Journal of Machine Learning Research 13 (2012) 1745-1808 Submitted 1/12; Published 6/12

# Variational Multinomial Logit Gaussian Process

**Kian Ming A. Chai** CKIANMIN@DSO.ORG.SG
*DSO National Laboratories*
*20 Science Park Drive*
*Singapore 118230*

**Editor:** Manfred Opper

## Abstract

Gaussian process prior with an appropriate likelihood function is a flexible non-parametric model for a variety of learning tasks. One important and standard task is multi-class classification, which is the categorization of an item into one of several fixed classes. A usual likelihood function for this is the multinomial logistic likelihood function. However, exact inference with this model has proved to be difficult because high-dimensional integrations are required. In this paper, we propose a variational approximation to this model, and we describe the optimization of the variational parameters. Experiments have shown our approximation to be tight. In addition, we provide data-independent bounds on the marginal likelihood of the model, one of which is shown to be much tighter than the existing variational mean-field bound in the experiments. We also derive a proper lower bound on the predictive likelihood that involves the Kullback-Leibler divergence between the approximating and the true posterior. We combine our approach with a recently proposed sparse approximation to give a variational sparse approximation to the Gaussian process multi-class model. We also derive criteria which can be used to select the inducing set, and we show the effectiveness of these criteria over random selection in an experiment.

**Keywords:** Gaussian process, probabilistic classification, multinomial logistic, variational approximation, sparse approximation

## 1. Introduction

Gaussian process (GP, Rasmussen and Williams, 2006) is attractive for non-parametric probabilistic inference because knowledge can be specified directly in the prior distribution through the mean and covariance function of the process. Inference can be achieved in closed form for regression under Gaussian noise, but approximation is necessary under other likelihoods. For binary classification with logistic and probit likelihoods, a number of approximations have been proposed and compared (Nickisch and Rasmussen, 2008). These are either Gaussian or factorial approximations to the posterior of the latent function values at the observed inputs. Compared to the binary case, progress is slight for multi-class classification. The main hurdle is the need for—and yet the lack of—accurate approximation to the multi-dimensional integration of the likelihood or the log-likelihood against Gaussians (Seeger and Jordan, 2004).

For multi-class classification with latent Gaussian process, different likelihood functions may be used: the multinomial logistic function (Williams and Barber, 1998; Gibbs, 1997; Seeger and Jordan, 2004), also called the soft-max (Bridle, 1989); the multinomial probit function (Girolami and Rogers, 2006); and the uniform noise model (Kim and Ghahramani, 2006). For inference, the exact posterior is usually approximated with a Gaussian or a factorial distribution, similar to the binary case. Different principles may be used to fit the approximation: Laplace approximation (Williams and Barber, 1998); assumed density filtering (Seeger and Jordan, 2004) and expectation propagation (Kim and Ghahramani, 2006); and variational approximation (Gibbs, 1997; Girolami and Rogers, 2006).

This paper addresses the variational approximation of the multinomial logit Gaussian process model, where the likelihood function is the multinomial logistic. In contrast with the variational mean-field approach of Girolami and Rogers (2006), where a factorial approximation is assumed from the onset, we use a full Gaussian approximation on the posterior of the latent function values. The approximation is fitted by minimizing the Kullback-Leibler divergence to the true posterior, which is known to be the same as maximizing a variational lower bound on the marginal likelihood. This procedure requires the expectation of the log-likelihood under the approximating distribution. This is intractable in general, so we introduce a bound on the expected log-likelihood and optimize this bound instead. This contrasts with the proposal by Gibbs (1997) to bound the multinomial logistic likelihood directly. Our bound on the expected log-likelihood is derived using a novel variational method that results in the multinomial logistic being associated with a mixture of Gaussians. Monte-Carlo simulations indicate that this bound is very tight in practice.

Our approach gives a lower bound on the marginal likelihood of the model. By fixing some variational parameters, we arrive at data-independent bounds on the marginal likelihood. These bounds depend only on the number of classes and kernel Gram matrix of the data, but not on the classifications in the data. On four UCI data sets, the one bound we evaluated is tighter than the variational mean-field bound (Girolami and Rogers, 2006).

Although the variational approximation provides a lower bound on the marginal likelihood, approximate prediction in the usual straightforward manner does not necessarily give a lower bound on the predictive likelihood. We show that a proper lower bound on the predictive likelihood can be obtained when we take into account the Kullback-Leibler divergence between the approximating and the true posterior. This perspective supports the minimization of the divergence as a criterion for approximate inference.

To address large data sets, we give a sparse approximation to the multinomial logit Gaussian process model. In a natural manner, this sparse approximation combines our proposed variational approximation with the variational sparse approximation that has been introduced for regression (Titsias, 2009a). The result maintains a variational lower bound on the marginal likelihood, which can be used to guide model learning. We also introduce scoring criteria for the selection of the inducing variables in the sparse approximation. Experiments indicate that the criteria are effective.

### 1.1 Overview

In Section 2, we describe the latent Gaussian process model with the multinomial logistic likelihood, and we give the variational lower bound on the marginal likelihood for approximate inference. The data-independent bounds on the marginal likelihood are developed in this section and so are the bounds for the predictive likelihood. In Section 3, we provide the necessary updates to optimize the variational bound. Sparse approximation is presented in Section 4. Section 5 looks at the sum-to-zero property that exists in our variational inference for certain covariance functions. This is the property that has been used in motivating several single-machine multi-class support vector machines (SVMs). Section 6 addresses model learning for the multinomial logit Gaussian process model. It also looks at the active selection of the inducing set for sparse approximation. Section 7 outlines the computational complexity of our approach. Related work is discussed in Section 8. Section 9 describes several experiments and gives the results. Among others, we compare the tightness of our variational approximation to the variational mean-field approximation (Girolami and Rogers, 2006), and the errors of our classification results with those given by four single-machine multi-class SVMs. Section 10 concludes and provides further discussions.

### 1.2 Notation

Vectors are represented by lower-case bold-faced letters, and matrices are represented by upper-case normal-faced letters. The transpose of matrix $X$ is denoted by $X^{\mathrm{T}}$. An asterisk $*$ in the superscript is used for the optimized value of a quantity or function. Sometimes it is used twice when optimized with respect to two variables. For example, if $h(x,y)$ is a function, $h^*(y)$ is $h(x,y)$ optimized over x, and $h^{**}$ is $h(x,y)$ optimized over $x$ and $y$. The dependency of a function on its variables is frequently suppressed when the context is clear: we write $h$ instead of $h(x,y)$ and $h^*$ instead of $h^*(y)$. In optimizing a function $h(x)$ over $x$, $x^{\mathrm{fx}}$ and $x^{\mathrm{NR}}$ refers to fixed-point update and Newton-Raphson update respectively, while $x^{\mathrm{cc}}$ refers to an update using the convex combination $x^{\mathrm{cc}} = (1-\eta)x_1 + \eta x_2$, where $\eta \in [0,1]$ is to be determined, and $x_1$ and $x_2$ are in the domain of optimization.

We use $\mathbf{x}_i$ for an input that has to be classified into one of $C$ classes. The class of $\mathbf{x}_i$ is denoted by $\mathbf{y}_i$ using the one-of-$C$ encoding. Hence, $\mathbf{y}_i$ is in the canonical basis of $\mathbb{R}^C$, which is the set $\{\mathbf{e}^c\}_{c=1}^{C}$, where $\mathbf{e}^c$ has one at the $c$th entry and zero everywhere else. Class index $c$ is used as superscript, while datum index $i$ is used as subscript. The $c$th entry in $\mathbf{y}_i$ is denoted by $y_i^c$, which is in $\{0,1\}$, and $\mathbf{x}_i$ belongs to the $c$th class if $y_i^c = 1$.

Both $\mathbf{x}_i$ and $\mathbf{y}_i$ are observed variables. Associated with each $y_i^c$ is a latent random function response $f_i^c$. For sparse approximation, we introduce another layer of latent variables, which we denote by $\mathbf{z}$ collectively. These are called the inducing variables. Other variables and functions associated with the sparse approximation are given a tilde $\sim$ accent. The asterisk subscript is used on $\mathbf{x}$, $\mathbf{y}$, $\mathbf{f}$ and $\mathbf{z}$ for two different purposes depending on the context: it is used to indicate a test input for predictive inference, and it is also used for a site under consideration for inclusion to the inducing set for sparse approximation.

We use $p$ to represent the probability density determined by the model and the data, including the case where the model involves sparsity. Any variational approximation to $p$ is denoted by $q$.

## 2. Model and Variational Inference

We recall the multinomial logit Gaussian process model (Williams and Barber, 1998) in Section 2.1. We add a simple generalization of the model to include the prior covariance between the latent functions. Bayesian inference with this model is outlined in Section 2.2; this is intractable. We provide variational bounds and approximate inference for the model in Section 2.3.

### 2.1 Model

For classifying or categorizing the $i$th input $\mathbf{x}_i$ into one of $C$ classes, we use a vector of $C$ indicator variables $\mathbf{y}_i \in \{\mathbf{e}^c\}$, wherein the $c$th entry, $y_i^c$, is one if $\mathbf{x}_i$ is in class $c$ and zero otherwise. We introduce $C$ latent functions, $f^1, \ldots, f^C$, on which we place a zero mean Gaussian process prior

$$\langle f^c(\mathbf{x}) f^{c'}(\mathbf{x}') \rangle = K^{\mathrm{c}}_{cc'} k^{\mathrm{x}}(\mathbf{x}, \mathbf{x}'), \tag{1}$$

where $K^{\mathrm{c}}_{cc'}$ is the $(c,c')$th entry of a $C$-by-$C$ positive semi-definite matrix $K^{\mathrm{c}}$ for modeling inter-function covariances, and $k^{\mathrm{x}}$ is a covariance function on the inputs. Let $f_i^c \overset{\text{def}}{=} f^c(\mathbf{x}_i)$. Given the vector of function values $\mathbf{f}_i \overset{\text{def}}{=} (f_i^1,\dots,f_i^C)^{\mathrm{T}}$ at $\mathbf{x}_i$, the likelihood for the class label is the multinomial logistic

$$p(y_i^c = 1|\mathbf{f}_i) \overset{\text{def}}{=} \frac{\exp f_i^c}{\sum_{c'=1}^C \exp f_i^{c'}}. \tag{2}$$

This can also be written as

$$p(\mathbf{y}_i|\mathbf{f}_i) = \frac{\exp \mathbf{f}_i^{\mathrm{T}} \mathbf{y}_i}{\sum_{c=1}^C \exp \mathbf{f}_i^{\mathrm{T}} \mathbf{e}^c}.$$

These two expressions for the likelihood function will be used interchangeably. We use the first expression when the interest on the class $c$ and the second when the interest is on $\mathbf{f}_i$.

The above model for the latent functions $f^c$s has been used previously for multi-task learning (Bonilla et al., 2008), where $f^c$ is the latent function for the $c$th task. Most prior works on multi-class Gaussian process (Williams and Barber, 1998; Seeger and Jordan, 2004; Kim and Ghahramani, 2006; Girolami and Rogers, 2006) have chosen $K^{\mathrm{c}}$ to be the $C$-by-$C$ identity matrix, so their latent functions are identical and independent. Williams and Barber (1998) have made this choice because the inter-function correlations are usually difficult to specify, although they have acknowledged that such correlations can be included in general. We agree with them on the difficulty, but we choose to address it by estimating $K^{\mathrm{c}}$ from observed data, as has been done for multi-task learning (Bonilla et al., 2008). If $K^{\mathrm{c}}$ is the identity matrix, then the block structure of the covariance matrix between the latent function values can be exploited to reduce computation (Seeger and Jordan, 2004).

The model in Equation 1 is known as the *separable model* for covariance. It is perhaps the simplest manner to involve inter-function correlations. One can also consider more involved models, such as those using convolution (Ver Hoef and Barry, 1998) and transformation (Lázaro-Gredilla and Figueiras-Vidal, 2009). Our presentation will mostly be general and applicable to these as well.

### 2.2 Exact Inference

Given a set of $n$ observations $\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^n$, we have an $nC$-vector $\mathbf{y}$ (resp. $\mathbf{f}$) of indicator variables (resp. latent function values) by stacking the $\mathbf{y}_i$s (resp. $\mathbf{f}_i$s). Let $X$ collects $\mathbf{x}_1,\dots,\mathbf{x}_n$. Dependencies on the inputs $X$ are suppressed henceforth unless necessary.

By Bayes' rule, the posterior over the latent function values is $p(\mathbf{f}|\mathbf{y}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f})/p(\mathbf{y})$, where $p(\mathbf{y}|\mathbf{f}) = \prod_i p(\mathbf{y}_i|\mathbf{f}_i)$ and $p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\mathrm{d}\mathbf{f}$. Inference for a test input $\mathbf{x}_*$ is performed in two steps. First we compute the distribution of latent function values at $\mathbf{x}_*$: $p(\mathbf{f}_*|\mathbf{y}) = \int p(\mathbf{f}_*|\mathbf{f})\,p(\mathbf{f}|\mathbf{y})\mathrm{d}\mathbf{f}$. Then we compute the posterior predictive probability of $\mathbf{x}_*$ being in class $c$, which is given by $p(y_*^c = 1|\mathbf{y}) = \int p(y_*^c = 1|\mathbf{f}_*)\,p(\mathbf{f}_*|\mathbf{y})\mathrm{d}\mathbf{f}_*$.

### 2.3 Variational Inference

The integrals needed in the exact inference steps are intractable due to the non-Gaussian likelihood $p(\mathbf{y}|\mathbf{f})$. To progress, we employ variational inference in the following manner. The posterior $p(\mathbf{f}|\mathbf{y})$ is approximated by the variational posterior $q(\mathbf{f}|\mathbf{y})$ by minimizing the Kullback-Leibler (KL) divergence

$$\mathrm{KL}\left(q(\mathbf{f}|\mathbf{y}) \,\|\, p(\mathbf{f}|\mathbf{y})\right) = \int q(\mathbf{f}|\mathbf{y}) \log \frac{q(\mathbf{f}|\mathbf{y})}{p(\mathbf{f}|\mathbf{y})} \mathrm{d}\mathbf{f}.$$

This is the difference between the log marginal likelihood $\log p(\mathbf{y})$ and a variational lower bound

$$\log Z_B = -\mathrm{KL}\left(q(\mathbf{f}|\mathbf{y}) \,\|\, p(\mathbf{f})\right) + \sum_{i=1}^{n} \ell_i(\mathbf{y}_i; q), \tag{3}$$

where

$$\ell_i(\mathbf{y}_i; q) \overset{\text{def}}{=} \int q(\mathbf{f}_i|\mathbf{y}) \log p(\mathbf{y}_i|\mathbf{f}_i)\, \mathrm{d}\mathbf{f}_i \tag{4}$$

is the expected log-likelihood of the $i$th datum under distribution $q$, and

$$q(\mathbf{f}_i|\mathbf{y}) = \int q(\mathbf{f}|\mathbf{y}) \prod_{j \neq i} \mathrm{d}\mathbf{f}_j \tag{5}$$

is the variational marginal distribution of $\mathbf{f}_i$; see Appendix B.1 for details. The Kullback-Leibler divergence component of $\log Z_B$ can be interpreted as the regularizing factor for the approximate posterior $q(\mathbf{f}|\mathbf{y})$, while the expected log-likelihood can be interpreted as the data fit component. The inequality $\log p(\mathbf{y}) \geq \log Z_B$ with $Z_B$ expressed as in Equation 3 has been given previously in the same context of variational inference for Gaussian latent models (Challis and Barber, 2011). It has also been used in the online learning setting (see Banerjee, 2006, and references therein).

For approximate inference on a test input $\mathbf{x}_*$, first we obtain the approximate posterior, which is $q(\mathbf{f}_*|\mathbf{y}) \overset{\text{def}}{=} \int p(\mathbf{f}_*|\mathbf{f})\, q(\mathbf{f}|\mathbf{y}) \mathrm{d}\mathbf{f}$. Then we obtain a lower bound to the approximate predictive probability for class $c$:

$$\begin{aligned} \log q(y_*^c = 1|\mathbf{y}) &\overset{\text{def}}{=} \log \int p(y_*^c = 1|\mathbf{f}_*)\, q(\mathbf{f}_*|\mathbf{y})\, \mathrm{d}\mathbf{f}_* \\ &\geq \int q(\mathbf{f}_*|\mathbf{y}) \log p(y_*^c = 1|\mathbf{f}_*)\, \mathrm{d}\mathbf{f}_* \\ &= \ell_*(y_*^c = 1; q), \end{aligned} \tag{6}$$

where the inequality is due to Jensen's inequality. The corresponding upper bound is obtained using the property of mutual exclusivity:

$$q(y_*^c = 1|\mathbf{y}) = 1 - \sum_{c' \neq c} q(y_*^{c'} = 1|\mathbf{y}) \leq 1 - \sum_{c' \neq c} \exp \ell_*(y_*^{c'} = 1; q). \tag{7}$$

The Bayes classification decision based on the upper bound is consistent with that based on the lower bound, since

$$\arg\max_c \left(1 - \sum_{c' \neq c} \exp \ell_*^{c'}\right) = \arg\max_c \left(1 - \sum_{c'=1}^{C} \exp \ell_*^{c'} + \exp \ell_*^c\right) = \arg\max_c \left(\exp \ell_*^c\right),$$

where we have written $\ell_*^c$ for $\ell_*(y_*^c = 1; q)$.

The variational inference procedure outlined here depends on the ability to compute expressions (a) $\mathrm{KL}\left(q(\mathbf{f}|\mathbf{y}) \,\|\, p(\mathbf{f})\right)$, (b) $q(\mathbf{f}_*|\mathbf{y})$ and (c) $\ell_i(\mathbf{y}_i; q)$. Expressions (a) and (b) can be made tractable by constraining $q(\mathbf{f}|\mathbf{y})$ to be a Gaussian density with mean $\mathbf{m}$ and covariance $V$, which are the variational parameters. For (c), we compute its lower bound instead, as detailed in the next section.

**Remark 1** *Approximate prediction using the approximate posterior as outlined above is the more common approach (see, for example, Rasmussen and Williams, 2006, §3.5). An alternative is to use $p(\mathbf{y}_*|\mathbf{y}) = p(\mathbf{y}_*, \mathbf{y})/p(\mathbf{y})$ directly. Lower bounds to the marginal likelihoods $p(\mathbf{y}_*, \mathbf{y})$ and $p(\mathbf{y})$ may replace the exact values if they are tight. However, this procedure is more expensive in general since an (approximate) marginal likelihood has to be computed for the training data together with the test data point for every test point.*

#### 2.3.1 VARIATIONAL BOUNDS FOR EXPECTED LOG-LIKELIHOOD

Equations 3 to 7 require the computation of the expected log-likelihood under $q(\mathbf{f}|\mathbf{y})$:

$$\ell(\mathbf{y};q) \overset{\text{def}}{=} \int q(\mathbf{f}|\mathbf{y}) \log p(\mathbf{y}|\mathbf{f})\, \mathrm{d}\mathbf{f}, \tag{8}$$

where we have suppressed the datum indices $i$ and $*$ here and henceforth for this section. In our setting, $q(\mathbf{f}|\mathbf{y})$ is a Gaussian density with mean $\mathbf{m}$ and covariance $V$, and we regard these parameters to be constant throughout this section. The subject of this section is lower bounds on $\ell(\mathbf{y};q)$. Two trivial lower bounds can be obtained by expanding $p(\mathbf{y}|\mathbf{f})$ and using the Jensen's inequality:

$$\ell(\mathbf{y};q) \geq \mathbf{m}^{\mathrm{T}}\mathbf{y} - \log \sum_{c=1}^{C} \exp\left[\mathbf{m}^{\mathrm{T}}\mathbf{e}^c + \frac{1}{2}(\mathbf{y}-\mathbf{e}^c)^{\mathrm{T}} V (\mathbf{y}-\mathbf{e}^c)\right], \tag{9}$$

$$\ell(\mathbf{y};q) \geq \mathbf{m}^{\mathrm{T}}\mathbf{y} - \log \sum_{c=1}^{C} \exp\left[\mathbf{m}^{\mathrm{T}}\mathbf{e}^c + \frac{1}{2}(\mathbf{e}^c)^{\mathrm{T}} V \mathbf{e}^c\right]. \tag{10}$$

These bounds can be very loose. In this section, we give a variational lower bound, and we have found this bound to be quite tight when the variational parameters are optimized. This bound exploits that if a prior $r(\mathbf{f})$ is a mixture of $C$ Gaussians with a particular set of parameters, then the corresponding posterior under the multinomial logistic likelihood is a $C$-variate Gaussian. We introduce this bound in terms of probability distributions and then express it in terms of variational parameters.

**Lemma 2** *Let $r(\mathbf{f}|\mathbf{y})$ be a $C$-variate Gaussian density with mean $\mathbf{a}$ and precision $W$, and let $\mathbf{a}^c$ be such that $W\mathbf{a}^c = W\mathbf{a} + \mathbf{e}^c - \mathbf{y}$. If $r(\mathbf{f}) = \sum_{c=1}^{C} \gamma^c r^c(\mathbf{f})$ is the mixture of $C$ Gaussians model on $\mathbf{f}$ with mixture proportions and components*

$$\gamma^c \overset{\text{def}}{=} \frac{\exp\left[\frac{1}{2}(\mathbf{a}^c)^{\mathrm{T}} W \mathbf{a}^c\right]}{\sum_{c'} \exp\left[\frac{1}{2}(\mathbf{a}^{c'})^{\mathrm{T}} W \mathbf{a}^{c'}\right]}, \qquad r^c(\mathbf{f}) \overset{\text{def}}{=} \frac{|W|^{1/2}}{(2\pi)^{C/2}} \exp\left[-\frac{1}{2}(\mathbf{f}-\mathbf{a}^c)^{\mathrm{T}} W (\mathbf{f}-\mathbf{a}^c)\right],$$

*and if*

$$r(\mathbf{y}) = \frac{\exp\left[\frac{1}{2}\mathbf{a}^{\mathrm{T}} W \mathbf{a}\right]}{\sum_{c=1}^{C} \exp\left[\frac{1}{2}(\mathbf{a}^c)^{\mathrm{T}} W \mathbf{a}^c\right]}, \tag{11}$$

*then*

$$\ell(\mathbf{y};q) \geq h(\mathbf{y};q,r) \overset{\text{def}}{=} \int q(\mathbf{f}|\mathbf{y}) \log r(\mathbf{f}|\mathbf{y}) \mathrm{d}\mathbf{f} + \log r(\mathbf{y}) - \log \sum_{c=1}^{C} \gamma^c \int q(\mathbf{f}|\mathbf{y}) r^c(\mathbf{f}) \mathrm{d}\mathbf{f}. \tag{12}$$

**Proof** The choice of notation used in the lemma will be clear from its proof. We begin with a variational posterior distribution $r(\mathbf{f}|\mathbf{y})$. Denote by $r(\mathbf{f})$ the corresponding prior distribution that gives this posterior when combined with the exact data likelihood $p(\mathbf{y}|\mathbf{f})$; that is

$$r(\mathbf{f}|\mathbf{y}) = p(\mathbf{y}|\mathbf{f})\, r(\mathbf{f}) / r(\mathbf{y}), \qquad \text{where} \qquad r(\mathbf{y}) \overset{\text{def}}{=} \int p(\mathbf{y}|\mathbf{f})\, r(\mathbf{f}) \mathrm{d}\mathbf{f}.$$

Rearranging for $p(\mathbf{y}|\mathbf{f})$ and putting back into $\ell(\mathbf{y};q)$ defined by (8) gives

$$\ell(\mathbf{y};q) = \int q(\mathbf{f}|\mathbf{y}) \log r(\mathbf{f}|\mathbf{y}) \mathrm{d}\mathbf{f} + \log r(\mathbf{y}) - \int q(\mathbf{f}|\mathbf{y}) \log r(\mathbf{f}) \mathrm{d}\mathbf{f}. \tag{13}$$

This is valid for any choice of distribution $r(\mathbf{f}|\mathbf{y})$, but let us choose it to be a $C$-variate Gaussian density with mean $\mathbf{a}$ and precision $W$. After some algebraic manipulation detailed in Appendix B.2, we obtain the expressions for $r(\mathbf{f})$ and $r(\mathbf{y})$ given in the lemma. We proceed with Jensen's inequality to move the logarithm outside the integral for the last term on the right of (13). This leads to the lower bound (12). ■

**Remark 3** *The first two terms in the expression for the expected log-likelihood $\ell(\mathbf{y};q)$ given by* (13) *are computable, since $r(\mathbf{f}|\mathbf{y})$ is Gaussian by definition, and $r(\mathbf{y})$ is given in* (11)*; however, the third term remains intractable since $r(\mathbf{f})$ is a mixture of Gaussians. Hence the additional step of using the Jensen's inequality is required to obtain the lower bound $h(\mathbf{y};q,r)$ in* (12) *that is computable.*

**Remark 4** *Lemma 2 depends only on the multinomial logistic likelihood function. It does not depend on the distribution $q(\mathbf{f}|\mathbf{y})$. In particular, $q(\mathbf{f}|\mathbf{y})$ can be non-Gaussian.*

**Lemma 5** *Let $W$ be a C-by-C positive semi-definite matrix, and let $\mathbf{a} \in \mathbb{R}^C$. Define $S \overset{\text{def}}{=} V^{-1} + W$, $\mathbf{b} \overset{\text{def}}{=} W(\mathbf{m}-\mathbf{a}) + \mathbf{y}$, and*

$$g^c(\mathbf{y};q,\mathbf{a},W) \overset{\text{def}}{=} \exp\left[\mathbf{m}^{\mathrm{T}}\mathbf{e}^c + \frac{1}{2}(\mathbf{b}-\mathbf{e}^c)^{\mathrm{T}}S^{-1}(\mathbf{b}-\mathbf{e}^c)\right]. \tag{14}$$

*Then*

$$\ell(\mathbf{y};q) \geq h(\mathbf{y};q,\mathbf{a},W) = \frac{C}{2} + \frac{1}{2}\log|SV| - \frac{1}{2}\operatorname{tr}SV + \mathbf{m}^{\mathrm{T}}\mathbf{y} - \log\sum_{c=1}^{C} g^c(\mathbf{y};q,\mathbf{a},W). \tag{15}$$

**Proof** This follows from Lemma 2 by expressing $h(\mathbf{y};q,r)$ in terms of parameters $W$ and $\mathbf{a}$; the derivation is in Appendix B.3. Matrix $W$ is allowed to be singular because our derivation does not involve the inversion of $W$; and the determinants of $W$ taken in $r(\mathbf{f}|\mathbf{y})$ and $r^c(\mathbf{f})$ directly cancel out by subtraction, so continuity arguments can be applied. ■

We can view $h$ given in (15) as parameterized either by $W$ and $\mathbf{a}$ or by $S$ and $\mathbf{b}$. For the latter view, the definitions of $S$ and $\mathbf{b}$ constrain their values. Therefore, the following seem necessary from the onset in order for the bound to be valid.

- $S \succeq V^{-1}$ so that $W$ is well-defined.
- If $W$ is rank-deficient, then $\mathbf{b}$ lies on the hyperplane passing through $\mathbf{y}$ and in the column space of $W$.

However, further analysis will show these constraints to be unnecessary for $h$ to be a lower bound. Consequently, we can view $h$ as a function of the pair $(\mathbf{b},S)$, regardless of there being a pair $(\mathbf{a},W)$ mapping to $(\mathbf{b},S)$. Before proceeding to the formal theorem, a few notations are necessary. Let

$$g^c(q,\mathbf{b},S) \overset{\text{def}}{=} \exp\left[\mathbf{m}^{\mathrm{T}}\mathbf{e}^c + \frac{1}{2}(\mathbf{b}-\mathbf{e}^c)^{\mathrm{T}}S^{-1}(\mathbf{b}-\mathbf{e}^c)\right] \tag{16}$$

be a function of the mean $\mathbf{m}$ of distribution $q$ and of $\mathbf{b}$ and $S \succ 0$. When the context is clear, we will suppress the parameters of $g^c$ for conciseness. Let

$$\bar{g}^c \overset{\text{def}}{=} g^c / \textstyle\sum_{c'=1}^{C} g^{c'}, \qquad\qquad \bar{\mathbf{g}} \overset{\text{def}}{=} (\bar{g}^1, \ldots, \bar{g}^C)^{\mathrm{T}}, \tag{17}$$

and let $\bar{G}$ be the diagonal matrix with $\bar{\mathbf{g}}$ along its diagonal. We further define

$$A \overset{\text{def}}{=} \textstyle\sum_{c=1}^{C} \bar{g}^c (\mathbf{b} - \mathbf{e}^c)(\mathbf{b} - \mathbf{e}^c)^{\mathrm{T}} = \mathbf{b}\mathbf{b}^{\mathrm{T}} - \mathbf{b}\bar{\mathbf{g}}^{\mathrm{T}} - \bar{\mathbf{g}}\mathbf{b}^{\mathrm{T}} + \bar{G} \succeq 0. \tag{18}$$

Matrix $A$ given above is a convex combination of $C$ positive semi-definite matrices of ranks one, so $A$ is positive semi-definite. Furthermore, $A \neq 0$. We will also suppress the dependency of $A$ on $\mathbf{m}$, $\mathbf{b}$ and $S$ for conciseness.

The lemmas necessary for the proof of the following theorem are in Appendix B.4.

**Theorem 6** *Let $S$ be a $C$-by-$C$ positive definite matrix, and let $\mathbf{b} \in \mathbb{R}^C$. Let*

$$h(\mathbf{y}; q, \mathbf{b}, S) \overset{\text{def}}{=} \frac{C}{2} + \frac{1}{2}\log|SV| - \frac{1}{2}\operatorname{tr} SV + \mathbf{m}^{\mathrm{T}}\mathbf{y} - \log\sum_{c=1}^{C} g^c(q, \mathbf{b}, S) \tag{19}$$

*be a function of $\mathbf{b}$ and $S$, where $g^c(q, \mathbf{b}, S)$ is given by* (16). *Then $\ell(\mathbf{y}; q) \geq h(\mathbf{y}; q, \mathbf{b}, S)$.*

**Proof** Let $(\mathbf{b}^*, \mathbf{S}^*) \overset{\text{def}}{=} \arg\max_{(\mathbf{b},S)} h(\mathbf{y}; q, \mathbf{b}, S)$. The joint concavity of $h$ in $\mathbf{b}$ and $S$ (Lemma 25) implies $h(\mathbf{y}; q, \mathbf{b}^*, S^*) > h(\mathbf{y}; q, \mathbf{b}, S)$ for any $\mathbf{b} \neq \mathbf{b}^*$ and $S \neq S^*$. Thus we only need to prove $\ell(\mathbf{y}; q) \geq h(\mathbf{y}; q, \mathbf{b}^*, S^*)$. Now, if there exists a pair $(\mathbf{a}^*, W^*)$ with $W^* \succeq 0$ such that $S^* = V^{-1} + W^*$ and $\mathbf{b}^* = W^*(\mathbf{m} - \mathbf{a}^*) + \mathbf{y}$, then the application of Lemma 5 completes the proof. To find such a pair, we first set $S^*$ and $W^*$ to the $S^{\mathrm{fx}}$ and the $W^{\mathrm{fx}}$ given by Lemma 28, then we show below that there exists an $\mathbf{a}^*$ under this setting.

Let $\bar{\mathbf{g}}^* \overset{\text{def}}{=} \bar{\mathbf{g}}(q, \mathbf{b}^*, S^*)$ and $\bar{G}^*$ be the diagonal matrix with $\bar{\mathbf{g}}^*$ along its diagonal. By Lemma 26, $\mathbf{b}^* = \bar{\mathbf{g}}^*$, so matrix $A$ simplifies to $A^*$ given by $A^* \overset{\text{def}}{=} \bar{G}^* - \bar{\mathbf{g}}^*(\bar{\mathbf{g}}^*)^{\mathrm{T}}$. Since $\bar{\mathbf{g}}^*$ is a probability vector, matrix $A^*$ is the covariance matrix of a multinomial distribution. The entries in $\bar{\mathbf{g}}^*$ are non-zero, so matrix $A^*$ is of rank $(C-1)$, and an eigenpair of $A^*$ is $(0, \mathbf{1})$ (see Watson, 1996). In other words, $\operatorname{null}(A^{**}) = \{\eta\mathbf{1} \mid \eta \in \mathbb{R}\}$. Using Lemma 28, we also have $\operatorname{null}(W^*) = \{\eta\mathbf{1} \mid \eta \in \mathbb{R}\}$. Since $(\mathbf{b}^* - \mathbf{y})^{\mathrm{T}}\mathbf{1} = 1 - 1 = 0$, we have $(\mathbf{b}^* - \mathbf{y}) \notin \operatorname{null}(W^*)$, unless $(\mathbf{b}^* - \mathbf{y}) = \mathbf{0}$. Equivalently, $(\mathbf{b}^* - \mathbf{y})$ is in the row space of $W^*$. Hence, there exists a vector $\mathbf{v}$ such that $W^*\mathbf{v} = \mathbf{b}^* - \mathbf{y}$. We let $\mathbf{a}^* \overset{\text{def}}{=} \mathbf{m} - \mathbf{v}$ to complete the proof. ■

There are two properties that $W^*$ obeys: $\operatorname{null}(W^*) = \{\eta\mathbf{1} \mid \eta \in \mathbb{R}\}$ and $W^* \succeq 0$. One parametrization of $W$ that always satisfies these properties is

$$W \overset{\text{def}}{=} M - M\mathbf{1}\mathbf{1}^{\mathrm{T}}M / \mathbf{1}^{\mathrm{T}}M\mathbf{1}, \qquad\qquad \text{where } M \succ 0. \tag{20}$$

The proof for the null space is straightforward, while the proof for positive definiteness is an application of Theorem 7.7.7(a) by Horn and Johnson (1985). If $M$ is a diagonal positive definite matrix, then the parametrization proposed by Seeger and Jordan (2004) is obtained. Further constraining the diagonal to sum to one gives the parametrization resultant from the Laplace approximation (Williams and Barber, 1998; Rasmussen and Williams, 2006). A diagonal $M$ is appealing because it entails that $W$ is the covariance of the multinomial or the Dirichlet distribution, which matches the likelihood. However, our experience has shown that a diagonal $M$ is far from optimum for our bounds. Therefore, we shall let $W$ vary freely but be subjected directly to the two properties stated at the beginning of this paragraph. There are two reasons for the non-optimality. First, the variational prior $r(\mathbf{f})$ in Lemma 2 is a mixture of Gaussian distributions and not a Dirichlet distribution. Second, the use of Jensen's inequality in Lemma 5 weaken the interpretation of $W$ as the covariance of the variational posterior $r(\mathbf{f}|\mathbf{y})$. Nonetheless, since the null space of $W^*$ is the line $\{\eta\mathbf{1} \mid \eta \in \mathbb{R}\}$, the optimized variational posterior satisfies the invariance $r(\mathbf{f}|\mathbf{y}) = r(\mathbf{f}+\eta\mathbf{1}|\mathbf{y})$, $\eta \in \mathbb{R}$. This is a pleasant property because the likelihood satisfies the same invariance: $p(\mathbf{y}|\mathbf{f}) = p(\mathbf{y}|\mathbf{f}+\eta\mathbf{1})$.

The significance of Theorem 6 over Lemma 5 is in the practical aspects of variational inference:

1. Maximizing $h$ with respect to $V$ does not involve the function $g^c$.

2. A block coordinate approach to optimization can be used, since we can optimize with respect to $V$ and to $S$ alternately, without ensuring $S \succeq V^{-1}$ when optimizing for $V$.

3. The vector $\mathbf{y}$ of observed classifications does not appear in the definition of $g^c$ given by Equation 16, in contrast to Equation 14.

Let us emphasis the second point listed above. In place of definitions (16) and (19) for functions $g^c$ and $h$, suppose we had used

$$g^c(\mathbf{b}',S) \overset{\text{def}}{=} \exp\frac{1}{2}(\mathbf{b}'-\mathbf{e}^c)^{\mathrm{T}}S^{-1}(\mathbf{b}'-\mathbf{e}^c),$$

$$h(\mathbf{y};q,\mathbf{b},S) \overset{\text{def}}{=} \frac{C}{2}+\frac{1}{2}\log|SV|-\frac{1}{2}\operatorname{tr}S(V+\mathbf{m}\mathbf{m}^{\mathrm{T}})+\mathbf{m}^{\mathrm{T}}(\mathbf{y}-\mathbf{b}')-\log\sum_{c=1}^{C}g^c(q,\mathbf{b}',S)$$

as functions of $\mathbf{b}'$ and $S \succ 0$. This is obtained from Lemma 2 by substituting in $S \overset{\text{def}}{=} V^{-1}+W$ and $\mathbf{b}' \overset{\text{def}}{=} -V^{-1}\mathbf{m}-W\mathbf{a}+\mathbf{y}$. This formulation of $h$ is jointly concave in $\mathbf{b}'$ and $S$, so there should be no computation difficulties in optimization. Unfortunately, this formulation does not guarantee $S \succeq V^{-1}$ when the optimization is done without constraints. This is in contrast with the formulation in Theorem 6, for which validity is guaranteed by Lemma 28.

