Journal of Machine Learning Research 9 (2008) 2035-2078 Submitted 8/07; Revised 4/08; Published 10/08

Approximations for Binary Gaussian Process Classification

Hannes Nickisch HN@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics
Spemannstraße 38
72076 Tübingen, Germany

Carl Edward Rasmussen* CER54@CAM.AC.UK
Department of Engineering
University of Cambridge
Trumpington Street
Cambridge, CB2 1PZ, UK

Editor: Carlos Guestrin

Abstract

We provide a comprehensive overview of many recent algorithms for approximate inference in Gaussian process models for probabilistic binary classification. The relationships between several approaches are elucidated theoretically, and the properties of the different algorithms are corroborated by experimental results. We examine both 1) the quality of the predictive distributions and 2) the suitability of the different marginal likelihood approximations for model selection (selecting hyperparameters) and compare to a gold standard based on MCMC. Interestingly, some methods produce good predictive distributions although their marginal likelihood approximations are poor. Strong conclusions are drawn about the methods: The Expectation Propagation algorithm is almost always the method of choice unless the computational budget is very tight. We also extend existing methods in various ways, and provide unifying code implementing all approaches.

Keywords: Gaussian process priors, probabilistic classification, Laplaces's approximation, expectation propagation, variational bounding, mean field methods, marginal likelihood evidence, MCMC

1. Introduction

Gaussian processes (GPs) can conveniently be used to specify prior distributions for Bayesian inference. In the case of regression with Gaussian noise, inference can be done simply in closed form, since the posterior is also a GP. For non-Gaussian likelihoods, such as e.g., in binary classification, exact inference is analytically intractable.

One prolific line of attack is based on approximating the non-Gaussian posterior with a tractable Gaussian distribution. One might think that finding such an approximating GP is a well-defined problem with a largely unique solution. However, we find no less than three different types of solution in the recent literature: Laplace Approximation (LA) (Williams and Barber, 1998), Expectation Propagation (EP) (Minka, 2001a) and Kullback-Leibler divergence (KL) minimization (Opper and Archambeau, 2008) comprising Variational Bounding (VB) (Gibbs and MacKay, 2000) as a special

*. Also at Max Planck Institute for Biological Cybernetics, Spemannstraße 38, 72076 Tübingen, Germany.

case. Another approach is based on a factorial approximation, rather than a Gaussian (Csató et al., 2000).

Practical applications reflect the richness of approximate inference methods: LA has been used for sequence annotation (Altun et al., 2004) and prostate cancer prediction (Chu et al., 2005), EP for affect recognition (Kapoor and Picard, 2005), VB for weld cracking prognosis (Gibbs and MacKay, 2000), Label Regression (LR) serves for object categorization (Kapoor et al., 2007) and MCMC sampling is applied to rheuma diagnosis (Schwaighofer et al., 2002). Brain computer interfaces (Zhong et al., 2008) even rely on several (LA, EP, VB) methods.

In this paper, we compare these different approximations and provide insights into the strengths and weaknesses of each method, extending the work of Kuss and Rasmussen (2005) in several directions: We cover many more approximation methods (VB,KL,FV,LR), put all of them in common framework and provide generic implementations dealing with both the logistic and the cumulative Gaussian likelihood functions and clarify the aspects of the problem causing difficulties for each method. We derive Newton's method for KL and VB. We show how to accelerate MCMC simulations. We highlight numerical problems, comment on computational complexity and supply runtime measurements based on experiments under a wide range of conditions, including different likelihood and different covariance functions. We provide deeper insights into the methods behavior by systematically linking them to each other. Finally, we review the tight connections to methods from the literature on Statistical Physics, including the TAP approximation and TAPnaive.

The quantities of central importance are the quality of the probabilistic predictions and the suitability of the approximate marginal likelihood for selecting parameters of the covariance function (hyperparameters). The marginal likelihood for any Gaussian approximate posterior can be lower bounded using Jensen's inequality, but the specific approximation schemes also come with their own marginal likelihood approximations.

We are able to draw clear conclusions. Whereas every method has good performance under some circumstances, only a single method gives consistently good results. We are able to theoretically corroborate our experimental findings; together this provides solid evidence and guidelines for choosing an approximation method in practice.

2. Gaussian Processes for Binary Classification

We describe probabilistic binary classification based on Gaussian processes in this section. For a graphical model representation see Figure 1 and for a 1d pictorial description consult Figure 2. Given data points $\mathbf{x}_i$ from a domain $\mathcal{X}$ with corresponding class labels $y_i \in \{-1,+1\}$, one would like to predict the class membership probability for a test point $\mathbf{x}_*$. This is achieved by using a *latent function* f whose value is mapped into the unit interval by means of a sigmoid function $\mathrm{sig}:\mathbb{R}\to[0,1]$ such that the class membership probability $\mathbb{P}(y=+1|\mathbf{x})$ can be written as $\mathrm{sig}(f(\mathbf{x}))$. The class membership probability must normalize $\sum_y \mathbb{P}(y|\mathbf{x})=1$, which leads to $\mathbb{P}(y=+1|\mathbf{x})=1-\mathbb{P}(y=-1|\mathbf{x})$. If the sigmoid function satisfies the point symmetry condition $\mathrm{sig}(t)=1-\mathrm{sig}(-t)$, the *likelihood* can be compactly written as

$$\mathbb{P}(y|\mathbf{x}) \quad = \quad \mathrm{sig}(y\cdot f(\mathbf{x}))\,.$$

In this paper, two point symmetric sigmoids are considered

$$\begin{aligned} \mathrm{sig}_{\mathrm{logit}}(t) &:= \frac{1}{1+e^{-t}} \\ \mathrm{sig}_{\mathrm{probit}}(t) &:= \int_{-\infty}^{t} \mathcal{N}(\tau|0,1)\mathrm{d}\tau. \end{aligned}$$

The two functions are very similar at the origin (showing locally linear behavior around $\mathrm{sig}(0) = 1/2$ with slope $1/4$ for $\mathrm{sig}_{\mathrm{logit}}$ and $1/\sqrt{2\pi}$ for $\mathrm{sig}_{\mathrm{probit}}$) but differ in how fast they approach $0/1$ when t goes to infinity. For large negative values of t, we have the asymptotics

$$\mathrm{sig}_{\mathrm{logit}}(t) \approx \exp(-t) \quad \text{and} \quad \mathrm{sig}_{\mathrm{probit}}(t) \approx \exp(-\frac{1}{2}t^2 + 0.158t - 1.78), \text{ for } t \ll 0.$$

Linear decay of $\ln(\mathrm{sig}_{\mathrm{logit}})$ corresponds to a weaker penalty for wrongly classified examples than the quadratic decay of $\ln(\mathrm{sig}_{\mathrm{probit}})$.

For notational convenience, the following shorthands are used: The matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ of size $n \times d$ collects the training points, the vector $\mathbf{y} = [y_1, \ldots, y_n]^\top$ of size $n \times 1$ collects the target values and latent function values are summarized by $\mathbf{f} = [f_1, \ldots, f_n]^\top$ with $f_i = f(\mathbf{x}_i)$. Observed data is written as $\mathcal{D} = \{(\mathbf{x}_i, y_i) \,|\, i = 1, \ldots, n\} = (\mathbf{X}, \mathbf{y})$. Quantities carrying an asterisk refer to test points, that is, $\mathbf{f}_*$ contains the latent function values for test points $[\mathbf{x}_{*,1}, \ldots, \mathbf{x}_{*,m}] = \mathbf{X}_* \subset \mathcal{X}$. Covariances between latent values $\mathbf{f}$ and $\mathbf{f}_*$ at data points $\mathbf{x}$ and $\mathbf{x}_*$ follow the same notation, namely $[\mathbf{K}_{**}]_{ij} = k(\mathbf{x}_{*,i}, \mathbf{x}_{*,j})$, $[\mathbf{K}_*]_{ij} = k(\mathbf{x}_i, \mathbf{x}_{*,j})$, $[\mathbf{k}_*]_i = k(\mathbf{x}_i, \mathbf{x}_*)$ and $k_{**} = k(x_*, x_*)$, where $[\mathbf{A}]_{ij}$ denotes the entry A_{ij} of the matrix $\mathbf{A}$.

Given the latent function f, the class labels are assumed to be Bernoulli distributed and independent random variables, which gives rise to a *factorial likelihood*, factorizing over data points (see Figure 1)

$$\mathbb{P}(\mathbf{y}|f) \;=\; \mathbb{P}(\mathbf{y}|\mathbf{f}) \;=\; \prod_{i=1}^{n} \mathbb{P}(y_i|f_i) \;=\; \prod_{i=1}^{n} \mathrm{sig}(y_i f_i). \tag{1}$$

A GP (Rasmussen and Williams, 2006) is a stochastic process fully specified by a *mean function* $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ and a positive definite *covariance function* $k(\mathbf{x}, \mathbf{x}') = \mathbb{V}[f(\mathbf{x}), f(\mathbf{x}')]$. This means that a random variable $f(\mathbf{x})$ is associated to every $\mathbf{x} \in \mathcal{X}$, such that for any set of inputs $\mathbf{X} \subset \mathcal{X}$, the joint distribution $\mathbb{P}(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\mathbf{m}_0, \mathbf{K})$ is Gaussian with mean vector $\mathbf{m}_0$ and covariance matrix $\mathbf{K}$. The mean function and covariance functions may depend on additional *hyperparameters* $\boldsymbol{\theta}$. For notational convenience we will assume $m(x) \equiv 0$ throughout. Thus, the elements of $\mathbf{K}$ are $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j, \boldsymbol{\theta})$.

