Journal of Machine Learning Research 14 (2013) 75-109 Submitted 2/12; Revised 9/12; Published 1/13

Nested Expectation Propagation for Gaussian Process Classification with a Multinomial Probit Likelihood

Jaakko Riihimäki JAAKKO.RIIHIMAKI@AALTO.FI
Pasi Jylänki PASI.JYLANKI@AALTO.FI
Aki Vehtari AKI.VEHTARI@AALTO.FI
Department of Biomedical Engineering and Computational Science
Aalto University School of Science
P.O. Box 12200
FI-00076 Aalto
Finland

Editor: Neil Lawrence

Abstract

This paper considers probabilistic multinomial probit classification using Gaussian process (GP) priors. Challenges with multiclass GP classification are the integration over the non-Gaussian posterior distribution, and the increase of the number of unknown latent variables as the number of target classes grows. Expectation propagation (EP) has proven to be a very accurate method for approximate inference but the existing EP approaches for the multinomial probit GP classification rely on numerical quadratures, or independence assumptions between the latent values associated with different classes, to facilitate the computations. In this paper we propose a novel nested EP approach which does not require numerical quadratures, and approximates accurately all between-class posterior dependencies of the latent values, but still scales linearly in the number of classes. The predictive accuracy of the nested EP approach is compared to Laplace, variational Bayes, and Markov chain Monte Carlo (MCMC) approximations with various benchmark data sets. In the experiments nested EP was the most consistent method compared to MCMC sampling, but in terms of classification accuracy the differences between all the methods were small from a practical point of view.

Keywords: Gaussian process, multiclass classification, multinomial probit, approximate inference, expectation propagation

1. Introduction

Gaussian process (GP) priors enable flexible model specification for Bayesian classification. In multiclass GP classification, the posterior inference is challenging because each target class increases the number of unknown latent variables by the number of observations n. Typically, independent GP priors are set for the latent values for each class and this is assumed throughout this paper. Since all the latent values depend on each other through the likelihood, they become a posteriori dependent, which can rapidly lead to computationally unfavorable scaling as the number of classes c grows. A cubic scaling in c is prohibitive, and from a practical point of view, a desired complexity is $O(cn^3)$ which is typical for the most existing approaches for multiclass GP classification. The cubic scaling with respect to the number of data points is standard for full GP priors, and to reduce this n^3 complexity, sparse approximations can be used, but these are not considered in this paper.

As an additional challenge, the posterior inference is analytically intractable because the likelihood term related to each observation is non-Gaussian and depends on multiple latent values (one for each class).

A Markov chain Monte Carlo (MCMC) approach for multiclass GP classification with a softmax likelihood (also called a multinomial logistic likelihood) was described by Neal (1998). Sampling of the latent values with the softmax model is challenging because the dimensionality is often high and standard methods such as the Metropolis-Hastings and Hamiltonian Monte Carlo algorithms require tuning of the step size parameters. Later Girolami and Rogers (2006) proposed an alternative approach based on the multinomial probit likelihood which can be augmented with auxiliary latent variables. This enables a convenient Gibbs sampling framework in which the latent function values are conditionally independent between classes and normally distributed. If the hyperparameters are sampled, one MCMC iteration scales as $O(cn^3)$ which can become computationally expensive for large n because thousands of posterior draws may be required to obtain uncorrelated posterior samples, and strong dependency between the hyperparameters and latent values can cause slow mixing of the chains.

To speed up the inference, Williams and Barber (1998) used the Laplace approximation (LA) to approximate the non-Gaussian posterior distribution of the latent function values with a tractable Gaussian distribution. Conveniently the LA approximation with the softmax likelihood leads to an efficient representation of the approximative posterior covariance scaling as $O((c+1)n^3)$, which facilitates considerably the predictions and gradient-based type-II maximum a posteriori (MAP) estimation of the covariance function hyperparameters. Later Girolami and Rogers (2006) proposed a factorized variational Bayes approximation (VB) for the augmented multinomial probit model. Assuming the latent values and the auxiliary variables a posteriori independent, a computationally efficient posterior approximation scheme is obtained. If the latent processes related to each class share the same fixed hyperparameters, VB requires only one $O(n^3)$ matrix inversion per iteration step compared to LA in which $c+1$ such inverses are required in each iteration. Recently, Chai (2012) proposed an alternative variational bounding approximation for the multinomial logistic likelihood, which results in $O(c^3n^3)$ base scaling. To reduce the computational complexity, sparse approximations were determined by active inducing set selection.

Expectation propagation (EP) is the method of choice in binary GP classification where it has been found very accurate with a reasonable computational cost (Kuss and Rasmussen, 2005; Nickisch and Rasmussen, 2008). Two types of EP approximations have been considered for the multiclass setting; the first assuming the latent values from different classes a posteriori independent (IEP) and the second assuming them fully correlated (Seeger and Jordan, 2004; Seeger et al., 2006; Girolami and Zhong, 2007). Incorporating the full posterior couplings requires evaluating the non-analytical moments of c-dimensional tilted distributions which Girolami and Zhong (2007) approximated with Laplace's method resulting in an approximation scheme known as Laplace propagation described by Smola et al. (2004). Earlier Seeger and Jordan (2004) proposed an alternative approach where the full posterior dependencies were approximated by enforcing a similar structure for the posterior covariance as in LA using the softmax likelihood. This enables a posterior representation scaling as $O((c+1)n^3)$ but the proposed implementation requires a c-dimensional numerical quadrature and double-loop optimization to obtain a restricted-form site covariance approximation for each likelihood term (Seeger and Jordan, 2004).[1] To reduce the computational demand of EP,

1. Seeger and Jordan (2004) achieve also a linear scaling in the number of training points but we omit sparse approaches here.

factorized posterior approximations were proposed by both Seeger et al. (2006) and Girolami and Zhong (2007). Both approaches omit the between-class posterior dependencies of the latent values which results in a posterior representation scaling as $O(cn^3)$. The approaches rely on numerical two-dimensional quadratures for evaluating the moments of the tilted distributions with the main difference being that Seeger et al. (2006) used fewer two-dimensional quadratures for computational speed-up.

A different EP approach for the multiclass setting was described by Kim and Ghahramani (2006) who adopted the threshold function as an observation model. Each threshold likelihood term factorizes into $c-1$ terms dependent on only two latent values. This property can be used to transform the inference onto an equivalent non-redundant model which includes $n(c-1)$ unknown latent values with a Gaussian prior and a likelihood consisting of $n(c-1)$ factorizing terms. It follows that standard EP methodology for binary GP classification (Rasmussen and Williams, 2006) can be applied for posterior inference but a straightforward implementation results in a posterior representation scaling as $O((c-1)^3n^3)$ and means to improve the scaling are not discussed by Kim and Ghahramani (2006). Contrary to the usual EP approach of maximizing the marginal likelihood approximation, Kim and Ghahramani (2006) determined the hyperparameters by maximizing a lower bound on the log marginal likelihood in a similar way as is done in the expectation maximization (EM) algorithm. Recently Hernández-Lobato et al. (2011) introduced a robust generalization of the multiclass GP classifier with a threshold likelihood by incorporating n additional binary indicator variables for modeling possible labeling errors. Efficiently scaling EP inference is obtained by making the IEP assumption.

In this paper, we focus on the multinomial probit model and describe an efficient quadrature-free nested EP approach for multiclass GP classification that scales as $O((c+1)n^3)$. The proposed EP method takes into account all the posterior covariances between the latent variables, and the posterior computations scale as efficiently as in the LA approximation. We validate the proposed nested EP algorithm with several experiments. First, we compare the nested EP algorithm to various quadrature-based EP methods with respect to the approximate marginal distributions of the latent values and class probabilities with fixed hyperparameter values, and show that nested EP achieves similar accuracy in a computationally efficient manner. Using the nested EP algorithm, we study visually the utility of the full EP approximation over IEP, and compare their convergence properties. Second, we compare nested EP and IEP to other Gaussian approximations (LA and VB). We visualize the accuracy of the approximate marginal distributions with respect to MCMC, illustrate the suitability of the respective marginal likelihood approximations for type-II MAP estimation of the covariance function hyperparameters, and discuss their computational complexities. Finally, we compare the predictive performance of all these methods with estimation of the hyperparameters using several real-world data sets. Since LA is known to be fast, we also test whether the predictive probability estimates of LA can be further improved using Laplace's method as described by Tierney and Kadane (1986).

2. Gaussian Processes for Multiclass Classification

We consider a classification problem consisting of d-dimensional input vectors $\mathbf{x}_i$ associated with target class labels $y_i \in \{1,\ldots,c\}$, where $c > 2$, for $i = 1,\ldots,n$. All the class labels are collected in the $n \times 1$ target vector $\mathbf{y}$, and all the covariate vectors are collected in the matrix $X = [\mathbf{x}_1,\ldots,\mathbf{x}_n]^T$ of size $n \times d$. Given the latent function values $\mathbf{f}_i = \left[f_i^1, f_i^2, \ldots, f_i^c\right]^T = \mathbf{f}(\mathbf{x}_i)$ at the observed input

locations $\mathbf{x}_i$, the observations y_i are assumed independently and identically distributed as defined by the observation model $p(y_i|\mathbf{f}_i)$. The latent vectors related to all the observations are denoted by $\mathbf{f} = \left[f_1^1,\ldots,f_n^1,f_1^2,\ldots,f_n^2,\ldots,f_1^c,\ldots,f_n^c\right]^T$.

Our goal is to predict the class membership for a new input vector $\mathbf{x}_*$ given the observed data $\mathcal{D} = \{X,\mathbf{y}\}$, which is why we need to make some assumptions on the unknown function $f(\mathbf{x})$. We set a priori independent zero-mean Gaussian process priors on the latent values related to each class, which is the usual assumption in multiclass GP classification (see, for example, Williams and Barber, 1998; Seeger and Jordan, 2004; Rasmussen and Williams, 2006; Girolami and Zhong, 2007). This specification results in the following zero-mean Gaussian prior for $\mathbf{f}$:

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

where K is a $cn \times cn$ block-diagonal covariance matrix with matrices $K^1,K^2,\ldots,K^c$ (each of size $n \times n$) on its diagonal. Element $K^k_{i,j}$ of the k'th covariance matrix defines the prior covariance between the function values f_i^k and f_j^k, which is defined by the covariance function $\kappa(\mathbf{x}_i,\mathbf{x}_j)$, that is, $K^k_{i,j} = \kappa(\mathbf{x}_i,\mathbf{x}_j) = \mathrm{Cov}\left[f_i^k,f_j^k\right]$ for the latent values related to class k. A common choice for the covariance function is the squared exponential

$$\kappa_{\mathrm{se}}(\mathbf{x}_i,\mathbf{x}_j|\theta) = \sigma^2 \exp\left(-\frac{1}{2}\sum_{k=1}^{d} l_k^{-2}(x_{i,k}-x_{j,k})^2\right),$$

where $x_{i,k}$ is the k'th component of $\mathbf{x}_i$, and $\theta = \{\sigma^2,l_1,\ldots,l_d\}$ collects the hyperparameters governing the smoothness properties of latent functions. The magnitude parameter σ^2 controls the overall variance of the unknown function values, and the lengthscale parameters $l_1,\ldots,l_d$ control the smoothness of the latent function by defining how fast the correlation decreases in each input dimension. The framework allows separate covariance functions or hyperparameters for different classes but throughout this work, for simplicity, we use the squared exponential covariance function with the same θ for all classes.

In this paper, we consider two different observation models: the softmax model

$$p(y_i|\mathbf{f}_i) = \frac{\exp(f_i^{y_i})}{\sum_{j=1}^{c}\exp(f_i^j)}, \tag{1}$$

and the multinomial probit model

$$p(y_i|\mathbf{f}_i) = \mathrm{E}_{p(u_i)}\left\{\prod_{j=1,j\neq y_i}^{c}\Phi(u_i+f_i^{y_i}-f_i^j)\right\}, \tag{2}$$

where the auxiliary variable u_i is distributed as $p(u_i) = \mathcal{N}(u_i|0,1)$, and $\Phi(x)$ denotes the cumulative density function of the standard normal distribution. The softmax and multinomial probit models are multiclass generalizations of the logistic and the probit models respectively.

By applying Bayes' theorem, the conditional posterior distribution of the latent values can be written as

$$p(\mathbf{f}|\mathcal{D},\theta) = \frac{1}{Z}p(\mathbf{f}|X,\theta)\prod_{i=1}^{n}p(y_i|\mathbf{f}_i), \tag{3}$$

where $Z = p(\mathbf{y}|X,\theta) = \int p(\mathbf{f}|X,\theta)\prod_{i=1}^{n} p(y_i|\mathbf{f}_i)d\mathbf{f}$ is known as the marginal likelihood of θ. Both observation models result in an analytically intractable posterior distribution and therefore approximate methods are needed for integration over the latent variables. Different approximate methods are more suitable for a particular likelihood function because of the convenience of implementation: the softmax is preferable for LA because of the efficient structure and computability of the partial derivatives (Williams and Barber, 1998), while the multinomial probit is preferable for VB, EP and Gibbs sampling because of the convenient auxiliary variable representations (Girolami and Rogers, 2006; Girolami and Zhong, 2007).

3. Approximate Inference Using Expectation Propagation

In this section, we first give a general description of EP for multiclass GP classification and review some existing approaches. Then we present a novel nested EP approach for the multinomial probit model.

3.1 Expectation Propagation for Multiclass GP Classification

Expectation propagation is an iterative algorithm for approximating integrals over functions that factor into simple terms (Minka, 2001b). Using EP the posterior distribution (3) can be approximated with

$$q_{\mathrm{EP}}(\mathbf{f}|\mathcal{D},\theta) = \frac{1}{Z_{\mathrm{EP}}} p(\mathbf{f}|X,\theta)\prod_{i=1}^{n}\tilde{t}_i(\mathbf{f}_i|\tilde{Z}_i,\tilde{\boldsymbol{\mu}}_i,\tilde{\Sigma}_i), \tag{4}$$

where $\tilde{t}_i(\mathbf{f}_i|\tilde{Z}_i,\tilde{\boldsymbol{\mu}}_i,\tilde{\Sigma}_i) = \tilde{Z}_i\mathcal{N}(\mathbf{f}_i|\tilde{\boldsymbol{\mu}}_i,\tilde{\Sigma}_i)$ are local likelihood term approximations parameterized with scalar normalization terms $\tilde{Z}_i$, $c\times 1$ site location vectors $\tilde{\boldsymbol{\mu}}_i$, and $c\times c$ site covariances $\tilde{\Sigma}_i$. In the algorithm, first the site approximations are initialized, and then each site is updated in turns. The update for the i'th site is done by first removing the site term from the marginal posterior which gives the cavity distribution

$$q_{-i}(\mathbf{f}_i) = \mathcal{N}(\mathbf{f}_i|\boldsymbol{\mu}_{-i},\Sigma_{-i}) \propto q(\mathbf{f}_i|\mathcal{D},\theta)\tilde{t}(\mathbf{f}_i)^{-1}.$$

The cavity distribution is then combined with the exact i'th likelihood term $p(y_i|\mathbf{f}_i)$ to form the non-Gaussian tilted distribution

$$\hat{p}(\mathbf{f}_i) = \hat{Z}_i^{-1} q_{-i}(\mathbf{f}_i)p(y_i|\mathbf{f}_i), \tag{5}$$

which is assumed to encompass more information about the true marginal distribution. Next a Gaussian approximation $\hat{q}(\mathbf{f}_i)$ is determined for $\hat{p}(\mathbf{f}_i)$ by minimizing the Kullback-Leibler (KL) divergence $\mathrm{KL}(\hat{p}(\mathbf{f}_i)||\hat{q}(\mathbf{f}_i))$, which for a Gaussian $\hat{q}(\mathbf{f}_i)$ is equivalent to matching the first and second moments of $\hat{q}(\mathbf{f}_i)$ with the corresponding moments of $\hat{p}(\mathbf{f}_i)$. Finally, the parameters of the i'th site are updated so that the mean and covariance of $q(\mathbf{f}_i)$ are consistent with $\hat{q}(\mathbf{f}_i)$. After updating the site parameters, the posterior distribution (4) is updated. This can be done either in a sequential way, where immediately after each site update the posterior is refreshed using a rank-c update, or in a parallel way (see, for example, van Gerven et al., 2009), where the posterior is refreshed only after all the site approximations have been updated once. This procedure is repeated until convergence, that is, until all the marginal distributions $q(\mathbf{f}_i)$ are consistent with $\hat{p}(\mathbf{f}_i)$.