The bound $h$ as defined in Theorem 6 is maximized by finding the stationary points with respect to variational parameters $\mathbf{b}$ and $S$. Computation can be reduced when the bound is relaxed through fixing or constraining these parameters. Two choices for $S$ are convenient: $I$ and $V^{-1}$. Fixing $S$ to $V^{-1}$ is expected to be a better choice since its optimal value is between $V^{-1}$ and $V^{-1}+A$ (Lemma 27). This gives the relaxed bound

$$h(\mathbf{y};q,\mathbf{b},V^{-1}) = \mathbf{m}^{\mathrm{T}}\mathbf{y}-\log\sum_{c=1}^{C}\exp\left[\mathbf{m}^{\mathrm{T}}\mathbf{e}^c+\frac{1}{2}(\mathbf{b}-\mathbf{e}^c)^{\mathrm{T}}V(\mathbf{b}-\mathbf{e}^c)\right].$$

For the case where $q$ is non-correlated Gaussians, that is, where $V$ is a diagonal matrix, we obtain the bound that has been proposed for variational message passing (Knowles and Minka, 2011, Equation 12). We can also choose to fix $\mathbf{b}$ to $\mathbf{y}$, giving

$$h(\mathbf{y};q,\mathbf{y},V^{-1}) = \mathbf{m}^{\mathrm{T}}\mathbf{y}-\log\sum_{c=1}^{C}\exp\left[\mathbf{m}^{\mathrm{T}}\mathbf{e}^c+\frac{1}{2}(\mathbf{y}-\mathbf{e}^c)^{\mathrm{T}}V(\mathbf{y}-\mathbf{e}^c)\right].$$

This is the bound (9) obtained using Jensen's inequality directly. Setting $\mathbf{b}$ to $\mathbf{0}$ instead of $\mathbf{y}$ gives

$$h(\mathbf{y};q,\mathbf{0},V^{-1}) = \mathbf{m}^{\mathrm{T}}\mathbf{y}-\log\sum_{c=1}^{C}\exp\left[\mathbf{m}^{\mathrm{T}}\mathbf{e}^c+\frac{1}{2}(\mathbf{e}^c)^{\mathrm{T}}V\mathbf{e}^c\right],$$

which is the bound (10) also obtained using Jensen's inequality directly. Therefore, the bound $\max_{(\mathbf{b},S)} h(\mathbf{y};q,\mathbf{b},S)$ is provably at least as tight as the Jensen's inequality bounds. Other choices for $S$ and $\mathbf{b}$ give different lower bounds on $\max_{(\mathbf{b},S)} h(\mathbf{y};q,\mathbf{b},S)$.

Thus far we have delved into lower bounds for $\ell(\mathbf{y};q)$ defined by Equation 8. Of independent interest is the following upper bound that is proved in Appendix B.5:

**Lemma 7** $\ell(\mathbf{y};q) \leq \log p(\mathbf{y}|\mathbf{m}) \overset{\text{def}}{=} \mathbf{m}^{\mathrm{T}}\mathbf{y} - \log\sum_{c=1}^{C} \exp \mathbf{m}^{\mathrm{T}}\mathbf{e}^{c}$.

### 2.3.2 Variational Bounds for Marginal Likelihood

To consolidate, the log marginal likelihood is lower bounded via the sequence

$$\log p(\mathbf{y}) \geq \log Z_B \geq \log Z_h \overset{\text{def}}{=} -\mathrm{KL}(q(\mathbf{f}|\mathbf{y}) \,\|\, p(\mathbf{f})) + \sum_{i=1}^{n} h(\mathbf{y}_i;q_i,\mathbf{b}_i,S_i),$$

where the datum subscript $i$ is reintroduced. The aim is to optimize the last lower bound. Recall that $\mathbf{m}$ and $V$ are the mean and covariance of the variational posterior $q(\mathbf{f}|\mathbf{y})$. Also recall that the prior distribution on $\mathbf{f}$ is given by the Gaussian process prior stated in Section 2.1, so $\mathbf{f}$ has zero mean and covariance $K \overset{\text{def}}{=} K^{\mathrm{x}} \otimes K^{\mathrm{c}}$, where $K^{\mathrm{x}}$ is the $n$-by-$n$ matrix of covariances between the inputs $\mathbf{x}_1,\ldots,\mathbf{x}_n$. Using arguments similar to those used in proving Lemma 25, one can show that $\log Z_h$ is jointly concave in $\mathbf{m}$, $V$, $\{\mathbf{b}_i\}$ and $\{S_i\}$. We highlight this with the following proposition, where $\log Z_h$ is expressed explicitly in the variational parameters.

**Proposition 8** *Let $V$ be an $nC$-by-$nC$ positive definite matrix and let $\mathbf{m} \in \mathbb{R}^{nC}$. For $i = 1,\ldots n$, let $S_i$ be a $C$-by-$C$ positive definite matrix and let $\mathbf{b}_i \in \mathbb{R}^{C}$. Let*

$$\begin{aligned}\log Z_h = nC &+ \frac{1}{2}\log|K^{-1}V| - \frac{1}{2}\operatorname{tr} K^{-1}V - \frac{1}{2}\mathbf{m}^{\mathrm{T}}K^{-1}\mathbf{m} + \mathbf{m}^{\mathrm{T}}\mathbf{y} \\ &+ \frac{1}{2}\sum_{i=1}^{n}\Big(\log|S_iV_i| - \operatorname{tr} S_iV_i\Big) - \sum_{i=1}^{n}\log\sum_{c=1}^{C}\exp\left[\mathbf{m}_i^{\mathrm{T}}\mathbf{e}^{c} + \frac{1}{2}(\mathbf{b}_i - \mathbf{e}^{c})^{\mathrm{T}}S_i^{-1}(\mathbf{b}_i - \mathbf{e}^{c})\right], \end{aligned} \quad (21)$$

*where $V_i$ is the* ith *$C$-by-$C$ diagonal block of $V$, and $\mathbf{m}_i$ is the* ith *$C$-vector of $\mathbf{m}$. Then $\log Z_h$ is jointly concave in $\mathbf{m}$, $V$, $\{\mathbf{b}_i\}$ and $\{S_i\}$, and $\log p(\mathbf{y}) \geq \log Z_h$.*

Suitable choices of the variational parameters leads to the following two theorems that are proved in Appendix B.6.

**Theorem 9** *For a multinomial logit Gaussian process model where the latent process has zero mean and the covariance function induces the Gram matrix $K$, the average log-marginal-likelihood satisfies*

$$\begin{aligned}\frac{1}{n}\log p(\mathbf{y}) \geq\ & \frac{C}{2} + \frac{C}{2}\log\sigma_{\mathrm{v}}^2 - \frac{1}{2n}\log|K| - \frac{\sigma_{\mathrm{v}}^2}{2n}\operatorname{tr} K^{-1} \\ & - \frac{C-1}{2}\left[2\sqrt{\frac{\sigma_{\mathrm{v}}^2}{C} + \frac{1}{4}} - \log\left(\sqrt{\frac{\sigma_{\mathrm{v}}^2}{C} + \frac{1}{4}} + \frac{1}{2}\right) - 1\right] - \log C \\ >\ & \frac{C}{2} + \frac{C}{2}\log\sigma_{\mathrm{v}}^2 - \frac{1}{2n}\log|K| - \frac{\sigma_{\mathrm{v}}^2}{2n}\operatorname{tr} K^{-1} - \frac{\sigma_{\mathrm{v}}^2}{2} - \log C\end{aligned}$$

*for every $\sigma_{\mathrm{v}}^2 > 0$.*

**Theorem 10** *For a multinomial logit Gaussian process model where the latent process has zero mean, the covariance function is $k((\mathbf{x},c),(\mathbf{x}',c')) = \sigma^2\delta(c,c')k^{\mathrm{x}}(\mathbf{x},\mathbf{x}')$ and $k^{\mathrm{x}}$ is a correlation function, that is, $k^{\mathrm{x}}(\mathbf{x},\mathbf{x}) = 1$, the average log-marginal-likelihood satisfies*

$$\begin{aligned}\frac{1}{n}\log p(\mathbf{y}) &\geq -\frac{C-1}{2}\left[2\sqrt{\frac{\sigma^2}{C}+\frac{1}{4}} - \log\left(\sqrt{\frac{\sigma^2}{C}+\frac{1}{4}}+\frac{1}{2}\right) - 1\right] - \log C \\ &> -\sigma^2/2 - \log C.\end{aligned}$$

The bounds in the theorems do not dependent on the observed classes $\mathbf{y}$ because they have been "zeroed-out" by setting $\mathbf{m} = \mathbf{0}$. For the setting in Theorem 10, the lower bound in Theorem 10 is always tighter than that in Theorem 9 because the first four terms within the latter is the negative of a Kullback-Leibler divergence, which is always less than zero. One may imagine that this bound is rather loose. However, we will show in experiments in Section 9.1 that even this is better than the optimized variational mean-field lower bound (Girolami and Rogers, 2006).

**Remark 11** *Theorem 10 is consistent with and generalizes the calculations previously obtained for binary classification and in certain limits of the length-scales of the model (Nickisch and Rasmussen, 2008, Appendix B). Our result is also more general because it includes the latent scale $\sigma^2$ of the model.*

### 2.3.3 PREDICTIVE DENSITY: APPROXIMATION AND BOUNDS

According to the Gaussian process prior model specified in Section 2.1, the $C$ latent function values $\mathbf{f}_*$ of a test input $\mathbf{x}_*$ and the latent function values of the $n$ observed data have prior

$$\begin{pmatrix}\mathbf{f}\\ \mathbf{f}_*\end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} K & K_* \\ K_*^{\mathrm{T}} & K_{**}\end{pmatrix}\right),$$

where $K_* \stackrel{\text{def}}{=} \mathbf{k}_*^{\mathrm{x}} \otimes K^{\mathrm{c}}$, $K_{**} \stackrel{\text{def}}{=} k^{\mathrm{x}}(\mathbf{x}_*,\mathbf{x}_*)K^{\mathrm{c}}$, and $\mathbf{k}_*^{\mathrm{x}}$ is the vector of covariances between the observed inputs $X$ and the test input $\mathbf{x}_*$. After the variational posterior $q(\mathbf{f}|\mathbf{y}) = \mathcal{N}(\mathbf{f}|\mathbf{m},V)$ has been obtained by maximizing the lower bound $\log Z_h$ in Proposition 8, we can obtain the approximate posterior at the test input $\mathbf{x}_*$:

$$q(\mathbf{f}_*|\mathbf{y}) \stackrel{\text{def}}{=} \int p(\mathbf{f}_*|\mathbf{f})q(\mathbf{f}|\mathbf{y})\mathrm{d}\mathbf{f} = \mathcal{N}(\mathbf{f}_*|\mathbf{m}_*,V_*),$$

where $\mathbf{m}_* \stackrel{\text{def}}{=} K_*^{\mathrm{T}}K^{-1}\mathbf{m}$ and $V_* \stackrel{\text{def}}{=} K_{**} - K_*^{\mathrm{T}}K^{-1}K_* + K_*^{\mathrm{T}}K^{-1}VK^{-1}K_*$.

The approximation to the posterior predictive density of $\mathbf{y}_*$ at $\mathbf{x}_*$ is

$$\begin{aligned}\log p(y_*^c = 1|\mathbf{y}) \approx \log q(y_*^c = 1|\mathbf{y}) &\stackrel{\text{def}}{=} \log\int p(y_*^c = 1|\mathbf{f}_*)\, q(\mathbf{f}_*|\mathbf{y})\,\mathrm{d}\mathbf{f}_* &(22)\\ &\geq \ell_*(y_*^c = 1; q) \\ &\geq \max_{\mathbf{b}_*,S_*} h(\mathbf{e}^c; q_*, \mathbf{b}_*, S_*), &(23)\end{aligned}$$

where $\ell_*(y_*^c = 1;q) = \int q(\mathbf{f}_*|\mathbf{y})\log p(y_*^c = 1|\mathbf{f}_*)\,\mathrm{d}\mathbf{f}_*$, and $q_*$ in the last expression refers to $q(\mathbf{f}_*|\mathbf{y})$. Expanding $h$ using definitions (16) and (19) gives

$$\log p(y_*^c \,=\, 1|\mathbf{y}) \;\gtrsim\; \frac{C}{2} + \mathbf{m}_*^{\mathrm{T}}\mathbf{e}^c + \max_{\mathbf{b}_*,S_*}\left(\frac{1}{2}\log|S_*V_*| - \frac{1}{2}\operatorname{tr}S_*V_* - \log\sum_{c'=1}^{C} g^{c'}(q_*,\mathbf{b}_*,S_*)\right), \quad (24)$$

where the probed class $\mathbf{e}^c$ is used only outside the max operator. Hence maximization needs to be only done once instead of $C$ times. Moreover, if one is interested only in the classification decision, then one may simply compare the *re-normalized probabilities*

$$\tilde{p}(y_*^c = 1|\mathbf{y}) = p(y_*^c = 1|\mathbf{m}_*) \stackrel{\text{def}}{=} \frac{\exp(m_*^c)}{\sum_{c'=1}^C \exp(m_*^{c'})}. \tag{25}$$

In this case, no maximization is required, and class prediction is faster. The faster prediction is possible because we have used the lower bound (23) for making classification decisions. These classification decisions do not match those given by $q(y_*^c|\mathbf{y})$ in general (Rasmussen and Williams, 2006, Section 3.5 and Exercise 3.10.3). In addition to the normalization across the $C$ classes, the predictive probability $\tilde{p}(y_*^c = 1|\mathbf{y})$ is also an upper bound on $\exp \ell_*(y_*^c = 1; q)$ because of Lemma 7.

The relation in Equation 24 is an approximate inequality ($\gtrsim$) instead of a proper inequality ($\geq$) due to the approximation to $\log q(y_*^c = 1|\mathbf{y})$ in Equation 22. As far as we are aware, this approximation is currently used throughout the literature for Gaussian process classification (Rasmussen and Williams 2006, Equations 3.25, 3.40 & 3.41 and 3.62; Nickisch and Rasmussen 2008, Equation 16). In order to obtain a proper inequality, we will show that the Kullback-Leibler divergence from the approximate posterior to the true posterior has to be accounted for.

First, we generalize and consider a set of $n_*$ test inputs $X_* \stackrel{\text{def}}{=} \{\mathbf{x}_{*1}, \ldots, \mathbf{x}_{*n_*}\}$. The following theorem, which give proper lower bounds, is proved in Appendix B.7.

**Theorem 12** *The log joint predictive probability for $\mathbf{x}_{*j}$ to be in class $c_j$ ($j = 1 \ldots n_*$) has lower bounds*

$$\begin{aligned}
\log p(\{y_{*j}^{c_j} = 1\}_{j=1}^{n_*}|\mathbf{y}) &\geq \sum_{j=1}^{n_*} \int q(\mathbf{f}_{*j}|\mathbf{y}) \log p(y_{*j}^{c_j} = 1|\mathbf{f}_{*j})\, d\mathbf{f}_{*j} - \mathrm{KL}(q(\mathbf{f}|\mathbf{y}) \| p(\mathbf{f}|\mathbf{y})) \\
&\geq \sum_{j=1}^{n_*} \max_{\mathbf{b}_{*j}, S_{*j}} h(\mathbf{e}^{c_j}; q_{*j}, \mathbf{b}_{*j}, S_{*j}) + \log Z_B - \log p(\mathbf{y}) \\
&\geq \sum_{j=1}^{n_*} \max_{\mathbf{b}_{*j}, S_{*j}} h(\mathbf{e}^{c_j}; q_{*j}, \mathbf{b}_{*j}, S_{*j}) + \log Z_h - \log p(\mathbf{y}).
\end{aligned}$$

In the first bound, the computation of the Kullback-Leibler divergence is intractable, but it is precisely this quantity that we have sought to minimize in the beginning, in Section 2.3. This implies that this divergence is a correct quantity to minimize in order to tighten the lower bound on the predictive probabilities. For one test input $\mathbf{x}_*$,

$$\log p(y_*^c = 1|\mathbf{y}) \geq \max_{\mathbf{b}_*, S_*} h(\mathbf{e}^c; q_*, \mathbf{b}_*, S_*) + \log Z_h - \log p(\mathbf{y}).$$

Because $\log Z_B$, $\log Z_h$ and $\log p(\mathbf{y})$ are independent of the probed class $\mathbf{e}^c$ at $\mathbf{x}_*$, the classification decision and the re-normalized probabilities (25) are also based on a true lower bound to the predictive probability.

Dividing the last bound in Theorem 12 by $n_*$ gives

$$\frac{1}{n_*} \log p(\{y_{*j}^{c_j} = 1\}_{j=1}^{n_*}|\mathbf{y}) \geq \frac{1}{n_*} \sum_{j=1}^{n_*} \max_{\mathbf{b}_{*j}, S_{*j}} h(\mathbf{e}^{c_j}; q_{*j}, \mathbf{b}_{*j}, S_{*j}) + \frac{1}{n_*} \Big[ \log Z_h - \log p(\mathbf{y}) \Big].$$

The term $\log Z_h - \log p(\mathbf{y})$ is a constant independent of $n_*$, so the last term diminishes when $n_*$ is large. In contrast the other two terms on either side of the inequality remain significant because each of them is the sum of $n_*$ summands. Hence, for large $n_*$, the last term can practically be ignored to give a computable lower bound on the average log predictive probability.

## 3. Variational Bound Optimization

To optimize the lower bound $\log Z_h$ during learning, we choose a block coordinate approach, where we optimize with respect to the variational parameters $\{\mathbf{b}_i\}$, $\{S_i\}$, $\mathbf{m}$ and $V$ in turn. For prediction, we only need to optimize $h$ with respect to the variational parameters $\mathbf{b}_*$ and $S_*$ for the test input $\mathbf{x}_*$.

### 3.1 Parameter $\mathbf{b}_i$

Parameters $\mathbf{b}_i$ and $S_i$ are contained within $h(\mathbf{y}_i; q_i, \mathbf{b}_i, S_i)$, so we only need to consider this function. For clarity, we suppress the datum subscript $i$ and the parameters for $h$ and $g^c$. The partial gradient with respect to $\mathbf{b}$ is $-S^{-1}(\mathbf{b}-\bar{\mathbf{g}})$, where $\bar{\mathbf{g}}$ is defined by Equation 17. Setting the gradient to zero gives the fixed-point update $\mathbf{b}^{\text{fx}} = \bar{\mathbf{g}}$, where $\bar{\mathbf{g}}$ is evaluated at the previous value of $\mathbf{b}$. This says that the optimal value $\mathbf{b}^*$ lies on the $C$-simplex, so a sensible initialization for $\mathbf{b}$ is a point therein. When the fixed-point update does not improve the lower bound $h$, we use to the Newton-Raphson update, which incorporates the Hessian

$$\frac{\partial^2 h}{\partial \mathbf{b}\, \partial \mathbf{b}^{\mathrm{T}}} = -S^{-1} - S^{-1}\left(\bar{G} - \bar{\mathbf{g}}\bar{\mathbf{g}}^{\mathrm{T}}\right) S^{-1},$$

where $\bar{G}$ is the diagonal matrix with $\bar{\mathbf{g}}$ along its diagonal. The Hessian is negative semi-definite, which is another proof that $h$ is a concave function of $\mathbf{b}$; see Lemma 25. The update is

$$\mathbf{b}^{\text{NR}} = \mathbf{b} - \eta \left( \frac{\partial^2 h}{\partial \mathbf{b}\, \partial \mathbf{b}^{\mathrm{T}}} \right)^{-1} \frac{\partial h}{\partial \mathbf{b}} = \mathbf{b} - \eta \left[ I + \left(\bar{G} - \bar{\mathbf{g}}\bar{\mathbf{g}}^{\mathrm{T}}\right) S^{-1} \right]^{-1} (\mathbf{b} - \bar{\mathbf{g}}),$$

where $\eta = 1$. This update may fail due to numerical errors in areas of high curvatures. In such a case, we search for an optimal $\eta \in [0,1]$ using the false position method.

### 3.2 Parameter $S_i$

Similar to $\mathbf{b}_i$, only $h(\mathbf{y}_i; q_i, \mathbf{b}_i, S_i)$ needs to be considered for $S_i$, and the datum subscript $i$ is suppressed here. The partial gradient with respect to $S$ is given by Equation 59, from which we obtain the implicit Equation 60. Let $V$ factorizes to $LL^{\mathrm{T}}$, where $L$ is non-singular since $V \succ 0$. Using $A$ given by (18) at the current value of $S$, a fixed-point update for $S$ is

$$S^{\text{fx}} = L^{-\mathrm{T}} P \tilde{\Lambda} P^{\mathrm{T}} L^{-1}, \qquad\qquad \tilde{\Lambda} \overset{\text{def}}{=} (\Lambda + I/4)^{1/2} + I/2,$$

where $P\Lambda P^{\mathrm{T}}$ is the eigen-decomposition of $L^{\mathrm{T}} A L$; see the proof of Lemma 28 in Appendix B.4.

The fixed-point update $S^{\text{fx}}$ may fail to improve the bound. We may fall-back on the Newton-Raphson update for $S$ that uses gradient (59) and a $C^2$-by-$C^2$ Hessian matrix. However, this can be rather involved since it needs to ensure that $S$ stays positive definite. An alternative, which we prefer, is to perform line-search in a direction that guarantees positive definiteness. To this end, let $S = S^{\text{cc}} \overset{\text{def}}{=} (1-\eta)S + \eta S^{\text{fx}}$, and we search for a $\eta \in [0,1]$ that optimizes the bound using the false position method. Appendix C.1 gives the details.

### 3.3 Parameter m, and Joint Optimization with b

We now optimize the bound $\log Z_h$ with respect to $\mathbf{m}$. Here, the datum subscript $i$ is reintroduced. Let $\mathbf{y}$ (resp. $\bar{\mathbf{g}}$) be the $nC$-vector obtained by stacking the $\mathbf{y}_i$s (resp. $\bar{\mathbf{g}}_i$s). Let $\bar{G}$ be the $nC$-by-$nC$ diagonal matrix with $\bar{\mathbf{g}}$ along its diagonal, and let $\tilde{G}$ be the $nC$-by-$nC$ block diagonal matrix with $\bar{\mathbf{g}}_i\bar{\mathbf{g}}_i^{\mathrm{T}}$ as the $i$th block. The gradient and Hessian with respect to $\mathbf{m}$ are

$$\frac{\partial \log Z_h}{\partial \mathbf{m}} = -K^{-1}\mathbf{m} + \mathbf{y} - \bar{\mathbf{g}}, \qquad \frac{\partial^2 \log Z_h}{\partial \mathbf{m} \partial \mathbf{m}^{\mathrm{T}}} = -K^{-1} - \left(\bar{G} - \tilde{G}\right). \tag{26}$$

The Hessian is negative semi-definite; this is another proof that $\log Z_h$ is concave in $\mathbf{m}$. The fixed-point update $\mathbf{m}^{\mathrm{fx}} = K(\mathbf{y} - \bar{\mathbf{g}})$ can be obtained by setting the gradient to zero. This update may fail to give a better bound. One remedy is to use the Newton-Raphson update. Alternatively the concavity in $\mathbf{m}$ can be exploited to optimize with respect to $\eta \in [0,1]$ in $\mathbf{m}^{\mathrm{cc}} \overset{\mathrm{def}}{=} (1-\eta)\mathbf{m} + \eta\mathbf{m}^{\mathrm{fx}}$, such as is done for the parameters $S_i$s in the previous section. Here we will give a combined update for $\mathbf{m}$ and the $\mathbf{b}_i$s that can be used during variational learning. This update avoids inverting $K$, which can be ill-conditioned.

The gradient in (26) implies the self-consistent equation $\mathbf{m}^* = K(\mathbf{y} - \bar{\mathbf{g}}^*)$ at the maximum, where $\bar{\mathbf{g}}^*$ is $\bar{\mathbf{g}}$ evaluated the the optimum parameters. From Lemma 26, another self-consistent equation is $\mathbf{b}^* = \bar{\mathbf{g}}^*$, where the $nC$-vector $\mathbf{b}^*$ is obtained by stacking all the $\mathbf{b}_i$s. Combining these two equations gives $\mathbf{m}^* = K(\mathbf{y} - \mathbf{b}^*)$, which is a bijection between $\mathbf{b}^*$ and $\mathbf{m}^*$ if $K$ has full rank. For the sparse approximation that will be introduced later, $K$ will be replaced by the "fat" matrix $K_{\mathrm{f}}$, which is column-rank deficient. There, the mapping from $\mathbf{b}^*$ to $\mathbf{m}^*$ becomes many-to-one. With this in mind, instead of letting $\mathbf{m}$ be a variational parameter, we fix it to be a function of $\mathbf{b}$, that is,

$$\mathbf{m} = K(\mathbf{y} - \mathbf{b}), \tag{27}$$

and we optimize over $\mathbf{b}$ instead. The details are in Appendix C.2.

This joint update for $\mathbf{m}$ and the $\mathbf{b}_i$s can be used for variational learning. This, however, does not make the update in Section 3.1 redundant: that is still required during approximate prediction, where the $\mathbf{b}_*$ for the test input $\mathbf{x}_*$ still needs to be optimized over even though $\mathbf{m}$ is fixed after learning.

### 3.4 Parameter $V$

For the gradient with respect to $V$ we have

$$\frac{\partial h_i}{\partial V_i} = \frac{1}{2}V_i^{-\mathrm{T}} - \frac{1}{2}S_i^{\mathrm{T}} = -\frac{1}{2}W_i^{\mathrm{T}}, \qquad \frac{\partial \log Z_h}{\partial V} = \frac{1}{2}V^{-1} - \frac{1}{2}K^{-1} - \frac{1}{2}W^{-1},$$

where $W_i \overset{\mathrm{def}}{=} S_i - V_i^{-1}$, and $W$ is the block diagonal matrix of the $W_i$s. Here, function $h_i$ is regarded as parameterized by $S_i$ (as in Theorem 6) rather than by $W_i$ (as in Lemma 25). Using gradient $\partial \log Z_h / \partial V$ directly as a search direction to update $V$ is undesirable for two reasons. First, it may not preserve the positive-definiteness of $V$. Second, it requires $K$ to be inverted, and this can cause numerical issues for some covariance functions such as the squared exponential covariance function, which has exponentially vanishing eigenvalues.

We propose to let $V$ follow the trajectory along a modified gradient, where $W$ is regarded fixed instead of depending on $V$. To explain, we recall that $\log Z_h \overset{\mathrm{def}}{=} -\mathrm{KL}(q(\mathbf{f}|\mathbf{y})\|p(\mathbf{f})) + h$, where

$h \overset{\text{def}}{=} \sum_{i=1}^{n} h_i$ is the sum of functions each concave in $V$. The modified gradient holds the gradient contribution from $h$ constant at the value at the initial $V$ while the gradient contribution from the Kullback-Leibler divergence varies along the trajectory. We follow the trajectory until the modified gradient is zero. Let this point be $V^{\text{fx}}$. Then

$$\frac{1}{2}(V^{\text{fx}})^{-1} - \frac{1}{2}K^{-1} - \frac{1}{2}W^{-1} = 0, \qquad \text{or} \qquad V^{\text{fx}} = \left(K^{-1} + W\right)^{-1}. \tag{28}$$

The equation on the right can be used as a naïve fix-point update.

The trajectory following this modified gradient will diverge from the trajectory following the exact gradient, so there is no guarantee that $V^{\text{fx}}$ gives an improvement over $V$. To remedy, we follow the strategy used for updating $S$: we use $V^{\text{cc}} \overset{\text{def}}{=} (1-\eta)V + \eta V^{\text{fx}}$ and optimize with respect to $\eta \in [0,1]$. Matrix $V^{\text{cc}}$ is guaranteed to be positive definite, since it is a convex combination of two positive definite matrices. Details are in Appendix C.3.

## 4. Sparse Approximation

The variational approach for learning multinomial logit Gaussian processes discussed in the previous sections has transformed an intractable integral problem into a tractable optimization problem. However, the variational approach is still expensive for large data sets because the computational complexity of the matrix operations is $O(C^3n^3)$, where $n$ is the size of the observed set and $C$ is the number of classes. One popular approach to reduce the complexity is to use sparse approximations: only $s \ll n$ data inputs or sites are chosen to be used within a complete but smaller Gaussian process model, and information for the rest of the observations are induced via these $s$ sites. Each of the $s$ data sites is called an *inducing site*, and the associated random variables $\mathbf{z}$ are called the *inducing variables*. We use the term *inducing set* to mean either the inducing sites or the inducing variables or both. The selection of the inducing set is seen as a model selection problem (Snelson and Ghahramani, 2006; Titsias, 2009a) and will be addressed in Section 6.1.

We seek a sparse approximation will lead to a lower bound on the true marginal likelihood. This approach has been proposed for Gaussian process regression (Titsias, 2009a), and it will facilitate the search for the inducing set later. Recall that the inducing variables at the $s$ inducing sites are denoted by $\mathbf{z} \in \mathbb{R}^s$. We retain $\mathbf{f}$ for the $nC$ latent function values associated with the $n$ observed data $(X, \mathbf{y})$. In general, the inducing variables $\mathbf{z}$ need not be chosen from the latent function values $\mathbf{f}$, so our presentation will treat them as distinct.

The Gaussian prior over the latent values $\mathbf{f}$ is extended to the inducing variables $\mathbf{z}$ to give a Gaussian joint prior $p(\mathbf{f}, \mathbf{z})$. Let $p(\mathbf{f}, \mathbf{z}|\mathbf{y})$ be the true joint posterior of the latent and inducing variables is given the observed data. This posterior is non-Gaussian because of the multinomial logistic likelihood function, and it is intractable to calculate this posterior as is in the non-sparse case. The approximation $q(\mathbf{f}, \mathbf{z}|\mathbf{y})$ to the exact posterior is performed in two steps. In the first step, we let $q(\mathbf{f}, \mathbf{z}|\mathbf{y})$ be a Gaussian distribution. This is a natural choice which follows from the non-sparse case. In the second step, we use the factorization

$$q(\mathbf{f}, \mathbf{z}|\mathbf{y}) \overset{\text{def}}{=} p(\mathbf{f}|\mathbf{z})\, q(\mathbf{z}|\mathbf{y}), \tag{29}$$

where $p(\mathbf{f}|\mathbf{z})$ is the marginal of $\mathbf{f}$ from the prior $p(\mathbf{f}, \mathbf{z})$. The same approximation has been used in the sparse approximation for regression (Titsias, 2009a, paragraph before Equation 7). This approximation makes clear the role of inducing variables $\mathbf{z}$ as the conduit of information from $\mathbf{y}$ to $\mathbf{f}$. Under this approximate posterior, we have the bound

$$\log p(\mathbf{y}) \geq \log \tilde{Z}_B = -\mathrm{KL}(q(\mathbf{z}|\mathbf{y}) \| p(\mathbf{z})) + \sum_{i=1}^{n} \ell_i(\mathbf{y}_i; q),$$

where $\ell_i(\mathbf{y}_i; q) \overset{\text{def}}{=} \int q(\mathbf{f}_i|\mathbf{y}) \log p(\mathbf{y}_i|\mathbf{f}_i)\, \mathrm{d}\mathbf{f}_i$, and $q(\mathbf{f}_i|\mathbf{y})$ is the marginal distribution of $\mathbf{f}_i$ from the joint distribution $q(\mathbf{f}, \mathbf{z}|\mathbf{y})$; see Appendix B.1. The reader may wish to compare with Equations 3, 4 and 5 for the non-sparse variational approximation.

Similar to the dissection of $\log Z_B$ after Equation 3, the Kullback-Leibler divergence component of $\log \tilde{Z}_B$ can be interpreted as the regularizing factor for the approximate posterior $q(\mathbf{z}|\mathbf{y})$, while the expected log-likelihood can be interpreted as the data fit component. This dissection provides three insights into the sparse formulation. First, the specification of $p(\mathbf{z})$ is part of the model and not part of the approximation—the approximation step is in the factorization (29). Second, the Kullback-Leibler divergence term involves only the inducing variables $\mathbf{z}$ and *not* the latent variables $\mathbf{f}$. Hence, the regularizing is on the approximate posterior of $\mathbf{z}$ and not on that of $\mathbf{f}$. Third, the involvement of $\mathbf{f}$ is confined to the data fit component in a two step process: generating $\mathbf{f}$ from $\mathbf{z}$ and then generating $\mathbf{y}$ from $\mathbf{f}$.

Applying Theorem 6 on the $\ell_i$s gives

$$\log \tilde{Z}_B \geq \log \tilde{Z}_h \overset{\text{def}}{=} -\mathrm{KL}(q(\mathbf{z}|\mathbf{y}) \| p(\mathbf{z})) + \sum_{i=1}^{n} h(\mathbf{y}_i; q_i, \mathbf{b}_i, S_i), \tag{30}$$

where $h$ is defined by Equation 19, and the $q_i$ within $h$ is the marginal distribution $q(\mathbf{f}_i|\mathbf{y})$.

We now examine $\log \tilde{Z}_h$ using the parameters of the distributions. Let the joint prior be

$$p\left(\begin{pmatrix} \mathbf{z} \\ \mathbf{f} \end{pmatrix}\right) = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} K & K_{\mathrm{f}} \\ K_{\mathrm{f}}^{\mathrm{T}} & K_{\mathrm{ff}} \end{pmatrix}\right).$$

One can generalize the prior for $\mathbf{z}$ to have a non-zero mean, but the above suffice for our purpose and simplifies the presentation. In the case where an inducing variable $z_i$ coincide with a latent variable $f_j^c$, we can "tie" them by setting their prior correlation to one. The marginal distribution $p(\mathbf{f})$ is the Gaussian process prior of the model, but we are now using $K_{\mathrm{ff}}$ to denote the covariance induced by $K^{\mathrm{c}}$ and $k^{\mathrm{x}}(\cdot,\cdot)$ while reserving $K$ for the covariance of $\mathbf{z}$. This facilitates comparison to the expressions for the non-sparse approximation.

For the approximate posterior, let $q(\mathbf{z}|\mathbf{y}) = \mathcal{N}(\mathbf{m}, V)$, so $\mathbf{m}$ and $V$ are the variational parameters of the approximation. Then $q(\mathbf{f}|\mathbf{y})$ is Gaussian with mean and covariance

$$\mathbf{m}_{\mathrm{f}} = K_{\mathrm{f}}^{\mathrm{T}} K^{-1} \mathbf{m}, \qquad V_{\mathrm{f}} = K_{\mathrm{ff}} - K_{\mathrm{f}}^{\mathrm{T}} K^{-1} K_{\mathrm{f}} + K_{\mathrm{f}}^{\mathrm{T}} K^{-1} V K^{-1} K_{\mathrm{f}}. \tag{31}$$

Therefore, the lower bound on the log marginal likelihood is

$$\log \tilde{Z}_h = \frac{s}{2} + \frac{1}{2} \log |K^{-1}V| - \frac{1}{2} \operatorname{tr} K^{-1} V - \frac{1}{2} \mathbf{m}^{\mathrm{T}} K^{-1} \mathbf{m} \\ + \frac{nC}{2} + \mathbf{m}_{\mathrm{f}}^{\mathrm{T}} \mathbf{y} + \frac{1}{2} \sum_{i=1}^{n} \Big( \log |S_i V_{\mathrm{f}i}| - \operatorname{tr} S_i V_{\mathrm{f}i} \Big) - \sum_{i=1}^{n} \log \sum_{c=1}^{C} g_i^c, \tag{32}$$

where $V_{\mathrm{f}i}$ is the $i$th diagonal $C$-by-$C$ block matrix of $V_{\mathrm{f}}$ and

$$g_i^c \overset{\mathrm{def}}{=} \exp\left[\mathbf{m}_{\mathrm{f}i}^{\mathrm{T}}\mathbf{e}^c + \frac{1}{2}(\mathbf{b}_i - \mathbf{e}^c)^{\mathrm{T}} S_i^{-1}(\mathbf{b}_i - \mathbf{e}^c)\right].$$

**Remark 13** *It is not necessary for* $\mathbf{z}$ *to be drawn from the latent Gaussian process prior directly. Therefore, the covariance* $K$ *of* $\mathbf{z}$ *need not be given by the covariance functions* $K^{\mathrm{c}}$ *and* $k^{\mathrm{x}}(\cdot,\cdot)$ *of the latent Gaussian process model. In fact,* $\mathbf{z}$ *can be any linear functional of draws from the latent Gaussian process prior (see, for example, Titsias, 2009b, Section 6). For example, it is almost always necessary to set* $\mathbf{z} = \mathbf{z}' + \epsilon$*, where* $\epsilon$ *is the isotropic noise, so that the matrix inversion of* $K$ *is not ill-conditioned. This matrix inversion cannot be avoided (without involving* $O(s^3)$ *computations) in the sparse approximation because of the need to compute* $K_{\mathrm{f}}^{\mathrm{T}} K^{-1} K_{\mathrm{f}}$*, which is the Nyström approximation to* $K_{\mathrm{ff}}$ *if* $\mathbf{z}' \equiv \mathbf{f}$ *(Williams and Seeger, 2001). One attractiveness of having a lower bound associated to the sparse approximation is that the noise variance of* $\epsilon$ *can be treated as a lower bounded variational parameter to be optimized (Titsias, 2009b, Section 6).*

**Remark 14** *The inducing variables* $\mathbf{z}$ *are associated with the latent values* $\mathbf{f}$ *and not with the observed data* $(X, \mathbf{y})$*. Therefore, it is not necessary to choose all the latent values* $f_i^1, \ldots, f_i^C$ *for any datum* $\mathbf{x}_i$*. One may choose the inducing sites to include, say,* $f_i^c$ *for datum* $\mathbf{x}_i$ *and* $f_j^{c'}$ *for datum* $\mathbf{x}_j$*, and to exclude* $f_i^{c'}$ *for datum* $\mathbf{x}_i$ *and* $f_j^c$ *for datum* $\mathbf{x}_j$*. This flexibility requires additional bookkeeping in the implementation.*

### 4.1 Comparing Sparse and Non-sparse Approximations

We can relate the bounds for the non-sparse and sparse approximations:

**Theorem 15** *Let*

$$\log Z_B^* \overset{\mathrm{def}}{=} \max_{q(\mathbf{f}|\mathbf{y})} \log Z_B, \quad \log Z_h^* \overset{\mathrm{def}}{=} \max_{q(\mathbf{f}|\mathbf{y})} \log Z_h, \quad \log \tilde{Z}_B^* \overset{\mathrm{def}}{=} \max_{q(\mathbf{z}|\mathbf{y})} \log \tilde{Z}_B, \quad \log \tilde{Z}_h^* \overset{\mathrm{def}}{=} \max_{q(\mathbf{z}|\mathbf{y})} \log \tilde{Z}_h,$$

*where the sparse bounds are for any inducing set. Then* $\log Z_B^* \geq \log \tilde{Z}_B^*$ *and* $\log Z_h^* \geq \log \tilde{Z}_h^*$*.*

The proof for the first inequality is given in Appendix B.1.1, while the second inequality is a consequence of Proposition 17 derived in Section 6.1. Though intuitive, the second inequality is not obvious because of the additional maximization over the variational parameters $\{\mathbf{b}_i\}$ and $\{S_i\}$. The presented sparse approximation is optimal: if $\mathbf{z} \equiv \mathbf{f}$, then $\log \tilde{Z}_B = \log Z_B$ and $\log \tilde{Z}_h = \log Z_h$, and the sparse approximation becomes the non-sparse approximation.

