Journal of Machine Learning Research 6 (2005) 1679–1704 Submitted 8/05; Published 10/05

Assessing Approximate Inference for Binary Gaussian Process Classification

Malte Kuss KUSS@TUEBINGEN.MPG.DE
Carl Edward Rasmussen CARL@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics
Spemannstraße 38
72076 Tübingen, Germany

Editor: Ralf Herbrich

Abstract

Gaussian process priors can be used to define flexible, probabilistic classification models. Unfortunately exact Bayesian inference is analytically intractable and various approximation techniques have been proposed. In this work we review and compare Laplace's method and Expectation Propagation for approximate Bayesian inference in the binary Gaussian process classification model. We present a comprehensive comparison of the approximations, their predictive performance and marginal likelihood estimates to results obtained by MCMC sampling. We explain theoretically and corroborate empirically the advantages of Expectation Propagation compared to Laplace's method.

Keywords: Gaussian process priors, probabilistic classification, Laplace's approximation, expectation propagation, marginal likelihood, evidence, MCMC

1. Introduction

In recent years models based on Gaussian process (GP) priors have attracted much attention in the machine learning community. Whereas inference in the GP regression model with Gaussian noise can be done analytically, probabilistic classification using GPs is analytically intractable, see Rasmussen and Williams (2006) for a general overview. Several approaches to approximate Bayesian inference have been suggested, including Laplace's method, Expectation Propagation (EP), variational approximations and Markov chain Monte Carlo (MCMC) sampling, some of these in conjunction with generalisation bounds, online learning schemes and sparse approximations (e.g. Neal, 1998; Williams and Barber, 1998; Gibbs and MacKay, 2000; Opper and Winther, 2000; Csató and Opper, 2002; Seeger, 2002; Lawrence et al., 2003).

Despite the abundance of recent work on probabilistic GP classifiers, most experimental studies provide only anecdotal evidence, and no clear picture has yet emerged, as to when and why which algorithm should be preferred. Thus, from a practitioners point of view it is unclear what the method of choice is for probabilistic GP classification. In this work, we set out to understand and compare two of the most wide-spread approximations: Laplace's method and Expectation Propagation (EP). We also compare to a sophisticated, but computationally demanding MCMC scheme, which becomes exact in the limit of long running times. We do not address issues of sparsification but stick to comparing the two types of approximation.

We examine two aspects of the approximation schemes: Firstly the accuracy of approximations to the marginal likelihood which is of central importance for model selection and model comparison.

In any practical application of GPs in classification (usually multiple) parameters of the covariance function (hyper-parameters) have to be handled. Bayesian model selection provides a consistent framework for setting such parameters. Therefore, it is essential to evaluate the accuracy of the marginal likelihood approximations as a function of the hyper-parameters, in order to assess the practical usefulness of the approach. The related question of whether the marginal likelihood correlates well with the generalisation performance cannot be answered in general but depends on the appropriateness of the model for a given data set. However, we do assess this empirically for two data sets.

Secondly, we need to assess the quality of the approximate probabilistic predictions. In the past, the probabilistic nature of the GP predictions has not received much attention, the focus being mostly on classification error *rates*. This unfortunate state of affairs is caused primarily by typical benchmarking problems being considered outside of a realistic context. The ability of a classifier to produce class probabilities or confidences, have obvious relevance in most areas of application, e.g. medical diagnosis and ROC analysis. We evaluate the predictive distributions of the approximate methods, and compare to the MCMC gold standard.

2. The Gaussian Process Model for Binary Classification

In this section we describe the Gaussian process model for binary classification (GPC). Let $y \in \{-1, 1\}$ denote the class label corresponding to an input $\mathbf{x}$. The GPC model is discriminative in the sense that it models $p(y|\mathbf{x})$ which for fixed $\mathbf{x}$ is a Bernoulli distribution. The probability of success $p(y=1|\mathbf{x})$ is related to an unconstrained latent function $f(\mathbf{x})$ which is mapped to the unit interval by a sigmoidal transformation, e.g. the *logit* or the *probit*. Both mappings are relatively similar around zero but show different tail behaviour. We will not examine the difference in this study. For reasons of analytic convenience (for the EP algorithm) we exclusively use the probit model $p(y=1|\mathbf{x}) = \Phi(f(\mathbf{x}))$, where Φ denotes the cumulative density function of the standard normal distribution.

In the GPC model Bayesian inference is performed about the latent function f in the light of observed data $\mathcal{D} = \{(y_i, \mathbf{x}_i)|i=1,\ldots,m\}$. Let $f_i = f(\mathbf{x}_i)$ and $\mathbf{f} = [f_1,\ldots,f_m]^\top$ be shorthand for the values of the latent function and $\mathbf{y} = [y_1,\ldots,y_m]^\top$ and $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_m]^\top$ collect the class labels and inputs respectively.

Given the latent function, the class labels are independent Bernoulli variables, so the joint likelihood factorises:

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{m} p(y_i|f_i) \tag{1}$$

and depends on f only through its value at the corresponding observed inputs. For the probit model the individual likelihood terms become $p(y_i|f_i) = \Phi(y_i f_i)$, due to the symmetry of Φ.

As prior over functions f we use a zero-mean Gaussian process (GP) prior (O'Hagan, 1978). A GP is a stochastic process where each input $\mathbf{x}$ has an associated random variable $f(\mathbf{x})$. The joint distribution of function values corresponding to any set of inputs $\mathbf{X}$ is multivariate Gaussian $p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{K})$. The covariance matrix is defined element-wise, $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j, \boldsymbol{\theta})$ where k is a positive definite covariance function parameterised by $\boldsymbol{\theta}$. Note that by choosing a covariance function we introduce *hyper-parameters* $\boldsymbol{\theta}$ to the prior. The zero-mean GP prior encodes that *a priori* $p(y=1|\mathbf{x}) = 1/2$ and certain further beliefs about the characteristics of the latent function.

For details on covariance functions and their implications on the prior over functions see for example Abrahamsen (1997) or Rasmussen and Williams (2006, ch. 4).

Using Bayes' rule the posterior distribution over the latent function values $\mathbf{f}$ for given hyper-parameters $\boldsymbol{\theta}$ becomes:

$$p(\mathbf{f}|\mathcal{D},\boldsymbol{\theta}) = \frac{p(\mathbf{y}|\mathbf{f})\,p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})}{p(\mathcal{D}|\boldsymbol{\theta})} = \frac{\mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{K})}{p(\mathcal{D}|\boldsymbol{\theta})}\prod_{i=1}^{m}\Phi(y_i f_i) \tag{2}$$

which is non-Gaussian. Properties of the posterior will be described in Section 5.

The main purpose of classification models is to predict the class label y_* for test inputs $\mathbf{x}_*$. The distribution of the latent function value can be computed by marginalisation:

$$p(f_*|\mathcal{D},\boldsymbol{\theta},\mathbf{x}_*) = \int p(f_*|\mathbf{f},\mathbf{X},\boldsymbol{\theta},\mathbf{x}_*)p(\mathbf{f}|\mathcal{D},\boldsymbol{\theta})d\mathbf{f}, \tag{3}$$

and by computing the expectation:

$$p(y_*|\mathcal{D},\boldsymbol{\theta},\mathbf{x}_*) = \int p(y_*|f_*)p(f_*|\mathcal{D},\boldsymbol{\theta},\mathbf{x}_*)df_* \tag{4}$$

the predictive distribution is obtained, which is again a Bernoulli distribution. The first term in the right hand side of equation (3) is Gaussian and obtained by conditioning the joint Gaussian prior distribution.

Unfortunately, neither the posterior eq. (2) $p(\mathbf{f}|\mathcal{D},\boldsymbol{\theta})$, the predictive distribution eq. (4) $p(y_* = 1|\mathcal{D},\boldsymbol{\theta},\mathbf{x}_*)$ nor the marginal likelihood eq. (7) $p(\mathcal{D}|\boldsymbol{\theta})$ can be computed analytically, so approximations are needed. For the GPC model approximations are either based on a Gaussian approximation $q(\mathbf{f}|\mathcal{D},\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{A})$ to the posterior $p(\mathbf{f}|\mathcal{D},\boldsymbol{\theta})$ or involve Markov chain Monte Carlo (MCMC) sampling.

A key insight is that a Gaussian approximation to the posterior implies a GP approximation to the posterior process, which gives rise to an approximate predictive distribution for test cases. Introducing the approximate Gaussian posterior into eq. (3) gives the approximate posterior $q(f_*|\mathcal{D},\boldsymbol{\theta},\mathbf{x}_*) = \mathcal{N}(f_*|\mu_*,\sigma_*^2)$, with mean and variance:

$$\mu_* = \mathbf{k}_*^\top\mathbf{K}^{-1}\mathbf{m} \tag{5a}$$

$$\sigma_*^2 = k(\mathbf{x}_*,\mathbf{x}_*) - \mathbf{k}_*^\top(\mathbf{K}^{-1} - \mathbf{K}^{-1}\mathbf{A}\mathbf{K}^{-1})\mathbf{k}_*, \tag{5b}$$

where $\mathbf{k}_* = [k(\mathbf{x}_1,\mathbf{x}_*),\ldots,k(\mathbf{x}_m,\mathbf{x}_*)]^\top$ is a vector of prior covariances between $\mathbf{x}_*$ and the training inputs $\mathbf{X}$. For the probit likelihood the approximate predictive probability (4) of $\mathbf{x}_*$ belonging to class 1 can be computed analytically:

$$q(y_*=1|\mathcal{D},\boldsymbol{\theta},\mathbf{x}_*) = \int \Phi(f_*)\,\mathcal{N}(f_*|\mu_*,\sigma_*^2)df_* = \Phi\Big(\frac{\mu_*}{\sqrt{1+\sigma_*^2}}\Big). \tag{6}$$

The parameters $\mathbf{m}$ and $\mathbf{A}$ of the posterior approximation can be found using Laplace's method (Section 3) or by Expectation Propagation (Section 4).

We have introduced the hyper-parameters $\boldsymbol{\theta}$ which we considered to be fixed. Typically very little information about these parameters is available *a priori*. In principle inference should be done jointly over f and $\boldsymbol{\theta}$ which can only be approximated using Markov chain Monte Carlo sampling.

However, a model selection approach can be implemented by selecting $\boldsymbol{\theta}$ maximising the marginal likelihood (evidence):

$$p(\mathcal{D}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})\, p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta}) d\mathbf{f} \tag{7}$$

which can be understood as a measure of the agreement between the model and observed data (Kass and Raftery, 1995; MacKay, 1999). This approach is called maximum likelihood II (ML-II) type hyper-parameter estimation and motivates the need for computing the marginal likelihood. Laplace's method as well as Expectation Propagation provide an approximation to the marginal likelihood (7) and so approximate ML-II hyper-parameter estimation can be implemented in both approximation schemes.

