Tilted Variational Bayes

James Hensman
Department of Computer Science[1]
University of Sheffield, UK

Max Zwießele
Department of Computer Science[1]
University of Sheffield, UK

Neil D. Lawrence
Department of Computer Science[1]
University of Sheffield, UK

Abstract

We present a novel method for approximate inference. Using some of the constructs from expectation propagation (EP), we derive a lower bound of the marginal likelihood in a similar fashion to variational Bayes (VB). The method combines some of the benefits of VB and EP: it can be used with light-tailed likelihoods (where traditional VB fails), and it provides a lower bound on the marginal likelihood. We apply the method to Gaussian process classification, a situation where the Kullback-Leibler divergence minimized in traditional VB can be infinite, and to robust Gaussian process regression, where the inference process is dramatically simplified in comparison to EP.

Code to reproduce all the experiments can be found at `github.com/SheffieldML/TVB`.

1 Introduction

The calculus of uncertainty requires Bayesian inference for obtaining the posterior distribution over a set of latent variables, $\mathbf{x}$, given a set of data observations $\mathbf{y}$. One common formalism is a prior over the latent variables, $p(\mathbf{x})$ and a set of independent likelihoods, $p(y_i|x_i)$, giving a joint distribution of the form $\prod_{i=1}^{n} p(y_i|x_i)p(\mathbf{x})$. Inference over the latent variables requires computation of the posterior distribution, $p(\mathbf{x}|\mathbf{y})$. However, for many combinations of likelihood and prior computing this posterior, and the associated marginal likelihood for $p(\mathbf{y})$ is intractable, and we must proceed with approximate methods such as variational Bayes [Wainwright and Jordan, 2008], expectation propagation [Minka, 2001] or Monte Carlo sampling schemes [Gilks et al., 1996].

The approach of expectation propagation (EP) has become particularly popular for models of this type, perhaps due to the empirical performance of these approaches in domains such as Gaussian process classification Kuss and Rasmussen [2005], where EP has become the de facto method of choice. EP proceeds through replacing the non-conjugate likelihoods terms with conjugate ones using a straightforward moment-matching procedure.

However, general implementation of EP is not without its problems. The original EP algorithm offers no guarantee of convergence, and more recent contributions surrounding the double-loop algorithm [Opper and Winther, 2004, Seeger and Nickisch, 2010] may require bespoke implementations that are specific to one problem. See for example Jylänki et al. [2011], where considerable effort is required for implementation of EP for Gaussian process regression with a Student-t likelihood.

Even when EP does provably converge, it is difficult to describe to what it is converging. Whilst expectation consistency [Opper and Winther, 2004] gives some guarantee of EP's performance, the minimisation of a KL divergence in VB is an attractive property.

In the case where the model has hyper-parameters which are to be optimized alongside approximate inference of latent variables, EP provides an approximation to the marginal likelihood which can be used as an objective function for the hyper-parameters. To evaluate the marginal likelihood and its gradient, EP must be run to convergence. In contrast, VB provides a lower bound on the marginal likelihood, which can be used to adjust the hyper-parameters whilst performing approximate inference. The guarantee of a lower bound on the marginal likelihood is attractive when integrating the model with a wider inference framework, or performing novel optimization of the approximation, e.g. Stochastic or conjugate methods [Hoffman et al.,

[1] Also at Sheffield Institute for Translational Neuroscience, SITraN.

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

2013, Honkela et al., 2010, Hensman et al., 2012].

Variational Bayesian approaches can be broadly characterized as either *free-form* variational minimization methods [Waterhouse et al., 1996], where the functional form of the approximating posterior emerges naturally after specific factorization assumptions are made, or *explicit* variational approximations, where the functional form is imposed. Our approach is a particular type of explicit approximation where we explicitly include the data observation (through its likelihood) in the approximation.

The main novelty in our approach is a particular form for the variational approximation that explicitly includes the likelihood. Similar in spirit to Opper and Winther [2004], we use *two* distributions which communicate. Rather than providing an approximation procedure akin to the removal/inclusion of a datum cf. EP, we simply provide the bound on the marginal likelihood and its gradient with respect to some variational parameters, alongside gradients with respect to hyper-parameters, and optimise the problem using our preferred optimisation routine. The implementation of the procedure is thus extremely simple, and unlike EP we do not need to converge to a temporary solution before adjusting the hyper-parameters.

2 Tilted Variational Bayes

We will consider the following form of probabilistic model

$$p(\mathbf{y}, \mathbf{x}) = p(\mathbf{x}) \prod_{i=1}^{n} p(y_i \,|\, x_i) \tag{1}$$

Where $\mathbf{y} = \{y_i\}_{i=1}^n$ is a vector of observed data points and $\mathbf{x} = \{x_i\}_{i=1}^n$ is a vector of latent variables. In our examples of Gaussian process models, $p(\mathbf{x})$ is a multivariate normal distribution over the values of a function at some observed points, and $p(y_i \,|\, x_i)$ is some non-Gaussian likelihood, but our work extends to a wide range of other models also, and so we maintain a general notation. This class of models is the same as considered by Opper and Winther [2004], and is easily generalized to those considered by Seeger and Nickisch [2010].

In the case where the likelihoods $p(y_i \,|\, x_i)$ and the prior $p(\mathbf{x})$ form a conjugate pair, Bayesian inference for the posterior $p(\mathbf{x} \,|\, \mathbf{y})$ is tractable. When inference is not tractable, approximate inference schemes can be used.