In binary GP classification, determining the moments of the tilted distribution requires solving only one-dimensional integrals, and assuming the probit likelihood function, these univariate integrals can be computed efficiently without numerical quadratures. In the multiclass setting, the problem is how to evaluate the multi-dimensional integrals which are required to determine the moments of the tilted distributions (5). Girolami and Zhong (2007) approximated these moments using the Laplace approximation which results in an algorithm called Laplace propagation (Smola et al., 2004). The problem with the LA approach is that the mean is replaced with the mode of the distribution and the covariance with the inverse Hessian of the log density at the mode. Because of the skewness of the tilted distribution caused by the likelihood function, the LA method can lead to inaccurate mean and covariance estimates in which case the resulting posterior approximation does not correspond to the full EP solution. Seeger and Jordan (2004) estimated the tilted moments using multi-dimensional quadratures, but this becomes computationally demanding when c increases, and to achieve a posterior representation scaling linearly in c, they do an additional optimization step to obtain a constrained site precision matrix for each likelihood term approximation.

Computations can be facilitated by using the IEP approximation where explicit between-class posterior dependencies are omitted. This simplification enables posterior computations scaling linearly in c. The existing approaches for the multinomial probit rely on multiple numerical quadratures for each site update; the implementation of Girolami and Zhong (2007) requires a total of $2c+1$ two-dimensional numerical quadratures for each likelihood term, whereas Seeger et al. (2006) described an alternative approach where only two two-dimensional and $2c-1$ one-dimensional quadratures are needed. Later, we will demonstrate that compared to the full EP approximation, IEP underestimates the uncertainty on the latent values and in practice it may require more iterations than full EP for convergence especially if the hyperparameter setting results in strong between-class posterior couplings.

3.2 Efficiently Scaling Quadrature-Free Implementation

In this section, we present a novel nested EP approach for multinomial probit classification that does not require numerical quadratures or sampling for estimation of the tilted moments and predictive probabilities. The method also leads simultaneously to low-rank site approximations which retain all posterior couplings but results in linear computational scaling with respect to the number of target classes c. Using the proposed nested EP approach a quadrature-free IEP approximation can also be formed with similar computational complexity as the full EP approximation.

3.2.1 QUADRATURE-FREE NESTED EXPECTATION PROPAGATION

Here we use the multinomial probit as the likelihood function because its product form consisting of cumulative Gaussian factors is computationally more suitable for EP than the sum of exponential terms in the softmax likelihood. Given the mean $\boldsymbol{\mu}_{-i}$ and the covariance Σ_{-i} of the cavity distribution, we need to determine the normalization factor $\hat{Z}_i$, mean vector $\hat{\boldsymbol{\mu}}_i$, and covariance matrix $\hat{\Sigma}_i$ of the tilted distribution

$$\hat{p}(\mathbf{f}_i) = \hat{Z}_i^{-1} \mathcal{N}(\mathbf{f}_i|\boldsymbol{\mu}_{-i}, \Sigma_{-i}) \int \mathcal{N}(u_i|0,1) \left(\prod_{j=1, j\neq y_i}^{c} \Phi(u_i + f_i^{y_i} - f_i^j) \right) du_i, \tag{6}$$

which requires solving non-analytical $(c+1)$-dimensional integrals over $\mathbf{f}_i$ and u_i. Instead of quadrature methods (Seeger and Jordan, 2004; Seeger et al., 2006; Girolami and Zhong, 2007), we use EP

to approximate these integrals. At first, this approach may seem computationally very demanding since individual EP approximations are required for each of the n sites. However, it turns out that these inner EP approximations can be updated incrementally between the outer EP loops. This scheme also leads naturally to an efficiently scaling representation for the site precisions $\tilde{\Sigma}_i^{-1}$.

To form a computationally efficient EP algorithm for approximating the tilted moments, it is helpful to consider the joint distribution of $\mathbf{f}_i$ and the auxiliary variable u_i arising from (6). Defining $\mathbf{w}_i = [\mathbf{f}_i^T, u_i]^T$ and removing the marginalization over u_i results in the following augmented tilted distribution:

$$\hat{p}(\mathbf{w}_i) = \hat{Z}_i^{-1} \mathcal{N}(\mathbf{w}_i | \boldsymbol{\mu}_{\mathrm{w}_i}, \Sigma_{\mathrm{w}_i}) \prod_{j=1, j\neq y_i}^{c} \Phi(\mathbf{w}_i^T \tilde{\mathbf{b}}_{i,j}), \tag{7}$$

where $\boldsymbol{\mu}_{\mathrm{w}_i} = [\boldsymbol{\mu}_{-i}^T, 0]^T$ and Σ_{w_i} is a block-diagonal matrix formed from Σ_{-i} and 1. Denoting the j'th unit vector of the c-dimensional standard basis by $\mathbf{e}_j$, the auxiliary vectors $\tilde{\mathbf{b}}_{i,j}$ can be written as $\tilde{\mathbf{b}}_{i,j} = [(\mathbf{e}_{y_i} - \mathbf{e}_j)^T, 1]^T$. The normalization term $\hat{Z}_i$ is the same for $\hat{p}(\mathbf{f}_i)$ and $\hat{p}(\mathbf{w}_i)$, and it is defined by $\hat{Z}_i = \int \mathcal{N}(\mathbf{w}_i | \boldsymbol{\mu}_{\mathrm{w}_i}, \Sigma_{\mathrm{w}_i}) \prod_{j\neq y_i} \Phi(\mathbf{w}_i^T \tilde{\mathbf{b}}_{i,j}) d\mathbf{w}_i$. The other quantities of interest, $\hat{\boldsymbol{\mu}}_i$ and $\hat{\Sigma}_i$, are equal to the marginal mean and covariance of the first c components of $\mathbf{w}_i$ with respect to $\hat{p}(\mathbf{w}_i)$.

The augmented distribution (7) is of similar functional form as the posterior distribution resulting from a linear binary classifier with a multivariate Gaussian prior on the weights $\mathbf{w}_i$ and a probit likelihood function. Therefore, the moments of (7) can be approximated with EP similarly as in linear classification (see, for example, Qi et al., 2004) or by applying the general EP formulation for latent Gaussian models described by Cseke and Heskes (2011, Appendix C). For clarity, we have summarized a computationally efficient implementation of the algorithm in Appendix A. The augmented tilted distribution (7) is approximated with

$$\hat{q}(\mathbf{w}_i) = Z_{\hat{q}_i}^{-1} \mathcal{N}(\mathbf{w}_i | \boldsymbol{\mu}_{\mathrm{w}_i}, \Sigma_{\mathrm{w}_i}) \prod_{j=1, j\neq y_i}^{c} \tilde{Z}_{\hat{q}_{i,j}} \mathcal{N}(\mathbf{w}_i^T \tilde{\mathbf{b}}_{i,j} | \tilde{\alpha}_{i,j}^{-1} \tilde{\beta}_{i,j}, \tilde{\alpha}_{i,j}^{-1}) d\mathbf{w}_i, \tag{8}$$

where the cumulative Gaussian functions are approximated with scaled Gaussian site functions and the normalization constant $\hat{Z}_i$ is approximated with $Z_{\hat{q}_i}$. From now on the site parameters of $\hat{q}(\mathbf{w}_i)$ in their natural exponential form are denoted by $\tilde{\boldsymbol{\alpha}}_i = [\tilde{\alpha}_{i,j}]_{j\neq y_i}^T$ and $\tilde{\boldsymbol{\beta}}_i = [\tilde{\beta}_{i,j}]_{j\neq y_i}^T$.

Note that the probit terms in Equation (7) depend on the unknown latents $\mathbf{f}_i$ only through the linear transformation $\mathbf{g}_i = \tilde{B}_i^T \mathbf{w}_i$, where $\tilde{B}_i = [\tilde{\mathbf{b}}_{i,j}]_{j\neq y_i}$, that is $g_i^j = f_i^{y_i} - f_i^j + u_i$. This relation implies that the likelihood of $\mathbf{f}_i$ increases as the latent value associated with the correct class y_i increases compared to the latents associated with the other classes. Integration over the auxiliary variable u_i results from the conic truncation of the latent variable representation of the multinomial probit model (see, for example, Girolami and Rogers, 2006). This relationship between $\mathbf{w}_i$ and $\mathbf{g}_i$ has two important computational consequences. First, the fully-coupled nested EP solution can be computed by propagating scalar moments of g_i^j which requires solving only one-dimensional integrals because each probit factor in the augmented tilted distribution depends only on the scalar g_i^j (see Appendix A and references therein). Second, it can be shown that the exact mean and covariance of $\mathbf{w}_i \sim \hat{p}(\mathbf{w}_i)$ can be solved from the respective moments of $\mathbf{g}_i$ whose distribution is obtained by $\mathbf{g}_i = \tilde{B}_i^T \mathbf{w}_i$ on Equation (7). Because the dimension of $\mathbf{g}_i$ is $c-1$ we can form computationally cheaper quadrature-based estimates of the tilted moments as described in Section 3.3. We will also use the approximate marginal moments of $\mathbf{g}_i$ to visualize differences in the predictive accuracy of EP and IEP approximations in Section 5.2.

3.2.2 EFFICIENTLY SCALING REPRESENTATION

In this section we show that the approximation (8) leads to matrix computations scaling as $O((c+1)n^3)$ in the evaluation of the moments of the approximate posterior (4). The idea is to show that the site precision matrix $\tilde{\Sigma}_i^{-1}$ resulting from the EP update step with $\hat{\Sigma}_i$ derived from (8) has a similar structure with the Hessian matrix of $\log p(y_i|\mathbf{f}_i)$ in the Laplace approximation (Williams and Barber, 1998; Seeger and Jordan, 2004; Rasmussen and Williams, 2006).

The approximate marginal covariance of $\mathbf{f}_i$ derived from (8) is given by

$$\hat{\Sigma}_i = H^T \left(\Sigma_{\mathrm{w}_i}^{-1} + \tilde{B}_i \tilde{T}_i \tilde{B}_i^T\right)^{-1} H, \tag{9}$$

where the matrix $\tilde{T}_i = \mathrm{diag}(\tilde{\boldsymbol{\alpha}}_i)$ is diagonal,[2] and $H^T = \begin{bmatrix} I_c & \mathbf{0} \end{bmatrix}$ picks up the desired components of $\mathbf{w}_i$, that is, $\mathbf{f}_i = H^T \mathbf{w}_i$. Using the matrix inversion lemma and denoting $B_i = H^T \tilde{B}_i = \mathbf{e}_{y_i} \mathbf{1}^T - E_{-y_i}$, where $E_{-y_i} = [\mathbf{e}_j]_{j \neq y_i}$ and $\mathbf{1}$ is a $(c-1) \times 1$ vector of ones, we can write the tilted covariance as

$$\begin{aligned} \hat{\Sigma}_i &= \Sigma_{-i} - \Sigma_{-i} B_i (\tilde{T}_i^{-1} + \mathbf{1}\mathbf{1}^T + B_i^T \Sigma_{-i} B_i)^{-1} B_i^T \Sigma_{-i} \\ &= (\Sigma_{-i}^{-1} + B_i (\tilde{T}_i^{-1} + \mathbf{1}\mathbf{1}^T)^{-1} B_i^T)^{-1}. \end{aligned} \tag{10}$$

Because in the moment matching step of the EP algorithm the site precision matrix is updated as $\tilde{\Sigma}_i^{-1} = \hat{\Sigma}_i^{-1} - \Sigma_{-i}^{-1}$, we can write

$$\tilde{\Sigma}_i^{-1} = B_i (\tilde{T}_i^{-1} + \mathbf{1}\mathbf{1}^T)^{-1} B_i^T = B_i (\tilde{T}_i - \tilde{\boldsymbol{\alpha}}_i (1 + \mathbf{1}^T \tilde{\boldsymbol{\alpha}}_i)^{-1} \tilde{\boldsymbol{\alpha}}_i^T) B_i^T. \tag{11}$$

Since B_i is a $c \times (c-1)$ matrix, we see that $\tilde{\Sigma}_i^{-1}$ is of rank $c-1$ and therefore a straightforward implementation based on (11) would result into $O((c-1)^3 n^3)$ scaling in the posterior update. However, a more efficient representation can be obtained by simplifying (11) further. Writing $B_i = -A_i E_{-y_i}$, where $A_i = [I_c - \mathbf{e}_{y_i} \mathbf{1}_c^T]$ and $\mathbf{1}_c$ is a $c \times 1$ vector of ones, we get

$$\tilde{\Sigma}_i^{-1} = A_i \left(E_{-y_i} \tilde{T}_i E_{-y_i}^T - \boldsymbol{\pi}_i (\mathbf{1}_c^T \boldsymbol{\pi}_i)^{-1} \boldsymbol{\pi}_i^T\right) A_i^T,$$

where we have defined $\boldsymbol{\pi}_i = E_{-y_i} \tilde{\boldsymbol{\alpha}}_i + \mathbf{e}_{y_i}$ and used $B_i \tilde{\boldsymbol{\alpha}}_i = -A_i \boldsymbol{\pi}_i$. Since $A_i \mathbf{e}_{y_i} = \mathbf{0}$ we can add $\mathbf{e}_{y_i} \mathbf{e}_{y_i}^T$ to the first term inside the brackets to obtain

$$\tilde{\Sigma}_i^{-1} = A_i \Pi_i A_i^T = \Pi_i, \quad \text{where} \quad \Pi_i = \mathrm{diag}(\boldsymbol{\pi}_i) - (\mathbf{1}_c^T \boldsymbol{\pi}_i)^{-1} \boldsymbol{\pi}_i \boldsymbol{\pi}_i^T. \tag{12}$$

The second equality can be explained as follows. Matrix Π_i is of similar form with the precision contribution of the i'th likelihood term, $W_i = -\nabla^2_{\mathbf{f}_i} \log p(y_i|\mathbf{f}_i)$, in the Laplace algorithm (Williams and Barber, 1998), and it has one eigenvector, $\mathbf{1}_c$, with zero eigenvalue: $\Pi_i \mathbf{1}_c = \mathbf{0}$. It follows that $A_i \Pi_i = (I_c - \mathbf{e}_{y_i} \mathbf{1}_c^T) \Pi_i = \Pi_i - \mathbf{e}_{y_i} \mathbf{0}^T = \Pi_i$ and therefore $\tilde{\Sigma}_i^{-1} = \Pi_i$. Matrix Π_i is also precisely of the same form as the a priori constrained site precision block that Seeger and Jordan (2004) determined by double-loop optimization of $\mathrm{KL}(\hat{q}(\mathbf{f}_i)||q(\mathbf{f}_i))$.