3. Laplace's Method

Williams and Barber (1998) describe Laplace's method to find a Gaussian $\mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{A})$ approximation to the posterior over latent function values (2) for fixed $\boldsymbol{\theta}$ (although they use the *logit* likelihood). Let $\ln \mathcal{L}(\mathbf{f}) \;=\; \ln p(\mathbf{y}|\mathbf{f})$ denote the log likelihood and:

$$\ln Q(\mathbf{f}|\mathcal{D},\boldsymbol{\theta}) = \ln \mathcal{L}(\mathbf{f}) - \frac{1}{2}\ln|\mathbf{K}| - \frac{1}{2}\mathbf{f}^\top \mathbf{K}^{-1}\mathbf{f} - \frac{m}{2}\ln(2\pi) \tag{8}$$

the unnormalised log posterior. Laplace's approximation is found by a second order Taylor expansion:

$$\ln Q(\mathbf{f}|\mathcal{D},\boldsymbol{\theta}) \simeq \ln Q(\mathbf{m}) - \frac{1}{2}(\mathbf{m}-\mathbf{f})^\top \mathbf{A}^{-1}(\mathbf{m}-\mathbf{f}) \tag{9}$$

around the mode of the (log) posterior:

$$\mathbf{m} = \underset{\mathbf{f}\in\mathbb{R}^m}{\operatorname{argmax}}\; \ln Q(\mathbf{f}|\mathcal{D},\boldsymbol{\theta}). \tag{10}$$

Since both the likelihood and the prior are log-concave the posterior is also log-concave and unimodal. Let:

$$\begin{aligned} \nabla_{\mathbf{f}} \ln Q &= \nabla_{\mathbf{f}} \ln \mathcal{L}(\mathbf{f}) - \mathbf{K}^{-1}\mathbf{f} && \text{(11a)}\\ \nabla\nabla_{\mathbf{f}} \ln Q &= \nabla\nabla_{\mathbf{f}} \ln \mathcal{L}(\mathbf{f}) - \mathbf{K}^{-1} && \text{(11b)} \end{aligned}$$

denote the gradient and the Hessian. The mode is conveniently found using Newton's method, iterating:

$$\mathbf{f} \;\leftarrow\; \mathbf{f} - (\nabla\nabla_{\mathbf{f}} \ln Q(\mathbf{f}))^{-1}\, \nabla_{\mathbf{f}} \ln Q(\mathbf{f}), \tag{12}$$

which usually converges rapidly to $\mathbf{m}$. The covariance matrix:

$$\mathbf{A} \;=\; -\big(\nabla\nabla_{\mathbf{f}} \ln Q(\mathbf{m})\big)^{-1} \;=\; (\mathbf{K}^{-1} + \mathbf{W})^{-1} \tag{13}$$

is approximated by the curvature at the mode, equal to the negative inverse Hessian, where $\mathbf{W} = -\nabla\nabla_{\mathbf{f}} \ln \mathcal{L}$.

This approximation also facilitates an approximation to the marginal likelihood:

$$p(\mathcal{D}|\boldsymbol{\theta}) \;=\; \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta}) d\mathbf{f} \;=\; \int \exp(\ln Q(\mathbf{f})) d\mathbf{f}. \tag{14}$$

Algorithm 1 Laplace's approximation for GPC

```
Given: $\boldsymbol{\theta}$, $\mathcal{D}$, $\mathbf{x}_*$
Initialise $\mathbf{f}$ (e.g. $\mathbf{f} \leftarrow \mathbf{0}$), compute $\mathbf{K}$ from $\boldsymbol{\theta}$ and $\mathbf{X}$
repeat
    $\mathbf{f} \leftarrow \mathbf{f} - (\nabla\nabla_{\mathbf{f}} \ln \mathcal{Q}(\mathbf{f}))^{-1} \nabla_{\mathbf{f}} \ln \mathcal{Q}(\mathbf{f})$
until convergence of $\mathbf{f}$
$\mathbf{m} \leftarrow \mathbf{f}$
$\mathbf{A} \leftarrow (\mathbf{K}^{-1} - \nabla\nabla_{\mathbf{f}} \ln \mathcal{Q}(\mathbf{m}))^{-1}$
Compute log marginal likelihood $\ln q(\mathcal{D}|\boldsymbol{\theta})$ by (15), and predictions $q(y_* = 1|\mathcal{D}, \boldsymbol{\theta}, \mathbf{x}_*)$ using (6).
```

Substituting $\ln \mathcal{Q}$ by its Taylor approximation (9) the Gaussian integral can be solved. The resulting approximate log marginal likelihood is:

$$\ln p(\mathcal{D}|\boldsymbol{\theta}) \;\simeq\; \ln q(\mathcal{D}|\boldsymbol{\theta}) \;=\; \ln \mathcal{Q}(\mathbf{m}) + \frac{m}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{A}| \tag{15}$$

and the derivative of this quantity w.r.t. $\boldsymbol{\theta}$ can be derived and used for optimisation (e.g. using conjugate gradient methods) in an ML-II type setting. See Algorithm 1 for an overview and Appendix A for details about our implementation.

4. Expectation Propagation

Minka (2001) proposed the iterative Expectation Propagation (EP) algorithm which can by applied to GPC. EP finds a Gaussian approximation $q(\mathbf{f}|\mathcal{D},\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{A})$ to the posterior $p(\mathbf{f}|\mathcal{D},\boldsymbol{\theta})$ by moment matching of approximate marginal distributions. The starting point is an approximation mimicking the factorising structure:

$$p(\mathbf{f}|\mathcal{D},\boldsymbol{\theta}) \;=\; \frac{p(\mathbf{f}|X,\boldsymbol{\theta})}{p(\mathcal{D}|\boldsymbol{\theta})}\prod_{i=1}^{m} p(y_i|f_i) \;\simeq\; \frac{p(\mathbf{f}|X,\boldsymbol{\theta})}{q(\mathcal{D}|\boldsymbol{\theta})}\prod_{i=1}^{m} t(f_i,\mu_i,\sigma_i^2,Z_i) \;=\; q(\mathbf{f}|\mathcal{D},\boldsymbol{\theta}), \tag{16}$$

where throughout we use p to denote exact quantities and q approximations, and the terms:

$$t(f_i,\mu_i,\sigma_i^2,Z_i) \;=\; Z_i\mathcal{N}(f_i|\mu_i,\sigma_i^2) \tag{17}$$

are called *site functions*. Note that the site functions are approximating the likelihood (which normalizes over observations y_i), with a Gaussian in f_i, so we cannot expect the site functions to normalize, hence the explicit term Z_i is necessary. For notational convenience we hide the *site parameters* μ_i, σ_i^2 and Z_i and write $t(f_i)$ instead. From (17) the Gaussian approximation (16) has mean and covariance:

$$q(\mathbf{f}|\mathcal{D},\boldsymbol{\theta}) \;=\; \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{A}), \;\text{ where }\; \mathbf{m} \;=\; \mathbf{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \;\text{ and }\; \mathbf{A} \;=\; (\mathbf{K}^{-1}+\boldsymbol{\Sigma}^{-1})^{-1}, \tag{18}$$

where $\boldsymbol{\mu} = (\mu_1,\dots,\mu_m)^\top$ and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2,\dots,\sigma_m^2)$ collect site function parameters. The EP algorithm iteratively visits each site function in turn, and adjusts the site parameters to match moments of an approximation to the posterior marginals. The kth moment of f_i under the posterior is:

$$\langle f_i^k \rangle \;=\; \frac{1}{p(\mathcal{D}|\boldsymbol{\theta})}\int f_i^k p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})d\mathbf{f} \;=\; \frac{1}{p(\mathcal{D}|\boldsymbol{\theta})}\int f_i^k p(y_i|f_i)\,p_{\setminus i}(f_i)d\,f_i \tag{19}$$

where:

$$p_{\backslash i}(f_i) \;=\; \int \prod_{j\neq i} p(y_j|f_j)p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})d\mathbf{f}^{\backslash i} \tag{20}$$

is called the *cavity distribution* and $\mathbf{f}^{\backslash i}$ denotes $\mathbf{f}$ without f_i. The marginalisation required to compute the exact cavity distribution is intractable for the GPC model. The key step in the EP algorithm is to replace the intractable exact cavity distribution with a tractable approximation based on the site functions:

$$q_{\backslash i}(f_i) \;=\; \int \prod_{j\neq i} t(f_j)p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})d\mathbf{f}^{\backslash i}. \tag{21}$$

The approximate cavity function comes in the form of an unnormalised Gaussian $q_{\backslash i}(f_i) \propto \mathcal{N}(f_i|\mu_{\backslash i},\sigma^2_{\backslash i})$. Multiplying both sides by $t(f_i)$:

$$q_{\backslash i}(f_i)t(f_i) \;=\; \int \mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{K})\prod_{j=1}^{m} t(f_j)d\mathbf{f}^{\backslash i} \;\propto\; \mathcal{N}(f_i|m_i,\mathbf{A}_{ii}), \tag{22}$$

and basic Gaussian identities give the parameters:

$$\sigma^2_{\backslash i} = \left((\mathbf{A}_{ii})^{-1} - \sigma_i^{-2}\right)^{-1} \quad \text{and} \quad \mu_{\backslash i} = \sigma^2_{\backslash i}\left(\frac{m_i}{\mathbf{A}_{ii}} - \frac{\mu_i}{\sigma_i^2}\right), \tag{23}$$

of the approximate cavity function.

The core idea of EP is to adjust the site parameters μ_i, σ_i and Z_i so that the approximate posterior marginal using the exact likelihood approximates as well as possible the posterior marginal based on the site function:

$$q_{\backslash i}(f_i)p(y_i|f_i) \;\simeq\; q_{\backslash i}(f_i)t(f_i,\mu_i,\sigma_i^2,Z_i) \tag{24}$$

by matching the zeroth, first and second moments. Recall that matching of moments minimizes Kullback-Leibler (KL) divergence.[1] For the probit likelihood $p(y_i|f_i) = \Phi(y_if_i)$ the $k = 0,1,2$ moments of the left hand side can be computed analytically

$$m_0 \;=\; \Phi\Big(\frac{y\mu_{\backslash i}}{\sqrt{1+\sigma^2_{\backslash i}}}\Big) \;=\; \Phi(z), \tag{25a}$$

$$m_1 \;=\; \mu_{\backslash i} + \frac{\sigma^2_{\backslash i}\mathcal{N}(z|0,1)}{\Phi(z)y\sqrt{1+\sigma^2_{\backslash i}}}, \tag{25b}$$

$$m_2 \;=\; 2\mu_{\backslash i}m_1 - \mu^2_{\backslash i} + \sigma^2_{\backslash i} - \frac{z\sigma^4_{\backslash i}\mathcal{N}(z|0,1)}{\Phi(z)(1+\sigma^2_{\backslash i})}, \tag{25c}$$

where $z = y\mu_{\backslash i}/\sqrt{1+\sigma^2_{\backslash i}}$. By equating these moments with those of the right hand side of (24) the update equations for the site parameters become

$$\sigma_i^2 \;=\; \left((m_2 - m_1^2)^{-1} - \sigma^{-2}_{\backslash i}\right)^{-1}, \tag{26a}$$

$$\mu_i \;=\; \sigma_i^2\Big(m_1(\sigma^{-2}_{\backslash i} + \sigma_i^{-2}) - \frac{\mu_{\backslash i}}{\sigma^2_{\backslash i}}\Big), \tag{26b}$$

$$Z_i \;=\; m_0\sqrt{2\pi(\sigma^2_{\backslash i}+\sigma_i^2)}\exp\left(\frac{(\mu_i - \mu_{\backslash i})^2}{2(\sigma^2_{\backslash i}+\sigma_i^2)}\right). \tag{26c}$$

[1] Although, the classical KL argument only applies to the first and second (and higher) moments for *normalized* distributions, it seems natural also to match zeroth moment.