By application of Bayes' rule, one gets an expression for the *posterior* distribution over the latent values $\mathbf{f}$

$$\mathbb{P}(\mathbf{f}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}) \;=\; \frac{\mathbb{P}(\mathbf{y}|\mathbf{f})\,\mathbb{P}(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})}{\int \mathbb{P}(\mathbf{y}|\mathbf{f})\,\mathbb{P}(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})\,\mathrm{d}\mathbf{f}} \;=\; \frac{\mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K})}{\mathbb{P}(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})} \prod_{i=1}^{n} \mathrm{sig}(y_i f_i), \tag{2}$$

where $Z = \mathbb{P}(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \int \mathbb{P}(\mathbf{y}|\mathbf{f})\,\mathbb{P}(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})\,\mathrm{d}\mathbf{f}$ denotes the *marginal likelihood* or *evidence* for the hyperparameter $\boldsymbol{\theta}$. The joint prior over training and test latent values $\mathbf{f}$ and $\mathbf{f}_*$ given the corresponding inputs is

$$\mathbb{P}(\mathbf{f}_*, \mathbf{f}|\mathbf{X}_*, \mathbf{X}, \boldsymbol{\theta}) \;=\; \mathcal{N}\left(\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \middle| \mathbf{0}, \begin{bmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^\top & \mathbf{K}_{**} \end{bmatrix} \right).$$

When making predictions, we marginalize over the training set latent variables

$$\mathbb{P}(\mathbf{f}_*|\mathbf{X}_*, \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}) = \int \mathbb{P}(\mathbf{f}_*, \mathbf{f}|\mathbf{X}_*, \mathbf{y}, \mathbf{X}, \boldsymbol{\theta})\, \mathrm{d}\mathbf{f} = \int \mathbb{P}(\mathbf{f}_*|\mathbf{f}, \mathbf{X}_*, \mathbf{X}, \boldsymbol{\theta})\, \mathbb{P}(\mathbf{f}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta})\, \mathrm{d}\mathbf{f}, \tag{3}$$

where the joint posterior is factored into the product of the posterior and the conditional prior

$$\mathbb{P}(\mathbf{f}_*|\mathbf{f}, \mathbf{X}_*, \mathbf{X}, \boldsymbol{\theta}) \;=\; \mathcal{N}\left(\mathbf{f}_*|\mathbf{K}_*^\top \mathbf{K}^{-1}\mathbf{f}, \mathbf{K}_{**} - \mathbf{K}_*^\top \mathbf{K}^{-1}\mathbf{K}_*\right).$$

Finally, the predictive class membership probability $p_* := \mathbb{P}(y_* = 1|\mathbf{x}_*, \mathbf{y}, \mathbf{X}, \boldsymbol{\theta})$ is obtained by averaging out the test set latent variables

$$\mathbb{P}(y_*|\mathbf{x}_*, \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}) = \int \mathbb{P}(y_*|f_*)\, \mathbb{P}(f_*|\mathbf{x}_*, \mathbf{y}, \mathbf{X}, \boldsymbol{\theta})\, df_* \;=\; \int \mathrm{sig}(y_* f_*)\, \mathbb{P}(f_*|\mathbf{x}_*, \mathbf{y}, \mathbf{X}, \boldsymbol{\theta})\, \mathrm{d}f_*. \tag{4}$$

The integral is analytically tractable for $\mathrm{sig}_{\mathrm{probit}}$ (Rasmussen and Williams, 2006, Ch. 3.9) and can be efficiently approximated for $\mathrm{sig}_{\mathrm{logit}}$ (Williams and Barber, 1998, App. A).

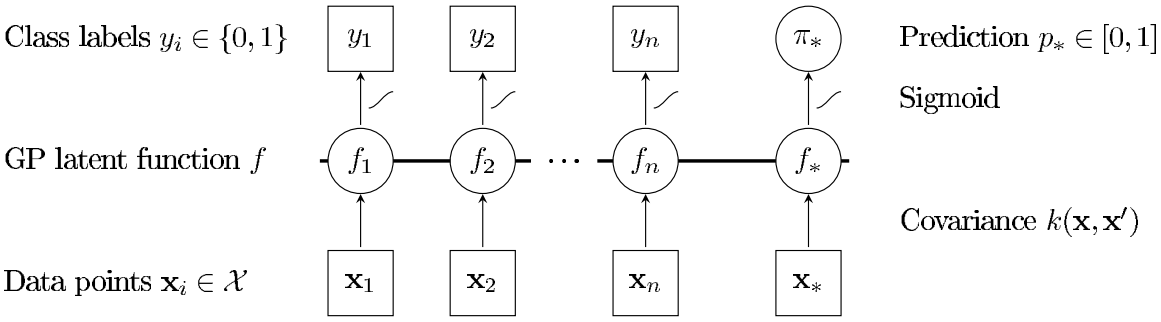


Figure 1: Graphical Model for binary Gaussian process classification: Circles represent unknown quantities, squares refer to observed variables. The horizontal thick line means fully connected latent variables. An observed label y_i is conditionally independent of all other nodes given the corresponding latent variable f_i. Labels y_i and latent function values f_i are connected through the sigmoid likelihood; all latent function values f_i are fully connected, since they are drawn from the same GP. The labels y_i are binary, whereas the prediction p_* is a probability and can thus have values from the whole interval $[0,1]$.

2.1 Stationary Covariance Functions

In preparation for the analysis of the approximation schemes described in this paper, we investigate some simple properties of the posterior for stationary covariance functions in different regimes

encountered in classification. Stationary covariances of the form $k(\mathbf{x},\mathbf{x}',\boldsymbol{\theta}) = \sigma_f^2 g(|\mathbf{x}-\mathbf{x}'|/\ell)$ with $g:\mathbb{R}\to\mathbb{R}$ a monotonously decreasing function[1] and $\boldsymbol{\theta} = \{\sigma_f, \ell\}$ are widely used. The following section supplies a geometric intuition of that specific prior in the classification scenario by analyzing the limiting behavior of the covariance matrix $\mathbf{K}$ as a function of the length scale ℓ and the limiting behavior of the likelihood as a function of the latent function scale σ_f. A pictorial illustration of the setting is given in Figure 3.

2.1.1 LENGTH SCALE

Two limiting cases of "ignorance with respect to the data" with marginal likelihood $Z = 2^{-n}$ can be distinguished, where $\mathbb{1} = [1,\dots 1]^\top$ and $\mathbf{I}$ is the identity matrix (see Appendix B.1)

$$\begin{aligned}\lim_{\ell\to 0}\mathbf{K} &= \sigma_f^2\mathbf{I},\\ \lim_{\ell\to\infty}\mathbf{K} &= \sigma_f^2\mathbb{1}\mathbb{1}^\top.\end{aligned}$$

For very small length scales ($\ell\to 0$), the prior is simply isotropic as all points are deemed to be far away from each other and the whole model factorizes. Thus, the (identical) posterior moments can be calculated dimension-wise. (See Figure 3, regimes 1, 4 and 7.)

For very long length scales ($\ell\to\infty$), the prior becomes degenerate as all datapoints are deemed to be close to each other and takes the form of a cigar along the hyper-diagonal. (See Figure 3, regimes 3, 6 and 9.) A 1d example of functions drawn from GP priors with different lengthscales ℓ is shown in Figure 2 on the left. The lengthscale has to be suited to the data; if chosen too small, we will overfit, if chosen too high underfitting will occur.

2.1.2 LATENT FUNCTION SCALE

The sigmoid likelihood function $\mathrm{sig}(y_i f_i)$ measures the agreement of the signs of the latent function and the label in a smooth way, that is, values close to one if the signs of y_i and f_i are the same and $|f_i|$ is large, and values close to zero if the signs are different and $|f_i|$ is large. The latent function scale σ_f of the data can be moved into the likelihood $\tilde{\mathrm{sig}}_{\sigma_f}(t) = \mathrm{sig}(\sigma_f^2 t)$, thus σ_f models the steepness of the likelihood and finally the smoothness of the agreement by interpolation between the two limiting cases "ignorant" and "hard cut"

$$\begin{aligned}\lim_{\sigma_f\to 0}\mathrm{sig}(t) &\equiv \frac{1}{2} \quad \text{"ignorant",}\\ \lim_{\sigma_f\to\infty}\mathrm{sig}(t) &\equiv \mathrm{step}(t) := \begin{cases} 0, t<0; & \frac{1}{2}, t=0; & 1, 0<t \end{cases} \quad \text{"hard cut".}\end{aligned}$$

In the case of very small latent scales ($\sigma_f\to 0$), the likelihood is flat causing the posterior to equal the prior. The marginal likelihood is again $Z = 2^{-n}$. (See Figure 3, regimes 7, 8 and 9.)