In a similar fashion, we can determine a simple formula for the natural location parameter $\tilde{\boldsymbol{\nu}}_i = \tilde{\Sigma}_i^{-1} \tilde{\boldsymbol{\mu}}_i$ as a function of $\tilde{\boldsymbol{\alpha}}_i$ and $\tilde{\boldsymbol{\beta}}_i$. The marginal mean of $\mathbf{f}_i$ with respect to $\hat{q}(\mathbf{w}_i)$ is given by

$$\hat{\boldsymbol{\mu}}_i = H_i^T \left(\Sigma_{\mathrm{w}_i}^{-1} + \tilde{B}_i \tilde{T}_i \tilde{B}_i^T\right)^{-1} \left(\Sigma_{\mathrm{w}_i}^{-1} \boldsymbol{\mu}_{\mathrm{w}_i} + \tilde{B}_i \tilde{\boldsymbol{\beta}}_i\right), \tag{13}$$

2. We use the following notation: $\mathrm{diag}(\mathbf{a})$ with a vector argument is a square matrix with $\mathbf{a}$ on the main diagonal, and $\mathrm{diag}(A)$ with a matrix argument is a column vector containing the diagonal elements of the matrix A.

which we can write using the matrix inversion lemma as

$$\hat{\boldsymbol{\mu}}_i = \hat{\Sigma}_i \Sigma_{-i}^{-1} \boldsymbol{\mu}_{-i} + \Sigma_{-i} B_i (\tilde{T}_i^{-1} + \mathbf{1}\mathbf{1}^T + B_i^T \Sigma_{-i} B_i)^{-1} \tilde{T}_i^{-1} \tilde{\boldsymbol{\beta}}_i. \tag{14}$$

Using the update formula $\tilde{\boldsymbol{\nu}}_i = \hat{\Sigma}_i^{-1} \hat{\boldsymbol{\mu}}_i - \Sigma_{-i}^{-1} \boldsymbol{\mu}_{-i}$ resulting from the EP moment matching step and simplifying further with the matrix inversion lemma, the site location $\tilde{\boldsymbol{\nu}}_i$ can be written as

$$\tilde{\boldsymbol{\nu}}_i = B_i \left(\tilde{\boldsymbol{\beta}}_i - \tilde{\boldsymbol{\alpha}}_i a_i \right) = a_i \boldsymbol{\pi}_i - E_{-y_i} \tilde{\boldsymbol{\beta}}_i, \tag{15}$$

where $a_i = (\mathbf{1}^T \tilde{\boldsymbol{\beta}}_i)/(\mathbf{1}_c^T \boldsymbol{\pi}_i)$. The site precision vector $\tilde{\boldsymbol{\nu}}_i$ is orthogonal with $\mathbf{1}_c$, that is, $\mathbf{1}_c^T \tilde{\boldsymbol{\nu}}_i = 0$, which is congruent with (12). Note that with results (12) and (15), the mean and covariance of the approximate posterior (4) can be evaluated using only $\tilde{\boldsymbol{\alpha}}_i$ and $\tilde{\boldsymbol{\beta}}_i$. It follows that the posterior (predictive) means and covariances as well as the marginal likelihood can be evaluated with similar computational complexity as with the Laplace approximation (Williams and Barber, 1998; Rasmussen and Williams, 2006). For clarity the main components are summarized in Appendix B. The IEP approximation in our implementation is formed by matching the i'th marginal covariance with $\mathrm{diag}(\mathrm{diag}(\hat{\Sigma}_i))$, and the corresponding mean with $\hat{\boldsymbol{\mu}}_i$.

3.2.3 EFFICIENT IMPLEMENTATION

Approximating the tilted moments using inner EP for each site may appear too slow for larger problems because typically several iterations are required to achieve convergence. However, the number of inner-loop iterations can be reduced by storing the site parameters $\tilde{\boldsymbol{\alpha}}_i$ and $\tilde{\boldsymbol{\beta}}_i$ after each inner EP run and continuing from the previous values in the next run. This framework where the inner site parameters $\tilde{\boldsymbol{\alpha}}_i$ and $\tilde{\boldsymbol{\beta}}_i$ are updated iteratively instead of $\tilde{\boldsymbol{\mu}}_i$ and $\tilde{\Sigma}_i$, can be justified by writing the posterior approximation (4) using the approximative site terms from (8):

$$q(\mathbf{f}|\mathcal{D},\theta) \propto p(\mathbf{f}|X,\theta) \prod_{i=1}^{n} \int \mathcal{N}(u_i|0,1) \prod_{j=1, j\neq y_i}^{c} \tilde{Z}_{\hat{q}_i,j} \mathcal{N}(u_i + f_i^{y_i} - f_i^j | \tilde{\alpha}_{i,j}^{-1} \tilde{\beta}_{i,j}, \tilde{\alpha}_{i,j}^{-1}) du_i. \tag{16}$$

Calculating the Gaussian integral over u_i leads to the same results for $\tilde{\boldsymbol{\mu}}_i$ and $\tilde{\Sigma}_i$ as derived earlier (Equations 12 and 15). Apart from the integration over the auxiliary variables u_i, Equation (16) resembles an EP approximation where $n(c-1)$ probit terms of the form $\Phi(u_i + f_i^{y_i} - f_i^j)$ are approximated with Gaussian site functions. In accordance with the standard EP framework we form the cavity distribution $q_{-i}(\mathbf{f}_i)$ by removing $c-1$ sites from (16) and subsequently refine $\tilde{\boldsymbol{\alpha}}_i$ and $\tilde{\boldsymbol{\beta}}_i$ using the mean and covariance of the tilted distribution (6). If we alternatively expand only the i'th site approximation with respect to u_i and write the corresponding marginal approximation as

$$q(\mathbf{f}_i|\mathcal{D},\theta) \propto q_{-i}(\mathbf{f}_i) \int \mathcal{N}(u_i|0,1) \prod_{j=1, j\neq y_i}^{c} \tilde{Z}_{\hat{q}_i,j} \mathcal{N}(u_i + f_i^{y_i} - f_i^j | \tilde{\alpha}_{i,j}^{-1} \tilde{\beta}_{i,j}, \tilde{\alpha}_{i,j}^{-1}) du_i, \tag{17}$$

we can consider updating only one of the approximative terms in (17) at a time. This is equivalent to starting the inner EP iterations with the values of $\tilde{\boldsymbol{\alpha}}_i$ and $\tilde{\boldsymbol{\beta}}_i$ from the previous outer-loop iteration instead of a zero initialization which is customary to standard EP implementations. In our experiments, only one inner-loop iteration per site was found sufficient for convergence with comparable number of outer-loop iterations, which results in significant computational savings in the tilted moment evaluations.

The previous interpretation of the algorithm is also useful for defining damping (Minka and Lafferty, 2002), which is commonly used to improve the numerical stability and convergence of EP. In damping the site parameters in their natural exponential forms are updated to a convex combination of the old and new values. Damping cannot be directly applied on the site precision matrix $\Pi_i = \tilde{\Sigma}_i^{-1}$ because the constrained form of the site precision (12) is lost. Instead we damp the updates on $\tilde{\boldsymbol{\alpha}}_i$ and $\tilde{\boldsymbol{\beta}}_i$ which preserves the desired structure. This can be justified with the same arguments as in the previous paragraph where we considered updating only one of the approximative terms in (17) at a time. Convergence of the nested EP algorithm with full posterior couplings using this scheme is illustrated with different damping levels in Section 5.4.

3.3 Quadrature-Based Full EP Implementation

A challenge in forming the fully-coupled EP approximation using numerical quadratures is how to obtain a site precision structure, which results in efficiently scaling posterior computations. Seeger and Jordan (2004) used c-dimensional Gauss-Hermite rules and determined a similar site precision matrix as in Equation (12) by optimizing $\mathrm{KL}(\hat{p}_i(\mathbf{f}_i)||q(\mathbf{f}_i))$. In this section, we use the ideas from Section 3.2 to form a simpler fully-coupled EP algorithm that uses similar approximate site precision structures determined directly using $(c-1)$-dimensional quadratures instead of separate optimizations.

We use the previously defined transformation $\mathbf{g}_i = \tilde{B}_i^T \mathbf{w}_i$, where $\mathbf{w}_i \sim \hat{p}(\mathbf{w}_i)$, and denote the tilted mean vector and covariance matrix of $\mathbf{w}_i$ with $\hat{\boldsymbol{\mu}}_{\mathrm{w}_i}$ and $\hat{\Sigma}_{\mathrm{w}_i}$. Analogously, we denote the corresponding moments of $\mathbf{g}_i$ resulting from the transformation with $\hat{\boldsymbol{\mu}}_{\mathrm{g}_i}$ and $\hat{\Sigma}_{\mathrm{g}_i}$. Making the transformation on (7) and differentiating twice with respect to $\boldsymbol{\mu}_{\mathrm{w}_i}$, it can be shown that the following relation holds between the exact covariance matrices of the random vectors $\mathbf{w}_i$ and $\mathbf{g}_i$:

$$\hat{\Sigma}_{\mathrm{g}_i} = \tilde{B}_i^T \hat{\Sigma}_{\mathrm{w}_i} \tilde{B}_i = \tilde{B}_i^T (\Sigma_{\mathrm{w}_i}^{-1} + \tilde{B}_i \Lambda_i \tilde{B}_i^T)^{-1} \tilde{B}_i, \tag{18}$$

where Σ_{w_i} is the cavity covariance of $\mathbf{w}_i$. Solving Λ_i from (18) gives

$$\Lambda_i = \hat{\Sigma}_{\mathrm{g}_i}^{-1} - \Sigma_{\mathrm{g}_i}^{-1}, \tag{19}$$

where $\Sigma_{\mathrm{g}_i} = B_i^T \Sigma_{-i} B_i + \mathbf{1}\mathbf{1}^T$, and $\hat{\Sigma}_{\mathrm{g}_i}$ can be estimated with a $(c-1)$-dimensional quadrature rule. The marginal tilted covariance of $\mathbf{f}_i$ can be computed from $\hat{\Sigma}_{\mathrm{w}_i}$ similarly as in Equations (9) and (10), and the corresponding site precision matrix $\tilde{\Sigma}_i^{-1}$ can be computed as in Equation (11) with Λ_i now in place of $\tilde{T}_i$. This gives the following site precision structure

$$\tilde{\Sigma}_i^{-1} = B_i (\Lambda_i^{-1} + \mathbf{1}\mathbf{1}^T)^{-1} B_i^T,$$

which depends only on Λ_i. The form of the site precision is similar to nested EP, except that now Λ_i is a full matrix, which would result in the unfavorable $O((c-1)^3 n^3)$ posterior scaling. Therefore, we approximate Λ_i with its diagonal to get the same structure as in Equation (12), where now $\tilde{\Lambda}_i = \mathrm{diag}(\mathrm{diag}(\Lambda_i))$ is used instead of $\tilde{T}_i$. This results in posterior computations scaling linearly in c similarly as with the full nested EP approach.

To estimate the site location parameter $\tilde{\boldsymbol{\nu}}_i$ using quadratures, we proceed in the same way as for the site precision. Making the transformation on (7) and differentiating once with respect to $\boldsymbol{\mu}_{\mathrm{w}_i}$, it can be shown that the tilted means of $\mathbf{w}_i$ and $\mathbf{g}_i$ are related according to

$$\hat{\boldsymbol{\mu}}_{\mathrm{g}_i} = \tilde{B}_i^T \hat{\boldsymbol{\mu}}_{\mathrm{w}_i} = \tilde{B}_i^T (\Sigma_{\mathrm{w}_i}^{-1} + \tilde{B}_i \Lambda_i \tilde{B}_i^T)^{-1} (\Sigma_{\mathrm{w}_i}^{-1} \boldsymbol{\mu}_{\mathrm{w}_i} + \tilde{B}_i \boldsymbol{\xi}_i), \tag{20}$$

where $\hat{\boldsymbol{\mu}}_{\mathrm{w}_i}$ has similar form as in Equation (13). The vector $\boldsymbol{\xi}_i$ corresponds to $\tilde{\boldsymbol{\beta}}_i$ in nested EP, and we can solve it from (20), which results in

$$\boldsymbol{\xi}_i = \hat{\Sigma}_{\mathrm{g}_i}^{-1} \hat{\boldsymbol{\mu}}_{\mathrm{g}_i} - \Sigma_{\mathrm{g}_i}^{-1} \boldsymbol{\mu}_{\mathrm{g}_i}, \tag{21}$$

where $\boldsymbol{\mu}_{\mathrm{g}_i} = B_i^T \boldsymbol{\mu}_{-i}$, and $\hat{\boldsymbol{\mu}}_{\mathrm{g}_i}$ can be estimated using a $(c-1)$-dimensional quadrature. If Λ_i is approximated with its diagonal, we have to substitute $\hat{\Sigma}_{\mathrm{g}_i}^{-1} = \tilde{\Lambda}_i + \Sigma_{\mathrm{g}_i}^{-1}$ in Equation (21), which results from the diagonal approximation of Λ_i made in Equation (19). In the same way as in Equations (13)-(15), we get the following expression for the site location

$$\tilde{\boldsymbol{\nu}}_i = B_i(\boldsymbol{\xi}_i - \tilde{\Lambda}_i \mathbf{1}(1 + \mathbf{1}^T \tilde{\Lambda}_i \mathbf{1})^{-1} \mathbf{1}^T \boldsymbol{\xi}_i),$$

which depends only on $\boldsymbol{\xi}_i$ and $\tilde{\Lambda}_i$. This site location structure is similar to nested EP (15) with $\boldsymbol{\xi}_i$ in place of $\tilde{\boldsymbol{\beta}}_i$. Using these results, a quadrature-based full EP algorithm can be implemented in the same way as the outer-loop of nested EP. Later in Section 5.1, we validate this approximate $(c-1)$-dimensional quadrature approach by comparing the tilted moments to those of a more expensive straightforward $(c+1)$-dimensional full quadrature solution.

4. Other Approximations for Bayesian Inference

In this section we discuss all the other approximations considered in this paper for multiclass GP classification. First we give a short description of the LA method. Then we show how it can be improved upon by computing corrections to the marginal predictive densities using Laplace's method as described by Tierney and Kadane (1986). Finally, we briefly summarize the MCMC and VB approximations.

4.1 Laplace Approximation

In the Laplace approximation a second order Taylor expansion of $\log p(\mathbf{f}|\mathcal{D},\theta)$ is made around the posterior mode $\hat{\mathbf{f}}$ which can be determined using Newton's method as described by Williams and Barber (1998) and Rasmussen and Williams (2006). This results in the posterior approximation

$$q_{\mathrm{LA}}(\mathbf{f}|\mathcal{D},\theta) = \mathcal{N}\left(\mathbf{f}|\hat{\mathbf{f}}, (K^{-1} + W)^{-1}\right),$$

where $W = -\nabla^2_{\mathbf{f}} \log p(\mathbf{y}|\mathbf{f})|_{\mathbf{f}=\hat{\mathbf{f}}}$ and in which $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^n p(y_i|\mathbf{f}_i)$. With the softmax likelihood (1), the submatrix of W related to each observation will have a similar structure with Π_i in (12), which enables efficient posterior computations that scale linearly in c as already discussed in the case of EP.

4.1.1 IMPROVING MARGINAL POSTERIOR DISTRIBUTIONS

In Gaussian process classification, the LA and EP methods can be used to efficiently form a multivariate Gaussian approximation for the posterior distribution of the latent values. Recently, motivated by the earlier ideas of Tierney and Kadane (1986), two methods have been proposed for improving the marginal posterior distributions in latent Gaussian models; one based on subsequent use of Laplace's method (Rue et al., 2009), and one based on EP (Cseke and Heskes, 2011). Because in classification the focus is not on the predictive distributions of the latent values but on the predictive probabilities related to a test input $\mathbf{x}_*$, applying these methods would require additional

numerical integration over the improved posterior approximation of the corresponding latent value $\mathbf{f}_* = \mathbf{f}(\mathbf{x}_*)$. In the multiclass setting integration over a multi-dimensional space is required which becomes computationally demanding to perform, for example, in a grid if c is large. To avoid this integration, we test computing the corrections directly for the predictive class probabilities following another approach presented by Tierney and Kadane (1986). A related idea for approximating the predictive distribution of linear model coefficients directly with a deterministic approximation has been discussed by Snelson and Ghahramani (2005).

The posterior mean of a smooth and positive function $h(f)$ is given by

$$\mathrm{E}[h(f)] = \frac{\int h(f)p(y|f)p(f)df}{\int p(y|f)p(f)df}, \tag{22}$$

where $p(y|f)$ is the likelihood function and $p(f)$ is the prior distribution. Tierney and Kadane (1986) proposed to approximate both integrals in (22) separately with Laplace's method. This approach can be readily applied for approximating the posterior predictive probabilities $p(y_*|\mathbf{x}_*)$ of class memberships $y_* \in \{1,...,c\}$ which are given by

$$p(y_*|\mathbf{x}_*, \mathcal{D}) = \frac{1}{Z}\iint p(y_*|\mathbf{f}_*)p(\mathbf{f}_*|\mathbf{f}, \mathbf{x}_*, X)p(\mathbf{f}|X)p(\mathbf{y}|\mathbf{f})d\mathbf{f}d\mathbf{f}_*, \tag{23}$$

where $Z = \iint p(\mathbf{f}_*|\mathbf{f}, \mathbf{x}_*, X)p(\mathbf{f}|X)p(\mathbf{y}|\mathbf{f})d\mathbf{f}d\mathbf{f}_* = \int p(\mathbf{f}|X)p(\mathbf{y}|\mathbf{f})d\mathbf{f}$ is the marginal likelihood. With a fixed class label y_* the integrals can be approximated by a straightforward application of either LA or EP, which is already done for the marginal likelihood Z in the standard approximations. The LA method can be used for smooth and positive functions such as the softmax whereas EP is applicable for a wider range of models.