2.1 Pseudo Data

In VB, the posterior distribution is approximated by selecting a member of some family of distributions which minimises the Kullback Leibler divergence from the approximation $q(\mathbf{x})$ to the true posterior $p(\mathbf{x} \,|\, \mathbf{y})$. In EP, the *factors* corresponding to the likelihood are replaced by approximate factors which ensure conjugacy to the prior, and thus tractability. For example if the prior is Gaussian, the factors are

$$t_i(x_i) = Z_i \mathcal{N}(x_i | \mu_i, \sigma_i^2) \tag{2}$$

and the parameters Z_i, μ_i, and σ_i^2 are iteratively updated until convergence. The approximate posterior is then

$$p(\mathbf{x} \,|\, \mathbf{y}) \approx \frac{1}{Z_{EP}} p(\mathbf{x}) \prod_{i=1}^{n} t_i(x_i) \ , \tag{3}$$

where $Z_{\text{EP}} = \int p(\mathbf{x}) \prod_{i=1}^n t_i(x_i) \, \mathrm{d}\mathbf{x}$ is the required normaliser.

Here we prefer to interpret these factors as *pseudo-data*. We denote these data $\widetilde{\mathbf{y}} = \{\widetilde{y}_i\}_{i=1}^n$, define some conjugate likelihood for them $p(\widetilde{y}_i \,|\, x_i)$, and use them in a similar fashion to the EP factors. The pseudo-data follow all the usual probabilistic rules, including normalisation, which we find helpful in constructing our algorithm. The approximation to the posterior using these pseudo-data is then

$$p(\mathbf{x} \,|\, \widetilde{\mathbf{y}}) = \frac{p(\mathbf{x}) \prod_{i=1}^n p(\widetilde{y}_i \,|\, x_i)}{p(\widetilde{\mathbf{y}})} \ . \tag{4}$$

We no longer have the normalising factors Z_i, which it turns out do not affect the posterior, but the marginal likelihood: similar terms appear in our bound on the marginal likelihood, and we find the ability to deal with the pseudo-data in probabilistic language helpful.

In EP, two important one-dimensional distributions are the *cavity* distribution and the *tilted* distribution. In terms of the pseudo-data, the cavity distribution can be written as

$$c(x_i) = p(x_i \,|\, \widetilde{\mathbf{y}}_{\setminus i}) \ , \tag{5}$$

where $\setminus i$ stands for 'all indices except i'. The tilted distribution is given by multiplying this cavity by the likelihood and re-normalizing:

$$q(x_i) = \frac{p(y_i \,|\, x_i) p(x_i \,|\, \widetilde{\mathbf{y}}_{\setminus i})}{\int p(y_i \,|\, x_i) p(x_i \,|\, \widetilde{\mathbf{y}}_{\setminus i}) \, \mathrm{d}x_i} = p(x_i \,|\, y_i, \widetilde{\mathbf{y}}_{\setminus i}) \ . \tag{6}$$

EP then proceeds by a moment-matching scheme which updates the parameters of the pseudo-likelihood (or equivalent EP factor) $p(\widetilde{y}_i \,|\, x_i)$ so as to minimise the *marginal* KL divergences:

$$\mathrm{KL}\left[q(x_i) || p(x_i \,|\, \widetilde{\mathbf{y}})\right] \ , \tag{7}$$

where the expectation is taken under the *tilted* distribution, as in our proposal.

2.2 Bounding the Marginal Likelihood

We have given the tilted distribution the moniker q since it is this distribution which we will use for averaging in our tilted VB scheme, defining the factorising distribution $q(\mathbf{x}) = \prod_{i=1}^{n} q(x_i)$. We emphasise that q includes the data. We proceed as for normal variational inference, but with the tilted distribution q where we would normally find a simpler approximation to the posterior. First observe that Bayes' rule can be re-arranged to give

$$p(\mathbf{y}) = \frac{p(\mathbf{y}\,|\,\mathbf{x})p(\mathbf{x})}{p(\mathbf{x}\,|\,\mathbf{y})}\frac{q(\mathbf{x})}{q(\mathbf{x})} \ , \tag{8}$$

take logarithms and then the expectation under q so that

$$\log p(\mathbf{y}) = \mathbb{E}_{q(\mathbf{x})}\left[\log \frac{p(\mathbf{y}\,|\,\mathbf{x})p(\mathbf{x})}{q(\mathbf{x})}\right] + \mathrm{KL}\left[q(\mathbf{x})||p(\mathbf{x}\,|\,\mathbf{y})\right] \ . \tag{9}$$

In exactly the same fashion as VB, we can now drop the KL term and obtain a lower bound on the marginal likelihood which serves as an objective both in improving the approximation (by adapting $\widetilde{\mathbf{y}}$ and any parameters of $p(\widetilde{\mathbf{y}}\,|\,\mathbf{x})$) and also for optimising hyper-parameters.