### 4.2 Optimization

The sparse approximation requires optimizing $\log \tilde{Z}_h$ with respect to the variational parameters: mean $\mathbf{m}$ and covariance $V$ of the inducing variables; and $\{\mathbf{b}_i\}$ and $\{S_i\}$ for the lower bound on the expected log-likelihood. The optimization with respect to $\{S_i\}$ is the same as that for the non-sparse approximation, but using $\mathbf{m}_{\mathrm{f}}$ and $V_{\mathrm{f}}$ for the mean and covariance of the latent variables. For $\{\mathbf{b}_i\}$ and $\mathbf{m}$, a joint optimization akin to that for the non-sparse approximation described in Section 3.3 can be used. Let $\mathbf{b}$ be the $nC$-vector that is the stacking of the $\mathbf{b}_i$s. At the saddle point with respect to $\{\mathbf{b}_i\}$ and $\mathbf{m}$, we have the self-consistent equations $\mathbf{b}_i^* = \bar{\mathbf{g}}_i^*$ and $\mathbf{m}^* = K_{\mathrm{f}}(\mathbf{y} - \bar{\mathbf{g}}^*)$, from which is obtained the linear mapping $\mathbf{m}^* = K_\mathrm{f}(\mathbf{y} - \mathbf{b}^*)$. For sparse approximation, matrix $K_\mathrm{f}$ has more columns than rows (that is, a "fat" matrix), so the linear mapping from $\mathbf{b}^*$ to $\mathbf{m}^*$ is many to one. Hence we substitute the constraint $\mathbf{m} = K_\mathrm{f}(\mathbf{y} - \mathbf{b})$ into the bound and optimize over $\mathbf{b}$. The optimization is similar to non-sparse case, and is detailed in Appendix C.2.1.

For $V$, the approach in Section 3.4 for the non-sparse approximation is followed. The gradients with respect to $V$ are

$$\begin{aligned} \frac{\partial h_i}{\partial V} &= \frac{1}{2} K^{-1} K_{\mathrm{f}i} \left( V_{\mathrm{f}i}^{-1} - S_i \right) K_{\mathrm{f}i}^\mathrm{T} K^{-1} \\ &= -\frac{1}{2} K^{-1} K_{\mathrm{f}i} W_{\mathrm{f}i} K_{\mathrm{f}i}^\mathrm{T} K^{-1}; \end{aligned} \qquad \begin{aligned} \frac{\partial \log \tilde{Z}_h}{\partial V} &= \frac{1}{2} V^{-1} - \frac{1}{2} K^{-1} - \frac{1}{2} K^{-1} K_\mathrm{f} W_\mathrm{f} K_\mathrm{f}^\mathrm{T} K^{-1} \\ &= \frac{1}{2} V^{-1} - \frac{1}{2} K^{-1} - \frac{1}{2} W, \end{aligned}$$

where $W_{\mathrm{f}i} \overset{\mathrm{def}}{=} S_i - V_{\mathrm{f}i}^{-1}$; matrix $W_\mathrm{f}$ is block diagonal with $W_{\mathrm{f}i}$ as its $i$th block; and we have introduced $W \overset{\mathrm{def}}{=} K^{-1} K_\mathrm{f} W_\mathrm{f} K_\mathrm{f}^\mathrm{T} K^{-1}$. The fixed point update for $V$ is

$$V^{\mathrm{fx}} = \left( K^{-1} + W \right)^{-1}, \tag{33}$$

which is obtained by setting $\partial Z_h / \partial V$ at $V^{\mathrm{fx}}$ to zero. This update is of the same character as Equation 28 for the non-sparse case. In the case where $V^{\mathrm{fx}}$ does not yield an improvement to the objective $\log \tilde{Z}_h$, we search for a $V^{\mathrm{cc}} \overset{\mathrm{def}}{=} (1 - \eta) V + \eta V^{\mathrm{fx}}$, $\eta \in [0, 1]$, using the false position method along $\eta$. Further details can be found in Appendix C.3.1.

## 5. On the Sum-to-zero Property

For many single-machine multi-class support vector machines (SVMs, Vapnik 1998; Bredensteiner and Bennett 1999; Guermeur 2002; Lee, Lin, and Wahba 2004), the sum of the predictive functions over the classes is constrained to be zero everywhere. For these SVMs, the constraint ensures the uniqueness of the solution (Guermeur, 2002). The lack of uniqueness without constraint is similar the non-identifiability of parameters in the multinomial probit model in statistics (see Geweke, Keane, and Runkle, 1994, and references therein). For multi-class classification with Gaussian process prior and multinomial logistic likelihood, the redundancy in representation has been acknowledged, but typically uniqueness has not been enforced to avoid arbitrary asymmetry in the prior (Williams and Barber, 1998; Neal, 1998). An exception is the work by Kim and Ghahramani (2006), where a linear transformation of the latent functions has been used to remove the redundancy. In this section, we show that such *sum-to-zero* property is present in the optimal variational posterior under certain common settings.

Recall from Equation 27 in Section 3.3 that $\mathbf{m} = K(\mathbf{y} - \mathbf{b})$ when the lower bound $Z_h$ is optimized. Let $\boldsymbol{\alpha} \overset{\mathrm{def}}{=} K^{-1}\mathbf{m}$. Then the set of self-consistent equations at stationary gives $\boldsymbol{\alpha}_i = \mathbf{y}_i - \mathbf{b}_i$, where $\boldsymbol{\alpha}_i$ is the $i$th $C$-dimensional sub-vector of $\boldsymbol{\alpha}$. Since $\mathbf{b}_i = \bar{\mathbf{g}}_i$ at stationary, and $\bar{\mathbf{g}}_i$ is a probability vector, it follows that

$$\forall i \qquad \sum_{c=1}^{C} \alpha_i^c = 0, \qquad \text{and} \qquad \alpha_i^c = \begin{cases} -b_i^c \in \,]{-1}, 0[ & \text{if } y_i^c = 0 \\ 1 - b_i^c \in \,]0, 1[ & \text{if } y_i^c = 1. \end{cases}$$

Consider an input $\mathbf{x}_*$, which may be in the observed set. Let $\mathbf{k}_*^\mathrm{x}$ be the vector of covariances to all the other inputs under the covariance function $k^\mathrm{x}$. The posterior latent mean of $\mathbf{f}_*$ at $\mathbf{x}_*$ under separable covariance (1) is

$$\mathbf{m}_* = \left((\mathbf{k}^{\mathrm{x}}_*)^{\mathrm{T}} \otimes K^{\mathrm{c}}\right) \boldsymbol{\alpha} = \sum_{i=1}^{n} \left(k^{\mathrm{x}}(\mathbf{x}_*, \mathbf{x}_i) K^{\mathrm{c}}\right) \boldsymbol{\alpha}_i = K^{\mathrm{c}} \sum_{i=1}^{n} k^{\mathrm{x}}(\mathbf{x}_*, \mathbf{x}_i) \boldsymbol{\alpha}_i. \tag{34}$$

Consider the common case where $K^{\mathrm{c}} = I$. Then the posterior latent mean for the $c$th class is $m^c_* = \sum_{i=1}^{n} k^{\mathrm{x}}(\mathbf{x}_*, \mathbf{x}_i)\alpha^c_i$, and the covariance from the $i$th datum has a positive contribution if it is from the $c$th class and a negative contribution otherwise. Moreover, the sum of the latent means is

$$\mathbf{1}^{\mathrm{T}}\mathbf{m}_* = \mathbf{1}^{\mathrm{T}} \sum_{i=1}^{n} k^{\mathrm{x}}(\mathbf{x}_*, \mathbf{x}_i) \boldsymbol{\alpha}_i = \sum_{i=1}^{n} k^{\mathrm{x}}(\mathbf{x}_*, \mathbf{x}_i) \mathbf{1}^{\mathrm{T}} \boldsymbol{\alpha}_i = 0. \tag{35}$$

Hence that the sum of the latent means for any datum, whether observed or novel, is constant at zero. We call this the sum-to-zero property.

The sum-to-zero property is also present, but in a different way, when

$$K^{\mathrm{c}} = M - M\mathbf{1}\mathbf{1}^{\mathrm{T}}M/\mathbf{1}^{\mathrm{T}}M\mathbf{1}, \tag{36}$$

where $M$ is $C$-by-$C$ and positive semi-definite. This is reminiscent of Equation 20, which gives a similar parametrization for $W_*$. Using the rightmost expression in (34) for $\mathbf{m}_*$, we find that $\mathbf{1}^{\mathrm{T}}\mathbf{m}_* = 0$ because $K^{\mathrm{c}}\mathbf{1} = \mathbf{0}$. This is in contrast with (35) for $K^{\mathrm{c}} = I$, where the sum-to-zero property holds because $\mathbf{1}^{\mathrm{T}}\boldsymbol{\alpha}_i = 0$.

Setting $K^{\mathrm{c}}$ via (36) leads to a degenerate Gaussian process, since the matrix will have a zero eigenvalue even if $M$ is strictly positive definite. Since degeneracy is usually not desirable, we add to (36) the term $\eta I$, where $\eta > 0$:

$$K^{\mathrm{c}} = M - M\mathbf{1}\mathbf{1}^{\mathrm{T}}M/\mathbf{1}^{\mathrm{T}}M\mathbf{1} + \eta I. \tag{37}$$

This not only ensures that $K^{\mathrm{c}}$ is positive definite but also preserves the sum-to-zero property. The parametrization effectively constrains the least dominant eigenvector of $K^{\mathrm{c}}$ to $\mathbf{1}/\sqrt{C}$.

### 5.1 The Sum-to-zero Property in Sparse Approximation

The sum-to-zero property is also present in sparse approximation when the inducing variables are such that if $f^c_i$ is an inducing variable, then so are $f^1_i, \ldots, f^C_i$. That is, the $C$ latent variables associated with any input $\mathbf{x}_i$ are either omitted or included together in the inducing set. The sparsity of single-machine multi-class SVMs is of this nature. Let $t$ be the number of inputs for which their latent variables are included.

Under the separable covariance model (1), covariance between the inducing variables and the latent variables is the Kronecker product $K_{\mathrm{f}} = K^{\mathrm{x}}_{\mathrm{f}} \otimes K^{\mathrm{c}}$, where $K^{\mathrm{x}}_{\mathrm{f}}$ is the covariance on the inputs only. The stationary point of the lower bound $\tilde{Z}_h$ in the sparse approximation has the self-consistent equation $\mathbf{m} = K_{\mathrm{f}}(\mathbf{y} - \bar{\mathbf{g}})$; see Section 4.2. As before, let $\boldsymbol{\alpha} \overset{\mathrm{def}}{=} K^{-1}\mathbf{m}$. The Gram matrix $K$ is the Kronecker product $K^{\mathrm{x}} \otimes K^{\mathrm{c}}$ under the separable covariance model. Hence $\boldsymbol{\alpha} = \left((K^{\mathrm{x}})^{-1}K^{\mathrm{x}}_{\mathrm{f}} \otimes I\right)(\mathbf{y} - \bar{\mathbf{g}})$ using the mixed-product property. Vector $\boldsymbol{\alpha}$ is the stacking of vectors $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_t$, where each $\boldsymbol{\alpha}_j$ is for one of the $t$ inputs with their latent variables in the inducing set and can be expressed as

$$\boldsymbol{\alpha}_j = \sum_{i=1}^{n} \left((K^{\mathrm{x}})^{-1}K^{\mathrm{x}}_{\mathrm{f}}\right)_{ji} (\mathbf{y}_i - \bar{\mathbf{g}}_i).$$

One finds that $\mathbf{1}^{\mathrm{T}}\boldsymbol{\alpha}_j = 0$, and the discussion for the non-sparse case applies similarly from Equation 34 onwards.

## 6. Model Learning

Model learning in a Gaussian process model is achieved by maximizing the marginal likelihood with respect to the parameters $\boldsymbol{\theta}$ of the covariance function. In the case of variational inference, the lower bound on the marginal likelihood is maximized instead. For the non-sparse variational approximation to the multinomial logit Gaussian process, this is

$$\log Z_h^*(\boldsymbol{\theta}) \overset{\text{def}}{=} \max_{\mathbf{m},V,\{\mathbf{b}_i\},\{S_i\}} \log Z_h(\mathbf{m},V,\{\mathbf{b}_i\},\{S_i\};X,\mathbf{y},\boldsymbol{\theta}),$$

which is the maximal lower bound on log marginal likelihood on the observed data $(X,\mathbf{y})$. The maximization is achieved by ascending the gradient

$$\begin{aligned}\frac{\mathrm{d}\log Z_h^*}{\mathrm{d}\theta_j} &= -\frac{1}{2}\operatorname{tr}\left(K^{-1}\frac{\partial K}{\partial\theta_j}\right) + \frac{1}{2}\operatorname{tr}\left(K^{-1}VK^{-1}\frac{\partial K}{\partial\theta_j}\right) + \frac{1}{2}\mathbf{m}^{\mathrm{T}}K^{-1}\frac{\partial K}{\partial\theta_j}K^{-1}\mathbf{m} \\ &= \frac{1}{2}\operatorname{tr}\left(\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathrm{T}} - K^{-1} + K^{-1}VK^{-1}\right)\frac{\partial K}{\partial\theta_j}\right),\end{aligned}$$

where $\boldsymbol{\alpha} \overset{\text{def}}{=} K^{-1}\mathbf{m}$. This gradient is also the partial and explicit gradient of $\log Z_h$ with respect to $\theta_j$. The implicit gradients via the variational parameters are not required since the derivative of $\log Z_h$ with respect to each of them is zero at the fixed point $\log Z_h^*$.

For the sparse approximation, we differentiate $\log\tilde{Z}_h^*$—the optimized bound on the log marginal likelihood for the sparse case given by Equation 32—with respect to the covariance function parameter $\theta_j$. The derivation in Appendix C.4 gives

$$\begin{aligned}\frac{\mathrm{d}\log\tilde{Z}_h^*}{\mathrm{d}\theta_j} = &-\frac{1}{2}\operatorname{tr}\left(\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathrm{T}} - K^{-1} + K^{-1}VK^{-1} + W\right)\frac{\partial K}{\partial\theta_j}\right) \\ &+ \operatorname{tr}\left(\left((\mathbf{y}-\bar{\mathbf{g}})\boldsymbol{\alpha}^{\mathrm{T}} + W_{\mathrm{f}}K_{\mathrm{f}}^{\mathrm{T}}\left(K^{-1} - K^{-1}VK^{-1}\right)\right)\frac{\partial K_{\mathrm{f}}}{\partial\theta_j}\right) - \frac{1}{2}\operatorname{tr}\left(W_{\mathrm{f}}\frac{\partial K_{\mathrm{ff}}}{\partial\theta_j}\right),\end{aligned}$$

where $\boldsymbol{\alpha} \overset{\text{def}}{=} K^{-1}\mathbf{m}$, and matrices $W_{\mathrm{f}}$ and $W$ are defined in Section 4.2.

The selection of the inducing set in sparse approximation can also be seen as a model learning problem (Snelson and Ghahramani, 2006; Titsias, 2009a).[1] This is addressed in the reminder of this section.

### 6.1 Active Inducing Set Selection

The quality of the sparse approximation depends on the set of inducing sites. Prior works have suggested using scores to greedily and iteratively add to the set. The Informative Vector Machine (IVM, Lawrence et al. 2003) and its generalization to multiple classes (Seeger and Jordan, 2004) use the differential entropy, which is the amount of additional information to the posterior. Alternatives based on the data likelihood have also been proposed (Girolami and Rogers, 2006; Henao and Winther, 2010). However, since our aim has always been to maximize the marginal likelihood $p(\mathbf{y})$ of the observed data, it is natural to choose the inducing sites that effect the most increase in the marginal likelihood. The same thought is behind the scoring for greedy selection in the

[1] However, in the strict sense, the exact model is fixed during the selection of the inducing set: the object that is learned is the approximating model.

sparse variational approximation to Gaussian process regression (Titsias, 2009a). For multi-class classification, it is too expensive to compute the exact increase in the marginal likelihood. Instead, we use the lower bound on the increment to (the lower bound on) the marginal likelihood.

Throughout, the asterisk $*$ will be used to subscript variables pertaining to the newly introduced inducing site $\tilde{\mathbf{x}}_*$. Given the current set of inducing sites $\tilde{X}$, the inclusion of $\tilde{\mathbf{x}}_*$ gives the new set $\tilde{X}_*$. The function values at $\tilde{\mathbf{x}}_*$, $\tilde{X}$ and $\tilde{X}_*$ are denoted by $z_*$, $\mathbf{z}$ and $\mathbf{z}_* \overset{\text{def}}{=} (\mathbf{z}^{\mathrm{T}}, z_*)^{\mathrm{T}}$. There is only one random scalar variable $z_*$ at the inducing site $\tilde{\mathbf{x}}_*$. In contrast, there is a random $C$-vector $\mathbf{f}_i$ at an observed input $\mathbf{x}_i$; see Remark 14. Hence there are $C$ potential inducing sites from a single observed site $\mathbf{x}_i$: $\tilde{\mathbf{x}}_* \in \{(\mathbf{x}_i, 1), \ldots, (\mathbf{x}_i, C)\}$.

We aim to select $\tilde{\mathbf{x}}_*$ that maximizes the increase in the optimized lower bounds on the marginal likelihood: $d(\tilde{\mathbf{x}}_*; \tilde{X}) \overset{\text{def}}{=} \log \tilde{Z}_h^*(\tilde{X}_*) - \log \tilde{Z}_h^*(\tilde{X})$, where $\tilde{X}_* \overset{\text{def}}{=} \{\tilde{\mathbf{x}}_*\} \cup \tilde{X}$, and

$$\log \tilde{Z}_h^*(\tilde{X}_*) \overset{\text{def}}{=} \max_{\mathbf{m}_*, V_*, \{\mathbf{b}_{*i}\}, \{S_{*i}\}} \log \tilde{Z}_h(\mathbf{m}_*, V_*, \{\mathbf{b}_{*i}\}, \{S_{*i}\}; \tilde{X}_*),$$
$$\log \tilde{Z}_h^*(\tilde{X}) \overset{\text{def}}{=} \max_{\mathbf{m}, V, \{\mathbf{b}_i\}, \{S_i\}} \log \tilde{Z}_h(\mathbf{m}, V, \{\mathbf{b}_i\}, \{S_i\}; \tilde{X}).$$

In words, $\tilde{Z}_h^*(\tilde{X})$ is the optimized lower bound on marginal likelihood with the current inducing set $\tilde{X}$, while $\tilde{Z}_h^*(\tilde{X}_*)$ is the optimized lower bound with the proposed new inducing set $\tilde{X}_*$. Because $\tilde{Z}_h$ combines the Kullback-Leibler divergence of the prior from the approximate posterior and the sum of the lower bounds on the expected log-likelihoods, $d(\tilde{\mathbf{x}}_*; \tilde{X})$ includes both the change in the approximate posterior and the effect of this change in explaining the observed data.

Computing $d(\tilde{\mathbf{x}}_*; \tilde{X})$ involves $\tilde{Z}_h^*(\tilde{X}_*)$, and this can be computationally expensive. A more viable alternative is to lower bound $d(\tilde{\mathbf{x}}_*; \tilde{X})$ by fixing selected variational parameters in $\tilde{Z}_h(\cdots; \tilde{X}_*)$ to the optimal ones from $\tilde{Z}_h^*(\tilde{X})$, which has already been computed. Let

$$\{\mathbf{m}, V, \{\mathbf{b}_i\}, \{S_i\}\} \overset{\text{def}}{=} \arg \log \tilde{Z}_h^*(\tilde{X}).$$

For the inducing set $\tilde{X}_*$, we set the prior on the inducing and latent variables, and the approximate posterior on the inducing variables to

$$p\left(\begin{pmatrix} \mathbf{z}_* \\ \mathbf{f} \end{pmatrix}\right) \overset{\text{def}}{=} \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} K_* & K_{\mathrm{f}*} \\ K_{\mathrm{f}*}^{\mathrm{T}} & K_{\mathrm{ff}} \end{pmatrix}\right), \qquad q(\mathbf{z}_* \mid \mathbf{y}) \overset{\text{def}}{=} \mathcal{N}(\mathbf{m}_*, V_*),$$

where

$$K_* \overset{\text{def}}{=} \begin{pmatrix} K & \mathbf{k}_* \\ \mathbf{k}_*^{\mathrm{T}} & k_{**} \end{pmatrix}, \qquad K_{\mathrm{f}*} \overset{\text{def}}{=} \begin{pmatrix} K_{\mathrm{f}} \\ \mathbf{k}_{\mathrm{f}*}^{\mathrm{T}} \end{pmatrix}, \qquad \mathbf{m}_* \overset{\text{def}}{=} \begin{pmatrix} \mathbf{m} \\ m_* \end{pmatrix}, \qquad V_* \overset{\text{def}}{=} \begin{pmatrix} V & \mathbf{v}_* \\ \mathbf{v}_*^{\mathrm{T}} & v_{**} \end{pmatrix}. \tag{38}$$

The above choice of posterior fixes the mean and the covariance of $\mathbf{z}$ for $\tilde{X}$ to the mean $\mathbf{m}$ and covariance $V$ in $\log \tilde{Z}_h^*(\tilde{X})$. Further setting $\{\mathbf{b}_{*i}\} \equiv \{\mathbf{b}_i\}$ and $\{S_{*i}\} \equiv \{S_i\}$, the additional variational parameters are those in the posterior of the inducing points for the additional site $\tilde{\mathbf{x}}_*$. Since we are optimizing over only a subset of the possible parameters, we obtain a lower bound on $d(\tilde{\mathbf{x}}_* | \tilde{X})$:

$$d(\tilde{\mathbf{x}}_* | \tilde{X}) \geq d_1(\tilde{\mathbf{x}}_* | \tilde{X}) \overset{\text{def}}{=} \max_{m_*, v_{**}, \mathbf{v}_*} \log \tilde{Z}_h(\mathbf{m}_*, V_*, \{\mathbf{b}_i\}, \{S_i\}; \tilde{X}_*) - \log \tilde{Z}_h^*(\tilde{X}), \tag{39}$$

where $\mathbf{m}_*$ and $V_*$ are as defined in Equation 38. By separating $\log \tilde{Z}_h$ into its summands expressed in Equation 30, we write

$$d_1(\tilde{\mathbf{x}}_* | \tilde{X}) = \max_{m_*, v_{**}, \mathbf{v}_*} \left( d_{\mathrm{KL}}(m_*, v_{**}, \mathbf{v}_*, \tilde{\mathbf{x}}_* | \tilde{X}) + \sum_{i=1}^{n} d_h^i(m_*, v_{**}, \mathbf{v}_*, \tilde{\mathbf{x}}_* | \tilde{X}) \right),$$

where

$$d_{\mathrm{KL}}(m_*, v_{**}, \mathbf{v}_*, \tilde{\mathbf{x}}_*|\tilde{X}) \overset{\mathrm{def}}{=} -\mathrm{KL}(q(\mathbf{z}_* \mid \mathbf{y}) \,\|\, p(\mathbf{z}_*)) + \mathrm{KL}(q(\mathbf{z} \mid \mathbf{y}) \,\|\, p(\mathbf{z})), \tag{40}$$

$$d_h^i(m_*, v_{**}, \mathbf{v}_*, \tilde{\mathbf{x}}_*|\tilde{X}) \overset{\mathrm{def}}{=} h(\mathbf{y}_i; q_{*i}, \mathbf{b}_i, S_i) - h(\mathbf{y}_i; q_i, \mathbf{b}_i, S_i). \tag{41}$$

The expressions for $d_{\mathrm{KL}}$ and $d_h^i$ in terms of the variational parameters $m_*$, $v_{**}$ and $\mathbf{v}_*$ are given in Appendix D.1. On inspecting these expressions, we find that the contributions from $m_*$ and $(\mathbf{v}_*, v_{**})$ are decoupled in objective function $d_{\mathrm{KL}} + \sum d_h$ within $d_1$, so the search for the optimal $m_*$ and $(\mathbf{v}_*, v_{**})$ are can be perform separately. Moreover, $d_{\mathrm{KL}} + \sum d_h$ is concave in $m_*$ and $v_{**}$ but not necessarily concave in $\mathbf{v}_*$. These findings are elaborated in Appendix D.2, which also gives the gradient updates for $\mathbf{m}_*$ and $v_{**}$ (given a fixed $\mathbf{v}_*$).

The non-concavity in $\mathbf{v}_*$ makes the maximization in $d_1$ less feasible. To make progress, we fix $\mathbf{v}_*$ to be that which maximizes only $d_{\mathrm{KL}}$. This gives $\mathbf{v}_* = VK^{-1}\mathbf{k}_*$ and leads to a second lower bound. This lower bound is non-trivial in the sense that it is non-negative. This is established in following lemma, which leads to another proposition.

**Lemma 16** *Let functions $d_1$, $d_{\mathrm{KL}}$ and $d_h^i$ be as defined in Equations 39, 40 and 41, and let*

$$d_2(\tilde{\mathbf{x}}_*|\tilde{X}) \overset{\mathrm{def}}{=} \max_{m_*, v_{**}} \left( d_{\mathrm{KL}}(m_*, v_{**}, VK^{-1}\mathbf{k}_*, \tilde{\mathbf{x}}_*|\tilde{X}) + \sum_{i=1}^{n} d_h^i(m_*, v_{**}, VK^{-1}\mathbf{k}_*, \tilde{\mathbf{x}}_*|\tilde{X}) \right). \tag{42}$$

*Then* $0 \le d_2(\tilde{\mathbf{x}}_*|\tilde{X}) \le d_1(\tilde{\mathbf{x}}_*|\tilde{X})$.

**Proof** Function $d_2$ is upper bounded by $d_1$ because it maximizes over a subset of the variational parameters in $d_1$. For non-negativity, we observe that the objective function within $d_2$ is zero when we set $m_* = \mathbf{k}_*^{\mathrm{T}} K^{-1}\mathbf{m}$ and $v_{**} = k_{**} - \mathbf{k}_*^{\mathrm{T}} K^{-1}\mathbf{k}_* + \mathbf{k}_*^{\mathrm{T}} K^{-1} V K^{-1}\mathbf{k}_*$. ■

**Proposition 17** *For the sparse variational approximation to the multinomial logit Gaussian process, any site added to the inducing set can never decrease the lower bound $\tilde{Z}_h^*$ to the marginal likelihood.*

This proposition is analogous one for Gaussian process regression (Titsias, 2009a, Proposition 1). Hence, we can interleave the greedy selection of inducing sites with hyper-parameters optimization (Titsias, 2009a, Section 3.1). One might have thought that this proposition is trivial because an additional inducing variable increases the flexibility of the variational model. Such an argument would have worked if we had compared the exact marginal likelihood $p(\mathbf{y})$ or the optimized variational lower bound $\tilde{Z}_B^*$. It would not have worked here because the optimized lower bound $\tilde{Z}_h^*$ is used here.

### 6.1.1 SUBSAMPLING AND FILTERING

Computation of $d_2$ for every possible site requires the full Gram matrix. This is because the required vector $\mathbf{k}_{\mathrm{f}*}$ for the site $\tilde{\mathbf{x}}_* = (\mathbf{x}_*, c)$ under consideration is the covariance from $\mathbf{x}_*$ to all the other observed data. This may be undesirable when covariance function is expensive to evaluate. In this case, we propose to approximate $\sum_{i=1}^{n} d_h^i$, which is over the whole data set, with one that is computed over a subset $\mathcal{S}$:

$$d_3(\tilde{\mathbf{x}}_*, \mathcal{S}|\tilde{X}) \overset{\mathrm{def}}{=} \max_{m_*, v_{**}} \left( d_{\mathrm{KL}}(m_*, v_{**}, VK^{-1}\mathbf{k}_*, \tilde{\mathbf{x}}_*|\tilde{X}) + \frac{n}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} d_h^i(m_*, v_{**}, VK^{-1}\mathbf{k}_*, \tilde{\mathbf{x}}_*|\tilde{X}) \right), \tag{43}$$

where $\mathcal{S}$ is the set of indices of the data to evaluate against. By partitioning the observed data appropriately, the number of covariance function evaluations can be reduced. This $d_3$ score can be used to directly choose the site to be added to the inducing set. Alternatively, it can be used as a filtering step so that only sites with high $d_3$ scores are further evaluated using the more expensive $d_2$ score function.

## 7. Computational Complexity

We now discuss the computation complexity of the approximate inference. For the non-sparse approximate inference, $O(n^3C^3)$ computations are required per iteration of the variational bound optimization, where $n$ is the size of the observed data set and $C$ is the number of classes. For the sparse approximate inference, $O(nC^3 + nCs^2 + nC^2s + s^3)$ computations are required per iteration, where $s$ is the number of inducing variables. This complexity is when we exploit the block diagonal structure of the variables. The complexity of computing the set of $n$ $d_2$-scores for active inducing set selection is the same. For probabilistic prediction with the posterior using sparsity, computing the lower bounds (24) to the predictive probabilities requires $O(C^3 + Cs^2 + C^2s)$ computations per datum, while computing the re-normalized probabilities (25) needs $O(Cs)$, which is less. For prediction with the posterior in the non-sparse case, these are $O(nC^3 + n^2C)$ and $O(nC^2)$ respectively.

One might have thought that complexity can be improved if $K^{\mathrm{c}} = I$ so that $K$ is block diagonal (after re-ordering) in the non-sparse approximation. However, we have not been able to exploit this structure. This is because computing $K^{-1} + W$ in Equation 28 involves $W$ that is block diagonal with a different ordering, which essentially destroys the structure (Seeger and Jordan, 2004).

For the sparse approximate inference, the direct complexity with respect to $n$ is linear. If we let $s \sim \log Cn$, then the overall complexity is $O(n \log^2 n)$ in $n$. Let us now consider three regimes depending on $C$. For $n \ll C$, we opine that some clustering process may be more appropriate than the classification model consider here. For $C \ll s$, the dominant complexity is $O(nCs^2)$. For $s \ll C \ll n$, the dominant complexity is $O(nC^3)$, which is for optimizing the variational parameters $\mathbf{b}_i$ and $S_i$ for each of the $n$ observed data.[2] Reducing the cubic complexity in $C$ requires constraining the variational parameters. In particular, one may constrain $S_i = (V_{\mathrm{f}i})^{-1} + W_i$ where $W_i = \gamma_i(\Pi_i - \boldsymbol{\pi}_i\boldsymbol{\pi}_i^{\mathrm{T}})$, $\gamma_i > 0$ and $\boldsymbol{\pi}_i$ is a probability vector. As remarked upon after Theorem 6, we have found that this constraint gives bounds that are quite loose. In addition, our present opinion is that effective inference with such a small inducing set may require rather strong correlations in both $K^{\mathrm{c}}$ and $k^{\mathrm{x}}(\cdot,\cdot)$ of the prior. We defer further investigation in the regime $s \ll C \ll n$ to future work.

In the $C \ll s$ regime, Seeger and Jordan (2004) and Girolami and Rogers (2006) have reported $O(nCs^2)$ computational complexity. In their cases, however, this complexity includes both the inference with a subset of the observed data and the active selection of the subset. Direct comparison with our approach can be misleading: the $O(nCs^2)$ in the preceding paragraph does not include active selection, but it does include projecting from the inducing variables to the entire set of observed data in the sparse approximate inference. Including the cost of greedy active selection up to $s$ inducing variables gives $O(nCs^3)$.

[2] In this regime, one should optimize the $\mathbf{b}_i$s separately. This is cheaper than optimizing $\mathbf{m}$ and $\mathbf{b}$ jointly.

| | Model | | Approximating Posterior | | Likelihood Approximation | |
|---|---|---|---|---|---|---|
| Citation | prior | likelihood | Family | Principle | Learning | Prediction |
| Williams and Barber (1998) | i.i.d. | logistic | Gaussian | Laplace | Exact | Monte Carlo |
| Neal (1998) | i.i.d. | logistic | Samples | MCMC | MCMC | MCMC |
| Gibbs (1997) | independent | logistic | Factored Gaussian | Variational | Variational | Analytic approximation |
| Seeger and Jordan (2004) | i.i.d. | logistic | Gaussian | ADF | Quadrature | Quadrature |
| Kim and Ghahramani (2006) | i.i.d. | uniform | Gaussian | EP | EP | |
| Girolami and Rogers (2006) | i.i.d. | probit | Factored Gaussian | Variational | Monte Carlo | Monte Carlo |
| This paper | separable | logistic | Gaussian | Variational | Variational | Variational |

Table 1: Existing works in multi-class Gaussian processes and their different aspects. In this paper, the likelihood approximation is for the expectation of the log-likelihood.

## 8. Related Work

We now discuss related works on multi-class Gaussian process classification. Table 1 tabulates different aspects of the existing related works that we know in the machine learning literature. Most consider the case where the latent functions are independent and identically distribution (i.i.d.), although Williams and Barber (1998) have seen no difficulty in extending to correlated latent functions. Gibbs (1997) has considered the case where the covariance functions of the prior latent Gaussian processes are independent and assumed to be from the same parametric family with possibility different parameters. In this paper, most results are applicable as long as the latent functions are jointly Gaussian, although at specific places we consider the separable covariance in Equation 1.

As with most existing works, our likelihood function is the multinomial logistic (2). Other likelihood functions are possible. In particular, one class of likelihood functions uses auxiliary independent random variables $u^c$s, $c = 1,\ldots,C$, and determine the class by $\arg\max_c u^c$, The multinomial logistic is in this class, and it is obtained when each auxiliary variable $u^c$ is Gumbel distributed with $p(u^c|f^c) = te^{-t}$, where $t \overset{\text{def}}{=} e^{-(u^c-f^c)}$ (McFadden, 1974). If $u^c \sim \mathcal{N}(f^c, 1)$, then the likelihood is the multinomial probit used by Girolami and Rogers (2006). If each auxiliary variable $u^c$ is supported only at $f^c$, we have the threshold likelihood function used by Kim and Ghahramani (2006). From this perspective, a model with the threshold likelihood function and prior covariance function $k^x(\mathbf{x},\mathbf{x}') + \delta(\mathbf{x},\mathbf{x}')$, where $\delta$ is the Kronecker delta function, is essentially the same as the model with the multinomial probit likelihood and prior covariance function $k^x(\mathbf{x},\mathbf{x}')$. Kim and Ghahramani (2006) have also used uniform noise (Angluin and Laird, 1988) with the threshold likelihood.