The integral on the right side of (23) is equivalent to the marginal likelihood resulting from a classification problem with one additional training point y_*. To compute the predictive probabilities for all classes, we evaluate this extended marginal likelihood consisting of $n+1$ observations with y_* fixed to one of the c possible class labels at a time. This is computationally demanding because several marginal likelihood evaluations are required for each test input. Additional modifications, for example, initializing the latent values to their predictive mean implied by standard LA, could be done to speed up the computations. Since further approximations can only be expected to reduce the accuracy of the predictions, we do not consider them in this paper, and focus only on the naive implementation due to its ease of use. Since LA is known to be fast, we test the goodness of the improved predictive probability estimates using only LA, and refer to the method as LA-TKP as an extension to the naming used by Cseke and Heskes (2011).

4.2 Markov Chain Monte Carlo

Because MCMC estimates become exact in the limit of infinite sample size, we use MCMC as a gold standard for measuring the performance of the other approximations. Depending on the likelihood, we use two different sampling techniques; scaled Metropolis-Hastings sampling for the softmax function, and Gibbs sampling for the multinomial probit function.

4.2.1 SCALED METROPOLIS-HASTINGS SAMPLING FOR SOFTMAX

To obtain samples from the posterior with the softmax likelihood, the following two steps are alternated. Given the hyperparameter values, the latent values are drawn from the conditional posterior

$p(\mathbf{f}|\mathcal{D},\theta)$ using the scaled Metropolis-Hastings sampling (Neal, 1998). Then, the hyperparameters can be drawn from the conditional posterior $p(\theta|\mathbf{f},\mathcal{D})$, for example, using the Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal, 1996).

4.2.2 GIBBS SAMPLING FOR MULTINOMIAL PROBIT

Girolami and Rogers (2006) described how to draw samples from the joint posterior using the Gibbs sampler. The multinomial probit likelihood (2) can be written in the form

$$p(y_i|\mathbf{f}_i) = \int \psi(v_i^{y_i} > v_i^k \forall k \neq y_i) \prod_{j=1}^{c} \mathcal{N}(v_i^j|f_i^j,1) d\mathbf{v}_i, \tag{24}$$

where $\mathbf{v}_i = [v_i^1,...,v_i^c]^T$ is a vector of auxiliary variables, and ψ is the indicator function whose value is one if the argument is true and zero otherwise. Gibbs sampling can then be employed by drawing samples alternately for all i from $p(\mathbf{v}_i|\mathbf{f}_i,y_i)$ which is a conic truncation of the multivariate Gaussian distribution, and from $p(\mathbf{f}|\mathbf{v},\theta)$ which is a multivariate Gaussian distribution. Given $\mathbf{v}$ and $\mathbf{f}$ the hyperparameters can be drawn, for example, using HMC.

4.3 Factorized Variational Approximation

A computationally convenient variational Bayesian approximation for $p(\mathbf{f}|\mathcal{D},\theta)$ can be formed by employing the auxiliary variable representation (24) of the multinomial probit likelihood. As shown by Girolami and Rogers (2006), assuming $\mathbf{f}$ a posteriori independent of $\mathbf{v}$ (which contains all $\mathbf{v}_i$) leads to the following approximation

$$q_{\mathrm{VB}}(\mathbf{v},\mathbf{f}|\mathcal{D},\theta) = q(\mathbf{v})q(\mathbf{f}) = \prod_{i=1}^{n} q(\mathbf{v}_i) \prod_{j=1}^{c} q(\mathbf{f}^j),$$

where the latent values associated with the j'th class, $\mathbf{f}^j$, are independent. The posterior approximation $q(\mathbf{f}^j)$ will be a multivariate Gaussian distribution, and $q(\mathbf{v}_i)$ a conic truncation of the multivariate Gaussian distribution (Girolami and Rogers, 2006). Given the hyperparameters, the parameters of $q(\mathbf{v})$ and $q(\mathbf{f})$ can be determined iteratively by maximizing a variational lower bound on the marginal likelihood. Each iteration step requires determining the expectations of $\mathbf{v}_i$ with respect to $q(\mathbf{v}_i)$ which can be obtained by either one-dimensional numerical quadratures or sampling methods. In our implementation, the hyperparameters θ are determined by maximizing the variational lower bound with fixed $q(\mathbf{v})$ and $q(\mathbf{f})$ similarly as in the maximization step of the EM algorithm.

5. Experiments

This section is divided into five parts. In Section 5.1 we compare nested EP to quadrature-based EP in cost and quality. In Section 5.2, we illustrate the differences of the nested EP and IEP approximations in a simple synthetic classification problem. In Section 5.3, we compare visually the quality of the approximate marginal distributions of $\mathbf{f}$, the marginal likelihood approximations and the predictive class probabilities between EP, IEP, VB and LA using a three-class real-world data set. In Section 5.4, we discuss the computational complexities of the different approximate methods, and in Section 5.5, we evaluate them in terms of predictive performance with estimation of the hyperparameters using several benchmark data sets.

5.1 Comparing Nested EP to Numerical Quadrature

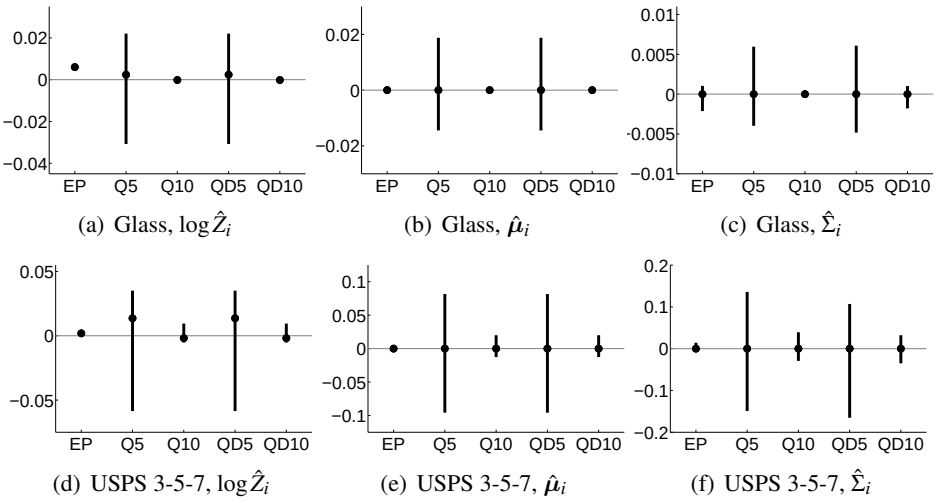
In this section we first validate the accuracy of the inner EP approximation and the full quadrature method described in Section 3.3 for estimation of the tilted moments. Then we compare the accuracy and numerical cost of the nested EP approximation to several quadrature-based EP implementations. In the comparisons, we use two different types of classification data: the Glass data set from the UCI Machine Learning Repository (Frank and Asuncion, 2010), and the USPS 3 vs. 5 vs. 7 data set from the US Postal Service (USPS) database (Hull, 1994). The USPS 3 vs. 5 vs. 7 data set is defined as a three class sub-problem from the USPS repartitioned handwritten digits data by considering classification of 3's vs. 5's vs. 7's.[3] The Glass data has six ($c = 6$) target classes but only a small number of observations ($n = 214$), whereas the USPS 3 vs. 5 vs. 7 data has only three ($c = 3$) target classes but a larger number of training points ($n = 1157$). See also Table 3.

In the first experiment, we examine the tilted moments after two parallel EP outer-loop iterations when the parameters of the cavity distributions are clearly different from their initialized values for all site terms. We fixed the hyperparameters of the squared exponential covariance function to $\log(\sigma^2) = 1$ and $\log(l) = 1$, where the small value of the magnitude parameter leads to a close-to-Gaussian posterior as will be discussed more in Section 5.3. The main reason for not using more difficult hyperparameter values (larger magnitude) in this experiment, was that we had stability problems in the actual EP algorithm using quadratures. Stability could be improved by increasing the number of quadrature points, but this became computationally too expensive with a larger number of classes.

As a baseline approximation, we compute the normalization factors $\hat{Z}_i$, the mean vectors $\hat{\boldsymbol{\mu}}_i$, and the covariance matrices $\hat{\Sigma}_i$ of the tilted distribution (6) for all $i = 1,\ldots,n$ using a $(c+1)$-dimensional Gauss-Hermite product rule with ten quadrature points in each dimension. We call this quadrature method QF10. This provides us a reference by which inner EP and the following four $(c-1)$-dimensional Gauss-Hermite quadrature methods (see Section 3.3 for further details) are assessed: Q5 using five and Q10 using ten quadrature points in each dimension with the full matrix Λ_i, and QD5 using five and QD10 using ten quadrature points in each dimension with the diagonal approximation $\tilde{\Lambda}_i$. Note that implementing an EP algorithm using QF10, Q5 or Q10, we would lose the linear posterior scaling in c. Figure 1 shows the pairwise differences of $\log \hat{Z}_i$, and all the entries of $\hat{\boldsymbol{\mu}}_i$ and $\hat{\Sigma}_i$ with respect to QF10. The mean values and the 95% intervals of the differences are illustrated. The normalization, mean and covariance are well calibrated for all the quadrature methods. Inner EP matches the mean and covariance accurately, but there is a small bias in the normalization term, probably due to the skewed tilted distributions. Variations of the pairwise differences are small with inner EP and the $(c-1)$-dimensional quadratures as long as there are enough quadrature points. Because QD10 agrees well with QF10, from now on, we use it to compute the tilted moments in the full quadrature solution, and refer to this algorithm as QEP.

In the second experiment, we compare the nested and quadrature EP algorithms in accuracy and computational cost. We use Gibbs sampling as a reference method by which nested EP and IEP, QEP, and quadrature-based IEP (QIEP) are measured. Both nested EP algorithms are implemented incrementally, so that only one inner-loop iteration per site is done at each outer-loop iteration step, which results in computational savings (see Section 3.2.3). For QIEP we use the implementation proposed by Seeger et al. (2006) with ten quadrature points for integration over the latent value from each class. We compare the absolute differences of class probabilities and latent means and

3. We use the same data partition as discussed by Rasmussen and Williams (2006).



(a) Glass, $\log \hat{Z}_i$ (b) Glass, $\hat{\boldsymbol{\mu}}_i$ (c) Glass, $\hat{\Sigma}_i$

(d) USPS 3-5-7, $\log \hat{Z}_i$ (e) USPS 3-5-7, $\hat{\boldsymbol{\mu}}_i$ (f) USPS 3-5-7, $\hat{\Sigma}_i$

Figure 1: A comparison of tilted moments after two parallel EP outer-loop iterations using the Glass and USPS 3 vs. 5 vs. 7 data sets. Using a $(c+1)$-dimensional Gauss-Hermite product rule with ten quadrature points in each dimension (QF10) as a baseline result, we compare inner EP and the following $(c-1)$-dimensional Gauss-Hermite quadrature methods: five- and ten-dimensional product rules with full Λ_i (Q5 and Q10) and diagonal $\tilde{\Lambda}_i$ (QD5 and QD10). The mean values and the 95% intervals of the pairwise differences of $\hat{Z}_i$, and all the entries of $\hat{\boldsymbol{\mu}}_i$ and $\hat{\Sigma}_i$ with respect to QF10 are shown. See the text for further details.

variances using the Glass and USPS 3 vs. 5 vs. 7 data sets with the same fixed hyperparameters as earlier. We split the Glass data set randomly into training and test parts, and use the predefined training and test parts for the USPS 3 vs. 5 vs. 7 data set. Table 1 reports the mean and maximum values of the element-wise differences with respect to Gibbs sampling after 30 outer-loop iterations of EP. Table 1 shows also the relative CPU times for training. From the table it can be seen that the differences in accuracy between the methods are small. For the Glass data the fully-coupled EP algorithms give slightly more accurate estimates for the mean and variances of the latents than the IEP algorithms do, but the class probabilities are in practice the same across all the methods. The main observation with the Glass data is that the CPU times of EP, IEP and QIEP are similar for practical purposes, but QEP is clearly slower due to the unfavorable scaling in c. We acknowledge that the performance differences in the relative CPU time are approximate and depend much on the implementation, but to reduce these effects the same outer-loop implementation was used for both nested and quadrature EP with the same fixed number of iterations. It is also worth to notice that QEP and EP have practically the same CPU times with the USPS 3 vs. 5 vs. 7 data where the number of target classes c is only three, but both of them are slower than the IEP algorithms due to larger n and the additional $n \times n$ inversion needed in the posterior update with the fully-coupled solutions (see also Section 5.4). We conclude that because the accuracy of nested and quadrature EP are similar, and we experienced some stability problems with the quadrature solutions in more difficult hyperparameter settings (larger values for the magnitude hyperparameter) and a

Glass		Training				Test			
		EP	QEP	IEP	QIEP	EP	QEP	IEP	QIEP
Latent means	mean	0.048	0.047	0.058	0.067	0.043	0.043	0.055	0.058
	max	0.181	0.186	0.190	0.217	0.176	0.177	0.193	0.181
Latent variances	mean	0.084	0.085	0.325	0.325	0.097	0.097	0.314	0.314
	max	0.553	0.560	0.656	0.667	0.571	0.572	0.643	0.655
Class probabilities	mean	0.004	0.004	0.004	0.004	0.003	0.004	0.004	0.004
	max	0.025	0.026	0.021	0.021	0.021	0.022	0.024	0.024
Relative CPU time		1.245	132.300	1.000	1.175				

USPS 3 vs. 5 vs. 7		Training				Test			
		EP	QEP	IEP	QIEP	EP	QEP	IEP	QIEP
Latent means	mean	0.065	0.065	0.065	0.065	0.006	0.006	0.006	0.006
	max	0.266	0.288	0.266	0.266	0.203	0.202	0.199	0.199
Latent variances	mean	0.167	0.167	0.167	0.167	0.215	0.215	0.216	0.215
	max	0.724	0.724	0.723	0.731	1.021	1.021	1.021	1.021
Class probabilities	mean	0.008	0.008	0.013	0.013	0.001	0.001	0.001	0.001
	max	0.040	0.043	0.059	0.058	0.022	0.024	0.035	0.035
Relative CPU time		2.542	2.602	1.000	1.001				

Table 1: A comparison of nested and quadrature-based EP in terms of accuracy and cost using the Glass and USPS 3 vs. 5 vs. 7 data sets. The table shows the element-wise mean and maximum absolute differences of the latent means and variances and the class probabilities for EP, QEP, IEP, and QIEP with respect to Gibbs sampling. See the text for further details.

small number of quadrature points (for example less than ten), from now on, we use nested EP and IEP implementations due to their stability and good computational scaling.

5.2 Illustrative Comparison of EP and IEP with Synthetic Data

In this section, we study the properties of the proposed nested EP and IEP approximations in a synthetic three-class classification problem with scalar inputs shown in Figure 2. The symbols x (class 1), + (class 2), and o (class 3) indicate the positions of $n = 15$ training inputs generated from three normal distributions with means -1, 2, and 3, and standard deviations 1, 0.5, and 0.5, respectively. The left-most observations from class 1 can be better separated from the others but the observations from classes 2 and 3 overlap more in the input space. We fixed the hyperparameters of the squared exponential covariance function at the corresponding MCMC means: $\log(\sigma^2) = 4.62$ and $\log(l) = 0.26$.