In the case of large latent scales ($\sigma_f \gg 1$), the likelihood approaches the step function. (See Figure 3, regimes 1, 2 and 3.) A further increase of the latent scale does not change the model anymore. The model is effectively the same for all σ_f above a threshold.

1. Furthermore, we require $g(0) = 1$ and $\lim_{t\to\infty} g(t) = 0$.

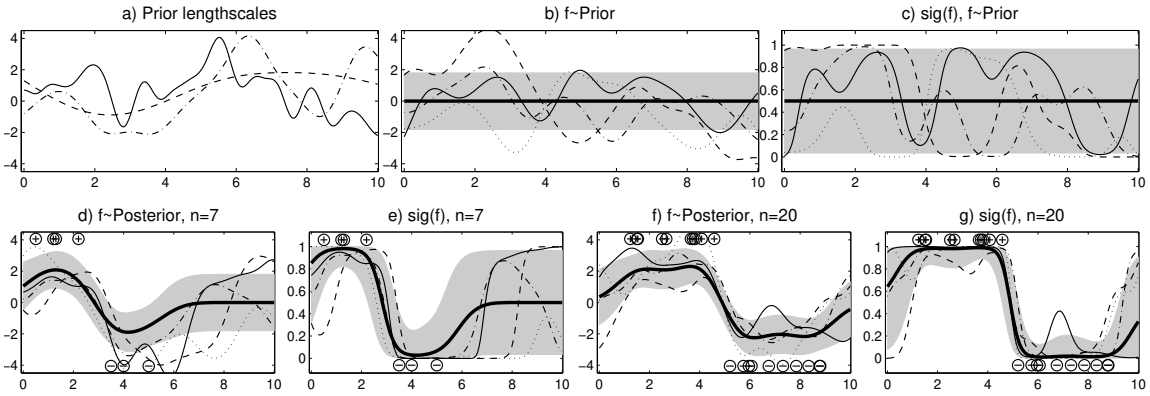


Figure 2: Pictorial illustration of binary Gaussian process classification in 1d: Plot a) shows 3 sample functions drawn from GPs with different lengthscales ℓ. Then, three pairs of plots show distributions over functions $f : \mathbb{R} \to \mathbb{R}$ and $\mathrm{sig}(f) : \mathbb{R} \to [0,1]$ occurring in GP classification. b+c) the prior, d+e) a posterior with $n = 7$ observations and f+g) a posterior with $n = 20$ observations along with the n observations with binary labels. The thick black line is the mean, the gray background is the $\pm$ standard deviation and the thin lines are sample functions. With more and more data points observed, the uncertainty is gradually shrunk. At the decision boundary the uncertainty is smallest.

2.2 Gaussian Approximations

Unfortunately, the posterior over the latent values (Equation 2) is not Gaussian due to the non-Gaussian likelihood (Equation 1). Therefore, the latent distribution (Equation 3), the predictive distribution (Equation 4) and the marginal likelihood Z cannot be written as analytical expressions. To obtain exact answers, one can resort to sampling algorithms (MCMC). However, if sig is concave in the logarithmic domain, the posterior can be shown to be unimodal motivating Gaussian approximations to the posterior. Five different Gaussian approximations corresponding to methods explained later onwards in the paper are depicted in Figure 4.

A quadratic approximation to the log likelihood $\phi(f_i) := \ln \mathbb{P}(y_i|f_i)$ at $\tilde{f}_i$

$$\phi(f_i) \;\approx\; \phi(\tilde{f}_i) + \phi'(\tilde{f}_i)(f_i - \tilde{f}_i) + \frac{1}{2}\phi''(\tilde{f}_i)(f_i - \tilde{f}_i)^2 \;=\; -\frac{1}{2} w_i f_i^2 + b_i f_i + \mathrm{const}_{f_i}$$

motivates the following approximate posterior $\mathbb{Q}(\mathbf{f}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta})$

$$\begin{aligned}
\ln \mathbb{P}(\mathbf{f}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}) \quad &\overset{(2)}{=} \quad -\frac{1}{2}\mathbf{f}^\top \mathbf{K}^{-1}\mathbf{f} + \sum_{i=1}^{n} \ln \mathbb{P}(y_i|f_i) + \mathrm{const}_{\mathbf{f}} \\
&\overset{\text{quad. approx.}}{\approx} \quad -\frac{1}{2}\mathbf{f}^\top \mathbf{K}^{-1}\mathbf{f} - \frac{1}{2}\mathbf{f}^\top \mathbf{W}\mathbf{f} + \mathbf{b}^\top \mathbf{f} + \mathrm{const}_{\mathbf{f}} \\
&\overset{\mathbf{m}:=(\mathbf{K}^{-1}+\mathbf{W})^{-1}\mathbf{b}}{=} \quad -\frac{1}{2}(\mathbf{f}-\mathbf{m})^\top \left(\mathbf{K}^{-1} + \mathbf{W}\right)(\mathbf{f}-\mathbf{m}) + \mathrm{const}_{\mathbf{f}} \\
&= \quad \ln \mathcal{N}(\mathbf{f}|\mathbf{m}, \mathbf{V}) \;=:\; \ln \mathbb{Q}(\mathbf{f}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}), \qquad (5)
\end{aligned}$$

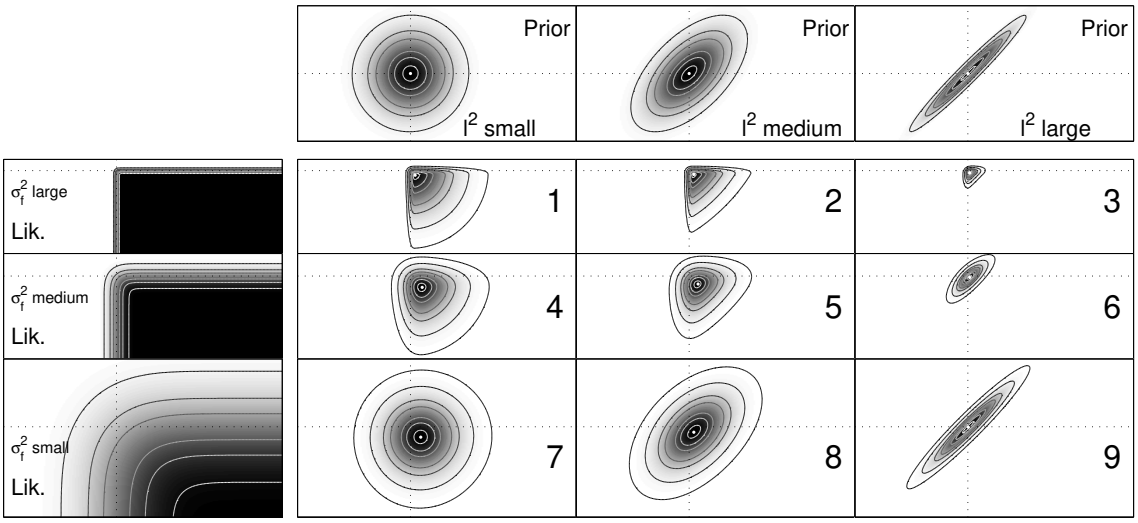


Figure 3: Gaussian Process Classification: Prior, Likelihood and exact Posterior: Nine numbered quadrants show posterior obtained by multiplication of different priors and likelihoods. The leftmost column illustrates the likelihood function for three different steepness parameters σ_f and the upper row depicts the prior for three different length scales ℓ. Here, we use σ_f as a parameter of the likelihood. Alternatively, rows correspond to "degree of Gaussianity" and columns stand for "degree of isotropy". The axes show the latent function values $f_1 = f(\mathbf{x}_1)$ and $f_2 = f(\mathbf{x}_2)$. A simple toy example employing the cumulative Gaussian likelihood and a squared exponential covariance $k(\mathbf{x},\mathbf{x}') = \sigma_f^2 \exp(-\|\mathbf{x}-\mathbf{x}'\|^2/2\ell^2)$ with length scales $\ln \ell = \{0, 1, 2.5\}$ and latent function scales $\ln \sigma_f = \{-1.5, 0, 1.5\}$ is used. Two data points $\mathbf{x}_1 = \sqrt{2}$, $\mathbf{x}_2 = -\sqrt{2}$ with corresponding labels $y_1 = 1$, $y_2 = -1$ form the data set.

where $\mathbf{V}^{-1} = \mathbf{K}^{-1} + \mathbf{W}$ and $\mathbf{W}$ denotes the precision of the effective likelihood (see Equation 7). It turns out that the methods discussed in the following sections correspond to particular choices of $\mathbf{m}$ and $\mathbf{V}$.

