsymbol{\theta}) = \log N\left(\langle \mathbf{f}\rangle | \mathbf{0}, \mathbf{K}+\beta^{-1}\mathbf{I}\right) - \sum_{n=1}^{N} \log N\left(\langle f_n\rangle | \langle \bar{f}_n\rangle, \beta^{-1}\right) + \sum_{n=1}^{N} \log Z_n$$

which will only require quadrature if Z_n requires quadrature (see Section 2.1).

4 Results

We performed a series of classification experiments with benchmark data sets to evaluate the performance of the PPA algorithm. For comparison we also include published results from the support vector machine on these data sets. In all our experiments we ordered updates as specified in Algorithm 1. Code for recreating our results is available on-line, for details see Appendix A.

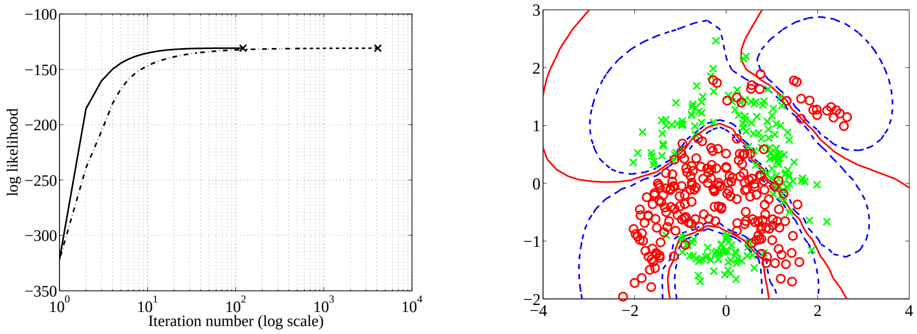
Algorithm 1 Optimisation of the Gaussian Process with PPA. Note that algorithmically it is still necessary to update $q(\bar{\mathbf{f}})$ for both variational and KL-corrected approaches as it is a pre-requisite for computation of each $q(f_n)$.

```
Inputs $\mathbf{X} = [\mathbf{x}_1, \ldots \mathbf{x}_N]^{\mathrm{T}}$ and outputs $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^{\mathrm{T}}$, a convergence tolerance, initial values for $\beta$ and $\boldsymbol{\theta}$.

E-Step — Iterate over the $q$-distributions
 Update $\langle \bar{\mathbf{f}} \rangle$ and $\langle \bar{\mathbf{f}}\bar{\mathbf{f}}^{\mathrm{T}} \rangle$.
 Calculate $\mathbf{g}$ and $\Gamma$ based on the given noise model.
 $n = 1 : N$
 Update $\langle f_n \rangle$ and $\langle f_n^2 \rangle$
 *
M-Step — Update the parameters
 Use gradient based optimisation for updating $\boldsymbol{\theta}$.
 For standard variational approach optimise (3), for KL-corrected approach optimise (9).
 Update $\beta$
 bound on likelihood changes by less than the convergence tolerance.
```

4.1 Convergence Speed

We first considered a synthetic data set **banana** [9]. This data set consists of two dimensional inputs sampled from Gaussian distributions. One hundred training/test partitions of the data are provided. We used the first partition to illustrate the improvements in training speed gained by using the KL-corrected objective function instead of the standard variational lower bound. The results are shown in Figure 2 (a). They show almost two orders of magnitude improvement in convergence in terms of iterations. These figures carry over into improvement in terms of timing as well. The final learnt decision boundary is shown in Figure 2 (b).



(a)

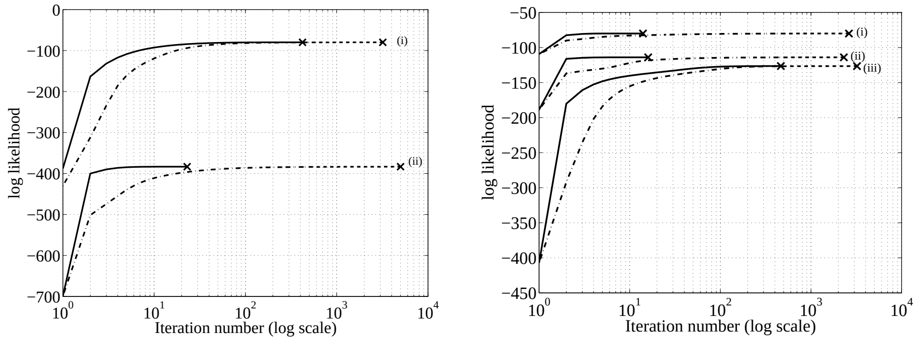
(b)

Figure 2. (a) Plot of log-likelihood vs iteration number (log-scale) for the KL-corrected objective function (solid line) and the standard variational bound (dashed line). KL-corrected requires 120 iterations for convergence while the standard variational approach requires 4102 iterations. The point of convergence for each line is marked on the plot with a cross. Note that both approaches converge to the same likelihood. (b) The resulting classification of the **banana** data set. Decision boundaries are given by solid lines, the dashed lines indicate contours at 0.25 and 0.75 probabilities.