Figure 2(a) shows the predictive probabilities of all tree classes estimated with EP, IEP and MCMC as a function of the input x. At the class boundaries, the methods give similar predictions but elsewhere MCMC is the most confident while IEP seems more conservative. The performance of EP is somewhere between MCMC and IEP, although the differences are small. To explain why

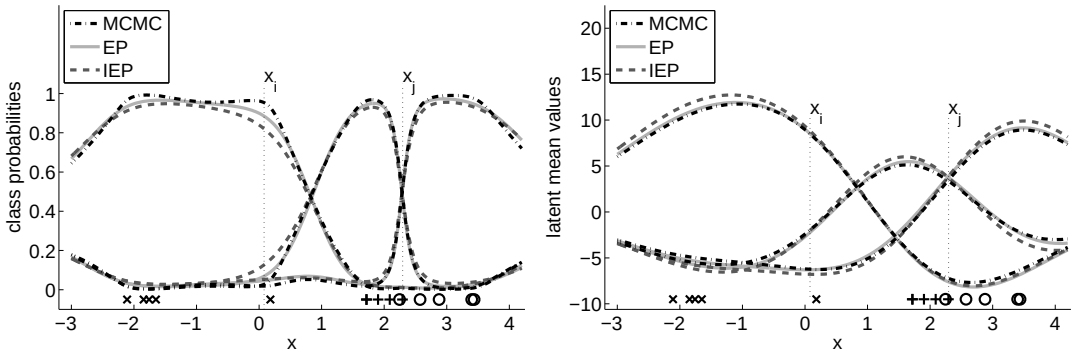


Figure 2: A synthetic one-dimensional example of a three class classification problem, where the MCMC, EP and IEP approximations are compared. The symbols x (class 1), + (class 2), and o (class 3) in the bottom of the plots indicate the positions of $n = 15$ observations. Plot (a) shows the predicted class probabilities, and (b) shows the predicted latent mean values for all three classes. The symbols x_i and x_j indicate two example positions, where the marginal distributions between the latent function values are illustrated in Figures 3 and 4. See the text for explanation.

the predictions differ, we look at the quality of the approximations made for the underlying $\mathbf{f}$. Figure 2(b) shows the approximated latent mean values which are similar at all input locations.

To illustrate the approximate posterior uncertainties of $\mathbf{f}$, we visualize two exemplary marginal distributions at locations x_i and x_j marked in Figure 2. The MCMC samples of f_i^1 and f_i^2 (the latents associated with classes 1 and 2 related to x_i) together with a smoothed density estimate are shown in Figure 3(a). The marginal distribution is non-Gaussian, and the latent values are more likely larger for class 1 than for class 2 indicating a larger predictive probability for class 1. The corresponding EP and IEP approximations are shown in Figures 3(b)-(c). EP captures the shape of the true marginal posterior distribution better than IEP. To illustrate the effect of these differences on the predictive probabilities, we show the unnormalized tilted distributions

$$\hat{p}(\mathbf{g}_i|\mathcal{D},x_i) = q(\mathbf{g}_i|\mathcal{D},x_i) \prod_{k=1,k\neq y_i}^{c} \Phi(g_i^k), \tag{25}$$

where the random vector $\mathbf{g}_i$ is defined in Section 3.2.1, and $q(\mathbf{g}_i|\mathcal{D},x_i)$ is the approximate marginal obtained from $q(\mathbf{f}_i|\mathcal{D},x_i)$ by a linear transformation. Note that the marginal predictive probability for class label y_i with the multinomial probit model (2) can be obtained by appropriately forming the transformation B_i and calculating the integral over $\mathbf{g}_i$ in (25). Figures 3(d)-(f) show the contours of the different approximations of $\hat{p}(\mathbf{g}_i|\mathcal{D},x_i)$ for $k \in \{2,3\}$, which for MCMC are obtained using a smoothed estimate of $q(\mathbf{g}_i|\mathcal{D},x_i)$ determined from transformed samples. The distributions are heavily skewed by the probit factors elsewhere than the upper-right quadrant. Compared to the MCMC estimate, IEP places more probability mass to the other quadrants, and therefore underestimates the predictive probability for class 1 more than EP. The approximate predictive probabilities are 0.95 for MCMC, 0.88 for EP, and 0.82 for IEP.

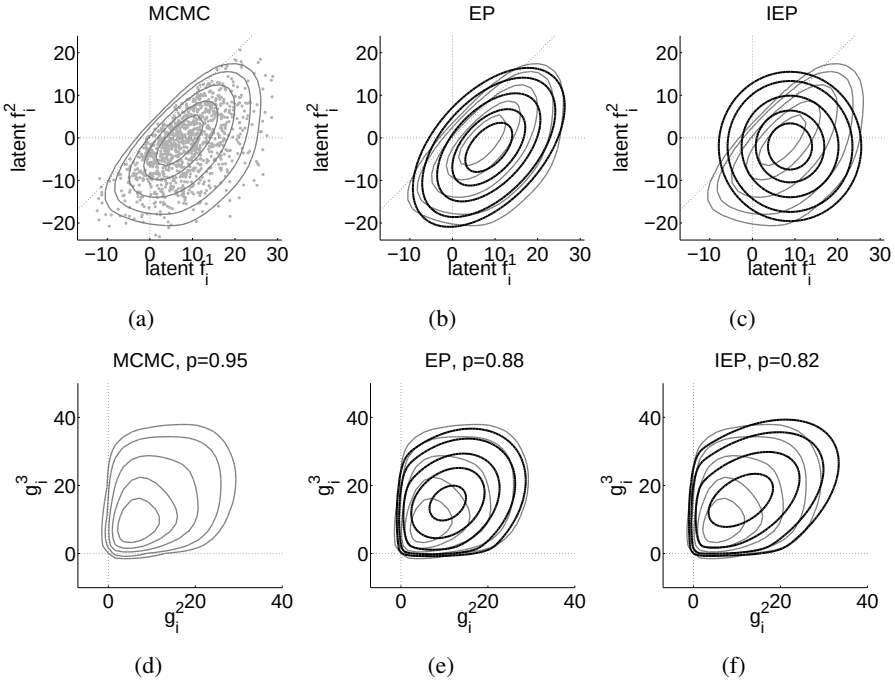


Figure 3: An example of a non-Gaussian marginal posterior distribution for the latent values related to the input x_i in the synthetic example shown in Figure 2. The first row shows the distribution for the latents f_i^1 and f_i^2. Plot (a) shows a scatter-plot of MCMC samples drawn from the posterior and estimated density contour levels which correspond to the areas that include approximately 95%, 90% 75%, 50%, and 25% of the probability mass. Plots (b) and (c) show the equivalent contour levels of the EP and IEP approximations (bold black lines) and the contour levels of the MCMC approximation (gray lines) for comparison. Plots (d)-(f) show contours of $\hat{p}(\mathbf{g}_i|\mathcal{D},x_i)$ for g_i^2 and g_i^3. The probability for class 1 is obtained by calculating the integral over $\mathbf{g}_i$, which results in approximately 0.95 for MCMC, 0.88 for EP, and 0.82 for IEP. See the text for explanation.

The second location x_j is near the class boundary, where all the methods give similar predictive probabilities, although the latent approximations can differ notably as shown in Figures 4(a)-(c), which visualize the marginal approximations for f_j^2 and f_j^3. EP is consistent with the MCMC estimate but due to the independence constraint IEP underestimates the uncertainty of this close-to-Gaussian but non-isotropic marginal distribution. Although Figures 4(d)-(f) show that IEP is more inaccurate than EP, the integral over the tilted distribution of $\mathbf{g}_j$ is in practice the same, since equal amount of probability mass is distributed on both sides of the diagonal in Figure 4(c). The predictive probability for class 2 is approximately 0.47 for all the methods.

5.3 Approximate Marginal Densities with Digit Classification Data

In this section, we compare the predictive performances and marginal likelihood approximations of EP, IEP, VB and LA using the USPS 3 vs. 5 vs. 7 data set, which consists of 1157 training points

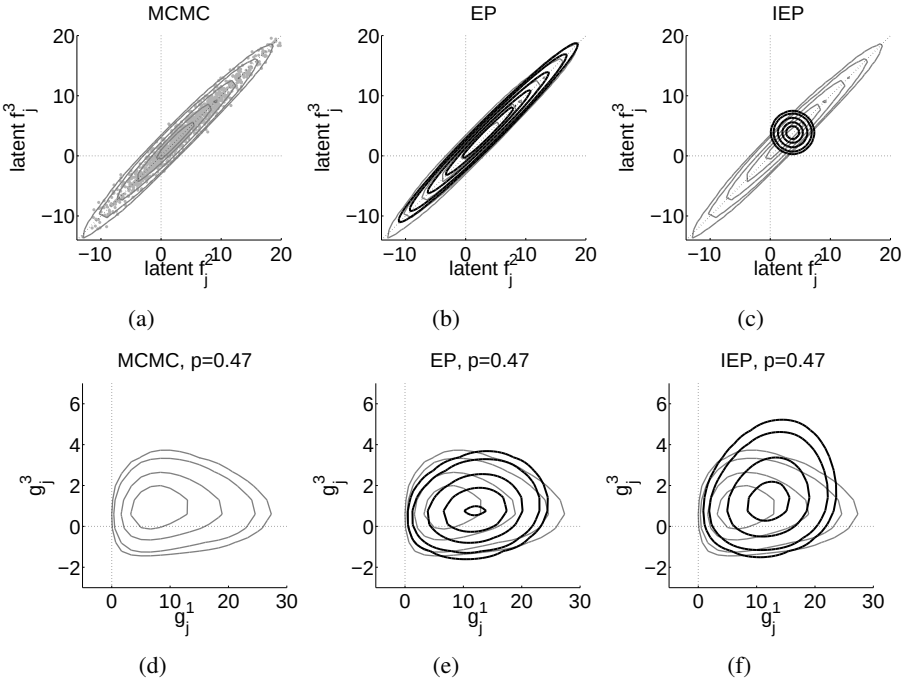


Figure 4: An example of a close-to-Gaussian but non-isotropic marginal posterior distribution for the latent values related to the input x_j in the synthetic example shown in Figure 2. The first row shows the distribution for the latents f_j^2 and f_j^3. Plot (a) shows a scatter-plot of MCMC samples drawn from the posterior and estimated density contour levels which correspond to the areas that include approximately 95%, 90% 75%, 50%, and 25% of the probability mass. Plots (b) and (c) show the equivalent contour levels of the EP and IEP approximations (bold black lines) and the contour levels of the MCMC approximation (gray lines) for comparison. Plots (d)-(f) show contours of $\hat{p}(\mathbf{g}_j|\mathcal{D},x_j)$ for g_j^1 and g_j^3. The probability for class 2 is obtained by calculating the integral over $\mathbf{g}_j$, which results in approximately 0.47 for all the methods. See the text for explanation.

and 1175 test points with 256 covariates. We fixed the hyperparameter values at $\log(\sigma^2) = 4$ and $\log(l) = 2$ which leads to skewed non-Gaussian marginal posterior distributions as will be illustrated shortly.

Figure 5 shows the predictive probabilities of the true class labels for all the approximate methods plotted against the MCMC estimate. The first row shows the training and the second row the test cases. Overall, EP gives the most accurate estimates while IEP slightly underestimates the probabilities for the training cases but performs well for the test cases. Both VB and LA underestimate the predictive probabilities for the test cases, but LA-TKP with the marginal corrections clearly improves the estimates of the LA approximation. Note that the LA methods use a different observation model, and therefore they are compared to the scaled Metropolis-Hastings sampling with the softmax model.

Figure 6 shows an example of the latent marginal posterior distributions for one training point with the correct class label being 2. For each method, the latent pairs (f_i^1, f_i^2), (f_i^1, f_i^3), and (f_i^2, f_i^3),

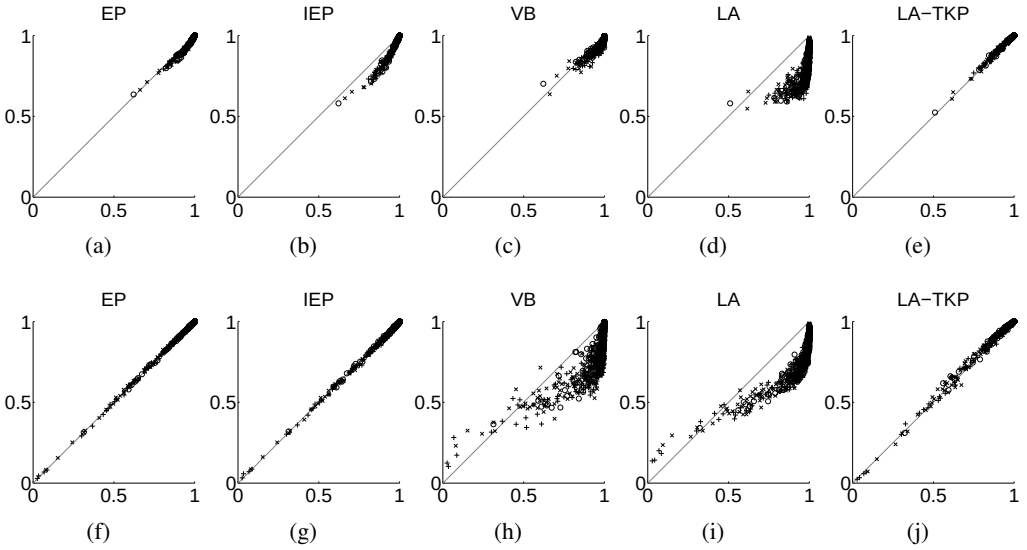


Figure 5: Class probabilities on the USPS 3 vs. 5 vs. 7 data. The MCMC estimates are shown on the x-axis and EP, IEP, VB, LA, and LA-TKP on the y-axis. The first row shows the predictive probabilities of the true class labels for the training points and the second row for the test points. The symbols (x, +, o) corresponds to the handwritten digit target classes 3, 5, and 7. The hyperparameters of the squared exponential covariance function were fixed at $\log(\sigma^2) = 4$ and $\log(l) = 2$.

are shown. The EP approximation agrees reasonably well with the MCMC samples. IEP underestimates the latent uncertainty, especially near the training inputs because of the skewing effect of the likelihood. This seems to affect more the predictive probabilities of the training points in Figure 5(b), which effect can also be seen in the previous example of Figure 2(a) further away from the decision boundary near the input x_i. Figure 6 shows that the VB method underestimates the latent uncertainty. The independence assumption of VB leads to an isotropic approximate distribution, and although the predictive probabilities for the training cases are somewhat consistent with MCMC, the predictions on the test data are less accurate (plots (c) and (h) in Figure 5). Note that the specific hyperparameter values are not optimal for VB, and these values are not supported by the marginal likelihood approximation of VB either, as will be visualized later in this section. The LA approximation captures some of the dependencies between the latent variables associated with different classes, but the joint mode of $\mathbf{f}$ is a poor estimate for the true mean, which causes inaccurate predictive probabilities (plots (d) and (i) in Figure 5). The VB mean estimate is also closer to LA than MCMC, although LA uses a different observation model.