Algorithm 2 EP for Gaussian process classification

```
Given: $\boldsymbol{\theta}$, $\mathcal{D}$, $\mathbf{x}_*$
Initialise: $\mathbf{A} \leftarrow \mathbf{K}$ and site parameters $\sigma_i^2$ and $\mu_i$
repeat
  for i=1,...,m do
    Compute parameters (23) of cavity
    Compute moments (25)
    Update the site parameters using (26)
    Update $\mathbf{m}$ and $\mathbf{A}$ according to (18)
  end for
until The site parameters converged
Compute log marginal likelihood $\ln q(\mathcal{D}|\boldsymbol{\theta})$ by (27), and predictions $q(y_* = 1|\mathcal{D}, \boldsymbol{\theta}, \mathbf{x}_*)$ using (6).
```

In the application of EP, one may generally not have a guarantee that the new site variance in (26a) is non-negative; however, in the GPC model with probit likelihood, one can show that variance is always positive. Once we have new values for μ_i and σ_i^2 we have to update $\mathbf{m}$ and $\mathbf{A}$ according to (18), which in practise is done using rank-one updates, to save computation.

The EP algorithm iteratively updates the site parameters as shown in Algorithm 2. Although we cannot prove the convergence of EP, we conjecture that it always converges for GPC with probit likelihood, and have never encountered an exception.

Finally the approximate log marginal likelihood can be obtained from the normalization of (16), giving

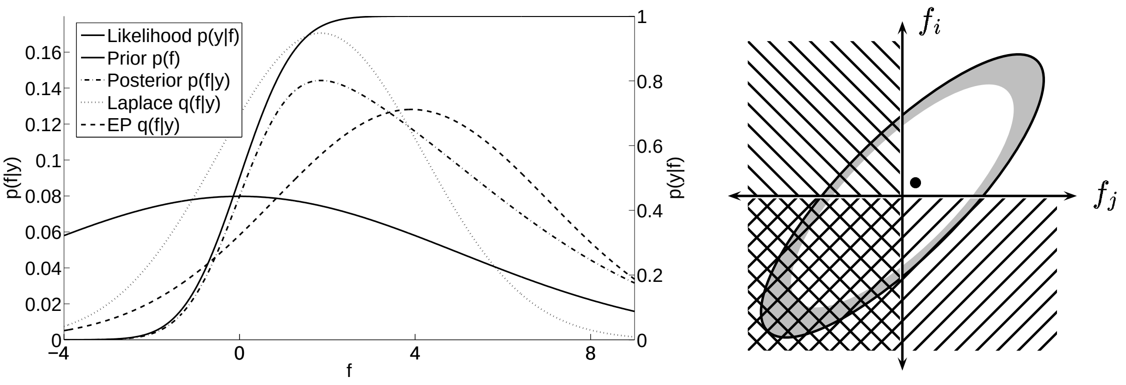
$$\begin{aligned} \ln p(\mathcal{D}|\boldsymbol{\theta}) \simeq \ln q(\mathcal{D}|\boldsymbol{\theta}) &= \ln \int q(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}) \prod_{i=1}^{m} t(f_i) d\mathbf{f} \\ &= \sum_{i=1}^{n} \ln Z_i - \frac{1}{2} \ln |\mathbf{K} + \boldsymbol{\Sigma}| - \frac{1}{2} \boldsymbol{\mu}^\top (\mathbf{K} + \boldsymbol{\Sigma})^{-1} \boldsymbol{\mu} - \frac{m}{2} \ln(2\pi). \end{aligned} \tag{27}$$

Perhaps this is not the standard way to compute an approximation to the marginal likelihood used elsewhere, but it seems the most natural given the approximation. The derivatives of the log marginal likelihood can be computed in order to implement ML-II parameter estimation of $\boldsymbol{\theta}$. Algorithm 2 summarises the computations, more details on implementing EP for GPC can be found in Appendix B.

5. Structural Properties of the Posterior

In the previous sections we described the GPC model and two alternative approximation schemes for finding a Gaussian approximation to the posterior. This section provides more details on the properties of the posterior which is compared to the structure of the respective approximations.

Figure 1(a) provides a one-dimensional illustration. The prior $\mathcal{N}(f|0, 5^2)$ combined with the probit likelihood $(y = 1)$ results in a skewed posterior. Intuitively, the likelihood cuts off the f values which have the opposite sign of y. The mode of the posterior remains relatively close to the origin, while the mass is placed over positive values in accordance with the observation. Laplace's approximation peaks at the posterior mode, but places far too much mass over negative values of f and too little over large positive values. The EP approximation attempts to match the first two



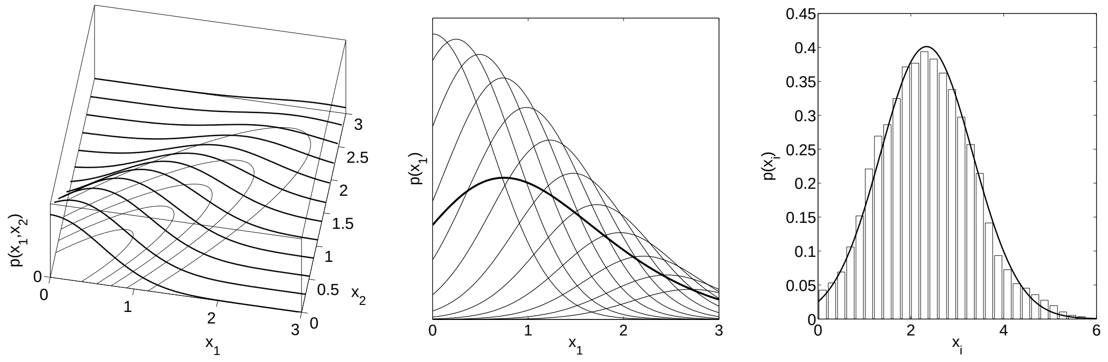
(a) (b)

Figure 1: Panel (a) provides a one-dimensional illustration of approximations. The prior $\mathcal{N}(f|0,5^2)$ combined with the probit likelihood $(y=1)$ results in a skewed posterior. The likelihood uses the right axis, all other curves use the left axis. In Panel (b) we caricature a high dimensional zero-mean Gaussian prior as an ellipse. The gray shadow indicates that for a high dimension Gaussian most of the mass lies in a thin shell. For large latent signals, the likelihood essentially cuts off regions which are incompatible with the training labels (hatched area), leaving the upper right orthant as the posterior. The dot represents the mode of the posterior, which is relatively unaffected by the truncation and remains close to the origin.

posterior moments, which results in a larger mean and a more accurate placement of probability mass compared to Laplace's approximation.

Structural properties of the posterior in higher dimensions can best be understood by examining its construction. The prior is a correlated m-dimensional Gaussian $\mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{K})$ centred at the origin. Each likelihood term $p(y_i|f_i)$ softly truncates the half-space from the prior that is incompatible with the observed label, see Figure 1(b). The resulting posterior is unimodal and skewed, similar to a multivariate Gaussian truncated to the orthant containing $\mathbf{y}$. The mode of the posterior remains close to the origin, while the mass is placed in accordance with the observed class labels. Additionally, high dimensional Gaussian distributions exhibit the property that most probability mass is contained in a thin ellipsoidal shell—depending on the covariance structure—away from the mean (MacKay, 2003, ch. 29.2). Intuitively this occurs since in high dimensions the volume grows extremely rapidly with the radius. As an effect the mode becomes less representative (typical) for the prior distribution as the dimension increases. For the GPC posterior this property persists: the mode of the posterior distribution stays relatively close to the origin, still being unrepresentative for the posterior distribution, while the mean moves to the mass of the posterior making mean and mode differ significantly.

As described, we cannot generally assume the posterior to be close to Gaussian, as in the often studied limit of low-dimensional parametric models with large amounts of data. Therefore in GPC we must be aware of making a Gaussian approximation to a non-Gaussian posterior. Laplace's approximation is centred around the mode of the posterior, which lies in the right orthant but too close



(a) (b) (c)

Figure 2: Panel (a) illustrates a bivariate normal distribution truncated to the positive quadrant. The lines describe slices through the probability density function for fixed x_2-values. Panel (b) shows the marginal distribution of $p(x_1)$ (thick line) obtained by (numerical) integration over x_2, which—intuitively speaking—corresponds to an averaging of the slices (thin lines) from Panel (a). Panel (c) shows a histogram of samples of a marginal distribution of an high-dimensional truncated Gaussian. The line describes a Gaussian with mean and variance estimated from the samples.

to the origin, such that the approximation will overlap with regions having practically zero posterior mass. As an effect the amplitude of the approximate latent posterior GP will be underestimated systematically, leading to overly cautious predictive distributions.

The EP approximation does not rely on a local expansion, but assumes that the marginal distributions of the posterior can be well approximated by Gaussians. As described above the posterior is similar to a high dimensional multivariate normal distribution truncated to one orthant. Although the posterior is skew and truncated, marginals of such a distribution can be relatively similar to a Gaussian.

As a low dimensional illustration the marginal distribution of a bivariate normal is shown in Figure 2(a-b). Depending on the covariance structure, the mode of the marginal distribution moves away from the origin and the distribution appear similar to a truncated univariate Gaussian.

In order to inspect the marginals of a truncated high-dimensional multivariate normal distribution we made an additional synthetic experiment. We constructed a 767 dimensional Gaussian $\mathcal{N}(\mathbf{x}|\mathbf{0},\mathbf{C})$ with a covariance matrix having one eigenvalue of 100 with eigenvector $\mathbf{1}$, and all other eigenvalues are 1. We then truncate this distribution such that all $\mathbf{x}_i \geq 0$. Note that the mode of the truncated Gaussian is still at zero, whereas the mean moved towards the remaining mass. Metropolis-Hastings sampling was used to generate samples from this truncated multivariate distribution. Figure 2(c) shows a normalised histogram of samples from a marginal distribution of one $\mathbf{x}_i$. The samples agree very well with a Gaussian approximation. Note that Laplace's method would be completely inappropriate for approximating a truncated multivariate normal distribution.

In order to validate the above arguments we will use Markov chain Monte Carlo methods to generate samples from the posterior and also to estimate the marginal likelihood.

6. Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) may be too slow for many practical applications, but has the advantage that it becomes exact in the limit of long runs. Thus, MCMC can provide a *gold standard* by which to measure the two analytic methods of the previous sections. Computing the predictions via an MCMC estimate of (3) and (4) is relatively straight forward and covered in Section 6.1.

Good MCMC estimates of the marginal likelihood are, however, notoriously difficult to obtain, being equivalent to the free-energy estimation problem in physics (Gelman and Meng, 1998). In Section 6.2 we explain the use of Annealed Importance Sampling (AIS), which can be seen as a sophisticated elaboration of Thermodynamic Integration, for this task.

6.1 Hybrid MCMC Sampling

Hybrid Monte Carlo (HMC) sampling as proposed by Duane et al. (1987) is a computationally efficient sampling technique which exploits gradient information of the target distribution. Detailed accounts are given by Neal (1993, ch. 5.2) and Liu (2001, ch. 9). MacKay (2003, ch. 30) also provides pseudo-code; we do not repeat the details here.

HMC can be used to generate samples from the posterior $p(\mathbf{f}|\boldsymbol{\theta},\mathcal{D})$, while only the unnormalised log posterior (8) and its derivatives are required. As described in the previous section, the exact posterior (2) takes the form of a (correlated) Gaussian (the GP prior), which is (softly) truncated by the constraints imposed by the training labels through the likelihood. To ease the sampling task by reducing correlations, we first do a linear transformation into new $\mathbf{g}=\mathbf{L}^{-1}\mathbf{f}$ variables, such that $\mathbf{g}$ is *white* w.r.t. $\mathbf{K}$, where $\mathbf{K}=\mathbf{L}\mathbf{L}^{\top}$ is the Cholesky decomposition. Given samples from the posterior, we generate test-latents from the Gaussian $p(f_*|\mathbf{f},\mathbf{X},\boldsymbol{\theta},\mathbf{x}_*)$ for use in a simple Monte Carlo estimate of (4).