With any of these likelihoods, exact inference is non-tractable and approximations must be used. Except for the work of Neal (1998) where the approximation is a set of samples obtained from Markov Chain Monte Carlo (MCMC), all existing works have used a Gaussian approximation to the true posterior. The approximating Gaussian can be determined through fitting using different principles: Laplace (Williams and Barber, 1998), assumed density filtering (ADF, Seeger and Jordan, 2004) and expectation propagation (EP, Kim and Ghahramani, 2006), and variational bounding (Gibbs, 1997; Girolami and Rogers, 2006).

This paper uses the variational approach, for which a lower bound on the marginal likelihood can be obtained. However, the variational approach used in this paper differs from those in existing works (Gibbs, 1997; Girolami and Rogers, 2006). Gibbs has placed Gaussian-type bounds that factorizes over the classes on the multinomial logistic likelihood functions. Since the prior also factorizes over the classes, the approximate posterior factorizes similarly. Girolami and Rogers have constrained the approximating Gaussian to factorize over classes from the onset, and have proceeded to use variational mean field to obtain the factors. In contrast, the approximating Gaussian in this paper does not factorize over classes. We begin from an unconstrained Gaussian. This is followed by the Kullback-Leibler divergence and a bound on the expected log-likelihood (Theorem 6). Neither of these steps needs factorization over classes.

In general, the approximating Gaussian has covariance $(K^{-1}+W)^{-1}$, where $W$ is a block diagonal matrix of $n$ $C$-by-$C$ blocks. Let $W_i$ be the $i$th $C\times C$ block in $W$. The matrix $W$ is diagonal when the assumed likelihood factorizes over classes and data (Gibbs, 1997; Girolami and Rogers, 2006). Let $\boldsymbol{\pi}_i$ be a probability vector, and let $\Pi_i$ be the diagonal matrix with $\boldsymbol{\pi}_i$ along its diagonal. Then the $i$th block $W_i$ of $W$ in the Laplace approximation (Williams and Barber, 1998) is $\Pi_i - \boldsymbol{\pi}_i\boldsymbol{\pi}_i^{\mathrm{T}}$, where the $c$th element in $\boldsymbol{\pi}_i$ is the multinomial logistic $p(y_i^c|\mathbf{f}_i)$. This parametrization of $W_i$ follows directly from fitting principle of Laplace approximation. If computational time complexity is important, one can also use the parameterization $W_i \stackrel{\text{def}}{=} \gamma_i\left(\Pi_i - \boldsymbol{\pi}_i\boldsymbol{\pi}_i^{\mathrm{T}}\right)$, where $\gamma_i > 0$ and $\boldsymbol{\pi}_i$ are to be estimated, to obtain the same computational complexity as factorized mean-field (Seeger and Jordan, 2004). If computational time complexity is not a major consideration, any positive definite $W_i$ can be used for a tighter approximation (see Kim and Ghahramani, 2006, for example). In the present paper, each block $W_i$ is determined by optimizing the expected log-likelihood of the $i$th datum. The optimized $W_i$ has null space $\{\eta\mathbf{1} \mid \eta \in \mathbb{R}\}$. Further discussions in relation to the works of Williams and Barber (1998) and Seeger and Jordan (2004) have been given after Theorem 6.

The approximate predictive probability in the multi-class Gaussian process model is the expected likelihood under the approximate posterior. This is intractable and approximations are needed. Two common approaches are Monte Carlo (Williams and Barber, 1998; Neal, 1998; Girolami and Rogers, 2006) and traditional numerical integration (Seeger and Jordan, 2004). Analytic approximation has also been used (Gibbs, 1997). In this paper, we have given a variational approximation of the expected log-likelihood through Theorem 6. In Section 9.1.1, we will see that this is quite tight on average.

In previous works, sparsity in multi-class Gaussian processes is achieved by performing inference using only a subset of the observed data (SoD), which can be selected actively (Seeger and Jordan, 2004; Girolami and Rogers, 2006). In contrast, the sparsity in this paper is achieved through using the subset to induce the entire set. Quiñonero-Candela et al. (2007) have discussed in detail the SoD approach and the more general inducing approaches in the context of regression. We use the variational approach for both sparse approximation and active selection of the subset. For regression, this approach has been shown to have several desirable characteristics over the other approaches (Titsias, 2009a).

## 9. Experiments and Results

We evaluate our approach to multi-class logit Gaussian process classification in various aspects. In Section 9.1, we compare the bounds in the marginal likelihood and the predictive likelihood provided by our variational approach with those provided by the variational mean-field approximation

| Name | train set | test set | classes | attributes | task description |
|---|---|---|---|---|---|
| iris | 90 | 60 | 3 | 4 | determine class of iris plant |
| thyroid | 129 | 86 | 3 | 5 | diagnosis of a patient's thyroid |
| wine | 106 | 72 | 3 | 13 | determine the cultivar of wine |
| glass | 128 | 86 | 7 | 9 | determine the type of glass |

Table 2: Summary of the four UCI data sets used in our experiments. For the glass data, there is no instance of class "vehicle windows that are non-float processed" in the set.

to multinomial probit regression (Girolami and Rogers, 2006).[3] We also look at how the quality of our bounds vary with the prior variance of the latent process. In Section 9.2, we relate the logit to the probit, and we also look at the the prior correlation between the latent process. Section 9.3 investigates the effectiveness of active inducing set selection using the criteria proposed in Section 6.1. In Section 9.4, we compare with single-machine multi-class support vector machines.

For comparison, we use a tight approximation to the exact posterior of the multi-class logit and probit Gaussian process model. This is obtained by importance sampling where the proposal is the multivariate-$t$ distribution (Kotz and Nadarajah, 2004) with four degrees of freedom, centered at the mean $\mathbf{m}^*$ of our variational approximation to $p(\mathbf{f}|\mathbf{y})$ and with covariance $2K$. We have found this to be more effective than the Gibbs sampling used by Girolami and Rogers (2006) and the anneal importance sampling (Neal, 2001) used by Nickisch and Rasmussen (2008). Due to the central limit theorem, the Monte Carlo estimate $\hat{p}(\mathbf{y})$ on the marginal likelihood has distribution $\mathcal{N}(p(\mathbf{y}), \sigma^2/n_s)$, where $\sigma^2$ is the true variance of the importance weights and $n_s$ is the number of samples. When reporting the marginal likelihood estimate, we use $\hat{p}(\mathbf{y}) + 3.09\sigma/\sqrt{n_s}$ to upper bound $p(\mathbf{y})$ with probability 0.999, where $\sigma$ is estimated from the samples. In our experiments, $n_s = 100,000$ for each Monte Carlo run. Details are in Appendix F.[4]

Our experiments are conducted on four data sets from the UCI Machine Learning Repository (Frank and Asuncion, 2010): *iris*, *thyroid*, *wine* and *glass*. Following Girolami and Rogers (2006), for each data set, 60% is used for training and 40% for testing. Each input attribute is normalized to zero mean and unit variance on the training set. Our experiments are conducted with fifty such random splits for each data set. The summary statistics for the data sets are given in Table 2.

### 9.1 Comparing Variational Approaches

We evaluate our approach against variational mean-field (Girolami and Rogers, 2006) and importance sampling. We fix the latent random functions to be independent and identically distributed (i.i.d.); that is, $K^{\mathrm{c}} = I$. The covariance function on inputs $\mathbf{x}$ and $\mathbf{x}'$ in $\mathbb{R}^d$ is the unit variance squared-

[3] We use version 1.6.0 of the *R* package available at `http://www.bioconductor.org/packages/devel/bioc/html/vbmp.html`.

[4] We have also experimented with using the approximate posterior $q(\mathbf{f}|\mathbf{y})$ directly as the proposal distribution (Ghahramani and Beal, 2000a,b). The set of estimates obtained in this way is generally indistinguishable from that obtained using the multivariate-$t$ distribution. However, the latter comes with a convergence rate guarantee because the tail of the $t$ distribution is heavier than that of a Gaussian.

exponential covariance function

$$k^{\mathrm{x}}(\mathbf{x},\mathbf{x}') = \mathrm{usqexpard}(\mathbf{x},\mathbf{x}';\boldsymbol{\theta}) \overset{\mathrm{def}}{=} \exp\left(-\frac{1}{2}\sum_{j=1}^{d}\left(\frac{x_j - x'_j}{\theta_j}\right)^2\right), \tag{44}$$

where $x_j$ (resp. $x'_j$) is the $j$th dimensional of $\mathbf{x}$ (resp. $\mathbf{x}'$), $\theta_j$ is the length-scale of the $j$th dimension, and $\boldsymbol{\theta} \overset{\mathrm{def}}{=} (\theta_1,\ldots,\theta_d)^{\mathrm{T}}$ collects parameters. Each length-scale affects the influence of the dimension; this allows automatic relevance determination (ARD, Neal 1996).

Table 3 reports the results comparing the log marginal likelihood $\log Z$ or its bounds on the training sets, and the predictive error and the log joint predictive probability $\log p(\mathbf{y}_*)$ on the test sets. The means and standard deviations over the fifty partitions for each data set are given. We use MNL to denote multinomial logistic likelihood and MNP to denote multinomial probit likelihood. KL-MNL is our variational approach with multinomial logistic likelihood, and MF-MNP is the variational mean-field with multinomial probit likelihood (Girolami and Rogers, 2006). MC-MNP and MC-MNL are the Monte Carlo approximations using importance sampling. The marginal likelihood estimates for importance sampling are the high confidence upper bounds. Column *Theorem 10* gives theoretic lower bounds on the marginal likelihood for the logistic likelihood. Results with two sets of hyper-parameters are given: one from the variational mean parameter estimation for MF-MNP (Girolami and Rogers, 2006), and the other from the model learning for KL-MNL (Section 6). With either set of hyper-parameters, the prior latent process has unit variance due to the choice of covariance function. Our results for MC-MNP are consistent with, but tighter than, those reported by Girolami and Rogers (2006, Table 1, column *Gibbs Sampler*).

We first compare the marginal likelihoods on the training data along the rows headed by $\log Z$. For either set of hyper-parameters, our variational approach (KL-MNL) gives lower bounds that are very close to the high confidence upper bounds on the marginal likelihoods obtained by sampling (MC-MNL). In fact, it is this tightness that leads us to finally use importance sampling: with Gibbs sampling and annealed importance sampling, we are unable to obtain estimates that are larger than our lower bounds. For MF-MNP, we find that it consistently gives lower bounds that are looser than the theoretical ones under column *Theorem 10*. This suggests that the theoretic bounds may be useful as sanity checks for variational approaches, although we must qualify that the theoretic bounds are for the multinomial logistic likelihood rather than for the probit one. Using the reasoning to be outlined in Section 9.2.1, we obtain approximate theoretic bounds for the probit using $\sigma^2 = \pi^2/6$ in Theorem 10, and these bounds are $-139.84$, $-200.44$, $-164.70$ and $-331.06$ for the iris, thyroid, wine and glass data sets respectively. These are lower than the theoretical ones in Table 3, but are still higher the bounds given by MF-MNP. The looseness of the bounds given by MF-MNP is also evident when compared with the estimates from sampling (MC-MNP).

Next, we compare the log joint predictive probability $\log p(\mathbf{y}_*)$ on the test set. One set of results obtained using Monte Carlo is reported for MF-MNP, MC-MNP and MC-MNL. Two sets of results are reported for KL-MNL: the upper set uses the re-normalized probabilities given by Equation 25 and the lower set uses Equation 24. As discussed in Section 2.3.3, the probabilities based on Equation 24 are approximate lower bounds on the exact predictive probabilities, while the re-normalized probabilities are always larger than these lower bounds. In the table, we see that these two sets of predictive probabilities from KL-MNL bound the probabilities from MC-MNL rather tightly. For MF-MNP, its predictive probabilities are close to those given by sampling (MC-MNP). This suggests that variational mean-field, which couples of posterior means of the latent function

|  |  | MF-MNP-$\boldsymbol{\theta}$ |  |  |  | KL-MNL-$\boldsymbol{\theta}$ |  |
|---|---|---|---|---|---|---|---|
|  | Theorem 10 | MF-MNP | MC-MNP | KL-MNL | MC-MNL | KL-MNL | MC-MNL |
| **Iris** |  |  |  |  |  |  |  |
| $\log Z$ | $-125.28$ | $-187.32 \pm 1.71$ | $-27.71 \pm 1.74$ | $-32.63 \pm 1.60$ | $-32.48 \pm 1.60$ | $-31.46 \pm 1.36$ | $-31.32 \pm 1.37$ |
| Error |  | $2.66 \pm 1.27$ | $2.64 \pm 1.26$ | $2.66 \pm 1.26$ | $2.64 \pm 1.22$ | $2.28 \pm 1.25$ | $2.28 \pm 1.25$ |
| $\log p(\mathbf{y}_*)$ |  | $-10.35 \pm 1.87$ | $-10.53 \pm 1.89$ | $-11.27 \pm 1.79$ | $-12.61 \pm 1.70$ | $-9.45 \pm 1.38$ | $-10.90 \pm 1.35$ |
|  |  |  |  | $-13.15 \pm 1.82$ |  | $-11.38 \pm 1.48$ |  |
| **Thyroid** |  |  |  |  |  |  |  |
| $\log Z$ | $-179.57$ | $-270.63 \pm 3.60$ | $-41.54 \pm 3.70$ | $-47.15 \pm 3.49$ | $-46.97 \pm 3.49$ | $-45.13 \pm 2.85$ | $-44.95 \pm 2.85$ |
| Error |  | $7.84 \pm 2.54$ | $7.86 \pm 2.54$ | $7.92 \pm 2.75$ | $8.00 \pm 2.85$ | $6.44 \pm 2.92$ | $6.52 \pm 3.03$ |
| $\log p(\mathbf{y}_*)$ |  | $-22.02 \pm 4.57$ | $-22.10 \pm 4.61$ | $-23.08 \pm 4.58$ | $-24.29 \pm 4.27$ | $-20.55 \pm 4.58$ | $-22.13 \pm 4.19$ |
|  |  |  |  | $-25.67 \pm 4.66$ |  | $-23.85 \pm 4.64$ |  |
| **Wine** |  |  |  |  |  |  |  |
| $\log Z$ | $-147.56$ | $-222.63 \pm 1.91$ | $-36.41 \pm 2.07$ | $-42.56 \pm 1.96$ | $-42.38 \pm 1.96$ | $-41.18 \pm 1.74$ | $-41.01 \pm 1.74$ |
| Error |  | $4.88 \pm 2.74$ | $4.88 \pm 2.88$ | $4.96 \pm 2.73$ | $4.94 \pm 2.75$ | $3.22 \pm 1.83$ | $3.22 \pm 1.73$ |
| $\log p(\mathbf{y}_*)$ |  | $-16.19 \pm 3.84$ | $-16.27 \pm 3.94$ | $-17.19 \pm 3.83$ | $-19.01 \pm 3.59$ | $-14.47 \pm 2.03$ | $-16.74 \pm 1.98$ |
|  |  |  |  | $-20.37 \pm 3.96$ |  | $-18.02 \pm 2.28$ |  |
| **Glass** |  |  |  |  |  |  |  |
| $\log Z$ | $-300.61$ | $-827.58 \pm 6.46$ | $-150.23 \pm 6.88$ | $-158.16 \pm 5.77$ | $-157.53 \pm 5.79$ | $-154.74 \pm 5.08$ | $-154.08 \pm 5.04$ |
| Error |  | $33.72 \pm 4.03$ | $36.00 \pm 4.16$ | $34.20 \pm 4.03$ | $34.40 \pm 4.09$ | $32.62 \pm 4.09$ | $33.02 \pm 4.05$ |
| $\log p(\mathbf{y}_*)$ |  | $-89.63 \pm 6.15$ | $-95.62 \pm 8.78$ | $-92.78 \pm 6.09$ | $-94.79 \pm 5.50$ | $-88.82 \pm 5.49$ | $-91.31 \pm 5.00$ |
|  |  |  |  | $-101.00 \pm 5.93$ |  | $-97.97 \pm 5.58$ |  |

Table 3: Results with the usqexpard covariance function (44) on inputs and with i.i.d. latent functions. The log marginal likelihood $\log Z$ (or its bounds), the empirical error and the log joint predictive probability $\log p(\mathbf{y}_*)$ (or its bounds and approximations) are reported with means and standard deviations over 50 partitions. Theoretic lower bounds are given under Theorem 10. Methods with MNP after the dash uses of the multinomial probit likelihood, while those with MNL uses the multinomial logistic likelihood. MF-MNP is the variational mean-field method (Girolami and Rogers, 2006), KL-MNL is the variational approach of this paper, while MC-MNP and MC-MNL are importance sampling. Columns under MF-MNP-$\boldsymbol{\theta}$ use the estimated mean hyper-parameters for MF-MNP; those under KL-MNL-$\boldsymbol{\theta}$ use the hyper-parameters optimized for KL-MNL. Method KL-MNL reports two sets of approximations to $\log p(\mathbf{y}_*)$: the upper set uses the re-normalized probabilities given by Equation 25 and the lower set uses the lower bound in Equation 24.

values, is perhaps sufficient for accurate predictive probabilities. In addition, the figures suggest that the predictive probabilities by MF-MNP upper bound those by MC-MNP. Further analysis is needed to confirm this for the general case.

Finally, we compare errors on test data. For the MF-MNP-$\boldsymbol{\theta}$ hyper-parameters, we find the errors to be similar across all methods, although KL-MNL and MC-MNL, which use the multinomial logistic likelihood, give marginally more errors. However, with hyper-parameters optimized for KL-MNL, both KL-MNL and MC-MNL give less errors consistently. This suggests that model learning is better performed with a tight approximation to the marginal likelihood, as is provided by KL-MNL.

### 9.1.1 EFFECT OF PRIOR VARIANCE

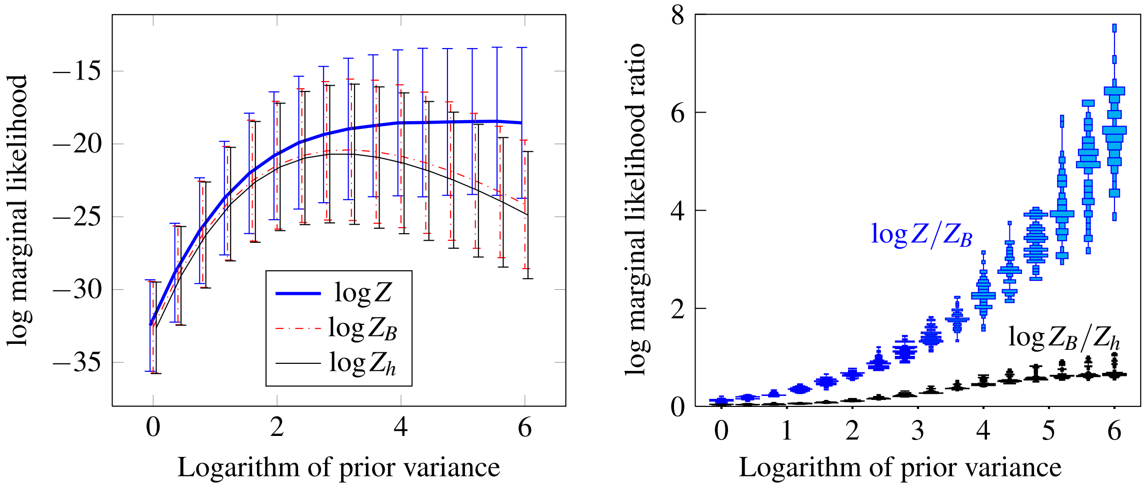
The results in Table 3 are where the latent processes have unit prior variance. Using the iris data set, we investigate the quality of our marginal likelihood bound when the prior variance increases. For each random partition, we fix the ARD hyper-parameters to that estimated for MF-MNP. The prior variance is then increased in steps. For each step, we obtain the marginal likelihoods using our variational inference and using importance sampling. The former is denoted by $Z_h$, and the latter by $Z$. Using Equation 3, we also obtain $Z_B$, which approximates the posterior with a Gaussian but computes the expected log-likelihood exactly. The Kullback-Leibler divergence is computed exactly, while $\mathcal{L} \overset{\text{def}}{=} \sum_{i=1}^{n} \ell_i(\mathbf{y}_i; q)$ is computed with Monte Carlo using $n_s = 100,000$ samples. For a sample $\mathbf{f}^{(s)}$ from the variational posterior, let $w^{(s)} \overset{\text{def}}{=} \sum_{i=1}^{n} \log p(\mathbf{y}_i|\mathbf{f}_i^{(s)})$. Then the Monte Carlo estimate of $\mathcal{L}$ is the sample mean $\bar{w}$ of the $w^{(s)}s$. We use the 99.9% confidence upper bound on $Z_B$ by estimating $\mathcal{L}$ with $\bar{w} + 3.09\sigma/\sqrt{n_s}$, where $\sigma^2$ is the sample variance of the $w^{(s)}s$.[5]

Figure 1 gives plots against the prior variance of the latent process. The left figure (a) plots the log marginal likelihoods while the right figure (b) gives the violin plots of the log ratios of the marginal likelihoods. The plots show that the quality of the bounds $Z_B$ and $Z_h$ decreases with prior variance. This is largely due to the Gaussian approximation to the posterior, which is given by $Z_B$, rather than the approximation $h$ to the expected likelihood: the violin plot for $\log Z_B/Z_h$ shows only slight increase as the prior variance increase, while the violin plot for $\log Z/Z_B$ increases more significantly. This illustrates the robustness of the approximation $h$ to the expected log-likelihood. The deterioration of Gaussian approximation to the posterior is also present in binary Gaussian process classification (Nickisch and Rasmussen, 2008, Figure 3). The intuition is that a higher prior variance allows the posterior latent process more flexibility to become less Gaussian.

## 9.2 Comparing Models

The availability of fairly tight approximations to the exact posterior opens the opportunity for model comparison on each model's own merit without being confounded by the gap that results from approximation. In this section, we investigate the Gaussian process models in two areas: the choice of likelihood and the choice of prior correlation between the functions.

[5] Since log-probability is unbounded, the true distribution of $w^{(s)}$ may not have finite variance. We eliminate this possibility empirically by verifying that the running sample variance has converged and that the estimated tail is not heavy (Koopman et al., 2009).



(a) Line plot of marginal likelihood

(b) Violin plot of log ratios of marginal likelihood

Figure 1: The quality of the lower bound on marginal likelihood for the iris data set by fixing the ARD hyper-parameters to those estimated for MF-MNP-$\boldsymbol{\theta}$ and then increasing the prior variance. Figure (a) plots the log marginal likelihood against the prior variance (on log-scale). The topmost curve $\log Z$ is for the marginal likelihood obtained using importance sampling; the middle curve $\log Z_B$ approximates the posterior of the latent process with a Gaussian; while the bottom curve $\log Z_h$ further approximates the expected log-likelihood. The error bars give 95% confidence interval computed over 50 random partitions of the data. The curves are translated slightly horizontally to reduce overlap in the error bars. Figure (b) plots the log marginal likelihood ratios against the prior variance (on log-scale) using the violin plot. The upper violin plot is for $\log Z/Z_B$ and the lower one is for $\log Z_B/Z_h$. These figures illustrate that the bound $\log Z_h$ becomes looser with increase in prior variance, and that this is mainly contributed by the Gaussian approximation to the posterior.

### 9.2.1 LIKELIHOOD

In Table 3, when we compare MC-MNP and MC-MNL on the set of hyper-parameters given by MF-MNP-$\boldsymbol{\theta}$, we see that the multinomial probit likelihood (MNP) fits the four training data sets better than the multinomial logistic likelihood (MNL). This difference can be explained by an equivalent model for each likelihood.

As outlined in Section 8, the Gaussian process latent model with covariance function $k^{\mathrm{x}}(\mathbf{x},\mathbf{x}')$ on the inputs and with the multinomial probit likelihood is equivalent to the model with covariance function $k^{\mathrm{x}}(\mathbf{x},\mathbf{x}') + \delta(\mathbf{x},\mathbf{x}')$ and the threshold likelihood function. This threshold likelihood partitions the space $\mathbb{R}^{nC}$ into orthants, one of which corresponds to the vector of observed data classes $\mathbf{y}$. Let us call this the $\mathbf{y}$-orthant. The marginal likelihood is hence the fraction of the prior probability mass in the $\mathbf{y}$-orthant. For a centered Gaussian prior, this fraction is determined by the correlation.

For multinomial logit likelihood, each auxiliary variable $u^c$ is Gumbel distributed around $f^c$ instead of Gaussian distributed; see Section 8. A moment matching approximation to the distribution of $u^c$ is a Gaussian cent-red at $f^c$ with variance $\pi^2/6$. Hence an approximation to the logit model is a Gaussian process latent model with covariance $k^{\mathrm{x}}(\mathbf{x},\mathbf{x}') + \delta(\mathbf{x},\mathbf{x}')\pi^2/6$ and with the threshold likelihood. As before, the marginal likelihood is the prior probability mass in the $\mathbf{y}$-orthant, and this is determined by the correlation.

The correlation functions of the equivalent models for the multinomial probit and multinomial logistic likelihoods are different. The former is obtained by removing the variance in $k^{\mathrm{x}}(\mathbf{x},\mathbf{x}') + \delta(\mathbf{x},\mathbf{x}')$, while the latter, $k^{\mathrm{x}}(\mathbf{x},\mathbf{x}') + \delta(\mathbf{x},\mathbf{x}')\pi^2/6$. One way to match the two correlation functions is to scale the original latent Gaussian process for the logit model by $\pi^2/6$, so that the equivalent covariance function becomes $\pi^2/6[k^{\mathrm{x}}(\mathbf{x},\mathbf{x}') + \delta(\mathbf{x},\mathbf{x}')]$. Consulting Figure 1a, the mean exact log marginal likelihood for the logit model on the iris data set is $\log Z \approx -0.28$ at $\log(\pi^2/6) \approx 1/2$ on the $x$-axis. This is consistent with the $-27.82$ under MC-MNP for the iris data set in Table 3.

### 9.2.2 PRIOR CORRELATION AMONG LATENT PROCESSES

It is common to assume prior independence among the latent functions for two reasons: to reduce computational complexity and to adhere to the principle of parsimony. In this section, we investigate if parsimony is a reason enough to exclude considering prior dependence among the latent functions. We evaluate on the four UCI data sets using the separable covariance structure in Equation 1.

For this evaluation, the covariance function on the inputs in $\mathbb{R}^d$ is the squared-exponential covariance function with equal length-scales along all the dimensions:

$$k^{\mathrm{x}}(\mathbf{x},\mathbf{x}') = \mathrm{sqexpiso}(\mathbf{x},\mathbf{x}';\sigma_{\mathrm{x}},\theta) \overset{\mathrm{def}}{=} \sigma_{\mathrm{x}}^2 \exp\left(-\frac{1}{2}\sum_{j=1}^{d}\left(\frac{x_j - x'_j}{\theta}\right)^2\right). \tag{45}$$

We consider five models $\mathcal{M}_1,\ldots,\mathcal{M}_5$ of covariances $K^{\mathrm{c}}$ between the latent functions for the classes. In each model, we keep the total variance $\mathrm{tr}\,K^{\mathrm{c}}$ to be constant at $C$.[6] The first model $\mathcal{M}_1$ is where the latent functions are i.i.d., so $K^{\mathrm{c}}$ is the $C$-by-$C$ identity matrix. The second is a diagonal matrix where the diagonal entries are positive and sum to $C$. Here, the latent functions remain a-priori independent, but they can have different variances while keeping the total variance the same as the first model. For the third and fourth models, we scale the $K^{\mathrm{c}}$ given by Equation 37 to have the same

[6] The total variance in $K^{\mathrm{c}}$ is then scaled by the $\sigma_{\mathrm{x}}^2$ in $k^{\mathrm{x}}$, so the total variance of the latent process at each datum is $C\sigma_{\mathrm{x}}^2$.

total variance as the first two models:

$$\tilde{K}^{\mathrm{c}} = M - M\mathbf{1}\mathbf{1}^{\mathrm{T}}M/\mathbf{1}^{\mathrm{T}}M\mathbf{1} + I, \qquad\qquad K^{\mathrm{c}} = \frac{C}{\operatorname{tr}\tilde{K}^{\mathrm{c}}}\tilde{K}^{\mathrm{c}}. \tag{46}$$

The equation for $\tilde{K}^{\mathrm{c}}$ omits the weight for the identity matrix because the normalization in $K^{\mathrm{c}}$ makes this unnecessary. Model $\mathcal{M}_3$ sets $M$ to be diagonal with positive diagonal entries; this is an attempt to approximate the correlation of multinomial or Dirichlet random variables. Model $\mathcal{M}_4$ allows $M$ to be any positive semi-definite matrix. Finally, in the fifth model, $K^{\mathrm{c}}$ is any positive definite matrix with $\operatorname{tr}K^{\mathrm{c}} = C$; this allows the latent functions to be correlated arbitrarily.

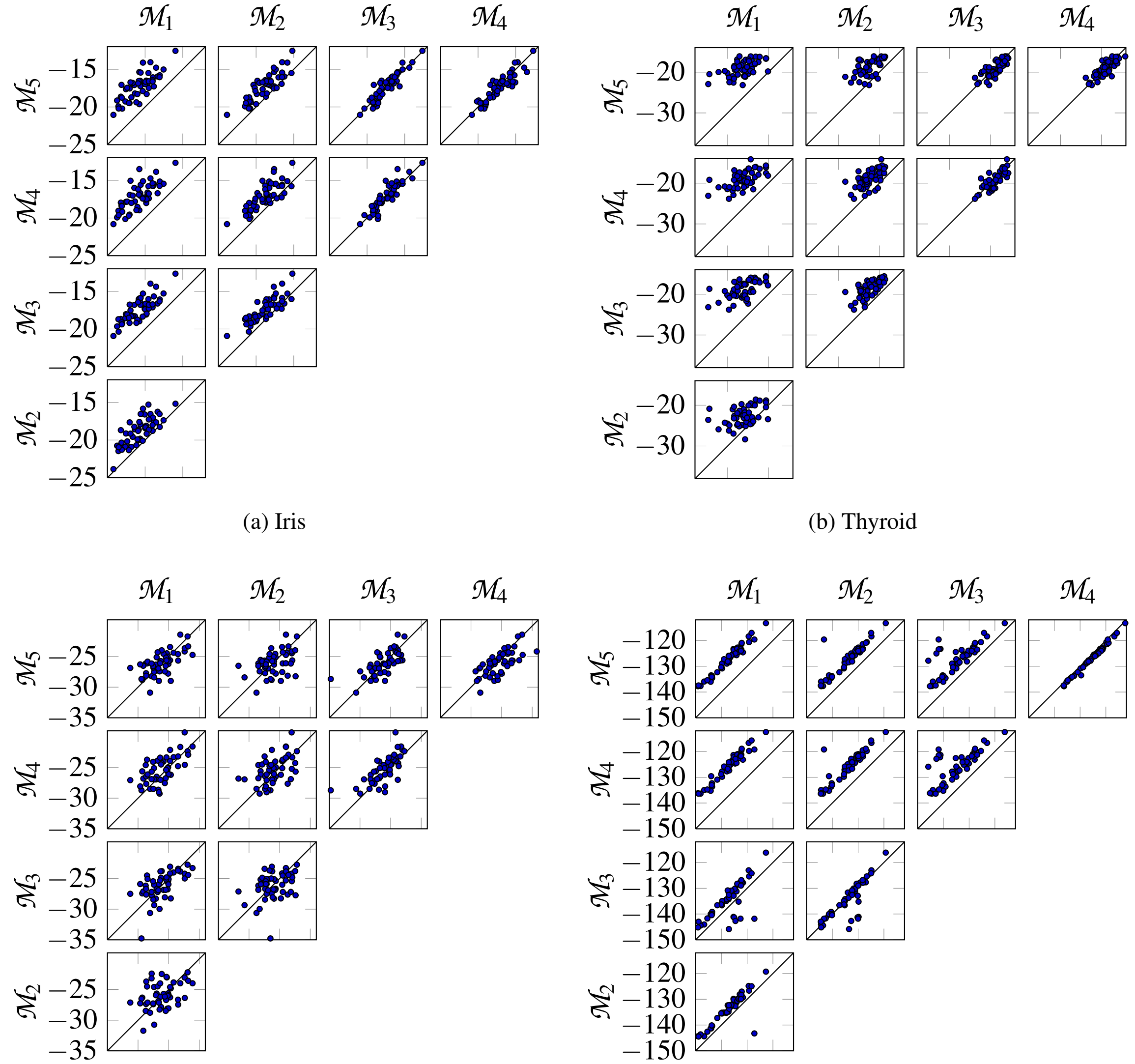
The first, third and fourth models satisfy the sum-to-zero property discussed in Section 5. Using $\mathcal{M}_i \supset \mathcal{M}_j$ to indicate that $\mathcal{M}_i$ is more expressive than $\mathcal{M}_j$, we have the ordering $\mathcal{M}_5 \supset \mathcal{M}_4 \supset \mathcal{M}_3 \supset \mathcal{M}_1$ and $\mathcal{M}_5 \supset \mathcal{M}_2 \supset \mathcal{M}_1$. The second model is not comparable with the third and fourth models in this ordering. The $K^{\mathrm{c}}$ for $\mathcal{M}_1$ is fixed and therefore parameter-free. The number of free parameters of $K^{\mathrm{c}}$ in models $\mathcal{M}_2$ and $\mathcal{M}_3$ are $C-1$. For $\mathcal{M}_4$ and $\mathcal{M}_5$, these are $C(C+1)/2-2$ and $C(C+1)/2-1$.

We estimate the parameters of the models in the following way. First, the hyper-parameters $\sigma_{\mathrm{x}}$ and $\theta$ for model $\mathcal{M}_1$ are optimized for the variational bound on marginal likelihood on the observed data. These two hyper-parameters are then considered fixed when optimizing matrix $K^{\mathrm{c}}$ for the other models using the variational bound.

After the hyper-parameters for each model have been estimated, we use sampling to obtain better estimates of the marginal likelihoods, errors and predictive likelihoods given the model and its hyper-parameters. This is done for each of the fifty partitions of the four UCI data sets. The sampling procedure is that outlined in the introduction to this section except for the glass data set, for which the Monte Carlo estimates to the marginal likelihood are lower than the variational lower bounds. There are two reasons for the lower estimates: (a) the sampling space is larger than in the other data sets because this data set has seven classes; and (b) for each data set partition, the prior variance is around 16, so the true posterior is conceivably less Gaussian and more different from the prior. To obtain Monte Carlo estimates that are better than the variational ones for this data set, we instead sample from the multivariate-$t$ distribution that has covariance $2V$ instead of $2K$, where $V$ is the covariance of the variational approximation.

Figure 2 gives a paired comparison between the models $\mathcal{M}_1, \ldots, \mathcal{M}_5$ based on their marginal likelihoods on the observed data. There are four sub-figures, one for each data set. Each graph is a scatter-plot, in which each point is for one partition of the data set named in the sub-caption, and the location of each point is the log marginal likelihoods of the two models indicated on the top and the left edge of the sub-figure. For a scatter-plot, if the points are mostly above the diagonal line, then model named on the left edge is better than the model named on the top edge. From Figure 2, we see no noticeable difference among the models for the wine data set, while we make the following four observations for the other data sets. (a) More free parameters generally results in better marginal likelihoods, as expected.[7] (b) Although $\mathcal{M}_2$ and $\mathcal{M}_3$ have the same number of free parameters, $\mathcal{M}_3$ generally gives better marginal likelihoods. (c) The marginal likelihoods of $\mathcal{M}_4$ and $\mathcal{M}_5$ are similar, showing that the additional free parameter in $\mathcal{M}_5$ over $\mathcal{M}_4$ is not useful. Observations (b) and (c) suggest that it is worthwhile to consider the sum-to-zero constraint, which is satisfied by $\mathcal{M}_3$ and $\mathcal{M}_4$ but not by $\mathcal{M}_2$ and $\mathcal{M}_5$.

[7] There are two reasons why more free parameters is *not always* better. First, the hyper-parameters are optimized using the variational approximations and not the true marginal likelihoods. Although our approximations are rather tight, the hyper-parameters may be sensitive to the remaining gaps in the approximations. Second, the marginal likelihood surface can be multi-modal with respect to the hyper-parameters, hence gradient ascent can be stuck at local maxima.



(a) Iris

(b) Thyroid

(c) wine

(d) glass

Figure 2: Paired comparisons of the marginal likelihood between five models of prior correlations $K^c$ between the latent functions: $\mathcal{M}_1$ gives i.i.d. latent functions; $\mathcal{M}_2$ gives independent latent functions with different variances; $\mathcal{M}_3$ allows the functions to be correlated using Equation 46 where $M$ is diagonal; $\mathcal{M}_4$ also uses Equation 46 but allows $M$ to be any positive semi-definite matrix; and $\mathcal{M}_5$ allows the functions to be correlated arbitrarily. For each model, $K^c$ is scaled such that the total variance $\operatorname{tr} K^c$ is constant $C$. Each figure is for the data set indicated in its caption. Each graph in a figure plots the log marginal likelihood ($\log Z$) of the model named at the left edge of the figure versus that named at the top edge. Each point in the scatter-plot is for one of the fifty random partitions of the data set. To ease comparison, the $x = y$ line is plotted in each graph. For example, each point in top left graph of Figure (a) is at the location $(x, y)$, where $x$ (resp. $y$) is the $\log Z$ for $\mathcal{M}_1$ (resp. $\mathcal{M}_5$) on one partition of the iris data set. All the points in this graph are above the $x = y$ line, so $\mathcal{M}_5$ gives better marginal likelihood than $\mathcal{M}_1$ for all the partitions.