Let us assume, we have found such a Gaussian approximation to the posterior with mean $\mathbf{m}$ and (co)variance $\mathbf{V}$. Consequently, the latent distribution for a test point becomes a tractable one-dimensional Gaussian $\mathbb{P}(f_*|\mathbf{x}_*, \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(f_*|\mu_*, \sigma_*^2)$ with the following moments (Rasmussen and Williams, 2006, p. 44 and 56):

$$\begin{array}{rclcrcl} \mu_* & = & \mathbf{k}_*^\top \mathbf{K}^{-1}\mathbf{m} = \mathbf{k}_*^\top \boldsymbol{\alpha}, & & \boldsymbol{\alpha} & = & \mathbf{K}^{-1}\mathbf{m}, \\ \sigma_*^2 & = & k_{**} - \mathbf{k}_*^\top \left(\mathbf{K}^{-1} - \mathbf{K}^{-1}\mathbf{V}\mathbf{K}^{-1}\right)\mathbf{k}_* & = & k_{**} - \mathbf{k}_*^\top \left(\mathbf{K} + \mathbf{W}^{-1}\right)^{-1}\mathbf{k}_*. & & \end{array} \tag{6}$$

Since Gaussians are closed under multiplication, one can—given the Gaussian prior $\mathbb{P}(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})$ and the Gaussian approximation to the posterior $\mathbb{Q}(\mathbf{f}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta})$—deduce the Gaussian factor $\mathbb{Q}(\mathbf{y}|\mathbf{f})$ such that $\mathbb{Q}(\mathbf{f}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}) \propto \mathbb{Q}(\mathbf{y}|\mathbf{f})\,\mathbb{P}(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})$. Consequently, this Gaussian factor can be thought of as an *effective likelihood*. Five different effective likelihoods, corresponding to methods discussed sub-

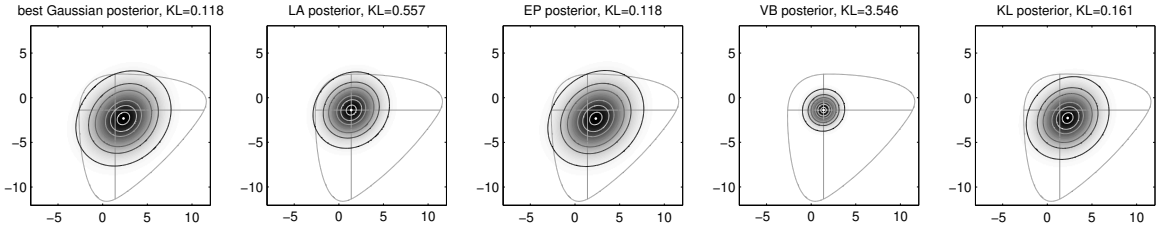


Figure 4: Five Gaussian Approximations to the Posterior (exact Posterior and mode in gray): Different Gaussian approximations to the exact posterior using the regime 2 setting of Figure 3 are shown. The exact posterior is represented in gray by a cross at the mode and a single equiprobability contour line. From left to right: The best Gaussian approximation (intractable) matches the moments of the true posterior, the Laplace approximation does a Taylor expansion around the mode, the EP approximation iteratively matches marginal moments, the variational method maximizes a lower bound on the marginal likelihood and the KL method minimizes the Kullback-Leibler to the exact posterior. The axes show the latent function values $f_1 = f(\mathbf{x}_1)$ and $f_2 = f(\mathbf{x}_2)$.

sequently in the paper, are depicted in Figure 5. By "dividing" the approximate Gaussian posterior (see Appendix B.2) by the true Gaussian prior we find the contribution of the effective likelihood $\mathbb{Q}(\mathbf{y}|\mathbf{f})$:

$$\mathbb{Q}(\mathbf{y}|\mathbf{f}) \propto \frac{\mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V})}{\mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{K})} \propto \mathcal{N}\left(\mathbf{f}|\,(\mathbf{K}\mathbf{W})^{-1}\mathbf{m}+\mathbf{m},\mathbf{W}^{-1}\right). \tag{7}$$

We see (also from Equation 5) that $\mathbf{W}$ models the precision of the effective likelihood. In general, $\mathbf{W}$ is a full matrix containing n^2 parameters.[2] However, all algorithms maintaining a Gaussian posterior approximation work with a diagonal $\mathbf{W}$ to enforce the effective likelihood to factorize over examples (as the true likelihood does, see Figure 1) in order to reduce the number of parameters. We are not aware of work quantifying the error made by this assumption.

2.3 Log Marginal Likelihood

Prior knowledge over the latent function f is encoded in the choice of a covariance function k containing hyperparameters $\boldsymbol{\theta}$. In principle, one can do inference jointly over f and $\boldsymbol{\theta}$ e.g., by sampling techniques. Another approach to model selection is maximum likelihood type II also known as the evidence framework (MacKay, 1992), where the hyperparameters $\boldsymbol{\theta}$ are chosen to maximize the marginal likelihood or evidence $\mathbb{P}(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})$. In other words, one maximizes the agreement between observed data and the model. Therefore, one has a strong motivation to estimate the marginal likelihood.

Geometrically, the marginal likelihood measures the volume of the prior times the likelihood. High volume implies a strong consensus between our initial belief and our observations. In GP classification, each data point $\mathbf{x}_i$ gives rise to a dimension f_i in latent space. The likelihood implements a mechanism, for smoothly restricting the posterior along the axis of f_i to the side corresponding

2. Numerical moment matching with $\mathbf{K} = \begin{bmatrix} 7 & 6 \\ 6 & 7 \end{bmatrix}$, $y_1 = y_2 = 1$ and $\mathrm{sig}_{\mathrm{probit}}$ leads to $\mathbf{W} = \begin{bmatrix} 0.142 & -0.017 \\ -0.017 & 0.142 \end{bmatrix}$.

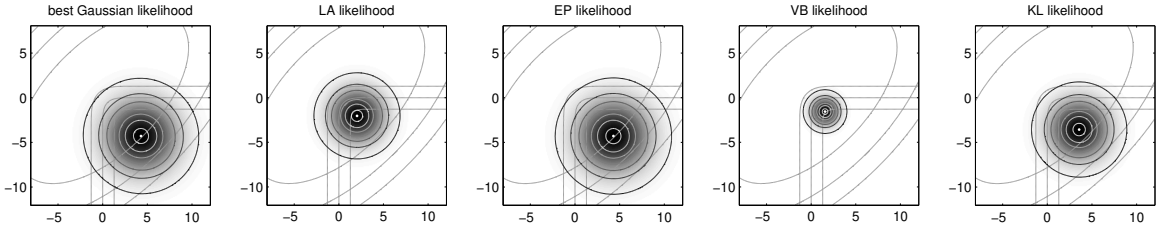


Figure 5: Five Effective Likelihoods (exact Prior/Likelihood in gray): A Gaussian approximation to the posterior induces a Gaussian effective likelihood (Equation 7). Different effective likelihoods are shown; order and setting are the same as described in Figure 4. The axes show the latent function values $f_1 = f(\mathbf{x}_1)$ and $f_2 = f(\mathbf{x}_2)$. The effective likelihood replaces the non-Gaussian likelihood (indicated by three gray lines). A good replacement behaves like the exact likelihood in regions of high prior density (indicated by gray ellipses). EP and KL yield a good coverage of that region. However LA and VB yield too concentrated replacements.

to the sign of y_i . Thus, the latent space $\mathbb{R}^n$ is softly cut down to the orthant given by the values in $\mathbf{y}$. The log marginal likelihood measures, what fraction of the prior lies in that orthant. Finally, the value $Z = 2^{-n}$ corresponds to the case, where half of the prior lies on either side along each axis in latent space. Consequently, successful inference is characterized by $Z > 2^{-n}$.

Some posterior approximations (Sections 3 and 4) provide an approximation to the marginal likelihood, other methods provide a lower bound (Sections 5 and 6). Any Gaussian approximation $\mathbb{Q}(\mathbf{f}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V})$ to the posterior $\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})$ gives rise to a lower bound Z_B to the marginal likelihood Z by application of Jensen's inequality. This bound has been used in the context of sparse approximations (Seeger, 2003).

$$\begin{aligned}\ln Z = \ln \mathbb{P}(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) \quad &= \quad \ln \int \mathbb{P}(\mathbf{y}|\mathbf{f})\,\mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})\,\mathrm{d}\mathbf{f} = \ln \int \mathbb{Q}(\mathbf{f}|\boldsymbol{\theta})\,\frac{\mathbb{P}(\mathbf{y}|\mathbf{f})\,\mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})}{\mathbb{Q}(\mathbf{f}|\boldsymbol{\theta})}\mathrm{d}\mathbf{f} \\ &\overset{\text{Jensen}}{\geq} \quad \int \mathbb{Q}(\mathbf{f}|\boldsymbol{\theta}) \ln \frac{\mathbb{P}(\mathbf{y}|\mathbf{f})\,\mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})}{\mathbb{Q}(\mathbf{f}|\boldsymbol{\theta})}\mathrm{d}\mathbf{f} =: \ln Z_B. \end{aligned} \tag{8}$$

Some algebra (Appendix B.3) leads to the following expression for $\ln Z_B$:

$$\underbrace{\sum_{i=1}^{n} \int \mathcal{N}(f|,0,1) \ln \mathrm{sig}\left(y_i\left\{\sqrt{V_{ii}}f + m_i\right\}\right)\mathrm{d}f}_{\text{1) data fit}} + \frac{1}{2}[n - \underbrace{\mathbf{m}^\top \mathbf{K}^{-1}\mathbf{m}}_{\text{2) data fit}} + \underbrace{\ln\left|\mathbf{V}\mathbf{K}^{-1}\right| - \mathrm{tr}\left(\mathbf{V}\mathbf{K}^{-1}\right)}_{\text{3) regularizer}}]. \tag{9}$$

Model selection means maximization of $\ln Z_B$. Term 1) is a sum of one-dimensional Gaussian integrals of sigmoid functions in the logarithmic domain with adjustable offset and steepness. The integrals can be numerically computed in an efficient way using Gauss-Hermite quadrature (Press et al., 1993, §4.5). As the sigmoid in the log domain takes only negative values, the first term will be negative. That means, maximization of the first term is done by shifting the log-sigmoid such that the high-density region of the Gaussian is multiplied by small values. Term 2) is the equivalent

of the data-fit term in GP regression (Rasmussen and Williams, 2006, Ch. 5.4.1). Thus, the first and the second term encourage fitting the data by favouring small variances V_{ii} and large means m_i having the same sign as y_i. The third term can be rewritten as $-\ln|\mathbf{I}+\mathbf{KW}| - \operatorname{tr}\left((\mathbf{I}+\mathbf{KW})^{-1}\right)$ and yields $-\sum_{i=1}^{n}\ln(1+\lambda_i)+\frac{1}{1+\lambda_i}$ with $\lambda_i \geq 0$ being the eigenvalues of $\mathbf{KW}$. Thus, term 3) keeps the eigenvalues of $\mathbf{KW}$ small, thereby favouring a smaller class of functions—this can be seen as an instance of Occam's razor.