6.2 Annealed Importance Sampling

The marginal likelihood (7) comes in the form of an m dimensional integral where m is the number of data points. A simple approach would be to use importance sampling with the EP or Laplace's approximation of the posterior as proposal distribution. However, for the GPC model the resulting importance weights show enormous variances, making simple importance sampling useless for this task (MacKay, 2003, ch. 29).

Neal (2001) describes Annealed Importance Sampling (AIS), which we will use to estimate the marginal likelihood in the GPC model. Instead of solving the integral (7) directly, a sequence of easier quantities is computed. We define:

$$Z_t \;=\; \int p(\mathbf{y}|\mathbf{f})^{\tau(t)}p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})d\mathbf{f} \tag{28}$$

where $\tau(t)$ is an inverse temperature schedule such that $\tau(0)=0$ and $\tau(T)=1$. The trick is to rewrite the marginal likelihood $Z=p(\mathcal{D}|\boldsymbol{\theta})$ as a fraction and expand:

$$Z \;=\; \frac{Z_T}{Z_0} \;=\; \frac{Z_T}{Z_{T-1}}\frac{Z_{T-1}}{Z_{T-2}}\cdots\frac{Z_1}{Z_0}, \tag{29}$$

Algorithm 3 Annealed Importance Sampling

```
Given: Temperature schedule $\tau$
for $r = 1, \ldots, R$ do
    Sample $\mathbf{f}_0$ from the prior $\mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K})$
    for $t = 1, \ldots, T$ do
        Sample $\mathbf{f}_t$ from $q(\mathbf{f}|\mathcal{D}, \tau(t), \boldsymbol{\theta})$ by HMC
        Compute $\ln(Z_t/Z_{t-1})$ using (31)
    end for
    Compute $Z_r$ using (32)
end for
Return $\ln Z = \ln\big(\frac{1}{R}\sum_{r=1}^{R} Z_r\big)$
```

where $Z_0 = 1$ since the prior normalises. Each term in (29) is approximated using importance sampling using samples from $q(\mathbf{f}|\mathcal{D}, \boldsymbol{\theta}, \tau(t)) \propto p(\mathbf{y}|\mathbf{f})^{\tau(t)} p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})$:

$$\frac{Z_t}{Z_{t-1}} = \int \frac{p(\mathbf{y}|\mathbf{f})^{\tau(t)} p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})}{p(\mathbf{y}|\mathbf{f})^{\tau(t-1)} p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})} q(\mathbf{f}|\mathcal{D}, \boldsymbol{\theta}, \tau(t-1)) d\mathbf{f} \tag{30a}$$

$$\simeq \frac{1}{S} \sum_{i=1}^{S} p(\mathbf{y}|\mathbf{f}_i)^{\tau(t)-\tau(t-1)} \tag{30b}$$

where $\mathbf{f}_i$ are samples from $q(\mathbf{f}|\mathcal{D}, \boldsymbol{\theta}, \tau(t))$, which we generate using HMC. Using a single sample $S = 1$ and a large number of temperatures, the log of each ratio is:

$$\ln(Z_t/Z_{t-1}) \simeq \big(\tau(t) - \tau(t-1)\big) \ln p(\mathbf{y}|\mathbf{f}_t) \tag{31}$$

where $\mathbf{f}_t$ is the only sample at temperature $\tau(t)$. Combining (29) with (31) we obtain the desired:

$$\ln Z \simeq \sum_{t=1}^{T} \ln(Z_t/Z_{t-1}). \tag{32}$$

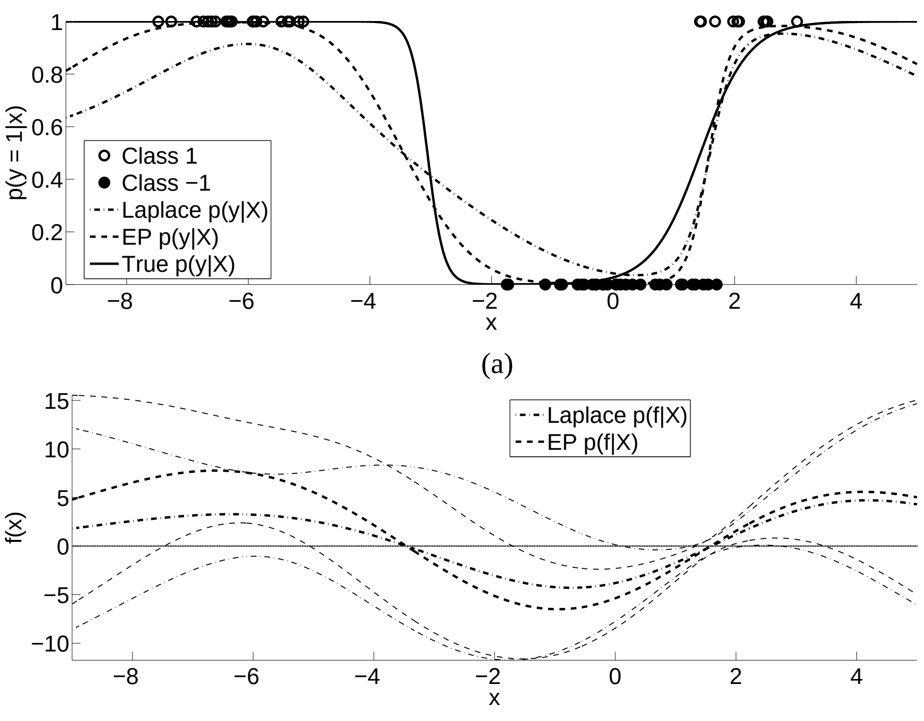
In all our experiments we use $\tau(t) = (t/T)^4$ for $t = 0, \ldots, 8000$. Using this temperature schedule we found that the sampling spends most of its efforts at temperatures with high variance of (31) such that the variance of (32) is relatively small. Note that this was only examined on the data sets we use below and only for certain values of $\boldsymbol{\theta}$. So far, we have described Thermodynamic Integration, which gives an unbiased estimate in the limit of slow temperature changes. In AIS the bias caused by finite temperature schedules is removed by combining multiple estimates by their geometric mean (see Algorithm 3). In the experiments we combine the estimates of $R = 3$ runs of Thermodynamic Integration.

7. Experiments

In this section we compare and inspect approximations for GPC using various benchmark data sets. The primary focus is not to optimise the absolute performance of GPC models but to compare the relative accuracy of approximations and to validate the arguments given in Section 5.

In all the GPC experiments we use a covariance function of the form:

$$k(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}) = \sigma^2 \exp\big(-\tfrac{1}{2\ell^2} \left\|\mathbf{x} - \mathbf{x}'\right\|^2\big), \tag{33}$$



(a)

(b)

Figure 3: Synthetic classification problem: Panel (a) illustrates the classification task, the generating $p(y|x)$ and two approximations thereof obtained by Laplace's method and EP. Panel (b) illustrates the approximate predictive distributions $p(f_*|\mathcal{D},\boldsymbol{\theta},x_*) \simeq \mathcal{N}(f_*|\mu_*,\sigma_*^2)$ of latent function values showing the mean μ_* and the range of $\pm 2\sigma_*$.

such that $\boldsymbol{\theta} = [\sigma, \ell]$. We refer to σ^2 as the signal variance and to ℓ as the characteristic length-scale. Note that for many classification tasks it may be reasonable to use an individual length scale parameter for every input dimension (ARD). Nevertheless, for the sake of presentability we use the above covariance function and we believe the conclusions to be independent of this choice.

Both analytic approximations have a computational complexity which is cubic $O(m^3)$ as common among non-sparse GP models due to inversions $m \times m$ matrices. In our implementations Laplace's method and EP need similar running times, on the order of a few minutes for several hundred data-points. Making AIS work efficiently requires some fine-tuning and a single estimate of $p(\mathcal{D}|\boldsymbol{\theta})$ can take several hours for data sets of a few hundred examples, but this could conceivably be improved upon.

7.1 Synthetic Classification Problem

The first experiment is a synthetic classification problem with scalar inputs. The observations for class 1 were generated from two normal distributions with means -6 and 2, each with a standard deviation of 0.8. For class -1 the mean is 0 and the same standard deviation was used.

We computed Laplace's and the EP approximation for the ML-II estimated value of $\boldsymbol{\theta}$ that maximised Laplace's approximation to the marginal likelihood (15). Note that this particular choice of $\boldsymbol{\theta}$ should be in favour of Laplace's method. Figure 3 shows the resulting classifiers and the underlying latent functions. In Figure 3(a) the approximations to $p(y|x)$ appear to be similar for positive x but we observe an appreciable discrepancy for negative values. Laplace's approximation gives an unreasonably high predictive uncertainty, which is caused by a significant overlap of the approximate predictive distribution $p(f_*|\mathcal{D},\boldsymbol{\theta},x_*) \simeq \mathcal{N}(f_*|\mu_*,\sigma_*^2)$ with zero as shown in Figure 3(b). However, note that both approximations agree on the sign of the predictive mean.

7.2 Ionosphere Data

The data consists of 351 examples in 34 dimensions. We standardised the inputs $\mathbf{X}$ to zero mean and unit variance. The training set is a random subset of size $m = 200$ leaving the remaining 151 instances out as a test set.

We do an exhaustive investigation on a regular 21×21 grid of values for the log hyper-parameters. For each $\boldsymbol{\theta}$ on the grid we compute the approximated log marginal likelihood by Laplace's method (15), EP (27) and AIS. Additionally we compute the predictive performance on the test set. As performance measure we use the average information in bits of the predictions about the test targets in excess of that of random guessing. Let $p^* = p(y_* = 1|\mathbf{x}_*)$ be the model's prediction, then we average:

$$I(p_i^*,y_i) = \tfrac{y_i+1}{2}\log_2(p_i^*) + \tfrac{1-y_i}{2}\log_2(1-p_i^*) + H \tag{34}$$

over all test cases, where H is the entropy of the training set labels. Results are shown in Figure 4.

For all three approximation techniques we see an agreement between marginal likelihood estimates and test performance, which justifies the use of ML-II parameter estimation. But the shape of the contours and the values differ between the methods. The contours for Laplace's method appear to be *slanted* compared to EP. The estimated marginal likelihood estimates of EP and AIS agree very well.[2] The EP predictions contain as much information about the test cases as the MCMC predictions and significantly more than for Laplace's method.

Note that for small signal variances (roughly $\ln(\sigma^2) < 0$) Laplace's method and EP give very similar results. A possible explanation is that for small signal variances the likelihood does not *truncate* the prior but only *down-weights* the tail that disagrees with the observation. As an effect the posterior will be less skewed and both approximations will lead to similar results.

7.3 USPS Digits

We define a binary sub-problem from the USPS digit data[3] by considering 3's vs. 5's. We repeated the experiments described in the previous section for a slightly modified grid of $\boldsymbol{\theta}$. Comparing the results shown in Figure 5 leads to similar results as mentioned above. The EP and MCMC results agree very well, given that the marginal likelihood comes as a 767 dimensional integral.

We now take a closer look at the approximations $q(\mathbf{f}|\mathcal{D},\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{A})$ for a given value of $\boldsymbol{\theta}$. We have chosen the values $\ln(\sigma) = 3.35$ and $\ln(\ell) = 2.85$ which are between the ML-II estimates of EP and Laplace's method. Comparing the respective means of the approximations in Figure 6(a) we

[2] Note that the agreement between the two seems to be limited by the accuracy of the AIS runs, as judged by the regularity of the contour lines; the tolerance is less than one unit on a (natural) log scale.