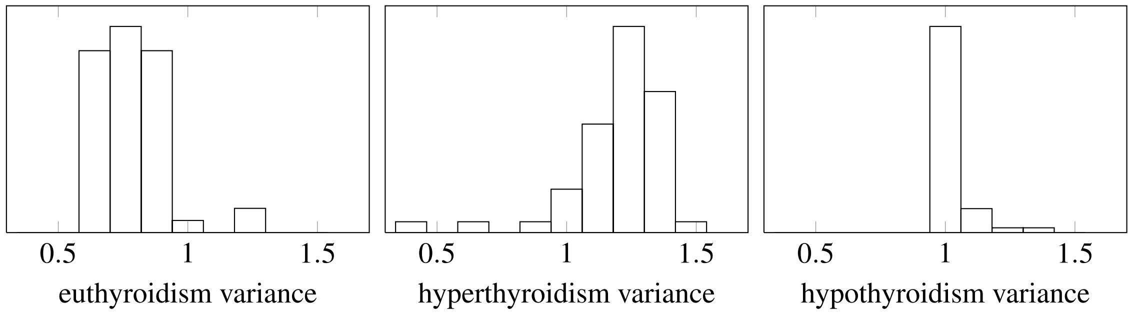


Figure 3: The histograms of the prior variances of the latent functions in $K^c$ estimated from the thyroid data under model $\mathcal{M}_4$. From left to right, we have histograms for euthyroidism, hyperthyroidism and hypothyroidism.

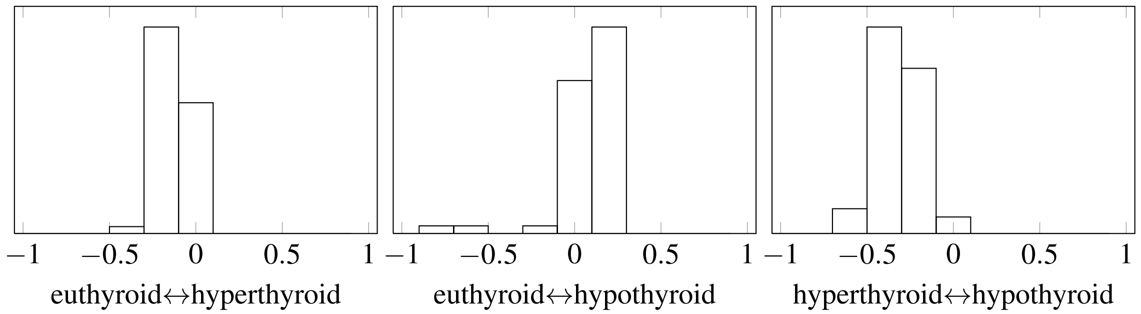


Figure 4: The histograms of the prior correlation of the latent functions estimated from the thyroid data under model $\mathcal{M}_4$. From left to right, we have histograms for the correlation between euthyroidism and hyperthyroidism, between euthyroidism and hypothyroidism, and between hyperthyroidism and hypothyroidism.

For the errors and predictive likelihoods, we observe no significant difference among the models, based on plots similar to those in Figure 2. For the purpose of prediction then, it seems sufficient to rely on the likelihood to provide the necessary posterior coupling between the latent functions. Nonetheless, we believe there may still be applications where prior coupling between the latent functions is helpful in prediction.

More insights into the various models can be obtained by looking at the estimated variances and correlations between the latent functions. As an example we shall use the thyroid data with model $\mathcal{M}_4$; examination with model $\mathcal{M}_3$ or model $\mathcal{M}_5$ gives similar conclusions. The task for the thyroid data is to predict the state of a subject's thyroid given the results of five different laboratory tests. This state can be one of three classes: euthyroidism (having normal functioning thyroid), hyperthyroidism (having overactive thyroid) and hypothyroidism (having underactive thyroid). Figure 3 gives the histogram of the prior variances of the latent functions in $K^c$ estimated by $\mathcal{M}_4$ over the fifty different partitions. From left to right, the histograms are for euthyroidism, hyperthyroidism and hypothyroidism. Each histogram is concentrated around a single mode. Bearing in mind that we have constrained the total variance in $K^c$ to be 3, the evidence in the data suggests that class hy-

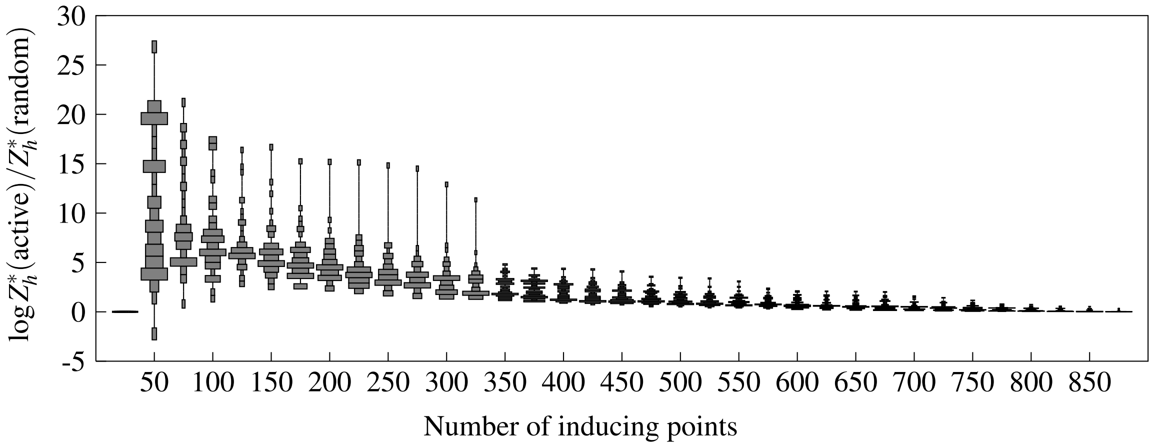


Figure 5: The performance of active selection over random selection as the number of inducing point is incremented in steps of 25. Each histogram is for the log ratios of $Z_h^*$ for active selection to random selection. The plot shows that active selection is better than random selection, though the advantage decreases with the number of inducing points.

perthyroidism varies more than class hypothyroidism, which varies more than class euthyroidism. Figure 4 gives the histogram of the correlations between the latent functions. From left to right, we have the correlations between euthyroidism and hyperthyroidism, between euthyroidism and hypothyroidism, and between hyperthyroidism and hypothyroidism. As for the variances, each histogram is concentrated around a single mode. The most significant result is the right histogram, which shows hyperthyroidism and hypothyroidism to be negatively correlated. This is intuitive: the two classes correspond to overproduction and underproduction of thyroid hormones respectively.

### 9.3 Active Inducing Set Selection

Sparse approximation is commonly used for efficient inference in large data sets. The quality of this approximation is dependent on the inducing set. In this section, we evaluate the effectiveness of criteria $d_2$ and $d_3$ (Equations 42 and 43 in Section 6.1) in selecting the inducing set actively. We do this by comparing with random selection on the glass data set.

We use the usqexpard covariance function (44) on the inputs and assume that the latent functions are i.i.d. For each training set partition, we fix the hyper-parameters to those optimized for our variational lower bound using the entire training set; these are the KL-MNL-$\boldsymbol{\theta}$ hyper-parameters estimated in Section 9.1. Given the training set partition and the hyper-parameters, the random approach selects $\delta$ sites to be added to the inducing set at each iteration. The active approach begins with the same $\delta$ random sites as the random approach, but subsequent choices of the $\delta$ sites are selected based on the $d_2$ and $d_3$ criteria. For each random variable induced by the training set, we use $d_3$ with subsample set of size $|\mathcal{S}|$. Next, we compute the $d_2$ scores of the $t$ random variables with the highest $d_3$ scores. The $\delta$ random variables having the highest $d_2$ scores computed in this manner will be added to the inducing set. We use $\delta = 25$, $|\mathcal{S}| = 5$ and $t = 40$ in the experiment.

The optimized variational lower bound $Z_h^*$ on the marginal likelihood for random selection is computed at each iteration for each training set partition. The same is computed for active selection. Figure 5 gives the violin plot of the log ratios of the $Z_h^*$ for the active selection to the $Z_h^*$ for the random selection. The horizontal axis gives the size of the inducing set—there are 896 potential inducing sites from the 7 types of glass and the 128 training **x**s. The histogram at each iteration is over the fifty random training set partitions.

At the first iteration with 25 inducing sites, the ratio is zero because both the random selection and the active selection begin with the same 25 random sites. At the second iteration with an additional 25 inducing sites, active selection usually provides higher $Z_h^*$, but it is possible for active selection to be worse than random selection. This is because the $d_2$ and $d_3$ criteria are designed for single inducing sites, so they are not optimal for selecting more than one site—25 in this experiment—at once. Nevertheless, in subsequent iterations, active selection always provides higher $Z_h^*$ than random selection. As the size of the inducing set increases, the benefit from active selection decreases because the value of any inducing site decreases.

### 9.4 Comparing with Single-machine Multi-class Support Vector Machines

Support vector machines (SVMs, Vapnik, 1998) are popular for classification and they have been known to give good classification accuracies in general. Although originally formulated for binary classification, several extensions have been proposed for multi-class classification. These extensions can be grouped roughly into two: one is to transform the multi-class problem into several binary class problems together with a decoding step; the other, called the single-machine approach, is to solve a single optimization problem for multiple classes, keeping to the broad principles of structural risk minimization (Vapnik, 1998). Comparisons between the two groups have been done by Rifkin and Klautau (2004). In this section, we compare our proposed variational approximation on multi-class Gaussian processes to four different single-machine multi-class SVMs using the MSVMpack package (Lauer and Guermeur, 2011). We denote the four single machines by WW (Vapnik, 1998; Weston and Watkins, 1999), CS (Crammer and Singer, 2001), LLW (Lee et al., 2004) and MSVM2 (Guermeur and Monfrini, 2011). The comparison is on the four UCI data sets.

For both Gaussian processes and SVMs, we use the sqexpiso covariance function or kernel (45). The signal variance $\sigma_{\text{x}}^2$ in the Gaussian process covariance function corresponds to the soft-margin trade-off parameter in the SVM objective functions, usually denoted by $C$. For the multi-class Gaussian processes, the parameters of the covariance function are estimated by optimizing our variational lower bound on the marginal likelihood. For the single-machine SVMs, the parameters $C$ and $\theta$ are estimated from a grid $(\log_{10} C, \theta) \in \{-2,-1,0,1,2,3\} \times \{0.1,1,5,10,15\}$ using five-fold cross validation on the training set.[8] This is repeated for each of the fifty train/test set partitions.

Table 4 gives the means and standard deviations of the errors over the fifty random train/test set partitions. The results for WW, CS and LLW are consistent with those reported by Weston and Watkins (1999), Hsu and Lin (2002) and Lee et al. (2004), when we take into perspective that their results are for 90%/10% train/test splits instead of the 60%/40% here. Comparing the errors for the Gaussian processes under column KL-MNL with the errors for the single-machine SVMs, we see that the Gaussian processes give better performances on the average. One reason is that model learning is achieved using continuous optimization with Gaussian processes, while only discrete

[8] When MSVMpack does not seem to converge on its stopping criterion for a parameter pair $(C,\theta)$, the learning is forced to terminate for that pair, and the validation score is computed based on the model at the point of termination.

| | | Single-machine multi-class support vector machines | | | |
|---|---|---|---|---|---|
| Data set | KL-MNL | WW | CS | LLW | MSVM2 |
| Iris | $2.18 \pm 1.42$ | $3.02 \pm 2.51$ | $3.22 \pm 4.34$ | $2.74 \pm 1.45$ | $2.88 \pm 1.47$ |
| Thyroid | $3.54 \pm 1.68$ | $4.42 \pm 3.45$ | $4.82 \pm 1.89$ | $5.28 \pm 2.08$ | $5.58 \pm 2.56$ |
| Wine | $1.40 \pm 0.90$ | $1.40 \pm 0.69$ | $2.20 \pm 1.34$ | $1.62 \pm 1.03$ | $1.50 \pm 0.95$ |
| Glass | $27.44 \pm 3.78$ | $29.30 \pm 10.70$ | $28.70 \pm 3.17$ | $29.54 \pm 9.72$ | $29.42 \pm 3.59$ |

Table 4: Errors of variational multinomial logit Gaussian process (column *KL-MNL*) and four variants of single-machine multi-class SVMs. The means and standard deviations of the errors over fifty partitions are reported. The sqexpiso covariance function/kernel (45) is used on the inputs, and there is no inter-class correlations between the functions.

| | Errors | | Number of support vectors | | | | |
|---|---|---|---|---|---|---|---|
| Data set | Sparse KL-MNL | SVM-WW | min | $Q_1$ | $Q_2$ | $Q_3$ | max |
| Iris | $2.14 \pm 1.39$ | $3.02 \pm 2.51$ | 12 | 17 | 20 | 24 | 36 |
| Thyroid | $7.38 \pm 5.22$ | $4.42 \pm 3.45$ | 2 | 3 | 4 | 7 | 39 |
| Wine | $1.32 \pm 0.91$ | $1.40 \pm 0.69$ | 4 | 10 | 25 | 34 | 37 |
| Glass | $28.42 \pm 4.05$ | $29.30 \pm 10.70$ | 8 | 18 | 56 | 65 | 75 |

Table 5: Errors of sparse variational multinomial logit Gaussian process and the WW variant of multi-class SVM. The means and standard deviations of the errors over fifty partitions are reported. The sqexpiso covariance function/kernel (45) is used on the inputs, and there is no inter-class correlations. The number of inducing variables for the sparse approximation is the number of support vectors given by WW times the number of classes. The last five columns give the statistics of number of support vectors over the partitions. Column *SVM-WW* duplicates column *WW* in Table 4.

optimization on the grid is used with the SVMs. Hence, the Gaussian processes can give finer parameter estimates.

The above uses the full (or non-sparse) approximation to the Gaussian process model. We also experiment with the sparse approximation. For each data set and each partition of the set, the target number inducing variables is fixed to the number of support vectors selected by WW multiplied by the number of classes. The initial inducing set is $C$ randomly chosen variables. Inducing variables are added using the strategy described in Section 9.3, but now with $\delta = C$, $|\mathcal{S}| = 5$ and $t = 20$. This expansion of the inducing set is alternated with one gradient-line-search to optimize the hyper-parameters of the model. After all the inducing variables are added, the hyper-parameters are further optimized. Table 5 reports the results repeated over the fifty partitions for each data set. The table also gives the minimum, maximum and quartiles $(Q_1, Q_2, Q_3)$ of the number of support vectors.

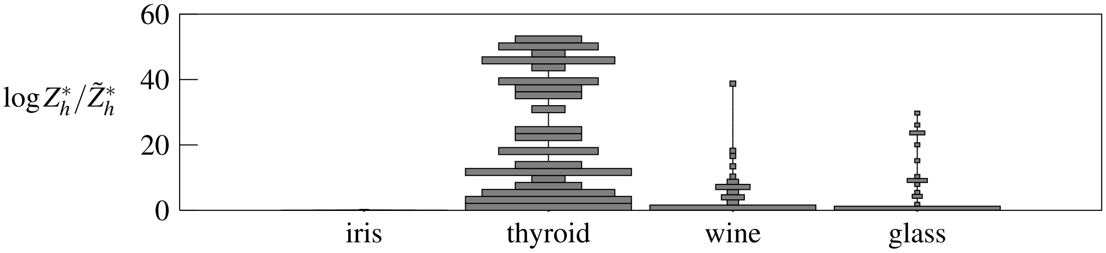


Figure 6: The quality of sparse approximations, at the same level of sparsity as the WW variant of SVM. Each histogram is for the log ratios of $Z_h^*$ for the full model to $\tilde{Z}_h^*$ for the sparse model. The approximation is mostly tight for the iris, wine and glass data sets, but is loose for the thyroid data. The histogram for the iris data concentrated at 0, so it is barely visible. The hyper-parameters for $Z_h^*$ and for $\tilde{Z}_h^*$ are different.

From Table 5, we see that the sparse Gaussian process model with active selection of inducing set compares favourably with the WW variant of SVM in terms of errors for three of the four data sets. The sparse Gaussian process model gives significantly more errors for the thyroid data. We believe there are two reasons for this. First, most of the fifty repetitions for the thyroid data have very small number inducing variables: the median is $4 \times 3 = 12$. Second, the sparsity in the Gaussian process is an *imposed* approximation to the full Gaussian process, while the sparsity in the SVM is a direct consequence of its objective function. Therefore, limited to a median of only 12 inducing variables, the sparse approximation is unsatisfactory. This is reflected in Figure 6, which gives the violin plot of the log ratios of the marginal likelihood for the full model to the sparse model. From the figure, we see that approximation of the Gaussian process model with the same level of sparsity as WW is unsatisfactory for the thyroid data. Finally, we remark that the full Gaussian process model gives $3.54 \pm 1.68$ errors; see Table 4.

## 10. Conclusion and Discussion

We have introduced a tractable variational approximation to the multinomial logit Gaussian processes for multi-class classification in Section 2.3, and we have provided the necessary updates to optimize this approximation in Section 3. Empirical results in Section 9.1 have indicated that our approximation is very faithful to the exact distribution, in contrast to the variational mean-field approximation (Girolami and Rogers, 2006). One key to the success of this approximation is Theorem 6, which gives a variational lower bound on the expected log-likelihood at each observation. In addition, bounds on the train data marginal likelihood and test data predictive likelihoods have been given in Sections 2.3.2 and 2.3.3, and these bounds have been shown to be supported by empirical results in Section 9.1.

In Section 4, the proposed variational approximation has been combined with the sparse variational approximation approach previously advocated for regression (Titsias, 2009a). This sparse approximation to the multinomial logit Gaussian processes has the property that incremental increases in the inducing set will lead to tighter bounds on the marginal likelihood. This property has been exploited in Section 6.1 to derive scores for potential inducing sites. An active selection strategy making greedy use of these scores has been compared favorably with random selection in Section 9.3.

The present paper is mostly independent of the covariance structure of the Gaussian process. Nevertheless, at various points, we have focused on the case where the covariance is separable into the covariance on the inputs and the covariance on the classes. Such separable covariance has been investigated previously in the context of multi-task Gaussian process regression (see, for example, Bonilla et al., 2008). In Section 5, we have looked into the cases where the optimized variational posterior satisfies the sum-to-zero property; this property is also present in many single-machine multi-class SVMs. In Section 9.2.2, we have compared several models of prior correlation between the latent functions. Although the experimental results are neither general nor conclusive against or for $K^{\mathrm{c}} = I$, further investigation into the thyroid data has suggested that useful knowledge can indeed be extracted if inter-latent-function correlations are permitted.

There are several possibilities building upon and extending this work. From the model perspective, it is worthwhile to have more interesting models in which latent functions can be related than, for example, the separable covariance of Equation 1. For this, covariance models developed for multi-task learning in the regression setting can be assessed for multinomial logit Gaussian process. Here, two questions specific to multi-class classification are of interest. First, should one consider models where a pair of latent functions are allowed to be positively correlated? On the one hand, the classes are *mutually exclusive*, so an increase in the probability of one class necessarily entails a decrease in the probability of another class when the probability of other classes are held constant; hence we can expect negative correlations between the latent functions. On the other hand, if there is a natural *hierarchical* structure to the classes, then the probability of two classes can rise in tandem against the probability of the other classes; hence we may also find positive correlations. The second question is: should the set of length-scales of the latent functions be the same? To argue for the same set of length-scales, one may say that a *single* property of the given object $\mathbf{x}$ is being predicted. The counter argument is that there are *different* values for this property, and the latent function for each value may demand its own set of length-scales.

From the variational approximation perspective, further constraints can be placed on the any of the variational parameters: $\mathbf{m}$, $V$, the $\mathbf{b}_i$s and the $S_i$s. Some constraints will lead to more efficient algorithms though with less faithful approximations, and trade-offs between the two conflicting goals will have to be examined. From a purely algorithmic perspective, more efficient updates than the ones presented in Section 3 can be explored.

Theorem 6 gives a variational lower bound on the expected log-likelihood at each observation. We have seen that it is rather tight on the average in Section 9.1. This bound can be applied to on-line multi-class classification under the assumed density filtering framework, following a prior work on sparse on-line binary classification (Csató and Opper, 2002).

## Acknowledgments

We thank Chris Williams, Hai Leong Chieu and the reviewers for their helpful comments and suggestions. We are grateful to DSO National Laboratories, Singapore, for the permission to publish this work. This work is funded by DR-Tech, Singapore.

## Appendix A. Mathematical Preliminaries

We provide general results required in the proofs for the main results of this paper.

### A.1 Gaussians

**Lemma 18** *Let $\mathbf{x}_1 \in \mathbb{R}^{n_1}$ and $\mathbf{x}_2 \in \mathbb{R}^{n_2}$ be random vectors with two jointly normal distributions $p(\mathbf{x}_1, \mathbf{x}_2) \overset{\text{def}}{=} \mathcal{N}(\mathbf{n}, U)$ and $q(\mathbf{x}_1, \mathbf{x}_2) \overset{\text{def}}{=} \mathcal{N}(\mathbf{m}, V)$, where the parameters are partitioned as*

$$\mathbf{n} \overset{\text{def}}{=} \begin{pmatrix} \mathbf{n}_1 \\ \mathbf{n}_2 \end{pmatrix}, \qquad U \overset{\text{def}}{=} \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix}, \qquad \mathbf{m} \overset{\text{def}}{=} \begin{pmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{pmatrix}, \qquad V \overset{\text{def}}{=} \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}.$$

*The difference between the Kullback-Leibler divergences on $\mathbf{x} \overset{\text{def}}{=} (\mathbf{x}_1^{\mathrm{T}}, \mathbf{x}_2^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{x}_1$ is*

$$\begin{aligned} &\mathrm{KL}(p(\mathbf{x})\|q(\mathbf{x})) - \mathrm{KL}(p(\mathbf{x}_1)\|q(\mathbf{x}_1)) \\ &\qquad = -\frac{n_2}{2} - \frac{1}{2}\log\left|V_{2|1}^{-1}U_{2|1}\right| + \frac{1}{2}\operatorname{tr} V_{2|1}^{-1}\left(U_{2|1} + WU_{11}^{-1}W^{\mathrm{T}}\right) + \frac{1}{2}\mathbf{t}^{\mathrm{T}}V_{2|1}^{-1}\mathbf{t}, \end{aligned}$$

*where*

$$\begin{aligned} U_{2|1} &\overset{\text{def}}{=} U_{22} - U_{21}U_{11}^{-1}U_{12}, & V_{2|1} &\overset{\text{def}}{=} V_{22} - V_{21}V_{11}^{-1}V_{12}, \\ W &\overset{\text{def}}{=} U_{21} - V_{21}V_{11}^{-1}U_{11}, & \mathbf{t} &\overset{\text{def}}{=} (\mathbf{m}_2 - \mathbf{n}_2) - V_{21}V_{11}^{-1}(\mathbf{m}_1 - \mathbf{n}_1). \end{aligned}$$

**Proof** The conditional distributions $p(\mathbf{x}_2|\mathbf{x}_1)$ and $q(\mathbf{x}_2|\mathbf{x}_1)$ have means

$$\mathbf{n}_{2|1} \overset{\text{def}}{=} \mathbf{n}_2 + U_{21}U_{11}^{-1}(\mathbf{x}_1 - \mathbf{n}_1), \qquad \mathbf{m}_{2|1} \overset{\text{def}}{=} \mathbf{m}_2 + V_{21}V_{11}^{-1}(\mathbf{x}_1 - \mathbf{m}_1)$$

and covariances $U_{2|1}$ and $V_{2|1}$. The difference between the Kullback-Leibler divergences can be derived through the Kullback-Leibler divergence of these conditionals:

$$\begin{aligned} \mathrm{KL}(p(\mathbf{x})\|q(\mathbf{x})) - \mathrm{KL}(p(\mathbf{x}_1)\|q(\mathbf{x}_1)) &= \int p(\mathbf{x})\log\frac{p(\mathbf{x})/p(\mathbf{x}_1)}{q(\mathbf{x})/q(\mathbf{x}_1)}\mathrm{d}\mathbf{x} = \int p(\mathbf{x})\log\frac{p(\mathbf{x}_2|\mathbf{x}_1)}{q(\mathbf{x}_2|\mathbf{x}_1)}\mathrm{d}\mathbf{x} \\ = \frac{1}{2}\int p(\mathbf{x}_1)\left[-n_2 - \log\left|V_{2|1}^{-1}U_{2|1}\right| + \operatorname{tr} V_{2|1}^{-1}U_{2|1} + \left(\mathbf{m}_{2|1} - \mathbf{n}_{2|1}\right)^{\mathrm{T}} V_{2|1}^{-1}\left(\mathbf{m}_{2|1} - \mathbf{n}_{2|1}\right)\right]\mathrm{d}\mathbf{x}_1 \end{aligned}$$

In the last expression above, only the final quadratic term in the integrand depends on $\mathbf{x}_1$, so we can move the other terms out of the integral. For this final term, its integral under $p(\mathbf{x}_1)\mathrm{d}\mathbf{x}_1$ is $\operatorname{tr} V_{2|1}^{-1}\left(WU_{11}^{-1}W^{\mathrm{T}}\right) + \mathbf{t}^{\mathrm{T}}V_{2|1}^{-1}\mathbf{t}$. ■

### A.2 Matrix

**Proposition 19** *For positive semi-definite matrices $A$ and $B$ of the same order, we have $B \preceq A$ implies* $\mathrm{null}(A) \subseteq \mathrm{null}(B)$.

**Proof** Let $\mathbf{x} \in \mathrm{null}(A)$, then $A\mathbf{x} = \mathbf{0}$ by definition. Hence $\mathbf{x}^{\mathrm{T}}A\mathbf{x} = 0$. Since $0 \preceq B \preceq A$, we have $0 \leq \mathbf{x}^{\mathrm{T}}B\mathbf{x} \leq \mathbf{x}^{\mathrm{T}}A\mathbf{x} = 0$. Therefore $\mathbf{x}^{\mathrm{T}}B\mathbf{x} = 0$. This means $B\mathbf{x} = 0$, or $\mathbf{x} \in \mathrm{null}(B)$ (Horn and Johnson, 1985, Section 7.5, Problem 14). Thus $\mathbf{x} \in \mathrm{null}(A) \Longrightarrow \mathbf{x} \in \mathrm{null}(B)$. ■

**Lemma 20** *(Matrix determinant lemma) For an n-by-n non-singular matrix A and vectors* $\mathbf{u}$ *and* $\mathbf{v}$ *in* $\mathbb{R}^n$, $|A+\mathbf{u}\mathbf{v}^{\mathrm{T}}| = (1+\mathbf{v}^{\mathrm{T}}A^{-1}\mathbf{u})|A|$.

**Lemma 21** *Under the setting in Lemma 20*

$$|A+\mathbf{u}\mathbf{v}^{\mathrm{T}}+\mathbf{v}\mathbf{u}^{\mathrm{T}}| = \left[(1+\mathbf{v}^{\mathrm{T}}A^{-1}\mathbf{u})(1+\mathbf{u}^{\mathrm{T}}A^{-1}\mathbf{v}) - \mathbf{u}^{\mathrm{T}}A^{-1}\mathbf{u}\,\mathbf{v}^{\mathrm{T}}A^{-1}\mathbf{v})\right]|A|.$$

**Proof** Apply Lemma 20 twice; then use the Sherman-Morrison formula on $(A+\mathbf{u}\mathbf{v}^{\mathrm{T}})^{-1}$. ■

**Corollary 22** *If A is also symmetric, then*

$$|A+\mathbf{u}\mathbf{v}^{\mathrm{T}}+\mathbf{v}\mathbf{u}^{\mathrm{T}}| = \left[(1+\mathbf{u}^{\mathrm{T}}A^{-1}\mathbf{v})^2 - \mathbf{u}^{\mathrm{T}}A^{-1}\mathbf{u}\,\mathbf{v}^{\mathrm{T}}A^{-1}\mathbf{v})\right]|A|.$$

**Lemma 23** *Further, with* $a \in \mathbb{R}$, *we have*

$$|A+a\mathbf{u}\mathbf{u}^{\mathrm{T}}+\mathbf{u}\mathbf{v}^{\mathrm{T}}+\mathbf{v}\mathbf{u}^{\mathrm{T}}| = \left(\left(1+\mathbf{u}^{\mathrm{T}}A^{-1}\mathbf{v}\right)^2 - \mathbf{u}^{\mathrm{T}}A^{-1}\mathbf{u}\,\mathbf{v}^{\mathrm{T}}A^{-1}\mathbf{v} + a\mathbf{u}^{\mathrm{T}}A^{-1}\mathbf{u}\right)|A|.$$

**Proof** Let $X \stackrel{\text{def}}{=} A + a\mathbf{u}\mathbf{u}^{\mathrm{T}}$ and $b \stackrel{\text{def}}{=} 1 + a\mathbf{u}^{\mathrm{T}}A^{-1}\mathbf{u}$. We apply Corollary 22, then we use Lemma 20 on $|X|$ and the Sherman-Morrison formula on $X^{-1}$:

$$\begin{aligned}
&\left[(1+\mathbf{u}^{\mathrm{T}}X^{-1}\mathbf{v})^2 - \mathbf{u}^{\mathrm{T}}X^{-1}\mathbf{u}\,\mathbf{v}^{\mathrm{T}}X^{-1}\mathbf{v})\right]|X| \\
&= |A|b\left[\left(1+\frac{1}{b}\mathbf{u}^{\mathrm{T}}A^{-1}\mathbf{v}\right)^2 - \frac{1}{b}\mathbf{u}^{\mathrm{T}}A^{-1}\mathbf{u}\left(\mathbf{v}^{\mathrm{T}}A^{-1}\mathbf{v} - \frac{a}{b}\left(\mathbf{u}^{\mathrm{T}}A^{-1}\mathbf{v}\right)^2\right)\right] \\
&= |A|b\left[1+\frac{2}{b}\mathbf{u}^{\mathrm{T}}A^{-1}\mathbf{v} + \frac{1}{b}\left(\mathbf{u}^{\mathrm{T}}A^{-1}\mathbf{v}\right)^2 - \frac{1}{b}\mathbf{u}^{\mathrm{T}}A^{-1}\mathbf{u}\mathbf{v}^{\mathrm{T}}A^{-1}\mathbf{v}\right] \\
&= |A|\left[(1+\mathbf{u}^{\mathrm{T}}A^{-1}\mathbf{v})^2 + a\mathbf{u}^{\mathrm{T}}A^{-1}\mathbf{u} - \mathbf{u}^{\mathrm{T}}A^{-1}\mathbf{u}\mathbf{v}^{\mathrm{T}}A^{-1}\mathbf{v}\right].
\end{aligned}$$

The second step in the derivation applies the identity $1-(a/b)\mathbf{u}^{\mathrm{T}}A^{-1}\mathbf{u} = 1/b$. ■

**Theorem 24** *(Matrix quadratic equation, a special case of Potter 1966). Let A and B be two real symmetric n-by-n matrices such that* $A+B^2$ *is positive semi-definite and B is positive definite. Then the positive definite solution to the equation* $-X^2+BX+XB+A=0$ *is* $X = P\Lambda^{\frac{1}{2}}P^{\mathrm{T}} + B$, *where* $P\Lambda P^{\mathrm{T}}$ *is the eigen-decomposition of* $A+B^2$.

**Proof** The solution $X$ can be proved by direct substitution into the equation, or by completing the square, or by following a construction due to Potter (1966). For positive definiteness, since $B \succ 0$, we only require that $(A+B^2)$ is positive semi-definite. ■

## Appendix B. Bounds

This appendix provides the proofs for the bounds stated in the main text.

### B.1 Variational Lower Bound on the Marginal Likelihood

We derive the variational lower bounds $\tilde{Z}_B$ on the true marginal likelihood in the sparse case. Using Jensen's inequality, we have

$$\log p(\mathbf{y}) = \log \int p(\mathbf{y}, \mathbf{f}, \mathbf{z})\, \mathrm{d}\mathbf{z}\, \mathrm{d}\mathbf{f} = \log \int p(\mathbf{y}, \mathbf{f}, \mathbf{z}) \frac{q(\mathbf{f}, \mathbf{z}|\mathbf{y})}{q(\mathbf{f}, \mathbf{z}|\mathbf{y})}\, \mathrm{d}\mathbf{z}\, \mathrm{d}\mathbf{f} \geq \log \tilde{Z}_B,$$

$$\text{where} \qquad \log \tilde{Z}_B \stackrel{\text{def}}{=} \int q(\mathbf{f}, \mathbf{z}|\mathbf{y}) \log \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{z})}{q(\mathbf{f}, \mathbf{z}|\mathbf{y})}\, \mathrm{d}\mathbf{z}\, \mathrm{d}\mathbf{f}. \tag{47}$$

Given the model, the joint distribution $p(\mathbf{y}, \mathbf{f}, \mathbf{z})$ factorizes into $p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{z})p(\mathbf{z})$ exactly; there is no approximation involved. For the approximate posterior $q(\mathbf{f}, \mathbf{z}|\mathbf{y})$, however, the factorization $q(\mathbf{f}, \mathbf{z}|\mathbf{y}) = p(\mathbf{f}|\mathbf{z})q(\mathbf{z}|\mathbf{y})$ is assumed. Using these two factorizations, we have

$$\begin{aligned}
\log \tilde{Z}_B &= \int p(\mathbf{f}|\mathbf{z})q(\mathbf{z}|\mathbf{y}) \log \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{y})} \mathrm{d}\mathbf{z}\mathrm{d}\mathbf{f} \\
&= \int q(\mathbf{z}|\mathbf{y}) \left[ \log \frac{p(\mathbf{z})}{q(\mathbf{z}|\mathbf{y})} + \int p(\mathbf{f}|\mathbf{z}) \log p(\mathbf{y}|\mathbf{f}) \mathrm{d}\mathbf{f} \right] \mathrm{d}\mathbf{z} \\
&= -\mathrm{KL}(q(\mathbf{z}|\mathbf{y}) \| p(\mathbf{z})) + \int q(\mathbf{f}|\mathbf{y}) \log p(\mathbf{y}|\mathbf{f}) \mathrm{d}\mathbf{f},
\end{aligned}$$

where $q(\mathbf{f}|\mathbf{y}) \stackrel{\text{def}}{=} \int q(\mathbf{f}, \mathbf{z}|\mathbf{y}) \mathrm{d}\mathbf{z}$. Since the joint likelihood factorizes across the $n$ data points, this is also $\log \tilde{Z}_B = -\mathrm{KL}\left(q(\mathbf{z}|\mathbf{y}) \,\|\, p(\mathbf{z})\right) + \sum_{i=1}^n \ell_i(\mathbf{y}_i; q)$, where $\ell_i(\mathbf{y}_i; q) \stackrel{\text{def}}{=} \int q(\mathbf{f}_i|\mathbf{y}) \log p(\mathbf{y}_i|\mathbf{f}_i)\, \mathrm{d}\mathbf{f}_i$.

The bound $Z_B$ in the non-sparse case can be obtained in a similar manner with

$$\log Z_B \stackrel{\text{def}}{=} \int q(\mathbf{f}|\mathbf{y}) \log \frac{p(\mathbf{y}, \mathbf{f})}{q(\mathbf{f}|\mathbf{y})}\, \mathrm{d}\mathbf{f}. \tag{48}$$

#### B.1.1 Relation between Bounds for Non-sparse and Sparse Approximations

We show that the optimized non-sparse bound $\log Z_B^*$ is not smaller than the optimized sparse bound $\log \tilde{Z}_B^*$. We begin by constraining the approximate posterior in $\log Z_B$:

$$\log Z_B^* \stackrel{\text{def}}{=} \max_{q(\mathbf{f}|\mathbf{y})} \log Z_B \geq \max_{q(\mathbf{z}|\mathbf{y})} \log Z_B \quad \left(\text{where } q(\mathbf{f}|\mathbf{y}) = \textstyle\int p(\mathbf{f}|\mathbf{z})q(\mathbf{z}|\mathbf{y})\mathrm{d}\mathbf{z}\right). \tag{49}$$

We introduce an arbitrary distribution $r$ on $\mathbf{z}$ and use Jensen's inequality to get

$$\log p(\mathbf{y}, \mathbf{f}) = \log \int p(\mathbf{y}, \mathbf{f}, \mathbf{z}) \mathrm{d}\mathbf{z} = \log \int r(\mathbf{z}) \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{z})}{r(\mathbf{z})} \mathrm{d}\mathbf{z} \geq \int r(\mathbf{z}) \log \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{z})}{r(\mathbf{z})} \mathrm{d}\mathbf{z}.$$

The above inequality is substituted into $\log Z_B$ through its definition (48), and the result is applied to the leftmost expression in (49):

$$\log Z_B^* \geq \max_{q(\mathbf{z}|\mathbf{y})} \int q(\mathbf{f}|\mathbf{y}) r(\mathbf{z}) \log \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{z})}{q(\mathbf{f}|\mathbf{y}) r(\mathbf{z})}\, \mathrm{d}\mathbf{z}\, \mathrm{d}\mathbf{f} \quad \left(\text{where } q(\mathbf{f}|\mathbf{y}) = \textstyle\int p(\mathbf{f}|\mathbf{z})q(\mathbf{z}|\mathbf{y})\mathrm{d}\mathbf{z}\right).$$

This is for any $r(\mathbf{z})$. We choose $r(\mathbf{z}) \stackrel{\text{def}}{=} q(\mathbf{z}|\mathbf{f}, \mathbf{y}) = q(\mathbf{f}, \mathbf{z}|\mathbf{y})/q(\mathbf{f}|\mathbf{y})$ and cancel out $q(\mathbf{f}|\mathbf{y})$ to obtain

$$\log Z_B^* \geq \max_{q(\mathbf{z}|\mathbf{y})} \int q(\mathbf{f}, \mathbf{z}|\mathbf{y}) \log \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{z})}{q(\mathbf{f}, \mathbf{z}|\mathbf{y})}\, \mathrm{d}\mathbf{z}\, \mathrm{d}\mathbf{f}.$$

The objective on the right is $\log \tilde{Z}_B$ by definition (47).