Kuss and Rasmussen (2005) and Nickisch and Rasmussen (2008) discussed how a large value of the magnitude hyperparameter σ^2 can lead to a skewed posterior distribution in binary classification. In the multiclass setting, similar behavior can be seen in the marginal distributions as illustrated in Figures 3 and 6. A large σ^2 leads to a more widely distributed prior which in turn is skewed more strongly by the likelihood where it disagrees with the target class. In the previous comparison, the hyperparameter values were chosen to produce non-Gaussian marginal posterior distributions

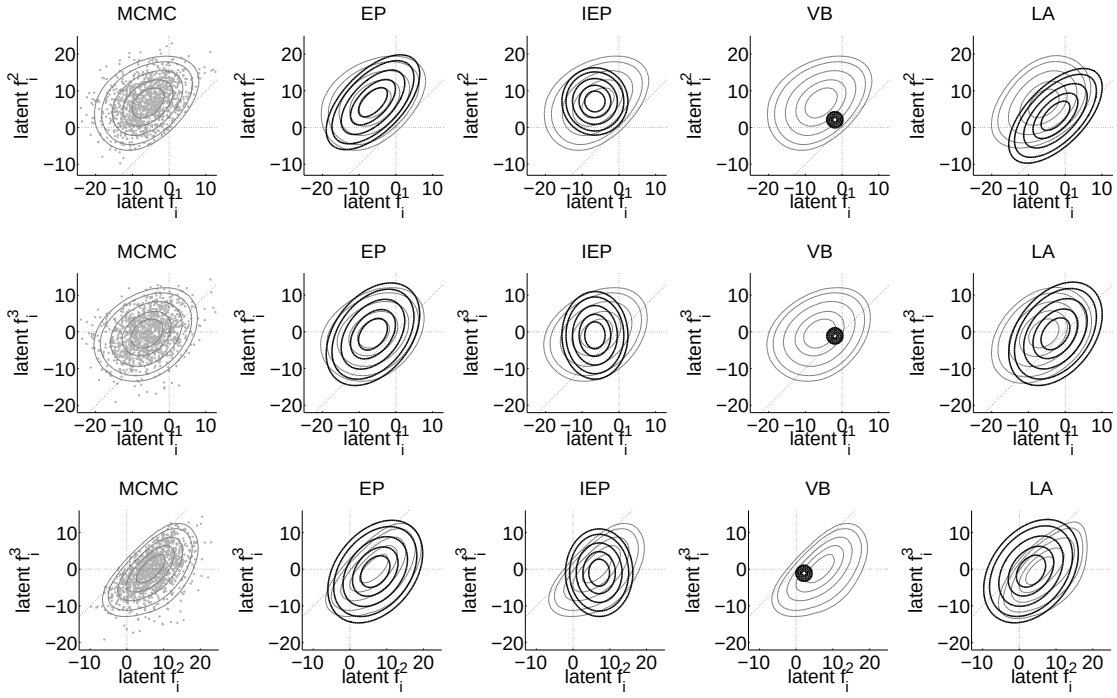


Figure 6: Marginal posterior distributions for one training point with the true class label being 2 on the USPS 3 vs. 5 vs. 7 data. Each row corresponds to one of the latent pairs (f_i^1, f_i^2), (f_i^1, f_i^3), and (f_i^2, f_i^3). The first column shows a scatter-plot of MCMC samples drawn from the posterior and estimated density contour levels which correspond to the areas that include approximately 95%, 90% 75%, 50%, and 25% of the probability mass. The rest of the columns show the equivalent contour levels of the EP, IEP, VB, and LA approximations (bold black lines) and the contour levels of the MCMC approximation (gray lines) for comparison. Note that the last column visualizes a different marginal distribution because LA uses the softmax likelihood. The hyperparameters of the squared exponential covariance function were fixed at $\log(\sigma^2) = 4$ and $\log(l) = 2$ to obtain a non-Gaussian posterior distribution.

for demonstration purposes. However, usually the hyperparameters are estimated by maximizing the marginal likelihood. Kuss and Rasmussen (2005) and Nickisch and Rasmussen (2008) studied the suitability of the marginal likelihood approximations for selecting hyperparameters in binary classification. They compared the calibration of predictive performance and the marginal likelihood estimates on a grid of hyperparameter values. In the following, we extend these comparisons to multiple classes with the USPS data set, for which similar considerations were done by Rasmussen and Williams (2006) with the LA method.

The upper row of Figure 7 shows the log marginal likelihood approximations for EP, IEP, and LA, and the lower bound on evidence for VB as a function of the log-lengthscale $\log(l)$ and log-magnitude $\log(\sigma^2)$ using the USPS 3 vs. 5 vs. 7 data. The middle row shows the log predictive densities evaluated on the test set, and the bottom row shows the corresponding classification accu-

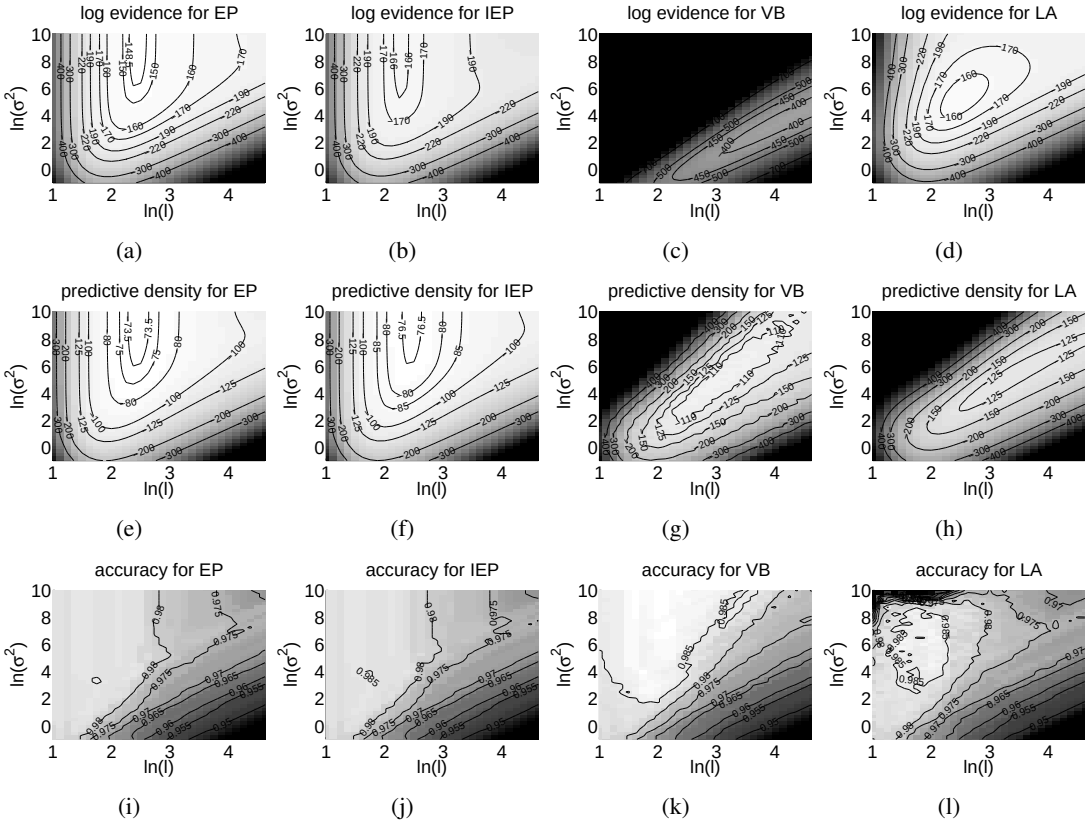


Figure 7: Marginal likelihood approximations and predictive performances as a function of the log-lengthscale $\log(l)$ and log-magnitude $\log(\sigma^2)$ for EP, IEP, VB, and LA on the USPS 3 vs. 5 vs. 7 data. The first row shows the log marginal likelihood approximations, the second row shows the log predictive densities in a test set, and the third row shows the classification accuracies in a test set.

racies. The marginal likelihood approximations and predictive densities for EP and IEP appear to be similar, but the maximum contour of the log marginal likelihood for IEP (the contour labeled with -166 in plot (b) of Figure 7) does not coincide with the maximum contour of the predictive density (the contour labeled with -76.5 in plot (f) of Figure 7), which is why a small bias can occur if the approximate marginal likelihood is used for selecting the hyperparameter values. With EP there is a good agreement between the maximum values in plots (a) and (e), and overall, the log predictive densities are higher than with the other approximations. The log predictive densities of VB and LA are small where $\log(\sigma^2)$ is large (regions where $q(\mathbf{f}|\mathcal{D},\theta)$ is likely to be non-Gaussian), but on the other hand, also the marginal likelihood approximations favor the areas of smaller $\log(\sigma^2)$ values.

There is a reasonable agreement with the marginal likelihood approximations and classification accuracies with EP and IEP, although the maximum accuracies are slightly lower than with VB and LA. The maximum accuracies are very high with VB, but the region of the highest accuracy does not agree with the region of the highest estimate of the marginal likelihood. With LA the marginal like-

lihood estimate is better calibrated in terms of classification accuracy, but the performance worsens when the posterior distribution is skewed with large values of $\log(\sigma^2)$.

5.4 Computational Complexity and Convergence

In this section we consider the computational complexities of the approximate methods for one iteration with fixed hyperparameter values. Note that the following discussion is only approximate, and that the practical efficiency of the algorithms depends much on implementations and the choices of convergence criteria.

Table 2 summarizes the approximate scaling of the number of computations as a function of n and c. EP and IEP refer to the fully-coupled and class-independent approximations, respectively, determined with the proposed nested EP algorithm. QEP refers to the quadrature-based fully-coupled solution using the diagonal approximation $\tilde{\Lambda}_i$ (see Section 3.3), QIEP refers to the quadrature-based class-independent approximation proposed by Seeger et al. (2006), and MCMC refers to Gibbs sampling with the multinomial probit model. The column Posterior complexity of Table 2 describes the overall scaling of the mean and covariance calculations related to the approximate conditional posterior of $\mathbf{f}$. The base computational cost resulting from the full GP prior scales as $O(n^3)$ due to the $n \times n$ matrix inversion (in practice computed using Cholesky decomposition), which is required c times for IEP, QIEP and MCMC, and one additional time for EP, QEP and LA due to incorporation of the between-class correlations. If the same prior covariance structure is used for all classes, VB has the lowest cost, because only one matrix inversion is required per iteration.

The column Likelihood complexity of Table 2 approximates the scaling of the number of calculations that are required besides the posterior mean and covariance evaluations (mainly likelihood related computations for one iteration). For both EP and IEP, this column describes the scaling of the computations needed for the tilted moment approximations done with the inner EP algorithm. For QEP the column summarizes the cost associated with a $(c-1)$-dimensional quadrature rule (denoted by n_{q}^{c-1}) required for the tilted moment evaluations, and for QIEP the cost of one- and two-dimensional quadratures (denoted by n_{q} and n_{q}^2 respectively) required under the independence assumption. For LA the column shows the number of calculations required for evaluating the first and second order derivatives of the softmax likelihood. Each VB iteration requires evaluating the expectations of the auxiliary variables either by a quadrature or sampling, and the cost of one such operation is denoted by n_{q} (for example, the number of quadrature design points). Gibbs sampling with the multinomial probit likelihood requires drawing from the conic truncation of a c-dimensional normal distribution for each observation, and the cost of one draw is denoted by n_{s}. QEP scales inefficiently in c, and is therefore limited to cases with a moderate number of target classes. The QIEP solution can be implemented efficiently because the same function evaluations can be used in all of the $2c-1$ one-dimensional quadratures and the number of two-dimensional quadratures does not depend on c. The cubic scaling in $c-1$ of the tilted moment evaluations in the nested EP and IEP algorithms can be alleviated by reducing the number of inner-loop iterations n_{in} as discussed in Section 3.2.3.

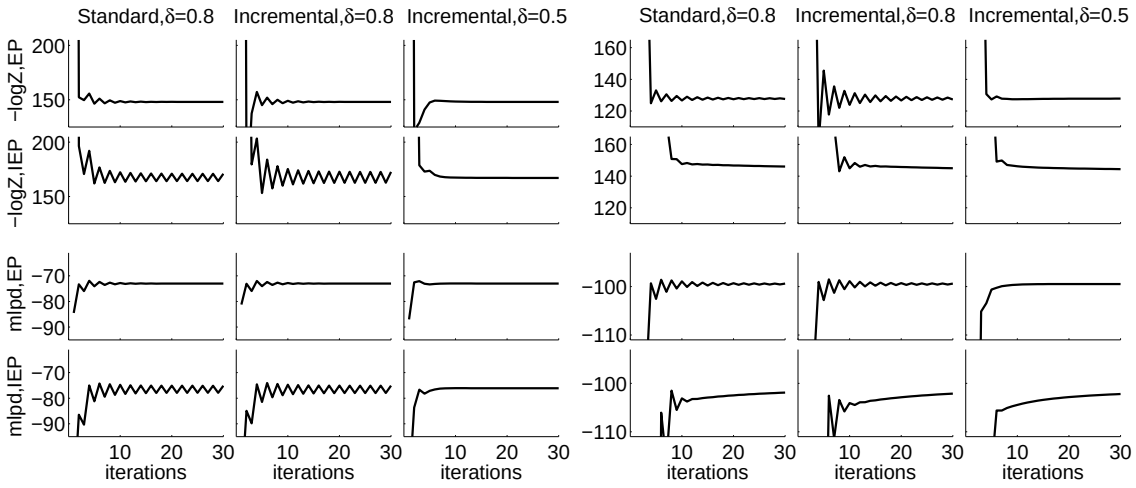
Using the USPS 3 vs. 5 vs. 7 data set, we measured the CPU time required for the posterior inference on $\mathbf{f}$ given nine different preselected hyperparameter values from the grid of Figure 7. With our implementations, LA was the fastest, and EP and VB were about three times more expensive than LA. Because of the efficient scaling (Table 2), VB should be much faster, and probably closer to the running time of LA. One reason for the slow performance may be our implementation based

Algorithm	Posterior complexity	Likelihood complexity
EP	$(c+1)n^3$	$nn_{\text{in}}(c-1)^3$
QEP	$(c+1)n^3$	nn_{q}^{c-1}
IEP	cn^3	$nn_{\text{in}}(c-1)^3$
QIEP	cn^3	$n((2c-1)n_{\text{q}}+2n_{\text{q}}^2)$
VB	n^3	$n(c-1)2n_{\text{q}}$
LA	$(c+1)n^3$	nc
MCMC (Gibbs sampling)	cn^3	ncn_{s}

Table 2: Approximate computational complexities of the various methods as a function of n and c for one iteration with fixed hyperparameters. The column Posterior complexity summarizes the scaling of the mean and covariance calculations related to the approximate conditional posterior of $\mathbf{f}$. The column Likelihood complexity approximates the scaling of the number of calculations required for additional likelihood related computations. The parameter n_{in} refers to the number of inner EP iterations in nested EP, n_{q}^c to the cost of a c-dimensional numerical quadrature, and n_{s} to the cost of sampling from a conic truncation of a c-variate Gaussian distribution.

on importance sampling steps, which may result in slower convergence due to fluctuations. The MCMC and LA-TKP approaches were overall very slow compared to LA. One iteration of MCMC is relatively cheap, but in our experiments thousands of posterior samples were required to obtain chains of sufficiently uncorrelated samples which is why MCMC was over hundred times slower than LA. LA-TKP requires roughly $c+1$ times the CPU time of LA for computing the predictions for each test input. Therefore, the computational cost of LA-TKP becomes quickly prohibitive as the number of test points increases.

In the CPU time comparisons across the range of hyperparameter values producing variety of skewed and non-isotropic posterior distributions, fully-coupled nested EP converged in fewer outer-loop iterations than nested IEP if the same convergence criteria were used. Figure 8 illustrates the difference in convergence with the USPS 3 vs. 5 vs. 7 and Glass data sets. We fixed the hyperparameters at $\log(\sigma^2) = 8$ and $\log(l) = 2.5$ which results in good predictive performances on the independent test data set with both methods for the USPS 3 vs. 5 vs. 7 data (see Figure 7). For both methods, the negative log marginal likelihood approximation $-\log Z_{\text{EP}}$ and the mean log predictive density (mlpd) in the test data set are shown after each iteration. Note that the converged EP approximation satisfying the moment matching conditions between $\hat{p}(\mathbf{f}_i)$ and $q(\mathbf{f}_i)$ corresponds to stationary points of an objective function similar to $-\log Z_{\text{EP}}$ (Minka, 2001b; Opper and Winther, 2005). The convergence is illustrated with a small amount of damping (damping factor $\delta = 0.8$) and with a larger amount of damping ($\delta = 0.5$). With the fully-coupled nested EP algorithm the damping is applied on the inner-EP site parameters $\tilde{\boldsymbol{\alpha}}_i$ and $\tilde{\boldsymbol{\beta}}_i$, whereas with IEP the damping is applied on the natural exponential site parameters $\tilde{\boldsymbol{\nu}}_i$ and $\tilde{\Sigma}_i^{-1}$. In the columns denoted Standard in Figure 8, the inner-loops of the nested EP and IEP algorithms are run until convergence at each outer-loop iteration, whereas in the rest of the columns (Incremental) only one inner-loop iteration per site is done at each outer-loop iteration. Recall from the previous discussion and from Table 2 that the



(a) USPS 3 vs. 5 vs. 7 (b) Glass

Figure 8: A convergence comparison between EP and IEP using parallel updates in the outer EP loop with the USPS 3 vs. 5 vs. 7 data (columns 1-3), and with the Glass data (columns 4-6). The first two rows show the negative log marginal likelihood estimates $-\log Z_{\mathrm{EP}}$ as a function of iterations for two different damping factors δ, and the bottom two rows show the corresponding mean log predictive density (mlpd) evaluated using a separate test data set. In the columns denoted Standard the inner-loops of the nested EP and IEP algorithms are run until convergence at each outer-loop iteration, whereas in the rest of the columns (Incremental) only one inner-loop iteration per site is done at each outer-loop iteration. The hyperparameters of the squared exponential covariance function were fixed at $\log(\sigma^2) = 8$ and $\log(l) = 2.5$ which results in a non-Gaussian posterior distribution.

incremental updates ($n_{\mathrm{in}} = 1$) reduce the computational burden of the inner-loop of the nested EP algorithm which scales as $O(n_{\mathrm{in}}(c-1)^3)$.