Furthermore, the bound

$$\ln Z_B = \int \mathbb{Q}(\mathbf{f}|\boldsymbol{\theta}) \ln \frac{\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})\,\mathbb{P}(\mathbf{y}|\mathbf{X})}{\mathbb{Q}(\mathbf{f}|\boldsymbol{\theta})} \mathrm{d}\mathbf{f} = \ln Z - \mathrm{KL}\left(\mathbb{Q}(\mathbf{f}|\boldsymbol{\theta}) \parallel \mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})\right) \qquad (10)$$

can be decomposed into the exact marginal likelihood minus the Kullback-Leibler (KL) divergence between the exact posterior and the approximate posterior. Thus by maximizing the lower bound $\ln Z_B$ on $\ln Z$, we effectively minimize the KL-divergence between $\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})$ and $\mathbb{Q}(\mathbf{f}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V})$. The bound is tight if and only if $\mathbb{Q}(\mathbf{f}|\boldsymbol{\theta}) = \mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})$.

3. Laplace Approximation (LA)

A second order Taylor expansion around the posterior mode $\mathbf{m}$ leads to a natural way of constructing a Gaussian approximation to the log-posterior $\boldsymbol{\Psi}(\mathbf{f}) = \ln\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})$ (Williams and Barber, 1998; Rasmussen and Williams, 2006, Ch. 3). The mode $\mathbf{m}$ is taken as the mean of the approximate Gaussian. Linear terms of $\boldsymbol{\Psi}$ vanish because the gradient at the mode is zero. The quadratic term of $\boldsymbol{\Psi}$ is given by the negative Hessian $\mathbf{W}$, which - due to the likelihood's factorial structure - turns out to be diagonal. The mode $\mathbf{m}$ is found by Newton's method.

3.1 Posterior

$$\begin{aligned}
\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta}) &\approx \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V}) = \mathcal{N}\left(\mathbf{f}|\mathbf{m},\left(\mathbf{K}^{-1}+\mathbf{W}\right)^{-1}\right), \\
\mathbf{m} &= \operatorname*{argmax}_{\mathbf{f}\in\mathbb{R}^n} \mathbb{P}(\mathbf{y}|\mathbf{f})\,\mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta}), \\
\mathbf{W} &= -\left.\frac{\partial^2 \ln\mathbb{P}(\mathbf{y}|\mathbf{f})}{\partial\mathbf{f}\partial\mathbf{f}^\top}\right|_{\mathbf{f}=\mathbf{m}} = -\left[\left.\frac{\partial^2 \ln\mathbb{P}(y_i|f_i)}{\partial f_i^2}\right|_{f_i=m_i}\right]_{ii}.
\end{aligned}$$

3.2 Log Marginal Likelihood

The unnormalized posterior $\mathbb{P}(\mathbf{y}|\mathbf{f})\,\mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})$ has its maximum $h = \exp(\boldsymbol{\Psi}(\mathbf{m}))$ at its mode $\mathbf{m}$, where the gradient vanishes. A Taylor expansion of $\boldsymbol{\Psi}$ is then given by $\boldsymbol{\Psi}(\mathbf{f}) \approx h - \frac{1}{2}(\mathbf{f}-\mathbf{m})^\top(\mathbf{K}^{-1}+\mathbf{W})(\mathbf{f}-\mathbf{m})$. Consequently, the log marginal likelihood can be approximated by plugging in the approximation of $\boldsymbol{\Psi}(\mathbf{f})$.

$$\begin{aligned}
\ln Z = \ln\mathbb{P}(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) &= \ln\int \mathbb{P}(\mathbf{y}|\mathbf{f})\,\mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})\,\mathrm{d}\mathbf{f} = \ln\int \exp(\boldsymbol{\Psi}(\mathbf{f}))\,\mathrm{d}\mathbf{f} \\
&\approx \ln h + \ln\int \exp\left(-\frac{1}{2}(\mathbf{f}-\mathbf{m})^\top\left(\mathbf{K}^{-1}+\mathbf{W}\right)(\mathbf{f}-\mathbf{m})\right)\mathrm{d}\mathbf{f} \\
&= \ln\mathbb{P}(\mathbf{y}|\mathbf{m}) - \frac{1}{2}\mathbf{m}^\top\mathbf{K}^{-1}\mathbf{m} + \frac{1}{2}\ln|\mathbf{I}+\mathbf{KW}|.
\end{aligned}$$

4. Expectation Propagation (EP)

EP (Minka, 2001b) is an iterative method to find approximations based on approximate marginal moments, which can be applied to Gaussian processes. See (Rasmussen and Williams, 2006, Ch. 3) for details. The individual likelihood terms are replaced by site functions $t_i(f_i)$ being unnormalized Gaussians

$$\mathbb{P}(y_i|f_i) \approx t_i\left(f_i,\mu_i,\sigma_i^2,Z_i\right) := Z_i\mathcal{N}\left(f_i|\mu_i,\sigma_i^2\right)$$

such that the approximate marginal moments of $\mathbb{Q}(f_i) := \int \mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{K})\prod_{j=1}^n Z_j\mathcal{N}\left(f_j|\mu_j,\sigma_j^2\right)\mathrm{d}\mathbf{f}_{\neg i}$ agree with the marginals of $\int \mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{K})\mathbb{P}(y_i|f_i)\prod_{j\neq i} Z_j\mathcal{N}\left(f_j|\mu_j,\sigma_j^2\right)\mathrm{d}\mathbf{f}_{\neg i}$ of the approximation based on the exact likelihood term $\mathbb{P}(y_j|f_j)$. That means, there are $3n$ quantities μ_i, σ_i^2 and Z_i to be iteratively optimized. Convergence of EP is not generally guaranteed, but there always exists a fixed-point for the EP updates in GP classification (Minka, 2001a). If the EP iterations converge, the solution obtained is a saddle point of a special energy function (Minka, 2001a). However, an EP update does not necessarily imply a decrease in energy. For our case of log-concave likelihood functions, we always observed convergence, but we are not aware of a formal proof.

4.1 Posterior

Based on these local approximations, the approximate posterior can be written as:

$$\begin{aligned}
\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta}) &\approx \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V}) = \mathcal{N}\left(\mathbf{f}|\mathbf{m},\left(\mathbf{K}^{-1}+\mathbf{W}\right)^{-1}\right),\\
\mathbf{W} &= \left[\sigma_i^{-2}\right]_{ii},\\
\mathbf{m} &= \mathbf{V}\mathbf{W}\boldsymbol{\mu} = \left[\mathbf{I}-\mathbf{K}\left(\mathbf{K}+\mathbf{W}^{-1}\right)^{-1}\right]\mathbf{K}\mathbf{W}\boldsymbol{\mu},\ \boldsymbol{\mu}=(\mu_1,\ldots,\mu_n)^\top.
\end{aligned}$$

4.2 Log Marginal Likelihood

>From the likelihood approximations, one can directly obtain an expression for the approximate log marginal likelihood

$$\begin{aligned}
\ln Z = \ln\mathbb{P}(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) &= \ln\int \mathbb{P}(\mathbf{y}|\mathbf{f})\,\mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})\,\mathrm{d}\mathbf{f}\\
&\approx \ln\int\prod_{i=1}^n t\left(f_i,\mu_i,\sigma_i^2,Z_i\right)\mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})\,\mathrm{d}\mathbf{f}\\
&= \sum_{i=1}^n \ln Z_i - \frac{1}{2}\boldsymbol{\mu}^\top\left(\mathbf{K}+\mathbf{W}^{-1}\right)^{-1}\boldsymbol{\mu} - \frac{1}{2}\ln\left|\mathbf{K}+\mathbf{W}^{-1}\right| - \frac{n}{2}\ln 2\pi\\
&= \sum_{i=1}^n \ln\frac{Z_i}{\sqrt{2\pi}} - \frac{1}{2}\mathbf{m}^\top\left(\mathbf{K}^{-1}+\mathbf{K}^{-1}\mathbf{W}^{-1}\mathbf{K}^{-1}\right)\mathbf{m} - \frac{1}{2}\ln\left|\mathbf{K}+\mathbf{W}^{-1}\right| =: \ln Z_{EP}.
\end{aligned}$$

The lower bound provided by Jensen's inequality Z_B (Equation 9) is known to be below the approximation Z_{EP} obtained by EP (Opper and Winther, 2005, page 2183). From $Z_{EP} \geq Z_B$ and $Z \geq Z_B$ it is not clear, which value one should use. In principle, Z_{EP} could be a bad approximation. However, our experimental findings and extensive Monte Carlo simulations suggest that Z_{EP} is very accurate.