As well as the synthetic set, **banana**, we tested the algorithm using seven other data sets from the **UCI**, **DELVE** and **STATLOG** benchmark repositories with partitions provided by [9]. To allow the classification error comparisons to be accurate, we mimicked the experimental setup found in [9] as far as possible. Each data set is presented as a binary classification problem and partitioned into 100 different training and test data sets. In [9] kernel parameters were chosen through running 5-fold cross validation on the first five realisations of each data set. The median of the parameters was then chosen. In PPA the marginal likelihood can be maximised to obtain the kernel parameters. Therefore, for these methods, no cross validation was used. We simply maximised the lower bound on the marginal likelihood for the first five data sets. The kernel parameters associated with the median RBF kernel width were then used for all data sets to compute the final results.

In Figure 3 we provide convergence plots for several of the data sets. Convergence plots and CPU timings were generated using the first partition of each data set. The final classification error is provided in Table 1 (a). Also included in this table for interest are the classification results reported by [9] for the SVM. The total time for convergence is given in Table 1 (b).

The experimental results show that a Gaussian Process with variational inference through PPA has broadly similar performance to the support vector machine (as we might expect).



(a) (b)

Figure 3. (a) Convergence plot for (i) **twonorm** data set and (ii) **German** data set. (b) convergence plot for (i) **titanic** data set, (ii) **breast-cancer** data set and (iii) **waveform** data set. Each plot shows the bound on the likelihood vs iteration number for the KL-corrected objective function (solid line) and the standard variational bound (dashed line).

5 Discussion

We have presented a correction to the standard variational bound in the context of Gaussian process models. The KL-corrected bound leads to an much improved speed up for variational learning, without losing the guarantee of convergence. In experiments on benchmark data, the bound lead to a speed increase for all our experiments. The lowest speed up was 5.89 times faster whilst the highest was 103 times faster.

There is potential for KL-correction to be applied in other models and not just when Gaussian likelihoods and priors are used. We discussed how updates with respect to the marginal variational approximations, $q\left(f_n\right)$, could also be done to optimise the KL-corrected bound, but we leave exploration of these updates and a study of the general conditions for which KL-correction can be applied to further work.

A On-line Source Code

The source code for re-running all the experiments detailed here is available online from `http://www.dcs.shef.ac.uk/~neil/ppa/`.

Table 1. (a) shows the classification error results of experiments with benchmark data sets compared to published results. The SVM results are taken from [9]. (b) displays CPU time comparisons for the experiments with benchmark data sets. Timings are given for the standard variational approach (STD) and the KL-corrected approach (KLC). The increase in speed is summarised by the speed up factor. Average speed up was 25.6.

DATASET	SVM	GP-PPA
BANANA	11.5 ± 0.7	10.9 ± 0.5
B. CANCER	26.0 ± 4.7	29.4 ± 5.0
DIABETES	23.5 ± 1.7	23.0 ± 2.0
GERMAN	23.6 ± 2.1	23.9 ± 2.0
HEART	16.0 ± 3.3	17 ± 3.0
TITANIC	22.4 ± 1.0	23.2 ± 0.3
TWONORM	3.0 ± 0.2	2.8 ± 0.3
WAVEFORM	9.9 ± 0.4	11.9 ± 0.4

(a)

	TIME		SPEED
DATASET	STD	KLC	UP
	$/10^3$s	$/10^3$s	FACTOR
BANANA	22.5	1.13	19.9
B. CANCER	4.10	0.187	21.9
DIABETES	34.2	3.92	8.75
GERMAN	111	1.08	103
HEART	2.77	0.153	18.1
TITANIC	1.94	0.0919	21.1
TWONORM	30.0	4.67	6.42
WAVENORM	36.3	6.16	5.89

(b)

References

1. Girolami, M., Rogers, S.: Variational bayesian multinomial probit regression with gaussian process priors. Neural Computation **18**(8) (2006) 1790–1817
2. King, N.J., Lawrence, N.D.: Variational inference in Gaussian processes via probabilistic point assimilation. Technical Report CS-05-06, The University of Sheffield, Department of Computer Science (2005)
3. O'Hagan, A.: Some Bayesian numerical analysis. In Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M., eds.: Bayesian Statistics 4, Valencia, Oxford University Press (1992) 345–363
4. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA (2006)
5. Waterhouse, S., MacKay, D.J.C., Robinson, T.: Bayesian methods for mixtures of experts. In Touretzky, D., Mozer, M., Hasselmo, M., eds.: Advances in Neural Information Processing Systems. Volume 8., Cambridge, MA, MIT Press (1996) 351–357
6. Seeger, M.: Bayesian model selection for support vector machines, Gaussian processes and other kernel classifiers. In Solla, S.A., Leen, T.K., Müller, K.R., eds.: Advances in Neural Information Processing Systems. Volume 12., Cambridge, MA, MIT Press (2000) 603–609
7. Minka, T.P.: A family of algorithms for approximate Bayesian inference. PhD thesis, Massachusetts Institute of Technology (2001)
8. Opper, M., Winther, O.: Gaussian processes for classification: Mean field algorithms. Neural Computation **12** (2000) 2655–2684
9. Rätsch, G., Onoda, T., Müller, K.R.: Soft margins for adaboost. Machine Learning **42**(3) (2001) 287–320