From Figure 8 it can be seen that the incremental updates require more damping than standard updates, but both update schemes seem to converge into the same solution. There is a clear difference in the amplitude of oscillations between the nested EP and IEP algorithms with the same damping level but this may be partly caused by the different parameterization. Compared to fully-coupled EP, there is a slow drift in $-\log Z_{\mathrm{EP}}$ and in the mlpd score even after 20 iterations with IEP, and the drift is more visible with the Glass data. One explanation for this behavior can be that the fully-coupled Gaussian distribution is more suitable approximating family for the true posterior (Minka, 2005), which is strongly non-Gaussian because of the large magnitude hyperparameter value, and has stronger between-class posterior dependencies induced through the likelihood terms because of the relatively large lengthscale.

5.5 Predictive Performance Across Data Sets with Hyperparameter Estimation

In this section we assess the predictive performances with estimation of the hyperparameters. We compare the performances of nested EP and IEP, VB, LA, LA-TKP, and Gibbs sampling with the multinomial probit model (MCMC) on various benchmark data sets. All the methods are compared

Data Set	n_{train}	n_{test}	Classes (c)	Covariates (d)	ARD
New-thyroid	215	215 (Ten-fold CV)	3	5	yes
Teaching	151	151 (Ten-fold CV)	3	5	yes
Glass	214	214 (Ten-fold CV)	6	9	yes
Wine	178	178 (Ten-fold CV)	3	13	yes
Image segmentation	210	2100	7	18	no
USPS 3 vs. 5 vs. 7	1157	1175	3	256	no
USPS 10-class	4649	4649	10	256	no

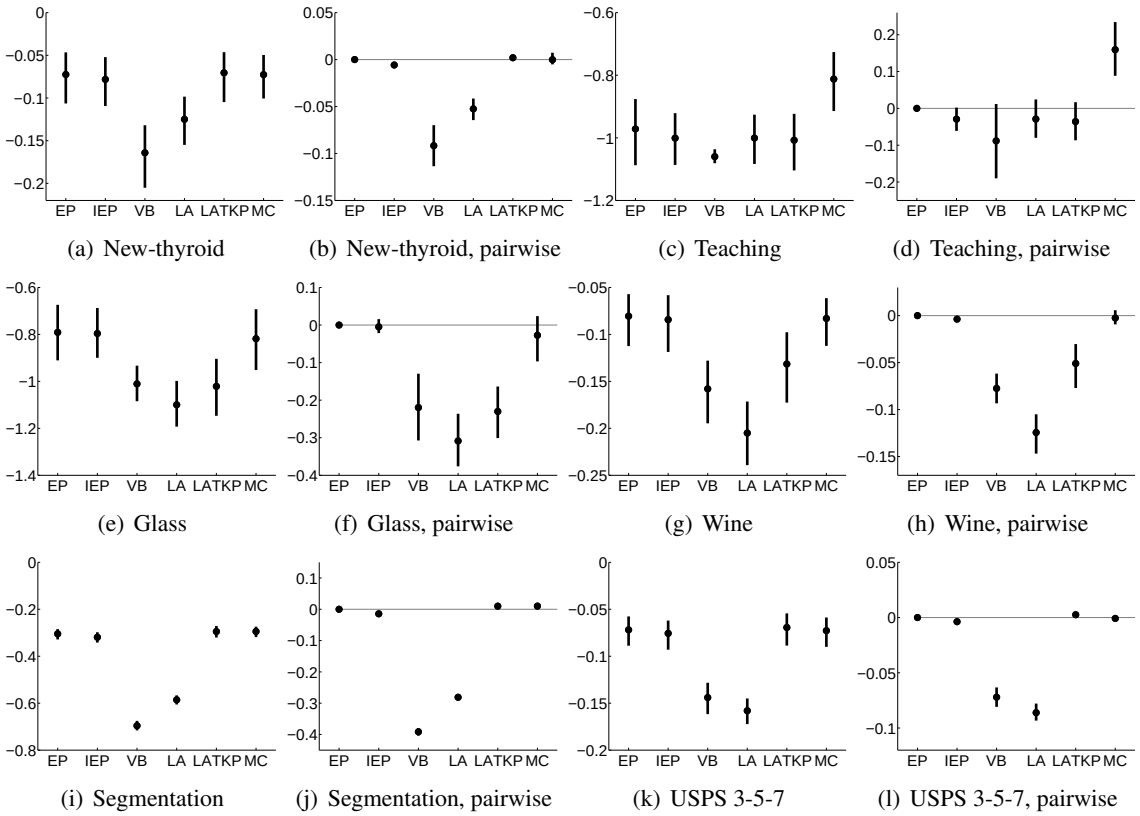
Table 3: Data sets used in the experiments.

using the USPS 3 vs. 5 vs. 7 data, and the following five UCI Machine Learning Repository (Frank and Asuncion, 2010) data sets: New-thyroid, Teaching, Glass, Wine, and Image segmentation. The comparisons are also done using the USPS 10-class data set, but only for EP, IEP, VB, and LA due to the large n. The main characteristics of the data sets are summarized in Table 3.

For all the data sets, we standardize the covariates to zero mean and unit variance, and use the squared exponential covariance function with the same hyperparameters for all classes. For the Image segmentation and USPS data sets we use a common lengthscale parameter for all dimensions. For other data sets we set individual lengthscale parameters for each input dimension (Automatic Relevance Determination, ARD, see, for example, Rasmussen and Williams, 2006). To avoid unnecessarily large hyperparameter values, we place a weakly informative prior on the lengthscale and magnitude parameters by choosing a half Student-t distribution with four degrees of freedom and variance equal to one hundred. With MCMC we sample the hyperparameters, and with the other methods, we use gradient-based type-II MAP estimation to select the hyperparameter values. The predictive performance is measured using a ten-fold cross-validation (CV) with four of the data sets, and using predetermined training and test sets with three of the data sets (see Table 3).

The first and third column in Figure 9 visualize the mean log predictive densities and their approximate 95% central credible intervals for six data sets estimated using the Bayesian bootstrap method as described by Vehtari and Lampinen (2002). To highlight the differences between the methods more clearly, we compute the pairwise differences of the log posterior predictive densities with respect to EP. The second and fourth column in Figure 9 show the mean values and the approximate 95% central credible intervals of the pairwise differences. The comparisons reveal that EP performs well when compared to MCMC; only in the Teaching and Image segmentation data sets MCMC is significantly better. IEP performs worse than EP in all the data sets except Teaching and Glass. The predictive densities of VB and LA are overall worse than EP, IEP or MCMC. LA-TKP improves the performance of LA with all the data sets except Teaching.

The first and third column in Figure 10 visualize the mean classification accuracies and their approximate 95% central credible intervals. The second and fourth column in Figure 10 show the pairwise mean differences of the classification accuracies together with the approximate 95% central credible intervals with respect to EP. Because the difference of the classification outcomes for each observation is a discrete variable with three possible values (worse, same, or better than EP), we use a multinomial model with a non-informative Dirichlet prior distribution for the comparison test. In a case where the method has exactly the same predictions as EP, a small circle is plotted at the mean value. The differences between all the methods are small. In the Teaching data set, where the overall



(a) New-thyroid (b) New-thyroid, pairwise (c) Teaching (d) Teaching, pairwise

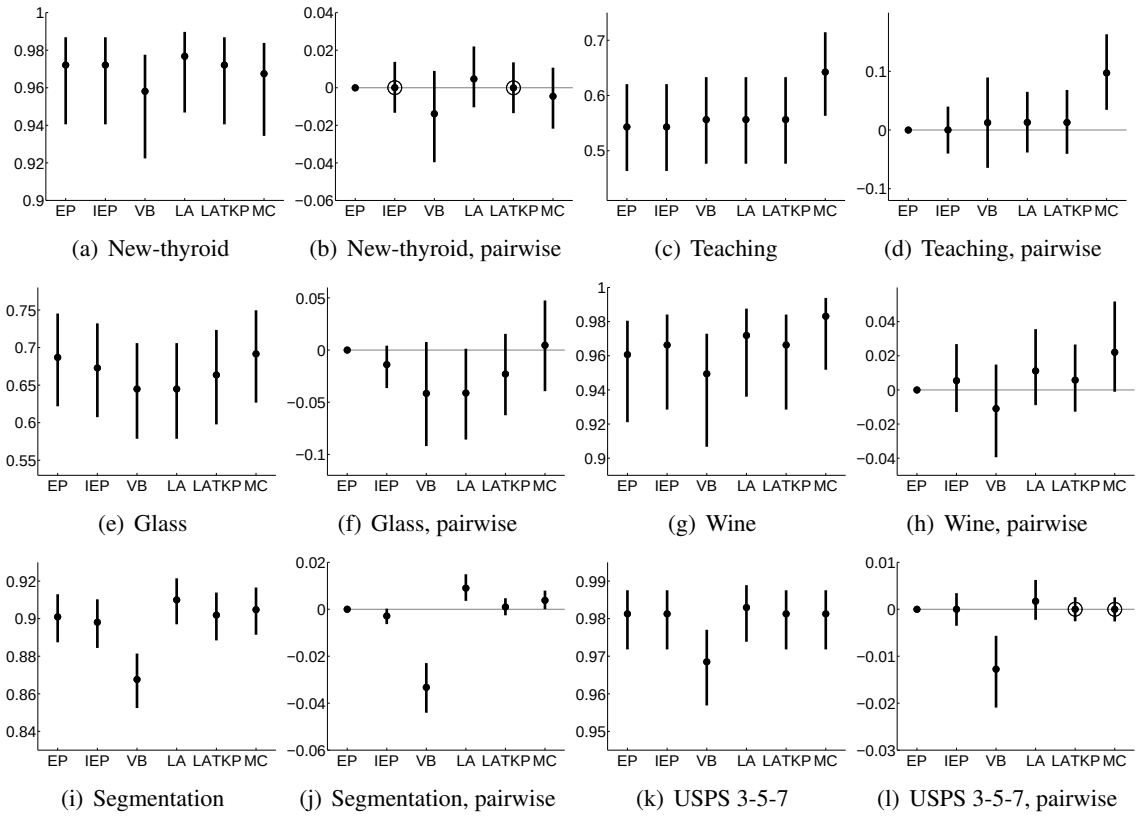
(e) Glass (f) Glass, pairwise (g) Wine (h) Wine, pairwise

(i) Segmentation (j) Segmentation, pairwise (k) USPS 3-5-7 (l) USPS 3-5-7, pairwise

Figure 9: The first and third column: The mean log predictive densities and their approximate 95% credible intervals for six data sets (see Table 3) using EP, IEP, VB, LA, LA-TKP, and MCMC with Gibbs sampling. The second and fourth column: The pairwise differences of the log predictive densities with respect to EP (mean + approximate 95% credible intervals). Values above zero indicate that a method is performing better than EP.

accuracy is the lowest, the MCMC estimate is significantly better than any other method. There is no statistically significant difference between EP and IEP; IEP performs slightly better in the Wine data set, but EP has a better accuracy in the Glass and Image segmentation data sets, which both have more than three target classes, and in which the overall classification accuracies are among the lowest. LA has a good classification accuracy, and performs better than EP in Image segmentation. A possible explanation for this is the different shape of the softmax likelihood function used by LA. If the classification accuracy is the only criterion, the LA-TKP correction seems unnecessary. VB has the lowest classification performance and is significantly worse than the other methods in the Image segmentation and USPS 3 vs. 5 vs. 7 data sets, which is probably caused by a worse estimate of the hyperparameter values.

Finally, we summarize the mlpd scores and classification accuracies of EP, IEP, VB, and LA with the USPS 10-class data set in Figure 11. Both EP approaches are significantly better than VB or LA with both measures. Considering the EP approaches, fully-coupled EP achieves a slightly better



(a) New-thyroid (b) New-thyroid, pairwise (c) Teaching (d) Teaching, pairwise

(e) Glass (f) Glass, pairwise (g) Wine (h) Wine, pairwise

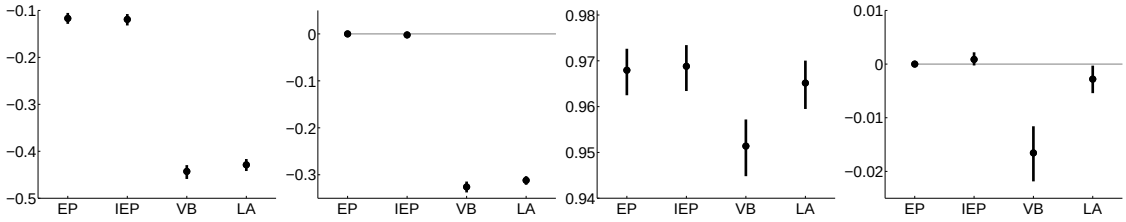
(i) Segmentation (j) Segmentation, pairwise (k) USPS 3-5-7 (l) USPS 3-5-7, pairwise

Figure 10: The first and third column: The classification accuracies and their approximate 95% credible intervals for six data sets (see Table 3) using EP, IEP, VB, LA, LA-TKP, and MCMC with Gibbs sampling. The second and fourth column: The pairwise differences of the classification accuracies with respect to EP (mean + approximate 95% credible intervals). Values above zero indicate that a method is performing better than EP. A small circle is plotted at the mean value if the predictions are exactly the same as with EP.

mlpd score, whereas IEP is slightly better in terms of classification accuracy, but the differences are not statistically significant.

6. Conclusions and Further Research

EP approaches for GP classification with the multinomial probit model have already been proposed by Seeger et al. (2006) and Girolami and Zhong (2007). In this paper, we have complemented their work with a novel quadrature-free nested EP algorithm that maintains all between-class posterior dependencies but still scales linearly in the number of classes. Our comparisons with fixed hyperparameters show that compared to quadrature-based EP algorithms, nested EP achieves similar accuracy, and its computational cost is comparable with a class-independent approximation whereas with full posterior couplings nested EP scales more efficiently. When the hyperparame-



(a) USPS (b) USPS, pairwise (c) USPS (d) USPS, pairwise

Figure 11: The mean log predictive densities (a) and classification accuracies (c) for the USPS 10-class data set (see Table 3) using EP, IEP, VB, and LA. The pairwise differences of the log predictive densities and the classification accuracies with respect to EP are shown in plots (b) and (d), respectively. In plots (b) and (d) values above zero indicate that a method is performing better than EP. In each plot, the mean values and approximate 95% central credible intervals are shown.

ters are determined by optimizing the marginal likelihood, nested EP is a consistent approximate method compared to full MCMC. In terms of predictive density, nested EP is close to MCMC, and more accurate compared to VB and LA, but if only the classification accuracy is concerned, all the approximations perform similarly. LA-TKP improves the predictive density estimates of LA but the computational cost becomes increasingly demanding if a larger number of predictions are needed.

In our comparisons the predictive accuracies of the full EP and IEP solutions obtained using the nested EP algorithm are similar for practical purposes. However, our visualizations show that the approximate marginal posterior distributions of the latent values provided by full EP are clearly more accurate, although the full nested EP solution can be calculated with similar computational burden as nested IEP. Because there is no convergence guarantee for the standard EP algorithm, it is worth to notice the differences in the convergence properties of full EP and IEP observed in our experiments. With the same hyperparameter values, nested IEP converged more slowly and required more damping than full nested EP. This can be due to slower propagation of information caused by the independence assumptions, and this behavior can get worse as the between-class posterior couplings get stronger with certain hyperparameter values. Given all these observations, we prefer full EP to IEP.