4.3 Thouless, Anderson & Palmer method (TAP)

Based on ideas rooted in Statistical Physics, one can approach the problem from a slightly different angle (Opper and Winther, 2000). Individual Gaussian approximations $\mathcal{N}(f_i|\mu_{\neg i},\sigma^2_{\neg i})$ are only made to predictive distributions $\mathbb{P}\left(f_i|\mathbf{x}_i,\mathbf{y}_{\setminus i},\mathbf{X}_{\setminus i},\boldsymbol{\theta}\right)$ for data points $\mathbf{x}_i$ that have been previously removed from the training set. Based on $\mu_{\neg i}$ and $\sigma^2_{\neg i}$ one can derive explicit expressions for $(\boldsymbol{\alpha},\mathbf{W}^{\frac{1}{2}})$, our parameters of interest.

$$\begin{aligned}\alpha_i &\approx \frac{\int \frac{\partial}{\partial f_i}\mathbb{P}(y_i|f_i)\,\mathcal{N}(f_i|\mu_{\neg i},\sigma^2_{\neg i})\mathrm{d}f_i}{\int \mathbb{P}(y_i|f_i)\,\mathcal{N}(f_i|\mu_{\neg i},\sigma^2_{\neg i})\mathrm{d}f_i},\\ \left[\mathbf{W}^{-1}\right]_{ii} &\approx \sigma^2_{\neg i}\left(\frac{1}{\alpha_i\left[\mathbf{K}\boldsymbol{\alpha}\right]_i}-1\right).\end{aligned} \tag{11}$$

In turn, the $2n$ parameters $(\mu_{\neg i},\sigma^2_{\neg i})$ can be expressed as a function of $\boldsymbol{\alpha}$, $\mathbf{K}$ and $\mathbf{W}^{\frac{1}{2}}$.

$$\begin{aligned}\sigma^2_{\neg i} &= 1/\left[\left(\mathbf{K}+\mathbf{W}^{-1}\right)^{-1}\right]_{ii}-\left[\mathbf{W}^{-1}\right]_{ii},\\ \mu_{\neg i} &= \left[\mathbf{K}\boldsymbol{\alpha}\right]_i-\sigma^2_{\neg i}\alpha_i.\end{aligned} \tag{12}$$

As a result, a system (Equations 11/12) of nonlinear equations in $\mu_{\neg i}$ and $\sigma^2_{\neg i}$ has to be solved by iteration. Each step involves a matrix inversion of cubic complexity. A faster "naïve" variant updating only n parameters has also been proposed (Opper and Winther, 2000) but it does not lead to the same fixed point. As in the FV algorithm (Section 7), a formal complex transformation leads to a simplified version by fixing $\sigma^2_{\neg i}=\mathbf{K}_{ii}$, called (TAPnaive) in the sequel.

Finally, for prediction, the predictive posterior $\mathbb{P}(f_*|\mathbf{x}_*,\mathbf{y},\mathbf{X},\boldsymbol{\theta})$ is approximated by a Gaussian $\mathcal{N}(f_*|\mu_*,\sigma^2_*)$ at a test point $\mathbf{x}_*$ based on the parameters $(\boldsymbol{\alpha},\mathbf{W}^{\frac{1}{2}})$ and according to equation (6).

A fixed-point of the TAP mean-field equations is also a fixed-point of the EP algorithm (Minka, 2001a). This theoretical result was confirmed in our numerical simulations. However, the EP algorithm is more practical and typically much faster. For this reason, we are not going to treat the TAP method as an independent algorithm in this paper.

5. KL-Divergence Minimization (KL)

In principle, we simply want to minimize a dissimilarity measure between the approximate posterior $\mathbb{Q}(\mathbf{f}|\boldsymbol{\theta})=\mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V})$ and the exact posterior $\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})$. One quantity to minimize is the KL-divergence

$$\mathrm{KL}\left(\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})\parallel\mathbb{Q}(\mathbf{f}|\boldsymbol{\theta})\right) = \int\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})\ln\frac{\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})}{\mathbb{Q}(\mathbf{f}|\boldsymbol{\theta})}\mathrm{d}\mathbf{f}.$$

Unfortunately, this expression is intractable. If instead, we measure the reverse KL-divergence, we regain tractability

$$\mathrm{KL}\left(\mathbb{Q}(\mathbf{f}|\boldsymbol{\theta})\parallel\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})\right) = \int\mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V})\ln\frac{\mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V})}{\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})}\mathrm{d}\mathbf{f}=:\mathrm{KL}(\mathbf{m},\mathbf{V}).$$

A similar approach has been followed for regression with Laplace or Cauchy noise (Opper and Archambeau, 2008). Finally, we minimize the following objective (see Appendix B.3) with respect to the variables $\mathbf{m}$ and $\mathbf{V}$. Constant terms have been dropped from the expression:

$$\mathrm{KL}(\mathbf{m},\mathbf{V}) \stackrel{c}{=} -\int \mathcal{N}(f)\left[\sum_{i=1}^{n} \ln \mathrm{sig}\left(\sqrt{v_{ii}}y_i f + m_i y_i\right)\right] \mathrm{d}f - \frac{1}{2}\ln|\mathbf{V}| + \frac{1}{2}\mathbf{m}^\top \mathbf{K}^{-1}\mathbf{m} + \frac{1}{2}\mathrm{tr}\left(\mathbf{K}^{-1}\mathbf{V}\right).$$

We refer to the first term of $\mathrm{KL}(\mathbf{m},\mathbf{V})$ as $a(\mathbf{m},\mathbf{V})$ to keep the expressions short. We calculate first derivatives and equate them with zero to obtain necessary conditions that have to be fulfilled at a local optimum $(\mathbf{m}^*,\mathbf{V}^*)$

$$\begin{aligned}
\frac{\partial \mathrm{KL}}{\partial \mathbf{m}} &= \frac{\partial a}{\partial \mathbf{m}} - \mathbf{K}^{-1}\mathbf{m} = \mathbf{0} \;\Rightarrow\; \mathbf{K}^{-1}\mathbf{m} = \frac{\partial a}{\partial \mathbf{m}} = \boldsymbol{\alpha}, \\
\frac{\partial \mathrm{KL}}{\partial \mathbf{V}} &= \frac{\partial a}{\partial \mathbf{V}} + \frac{1}{2}\mathbf{V}^{-1} - \frac{1}{2}\mathbf{K}^{-1} = \mathbf{0} \;\Rightarrow\; \mathbf{V} = \left(\mathbf{K}^{-1} - 2\frac{\partial a}{\partial \mathbf{V}}\right)^{-1} = \left(\mathbf{K}^{-1} - 2\boldsymbol{\Lambda}\right)^{-1}
\end{aligned}$$

which defines $\boldsymbol{\Lambda}$. If the approximate posterior is parametrized by $(\mathbf{m},\mathbf{V})$, there are in principle in the order of n^2 parameters. But if the necessary conditions for a local minimum are fulfilled (i.e., the derivatives $\partial\mathrm{KL}/\partial\mathbf{m}$ and $\partial\mathrm{KL}/\partial\mathbf{V}$ vanish), the problem can be re-parametrized in terms of $(\boldsymbol{\alpha},\boldsymbol{\Lambda})$. Since $\boldsymbol{\Lambda} = \partial a/\partial\mathbf{V}$ is a diagonal matrix (see Equation 17), the optimum is characterized $2n$ free parameters. This fact was already pointed out by Manfred Opper (personal communication) and Matthias Seeger (Seeger, 1999, Ch. 5.21, Eq. 5.3). Thus, a minimization scheme based on Newton iterations on the joint vector $\boldsymbol{\xi} := [\boldsymbol{\alpha}^\top, \boldsymbol{\Lambda}_{ii}]^\top$ takes $O(8\cdot n^3)$ operations. Details about the derivatives $\partial\mathrm{KL}/\partial\boldsymbol{\xi}$ and $\partial^2\mathrm{KL}/\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^\top$ are provided in Appendix A.2.

5.1 Posterior

Based on these local approximations, the approximate posterior can be written as:

$$\begin{aligned}
\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta}) &\approx \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V}) = \mathcal{N}\left(\mathbf{f}|\mathbf{m},\left(\mathbf{K}^{-1}+\mathbf{W}\right)^{-1}\right), \\
\mathbf{W} &= -2\boldsymbol{\Lambda}, \\
\mathbf{m} &= \mathbf{K}\boldsymbol{\alpha}.
\end{aligned}$$

5.2 Log Marginal Likelihood

Since the method inherently maximizes a lower bound on the marginal likelihood, this bound (Equation 9) is used as approximation to the marginal likelihood.