The lower bound in equation (9) initially appears difficult since it contains the expectation of the log of the likelihood function and the entropy of the tilted distribution, which may not be tractable. However, the difficult terms cancel as follows. Defining the lower bound on the marginal likelihood as $\mathcal{L}$ and expanding the definition of the tilted distribution:

$$\begin{aligned}
\mathcal{L} &\triangleq \mathbb{E}_{q(\mathbf{x})}\left[\log \frac{p(\mathbf{y}\,|\,\mathbf{x})p(\mathbf{x})}{q(\mathbf{x})}\right] \\
&= \mathbb{E}_{q(\mathbf{x})}\left[\log \frac{\prod_{i=1}^{n} p(y_i\,|\,x_i)p(\mathbf{x})}{\prod_i p(y_i\,|\,x_i)p(x_i\,|\,\widetilde{\mathbf{y}}_{\setminus i})/p(y_i\,|\,\widetilde{\mathbf{y}}_{\setminus i})}\right] \\
&= \sum_{i=1}^{n} \log p(y_i\,|\,\widetilde{\mathbf{y}}_{\setminus i}) + \mathbb{E}_{q(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{\prod_i p(x_i\,|\,\widetilde{\mathbf{y}}_{\setminus i})}\right] \ .
\end{aligned} \tag{10}$$

The requirements for computing the bound are then identical to those for EP. We require the zeroth moment of the tilted distribution ($p(y_i\,|\,\widetilde{\mathbf{y}}_{\setminus i})$), and the moments of the tilted distribution as appear in the prior $p(\mathbf{x})$ and the cavity distributions $p(x_i\,|\,\widetilde{\mathbf{y}}_{\setminus i})$.

Whilst the updating procedure of EP is superficially attractive, inference in our scheme is much simpler: we simply take derivatives of the bound with respect to the pseudo-data and any additional parameters of $p(\widetilde{y}_i\,|\,x_i)$. Approximate inference can then be performed using any off-the-shelf optimisation routine. This allows the entire approximation to be updated in parallel. Unlike EP, we do not need to wait for the algorithm to converge to a temporary solution before adjusting the hyper-parameters.

2.3 The Tilting Effect

Our method proposes to minimise the KL divergence from the tilted distribution $q(\mathbf{x})$ to the posterior. In standard VB, we proceed similarly, but usually with a fixed form for the approximating distribution q. To see the effect of tilting, consider the scalar case with a Gaussian prior $p(x) = \mathcal{N}(\mu, \sigma^2)$, and a light-tailed likelihood (e.g. Gaussian process classification). The variational KL divergence takes the expectation under the approximating distribution. The standard variational approach assumes a Gaussian form for q. Because this has heavier tails than the true posterior the variational KL divergence can increase very quickly. This is because the expectation is taken under the heavy tail in the area where the logarithm evaluated in the region of a light tail. The extreme of a light tail is the case where the likelihood has no support and the KL divergence goes to zero (e.g. when an inverse Heaviside link function is used).

2.4 Interpretation of the Bound

The lower bound (10) contains two terms, which we shall analyse in the following. The first term $\sum_{i=1}^{n} \log p(y_i\,|\,\widetilde{\mathbf{y}}_{\setminus i})$ appears as a data likelihood under an alternative model where the prior has been replaced with the factorising cavity distribution. The second term acts as a penalty for the cavity distribution being different from the prior. If the prior is itself a factorising distribution, then the cavity will become equal to the prior, and the bound on the marginal likelihood will be exact.

We find that the required derivatives require higher order moments of the tilted distribution than are required to simply compute the bound. We do not include a full derivation of the straightforward derivatives of the bound here, but we do note in the following some of the derivatives that lead to interesting insights into the bound.

Consider the derivative of the zeroth-moment of the i^{th} tilted distribution (as appear in the first term of the lower bound) $p(y_i\,|\,\widetilde{\mathbf{y}}_{\setminus i})$ with respect to another pseudo-data point $\widetilde{y}_j$. The pseudo-likelihoods are designed to be conjugate to the prior, so the i^{th} cavity distribution $p(x_i\,|\,\widetilde{\mathbf{y}}_{\setminus i})$ is tractable. Assume the cavity distribution is in the exponential family with sufficient statistic vector $\mathbf{g}(x_i)$ (see e.g. Wainwright and Jordan

[2008] for details):

$$p(x_i \,|\, \widetilde{\mathbf{y}}_{\setminus i}) = h(x_i) \exp\{\boldsymbol{\theta}_i^\top \mathbf{g}(x_i) - \Psi(\boldsymbol{\theta}_i)\} \ . \tag{11}$$

The parameters of the cavity $\boldsymbol{\theta}_i$ are a simple function of the *other* pseudo-likelihoods. Expanding the derivative using the chain rule gives

$$\begin{aligned}\frac{\mathrm{d}p(y_i \,|\, \widetilde{\mathbf{y}}_{\setminus i})}{\mathrm{d}\widetilde{\mathbf{y}}_j} &= \frac{\mathrm{d}\boldsymbol{\theta}_i}{\mathrm{d}\widetilde{y}_j}\frac{\mathrm{d}p(y_i \,|\, \widetilde{\mathbf{y}}_{\setminus i})}{\mathrm{d}\boldsymbol{\theta}_i} \\ &\propto \int p(y_i \,|\, x_i)\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}_i} p(x_i \,|\, \widetilde{\mathbf{y}}_{\setminus i})\,\mathrm{d}x_i \\ &\propto \int q(x_i)\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}_i}\log p(x_i \,|\, \widetilde{\mathbf{y}}_{\setminus i})\,\mathrm{d}x_i \\ &\propto \mathbb{E}_{q(x_i)}\big[\mathbf{g}(x_i)\big] - \mathbb{E}_{p(x_i \,|\, \widetilde{\mathbf{y}}_{\setminus i})}\big[\mathbf{g}(x_i)\big] \ .\end{aligned} \tag{12}$$

Here we have used the property that the derivative of $\Psi(\boldsymbol{\theta})$ is the expected value under the exponential family distribution. This result shows that the first term in the lower bound (10) is maximised when the moments of the cavity distribution match the moments of the tilted distribution.