Models in which each likelihood term related to a certain observation depends on multiple latent values, such as the multinomial probit, are challenging for EP because a straightforward quadrature-based implementation may become computationally infeasible unless independence assumptions between the latent values or some other simplifications are made. In the presented nested EP approach, we have applied inner EP approximations for each likelihood term within an outer EP framework in a computationally efficient manner. This approach could be applicable also for other similar multi-latent models which involve integral representations consisting of simple factorized functions each depending on linear transformations of the latent variables. For example, one straightforward extension would be linear multinomial probit regression with Gaussian priors on the weights.

A drawback with GP classifiers is the fundamental computational scaling $O(n^3)$ resulting from the prior structure. To speed up the inference in multiclass GP classification, sparse approximations

such as the informative vector machine (IVM) have been proposed (Seeger and Jordan, 2004; Girolami and Rogers, 2006; Seeger et al., 2006). IVM uses the information provided by all observations to form an active subset which is then used to form the posterior mean and covariance approximations. The presented EP approach could be extended to IVM in a similar fashion as described by Seeger and Jordan (2004). The accurate marginal approximations of full EP could be useful in determining the relative entropy measures used as a scoring criterion to select the active set. To speed up the computations, the inner EP site parameters could be updated iteratively even for the observations not in the active set in a similar fashion as described in Section 3.2. Recently, a similar approach to IVM called predictive active set selection (PASS-GP) has been proposed by Henao and Winther (2010) to lower the computational complexity in binary GP classification. PASS-GP uses the approximate cavity and cavity predictive distributions of EP to determine a representative active set. The proposed EP approach could prove useful when extending PASS-GP to multiple classes, because it provides accurate marginal predictive density estimates.

The presented fully-coupled nested EP approach for approximate inference with Gaussian process models is implemented in the free GPstuff software package and the code will be made available at `http://becs.aalto.fi/en/research/bayes/gpstuff/`.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments to improve the manuscript. The research has been funded by the Academy of Finland (grant 218248).

Appendix A. Approximating Tilted Moments Using EP

For convenience, we summarize the inner EP algorithm for approximating the tilted moments resulting from a multinomial probit likelihood. Essentially the same algorithm was presented by Minka (2001a) for classification with the Bayes point machine and later by Qi et al. (2004) for the binary probit classifier. To facilitate a computationally efficient implementation, the following algorithm description is written with an emphasis to reduce the number of vector and matrix operations in a similar fashion as in the general EP formulation presented by Cseke and Heskes (2011, Appendix C).

We want to approximate the normalization, mean and covariance of the tilted distribution

$$\hat{p}(\mathbf{w}_i) = \hat{Z}_i^{-1} \mathcal{N}(\mathbf{w}_i | \boldsymbol{\mu}_{\mathrm{w}_i}, \Sigma_{\mathrm{w}_i}) \prod_{j=1, j\neq y_i}^{c} \Phi(\mathbf{w}_i^T \tilde{\mathbf{b}}_{i,j}).$$

This is done using the EP algorithm which results in the following Gaussian approximation

$$\hat{q}(\mathbf{w}_i) = Z_{\hat{q}_i}^{-1} \mathcal{N}(\mathbf{w}_i | \boldsymbol{\mu}_{\mathrm{w}_i}, \Sigma_{\mathrm{w}_i}) \prod_{j=1, j\neq y_i}^{c} \tilde{Z}_j \mathcal{N}(\mathbf{w}_i^T \tilde{\mathbf{b}}_{i,j} | \tilde{\beta}_{i,j} \tilde{\alpha}_{i,j}^{-1}, \tilde{\alpha}_{i,j}^{-1}) = \mathcal{N}(\mathbf{w}_i | \boldsymbol{\mu}_{\hat{q}_i}, \Sigma_{\hat{q}_i}),$$

where we have used the natural parameters $\tilde{\alpha}_{i,j}$ (precision) and $\tilde{\beta}_{i,j}$ (location) for the site approximations. The index i denotes the i'th observation, and to clarify the notation below, we leave out this index from the inner EP terms. In the first outer-loop, the site parameters $\tilde{\alpha}$ and $\tilde{\beta}$ are initialized to zero, $\boldsymbol{\mu}_{\hat{q}_i}$ to $\boldsymbol{\mu}_{\mathrm{w}_i}$, and $\Sigma_{\hat{q}_i}$ to Σ_{w_i}. After the first outer-loop, these parameters are initialized to their

last values from the previous outer-loop iteration for speed-up. The following steps are repeated for all $j = \{1,\ldots,c | j \neq y_i\}$ until convergence.

1. Cavity evaluations:

$$\begin{aligned} v_{-j} &= (v_j^{-1} - \tilde{\alpha}_j)^{-1}, \\ m_{-j} &= v_{-j}(v_j^{-1} m_j - \tilde{\beta}_j), \end{aligned}$$

   where scalars $v_j = \tilde{\mathbf{b}}_{i,j}^T \Sigma_{\hat{q}_i} \tilde{\mathbf{b}}_{i,j}$ and $m_j = \tilde{\mathbf{b}}_{i,j}^T \boldsymbol{\mu}_{\hat{q}_i}$ correspond to the marginal distribution of latent $g_i^j = \mathbf{w}_i^T \tilde{\mathbf{b}}_{i,j}$.

2. Tilted moments for g_i^j:

$$\begin{aligned} \hat{Z}_j &= \Phi(z_j), \\ \hat{m}_j &= \rho_j v_{-j} + m_{-j}, \\ \hat{v}_j &= v_{-j} - v_{-j}^2 \gamma_j, \end{aligned}$$

   where $z_j = m_{-j}(1+v_{-j})^{-1/2}$, $\rho_j = \frac{\mathcal{N}(z_j|0,1)}{\Phi(z_j)}(1+v_{-j})^{-1/2}$ and $\gamma_j = \rho_j^2 + z_j \rho_j (1+v_{-j})^{-1/2}$.

3. Site updates with damping:

$$\begin{aligned} \Delta\tilde{\alpha}_j &= \delta(\hat{v}_j^{-1} - v_j^{-1}), \\ \Delta\tilde{\beta}_j &= \delta(\hat{v}_j^{-1}\hat{m}_j - v_j^{-1} m_j), \end{aligned}$$

   where $\delta \in (0,1]$ is the damping factor.

4. Rank-1 covariance update:

$$\begin{aligned} \Sigma_{\hat{q}_i}^{\text{new}} &= \Sigma_{\hat{q}_i} - \boldsymbol{\phi}_j (1 + \Delta\tilde{\alpha}_j v_j)^{-1} \Delta\tilde{\alpha}_j \boldsymbol{\phi}_j^T, \\ \boldsymbol{\mu}_{\hat{q}_i}^{\text{new}} &= \boldsymbol{\mu}_{\hat{q}_i} + \boldsymbol{\phi}_j (1 + \Delta\tilde{\alpha}_j v_j)^{-1} (\Delta\tilde{\beta}_j - \Delta\tilde{\alpha}_j m_j), \end{aligned}$$

   where $\boldsymbol{\phi}_j = \Sigma_{\hat{q}_i} \tilde{\mathbf{b}}_{i,j}$.

Alternatively, the rank-1 updates of step 4 could be replaced by only one parallel covariance update after each sweep over the sites indexed by j.

After convergence, we evaluate the normalization $Z_{\hat{q}_i}$ of the tilted distribution as

$$\begin{aligned} \log Z_{\hat{q}_i} &= \frac{1}{2}\boldsymbol{\mu}_{\hat{q}_i}^T \Sigma_{\hat{q}_i}^{-1} \boldsymbol{\mu}_{\hat{q}_i} + \frac{1}{2}\log|\Sigma_{\hat{q}_i}| - \frac{1}{2}\boldsymbol{\mu}_{\mathrm{w}_i}^T \Sigma_{\mathrm{w}_i}^{-1} \boldsymbol{\mu}_{\mathrm{w}_i} - \frac{1}{2}\log|\Sigma_{\mathrm{w}_i}| + \sum_{j=1, j\neq y_i}^{c} \log \hat{Z}_j \\ &\quad + \frac{1}{2}\sum_{j=1, j\neq y_i}^{c} \left(m_{-j}^2 v_{-j}^{-1} + \log v_{-j} \right) - \frac{1}{2}\sum_{j=1, j\neq y_i}^{c} \left(m_j^2 v_j^{-1} + \log v_j \right). \end{aligned}$$

Appendix B. Details of Posterior Computations

The site covariance can be written as $\tilde{T} = D - DR(R^T DR)^{-1}R^T D$, where D is a $cn \times cn$ diagonal matrix $D = \text{diag}\left[\pi_1^1, \ldots, \pi_n^1, \pi_1^2, \ldots, \pi_n^2, \ldots, \pi_1^c, \ldots, \pi_n^c\right]^T$, and R is a $cn \times n$ matrix consisting of identity matrices I_n stacked c times vertically. To compute predictions related to a test point $\mathbf{x}_*$, we need to first evaluate the mean and covariance of $\mathbf{f}_* = \left[f_*^1, f_*^2, \ldots, f_*^c\right]^T$ as

$$\begin{aligned} \mathrm{E}[\mathbf{f}_*] &= K_*\tilde{\boldsymbol{\nu}} - K_* M K\tilde{\boldsymbol{\nu}}, \quad (26) \\ \mathrm{Cov}[\mathbf{f}_*] &= K_{*,*} - K_* M K_*^T, \quad (27) \end{aligned}$$

where $\tilde{\boldsymbol{\nu}}$ contains all $\tilde{\boldsymbol{\nu}}_i$ in the same order with $\mathbf{f}$, $M = \tilde{T}(I_{cn} + K\tilde{T})^{-1}$, K_* is a $c \times cn$ covariance matrix between the test point and the training points, and $K_{*,*}$ is a $c \times c$ covariance matrix for the test point. The matrix M in Equations (26) and (27) can be evaluated using

$$M = B - BRP^{-1}R^T B,$$

where $B = D^{1/2}A^{-1}D^{1/2}$, $P = R^T BR$, and $A = I_{cn} + D^{1/2}KD^{1/2}$. To evaluate expressions involving M, we compute the Cholesky decompositions of P and the c diagonal blocks of A, which results in the scaling $O((c+1)n^3)$. The predictive mean and covariance can be computed using the block-diagonal structure of B and the sparse structure of K_*. Given $\mathrm{E}[\mathbf{f}_*]$ and $\mathrm{Cov}[\mathbf{f}_*]$, the integration over the posterior uncertainty of $\mathbf{f}_*$ required to compute the predictive class probabilities, is equivalent to the tilted moment evaluation, and can be approximated using the algorithm described in Appendix A.

To compute the marginal mean $\boldsymbol{\mu}_i$ (a vector of length c) and covariance Σ_i (a matrix of size $c \times c$) of the training latent $\mathbf{f}_i$ for all i during the posterior update step in the outer-loop iteration, we replace K_* and $K_{*,*}$ with K in Equations (26) and (27), and compute only the required blocks of the full posterior covariance matrix. After convergence of the outer EP algorithm, the marginal likelihood approximation of EP can be computed as

$$\begin{aligned} \log Z_{\mathrm{EP}} &= \frac{1}{2}\tilde{\boldsymbol{\nu}}^T\boldsymbol{\mu} - \frac{1}{2}\log|I_{cn} + K\tilde{T}| + \sum_{i=1}^{n}\log Z_{\hat{q}_i} + \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{\mu}_{-i}^T\Sigma_{-i}^{-1}\boldsymbol{\mu}_{-i} + \log|\Sigma_{-i}|\right) \\ &\quad - \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{\mu}_i^T\Sigma_i^{-1}\boldsymbol{\mu}_i + \log|\Sigma_i|\right), \quad (28) \end{aligned}$$

where $\boldsymbol{\mu}_{-i}$ and Σ_{-i} are the i'th cavity mean and covariance, and the posterior mean $\boldsymbol{\mu}$ contains all $\boldsymbol{\mu}_i$. The normalization terms $Z_{\hat{q}_i}$ are obtained from the inner EP algorithm described in Appendix A. Finally, the determinant term in (28) can be evaluated as

$$|I_{cn} + K\tilde{T}| = |A||R^T DR|^{-1}|P|.$$

The gradients of the log marginal likelihood with respect to θ can be obtained by calculating only the explicit derivatives of the first two terms of (28). The implicit derivatives with respect to the site parameters and cavity parameters (in their natural exponential forms) of the outer EP cancel each other out in the convergence (Opper and Winther, 2005; Seeger, 2005). Since the likelihood does not depend on any hyperparameters, the explicit derivatives of $\log Z_{\hat{q}_i}$ are zero. Also the implicit derivatives of $\log Z_{\hat{q}_i}$ with respect to the inner EP parameters cancel out because these terms are formed as marginal likelihood approximations with the inner EP, which also satisfies the same cancellation property of the EP algorithm.

References

Kian Ming A. Chai. Variational multinomial logit Gaussian process. *Journal of Machine Learning Research*, 13:1745–1808, 2012.

Botond Cseke and Tom Heskes. Approximate marginals in latent Gaussian models. *Journal of Machine Learning Research*, 12:417–454, 2011.

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

Andrew Frank and Arthur Asuncion. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, 2010. URL `http://archive.ics.uci.edu/ml`.

Mark Girolami and Simon Rogers. Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Computation*, 18:1790–1817, 2006.

Mark Girolami and Mingjun Zhong. Data integration for classification problems employing Gaussian process priors. In *Advances in Neural Information Processing Systems 19*, pages 465–472. The MIT Press, 2007.

Ricardo Henao and Ole Winther. PASS-GP: Predictive active set selection for Gaussian processes. In *IEEE International Workshop on Machine Learning for Signal Processing*, pages 148–153, 2010.

Daniel Hernández-Lobato, José M. Hernández-Lobato, and Pierre Dupont. Robust multi-class Gaussian process classification. In *Advances in Neural Information Processing Systems 24*, pages 280–288, 2011.

Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994.

Hyun-Chul Kim and Zoubin Ghahramani. Bayesian Gaussian process classification with the EM-EP algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1948–1959, 2006.

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

Thomas P. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology, 2001a.

Thomas P. Minka. Expectation Propagation for approximative Bayesian inference. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 362–369. Morgan Kaufmann, San Francisco, CA, 2001b.

Thomas P. Minka. Divergence measures and message passing. Technical report, Microsoft Research, Cambridge, 2005.

Thomas P. Minka and John Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 352–359. Morgan Kaufmann, San Francisco, CA, 2002.

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, 1996.

Radford M. Neal. Regression and classification using Gaussian process priors (with discussion). In *Bayesian Statistics 6*, pages 475–501. Oxford University Press, 1998.

Hannes Nickisch and Carl E. Rasmussen. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078, 2008.

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

Yuan Qi, Thomas P. Minka, Rosalind W. Picard, and Zoubin Ghahramani. Predictive automatic relevance determination by expectation propagation. In *Proceedings of the 21st International Conference on Machine Learning*, pages 671–678, 2004.

Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society (Series B)*, 71(2):319–392, 2009.

Matthias Seeger. Expectation propagation for exponential families. Technical report, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2005.

Matthias Seeger and Michael Jordan. Sparse Gaussian process classification with multiple classes. Technical report, University of California, Berkeley, CA, 2004.

Matthias Seeger, Neil Lawrence, and Ralf Herbrich. Efficient nonparametric Bayesian modelling with sparse Gaussian process approximations. Technical report, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2006.

Alexander Smola, Vishy Vishwanathan, and Eleazar Eskin. Laplace propagation. In *Advances in Neural Information Processing Systems 16*. The MIT Press, 2004.

Edward Snelson and Zoubin Ghahramani. Compact approximations to Bayesian predictive distributions. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 840–847, 2005.

Luke Tierney and Joseph B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

Marcel van Gerven, Botond Cseke, Robert Oostenveld, and Tom Heskes. Bayesian source localization with the multivariate Laplace prior. In *Advances in Neural Information Processing Systems 22*, pages 1901–1909, 2009.

Aki Vehtari and Jouko Lampinen. Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, 14(10):2439–2468, 2002.

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