### B.2 Derivation of $r(\mathbf{f})$ and $r(\mathbf{y})$ for Lemma 2

Let $r(\mathbf{f})$ be a prior distribution such that the posterior

$$r(\mathbf{f}|\mathbf{y}) = p(\mathbf{y}|\mathbf{f})\, r(\mathbf{f})/r(\mathbf{y}), \qquad \text{where } r(\mathbf{y}) \overset{\text{def}}{=} \int p(\mathbf{y}|\mathbf{f})\, r(\mathbf{f}) \mathrm{d}\mathbf{f},$$

is a $C$-variate Gaussian density on $\mathbf{f}$ with mean $\mathbf{a}$ and precision $W$. Rearranging gives

$$\frac{r(\mathbf{f})}{r(\mathbf{y})} = \frac{r(\mathbf{f}|\mathbf{y})}{p(\mathbf{y}|\mathbf{f})} = \sum_{c=1}^{C} \frac{|W|^{1/2}}{(2\pi)^{C/2}} \exp -\frac{1}{2}\Big[(\mathbf{f}-\mathbf{a})^{\mathrm{T}} W(\mathbf{f}-\mathbf{a}) - 2(\mathbf{e}^c - \mathbf{y})^{\mathrm{T}}\mathbf{f})\Big]. \tag{50}$$

Let $\mathbf{a}^c$ be such that

$$W\mathbf{a}^c = W\mathbf{a} + \mathbf{e}^c - \mathbf{y}, \tag{51}$$

and define

$$r^c(\mathbf{f}) \overset{\text{def}}{=} \frac{|W|^{1/2}}{(2\pi)^{C/2}} \exp\left[-\frac{1}{2}(\mathbf{f}-\mathbf{a}^c)^{\mathrm{T}} W(\mathbf{f}-\mathbf{a}^c)\right].$$

By completing the square the terms within the brackets of (50), we obtain

$$r(\mathbf{f}) = r(\mathbf{y}) \exp\left[-\frac{1}{2}\mathbf{a}^{\mathrm{T}} W\mathbf{a}\right] \sum_{c=1}^{C} \exp\left[\frac{1}{2}(\mathbf{a}^c)^{\mathrm{T}} W\mathbf{a}^c\right] r^c(\mathbf{f}).$$

This is a mixture of Gaussians model, so let $r(\mathbf{f}) = \sum_c \gamma^c r^c(\mathbf{f})$. Normalization gives

$$r(\mathbf{y}) = \frac{\exp\left[\frac{1}{2}\mathbf{a}^{\mathrm{T}} W\mathbf{a}\right]}{\sum_{c=1}^{C} \exp\left[\frac{1}{2}(\mathbf{a}^c)^{\mathrm{T}} W\mathbf{a}^c\right]}, \qquad \gamma^c \overset{\text{def}}{=} \frac{\exp\left[\frac{1}{2}(\mathbf{a}^c)^{\mathrm{T}} W\mathbf{a}^c\right]}{\sum_{c'} \exp\left[\frac{1}{2}(\mathbf{a}^{c'})^{\mathrm{T}} W\mathbf{a}^{c'}\right]}. \tag{52}$$

### B.3 Derivation of Lower Bound $h$ on the Expected Log-likelihood for Lemma 5

Recall from (12) that

$$h(\mathbf{y};q,r) \overset{\text{def}}{=} \int q(\mathbf{f}|\mathbf{y}) \log r(\mathbf{f}|\mathbf{y}) \mathrm{d}\mathbf{f} + \log r(\mathbf{y}) - \log \sum_{c=1}^{C} \gamma^c \int q(\mathbf{f}|\mathbf{y}) r^c(\mathbf{f}) \mathrm{d}\mathbf{f}. \tag{53}$$

We simplify the first two terms on the right:

$$\int q(\mathbf{f}|\mathbf{y}) \log r(\mathbf{f}|\mathbf{y}) \mathrm{d}\mathbf{f} = \frac{1}{2}\Big(-C\log 2\pi + \log|W| - (\mathbf{m}-\mathbf{a})^{\mathrm{T}} W(\mathbf{m}-\mathbf{a})^{\mathrm{T}} - \operatorname{tr} WV\Big), \tag{54}$$

$$\log r(\mathbf{y}) = \frac{1}{2}\mathbf{a}^{\mathrm{T}} W\mathbf{a} - \log \sum_{c=1}^{C} \exp\left[\frac{1}{2}(\mathbf{a}^c)^{\mathrm{T}} W\mathbf{a}^c\right]. \tag{55}$$

For the third term on the right of (53), we introduce $S \overset{\text{def}}{=} V^{-1} + W$ and use

$$\begin{aligned}&\int q(\mathbf{f}|\mathbf{y}) r^c(\mathbf{f}) \mathrm{d}\mathbf{f} \\ &= \sqrt{\frac{|W|}{(2\pi)^C |V|\,|S|}} \exp -\frac{1}{2}\Big[\mathbf{m}^{\mathrm{T}} V^{-1}\mathbf{m} + (\mathbf{a}^c)^{\mathrm{T}} W\mathbf{a}^c - \big(V^{-1}\mathbf{m} + W\mathbf{a}^c\big)^{\mathrm{T}} S^{-1} \big(V^{-1}\mathbf{m} + W\mathbf{a}^c\big)\Big]\end{aligned} \tag{56}$$

to obtain

$$\log \sum_{c=1}^{C} \gamma^c \int q(\mathbf{f}|\mathbf{y}) r^c(\mathbf{f}) d\mathbf{f} = -\log \sum_{c=1}^{C} \exp\left[\frac{1}{2}(\mathbf{a}^c)^{\mathrm{T}} W \mathbf{a}^c\right] + \frac{1}{2}\log|W| - \frac{C}{2}\log 2\pi - \frac{1}{2}\log|SV|$$
$$+\log \sum_{c=1}^{C} \exp -\frac{1}{2}\left\{\mathbf{m}^{\mathrm{T}} V^{-1}\mathbf{m} - \left(V^{-1}\mathbf{m} + W\mathbf{a}^c\right)^{\mathrm{T}} S^{-1}\left(V^{-1}\mathbf{m} + W\mathbf{a}^c\right)\right\}, \quad (57)$$

where the first term is from the denominator of $\gamma^c$ (Equation 52) and the second to fourth terms are from the first factor in Equation 56. At present, let us focus on the term within the braces in the above equation. Let $\mathbf{b} \overset{\text{def}}{=} W(\mathbf{m}-\mathbf{a}) + \mathbf{y}$ and use the identity $W\mathbf{a}^c = W\mathbf{a} + \mathbf{e}^c - \mathbf{y}$ (Equation 51) and definition $S \overset{\text{def}}{=} V^{-1} + W$. Then

$$\begin{aligned}
&\mathbf{m}^{\mathrm{T}} V^{-1}\mathbf{m} - \left(V^{-1}\mathbf{m} + W\mathbf{a}^c\right)^{\mathrm{T}} S^{-1}\left(V^{-1}\mathbf{m} + W\mathbf{a}^c\right) \\
&= \mathbf{m}^{\mathrm{T}} V^{-1}\mathbf{m} - \left(V^{-1}\mathbf{m} + W\mathbf{a} + \mathbf{e}^c - \mathbf{y}\right)^{\mathrm{T}} S^{-1}\left(V^{-1}\mathbf{m} + W\mathbf{a} + \mathbf{e}^c - \mathbf{y}\right) \\
&= \mathbf{m}^{\mathrm{T}} V^{-1}\mathbf{m} - \left(V^{-1}\mathbf{m} + W\mathbf{m} + \mathbf{e}^c - \mathbf{b}\right)^{\mathrm{T}} S^{-1}\left(V^{-1}\mathbf{m} + W\mathbf{m} + \mathbf{e}^c - \mathbf{b}\right) \\
&= \mathbf{m}^{\mathrm{T}} V^{-1}\mathbf{m} - \left(S\mathbf{m} + \mathbf{e}^c - \mathbf{b}\right)^{\mathrm{T}} S^{-1}\left(S\mathbf{m} + \mathbf{e}^c - \mathbf{b}\right) \\
&= \mathbf{m}^{\mathrm{T}} V^{-1}\mathbf{m} - \mathbf{m}^{\mathrm{T}} S\mathbf{m} - 2\mathbf{m}^{\mathrm{T}}(\mathbf{e}^c - \mathbf{b}) - (\mathbf{e}^c - \mathbf{b})^{\mathrm{T}} S^{-1}(\mathbf{e}^c - \mathbf{b}) \\
&= -\mathbf{m}^{\mathrm{T}} W\mathbf{m} + 2\mathbf{m}^{\mathrm{T}} W(\mathbf{m}-\mathbf{a}) + 2\mathbf{m}^{\mathrm{T}}\mathbf{y} - 2\mathbf{m}^{\mathrm{T}}\mathbf{e}^c - (\mathbf{b} - \mathbf{e}^c)^{\mathrm{T}} S^{-1}(\mathbf{b} - \mathbf{e}^c) \\
&= \mathbf{m}^{\mathrm{T}} W\mathbf{m} - 2\mathbf{m}^{\mathrm{T}} W\mathbf{a} + 2\mathbf{m}^{\mathrm{T}}\mathbf{y} - 2\mathbf{m}^{\mathrm{T}}\mathbf{e}^c - (\mathbf{b} - \mathbf{e}^c)^{\mathrm{T}} S^{-1}(\mathbf{b} - \mathbf{e}^c) \\
&= (\mathbf{m}-\mathbf{a})^{\mathrm{T}} W(\mathbf{m}-\mathbf{a}) - \mathbf{a}^{\mathrm{T}} W\mathbf{a} + 2\mathbf{m}^{\mathrm{T}}\mathbf{y} - 2\mathbf{m}^{\mathrm{T}}\mathbf{e}^c - (\mathbf{b} - \mathbf{e}^c)^{\mathrm{T}} S^{-1}(\mathbf{b} - \mathbf{e}^c) \\
&= (\mathbf{m}-\mathbf{a})^{\mathrm{T}} W(\mathbf{m}-\mathbf{a}) - \mathbf{a}^{\mathrm{T}} W\mathbf{a} + 2\mathbf{m}^{\mathrm{T}}\mathbf{y} - 2\log g^c(\mathbf{y}; q, r),
\end{aligned}$$

where $g^c(\mathbf{y}; q, r) \overset{\text{def}}{=} \exp\left[\mathbf{m}^{\mathrm{T}}\mathbf{e}^c + \frac{1}{2}(\mathbf{b} - \mathbf{e}^c)^{\mathrm{T}} S^{-1}(\mathbf{b} - \mathbf{e}^c)\right]$. By pulling out the terms independent of the dummy variable $c$ in the last term of (57), we can rewrite

$$\log \sum_{c=1}^{C} \gamma^c \int q(\mathbf{f}|\mathbf{y}) r^c(\mathbf{f}) d\mathbf{f} = -\log \sum_{c=1}^{C} \exp\left[\frac{1}{2}(\mathbf{a}^c)^{\mathrm{T}} W \mathbf{a}^c\right] + \frac{1}{2}\log|W| - \frac{C}{2}\log 2\pi$$
$$-\frac{1}{2}\log|SV| - \frac{1}{2}(\mathbf{m}-\mathbf{a})^{\mathrm{T}} W(\mathbf{m}-\mathbf{a}) + \frac{1}{2}\mathbf{a}^{\mathrm{T}} W\mathbf{a} - \mathbf{m}^{\mathrm{T}}\mathbf{y} + \log \sum_{c=1}^{C} g^c(\mathbf{y}; q, r). \quad (58)$$

Finally, putting (54), (55) and (58) into (53) and cancelling terms yields

$$h(\mathbf{y}; q, r) = \frac{C}{2} + \frac{1}{2}\log|SV| - \frac{1}{2}\operatorname{tr} SV + \mathbf{m}^{\mathrm{T}}\mathbf{y} - \log \sum_{i=1}^{C} g^c(\mathbf{y}; q, r).$$

Since distribution $r(\mathbf{f}|\mathbf{y})$ is completely determined by its mean $\mathbf{a}$ and precision $W$, we may use these parameters instead of $r$ in our notation; that is, $h(\mathbf{y}; q, \mathbf{a}, W)$ instead of $h(\mathbf{y}; q, r)$.

### B.4 Lemmas to Prove Theorem 6

This section collects the necessary lemmas to prove Theorem 6. Function $g^c(q, \mathbf{b}, S)$ and function $h(\mathbf{y}; q, \mathbf{b}, S)$ are given by Equations 16 and 19 in the main text, while variables $\bar{\mathbf{g}}$ and $A$ are defined by Equations 17 and 18.

**Lemma 25** *Function $h$ is jointly concave in $\mathbf{b}$ and $S$.*

**Proof** The following facts are used: (a) the log-determinant term is concave in $S$ (Horn and Johnson, 1985, Theorem 7.6.7); (b) the matrix trace term is both concave and convex in $S$; (c) the quadratic term in the exponent of $g^c$ is jointly convex in $S$ and $\mathbf{b}$ (Ando, 1979); and (d) the sum of log-convex functions is log-convex. ■

**Lemma 26** *The maximum of $h$ given $S$ with respect to $\mathbf{b}$ is at $\mathbf{b} = \mathbf{b}^*$ that satisfies $\mathbf{b}^* = \bar{\mathbf{g}}^*$, where $\bar{\mathbf{g}}^*$ is obtained by evaluating $\bar{\mathbf{g}}$ at $\mathbf{b}^*$.*

**Proof** Proved by setting the gradient $\partial h/\partial \mathbf{b}$ to zero. ■

**Lemma 27** *The maximum of $h$ given $\mathbf{b}$ with respect to $S$ is at $S = S^*$ that satisfies the implicit equation $-S^*VS^* + S^* + A^* = 0$, where $A^* \neq 0$ is $A$ evaluated at $S^*$.*

**Proof** Proved by equating the gradient

$$\frac{\partial h}{\partial S} = -\frac{1}{2}V + \frac{1}{2}S^{-1} + \frac{1}{2}S^{-1}AS^{-1} \tag{59}$$

to zero and pre- and post-multiplying both sides by $S$ (valid since $S \succ 0$ by definition). ■

**Lemma 28** *Let $A \neq 0$, $A \succeq 0$ and $V \succ 0$. Let $S^{\text{fx}}$ be the fixed point given implicitly by*

$$-S^{\text{fx}}VS^{\text{fx}} + S^{\text{fx}} + A = 0. \tag{60}$$

*Then $V^{-1} \preceq S^{\text{fx}} \preceq V^{-1} + A$, and $S \neq V^{-1}$ and $S^{\text{fx}} \neq V^{-1} + A$; that is, there exists a matrix $W$ satisfying $0 \preceq W^{\text{fx}} \preceq A$ and $W \notin \{0, A\}$ such that $S^{\text{fx}} = V^{-1} + W^{\text{fx}}$. Furthermore,* $\text{null}(W^{\text{fx}}) = \text{null}(A)$.

**Proof** Let $V$ factorizes to $LL^{\text{T}}$, where $L$ is non-singular; for example, matrix $L$ can be the lower Cholesky factor of $V$. We pre- and post-multiply Equation 60 by $L^{\text{T}}$ and $L$ to obtain the equation $-(L^{\text{T}}S^{\text{fx}}L)(L^{\text{T}}S^{\text{fx}}L) + (L^{\text{T}}S^{\text{fx}}L) + L^{\text{T}}AL = 0$. This is a matrix quadratic equation in $L^{\text{T}}S^{\text{fx}}L$, so we use Theorem 24 to reach the solution

$$L^{\text{T}}S^{\text{fx}}L = P\tilde{\Lambda}P^{\text{T}}, \qquad \tilde{\Lambda} \overset{\text{def}}{=} (\Lambda + I/4)^{1/2} + I/2, \tag{61}$$

where $P\Lambda P^{\text{T}}$ is the eigen-decomposition of $L^{\text{T}}AL$. Matrix $A$ is positive semi-definite, so similarly is $L^{\text{T}}AL$ (Horn and Johnson, 1985, Observation 7.7.2) and $L^{\text{T}}AL + I/4$. Therefore, $L^{\text{T}}S^{\text{fx}}L$ is positive definite; see Theorem 24. Since $L$ is non-singular, we can write $S^{\text{fx}} = L^{-\text{T}}\left(L^{\text{T}}S^{\text{fx}}L\right)L^{-1}$, so $S^{\text{fx}}$ is positive definite (Horn and Johnson, 1985, Observation 7.7.2). Define $W^{\text{fx}} \overset{\text{def}}{=} S^{\text{fx}} - V^{-1}$, then

$$W^{\text{fx}} = L^{-\text{T}}\left(L^{\text{T}}S^{\text{fx}}L\right)L^{-1} - (LL^{\text{T}})^{-1} = L^{-\text{T}}\left(L^{\text{T}}S^{\text{fx}}L - I\right)L^{-1} = L^{-\text{T}}P\left(\tilde{\Lambda} - I\right)P^{\text{T}}L^{-1},$$

where (61) is used. Since the least diagonal value in $\tilde{\Lambda}$ is one, $W^{\text{fx}}$ is positive semi-definite, so $S^{\text{fx}} \succeq V^{-1}$. Moreover, since $V \neq 0$ and $A \neq 0$, so $L^{\text{T}}AL \neq 0$, $\Lambda \neq 0$, $\tilde{\Lambda} \neq I$ and $W^{\text{fx}} \neq 0$. Hence $S^{\text{fx}} \neq V^{-1}$. Substitute $S^{\text{fx}} = V^{-1} + W^{\text{fx}}$ into (60) and rearranging gives

$$W^{\text{fx}} = A - W^{\text{fx}} V W^{\text{fx}} \preceq A. \tag{62}$$

Thus $S^{\text{fx}} \preceq V^{-1} + A$. Moreover, $W^{\text{fx}} \neq 0$ and $V \neq 0$ shows that $S^{\text{fx}} \neq V^{-1} + A$.

We now prove $\text{null}(W^{\text{fx}}) = \text{null}(A)$. Already, $W^{\text{fx}} \preceq A$ gives $\text{null}(A) \subseteq \text{null}(W)$ with Proposition 19, so it remains to proof $\text{null}(W^{\text{fx}}) \subseteq \text{null}(A)$. Let $\mathbf{x} \in \text{null}(W^{\text{fx}})$. Post-multiply both sides of the equality in (62) by $\mathbf{x}$ and use $W^{\text{fx}}\mathbf{x} = \mathbf{0}$ to give $A\mathbf{x} = \mathbf{0}$. Thus $\mathbf{x} \in \text{null}(A)$. ■

### B.5 Proof of Lemma 7

We introduce $u(\eta)$, where $u(0) = \ell(\mathbf{y}; q)$ and $u(1) = \log p(\mathbf{y}|\mathbf{m})$, and obtain its first two derivatives:

$$\begin{aligned}
u(\eta) &\overset{\text{def}}{=} \int q(\mathbf{f}|\mathbf{y}) \left( [(1-\eta)\mathbf{f} + \eta\mathbf{m}]^{\text{T}} \mathbf{y} - \log \sum_{c=1}^{C} \exp\left[(1-\eta)\mathbf{f} + \eta\mathbf{m}\right]^{\text{T}} \mathbf{e}^c \right) \mathrm{d}\mathbf{f}, \\
\frac{\mathrm{d}u}{\mathrm{d}\eta} &= \int q(\mathbf{f}|\mathbf{y}) \left( [\mathbf{m} - \mathbf{f}]^{\text{T}} \mathbf{y} - [\mathbf{m} - \mathbf{f}]^{\text{T}} \boldsymbol{\pi}_\eta \right) \mathrm{d}\mathbf{f} = -\int q(\mathbf{f}|\mathbf{y}) \, [\mathbf{m} - \mathbf{f}]^{\text{T}} \boldsymbol{\pi}_\eta \, \mathrm{d}\mathbf{f}, \\
\frac{\mathrm{d}^2 u}{\mathrm{d}\eta^2} &= -\int q(\mathbf{f}|\mathbf{y}) \, [\mathbf{m} - \mathbf{f}]^{\text{T}} \left( \Pi_\eta - \boldsymbol{\pi}_\eta \boldsymbol{\pi}_\eta^{\text{T}} \right) [\mathbf{m} - \mathbf{f}] \, \mathrm{d}\mathbf{f},
\end{aligned}$$

where

$$\pi_\eta^c \overset{\text{def}}{=} \frac{\exp\left[(1-\eta)\mathbf{f} + \eta\mathbf{m}\right]^{\text{T}} \mathbf{e}^c}{\sum_{c'=1}^{C} \exp\left[(1-\eta)\mathbf{f} + \eta\mathbf{m}\right]^{\text{T}} \mathbf{e}^{c'}}, \qquad \boldsymbol{\pi}_\eta \overset{\text{def}}{=} \left( \pi_\eta^1, \ldots, \pi_\eta^C \right)^{\text{T}},$$

and $\Pi_\eta$ is a diagonal matrix with $\boldsymbol{\pi}_\eta$ along its diagonal. The first derivative $\mathrm{d}u/\mathrm{d}\eta$ at $\eta = 1$ is zero because $\boldsymbol{\pi}_1$ is independent of $\mathbf{f}$ and the mean of $\mathbf{f}$ under $q(\mathbf{f}|\mathbf{y})$ is $\mathbf{m}$. Moreover, the second derivative $\mathrm{d}^2u/\mathrm{d}\eta^2$ is non-positive because the matrix within the parentheses is positive semi-definite. Hence $u$ is concave in $\eta$, and a maximal is $u(1) = \log p(\mathbf{y}|\mathbf{m})$ where the gradient is zero. ■

### B.6 A Data-independent Lower Bound on the Marginal Likelihood

We first introduce a bound on $h(\mathbf{y}; q, \mathbf{b}, S)$ when the variational posterior is chosen to be an isotropic Gaussian.

**Lemma 29** *Let $q \equiv \mathcal{N}(\mathbf{m}, \sigma_{\text{v}}^2 I)$, and let $h(\mathbf{y}; q, \mathbf{b}, S)$ be as defined by Equation 19. Then*

$$\max_{\mathbf{b},S} h(\mathbf{y}; q, \mathbf{b}, S) \geq -\frac{C-1}{2} \left[ 2\sqrt{\frac{\sigma_{\text{v}}^2}{C} + \frac{1}{4}} - \log\left( \sqrt{\frac{\sigma_{\text{v}}^2}{C} + \frac{1}{4}} + \frac{1}{2} \right) - 1 \right] - \frac{1}{2}\log C + \frac{1}{2} \log \frac{\exp 2\mathbf{m}^{\text{T}}\mathbf{y}}{\sum_{c=1}^{C} \exp 2\mathbf{m}^{\text{T}}\mathbf{e}^c}.$$

*with equality when* $\mathbf{m} = \mathbf{0}$*. This is a decreasing function of* $C$ *and* $\sigma_{\mathrm{v}}^2$*. Moreover*

$$\max_{\mathbf{b},S} h(\mathbf{y};q,\mathbf{b},S) > -\frac{1}{2}\sigma_{\mathrm{v}}^2 - \frac{1}{2}\log C + \frac{1}{2}\log\frac{\exp 2\mathbf{m}^{\mathrm{T}}\mathbf{y}}{\sum_{c=1}^{C}\exp 2\mathbf{m}^{\mathrm{T}}\mathbf{e}^c}.$$

**Proof** Let $\tilde{g}^c(\mathbf{b},S) \overset{\text{def}}{=} \exp(\mathbf{b}-\mathbf{e}^c)^{\mathrm{T}}S^{-1}(\mathbf{b}-\mathbf{e}^c)$. Using Cauchy-Schwarz inequality on $\sum_{c=1}^{C} g^c$ gives

$$\log\sum_{c=1}^{C} g^c \leq \frac{1}{2}\log\sum_{c=1}^{C}\exp 2\mathbf{m}^{\mathrm{T}}\mathbf{e}^c + \frac{1}{2}\log\sum_{c=1}^{C}\tilde{g}^c.$$

We use this inequality together with the choice of distribution $q$, which has variance $\sigma_{\mathrm{v}}^2$ in all directions, to obtain $h(\mathbf{y};q,\mathbf{b},S) \geq \tilde{h}(\mathbf{y};q,\mathbf{b},S)$, where

$$\tilde{h}(\mathbf{y};q,\mathbf{b},S) \overset{\text{def}}{=} \frac{C}{2} + \frac{C}{2}\log\sigma_{\mathrm{v}}^2 + \frac{1}{2}\log|S| - \frac{\sigma_{\mathrm{v}}^2}{2}\operatorname{tr}S - \frac{1}{2}\log\sum_{c=1}^{C}\tilde{g}^c(\mathbf{b},S) + \frac{1}{2}\log\frac{\exp 2\mathbf{m}^{\mathrm{T}}\mathbf{y}}{\sum_{c=1}^{C}\exp 2\mathbf{m}^{\mathrm{T}}\mathbf{e}^c}.$$

Let $\bar{\tilde{g}}^c \overset{\text{def}}{=} \tilde{g}^c/\sum_{c'=1}^{C}\tilde{g}^{c'}$ and $\bar{\tilde{\mathbf{g}}} \overset{\text{def}}{=} (\bar{\tilde{g}}^1,\ldots,\bar{\tilde{g}}^C)^{\mathrm{T}}$, where the arguments $(\mathbf{b},S)$ are suppressed in the notation. Let $\bar{\tilde{G}}$ be the diagonal matrix with $\bar{\tilde{\mathbf{g}}}$ along its diagonal. Also, define $\tilde{A} \overset{\text{def}}{=} \mathbf{b}\mathbf{b}^{\mathrm{T}} - \bar{\tilde{\mathbf{g}}}\mathbf{b}^{\mathrm{T}} - \mathbf{b}\bar{\tilde{\mathbf{g}}}^{\mathrm{T}} + \bar{\tilde{G}}$. These two definitions are analogous to the definitions of $\bar{\mathbf{g}}$ and $A$ in Equations 17 and 18.

Let the maximum of $\tilde{h}$ be at $(\mathbf{b}^*,S^*)$. It is straightforward to modify Lemmas 25, 26 and 27 for $\tilde{h}$. The modified Lemma 25 says that $(\mathbf{b}^*,S^*)$ is unique. The other two modified lemmas will give the self-consistent equations

$$\mathbf{b}^* = \bar{\tilde{\mathbf{g}}}(\mathbf{b}^*,S^*), \qquad\qquad -\sigma_{\mathrm{v}}^2(S^*)^2 + S^* + A^* = 0.$$

By symmetry, $\mathbf{b}^* = \mathbf{1}/C$ and $A^* = I/C - \mathbf{1}\mathbf{1}^{\mathrm{T}}/C^2$. An eigenpair of $A^*$ is $(0,\mathbf{1}/\sqrt{C})$; the other $(C-1)$ eigenpairs are $(1/C,\mathbf{u}_d)$, $d = 1\ldots(C-1)$, where $\mathbf{1}^{\mathrm{T}}\mathbf{u}_d = 0$ in addition to the orthonormal conditions. Let

$$\lambda \overset{\text{def}}{=} \sqrt{\sigma_{\mathrm{v}}^2/C + 1/4} + 1/2. \tag{63}$$

Since $\sigma_{\mathrm{v}}^2 S^* = (\sigma_{\mathrm{v}}^2 A^* + I/4)^{1/2} + I/2$ (see the proof for Lemma 28 in Appendix B.4), the eigenvalues of $S^*$ are $\sigma_{\mathrm{v}}^{-2}$ (with algebraic multiplicity one) and $\sigma_{\mathrm{v}}^{-2}\lambda$ (with algebraic multiplicity $C-1$), and the eigenvectors of $S^*$ are those of $A^*$. Thus the determinant and trace of $S^*$ can be readily obtained. With $\mathbf{b}^* = \mathbf{1}/C$, observe that

$$(\mathbf{b}^* - \mathbf{e}^c)^{\mathrm{T}}\mathbf{1}/\sqrt{C} = 0, \qquad\qquad (\mathbf{b}^* - \mathbf{e}^c)^{\mathrm{T}}\mathbf{u}_d = -u_{dc},$$

where $u_{dc}$ is the $c$th entry in the eigenvector $\mathbf{u}_d$. For the exponent of $\tilde{g}^c$, using $(S^*)^{-1}$ in its eigendecomposition and the two observations above gives $(\mathbf{b}^* - \mathbf{e}^c)^{\mathrm{T}}(S^*)^{-1}(\mathbf{b}^* - \mathbf{e}^c) = (\sigma_{\mathrm{v}}^2/\lambda)\sum_{d=1}^{C-1}u_{dc}^2$. But the eigenvectors of $S^*$ are orthonormal, so $\left((1/\sqrt{C})^2 + \sum_{d=1}^{C-1}u_{dc}^2\right)$ is unity. Hence, we have $(\mathbf{b}^* - \mathbf{e}^c)^{\mathrm{T}}(S^*)^{-1}(\mathbf{b}^* - \mathbf{e}^c) = (\sigma_{\mathrm{v}}^2/\lambda)(1 - 1/C)$. This is independent of $c$. Therefore

$$\begin{aligned}
&\max_{\mathbf{b},S}\tilde{h}(\mathbf{y};q,\mathbf{b},S)\\
&= \frac{C}{2} + \frac{C}{2}\log\sigma_{\mathrm{v}}^2 + \frac{1}{2}\log\sigma_{\mathrm{v}}^{-2}(\sigma_{\mathrm{v}}^{-2}\lambda)^{C-1} - \frac{\sigma_{\mathrm{v}}^2}{2}\left(\sigma_{\mathrm{v}}^{-2} + (C-1)\sigma_{\mathrm{v}}^{-2}\lambda\right)\\
&\qquad\qquad - \frac{1}{2}\log C\exp\left[\frac{\sigma_{\mathrm{v}}^2}{\lambda}\left(1 - \frac{1}{C}\right)\right] + \frac{1}{2}\log\frac{\exp 2\mathbf{m}^{\mathrm{T}}\mathbf{y}}{\sum_{c=1}^{C}\exp 2\mathbf{m}^{\mathrm{T}}\mathbf{e}^c}\\
&= \frac{C}{2} + \frac{C-1}{2}\log\lambda - \frac{1}{2}(1 + (C-1)\lambda) - \underline{\frac{1}{2}\frac{\sigma_{\mathrm{v}}^2}{\lambda}\left(1 - \frac{1}{C}\right)} - \frac{1}{2}\log C + \frac{1}{2}\log\frac{\exp 2\mathbf{m}^{\mathrm{T}}\mathbf{y}}{\sum_{c=1}^{C}\exp 2\mathbf{m}^{\mathrm{T}}\mathbf{e}^c}.
\end{aligned}$$

The underlined term can be simplified to $(C-1)(1-\lambda)/2$ by expressing $\sigma_{\mathrm{v}}^2$ in $\lambda$ using (63). Further simplification and substitution with the definition of $\lambda$ gives

$$\max_{\mathbf{b},S} \tilde{h}(\mathbf{y};q,\mathbf{b},S) = -\frac{C-1}{2}\left[2\sqrt{\frac{\sigma_{\mathrm{v}}^2}{C}+\frac{1}{4}} - \log\left(\sqrt{\frac{\sigma_{\mathrm{v}}^2}{C}+\frac{1}{4}}+\frac{1}{2}\right) - 1\right] \\ -\frac{1}{2}\log C + \frac{1}{2}\log\frac{\exp 2\mathbf{m}^{\mathrm{T}}\mathbf{y}}{\sum_{c=1}^{C}\exp 2\mathbf{m}^{\mathrm{T}}\mathbf{e}^{c}}.$$

Combining this with $h(\mathbf{y};q,\mathbf{b},S) \geq \tilde{h}(\mathbf{y};q,\mathbf{b},S)$ gives the the first inequality in the lemma statement. When $\mathbf{m}=\mathbf{0}$, we can obtain a modification of the proof using $\sum_{c=1}^{C} g^c$ directly without bounding through the Cauchy-Schwarz inequality. This modified proof shows that

$$\max_{\mathbf{b},S} h(\mathbf{y};q,\mathbf{b},S) = -\frac{C-1}{2}\left[2\sqrt{\frac{\sigma_{\mathrm{v}}^2}{C}+\frac{1}{4}} - \log\left(\sqrt{\frac{\sigma_{\mathrm{v}}^2}{C}+\frac{1}{4}}+\frac{1}{2}\right) - 1\right] - \log C. \qquad (\mathbf{m}=\mathbf{0})$$

The first term is a decreasing function of $C$, and we now show that this first term is bounded by $-\sigma_{\mathrm{v}}^2/2$ from below. Let $f(x) \stackrel{\mathrm{def}}{=} x^2-3x+2+\log x$. Then $f=0$ and $\mathrm{d}f/\mathrm{d}x=0$ at $x=1$, and $\mathrm{d}^2 f/\mathrm{d}x^2>0$ in the domain $x\geq 1$. Therefore, $f(x)\geq 0$ for all $x\geq 1$. Then, for function $f(x)$ we set $x \stackrel{\mathrm{def}}{=} \sqrt{\sigma_{\mathrm{v}}^2/C+1/4}+1/2$ and use $C-1<C$ to complete the proof after rearrangement. ■

**Proof (of Theorem 9)**

$$\begin{aligned}\log p(\mathbf{y}) \geq \max_{q,\{\mathbf{b}_i\},\{S_i\}} \log Z_h \geq \max_{\{\mathbf{b}_i\},\{S_i\}} \log Z_h\big|_{q(\mathbf{f}|\mathbf{y})=\mathcal{N}(\mathbf{0},\sigma_{\mathrm{v}}^2 I)} \\ = \frac{nC}{2}+\frac{nC}{2}\log\sigma_{\mathrm{v}}^2 - \frac{1}{2}\log|K| - \frac{\sigma_{\mathrm{v}}^2}{2}\operatorname{tr}K^{-1} + \sum_{i=1}^{n}\max_{\mathbf{b}_i,S_i} h(\mathbf{y}_i,\mathcal{N}(\mathbf{0},\sigma_{\mathrm{v}}^2 I),\mathbf{b}_i,S_i) \\ = \frac{nC}{2}+\frac{nC}{2}\log\sigma_{\mathrm{v}}^2 - \frac{1}{2}\log|K| - \frac{\sigma_{\mathrm{v}}^2}{2}\operatorname{tr}K^{-1} + n\max_{\mathbf{b},S} h(\mathbf{y},\mathcal{N}(\mathbf{0},\sigma_{\mathrm{v}}^2 I),\mathbf{b},S).\end{aligned}$$

Lemma 29 is then applied on $\max h$. ■

**Proof (of Theorem 10)**

$$\begin{aligned}\log p(\mathbf{y}) &\geq \max_{q,\{\mathbf{b}_i\},\{S_i\}} \log Z_h \\ &\geq \max_{\{\mathbf{b}_i\},\{S_i\}} \log Z_h\big|_{q(\mathbf{f})=p(\mathbf{f})} \\ &= \max_{\{\mathbf{b}_i\},\{S_i\}} \left[\frac{nC}{2}+\frac{1}{2}\sum_{i=1}^{n}\Big(\log|S_iK_i| - \operatorname{tr}S_iK_i\Big) - \sum_{i=1}^{n}\log\sum_{c=1}^{C}\exp\left[\frac{1}{2}(\mathbf{b}_i-\mathbf{e}^c)^{\mathrm{T}}S_i^{-1}(\mathbf{b}_i-\mathbf{e}^c)\right]\right].\end{aligned}$$

The same expression can be obtained by setting $V=K$ and $\mathbf{m}=\eta\mathbf{1}$ and maximizing the resultant expression with respect to $\eta$. For $K_1=K_2,\ldots K_n=K^{\mathrm{c}}$, we have

$$\frac{1}{n}\log p(\mathbf{y}) \geq \max_{\mathbf{b},S}\left[\frac{C}{2}+\frac{1}{2}\log|SK^{\mathrm{c}}| - \frac{1}{2}\operatorname{tr}SK^{\mathrm{c}} - \log\sum_{c=1}^{C}\exp\left[\frac{1}{2}(\mathbf{b}-\mathbf{e}^c)^{\mathrm{T}}S^{-1}(\mathbf{b}-\mathbf{e}^c)\right]\right].$$