[3] Because the training and test partitions in the original data differ significantly, we pooled cases and randomly divided them into new sets, with 767 cases for training and 773 for testing.

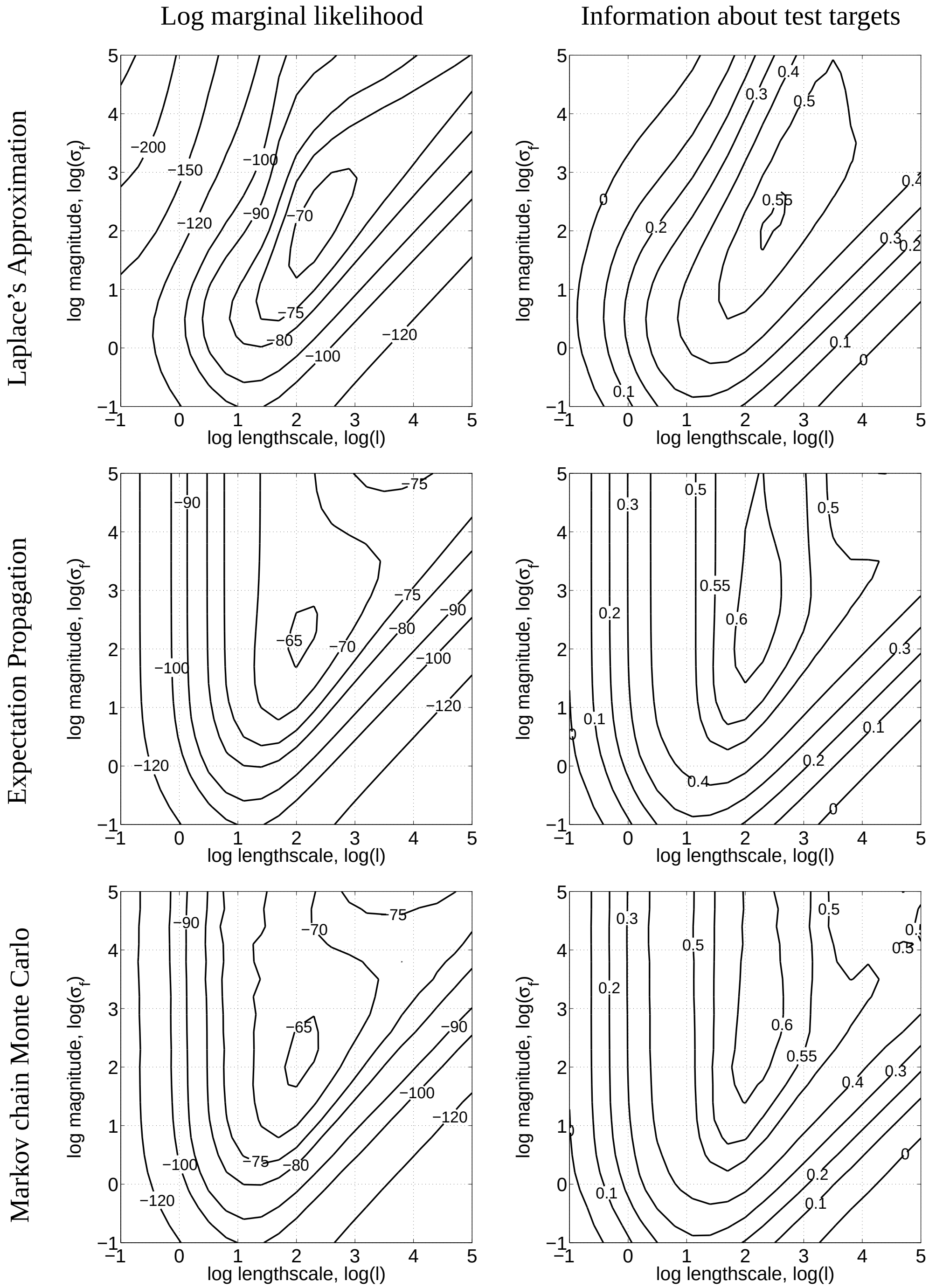


Figure 4: Comparison of marginal likelihood approximations and predictive performances for the Ionosphere data set. The first column shows the estimates of log marginal likelihood, while the second column shows the performance on the test set measured by the information about test targets in bits (34).

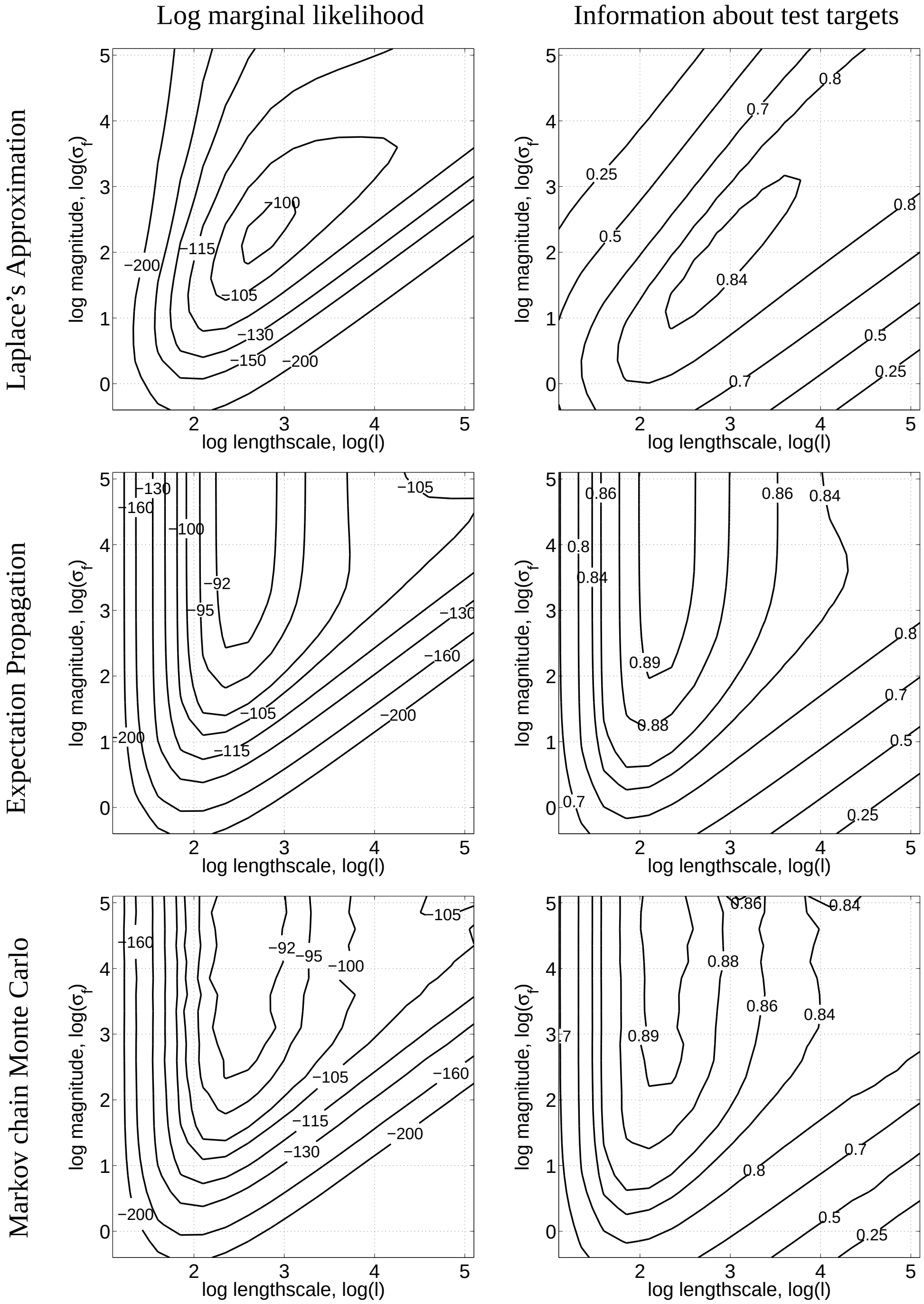
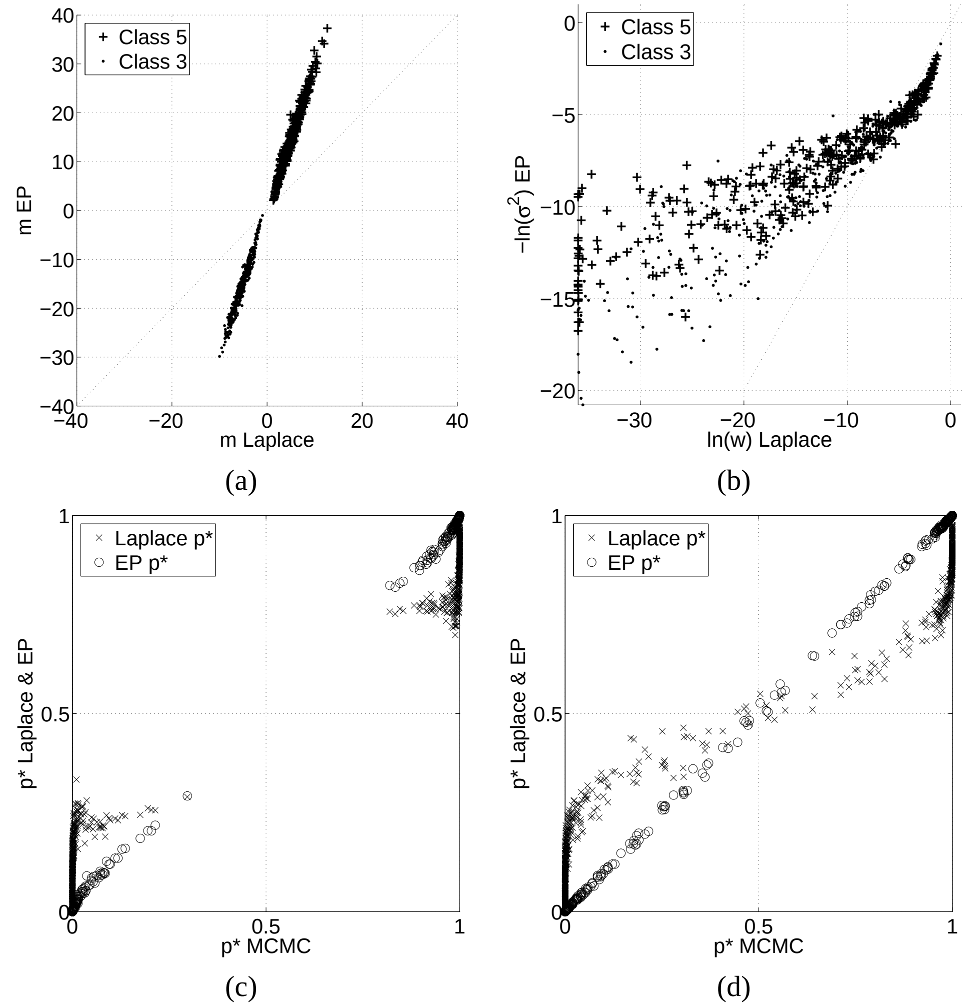


Figure 5: Comparison of marginal likelihood approximations and predictive performances of the different methods for classifying 3's vs. 5's from the USPS image database. The plots are arranged as in Figure 4.