Now, consider that the second term in our bound can be written as a series of KL divergences:

$$\begin{aligned}\mathbb{E}_{q(x)}\left[\log \frac{p(\mathbf{x})}{\prod_i p(x_i \,|\, \widetilde{\mathbf{y}}_{\setminus i})}\right] &= \sum_{i=1}^{n} \mathrm{KL}[q(x_i)||p(x_i \,|\, \widetilde{\mathbf{y}}_{\setminus i})] \\ &\quad - \mathrm{KL}[q(\mathbf{x})||p(\mathbf{x})]\end{aligned} \tag{13}$$

When the moments match for the first term, they also minimise the KL divergences in the second term, aside from the final KL divergence which is the usual variational term which ensures that the approximation is close to the prior.

3 The Second Approximate Distribution

The distribution $p(\mathbf{x} \,|\, \widetilde{\mathbf{y}})$ is of the same form as that used in EP. Relating our pseudo-data likelihoods $p(\widetilde{\mathbf{y}} \,|\, \mathbf{x})$ to the EP factors once more, we see that the two methods have the same posterior, though in TVB we present a different method for selecting $p(\widetilde{\mathbf{y}} \,|\, \mathbf{x})$.

To see why using $p(\mathbf{x} \,|\, \widetilde{\mathbf{y}})$ as an approximate posterior is a valid choice, consider prediction for a new datum $y_\star$ related to the rest of the model through latent variable $x_\star$ and $p(x_\star \,|\, \mathbf{x})$. If we have the true posterior $p(\mathbf{x} \,|\, \mathbf{y})$, then we simply do

$$p(y_\star|\mathbf{y}) = \int p(y_\star \,|\, x_\star) \int p(x_\star \,|\, \mathbf{x}) p(\mathbf{x} \,|\, \mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}x_\star. \tag{14}$$

The inner integral is of course intractable, and so we replace the posterior with the approximation $p(\mathbf{x} \,|\, \widetilde{\mathbf{y}})$. The inner integral results in $p(x_\star \,|\, \widetilde{\mathbf{y}})$, which is the cavity distribution for this additional datum, and the whole integral results in $p(y_\star|\widetilde{\mathbf{y}})$. In other words, replacing the posterior with $p(\mathbf{x}|\widetilde{\mathbf{y}})$ means treating additional data in the same fashion as the training data.

3.1 Relationship to Expectation Consistency

Our method bears some resemblance to the expectation consistency (EC) method of [Opper and Winther, 2005]. EC is also based on several global approximations, with one taking the form of the likelihood (as our $q(\mathbf{x})$) and one taking a conjugate form (similar to our $p(\mathbf{x} \,|\, \widetilde{\mathbf{y}})$). Indeed, they mention a variational bound which bears a striking resemblance to ours, though they discard this method before deriving EC. EC proceeds by matching the moments of the distributions so as to find a saddle point of a free-energy function. By contrast, our distributions are intrinsically linked by definition.

4 Approximating the Marginal Likelihood

Our method is based on a lower bound on the marginal likelihood, which we use for optimising the approximate factors as well as any hyper-parameters. Expectation propagation provides an approximation to the marginal likelihood which Kuss and Rasmussen [2005] and Jylänki et al. [2011] found empirically to be very accurate for Gaussian Process classification and robust regression. The approximation is

$$\mathcal{L}_{EP} = \log \int p(\mathbf{x}) \prod_{i=1}^{n} t(x_i)\,\mathrm{d}\mathbf{x}. \tag{15}$$

Given our interpretation of the approximate factors $t(x_i)$ as un-normalised pseudo-data, with $t(x_i) = Z_i p(\widetilde{y}_i \,|\, x_i)$, and noting that at convergence of EP $Z_i = p(\widetilde{y}_i \,|\, \widetilde{\mathbf{y}}_{\setminus i})$, the EP approximation to the log marginal likelihood is

$$\mathcal{L}_{EP} = \sum_{i=1}^{n} \log p(y_i \,|\, \widetilde{\mathbf{y}}_{\setminus i}) + \log p(\widetilde{\mathbf{y}}) \tag{16}$$

using appendix A to expand the variational bound, we have

$$\begin{aligned}\mathcal{L}_{EP} = \log p(\mathbf{y}) - \mathrm{KL}[q(\mathbf{x})||p(\mathbf{x} \,|\, \mathbf{y})] + \mathrm{KL}[q(\mathbf{x})||p(\mathbf{x} \,|\, \widetilde{\mathbf{y}})] \\ - \sum_{i=1}^{n} \mathrm{KL}[q(x_i)||p(x_i \,|\, \widetilde{\mathbf{y}})] + p(\widetilde{y}_i \,|\, \widetilde{\mathbf{y}}_{\setminus i}) \ .\end{aligned} \tag{17}$$

Compare this to the lower bound produced by eVB:

$$\mathcal{L}_{\text{TVB}} = \log p(\mathbf{y}) - \text{KL}[q(\mathbf{x})||p(\mathbf{x}\,|\,\mathbf{y})]. \tag{18}$$

The moment matching procedure of EP ensures that at a stationary point, the marginal KL divergences $\text{KL}[q(x_i)||p(x_i\,|\,\widetilde{\mathbf{y}})]$ are minimal, and so we expect the EP approximation to the marginal likelihood to be larger than our bound by $\text{KL}[q(\mathbf{x})||p(\mathbf{x}\,|\,\widetilde{\mathbf{y}})]$.

EP then gives a good approximation to the marginal likelihood when the two KL divergences (from the tilted to the posterior and from the tilted to the approximate posterior) are similar.