6. Variational Bounds (VB)

The following variational bounding method (Gibbs and MacKay, 2000) is a special case of the KL method. Instead of optimizing a bound on the joint (Eq. 8), they impose the bounding condition on each likelihood term individually. Here, we treat parametrization based on quadratic lower bounds on the individual likelihoods in the logarithmic domain. We first derive all calculations based on

general likelihoods. Individual likelihood bounds

$$\begin{aligned}\mathbb{P}(y_i|f_i) &\geq \exp\left(a_i f_i^2 + b_i y_i f_i + c_i\right), \forall f_i \in \mathbb{R} \,\forall i \\ \Rightarrow \mathbb{P}(\mathbf{y}|\mathbf{f}) &\geq \exp\left(\mathbf{f}^\top \mathbf{A}\mathbf{f} + (\mathbf{b}\odot\mathbf{y})^\top \mathbf{f} + \mathbf{c}^\top \mathbb{1}\right) =: \mathbb{Q}(\mathbf{y}|\mathbf{f},\mathbf{A},\mathbf{b},\mathbf{c}), \forall \mathbf{f} \in \mathbb{R}\end{aligned}$$

are defined in terms of coefficients a_i, b_i and c_i, where $\odot$ denotes the element-wise product of two vectors. This lower bound on the likelihood induces a lower bound on the marginal likelihood.

$$Z = \int \mathbb{P}(\mathbf{f}|\mathbf{X})\,\mathbb{P}(\mathbf{y}|\mathbf{f})\,\mathrm{d}\mathbf{f} \quad \geq \quad \int \mathbb{P}(\mathbf{f}|\mathbf{X})\,\mathbb{Q}(\mathbf{y}|\mathbf{f},\mathbf{A},\mathbf{b},\mathbf{c})\,\mathrm{d}\mathbf{f} = Z_B.$$

Carrying out the Gaussian integral

$$Z_B \quad = \quad \int \mathcal{N}(\mathbf{f}|0,\mathbf{K}) \exp\left(\mathbf{f}^\top \mathbf{A}\mathbf{f} + (\mathbf{b}\odot\mathbf{y})^\top \mathbf{f} + \mathbf{c}^\top \mathbb{1}\right) \mathrm{d}\mathbf{f}$$

leads to (see Appendix B.4)

$$\ln Z_B \quad = \quad \mathbf{c}^\top \mathbb{1} + \frac{1}{2}(\mathbf{b}\odot\mathbf{y})^\top \left(\mathbf{K}^{-1} - 2\mathbf{A}\right)^{-1} (\mathbf{b}\odot\mathbf{y}) - \frac{1}{2}\ln|\mathbf{I} - 2\mathbf{A}\mathbf{K}| \tag{13}$$

which can now be maximized with respect to the coefficients a_i, b_i and c_i. In order to get an efficient algorithm, one has to calculate the first and second derivatives $\partial \ln Z_B/\partial\boldsymbol{\varsigma}$, $\partial^2 \ln Z_B/\partial\boldsymbol{\varsigma}\partial\boldsymbol{\varsigma}^\top$ (as done in Appendix A.1). Hyperparameters can be optimized using the gradient $\partial \ln Z_B/\partial\boldsymbol{\theta}$.

6.1 Logit Bound

Optimizing the logistic likelihood function (Gibbs and MacKay, 2000), we obtain the necessary conditions

$$\begin{aligned}\mathbf{A}_{\boldsymbol{\varsigma}} &:= -\boldsymbol{\Lambda}_{\boldsymbol{\varsigma}}, \\ \mathbf{b}_{\boldsymbol{\varsigma}} &:= \frac{1}{2}\mathbb{1}, \\ \mathbf{c}_{\boldsymbol{\varsigma},i} &:= \varsigma_i^2 \lambda(\varsigma_i) - \frac{1}{2}\varsigma_i + \ln \mathrm{sig}_{\mathrm{logit}}(\varsigma_i)\end{aligned}$$

where we define $\lambda(\varsigma_i) = \left(2\mathrm{sig}_{\mathrm{logit}}(\varsigma_i) - 1\right)/(4\varsigma_i)$ and $\boldsymbol{\Lambda}_{\boldsymbol{\varsigma}} = [\lambda(\varsigma_i)]_{ii}$. This shows, that we only have to optimize with respect to n parameters $\boldsymbol{\varsigma}$. We apply Newton's method for this purpose. The bound is symmetric and tight at $\mathbf{f} = \pm\boldsymbol{\varsigma}$.

6.2 Probit Bound

For reasons of completeness, we derive similar expressions (Appendix B.5) for the cumulative Gaussian likelihood $\mathrm{sig}_{\mathrm{probit}}(f_i)$ with necessary conditions

$$\begin{aligned}\mathbf{a}_{\boldsymbol{\varsigma}} &:= -\frac{1}{2}\mathbb{1}, \\ \mathbf{b}_{\boldsymbol{\varsigma},i} &:= \varsigma_i + \frac{\mathcal{N}(\varsigma_i)}{\mathrm{sig}_{\mathrm{probit}}(\varsigma_i)}, \\ \mathbf{c}_{\boldsymbol{\varsigma},i} &:= \left(\frac{\varsigma_i}{2} - b_i\right)\varsigma_i + \ln\left(\mathrm{sig}_{\mathrm{probit}}(\varsigma_i)\right)\end{aligned} \tag{14}$$

which again depend only on a single vector of parameters we optimize using Newton's method. The bound is tight for $\mathbf{f} = \boldsymbol{\varsigma}$.

6.3 Posterior

Based on these local approximations, the approximate posterior can be written as

$$\begin{aligned}\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta}) &\approx \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V}) = \mathcal{N}\left(\mathbf{f}|\mathbf{m},\left(\mathbf{K}^{-1}+\mathbf{W}\right)^{-1}\right),\\ \mathbf{W} &= -2\mathbf{A}_{\boldsymbol{\varsigma}},\\ \mathbf{m} &= \mathbf{V}(\mathbf{y}\odot\mathbf{b}_{\boldsymbol{\varsigma}}) = \left(\mathbf{K}^{-1}-2\mathbf{A}_{\boldsymbol{\varsigma}}\right)^{-1}(\mathbf{y}\odot\mathbf{b}_{\boldsymbol{\varsigma}}),\end{aligned}$$

where we have expressed the posterior parameters directly as a function of the coefficients. Finally, we deal with an approximate posterior $\mathbb{Q}(\mathbf{f}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\mathbf{m}_{\boldsymbol{\varsigma}},\mathbf{V}_{\boldsymbol{\varsigma}})$ only depending on a vector $\boldsymbol{\varsigma}$ of n variational parameters and a mapping $\boldsymbol{\varsigma}\mapsto(\mathbf{m}_{\boldsymbol{\varsigma}},\mathbf{V}_{\boldsymbol{\varsigma}})$. In the KL method, every combination of values $\mathbf{m}$ and $\mathbf{W}$ is allowed, in the VB method, $\mathbf{m}_{\boldsymbol{\varsigma}}$ and $\mathbf{V}_{\boldsymbol{\varsigma}}$ cannot be chosen independently, since the have to be compatible with the bounding requirements. Therefore, the variational posterior is more constrained than the general Gaussian posterior and thus easier to optimize.

6.4 Log Marginal Likelihood

It turns out, that the approximation to the marginal likelihood (Equation 13) is often quite poor and the more general Jensen bound approach (Equation 9) is much tighter. In practice, one would have to evaluate both of them and keep the maximum value.

7. Factorial Variational Method (FV)

Instead of approximating the posterior $\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})$ by the closest Gaussian distribution, one can use the closest factorial distribution $\mathbb{Q}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta}) = \prod_i \mathbb{Q}(f_i)$, also called *ensemble learning* (Csató et al., 2000). Another kind of factorial approximation $\mathbb{Q}(\mathbf{f}) = \mathbb{Q}(\mathbf{f}^+)\mathbb{Q}(\mathbf{f}^-)$—a posterior factorizing over classes—is used in multi-class classification (Girolami and Rogers, 2006).

7.1 Posterior

As a result of free-form minimization of the Kullback-Leibler divergence $\mathrm{KL}(\mathbb{Q}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta}) \parallel \mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta}))$ by equating its functional derivative $\delta\mathrm{KL}/\delta\mathbb{Q}(f_i)$ with the zero function (Appendix B.6), one finds the best approximation to be of the following form:

$$\begin{aligned}\mathbb{Q}(f_i) &\propto \mathcal{N}\left(f_i\middle|\mu_i,\sigma_i^2\right)\mathbb{P}(y_i|f_i),\\ \mu_i &= m_i - \sigma_i^2\left[\mathbf{K}^{-1}\mathbf{m}\right]_i = [\mathbf{K}\boldsymbol{\alpha}]_i - \sigma_i^2\alpha_i,\\ \sigma_i^2 &= \left[\mathbf{K}^{-1}\right]_{ii}^{-1},\\ m_i &= \int f_i\mathbb{Q}(f_i)\,\mathrm{d}f_i.\end{aligned} \tag{15}$$

In fact, the best product distribution consists of a factorial Gaussian times the original likelihood. The Gaussian has the same moments as the Leave-One-Out prediction (Sundararajan and Keerthi, 2001). Since the posterior is factorial, the effective likelihood of the factorial approximation has an odd shape. It effectively has to annihilate the correlations in the prior, and these correlations are usually what allows learning to happen in the first place. However, the best fitting factorial is still able to ensure that the latent means have the right signs. Even though all correlations are neglected,

it is still possible that the model picks up the most important structure, since the expectations are coupled. Of course, at test time, it is essential that correlations are taken into account again using Equation 6, as it would otherwise be impossible to inject any knowledge into the predictive distribution. For predictions we use the Gaussian $\mathcal{N}(\mathbf{f}|\mathbf{m},\mathrm{Dg}(\mathbf{v}))$ instead of $\mathbb{Q}(\mathbf{f})$. This is a further approximation, but it allows to stay inside the Gaussian framework.