(a)

(b)

(c)

(d)

Figure 6: Comparison of approximations $q(\mathbf{f}|\mathcal{D},\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{A})$ for a given value of $\boldsymbol{\theta}$. Panel (a) shows a comparison of the means $\mathbf{m}_i$. In Panel (b) we compare the elements of the diagonal matrices $\mathbf{W}_{ii}$ and $\boldsymbol{\Sigma}_{ii}$. Panels (c) and (d) compare predictions p^* obtained by MCMC (abscissa) to predictions obtained from Laplace's method and EP (ordinate). Panel (c) shows predictions on training cases and (d) shows predictions on test cases.

see that the magnitude of the means from the Laplace approximation is much smaller than from EP. The relation appears to be roughly linear. In Figure 6(b) we compare the elements of $\mathbf{W}$ and $\boldsymbol{\Sigma}^{-1}$ which cause the difference in the approximations (13) and (18) of the posterior covariance matrix $\mathbf{A}$. We observe that the relatively large entries in $\mathbf{W}$ are larger than the corresponding entries in $\boldsymbol{\Sigma}^{-1}$, but in total $\mathbf{W}$ contains more small values than $\boldsymbol{\Sigma}^{-1}$. The exact effect on the posterior covariance is difficult to characterise due to the inversion, but intuitively the smaller the values the more the posterior covariance will be similar to the prior.

Figures 6(c-d) compare the predictive uncertainty p^* resulting from the respective approximations to MCMC predictions. For both training and test set we observe that EP and MCMC agree very well, while Laplace's method shows over-conservative predictions.

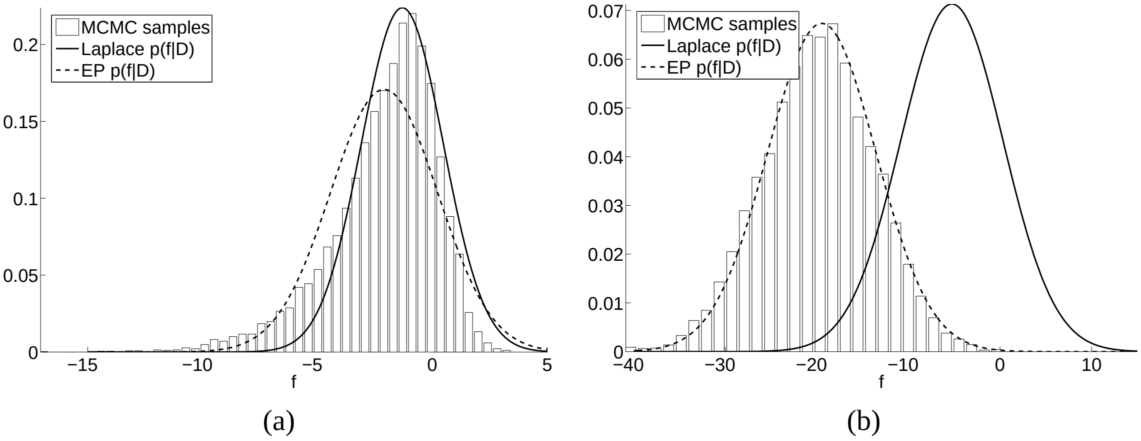


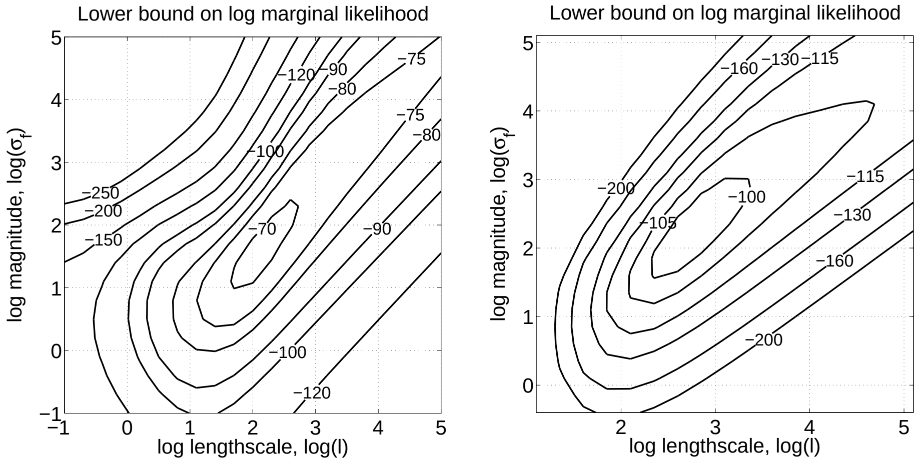
Figure 7: Two marginal distributions $p(f_i|\mathcal{D},\boldsymbol{\theta})$ from the posterior. For Panel (a) we picked the f_i for which the posterior marginal is maximally skewed (see again Figure 1). The true posterior is approximated by a normalised histogram of 9000 samples of f_i obtained by MCMC sampling. Panel (b) shows a case where EP and Laplace's approximation differ significantly.

We now inspect the marginal distributions $p(f_i|\mathcal{D},\boldsymbol{\theta})$ of single latent function values under the posterior approximation. We use hybrid MCMC to generate 9000 samples from the posterior of $\mathbf{f}$ for the above $\boldsymbol{\theta}$. For Laplace's method and EP the approximated distribution is $\mathcal{N}(f_i|\mathbf{m}_i,\mathbf{A}_{ii})$ where $\mathbf{m}$ and $\mathbf{A}$ are found by the respective approximation techniques.

In general we observe that the marginal distributions of MCMC samples agree very well with the respective marginal distributions of the EP approximation. This supports the claim made in Section 5 where we argued that the marginal distributions of the posterior can be very similar to Gaussians, even if the posterior is a skew distribution. For Laplace's approximation we find the mean to be underestimated and the marginal distributions to overlap with zero far more than the EP approximations. Figure 7(a) displays the marginal distribution and its approximations for which the MCMC samples show maximal skewness. Figure 7(b) shows a typical example where the EP approximation agrees very well with the MCMC samples. We show this particular example because under the EP approximation $q(y_i = 1|\mathcal{D},\boldsymbol{\theta}) < 0.1\%$ but Laplace's approximation gives $q(y_i = 1|\mathcal{D},\boldsymbol{\theta}) \simeq 18\%$.

7.4 Lower Bound Approximation

In the context of sparse EP approximations Seeger (2003) proposed a lower bound on the marginal likelihood. The bound is obtained from the EP approximation of the posterior using Jensen's in-



(a) Ionosphere

(b) USPS 3's vs. 5's

Figure 8: Lower bound on marginal likelihood. Panel (a) shows the lower bound eq. (35) on the marginal likelihood for the Ionosphere data set (compare to left column of Figure 4). Panel (b) shows the value of the lower bound for the USPS 3's vs. 5's (compare to left column of Figure 5)

equality:

$$\begin{aligned}
\ln p(\mathcal{D}|\boldsymbol{\theta}) &= \ln \int p(\mathbf{y}|\mathbf{f})\mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{K})d\mathbf{f} && \text{(35a)}\\
&\geq \int \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{A}) \ln \frac{p(\mathbf{y}|\mathbf{f})\mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{K})}{\mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{A})} d\mathbf{f} && \text{(35b)}\\
&= \sum_{i=1}^{m} \int \mathcal{N}(f_i|\mathbf{m}_i,\mathbf{A}_{ii}) \ln \Phi(y_i f_i) d f_i \\
&\quad -\frac{1}{2}\mathbf{m}^\top \mathbf{K}^{-1}\mathbf{m} - \frac{1}{2}\operatorname{tr}(\mathbf{K}^{-1}\mathbf{A}) + \frac{1}{2}\ln|\mathbf{K}^{-1}\mathbf{A}| + \frac{m}{2}\ . && \text{(35c)}
\end{aligned}$$

Note that the one dimensional integrals in eq. (35c) have to be solved using numerical integration methods.

In sparse EP methods the Gaussian approximation is based on only a subset of observations and so the evidence (27) may be a bad approximation of the total evidence since it does not take all available data into account. Assume that the m points are only a subset of of a total of m' observations. The lower bound (35c) can be extended to a lower bound on all m' observations by including all points in the one dimensional integrals over the individual log likelihood terms.

Several authors maximise this lower bound instead of maximising (27) for ML-II hyper-parameter estimation also in the case of non-sparse EP approximations, e.g. Chu and Ghahramani (2005). In Figure 8 we show the value of the lower bound as a function of the hyper-parameters for the Ionosphere and USPS data described in the previous sections (for the full EP approximation). Interestingly, for both data sets the lower bounds appear to be more similar to the approximate evidence obtained by Laplace's method than by EP (compare to the upper left panel in Figures 4 and 5 respectively). However, the maxima of the lower bounds correspond to sub-optimal predictive performances compared to the maxima of the approximate marginal likelihood (27) (compare to the second row in Figures 4 and 5 respectively). Therefore for non-sparse EP approximations the use of (27) seems advisable, which is also computationally advantageous.

7.5 Benchmark Data Sets

In this section we compare the performance of Laplace's method and Expectation Propagation for GPC on several well known benchmark problems for binary classification.

The *Ionosphere*, the *Wisconsin* Breast Cancer, and the *Sonar* data sets are taken from Hettich et al. (1998). The Leptograpsus *Crabs* and the *Pima* Indians Diabetes data sets were described by Ripley (1996). Note that for the Crabs data set we use the sex (not the colour) of the crabs as target variable. The largest data set in the comparison are the 3's vs. 5's from the USPS handwritten digits described above.

We standardise the inputs $\mathbf{X}$ to zero mean and unit variance. All data sets are randomly split into 10 folds of which one at a time is left out as a test set to measure the predictive performance of a model trained (or selected) on the remaining nine folds.

For GPC we implement model selection by ML-II hyper-parameter estimation. We use a conjugate gradient optimisation routine to find a minimum

$$\boldsymbol{\theta}_{\mathrm{ML}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \; -\ln q(\mathcal{D}|\boldsymbol{\theta}) \tag{36}$$

of the negative log marginal likelihood approximated by Laplace's method (15) and EP (27) respectively. For the respective $\boldsymbol{\theta}_{\mathrm{ML}}$ the approximations $\mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{A})$ are computed and predictions are made for the left out test set. From the predictive distributions the average information (34) is computed and averaged over the ten folds. Furthermore the average error rate E is reported, which equals the average percentage of erroneous class assignments if prediction is understood as a decision problem with symmetric costs (thresholding the predictive uncertainty at $1/2$).

In order to have a better absolute impression of the predictive performance we report the results of support vector machines (SVM) (Schölkopf and Smola, 2002). We use the LIBSVM implementation of C-SVM by Chang and Lin (2001) with a radial basis function kernel which is equivalent to the covariance function (33) up to the signal variance parameter. The values of the length scale parameter ℓ and the regularisation parameter C are found by an *inner loop* of 5-fold cross-validation on the nine training folds respectively. We manually refine the parameter grids and repeat the cross-validation procedure until the performance stabilises.

We use the technique described by Platt (2000) to estimate predictive probabilities from an SVM. This is implemented by fitting a sigmoidal mapping from the unthresholded output of the SVM to the unit interval. The parameters of the mapping are estimated on the test set in the inner loop of 5-fold cross-validation.

Results are summarised in Table 1. Comparing Laplace's method to EP the latter shows to be more accurate both in terms of error rate and information. While the error rates are relatively similar the predictive distribution obtained by EP shows to be more informative about the test targets. As to be expected by now, the length of the mean vector $\|\mathbf{m}\|$ shows much larger values for the EP approximations. Comparing EP and SVM the results are mixed.