5 Experiments

We envisage that our method can provide a viable alternative to EP. We therefore concentrate on applications where EP has been used extensively: Gaussian process models with non-conjugate likelihoods. For clarity we change notation: the latent variables (previously $\mathbf{x}$) now represent the values of a function $f(\mathbf{x})$ taken at some known *input* points $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$; $\mathbf{f} = \{f(\mathbf{x}_i)\}_{i=1}^n = \{f_i\}_{i=1}^n$. The prior is multivariate Gaussian over latent function values $p(\mathbf{f}) = \mathcal{N}(0, \mathbf{K})$ where the covariance matrix $\mathbf{K}$ has entries given by the covariance function:

$$\mathbf{K}_{[i,j]} = k(\mathbf{x}_i, \mathbf{x}_j) \tag{19}$$

In all our experiments, we have used the exponentiated quadratic covariance function with the addition of a diagonal constant:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp\{-\tfrac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{A}^{-1}(\mathbf{x}_i - \mathbf{x}_j)\} + \sigma_n^2 \delta(i,j). \tag{20}$$

The diagonal matrix $\mathbf{A}$ contains *lengthscale* parameters $\mathbf{A}_{i,i} = \ell_i$, which we collect together with the process variance σ^2 and the 'noise' variance σ_n^2 into a vector of hyper-parameters.

The likelihood $p(y_i\,|\,f_i)$ is some application-dependent *non*-Gaussian density. We replace the likelihoods with pseudo-likelihoods which are Gaussian, $p(\widetilde{\mathbf{y}}|f) = \prod_{i=1}^N \mathcal{N}(\widetilde{y}_i\,|\,f_i, \beta_i^{-1})$ (see section 2.1).

The hyper-parameters, pseudo-likelihoods and parameters of the true likelihood (if any) are collected together and optimised jointly using the L-BFGS-B algorithm [Zhu et al., 1997].

5.1 Gaussian Process Classification

One place where EP is ubiquitous is in building Gaussian process classifiers, following the work of Kuss and Rasmussen [2005], who showed empirically that EP works well both in terms of predictive performance and estimation of the marginal likelihood. Here, the data are binary labels $y_i \in \{0, 1\}$ and the likelihood is

$$p(y_i\,|\,f_i) = \phi(f_i)^{y_i}(1 - \phi(f_i))^{(1-y_i)} \quad . \tag{21}$$

The inverse-link function $\phi(f_i)$ transforms the Gaussian process function values into the range (0,1). Here we have used the Heaviside and probit functions:

$$\phi(f_i) = \begin{cases} 1, & \text{if } f_i \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad , \quad \phi(f_i) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{f_i^2}{\sqrt{2}}\right)\right] \quad . \tag{22}$$

As discussed, the light-tailed nature of this likelihood leads to a light-tailed posterior, which is difficult to approximate using standard variational methods. Figure 1 illustrates our approximation scheme for the case where two data are observed $y_i = 0$, with a correlated prior (the inputs are close under the exponentiated quadratic). The tilted distribution does a good job of capturing the form of the posterior by using the likelihood, whilst the Gaussian approximation $p(\mathbf{f}\,|\,\widetilde{\mathbf{y}})$ captures the requisite correlations. The case where the data are not in agreement with the prior is illustrated in Figure S.6 of the supplementary material.

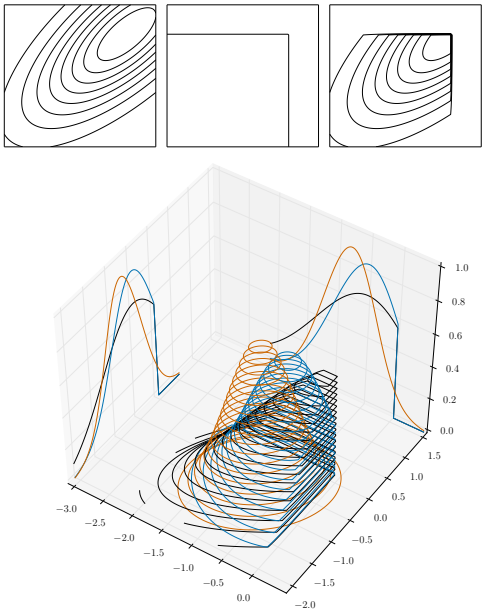


Figure 1: Approximating Gaussian Process classification with only two data points. The top frames show contours of the prior, likelihood and posterior of the Gaussian process: the Heaviside likelihood truncates the normal distribution. In the main frame, the posterior is shown in black contours, the tilted distribution in blue, and the Gaussian $p(\mathbf{f}\,|\,\widetilde{\mathbf{y}})$ is shown in orange. The marginals of all three distributions are shown at the edges of the plot.