For the choice of $K^{\mathrm{c}} \stackrel{\mathrm{def}}{=} \sigma^2 I$, we apply Lemma 29. ■

### B.7 Lower Bound on Predictive Probability: Proof of Theorem 12

For a set of $n_*$ test inputs $X_* \overset{\text{def}}{=} \{\mathbf{x}_{*1}, \ldots, \mathbf{x}_{*n_*}\}$, the log joint predictive probability for $\mathbf{x}_{*j}$ to be in class $c_j$ $(j = 1 \ldots n_*)$ is

$$\begin{aligned}
\log p(\{y_{*j}^{c_j} = 1\}_{j=1}^{n_*}|\mathbf{y}) &= \log \int p(\{y_{*j}^{c_j} = 1\}_{j=1}^{n_*}|\mathbf{f}_*)\, p(\mathbf{f}_*|\mathbf{y})\, d\mathbf{f}_* \\
&= \log \int p(\{y_{*j}^{c_j} = 1\}_{j=1}^{n_*}|\mathbf{f}_*)\, p(\mathbf{f}_*, \mathbf{f}|\mathbf{y}) d\mathbf{f}_*\, d\mathbf{f} \\
&= \log \int p(\{y_{*j}^{c_j} = 1\}_{j=1}^{n_*}|\mathbf{f}_*)\, q(\mathbf{f}_*, \mathbf{f}|\mathbf{y}) \frac{p(\mathbf{f}_*, \mathbf{f}|\mathbf{y})}{q(\mathbf{f}_*, \mathbf{f}|\mathbf{y})}\, d\mathbf{f}_* d\mathbf{f}.
\end{aligned}$$

Applying Jensen's inequality gives the inequality

$$\begin{aligned}
\log p(\{y_{*j}^{c_j} = 1\}_{j=1}^{n_*}|\mathbf{y}) &\geq \int q(\mathbf{f}_*, \mathbf{f}|\mathbf{y}) \log p(\{y_{*j}^{c_j} = 1\}_{j=1}^{n_*}|\mathbf{f}_*) \frac{p(\mathbf{f}_*, \mathbf{f}|\mathbf{y})}{q(\mathbf{f}_*, \mathbf{f}|\mathbf{y})}\, d\mathbf{f}_* d\mathbf{f} \\
&= \int q(\mathbf{f}_*|\mathbf{y}) \log p(\{y_{*j}^{c_j} = 1\}_{j=1}^{n_*}|\mathbf{f}_*)\, d\mathbf{f}_* - \mathrm{KL}(q(\mathbf{f}_*, \mathbf{f}|\mathbf{y}) \| p(\mathbf{f}_*, \mathbf{f}|\mathbf{y})).
\end{aligned}$$

Within the first term, the conditional joint predictive probability factorizes across the $\mathbf{x}_{*j}$s. The second term is the Kullback-Leibler divergence from $q(\mathbf{f}_*, \mathbf{f}|\mathbf{y})$ to $p(\mathbf{f}_*, \mathbf{f}|\mathbf{y})$, which can be written as

$$\begin{aligned}
\mathrm{KL}(q(\mathbf{f}_*, \mathbf{f}|\mathbf{y}) \| p(\mathbf{f}_*, \mathbf{f}|\mathbf{y})) &\overset{\text{def}}{=} \int q(\mathbf{f}, \mathbf{f}_*|\mathbf{y}) \log \frac{q(\mathbf{f}, \mathbf{f}_*|\mathbf{y})}{p(\mathbf{f}, \mathbf{f}_*|\mathbf{y})}\, d\mathbf{f}_* d\mathbf{f} \\
&= \int q(\mathbf{f}|\mathbf{y})\, p(\mathbf{f}_*|\mathbf{f}) \log \frac{q(\mathbf{f}|\mathbf{y})\, p(\mathbf{f}_*|\mathbf{f})}{p(\mathbf{f}|\mathbf{y})\, p(\mathbf{f}_*|\mathbf{f})}\, d\mathbf{f}_* d\mathbf{f} = \int q(\mathbf{f}|\mathbf{y}) \log \frac{q(\mathbf{f}|\mathbf{y})}{p(\mathbf{f}|\mathbf{y})}\, d\mathbf{f}, \qquad (64)
\end{aligned}$$

which is $\mathrm{KL}(q(\mathbf{f}|\mathbf{y}) \| p(\mathbf{f}|\mathbf{y}))$. Hence

$$\log p(\{y_{*j}^{c_j} = 1\}_{j=1}^{n_*}|\mathbf{y}) \geq \sum_{j=1}^{n_*} \int q(\mathbf{f}_{*j}|\mathbf{y}) \log p(y_{*j}^{c_j} = 1|\mathbf{f}_{*j})\, d\mathbf{f}_{*j} - \mathrm{KL}(q(\mathbf{f}|\mathbf{y}) \| p(\mathbf{f}|\mathbf{y})).$$

Theorem 6 can now be applied to each summand within the first term. For the second term, the KL-divergence is also $\log p(\mathbf{y}) - \log Z_B$, and $\log Z_B$ is lower bounded by $\log Z_h$. ■

**Remark 30** *Derivation 64 has been shown in (Seeger, 2002, Section 2.2) and (Rasmussen and Williams, 2006, Section 7.4.3), but there the exact prior has been used instead of the exact posterior. Our presentation closely follows (Rasmussen and Williams, 2006)'s.*

## Appendix C. Optimization

We provide details on the optimization with respect to the variational parameters $\mathbf{m}$, $V$, $\{\mathbf{b}_i\}$ and $\{S_i\}$ in Sections C.1 to C.3. In Section C.4, we give the derivation for the updates to the hyperparameters required for model learning in the sparse approximation.

Parameters $\mathbf{m}$ and $\mathbf{b}_i$s are updated together using Newton-Raphson in Section C.2. In regions of high-curvature, this update can be modified to include include a step-size $\eta$, the value of which can be determined using the method of false position.

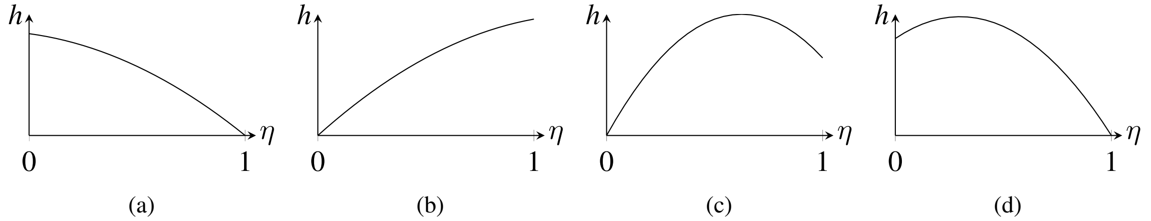


Figure 7: Four possible shapes of a segment of the concave function $h$ within the convex combination coefficient $\eta \in [0,1]$. The horizontal axis is along $\eta$, with $S^{\text{cc}} = S$ at $\eta = 0$ and $S^{\text{cc}} = S^{\text{fx}}$ at $\eta = 1$. The vertical axis is the variational lower bound $h(\mathbf{y}; q, \mathbf{b}, S)$. Cases (b) and (c) can be removed from consideration, since update $S^{\text{cc}}$ is only used when $h$ is higher at $S$ than at $S^{\text{fx}}$. Case (a) is eliminated by showing that the gradient with respect to $\eta$ at $\eta = 0$ is non-negative.

For $V$ and the $S_i$s, their fixed point updates given in Section 3 are computed and tested for improvement in the variational lower bound.[9] When the bound is worse at the fixed point updates, we search for the optimal convex combination between the previous value and the fixed-point update. For example, $S_i^{\text{cc}} = (1-\eta)S_i + \eta S_i^{\text{fx}}$, where $S_i$ is the previous value and $S_i^{\text{fx}}$ is the fixed point update. We optimize $\eta$ using the method of false position with end-point down-weighting. Sections C.1 and C.3 give the gradients with respect to $\eta$ and guarantee the existence of an optimal $\eta$.

### C.1 Optimization for $S$ along $\eta$

When the fixed-point $S^{\text{fx}}$ improves the bound over $S$, it is accepted as an update. Otherwise, we use $S^{\text{cc}} \overset{\text{def}}{=} (1-\eta)S + \eta S^{\text{fx}}$, and we search for a $\eta \in [0,1]$ that optimizes the bound using the false position method. Matrix $S^{\text{cc}}$ is guaranteed to be positive definite, since it is a convex combination of two positive definite matrices. Let $W = S - V^{-1}$ and $W^{\text{fx}} = S^{\text{fx}} - V^{-1}$. Matrices $W$ and and $W^{\text{fx}}$ are positive semi-definite; see Lemma 28.[10] Define the gradient $\Delta \overset{\text{def}}{=} \mathrm{d}S^{\text{cc}}/\mathrm{d}\eta = S^{\text{fx}} - S = W^{\text{fx}} - W$. The search gradient along $\eta$ is

$$\frac{\partial h(\mathbf{y}; q, \mathbf{b}, S^{\text{cc}})}{\partial \eta} = \frac{1}{2}\,\mathrm{tr}\left[\left\{-V + (S^{\text{cc}})^{-1} + (S^{\text{cc}})^{-1} A^{\text{cc}} (S^{\text{cc}})^{-1}\right\}\Delta\right],$$

where $A^{\text{cc}}$ is given by (18) evaluated at $S^{\text{cc}}$ (recall that $A$ depends on $g^c$, a function of $S$).

The optimal value of $\eta$ is found using the false position method, which requires the maximal to be bracketed within $S^{\text{cc}} = S$ and $S^{\text{cc}} = S^{\text{fx}}$. Figure 7 enumerates the four possible segments of a one-dimensional concave function. Let $\eta = 0$ at the start of the segment and $\eta = 1$ at the end. If update $S^{\text{cc}}$ is only used when $h(\mathbf{y}; q, \mathbf{b}, S) > h(\mathbf{y}; q, \mathbf{b}, S^{\text{fx}})$, then segments (b) and (c) need not be considered further. To show that the segment is always of the type given by Figure 7d, we require

[9] These fixed point updates guarantees positive definiteness, a property which is absent in straightforward gradient ascent. To guarantee positive definiteness, one may suppose a viable alternative is to update the Cholesky factors or eigenvectors and eigenvalues. Unfortunately, the variational lower bound is not concave with respect to these factorizations, so they cannot be used straightforwardly.

[10] In the beginning, if $W$ is not positive semi-definite, we can re-initialize it to be so, either by using a fixed positive semi-definite matrix, or by letting $W$ be $W^{\text{fx}}$ in the first iteration.

$\partial h/\partial\eta$ to be non-negative at $\eta = 0$. We proceed to show this. At $\eta = 0$, we have $S^{\mathrm{cc}} = S$ and $A^{\mathrm{cc}} = A$ evaluated at $S$. Furthermore, we have $A = S^{\mathrm{fx}}VS^{\mathrm{fx}} - S^{\mathrm{fx}}$ since $S^{\mathrm{fx}}$ satisfies (60). Thus

$$\begin{aligned}
\left.\frac{\partial h(\mathbf{y};q,\mathbf{b},S^{\mathrm{cc}})}{\partial\eta}\right|_{\eta=0} &= \frac{1}{2}\operatorname{tr}\left[\left(-V+S^{-1}+S^{-1}\left(S^{\mathrm{fx}}VS^{\mathrm{fx}}-S^{\mathrm{fx}}\right)S^{-1}\right)\Delta\right] \\
&= \frac{1}{2}\operatorname{tr}\left[S^{-1}\left(S^{\mathrm{fx}}VS^{\mathrm{fx}}-SVS-S^{\mathrm{fx}}+S\right)S^{-1}\Delta\right] \\
&= \frac{1}{2}\operatorname{tr}\left[S^{-1}\left(W^{\mathrm{fx}}VW^{\mathrm{fx}}-WVW\right)S^{-1}\Delta\right]+\frac{1}{2}\operatorname{tr}\left[S^{-1}\Delta S^{-1}\Delta\right].
\end{aligned}$$

The second term on the right of the equality is non-negative, so its removal gives

$$\begin{aligned}
\left.\frac{\partial h(\mathbf{y};q,\mathbf{b},S^{\mathrm{cc}})}{\partial\eta}\right|_{\eta=0} &\geq \frac{1}{2}\operatorname{tr}\left[S^{-1}\left(W^{\mathrm{fx}}VW^{\mathrm{fx}}-WVW\right)S^{-1}\Delta\right] \\
&= \frac{1}{2}\operatorname{tr}\left[S^{-1}\left(W^{\mathrm{fx}}VW^{\mathrm{fx}}-WVW+W^{\mathrm{fx}}VW-W^{\mathrm{fx}}VW\right)S^{-1}\Delta\right] \\
&= \frac{1}{2}\operatorname{tr}\left[S^{-1}\left(W^{\mathrm{fx}}VW-WVW+W^{\mathrm{fx}}VW^{\mathrm{fx}}-W^{\mathrm{fx}}VW\right)S^{-1}\Delta\right] \\
&= \frac{1}{2}\operatorname{tr}\left[S^{-1}\left(\Delta VW+W^{\mathrm{fx}}V\Delta\right)S^{-1}\Delta\right] \\
&= \frac{1}{2}\operatorname{tr}\left[S^{-1}\Delta VWS^{-1}\Delta\right]+\frac{1}{2}\operatorname{tr}\left[S^{-1}W^{\mathrm{fx}}V\Delta S^{-1}\Delta\right] \\
&\geq 0.
\end{aligned}$$

### C.2 Joint Optimization for m and b

Let $K \overset{\text{def}}{=} (K_1|K_2|\ldots|K_n)$ be a partition of $K$, where each $K_i$ is a $Cn$-by-$C$ matrix. We wish to optimize the variational lower bound $\log Z_h$ (21) with respect to $\mathbf{b}$ by setting $\mathbf{m} = K(\mathbf{y}-\mathbf{b})$. Call this particular setting of parameters $\log Z_h^*$. The gradient of $\log Z_h^*$ with respect to $\mathbf{b}$ including the indirect contribution from $\mathbf{m}$ is

$$\begin{aligned}
\frac{\partial\log Z_h^*}{\partial\mathbf{b}} &= \frac{\partial\log Z_h^*|_{\mathbf{m}\text{ constant}}}{\partial\mathbf{b}}+\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}\mathbf{b}}\frac{\partial\log Z_h^*|_{\mathbf{b}\text{ constant}}}{\partial\mathbf{m}} \\
&= -S^{-1}(\mathbf{b}-\bar{\mathbf{g}})-K\left(-K^{-1}\mathbf{m}+\mathbf{y}-\bar{\mathbf{g}}\right) \\
&= -(K+S^{-1})(\mathbf{b}-\bar{\mathbf{g}}).
\end{aligned}$$

Unlike case of per-datum update for $\mathbf{b}_i$, we find the fixed-point update setting $\mathbf{b}$ to $\bar{\mathbf{g}}$ ineffective. Therefore, we use the Newton-Raphson update. The required Hessian is

$$\frac{\partial^2\log Z_h^*}{\partial\mathbf{b}\partial\mathbf{b}^{\mathrm{T}}} = -(K+S^{-1})\left(I-\frac{\partial\bar{\mathbf{g}}}{\partial\mathbf{b}^{\mathrm{T}}}\right) = -(K+S^{-1})\left(I+(\bar{G}-\tilde{G})(K+S^{-1})\right).$$

The Hessian is negative definite, so $\log Z_h$ is concave in $\mathbf{b}$. The second order update is

$$\mathbf{b}^{\mathrm{NR}} = \mathbf{b}-\left(I+(\bar{G}-\tilde{G})(K+S^{-1})\right)^{-1}(\mathbf{b}-\bar{\mathbf{g}}). \tag{65}$$

#### C.2.1 Joint Optimization for m and b in Sparse Approximation

For sparse approximation, the update is similar to Equation 65, the only difference being the replacement of $K$ with $K_{\mathrm{f}}^{\mathrm{T}}K^{-1}K_{\mathrm{f}}$:

$$\mathbf{b}^{\mathrm{NR}} = \mathbf{b}-\left(I+(\bar{G}-\tilde{G})\left(K_{\mathrm{f}}^{\mathrm{T}}K^{-1}K_{\mathrm{f}}+S^{-1}\right)\right)^{-1}(\mathbf{b}-\bar{\mathbf{g}}).$$

### C.3 Optimization for $V$ along $\eta$

Let $W$ be the block diagonal matrix with the $i$th block given by $W_i \overset{\text{def}}{=} S_i - V_i^{-1} \succ 0$. When the fixed-point $V^{\text{fx}} = (K^{-1} + W)^{-1}$ improves the bound over $V$, it is accepted as an update. Otherwise, we use $V^{\text{cc}} \overset{\text{def}}{=} (1-\eta)V + \eta V^{\text{fx}}$, and we search for a $\eta \in [0,1]$ that optimizes the bound using the false position method. Matrix $V^{\text{cc}}$ is guaranteed to be positive definite, since it is a convex combination of two positive definite matrices. Let $\Delta \overset{\text{def}}{=} \mathrm{d}V^{\text{cc}}/\mathrm{d}\eta = V^{\text{fx}} - V$. Below, we shall make explicit that the lower bound $\log Z_h$ is parameterized by the covariance $V$ of the variational posterior. The search gradient is

$$\frac{\partial \log Z_h(V^{\text{cc}})}{\partial \eta} = \frac{1}{2}\operatorname{tr}\left((V^{\text{cc}})^{-1}\Delta\right) - \frac{1}{2}\operatorname{tr}\left(K^{-1}\Delta\right) + \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}\left((V_i^{\text{cc}})^{-1}\Delta_i\right) - \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}\left(S_i\Delta_i\right),$$

where $V_i^{\text{cc}}$ and $\Delta_i$ are the $i$th blocks along the diagonal of $V^{\text{cc}}$ and $\Delta$ respectively. The update $V^{\text{cc}}$ is only used when $\log Z_h(V) > \log Z_h(V^{\text{cc}})$. By arguments similar to those for the update $S^{\text{cc}}$ discussed in Appendix C.1, we can guarantee that there is a maximum between $V$ and $V^{\text{fx}}$ by showing that $\partial Z_h/\partial\eta$ is non-negative at $\eta = 0$:

$$\begin{aligned}
&\left.\frac{\partial \log Z_h(V^{\text{cc}})}{\partial \eta}\right|_{\eta=0} \\
&= \frac{1}{2}\operatorname{tr}\left(V^{-1}\Delta\right) - \frac{1}{2}\operatorname{tr}\left(K^{-1}\Delta\right) - \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}\left(W_i\Delta_i\right) \\
&= \frac{1}{2}\operatorname{tr}\left(V^{-1}\Delta\right) - \frac{1}{2}\operatorname{tr}\left(K^{-1}\Delta\right) - \frac{1}{2}\operatorname{tr}\left(W\Delta\right) && \text{(since } W \text{ is block diagonal)} \\
&= \frac{1}{2}\operatorname{tr}\left(V^{-1}\Delta\right) - \frac{1}{2}\operatorname{tr}\left(K^{-1}\Delta\right) - \frac{1}{2}\operatorname{tr}\left(\left((V^{\text{fx}})^{-1} - K^{-1}\right)\Delta\right) && \text{(since } V^{\text{fx}} = (K^{-1}+W)^{-1}\text{)} \\
&= \frac{1}{2}\operatorname{tr}\left(\left(V^{-1} - (V^{\text{fx}})^{-1}\right)\Delta\right) \\
&= \frac{1}{2}\operatorname{tr}\left((V^{\text{fx}})^{-1}(V^{\text{fx}} - V)V^{-1}\Delta\right) \\
&= \frac{1}{2}\operatorname{tr}\left((V^{\text{fx}})^{-1}\Delta V^{-1}\Delta\right) \\
&\geq 0.
\end{aligned}$$

#### C.3.1 Optimization for $V$ along $\eta$ in Sparse Approximation

We use the same strategy in the sparse approximation. Let the covariance of the inducing variables be $V^{\text{cc}} \overset{\text{def}}{=} (1-\eta)V + \eta V^{\text{fx}}$. The covariance of the latent variables is $V_{\text{f}}^{\text{cc}} = (1-\eta)V_{\text{f}} + \eta V_{\text{f}}^{\text{fx}}$. Let $\Delta \overset{\text{def}}{=} \mathrm{d}V^{\text{cc}}/\mathrm{d}\eta = V^{\text{fx}} - V$, and let $\Delta_{\text{f}} \overset{\text{def}}{=} \mathrm{d}V_{\text{f}}^{\text{cc}}/\mathrm{d}\eta = V_{\text{f}}^{\text{fx}} - V_{\text{f}} = K_{\text{f}}^{\text{T}}K^{-1}\Delta K^{-1}K_{\text{f}}$. The gradient along $\eta \in [0,1]$ for the false position update is

$$\frac{\partial \log \tilde{Z}_h(V^{\text{cc}})}{\partial \eta} = \frac{1}{2}\operatorname{tr}((V^{\text{cc}})^{-1}\Delta) - \frac{1}{2}\operatorname{tr}(K^{-1}\Delta) + \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}((V_{\text{f}i}^{\text{cc}})^{-1}\Delta_{\text{f}i}) - \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}(S_i\Delta_{\text{f}i}).$$

The proof that $\partial \log \tilde{Z}_h(V^{\text{cc}})/\partial\eta$ is non-negative at $\eta = 0$ follows the same reasoning as that for the non-sparse approximation.

### C.4 Hyper-parameter Estimation in Sparse Approximation

In this section, we give the gradients of the optimized variational lower bound $\tilde{Z}_h^*$ for the sparse case. First, we introduce

$$T \overset{\text{def}}{=} K^{-1} - K^{-1}VK^{-1}, \qquad \Gamma_j \overset{\text{def}}{=} K\left(\frac{\partial K^{-1}K_{\text{f}}}{\partial\theta_j}\right) = \frac{\partial K_{\text{f}}}{\partial\theta_j} - \frac{\partial K}{\partial\theta_j}K^{-1}K_{\text{f}}. \tag{66}$$

Then

$$\frac{\partial \mathbf{m}_{\text{f}}}{\partial\theta_j} = \Gamma_j^{\text{T}}K^{-1}\mathbf{m} = \Gamma_j^{\text{T}}\boldsymbol{\alpha}, \tag{67}$$

$$\frac{\partial V_{\text{f}}}{\partial\theta_j} = \frac{\partial K_{\text{ff}}}{\partial\theta_j} - K_{\text{f}}^{\text{T}}K^{-1}\frac{\partial K}{\partial\theta_j}K^{-1}K_{\text{f}} - \Gamma_j^{\text{T}}TK_{\text{f}} - K_{\text{f}}^{\text{T}}T\Gamma_j. \tag{68}$$

The gradient is

$$\begin{aligned}\frac{\mathrm{d}\log\tilde{Z}_h^*}{\mathrm{d}\theta_j} = &-\frac{1}{2}\operatorname{tr}\left(K^{-1}\frac{\partial K}{\partial\theta_j}\right) + \frac{1}{2}\operatorname{tr}\left(K^{-1}VK^{-1}\frac{\partial K}{\partial\theta_j}\right) + \frac{1}{2}\mathbf{m}^{\text{T}}K^{-1}\frac{\partial K}{\partial\theta_j}K^{-1}\mathbf{m}\\ &+\frac{\partial \mathbf{m}_{\text{f}}^{\text{T}}}{\partial\theta_j}\mathbf{y} + \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}\left((V_{\text{f}i})^{-1}\frac{\partial V_{\text{f}i}}{\partial\theta_j}\right) - \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}\left(S_i\frac{\partial V_{\text{f}i}}{\partial\theta_j}\right) - \sum_{i=1}^{n}\sum_{c=1}^{C}\bar{g}_i^c\frac{\partial \mathbf{m}_{\text{f}i}^c}{\partial\theta_j}\\ =&\ \frac{1}{2}\operatorname{tr}\left(\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^{\text{T}} - T\right)\frac{\partial K}{\partial\theta_j}\right) + \frac{\partial \mathbf{m}_{\text{f}}^{\text{T}}}{\partial\theta_j}(\mathbf{y}-\bar{\mathbf{g}}) - \frac{1}{2}\operatorname{tr}\left(W_{\text{f}}\frac{\partial V_{\text{f}}}{\partial\theta_j}\right),\end{aligned}$$

where $\boldsymbol{\alpha} \overset{\text{def}}{=} K^{-1}\mathbf{m}$, $W_{\text{f}i} \overset{\text{def}}{=} S_i - V_{\text{f}i}^{-1}$ and $W_{\text{f}}$ is a block diagonal matrix of the $W_{\text{f}i}$s. Let us investigate the second term in the last expression above. Using (67), the definition of $\Gamma_j$ in (66) and the identity $\mathbf{m} = K_{\text{f}}(\mathbf{y}-\bar{\mathbf{g}})$ at optimality (see Section 4.2), we have

$$\begin{aligned}\frac{\partial \mathbf{m}_{\text{f}}^{\text{T}}}{\partial\theta_j}(\mathbf{y}-\bar{\mathbf{g}}) = \boldsymbol{\alpha}^{\text{T}}\Gamma_j(\mathbf{y}-\bar{\mathbf{g}}) &= \boldsymbol{\alpha}^{\text{T}}\frac{\partial K_{\text{f}}}{\partial\theta_j}(\mathbf{y}-\bar{\mathbf{g}}) - \boldsymbol{\alpha}^{\text{T}}\frac{\partial K}{\partial\theta_j}K^{-1}K_{\text{f}}(\mathbf{y}-\bar{\mathbf{g}})\\ &= \boldsymbol{\alpha}^{\text{T}}\frac{\partial K_{\text{f}}}{\partial\theta_j}(\mathbf{y}-\bar{\mathbf{g}}) - \boldsymbol{\alpha}^{\text{T}}\frac{\partial K}{\partial\theta_j}\boldsymbol{\alpha}.\end{aligned}$$

We now turn to the trace expression in the gradient of $\log\tilde{Z}_h^*$. Using (68), the definition of $\Gamma_j$ in (66) and the invariance of trace under cyclic permutations, we obtain

$$\operatorname{tr}\left(W_{\text{f}}\frac{\partial V_{\text{f}}}{\partial\theta_j}\right) = \operatorname{tr}\left(W_{\text{f}}\frac{\partial K_{\text{ff}}}{\partial\theta_j}\right) - \operatorname{tr}\left(W\frac{\partial K}{\partial\theta_j}\right) - 2\operatorname{tr}\left(W_{\text{f}}K_{\text{f}}^{\text{T}}T\frac{\partial K_{\text{f}}}{\partial\theta_j}\right) + 2\operatorname{tr}\left(WKT\frac{\partial K}{\partial\theta_j}\right),$$

where we have used $W \overset{\text{def}}{=} K^{-1}K_{\text{f}}W_{\text{f}}K_{\text{f}}^{\text{T}}K^{-1}$. Further substituting the definition for $T$ from (66) into the last term and simplifying using $W = V^{-1} - K^{-1}$ at optimality (see Equation 33) gives

$$\operatorname{tr}\left(W_{\text{f}}\frac{\partial V_{\text{f}}}{\partial\theta_j}\right) = \operatorname{tr}\left(W_{\text{f}}\frac{\partial K_{\text{ff}}}{\partial\theta_j}\right) + \operatorname{tr}\left((W-2T)\frac{\partial K}{\partial\theta_j}\right) - 2\operatorname{tr}\left(W_{\text{f}}K_{\text{f}}^{\text{T}}T\frac{\partial K_{\text{f}}}{\partial\theta_j}\right).$$

Putting the simplifications back into the gradient of $\log\tilde{Z}_h^*$ gives

$$\frac{\mathrm{d}\log\tilde{Z}_h^*}{\mathrm{d}\theta_j} = -\frac{1}{2}\operatorname{tr}\left(\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^{\text{T}} - T + W\right)\frac{\partial K}{\partial\theta_j}\right) + \operatorname{tr}\left(\left((\mathbf{y}-\bar{\mathbf{g}})\boldsymbol{\alpha}^{\text{T}} + W_{\text{f}}K_{\text{f}}^{\text{T}}T\right)\frac{\partial K_{\text{f}}}{\partial\theta_j}\right) - \frac{1}{2}\operatorname{tr}\left(W_{\text{f}}\frac{\partial K_{\text{ff}}}{\partial\theta_j}\right).$$

## Appendix D. Selection of Inducing Sites

This appendix details the derivation of criterion $d_1$ used for selecting inducing sites actively.

### D.1 A Lower Bound on the Increase to the Marginal Likelihood Bound

Our objective is to add an inducing site $\tilde{\mathbf{x}}_*$ to the current inducing set $\tilde{X}$ so as to maximize the lower bound (30) on the increase in $\log \tilde{Z}_h$. The random variables at $\tilde{\mathbf{x}}_*$ and $\tilde{X}$ are denoted by $z_*$ and $\mathbf{z}$. Let $\mathbf{z}_* \stackrel{\text{def}}{=} (\mathbf{z}^{\mathrm{T}}, z_*)^{\mathrm{T}}$ and $\tilde{X}_* \stackrel{\text{def}}{=} \tilde{X} \cup \{\tilde{\mathbf{x}}_*\}$. The prior on $\mathbf{z}_*$ and $\mathbf{f}$ is

$$p\left(\begin{pmatrix}\mathbf{z}_* \\ \mathbf{f}\end{pmatrix}\right) \stackrel{\text{def}}{=} \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} K_* & K_{\mathrm{f}*} \\ K_{\mathrm{f}*}^{\mathrm{T}} & K_{\mathrm{ff}}\end{pmatrix}\right), \qquad \text{where} \qquad K_* \stackrel{\text{def}}{=} \begin{pmatrix} K & \mathbf{k}_* \\ \mathbf{k}_*^{\mathrm{T}} & k_{**}\end{pmatrix} \qquad K_{\mathrm{f}*} \stackrel{\text{def}}{=} \begin{pmatrix} K_{\mathrm{f}} \\ \mathbf{k}_{\mathrm{f}*}^{\mathrm{T}}\end{pmatrix}.$$

Let

$$\{\mathbf{m}, V, \{\mathbf{b}_i\}, \{S_i\}\} = \arg \max_{\mathbf{m}, V, \{\mathbf{b}_i\}, \{S_i\}} \log \tilde{Z}_h(\mathbf{m}, V, \{\mathbf{b}_i\}, \{S_i\}; \tilde{X}),$$

where $\mathbf{m}$ and $V$ are the mean and covariance of $\mathbf{z}$ in the approximate posterior using inducing set $\tilde{X}$. Let $\log \tilde{Z}_h^*(\tilde{X})$ be the optimal value of the objective function in the equation above. Then a lower bound on the increase is

$$d_1(\tilde{\mathbf{x}}_* | \tilde{X}) \stackrel{\text{def}}{=} \max_{m_*, v_{**}, \mathbf{v}_*} \log Z_h(\mathbf{m}_*, V_*, \{\mathbf{b}_i\}, \{S_i\}; \tilde{X}_*) - \log Z_h^*(\tilde{X}), \tag{69}$$

where the mean $\mathbf{m}_*$ and covariance $V_*$ of the approximate posterior on $\mathbf{z}_*$ are constrained:

$$\mathbf{m}_* \stackrel{\text{def}}{=} \begin{pmatrix}\mathbf{m} \\ m_*\end{pmatrix}, \qquad\qquad V_* \stackrel{\text{def}}{=} \begin{pmatrix} V & \mathbf{v}_* \\ \mathbf{v}_*^{\mathrm{T}} & v_{**}\end{pmatrix}. \tag{70}$$

Denote the posterior distribution of the latent function values $\mathbf{f}$ under the sparse approximation by $q_*(\mathbf{f}|\mathbf{y}) \stackrel{\text{def}}{=} q(\mathbf{f}|\mathbf{y}, \tilde{X}_*)$. This is the approximate posterior using the inducing sites $\tilde{X}_*$, while $q(\mathbf{f}|\mathbf{y}) \stackrel{\text{def}}{=} q(\mathbf{f}|\mathbf{y}, \tilde{X})$ is the posterior using $\tilde{X}$. The choice of the factored form of the approximate posterior in Equation 29 means that

$$q(\mathbf{f}|\mathbf{y}) = \int p(\mathbf{f}|\mathbf{z}) q(\mathbf{z}|\mathbf{y}) \mathrm{d}\mathbf{z}, \qquad\qquad q_*(\mathbf{f}|\mathbf{y}) = \int p(\mathbf{f}|\mathbf{z}_*) q(\mathbf{z}_*|\mathbf{y}) \mathrm{d}\mathbf{z}_*.$$

Expressions for the mean $\mathbf{m}_{\mathrm{f}}$ and covariance $V_{\mathrm{f}}$ of $\mathbf{f}$ under $q(\mathbf{f}|\mathbf{y})$ are given in Equation 31. The mean $\mathbf{m}_{\mathrm{f}*}$ and covariance $V_{\mathrm{f}*}$ of $\mathbf{f}$ under $q_*(\mathbf{f}|\mathbf{y})$ are

$$\begin{aligned} \mathbf{m}_{\mathrm{f}*} &= K_{\mathrm{f}*}^{\mathrm{T}} K_*^{-1} \mathbf{m}_* & V_{\mathrm{f}*} &= K_{\mathrm{ff}} - K_{\mathrm{f}*}^{\mathrm{T}} K_*^{-1} K_{\mathrm{f}*} + K_{\mathrm{f}*}^{\mathrm{T}} K_*^{-1} V_* K_*^{-1} K_{\mathrm{f}*} \\ &= \mathbf{m}_{\mathrm{f}} + \mu \boldsymbol{\kappa}; & &= V_{\mathrm{f}} - (\kappa - \chi) \boldsymbol{\kappa}\boldsymbol{\kappa}^{\mathrm{T}} + \boldsymbol{\psi}\boldsymbol{\kappa}^{\mathrm{T}} + \boldsymbol{\kappa}\boldsymbol{\psi}^{\mathrm{T}}, \end{aligned} \tag{71}$$

where

$$\begin{aligned} &\kappa \stackrel{\text{def}}{=} k_{**} - \mathbf{k}_*^{\mathrm{T}} K^{-1} \mathbf{k}_*, && \nu \stackrel{\text{def}}{=} v_{**} - \mathbf{v}_*^{\mathrm{T}} V^{-1} \mathbf{v}_*, && \chi \stackrel{\text{def}}{=} \nu + \boldsymbol{\nu}^{\mathrm{T}} V^{-1} \boldsymbol{\nu}, && \mu \stackrel{\text{def}}{=} m_* - \mathbf{k}_*^{\mathrm{T}} K^{-1} \mathbf{m}, \\ &\boldsymbol{\kappa} \stackrel{\text{def}}{=} (\mathbf{k}_{\mathrm{f}*} - K_{\mathrm{f}}^{\mathrm{T}} K^{-1} \mathbf{k}_*)/\kappa, && \boldsymbol{\nu} \stackrel{\text{def}}{=} \mathbf{v}_* - V K^{-1} \mathbf{k}_*, && \boldsymbol{\psi} \stackrel{\text{def}}{=} K_{\mathrm{f}}^{\mathrm{T}} K^{-1} \boldsymbol{\nu}. \end{aligned}$$

The two expressions in (71) relate the parameters for $q_*(\mathbf{f}|\mathbf{y})$ to those for $q(\mathbf{f}|\mathbf{y})$. The derivation uses the Banachiewicz inversion formula (Puntanen and Styan, 2005) on $(K_*)^{-1}$. The term

$(\kappa - \chi)$ in the expression for $V_{\mathrm{f}*}$ is non-negative because $K_* \succeq V_*$ at the stationary, which gives $(\kappa - \chi) = \begin{pmatrix} -K^{-1}\mathbf{k}_* & 1 \end{pmatrix} (K_* - V_*) \begin{pmatrix} -K^{-1}\mathbf{k}_* & 1 \end{pmatrix}^{\mathrm{T}} \geq 0$. The posterior covariance for $i$th data point under $q_*$ is the $i$th $C$-by-$C$ diagonal block matrix of $V_{\mathrm{f}*}$:

$$V_{\mathrm{f}*i} = V_{\mathrm{f}i} - (\kappa - \chi)\boldsymbol{\kappa}_i\boldsymbol{\kappa}_i^{\mathrm{T}} + \boldsymbol{\psi}_i\boldsymbol{\kappa}_i^{\mathrm{T}} + \boldsymbol{\kappa}_i\boldsymbol{\psi}_i^{\mathrm{T}}, \tag{72}$$

where $V_{\mathrm{f}i}$ is the $i$th $C$-by-$C$ diagonal block matrix of of $V_{\mathrm{f}}$, and $\boldsymbol{\kappa}_i$ (resp. $\boldsymbol{\psi}_i$) is the $i$th $C$-vector of $\boldsymbol{\kappa}$ (resp. $\boldsymbol{\psi}$). Using Lemma 23, we obtain

$$|V_{\mathrm{f}*i}| = \omega_i |V_{\mathrm{f}i}|, \qquad \text{where} \qquad \omega_i \stackrel{\mathrm{def}}{=} \left(1 + \boldsymbol{\kappa}_i^{\mathrm{T}} V_{\mathrm{f}i}^{-1} \boldsymbol{\psi}_i\right)^2 - \boldsymbol{\kappa}_i^{\mathrm{T}} V_{\mathrm{f}i}^{-1} \boldsymbol{\kappa}_i \left(\kappa - \chi + \boldsymbol{\psi}_i^{\mathrm{T}} V_{\mathrm{f}i}^{-1} \boldsymbol{\psi}_i\right). \tag{73}$$

Since $|V_{\mathrm{f}*i}| > 0$ and $|V_{\mathrm{f}i}| > 0$, so $\omega_i > 0$. We are now ready to express $d_1$ defined by (69) in terms of the parameters, separating $\log \tilde{Z}_h$ into its summands expressed in Equation 30:

$$d_1(\tilde{\mathbf{x}}_* | \tilde{X}) = \max_{m_*, v_{**}, \mathbf{v}_*} \left( d_{\mathrm{KL}}(m_*, v_{**}, \mathbf{v}_*, \tilde{\mathbf{x}}_* | \tilde{X}) + \sum_{i=1}^{n} d_h^i(m_*, v_{**}, \mathbf{v}_*, \tilde{\mathbf{x}}_* | \tilde{X}) \right),$$

where

$$\begin{aligned} d_{\mathrm{KL}}(m_*, v_{**}, \mathbf{v}_*, \tilde{\mathbf{x}}_* | \tilde{X}) &\stackrel{\mathrm{def}}{=} -\mathrm{KL}(q(\mathbf{z}_* \mid \mathbf{y}) \,\|\, p(\mathbf{z}_*)) + \mathrm{KL}(q(\mathbf{z} \mid \mathbf{y}) \,\|\, p(\mathbf{z})) \\ &= \frac{1}{2} + \frac{1}{2}\log\frac{\nu}{\kappa} - \frac{\chi}{2\kappa} - \frac{\mu^2}{2\kappa}; \\ d_h^i(m_*, v_{**}, \mathbf{v}_*, \tilde{\mathbf{x}}_* | \tilde{X}) &\stackrel{\mathrm{def}}{=} h(\mathbf{y}_i; q_{*i}, \mathbf{b}_i, S_i) - h(\mathbf{y}_i; q_i, \mathbf{b}_i, S_i) \\ &= \frac{1}{2}\log\omega_i + \frac{\kappa - \chi}{2}\boldsymbol{\kappa}_i^{\mathrm{T}} S_i \boldsymbol{\kappa}_i - \boldsymbol{\kappa}_i^{\mathrm{T}} S_i \boldsymbol{\psi}_i + \mu\boldsymbol{\kappa}_i^{\mathrm{T}}\mathbf{y}_i - \log\sum_{c=1}^{C} \bar{g}_i^c e^{\mu\boldsymbol{\kappa}_i^{\mathrm{T}}\mathbf{e}^c}. \end{aligned}$$

Lemma 18 is used to obtain the second expression for $d_{\mathrm{KL}}$, and (71) to (73) are used to obtain the second expression for $d_h^i$. The $q_{*i}$ in the definition of $d_h^i$ refers to the the marginal for $\mathbf{f}_i$ under $q_*(\mathbf{f}|\mathbf{y})$, while the $\bar{g}_i^c$s in the term for $d_h^i$ is evaluated under $q(\mathbf{f}|\mathbf{y})$.