At first sight it may seem surprising that Laplace's method gives relatively similar error rates compared to EP. Note that for both methods the error rate only depends on the sign of the latent

			Laplace			EP			SVM	
Data Set	m	n	E	I	$\|\|\mathbf{m}\|\|$	E	I	$\|\|\mathbf{m}\|\|$	E	I
Ionosphere	351	34	8.84	0.591	49.96	7.99	0.661	124.94	5.69	0.681
Wisconsin	683	9	3.21	0.804	62.62	3.21	0.805	84.95	3.21	0.795
Pima Indians	768	8	22.77	0.252	29.05	22.63	0.253	47.49	23.01	0.232
Crabs	200	7	2.0	0.682	112.34	2.0	0.908	2552.97	2.0	0.047
Sonar	208	60	15.36	0.439	26.86	13.85	0.537	15678.55	11.14	0.567
USPS 3 vs 5	1540	256	2.27	0.849	163.05	2.21	0.902	22011.70	2.01	0.918

Table 1: Results for benchmark data sets. The first three columns give the name of the data set, number of observation m and dimension of inputs n. For Laplace's method and EP the table reports the average error rate E, the average information I (34) and the average length $||\mathbf{m}||$ of the mean vector of the Gaussian approximation. For SVMs the error rate and the average information about the test targets are reported.

mean function (5a) at the test locations, which in turn depend on $\mathbf{m}$ only. Therefore the error rate is less sensitive to the accuracy of the approximation to the posterior, but of course depends on the ML-II estimated hyper-parameters, which differ between the methods. Also in the example shown in Figure 3(b) it can be observed that the latent mean functions differ but their sign matches very accurately.

For the Crabs data set all methods show the same error rate but the information content of the predictive distributions differs dramatically. For some test cases the SVM predicts the wrong class with large certainty. Because the mapping of the unthresholded output of the SVM to the predictive probability is estimated from a left out set, the mapping can be poor if too few errors are observed on this.

8. Conclusions

Our experiments reveal serious differences between Laplace's method and EP when used in GPC models. The results corroborate the considerations about the two approximations based on the structure of the posterior given in Section 5. Although only a handful of data sets have been used in the study, we believe the conclusions to be well-founded and generally valid.

From the structural properties of the posterior we described why Laplace's method systematically underestimates the mean $\mathbf{m}$. The resulting approximate posterior GP over latent functions will have too small amplitude, although the sign of the mean function will be mostly correct. As an effect Laplace's method gives over-conservative predictive probabilities, and diminished information about the test labels. This effect has been shown empirically on several real world examples. Large resulting discrepancies in the actual posterior probabilities were found, even at the training locations, which renders the predictive class probabilities produced under this approximation grossly inaccurate. Note, the difference becomes less dramatic if we only consider the classification error rates obtained by thresholding p^* at $1/2$. For this particular task, we have seen the sign of the latent function tends to be correct (at least at the training locations). However, the performance on benchmark data sets also revealed the error rates obtained by Laplace's method to be inferior to EP results.

The EP approximation has shown to give results very close to MCMC both in terms of predictive distributions and marginal likelihood estimates. We have shown and explained why the marginal distributions of the posterior can be well approximated by Gaussians.

Further, the marginal likelihood values obtained by Laplace's method and EP differ systematically which will lead to different results of ML-II hyper-parameter estimation. The discrepancies are similar for different tasks. We were able to exemplify that the EP approximation of the marginal likelihood is accurate. To show this we described how AIS can be used to obtain unbiased estimates of the marginal likelihood for Gaussian process models.

In the experiments summarised in Table 1 we compared the predictive accuracy of GPC to support vector machines. While the SVMs show a tendency to give lower error rates, the information contained in predictive distributions seems comparable. Conceptually GPC comes with the advantage that the Bayesian model selection can be used to set hyper-parameters by ML-II estimation, while the parameters of an SVM usually have to be set by cross-validation (gradient based methods exist, see e.g. Chapelle et al. (2002)).

In summary, we found that EP is the method of choice for approximate inference in binary GPC models, when the computational cost of MCMC is prohibitive. Very good agreement is achieved for both predictive probabilities and marginal likelihood estimates. In contrast, the Laplace approximation is so inaccurate that we advise against its use, especially when predictive probabilities are to be taken seriously.

Acknowledgments

Both authors acknowledge support by the German Research Foundation (DFG) through grant RA 1030/1. We would like to thank Tobias Pfingsten, Jeremy Hill and Matthias Seeger for comments and discussions.

Appendix A. Implementation of Laplace's Approximation

In Sections 3 we described Laplace's method for approximate inference in the GPC model and sketched the corresponding computations in Algorithm 1. In this appendix we describe our implementation of the method in more detail. See also the appendices of Williams and Barber (1998).

Computing Laplace's approximation $\mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{A})$ for given $\boldsymbol{\theta}$ the main computational effort is involved in finding the maximum of the unnormalised log posterior $\ln \mathcal{Q}$ (eq. (8)). Our implementation uses Newton's method to find the mode. In each Newton step the vector $\mathbf{f}$ is updated according to

$$\begin{aligned}\mathbf{f}^{t+1} &= \mathbf{f}^t - (\nabla\nabla_{\mathbf{f}} \ln \mathcal{Q}(\mathbf{f}^t))^{-1} \nabla_{\mathbf{f}} \ln \mathcal{Q}(\mathbf{f}^t) && (37a)\\ &= (\mathbf{K}^{-1} + \mathbf{W})^{-1}(\mathbf{W}\mathbf{f}^t + \nabla_{\mathbf{f}} \ln \mathcal{L}(\mathbf{f}^t)) && (37b)\end{aligned}$$

until convergence of $\mathbf{f}$ to the mode $\mathbf{m}$. To ensure convergence the update is accepted if the value of the target function increases, otherwise the the step size is shortened until $\ln \mathcal{Q}(\mathbf{f}^{t+1}) > \ln \mathcal{Q}(\mathbf{f}^t)$.

Computationally Newtons's method is dominated by the repeated inversion of the Hessian. Since $\mathbf{K}$ can be poorly conditioned we use the identity

$$(\mathbf{K}^{-1} + \mathbf{W})^{-1} = \mathbf{K} - \mathbf{K}\mathbf{W}^{\frac{1}{2}}(\mathbf{I} + \mathbf{W}^{\frac{1}{2}}\mathbf{K}\mathbf{W}^{\frac{1}{2}})^{-1}\mathbf{W}^{\frac{1}{2}}\mathbf{K} \tag{38}$$

such that only the well conditioned, positive definite matrix $(\mathbf{I}+\mathbf{W}^{\frac{1}{2}}\mathbf{K}\mathbf{W}^{\frac{1}{2}})$ has to be inverted. In our implementation the inverse is computed from a Cholesky decomposition of this matrix. Note that $\mathbf{W}$ is a diagonal matrix with positive entries, so computing $\mathbf{W}^{\frac{1}{2}}$ is trivial.

Note that implementing the Newton updates (37) only requires the *product* of the inverse Hessian times the gradient which can be computed more efficiently using an iterative conjugate gradient method (Golub and Van Loan, 1989, ch. 10).

Having found the mode $\mathbf{m}$ the marginal likelihood approximation (15) and its derivatives can be computed. The approximate marginal likelihood takes the form

$$\begin{aligned}\ln q(\mathcal{D}|\boldsymbol{\theta}) &= \ln Q(\mathbf{m})+\frac{m}{2}\ln(2\pi)+\frac{1}{2}\ln|\mathbf{A}| &\quad (39a)\\ &= \ln \mathcal{L}(\mathbf{m})-\frac{1}{2}\mathbf{m}^\top\mathbf{K}^{-1}\mathbf{m}-\frac{1}{2}\ln|\mathbf{I}+\mathbf{K}\mathbf{W}|\;. &\quad (39b)\end{aligned}$$

To avoid the direct inversion of $\mathbf{K}$ in the second term of (39b) we use the recurrence relation (37b). Let $\mathbf{a}=\mathbf{K}^{-1}\mathbf{m}$ then by substituting (38) into (37b) we obtain:

$$\mathbf{a}=(\mathbf{I}-\mathbf{W}^{\frac{1}{2}}(\mathbf{I}+\mathbf{W}^{\frac{1}{2}}\mathbf{K}\mathbf{W}^{\frac{1}{2}})^{-1}\mathbf{W}^{\frac{1}{2}}\mathbf{K})(\mathbf{W}\mathbf{m}+\nabla_{\mathbf{f}}\ln\mathcal{L}(\mathbf{m})) \tag{40}$$

such that $\mathbf{m}^\top\mathbf{K}^{-1}\mathbf{m}=\mathbf{m}^\top\mathbf{a}$. The determinant in eq. (39b) can be rewritten

$$\ln|\mathbf{I}+\mathbf{K}\mathbf{W}|=\ln\left|\mathbf{I}+\mathbf{W}^{\frac{1}{2}}\mathbf{K}\mathbf{W}^{\frac{1}{2}}\right| \tag{41}$$

and computed from the Cholesky decomposition, that was used to calculate the inverse in eq. (38). Note that if $\mathbf{M}=\mathbf{L}\mathbf{L}^\top$ is a Cholesky decomposition then $\ln|\mathbf{M}|=2\sum\ln\mathbf{L}_{ii}$.

During ML-II estimation (36) of hyper-parameters the approximate log marginal likelihood (39) is maximised as a function of $\boldsymbol{\theta}$. Our implementation is based on a conjugate gradient optimisation routine such that we also need to compute the derivatives of (39b) with respect to the elements of $\boldsymbol{\theta}$.

The dependency of the approximate marginal likelihood on $\boldsymbol{\theta}$ is two-fold:

$$\frac{\partial\ln q(\mathcal{D}|\boldsymbol{\theta})}{\partial\theta_i}=\sum_{k,l}\frac{\partial\ln q(\mathcal{D}|\boldsymbol{\theta})}{\partial\mathbf{K}_{kl}}\frac{\partial\mathbf{K}_{kl}}{\partial\theta_i}+\frac{\partial\ln q(\mathcal{D}|\boldsymbol{\theta})}{\partial\mathbf{m}^\top}\frac{\partial\mathbf{m}}{\partial\theta_i} \tag{42}$$

there is a direct dependency via the terms involving $\mathbf{K}$ and an implicit dependency through the change in $\mathbf{m}$ (see also Williams and Barber (1998, Appendix B)).

The explicit derivative of eq. (39b) due to the direct dependency of the covariance matrix is

$$\sum_{k,l}\frac{\partial\ln q(\mathcal{D}|\boldsymbol{\theta})}{\partial\mathbf{K}_{kl}}\frac{\partial\mathbf{K}_{kl}}{\partial\theta_i}=\frac{1}{2}\mathbf{m}^\top\mathbf{K}^{-1}\frac{\partial\mathbf{K}}{\partial\theta_i}\mathbf{K}^{-1}\mathbf{m}-\frac{1}{2}\operatorname{tr}\left((\mathbf{I}+\mathbf{K}\mathbf{W})^{-1}\frac{\partial\mathbf{K}}{\partial\theta_i}\mathbf{W}\right) \tag{43}$$

where the first term is computed using $\mathbf{a}$ (40) and the inverse in the second term can be rewritten as

$$(\mathbf{I}+\mathbf{K}\mathbf{W})^{-1}=\mathbf{I}-(\mathbf{K}^{-1}+\mathbf{W})^{-1}\mathbf{W} \tag{44}$$

where the inverse (38) is already known.