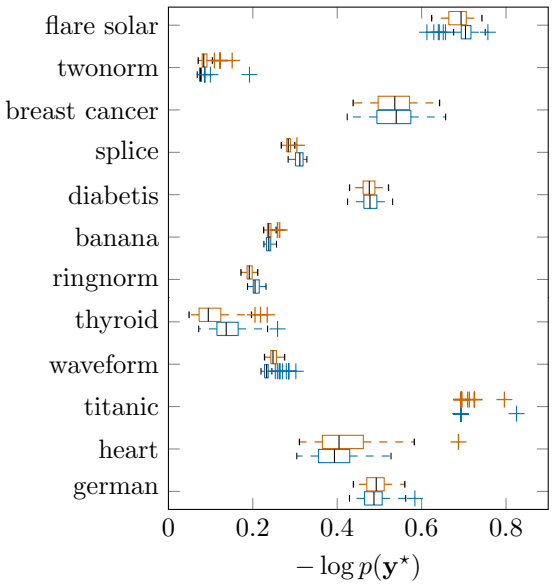


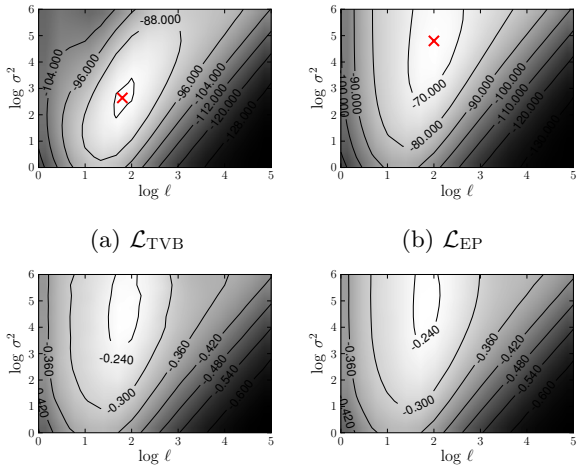
Figure 2: Whisker plots of the hold-out negative log probability (smaller is better) for our method (blue boxes, bottom) and for standard EP (orange, top) for the classification benchmark datasets described by Onoda et al. [2001].

Figure 2 examines the performance of our method in comparison to EP in application to the benchmark datasets used in Onoda et al. [2001]. The values of hold-out probability presented here are similar to those previously reported for EP [Naish-Guzman and Holden, 2007]. We find that in most cases TVB works as well as EP, occasionally slightly better (waveform, heart datasets) and occasionally slightly worse (thyroid, splice). Examining the (negative) log probability of hold-out data in this way examines the methods' abilites to make well calibrated probabilistic predictions. We note that in terms of simple hold-out error (taking a threshold at 0.5) the methods' performances are nearly identical, see supplementary Figure S.1a.

To investigate the minor discrepancies in performance between our method and EP, we follow Kuss and Rasmussen [2005] in examining the methods ability to estimate the marginal likelihood. Figure 3 recreates their experiment for the ionosphere dataset, varying the parameters of the exponentiated quadratic covariance function and plotting contours of the marginal likelihood estimation. For comparison, we also show the information (here log probability) of hold-out data. We see that EP's estimate of the marginal likelihood is more consistent with the hold-out likelihood in form, and the estimate of the hyper-parameters by type-II maximum likelihood leads to a region of high predictive density. The plots closely resemble those reported by Kuss and Rasmussen [2005] for EP. In terms of predictive density, the TVB method has near-identical performance to the EP method, with perhaps a slightly larger region of high density (see the contour at −0.24). However, the method mis-estimates the hyper-parameters, leading to suboptimal predictions for optimized hyper-parameters.

This effect does not appear to be universal. We recreated the experiment for the heart dataset, where our method shows superior hold-out performance to EP, the results are shown in Figure 4. Here, the TVB bound shows better consistency with the hold-out likelihood, and the method leads to a superior set of hyper-parameters, in terms of predictive density, to the EP method. Figures 3 and 4 are reproduced in larger form in the supplementary material, with both the Heaviside and probit likelihoods in each case.

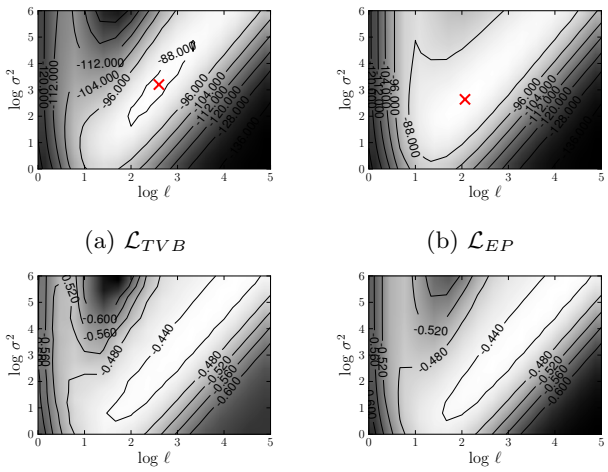
To further examine the effect of hyper-parameter estimation, we re-ran the hold-out experiments of Figure 2 using the hyper-parameters estimated by EP. The results are shown in supplementary Figure S.1b. We find that in most cases the approach is detrimental to performance.



(a) $\mathcal{L}_{\text{TVB}}$ (b) $\mathcal{L}_{\text{EP}}$

(c) $p(y_\star \,|\, \mathbf{y})$ (TVB) (d) $p(y_\star \,|\, \mathbf{y})$ (EP)

Figure 3: Top row: the marginal likelihood bound using TVB (a), and estimation of the marginal likelihood using EP (b) for the ionosphere dataset. Maxima are marked with a red cross. Bottom: average hold-out probability using the TVB (a) and EP (b). The horizontal axis shows log-lengthscale ℓ of the exponentiated quadratic, and the vertical axis shows the logarithmic variance σ^2. Here we have used the probit inverse-link function for compatibility with the results of Kuss and Rasmussen [2005] (see supplementary Figure S.3 for the same figure with the Heaviside inverse-linkfunction)



(a) $\mathcal{L}_{TVB}$ (b) $\mathcal{L}_{EP}$

(c) $p(y_\star \,|\, \mathbf{y})$ (TVB) (d) $p(y_\star \,|\, \mathbf{y})$ (EP)

Figure 4: Top row: the marginal likelihood bound using TVB (a), and estimation of the marginal likelihood using EP (b) for the heart dataset. The maximum is marked as red cross for both likelihood surfaces, respectively. Bottom: average hold-out information content using the TVB (a) and EP (b) in nats. The horizontal axis shows logarithmic lengthscales ℓ of the exponentiated quadratic, and the vertical axis shows the variance σ^2 in log-scale. We have used the probit link function for compatibility with the results of Kuss and Rasmussen [2005] (see supplementary Figure S.3 for the same figure with the Heaviside link function).