## D.2 Optimizing the Lower Bound on the Increase

Let

$$d_1(m_*, v_{**}, \mathbf{v}_*, \tilde{\mathbf{x}}_* | \tilde{X}) \stackrel{\mathrm{def}}{=} d_{\mathrm{KL}}(m_*, v_{**}, \mathbf{v}_*, \tilde{\mathbf{x}}_* | \tilde{X}) + \sum_{i=1}^{n} d_h^i(m_*, v_{**}, \mathbf{v}_*, \tilde{\mathbf{x}}_* | \tilde{X})$$

be the objective function within $d_1(\tilde{\mathbf{x}}_* | \tilde{X})$. Within this section, $d_1$ shall refer to this objective function instead of its maximum. The contributions from $m_*$ and $(\mathbf{v}_*, v_{**})$ are decoupled in this objective, so the search for the optimal $m_*$ and $(\mathbf{v}_*, v_{**})$ can be perform separately.

Instead of using $m_*$, $v_{**}$ and $\mathbf{v}_*$ as the variational parameters, we can treat $\mu$, $\nu$ and $\boldsymbol{\nu}$ as the variational parameters, and then define $m_{**}$, $v_{**}$ and $\mathbf{v}_*$ as functions of them:

$$m_* \stackrel{\mathrm{def}}{=} \mu + \mathbf{k}_*^{\mathrm{T}} K^{-1}\mathbf{m}, \qquad v_{**} \stackrel{\mathrm{def}}{=} \nu + \mathbf{v}_*^{\mathrm{T}} V^{-1} \mathbf{v}_*, \qquad \mathbf{v}_* \stackrel{\mathrm{def}}{=} \boldsymbol{\nu} + V K^{-1}\mathbf{k}_*.$$

This is valid and does not change the search space of the original variational parameters. During the optimization, the positive definiteness of $V_*$ (70) can be ensured by constraining the Schur complement $\nu$ to be positive (see Horn and Johnson 1985, Theorem 7.7.6). Under this re-parametrization, $d_1$ is concave in $\mu$ and $\nu$ but not necessarily concave in $\boldsymbol{\nu}$ because of the positive quadratic term within $\omega_i$ (73).

Below, we give the gradient updates for $\mu$ and $\nu$.

### D.2.1 Newton-Raphson Updates for $\mu$

Let $g^c_{*i} \overset{\text{def}}{=} \bar{g}^c_i \exp \mu \boldsymbol{\kappa}_i^{\mathrm{T}} \mathbf{e}^c$, $\bar{g}^c_{*i} \overset{\text{def}}{=} g^c_{*i} / \sum_{c'=1}^{C} g^{c'}_{*i}$, $\bar{\mathbf{g}}_{*i} \overset{\text{def}}{=} \left( \bar{g}^1_{*i}, \ldots, \bar{g}^C_{*i} \right)^{\mathrm{T}}$, $\bar{\mathbf{g}}_*$ be the stacking of $\bar{\mathbf{g}}_{*1}, \ldots, \bar{\mathbf{g}}_{*n}$, $\bar{G}_*$ be the diagonal matrix with $\bar{\mathbf{g}}_*$ down its diagonal, and $\tilde{G}_*$ be a $nC$-by-$nC$ block diagonal matrix where the $i$th block is $\bar{\mathbf{g}}_{*i} \bar{\mathbf{g}}_{*i}^{\mathrm{T}}$. The Newton-Raphson update for $\mu$ is obtained from the first and the second derivatives $\partial d_1 / \partial \mu = -\mu/\kappa + \boldsymbol{\kappa}^{\mathrm{T}} \mathbf{y} - \boldsymbol{\kappa}^{\mathrm{T}} \bar{\mathbf{g}}_*$ and $\partial^2 d_1 / \partial \mu^2 = -1/\kappa - \boldsymbol{\kappa}^{\mathrm{T}} (\bar{G}_* - \tilde{G}_*) \boldsymbol{\kappa}$.

### D.2.2 "Beyond" Newton Raphson Updates for $v_{**}$

We give an update for $\nu$ that converges faster than the Newton-Raphson update for $\log \nu$ when the optimal value is small, using a non-quadratic local approximation (Minka, 2002):

$$\tilde{d}_1(\nu) = \text{constant} + \frac{1}{2} \log \nu + \frac{n}{2} \log(\nu + a) - \frac{b}{2} \nu,$$

where $a$ and $b$ are parameters in the approximation. Within the approximation, $\nu$ is constrained to be positive due to the second term. By equating the first two derivatives of $d_1(\nu)$ to those of $\tilde{d}_1(\nu)$ at a given $\nu$, we obtain

$$a = \sqrt{\frac{n}{\sum_{i=1}^n \tau_i^2}} - \nu, \qquad\qquad b = \sqrt{n \textstyle\sum_{i=1}^n \tau_i^2} + \frac{1}{\kappa} + \boldsymbol{\kappa}^{\mathrm{T}} S \boldsymbol{\kappa} - \sum_{i=1}^n \tau_i,$$

where the positive branch of the square-root for $a$ is used so that $a + \nu$ remains positive. Fixing $a$ and $b$, the update for $\nu$ is obtained by equating the gradient of $\tilde{d}_1(\nu)$ at the updated point, say $\nu^{\mathrm{bNR}}$, to zero. This involve a quadratic equation, and we use its positive solution

$$\nu^{\mathrm{bNR}} = \frac{-(ab - n - 1) + \sqrt{(ab - n - 1)^2 + 4ab}}{2b}. \tag{74}$$

We prove that this update is guaranteed to be positive in Theorem 32 below.

**Lemma 31** $\tau_i \overset{\text{def}}{=} \boldsymbol{\kappa}_i^{\mathrm{T}} V_{\mathrm{f}i}^{-1} \boldsymbol{\kappa}_i / \omega_i < 1/\nu$.

**Proof** Define

$$\tilde{V}_* \overset{\text{def}}{=} \begin{pmatrix} V & \mathbf{v}_* \\ \mathbf{v}_*^{\mathrm{T}} & \mathbf{v}^{\mathrm{T}} V^{-1} \mathbf{v} \end{pmatrix},$$

which is positive semi-definite (Horn and Johnson, 1985, Theorem 7.7.6). Then $\tilde{V}_{\mathrm{f}*}$ below is positive definite since the covariance of the joint prior $p(\mathbf{z}_*, \mathbf{f})$ is positive definite.

$$\tilde{V}_{\mathrm{f}*} \overset{\text{def}}{=} K_{\mathrm{ff}} - K_{\mathrm{f}*}^{\mathrm{T}} K_*^{-1} K_{\mathrm{f}*} + K_{\mathrm{f}*}^{\mathrm{T}} K_*^{-1} \tilde{V}_* K_*^{-1} K_{\mathrm{f}*} = V_{\mathrm{f}} - (\kappa - (\chi - \nu)) \boldsymbol{\kappa} \boldsymbol{\kappa}^{\mathrm{T}} + \boldsymbol{\psi} \boldsymbol{\kappa}^{\mathrm{T}} + \boldsymbol{\kappa} \boldsymbol{\psi}^{\mathrm{T}}.$$

Similarly, the $i$th diagonal $C$-by-$C$ sub-matrix of $\tilde{V}_{\mathrm{f}*}$ given by

$$\tilde{V}_{\mathrm{f}*i} = V_{\mathrm{f}i} - (\kappa - (\chi - \nu)) \boldsymbol{\kappa}_i \boldsymbol{\kappa}_i^{\mathrm{T}} + \boldsymbol{\psi}_i \boldsymbol{\kappa}_i^{\mathrm{T}} + \boldsymbol{\kappa}_i \boldsymbol{\psi}_i^{\mathrm{T}}$$

is positive definite. Using Lemma 23, we obtain $|\tilde{V}_{\mathrm{f}*i}| = \tilde{\omega}_i |V_{\mathrm{f}i}|$, where

$$\tilde{\omega}_i \overset{\text{def}}{=} \left(1 + \boldsymbol{\kappa}_i^{\mathrm{T}} V_{\mathrm{f}i}^{-1} \boldsymbol{\psi}_i\right)^2 - \boldsymbol{\kappa}_i^{\mathrm{T}} V_{\mathrm{f}i}^{-1} \boldsymbol{\kappa}_i \left(\kappa - (\chi - \nu) + \boldsymbol{\psi}_i^{\mathrm{T}} V_{\mathrm{f}i}^{-1} \boldsymbol{\psi}_i\right)$$

is positive because both $|\tilde{V}_{\mathrm{f}*i}|$ and $|V_{\mathrm{f}i}|$ are positive. But $\tilde{\omega}_i = \omega_i - \nu \boldsymbol{\kappa}_i^{\mathrm{T}} V_{\mathrm{f}i}^{-1} \boldsymbol{\kappa}_i$. Thus

$$\omega_i = \tilde{\omega}_i + \nu \boldsymbol{\kappa}_i^{\mathrm{T}} V_{\mathrm{f}i}^{-1} \boldsymbol{\kappa}_i > \nu \boldsymbol{\kappa}_i^{\mathrm{T}} V_{\mathrm{f}i}^{-1} \boldsymbol{\kappa}_i.$$

So $\tau_i < 1/\nu$. ■

**Theorem 32** *Update $\nu^{bNR}$ given in Equation 74 is positive.*

**Proof** This update is guaranteed to be positive when $a$ and $b$ are both positive. Parameter $b$ is positive because $1/\kappa + \boldsymbol{\kappa}^{\mathrm{T}} S \boldsymbol{\kappa}$ is positive and $\sum_{i=1}^{n} \tau_i \le (n \sum_{i=1}^{n} \tau_i^2)^{1/2}$ by applying the Cauchy-Schwarz inequality. Parameter $a$ is positive because $\tau_i < 1/\nu$ from Lemma 31. ■

## Appendix E. Implementation Considerations

This appendix considers the details for an implementation of the variational bound optimization presented in this paper.

### E.1 Matrix Inversion in Update for b in Sparse Approximation

For the sparse approximation, the Newton-Raphson update for **b** given in Appendix C.2.1 requires inverting $X \overset{\text{def}}{=} I + (\bar{G} - \tilde{G}) \left( K_{\mathrm{f}}^{\mathrm{T}} K^{-1} K_{\mathrm{f}} + S^{-1} \right)$ of order $Cn$-by-$Cn$. To avoid $O(C^3 n^3)$ computation, we apply the Woodbury's inversion lemma thrice. Let $M \overset{\text{def}}{=} (S + \bar{G} - \tilde{G})^{-1}$, and $\tilde{L}_K \overset{\text{def}}{=} K_{\mathrm{f}}^{\mathrm{T}} L_K^{-\mathrm{T}}$, where $L_K$ is the lower Cholesky factor of $K$. Then

$$
\begin{aligned}
X^{-1} &= I - (\bar{G} - \tilde{G}) \left( \left( \tilde{L}_K \tilde{L}_K^{\mathrm{T}} + S^{-1} \right)^{-1} + (\bar{G} - \tilde{G}) \right)^{-1} \\
&= I - (\bar{G} - \tilde{G}) \left( M^{-1} - S \tilde{L}_K \left( I + \tilde{L}_K^{\mathrm{T}} S \tilde{L}_K \right)^{-1} \tilde{L}_K^{\mathrm{T}} S \right)^{-1} \\
&= I - (\bar{G} - \tilde{G}) \left( M + M S \tilde{L}_K \left( I + \tilde{L}_K^{\mathrm{T}} (S - SMS) \tilde{L}_K \right)^{-1} \tilde{L}_K^{\mathrm{T}} S M \right) \\
&= SM - (S - SMS) \tilde{L}_K \left( I + \tilde{L}_K^{\mathrm{T}} (S - SMS) \tilde{L}_K \right)^{-1} \tilde{L}_K^{\mathrm{T}} S M,
\end{aligned}
$$

where we have substituted $(\bar{G} - \tilde{G}) = M^{-1} - S$ to obtain the last expression.

### E.2 Better Conditioned Updates for $V$

In this section, we give better conditioned updates for the optimization of $V$.

#### E.2.1 NON-SPARSE CASE

Equation 28 in Section 3.4 gives the fixed-point update $V^{\mathrm{fx}} = (K^{-1} + W)^{-1}$ for the variational parameter $V$, where $W$ is rank deficient (see Lemma 28). We factorize $W = L_W L_W^{\mathrm{T}}$, and introduce $B \overset{\text{def}}{=} L_W^{\mathrm{T}} K L_W + I$ and $T \overset{\text{def}}{=} L_W B^{-1} L_W^{\mathrm{T}}$. Then the Woodbury's inversion lemma gives $V^{\mathrm{fx}} = K - KTK$. The optimal update is given by the best convex combination of $V$ and $V^{\mathrm{fx}}$. Let $T^{\mathrm{old}}$ be such that

$$
V = K - K T^{\mathrm{old}} K. \tag{75}
$$

The best convex combination is the one optimized over $\eta \in [0,1]$ in $V^{\mathrm{cc}} = K - K T^{\mathrm{cc}} K$, where $T^{\mathrm{cc}} \overset{\text{def}}{=} (1-\eta) T^{\mathrm{old}} + \eta T$. The update for $V^{\mathrm{cc}}$ implies that $T^{\mathrm{cc}}$ is the $T^{\mathrm{old}}$ for the next iteration, so (75) is always possible. Moreover, with $\Delta \overset{\text{def}}{=} \mathrm{d}V^{\mathrm{cc}}/\mathrm{d}\eta$, we also have $\Delta = K(T^{\mathrm{old}} - T)K$ and $K^{-1}\Delta = (T^{\mathrm{old}} - T)K$.

### E.2.2 SPARSE CASE

Equation 33 in Section 4.2 gives the fixed-point update in the sparse case:

$$V^{\text{fx}} = \left(K^{-1} + K^{-1}K_{\text{f}}W_{\text{f}}K_{\text{f}}^{\text{T}}K^{-1}\right)^{-1}.$$

If we were to proceed as for the non-sparse case using the Woodbury's inversion lemma, then the inversion of a $Cn$-by-$Cn$ matrix would be required. However, this is to be avoided in the sparse approximation, which aims to reduce time complexity. Instead, we compute $V^{\text{fx}} = L_K(I + \tilde{L}_K^{\text{T}}W_{\text{f}}\tilde{L}_K)^{-1}L_K^{\text{T}}$, where $L_K$ is the lower Cholesky factor of $K$, and $\tilde{L}_K \overset{\text{def}}{=} K_{\text{f}}^{\text{T}}L_K^{-\text{T}}$. This is more efficient and yet does not involve any inversion of $K$.

The computation of $V_{\text{f}}$ at this fixed point requires $T \overset{\text{def}}{=} K^{-1} - K^{-1}V^{\text{fx}}K^{-1}$. This can be done with the above formula for $V^{\text{fx}}$:

$$T = K^{-1} - L_K^{-\text{T}}\left(I + \tilde{L}_K^{\text{T}}W_{\text{f}}\tilde{L}_K\right)^{-1}L_K^{-1} = L_K^{-\text{T}}\left(I - \left(I + \tilde{L}_K^{\text{T}}W_{\text{f}}\tilde{L}_K\right)^{-1}\right)L_K^{-1}.$$

Hence

$$V_{\text{f}}^{\text{fx}} = K_{\text{ff}} - K_{\text{f}}^{\text{T}}TK_{\text{f}} = K_{\text{ff}} - \tilde{L}_K\left(I - \left(I + \tilde{L}_K^{\text{T}}W_{\text{f}}\tilde{L}_K\right)^{-1}\right)\tilde{L}_K^{\text{T}}.$$

## E.3 Initialization

Our variational lower bound (21) on the marginal likelihood is concave with respect to all the variational parameters, so the initialization of parameters does not affect the converged answer in theory. However, in practice, initialization is still important for two reasons. First, it can ensure that the matrices are better conditioned. Second, it can ensure that we start near to the converged answer, so that convergence is sooner.

For initialization, there are two cases to be considered. The easier case is during model learning when we can use the optimized variational parameters from the previous model to initialize the variational parameters of the current model. We shall omit details for this case. The more difficult case is when there is no previous model, usually when no model learning is involved or at the onset of model learning. In this section, we suggest a procedure for initialization in this case. The key idea behind our procedure is to locate the variational mean at the data and to use the same covariance at every input $\mathbf{x}_i$.

### E.3.1 COVARIANCES

From Equation 20 for the analysis of the proof of Theorem 6, a parametrization of $W_i$ that satisfies the two mentioned properties is $W_i \overset{\text{def}}{=} M - M\mathbf{1}\mathbf{1}^{\text{T}}M/\mathbf{1}^{\text{T}}M\mathbf{1}$, where $M$ is a $C$-by-$C$ positive definite matrix. Although we have noted there that using a diagonal $M$ is suboptimal, there is much appeal in such a setting for initialization because of the match with the likelihood terms. Hence we shall initialize with $W_i \overset{\text{def}}{=} \gamma\left(I/C - \mathbf{1}^{\text{T}}\mathbf{1}/C^2\right)$, for some $\gamma > 0$. The initial covariance $V$ of the variational posterior can be computed using Woodbury's inversion lemma on the fixed point equation $(K^{-1} + W)^{-1}$, where $W$ is the block diagonal matrix consisting or the $W_i$s.

### E.3.2 MEANS

Our initialization for the mean locate it at the data. To this end, let us recall a few invariances at the stationary point of the variational lower bound (21) on the marginal likelihood. For the $i$th datum,

$\mathbf{a}_i$ and $W_i$ are the parameters for the variational posterior $r(\mathbf{f}_i|\mathbf{y}_i)$ defined in Lemma 2. For the case of non-sparse approximation, we have the invariances

$$\begin{aligned} \mathbf{y}-\mathbf{b} &= W(\mathbf{a}-\mathbf{m}), && \text{(From definition in Lemma 5)} \\ \mathbf{m} &= K(\mathbf{y}-\mathbf{b}), && \text{(Section 3.3)} \\ V &= (K^{-1}+W)^{-1}, && \text{(Section 3.4)} \end{aligned}$$

where $\mathbf{y}$ (resp. $\mathbf{b}$, $\mathbf{a}$, $\mathbf{m}$) is the stacking of the $\mathbf{y}_i$s (resp. $\mathbf{b}_i$s, $\mathbf{a}_i$s, $\mathbf{m}_i$s) for each datum. Rearranging for $\mathbf{m}$ gives

$$\mathbf{m} = (I+KW)^{-1}KW\mathbf{a} = (K^{-1}+W)^{-1}W\mathbf{a} = VW\mathbf{a}. \tag{76}$$

We initialize $\mathbf{m}$ through an appropriate value for $\mathbf{a}$. Since $\mathbf{a}_i$ is the mean of $r(\mathbf{f}_i|\mathbf{y}_i)$, we choose to set $\mathbf{a}_i = \gamma(\mathbf{y}_i + (\mathbf{y}_i - \mathbf{1})/(C-1))$, for some fixed parameter $\gamma$. For example, if $\mathbf{x}_i$ is in the first class, then $\mathbf{a}_i = (\gamma, -\gamma/(C-1), \ldots, -\gamma/(C-1))^{\mathrm{T}}$. This locates the mean of $r(\mathbf{f}_i|\mathbf{y}_i)$ to be positive for the class given by the data and uniformly negative otherwise. Let $\boldsymbol{\alpha} \overset{\text{def}}{=} K^{-1}\mathbf{m}$. The initialization (76) satisfies the sum-to-zero property:

$$\mathbf{1}^{\mathrm{T}}\boldsymbol{\alpha} = \mathbf{1}^{\mathrm{T}}K^{-1}VW\mathbf{a} = \mathbf{1}^{\mathrm{T}}(I+WK)^{-1}W\mathbf{a} = \mathbf{1}^{\mathrm{T}}W(I+KW)^{-1}\mathbf{a} = \mathbf{0},$$

where the third equality applies Searle's Identity, and the last equality is because $\mathbf{1}$ is in the null-space of $W$. With $\gamma = 1$, the setting for $\mathbf{a}_i$ is the minimizer of the loss function in a multi-class SVM under the sum-to-zero constraint (Lee et al., 2004, Lemma 1).

Similarly, for sparse approximation, we have

$$\begin{aligned} \mathbf{y}-\mathbf{b} &= W_{\mathrm{f}}(\mathbf{a}-\mathbf{m}_{\mathrm{f}}), && \text{(From definition in Lemma 5)} \\ \mathbf{m}_{\mathrm{f}} &= K_{\mathrm{f}}^{\mathrm{T}}K^{-1}\mathbf{m}, && \text{(Section 4, Equation 31)} \\ \mathbf{m} &= K_{\mathrm{f}}(\mathbf{y}-\mathbf{b}), && \text{(Section 4.2)} \\ V &= (K^{-1}+K^{-1}K_{\mathrm{f}}W_{\mathrm{f}}K_{\mathrm{f}}^{\mathrm{T}}K^{-1})^{-1}. && \text{(Section 4.2)} \end{aligned}$$

Rearranging for $\mathbf{m}$ gives

$$\mathbf{m} + K_{\mathrm{f}}W_{\mathrm{f}}K_{\mathrm{f}}^{\mathrm{T}}K^{-1}\mathbf{m} = K_{\mathrm{f}}W_{\mathrm{f}}\mathbf{a} \quad\Longleftrightarrow\quad KV^{-1}\mathbf{m} = K_{\mathrm{f}}W_{\mathrm{f}}\mathbf{a} \quad\Longleftrightarrow\quad \mathbf{m} = VK^{-1}K_{\mathrm{f}}W_{\mathrm{f}}\mathbf{a}.$$

Initialization of $\mathbf{a}$ is done as in the non-sparse case.

## Appendix F. Importance Sampling

In this section, we describe how various quantities of interest can be computed using importance sampling. Let $p(\mathbf{f}|\mathbf{y})$ be the exact posterior of the latent function values at the observed data. This is obtained from Bayes' rule $p(\mathbf{f}|\mathbf{y}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f})/p(\mathbf{y})$, where $p(\mathbf{y})$ is the marginal likelihood, which is intractable to compute exactly. Let $p_s(\mathbf{f})$ be a proposal distribution. Our choice of $p_s(\mathbf{f})$ is the multivariate-$t$ distribution (Kotz and Nadarajah, 2004) with four degrees of freedom, centered at the that mean of the optimized variational approximation to $p(\mathbf{f}|\mathbf{y})$ and with covariance twice the covariance of the prior $p(\mathbf{f})$; that is

$$p_s(\mathbf{f}) = \frac{\Gamma((\nu+p)/2)}{((\pi\nu)^{p/2}\Gamma(\nu/2)|K|^{1/2}}\left[1+\frac{1}{\nu}(\mathbf{f}-\mathbf{m}^*)^{\mathrm{T}}K^{-1}(\mathbf{f}-\mathbf{m}^*)\right]^{(\nu+p)/2},$$

where $\nu = 4$, $p = nC$ is the dimension of $\mathbf{f}$, $K$ is the prior covariance of $\mathbf{f}$, and $\mathbf{m}^*$ is the mean of the optimized variational posterior. This choice of proposal ensures that $p(\mathbf{f}) \leq c\, p_s(\mathbf{f})$ for all $\mathbf{f}$ for some finite constant $c > 0$, which is a desideratum for importance samplers. It also locates the proposal at the estimated mean of the posterior.

Let $\mathbf{f}^{(s)}$ be a sample from the proposal, indexed by $s$ over $n_s$ samples. Its unnormalized weight $w^{(s)}$ is

$$w^{(s)} \overset{\text{def}}{=} \frac{p(\mathbf{y})p(\mathbf{f}^{(s)}|\mathbf{y})}{p_s(\mathbf{f}^{(s)})} = \frac{p(\mathbf{y}|\mathbf{f}^{(s)})p(\mathbf{f}^{(s)})}{p_s(\mathbf{f}^{(s)})},$$

which can be computed exactly for the multinomial logistic likelihood. A Monte Carlo estimate of $p(\mathbf{y})$ is $\hat{p}(\mathbf{y}) \overset{\text{def}}{=} \sum_s w^{(s)}/n_s$, which is the sample mean of the $w^{(s)}$s, because

$$p(\mathbf{y}) \overset{\text{def}}{=} \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\mathrm{d}\mathbf{f} = \int \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{p_s(\mathbf{f})} p_s(\mathbf{f})\mathrm{d}\mathbf{f} \approx \frac{1}{n_s}\sum_s w^{(s)}.$$

The strong law of large numbers says that $\hat{p}(\mathbf{y})$ converges to $p(\mathbf{y})$ almost surely as $n_s$ approaches infinity (Geweke, 2005, Theorem 4.2.2). The rate of convergence is given by the Lindeberg-Lévy central limit theorem (Geweke, 2005, Theorem 4.2.2)

$$\sqrt{n_s}\,(p(\mathbf{y}) - \hat{p}(\mathbf{y})) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2),$$

where $\sigma^2$ is the true variance of unnormalized weights. This variance exists for our choice of the proposal distribution because $p(\mathbf{f}) \leq c\, p_s(\mathbf{f})$ and the likelihood is bounded. This variance can be estimated from the samples $w^{(s)}$s. We use this convergence in distribution to compute a high probability upper bound to $p(\mathbf{y})$ based on the samples. Since, the weights and $p(\mathbf{y})$ are positive, one might be concerned that skewness has not been factored into the approximation. Then, one might consider using the $\chi^2$ approximation (Hall, 1983). However, our calculations have shown this to have negligible effect on the upper bound estimate because we have used $n_s = 100,000$ samples.

### F.1 Prediction

The normalized weight $\tilde{w}^{(s)}$ of $\mathbf{f}^{(s)}$ is estimated with

$$\tilde{w}^{(s)} \overset{\text{def}}{=} \frac{1}{n_s}\frac{p(\mathbf{f}^{(s)}|\mathbf{y})}{p_s(\mathbf{f}^{(s)})} = \frac{1}{n_s}\frac{w^{(s)}}{p(\mathbf{y})} \approx \frac{w^{(s)}}{\sum_{s'} w^{(s')}}.$$

For prediction at $\mathbf{x}_*$, the exact joint posterior of $(\mathbf{f}, \mathbf{f}_*)$ is $p(\mathbf{f}, \mathbf{f}_*|\mathbf{y}) = p(\mathbf{f}|\mathbf{y})p(\mathbf{f}_*|\mathbf{f})$. For the proposal distribution, we use $p_s(\mathbf{f}, \mathbf{f}_*) = p_s(\mathbf{f})p(\mathbf{f}_*|\mathbf{f})$, and a draw from the proposal follows this generative model. The normalized weight of sample $(\mathbf{f}, \mathbf{f}_*)^{(s)} \overset{\text{def}}{=} (\mathbf{f}^{(s)}, \mathbf{f}_*^{(s)})$ is

$$\tilde{w}_*^{(s)} \overset{\text{def}}{=} \frac{1}{n_s}\frac{p(\mathbf{f}^{(s)}|\mathbf{y})p(\mathbf{f}_*^{(s)}|\mathbf{f}^{(s)})}{p_s(\mathbf{f}^{(s)})p(\mathbf{f}_*^{(s)}|\mathbf{f}^{(s)})} = \tilde{w}^{(s)}.$$

The predictive probability is

$$p(\mathbf{y}_*|\mathbf{y}) = \int p(\mathbf{y}_*|\mathbf{f}_*)\, p(\mathbf{f}, \mathbf{f}_*|\mathbf{y})\, \mathrm{d}\mathbf{f}\,\mathrm{d}\mathbf{f}_* = \int p(\mathbf{y}_*|\mathbf{f}_*)\frac{p(\mathbf{f}, \mathbf{f}_*|\mathbf{y})}{p_s(\mathbf{f}, \mathbf{f}_*)} p_s(\mathbf{f}, \mathbf{f}_*)\, \mathrm{d}\mathbf{f}\,\mathrm{d}\mathbf{f}_* \approx \sum_s \tilde{w}_*^{(s)} p(\mathbf{y}_*|\mathbf{f}_*^{(s)}).$$

For the multinomial logistic likelihood, $p(\mathbf{y}_i|\mathbf{f}_i)$ and $p(\mathbf{y}_*|\mathbf{f}_*)$ can be computed readily. For the multinomial probit likelihood, we use the sampling approach (Girolami and Rogers, 2006) with twenty samples, which is sufficient when $n_s$ is large.

## References

T. Ando. Concavity of certain maps on positive definite matrices and applications to Hadamard products. *Linear Algebra and Its Applications*, 26:203–241, 1979.

D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2:343–370, 1988.

A. Banerjee. On Bayesian bounds. In *International Conference on Machine Learning*, 2006.

E. V. Bonilla, K. M. A. Chai, and C. K. I. Williams. Multi-task Gaussian process prediction. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20, pages 153–160. Curran Associates, Inc., 2008.

E. J. Bredensteiner and K. P. Bennett. Multicategory classification by support vector machines. *Computational Optimization Applications*, 12:53–79, January 1999.

J. S. Bridle. Probabilistic interpretation of feedforward classification network outputs with relationships to statistical pattern recognition. In *Neuro-computing: Algorithms, Architectures and Applications*, NATO ASI Series in Systems and Computer Science. Springer, 1989.

E. Challis and D. Barber. Concave Gaussian variational approximations for inference in large-scale Bayesian linear models. In D. Dunson and M. Dudk, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *JMLR: Workshop and Conference Proceedings Series*, 2011.

K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

L. Csató and M. Opper. Sparse online Gaussian processes. *Neural Computation*, 14:641–668, 2002.

A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL `http://archive.ics.uci.edu/ml`.

J. Geweke. *Contemporary Bayesian Econometrics and Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., New Jersey, 2005.

J. Geweke, M. P. Keane, and D. Runkle. Alternative computational approaches to inference in the multinomial probit model. *The Review of Economics and Statistics*, 76(4):609–32, 1994.

Z. Ghahramani and M. J. Beal. Graphical models and variational methods. In D. Saad and M. Opper, editors, *Advanced Mean Field methods — Theory and Practice*. MIT Press, 2000a.

Z. Ghahramani and M. J. Beal. Variational inference for Bayesian mixtures of factor analysers. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 449–455. MIT Press, Cambridge, MA, 2000b.

M. N. Gibbs. *Bayesian Gaussian Processes for Regression and Classification*. PhD thesis, University of Cambridge, Department of Physics, 1997.

M. Girolami and S. Rogers. Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Computation*, 18(8):1790–1817, 2006.

Y. Guermeur. Combining discriminant models with new multi-class SVMs. *Pattern Analysis and Applications*, 5(2):168–179, 2002.

Y. Guermeur and E. Monfrini. A quadratic loss multi-class SVM for which a radius-margin bound applies. *Informatica*, 22(1):73–96, 2011.

P. Hall. Chi squared approximations to the distribution of a sum of independent random variables. *The Annals of Probability*, 11(4):1028–1036, 1983.

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

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

H.-C. Kim and Z. Ghahramani. Bayesian Gaussian process classification with the EM-EP algorithm. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 28:1948–1959, 2006.

D. A. Knowles and T. P. Minka. Non-conjugate variational message passing for multinomial and binary regression. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24, pages 1701–1709. 2011.

S. J. Koopman, N. Shephard, and D. Creal. Testing the assumptions behind importance sampling. *Journal of Econometrics*, 149(1):2–11, 2009.

S. Kotz and S. Nadarajah. *Multivariate t Distributions and Their Applications*. Cambridge University Press, Cambridge, UK, 2004.

F. Lauer and Y. Guermeur. MSVMpack: a multi-class support vector machine package. *Journal of Machine Learning Research*, 12:2269–2272, 2011. `http://www.loria.fr/~lauer/MSVMpack`.

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

M. Lázaro-Gredilla and A. Figueiras-Vidal. Inter-domain Gaussian processes for sparse inference using inducing features. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22, pages 1087–1095. Curran Associates, Inc., 2009.

Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

D. McFadden. Conditional logit analysis of qualitative choice behavior. In P. Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, 1974.

T. P. Minka. Beyond Newton's method, 2002. URL `http://research.microsoft.com/en-us/um/people/minka/papers/minka-newton.pdf`.

R. M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, 1996.

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

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

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

J. E. Potter. Matrix quadratic solutions. *SIAM Journal on Applied Mathematics*, 14(3):496–501, May 1966.

S. Puntanen and G. P. H. Styan. Historical introduction: Issai Schur and the early development of the Schur complement. In F. Zhang, editor, *The Schur Complement and Its Applications*, Numerical Methods and Algorithms, pages 1–16. Springer, 2005.

J. Quiñonero-Candela, C. E. Rasmussen, and C. K. I. Williams. Approximation methods for Gaussian process regression. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, editors, *Large Scale Kernel Machines*, pages 203–223. MIT Press, 2007.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Massachusetts, 2006.

R. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

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

M. Seeger and M. I. Jordan. Sparse Gaussian process classification with multiple classes. Technical report, University of California at Berkeley, Department of Statistics, 2004.

E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18, pages 1257–1264, Cambridge, MA, 2006. MIT Press.

M. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *JMLR: Workshop and Conference Proceedings Series*, pages 567–574, 2009a.

M. Titsias. Variational model selection for sparse Gaussian process regression. Technical report, University of Manchester, School of Computer Science, 2009b. URL `http://www.cs.man.ac.uk/~mtitsias/papers/sparseGPv2.pdf`.

V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.

J. M. Ver Hoef and R. P. Barry. Constructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference*, 69(2):275–294, 1998.

G. S. Watson. Spectral decomposition of the covariance matrix of a multinomial. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1):289–291, 1996.

J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *Proceedings of the 7th European Symposium on Artificial Neural Network*, 1999.

C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In T. K. Leen, T. G. Diettrich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13, pages 682–688, Cambridge, MA, 2001. MIT Press.

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