The implicit derivative accounts for the dependency of eq. (39b) on $\boldsymbol{\theta}$ due to change in the mode $\mathbf{m}$. Differentiating eq. (39a) with respect to $\mathbf{m}$ reduces to $\partial\ln|\mathbf{A}|/\partial\mathbf{m}$ since $\mathbf{m}$ is the maximum of $\ln Q$ and therefore $\partial\ln Q/\partial\mathbf{m}$ vanishes.

$$\frac{\partial\ln q(\mathcal{D}|\boldsymbol{\theta})}{\partial\mathbf{m}^\top}\frac{\partial\mathbf{m}}{\partial\theta_i}\;=\;-\frac{1}{2}\frac{\partial|\mathbf{K}^{-1}+\mathbf{W}|}{\partial\mathbf{m}^\top}\frac{\partial\mathbf{m}}{\partial\theta_i}\;=\;-\frac{1}{2}(\mathbf{K}^{-1}+\mathbf{W})^{-1}\frac{\partial\mathbf{W}}{\partial\mathbf{m}^\top}\frac{\partial\mathbf{m}}{\partial\theta_i} \tag{45}$$

The dependency of $\mathbf{m}$ on $\boldsymbol{\theta}_i$ is obtained by differentiating (11a) at $\mathbf{m}$:

$$0 = \nabla_{\mathbf{f}} \ln \mathcal{L}(\mathbf{m}) - \mathbf{K}^{-1}\mathbf{m} \quad \Longrightarrow \quad \mathbf{m} = \mathbf{K}\nabla_{\mathbf{f}} \ln \mathcal{L}(\mathbf{m}) \tag{46}$$

so

$$\frac{\partial \mathbf{m}}{\partial \boldsymbol{\theta}_i} = \frac{\partial \mathbf{K}}{\partial \boldsymbol{\theta}_i}\nabla_{\mathbf{f}} \ln \mathcal{L}(\mathbf{m}) + \mathbf{K}\nabla\nabla_{\mathbf{f}} \ln \mathcal{L}(\mathbf{m})\frac{\partial \mathbf{m}}{\partial \boldsymbol{\theta}_i} = (\mathbf{I}+\mathbf{K}\mathbf{W})^{-1}\frac{\partial \mathbf{K}}{\partial \boldsymbol{\theta}_i}\nabla_{\mathbf{f}} \ln \mathcal{L}(\mathbf{m}) \tag{47}$$

and we have both terms necessary to compute the gradient (42).

To compute the predictive probability $p_* = p(y_* = 1|\mathbf{x}_*)$ for a test input $\mathbf{x}_*$ the predictive distribution (5) of the latent function value is $\mathcal{N}(f_*|\mu_*, \sigma_*^2)$ where

$$\begin{aligned} \mu_* &= \mathbf{k}_*^\top \mathbf{K}^{-1}\mathbf{m} = \mathbf{k}_*^\top \mathbf{a} && \text{(48a)} \\ \sigma_*^2 &= k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top \mathbf{W}^{\frac{1}{2}}(\mathbf{I} + \mathbf{W}^{\frac{1}{2}}\mathbf{K}\mathbf{W}^{\frac{1}{2}})^{-1}\mathbf{W}^{\frac{1}{2}}\mathbf{k}_* && \text{(48b)} \end{aligned}$$

and p_* can be computed from eq. (6).

Due to the Cholesky decomposition in (38) computing Laplace's approximation is $O(m^3)$. However, following the implementation we described in this section a Cholesky decomposition has to be computed once per Newton step and all other quantities can be computed from it in at most $O(m^2)$. The number of Newton steps necessary depends on the convergence criterion, the initialisation of $\mathbf{f}$ and the hyper-parameters $\boldsymbol{\theta}$.

Appendix B. Implementation of Expectation Propagation

In this appendix we describe details of our implementation of EP as described in Section 4 and summarised in Algorithm 2. See also the appendices of Seeger (2003).

In our implementation the site functions (17) are parameterised in terms of natural parameters σ_i^{-2} and $\sigma_i^{-2}\mu_i$. For given $\boldsymbol{\theta}$ the algorithm starts by initialising $\mathbf{A} = \mathbf{K}$ and $\sigma_i^{-2} = 0$ and $\sigma_i^{-2}\mu_i = 0$. The algorithm proceeds by updating the site parameters in random order. In each sweep every site function is updated following equations (23), (25), and (26). After each update of a site function the effect on $\mathbf{m}$ and $\mathbf{A}$ has to be computed according to (18). The change in $\mathbf{A}$ can be computed using a rank one update. Let δ be the change in σ_i^{-2} due to the update and $\mathbf{e}_i$ the vector whose ith entry is 1 and all other 0. The relation

$$(\mathbf{K}^{-1} + \boldsymbol{\Sigma}^{-1} + \delta\mathbf{e}_i\mathbf{e}_i^\top)^{-1} = \mathbf{A} - \mathbf{A}\mathbf{e}_i(\mathbf{A}_{ii} + \delta^{-1})^{-1}\mathbf{e}_i^\top\mathbf{A} \tag{49}$$

can be used to update $\mathbf{A}$. Each single update is $O(m^2)$ and repeated m times per sweep, such that the EP algorithm is $O(m^3)$ in time. Because of accumulating numerical errors, after a complete sweep over all site functions we recompute the matrix $\mathbf{A}$ from scratch. For numerical stability we rewrite

$$\mathbf{A} = (\mathbf{K}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1} = \mathbf{K} - \mathbf{K}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{I} + \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{K}\boldsymbol{\Sigma}^{-\frac{1}{2}})^{-1}\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{K} \tag{50}$$

and compute the inverse from the Cholesky decomposition of $(\mathbf{I} + \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{K}\boldsymbol{\Sigma}^{-\frac{1}{2}})$.

After convergence the approximate log marginal likelihood (27) can be computed and its partial derivatives with respect to the hyper-parameters:

$$\frac{\partial \ln q(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i} = -\frac{1}{2}\operatorname{tr}\left(\frac{\partial \mathbf{K}}{\partial \boldsymbol{\theta}_i}\Big((\mathbf{K}+\boldsymbol{\Sigma})^{-1} - (\mathbf{K}+\boldsymbol{\Sigma})^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^\top(\mathbf{K}+\boldsymbol{\Sigma})^{-1}\Big)\right) . \tag{51}$$

which do not depend on the Z_i (Seeger, 2005).

The inverse of $\mathbf{K}+\mathbf{\Sigma}$ can be computed from the inverse in eq. (50):

$$(\mathbf{K}+\mathbf{\Sigma})^{-1} = \mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{I}+\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{K}\mathbf{\Sigma}^{-\frac{1}{2}})^{-1}\mathbf{\Sigma}^{-\frac{1}{2}} \ . \tag{52}$$

For computing the log marginal likelihood (27) also the determinant $|\mathbf{K}+\mathbf{\Sigma}|$ has to be computed. By rewriting

$$\ln|\mathbf{K}+\mathbf{\Sigma}| = \ln(|\mathbf{\Sigma}||\mathbf{I}+\mathbf{\Sigma}^{-1}\mathbf{K}|) = \ln|\mathbf{\Sigma}|+\ln|\mathbf{I}+\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{K}\mathbf{\Sigma}^{-\frac{1}{2}}| \tag{53}$$

we obtain an expression in which the first term is a determinant of a diagonal matrix and the second term can be computed from the Cholesky decomposition that was used to compute the inverse in eq. (50).

To compute the predictive probability $p_* = p(y_* = 1|\mathbf{x}_*)$ for a test input $\mathbf{x}_*$ the predictive distribution (5) of the latent function value is $\mathcal{N}(f_*|\mu_*,\sigma_*^2)$ where

$$\mu_* = \mathbf{k}_*^\top(\mathbf{K}+\mathbf{\Sigma})^{-1}\boldsymbol{\mu} \tag{54a}$$

$$\sigma_*^2 = k(\mathbf{x}_*,\mathbf{x}_*) - \mathbf{k}_*^\top(\mathbf{K}+\mathbf{\Sigma})^{-1}\mathbf{k}_* \tag{54b}$$

and p_* can be computed from eq. (6).

The EP algorithm is of computational complexity $O(m^3)$ due to the computations for updating $\mathbf{A}$. However, per sweep the computation of $\mathbf{A}$ (50) and the m rank one updates sum to more computational effort compared to Laplace's method.

Using a covariance function of the form (33) for some data sets we observed numerical problems during ML-II hyper-parameter estimation because the optimisation algorithm asked to evaluate the marginal likelihood for extremely large signal variances σ^2. The problem stems from the property that for large values of σ^2 the marginal likelihood becomes insensitive to changes in σ^2. At this point it is recommended to take another look at Figure 1(b). Intuitively, for large signal variances the prior becomes more spread, such that the likelihood becomes more and more similar to a hard truncation. The marginal likelihood equals the probability mass of the prior in the orthant that is left after truncation. But the probability mass in any of the orthants remains constant if only the signal variance is changed for fixed correlation structure. This argument is based on the assumption that the likelihood implements a hard truncation, which is only an approximation, but this approximation becomes better the larger σ^2 is. Note that this insensitivity of the marginal likelihood with respect to changes in the signal variance can already be observed in the upper parts of of the marginal likelihood plots for EP in Figures 4 and 5. A possible solution to this problem is to limit $\sigma^2 < 10^5$, say, since we wouldn't typically expect any new interesting behaviour beyond this.

References

P. Abrahamsen. A review of Gaussian random fields and correlation functions. Technical Report 917, Norwegian Computing Center, Oslo, 1997.

C.-C. Chang and C.-J. Lin. *LIBSVM: A library for Support Vector Machines*, 2001. `http://www.csie.ntu.edu.tw/∼cjlin/libsvm`.

O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1):131–159, 2002.

W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041, 2005.

L. Csató and M. Opper. Sparse online Gaussian processes. *Neural Computation*, 14(2):641–669, 2002.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, 1998.

M. N. Gibbs and D. J. C. MacKay. Variational Gaussian process classifiers. *IEEE Transactions on Neural Networks*, 11(6):1458–1464, 2000.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. John Hopkins University Press, Baltimore, second edition, 1989.

S. Hettich, C. L. Blake, and C. J. Merz. UCI repository of machine learning databases, 1998. `http://www.ics.uci.edu/∼mlearn/MLRepository.html`.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90 (430):773–795, 1995.

N. Lawrence, M. Seeger, and R. Herbrich. Fast sparse Gaussian process methods: The informative vector machine. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 609–616, Cambridge, MA, 2003. The MIT Press.

J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, New York, 2001.

D. J. C. MacKay. Comparison of approximate methods for handling hyperparameters. *Neural Computation*, 11(5):1035–1068, 1999.

D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, UK, 2003.

T. P. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Department of Electrical Engineering and Computer Science, MIT, 2001.

R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993.

R. M. Neal. Regression and classification using Gaussian process priors. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 6*, pages 475–501. Oxford University Press, 1998.

R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 2001.

A. O'Hagan. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society, Series B*, 40(1):1–42, 1978.

M. Opper and O. Winther. Gaussian processes for classification: Mean-field algorithms. *Neural Computation*, 12(11):2655–2684, 2000.

J. C. Platt. Probabilities for SV machines. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schurmans, editors, *Advances in Large Margin Classifiers*, pages 61–73. The MIT Press, Cambridge, MA, 2000.

C. E. Rasmussen and C. K. I Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA, 2006. *In press*.

B. D. Ripley. *Pattern Recognition and Neural Newtorks*. Cambridge University Press, Cambridge, UK, 1996.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. The MIT Press, Cambridge, MA, 2002.

M. Seeger. PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002.

M. Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations*. PhD thesis, University of Edinburgh, 2003.

M. Seeger. Expectation propagation for exponential families, 2005. Note obtainable from `http://www.kyb.tuebingen.mpg.de/∼seeger`.

C. K. I. Williams and D. Barber. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.