5.2 Robust Gaussian Process Regression

Expectation propagation is the de-facto algorithm for GP classification, but for robust GP regression, where the noise on the data is assumed drawn from a heavy-tailed distribution, EP is more problematic. Jylänki et al. [2011] discusses in detail a scheme for robust GP regression using a student-t likelihood, and makes extensive comparisons with variational methods [Tipping and Lawrence, 2005]. Here we apply our methodology to the same task, but our treatment is dramatically shorter. We are guaranteed to find a solution, and though there may be local optima in the objective, we did not experience significant problems.

We begin with an illustrative example of regressing the winning times for the men's olympic marathon against time. In 1904, the race was *"run in brutally hot weather, over dusty roads, with horses and automobiles clearing the way and creating dust clouds"* [Wikipedia, 2013], leading to an outlier on the winning time. Figure 5 shows regression on the data using a Gaussian likelihood: the outlying data point affects the marginal likelihood, resulting in over-estimation of the noise variance and consequently over-estimation of the lengthscale. The method then fails to capture patterns in the data corresponding to war-time stagnation and post-war improvement. With robust regression using a student-t likelihood with 3 degrees of freedom, our approximation method ignores the outlying point and makes better estimates of the marginal likelihood, resulting in a (subjectively) better fit (Figure 5, middle). For comparison, we show the Laplace approximation for the same model (Figure 5, bottom), which also failed to estimate the hyper-parameters correctly.

See Figure 6 for another example from drawing a circle with an outlier prone robot arm. Robust regression is able to remove biases in the circle detection compared to normal GP, comparable with the orgininal variational solution [Tipping and Lawrence, 2005].

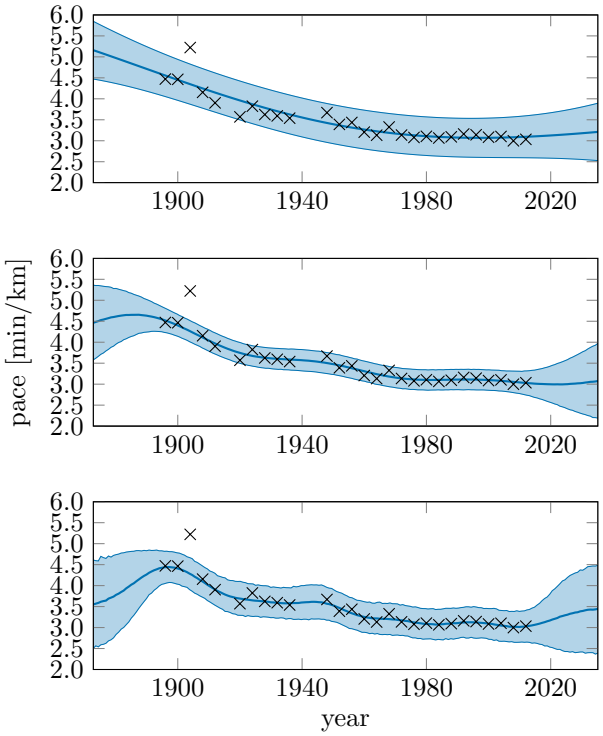
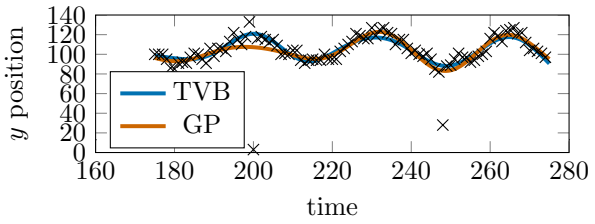
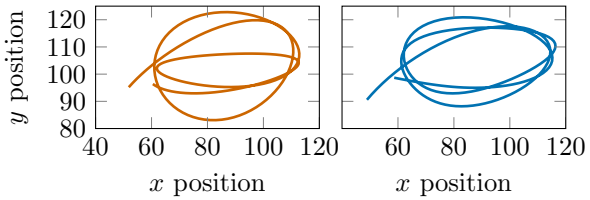


Figure 5: Robust regression for olympic marathon times. Top row shows standard GP regression using a Gaussian likelihood. Middle row is robust GP regression with a student-t likelihood using our method, and the bottom row is Laplace approximation of the same for comparison.

To illustrate the workings of our methodology for robust regression, we again turn to a trivial example where there are only two data points. Figure 7 shows the posterior and approxiate posterior for two datapoints which are correlated a-priori (proximal under the exponentiated quadratic covariance function), but take different values. The heavy-tailed nature of the likelihood (top middle frame) leads to a bi-modal pos-



(a) Robust regression for circle pen movement dataset. Normal GP (orange) is prone to outliers, whereas our method with robust regression finds the true underlying structure of pen movements.



(b) x position against y position of the pen over time as predicted by GP (left). Note the ellyptic distortion of the structure in the data. After using robust regression (on the right) the circular structure of the data is recreated.

Figure 6: Robust regression example on outlier prone xy-pen data from Tipping and Lawrence [2005].

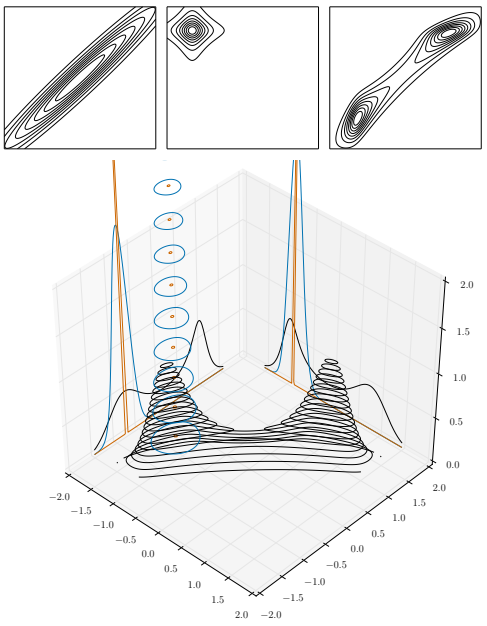


Figure 7: Robust Gaussian process regression with two data points. The two values of the function $\mathbf{f}$ are correlated under the prior (top left frame), but are observed at different values with a heavy-tailed likelihood (top centre frame), leading to a bimodal posterior (top right frame). The main frame shows the posterior, tilted and Gaussian distributions similarly to Figure 1.

terior. The tilted distribution is able to capture only one mode of the posterior. In this case, the problem is symmetric and re-initialising the variational parameters leads to selection of one mode at random.

It is interesting that in this case the Gaussian approximation has a very small variance. A similar effect is present even when the two-data points are consistent with the prior (see supplementary Figure S.7), leading to a uni-modal posterior. Returning to our discussion of this distribution, we posit that these variances are quite reasonable: when making a prediction at a new point we include the heavy-tailed likelihood.

6 Discussion

Expectation propagation is a general framework for inference that has proven particularly useful in the context of Gaussian process classification. However, there are two particular problems with EP. Firstly, it doesn't provide a strict lower bound on the likelihood, making its inclusion within a wider inference/optimization framework problematic. Secondly, the implementation of EP when the likelihood is heavy tailed can be very involved. Standard variational approaches do provide a strict lower bound but have been shown to perform poorly when likelihoods are light-tailed: the variational Kullback Leibler divergence in this case takes an expectation under a density that is heavier tailed than the true posterior. This leads to a very loose bound on the marginal likelihood.

Our contribution has been to introduce tilted variational Bayes, a new variational approach that combines the advantages of EP with the rigorous lower bound associated with variational inference. The innovation is to explicitly introduce the likelihood inside the variational posterior. This dependence ensures that the variational KL is taken under a light-tailed density when appropriate. The resulting algorithm is far simpler in implementation on heavy tailed likelihoods, offering the scope for a more unified approximating framework.

Empirically we found that for GP classification, there is little difference between our approach and EP. There is some difference in terms of the estimation of the marginal likelihood, but this is to be expected when obtaining a rigorous lower bound in place of an approximation.

Acknowledgements

JH was supported by MRC fellowship "Bayesian models of expression in the transcriptome for clinical RNA-Seq", MZ was supported by EU FP7-PEOPLE Project Ref 316861, and NL by EU FP7-KBBE Project Ref 289434 and EU FP7-HEALTH Project Ref 305626.

References

Walter R. Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, 1996.

James Hensman, Magnus Rattray, and Neil Lawrence. Fast variational inference in the conjugate exponential family. In *Advances in Neural Information Processing Systems 25*, pages 2897–2905, 2012.

M Hoffman, D Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

Antti Honkela, Tapani Raiko, Mikael Kuusela, Matti Tornio, and Juha Karhunen. Approximate riemannian conjugate gradient learning for fixed-form variational bayes. *The Journal of Machine Learning Research*, 9999:3235–3268, 2010.

Pasi Jylänki, Jarno Vanhatalo, and Aki Vehtari. Robust Gaussian process regression with a student-t likelihood. *The Journal of Machine Learning Research*, 12:3227–3257, 2011.

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

Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In Jack S. Breese and Daphne Koller, editors, *Uncertainty in Artificial Intelligence*, volume 17, San Francisco, CA, 2001. Morgan Kauffman.

Andrew Naish-Guzman and Sean Holden. The generalized fitc approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1064, 2007.

T Onoda, G Rätsch, and KR Müller. Soft margins for adaboost. *Machine Learning*, 42(3):287–320, 2001.

Manfred Opper and Ole Winther. Expectation consistent free energies for approximate inference. In *Advances in Neural Information Processing Systems*, pages 1001–1008, 2004.

Manfred Opper and Ole Winther. Expectation consistent approximate inference. *The Journal of Machine Learning Research*, 6:2177–2204, 2005.

Matthias W Seeger and Hannes Nickisch. Fast convergent algorithms for expectation propagation approximate bayesian inference. *arXiv preprint arXiv:1012.3584*, 2010.

Michael E Tipping and Neil D Lawrence. Variational inference for student-t models: Robust bayesian interpolation and generalised component analysis. *Neurocomputing*, 69(1):123–141, 2005.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

Steve Waterhouse, David J. C. MacKay, and Tony Robinson. Bayesian methods for mixtures of experts. In David Touretzky, Michael Mozer, and Mark Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 351–357, Cambridge, MA, 1996. MIT Press.

Wikipedia. 1904 Summer Olympics — Wikipedia, the free encyclopedia, 2013. URL `http://en.wikipedia.org/wiki/1904_Summer_Olympics#Marathon`. Online; accessed 1-November-2013.

Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.