Parameters μ_i and m_i are found by the following algorithm. Starting from $\mathbf{m}=\mathbf{0}$, iterate the following until convergence; (1) compute μ_i, (2) update m_i by taking a step in the direction towards m_i as given by Equation 15. Stepsizes are adapted.

7.2 Log Marginal Likelihood

Surprisingly, one can obtain a lower bound on the marginal likelihood (Csató et al., 2000):

$$\ln Z \quad \geq \quad \sum_{i=1}^{n} \ln \mathrm{sig}\left(\frac{y_i m_i}{\sigma_i}\right) - \frac{1}{2}\boldsymbol{\alpha}^\top \left(\mathbf{K} - \mathrm{Dg}(\left[\sigma_1^2,\ldots,\sigma_n^2\right]^\top)\right)\boldsymbol{\alpha} - \frac{1}{2}\ln|\mathbf{K}| + \sum_{i=1}^{n}\ln\sigma_i.$$

8. Label Regression Method (LR)

Classification has also been treated using label regression or least squares classification (Rifkin and Klautau, 2004). In its simplest form, this method simply ignores the discreteness of the class labels at the cost of not being able to provide proper probabilistic predictions. However, we treat LR as a heuristic way of choosing $\boldsymbol{\alpha}$ and $\mathbf{W}$, which allows us to think of it as yet another Gaussian approximation to the posterior allowing for valid predictions of class probabilities.

8.1 Posterior

After inference, according to Equation 6, the moments of the (Gaussian approximation to the) posterior GP can be written as $\mu_* = \mathbf{k}_*^\top\boldsymbol{\alpha}$ and $\sigma_*^2 = k_{**} - \mathbf{k}_*^\top\left(\mathbf{K}+\mathbf{W}^{-1}\right)^{-1}\mathbf{k}_*$. Fixing

$$\mathbf{W}^{-1} = \sigma_n^2\mathbf{I} \quad \text{and} \quad \boldsymbol{\alpha} = \left(\mathbf{K}+\mathbf{W}^{-1}\right)^{-1}\left(\mathbf{K}+\mathbf{W}^{-1}\right)\boldsymbol{\alpha} = \left(\mathbf{K}+\mathbf{W}^{-1}\right)^{-1}\mathbf{y},$$

we obtain GP regression from data points $\mathbf{x}_i \in \mathcal{X}$ to real labels $y_i \in \mathbb{R}$ with noise of variance σ_n^2 as a special case. In regression, the posterior moments are given by $\mu_* = \mathbf{k}_*^\top\left(\mathbf{K}+\sigma_n^2\mathbf{I}\right)^{-1}\mathbf{y}$ and $\sigma_*^2 = k_{**} - \mathbf{k}_*^\top\left(\mathbf{K}+\sigma_n^2\mathbf{I}\right)^{-1}\mathbf{k}_*$ (Rasmussen and Williams, 2006). The arbitrary scale of the discrete $\mathbf{y}$ can be absorbed by the hyperparameters. There is an additional parameter σ_n, describing the width of the effective likelihood. In experiments, we selected $\sigma_n \in [0.5, 2]$ to maximise the log marginal likelihood.

8.2 Log Marginal Likelihood

There are two ways of obtaining an estimate of the log marginal likelihood. One can simply ignore the binary nature and use the regression marginal likelihood $\ln Z_{\mathrm{reg}}$ as proxy for $\ln Z$—an approach we only mention but not use in the experiments

$$\ln Z_{\mathrm{reg}} \quad = \quad -\frac{1}{2}\boldsymbol{\alpha}^\top\left(\mathbf{K}+\sigma_n^2\mathbf{I}\right)\boldsymbol{\alpha} - \frac{1}{2}\ln\left|\mathbf{K}+\sigma_n^2\mathbf{I}\right| - \frac{n}{2}\ln 2\pi.$$

Alternatively, the Jensen bound (8) yields a lower bound $\ln Z \geq \ln Z_B$—which seems more in line with the classification scenario than $\ln Z_{\mathrm{reg}}$.

9. Relations Between the Methods

All considered approximations can be separated into *local* and *global methods*. Local methods exploit properties (such as derivatives) of the posterior at a special location only. Global methods minimize the KL-divergence $\mathrm{KL}(\mathbb{Q}||\mathbb{P}) = \int \mathbb{Q}(\mathbf{f}) \ln \mathbb{Q}(\mathbf{f}) / \mathbb{P}(\mathbf{f}) \, \mathrm{d}\mathbf{f}$ between the posterior $\mathbb{P}(\mathbf{f})$ and a tractable family of distributions $\mathbb{Q}(\mathbf{f})$. Often this methodology is also referred to as a variational algorithm.

assumption	relation	conditions	approx. posterior $\mathbb{Q}(\mathbf{f})$	name
$\mathbb{Q}(\mathbf{f}) = \mathcal{N}(\mathbf{f}\|\mathbf{m},\mathbf{V})$	$\rightarrow$	$\mathbf{m} = \operatorname{argmax}_{\mathbf{f}} \mathbb{P}(\mathbf{f})$ $\mathbf{W} = -\frac{\partial^2 \ln \mathbb{P}(\mathbf{y}\|\mathbf{f})}{\partial \mathbf{f} \partial \mathbf{f}^\top}$	$\mathcal{N}(\mathbf{f}\|\mathbf{m},(\mathbf{K}^{-1}+\mathbf{W})^{-1})$	LA
$\mathbb{Q}(\mathbf{f}) = \prod_i q_i(f_i)$	$\rightarrow$	$\frac{\delta \mathrm{KL}}{\delta q_i(f_i)} \equiv \mathbf{0}$	$\prod_i \mathcal{N}(f_i\|\mu_i,\sigma_i^2)\mathbb{P}(y_i\|f_i)$	FV
	$\searrow$			
		$\langle f_i^d \rangle_{q_i(f_i)} = \langle f_i^d \rangle_{\mathbb{Q}(f_i)}$	$\mathcal{N}\left(\mathbf{f}\|\mathbf{m},(\mathbf{K}^{-1}+\mathbf{W})^{-1}\right)$	EP
	$\nearrow$			
$\mathbb{Q}(\mathbf{f}) = \mathcal{N}(\mathbf{f}\|\mathbf{m},\mathbf{V})$	$\rightarrow$	$\frac{\partial \mathrm{KL}}{\partial \mathbf{V},\mathbf{m}} = \mathbf{0}$	$\mathcal{N}\left(\mathbf{f}\|\mathbf{m},(\mathbf{K}^{-1}+\mathbf{W})^{-1}\right)$	KL
	$\searrow$			
$\mathbb{P}(y_i\|f_i) \geq \mathcal{N}(f_i\|\mu_{\varsigma i},\sigma_{\varsigma_i}^2)$	$\rightarrow$	$\frac{\partial \mathrm{KL}}{\partial \varsigma_*} = \mathbf{0}$	$\mathcal{N}\left(\mathbf{f}\|\mathbf{m}_{\varsigma_*},(\mathbf{K}^{-1}+\mathbf{W}_{\varsigma_*})^{-1}\right)$	VB
$\mathbb{P}(y_i\|f_i) := \mathcal{N}(f_i\|y_i,\sigma_n^2)$	$\rightarrow$	$\mathbf{m} = (\mathbf{I}+\sigma_n^2\mathbf{K}^{-1})^{-1}\mathbf{y}$	$\mathcal{N}(\mathbf{f}\|\mathbf{m},(\mathbf{K}^{-1}+\sigma_n^{-2}\mathbf{I})^{-1})$	LR

The only local method considered is the LA approximation matching curvature at the posterior mode. Common tractable distributions for global methods include factorial and Gaussian distributions. They have their direct correspondent in the FV method and the KL method. Individual likelihood bounds make the VB method a more constrained and easier-to-optimize version of the KL method. Interestingly, EP can be seen in some sense as a hybrid version of FV and KL, combining the advantages of both methods. Within the Expectation Consistence framework (Opper and Winther, 2005), EP can be thought of as an algorithm that implicitly works with two distributions—a factorial and a Gaussian—having the same marginal moments $\langle f_i^d \rangle$. By means of iterative updates, one keeps these expectations consistent and produces a posterior approximation.

In the divergence measure and message passing framework (Minka, 2005), EP is cast as a message passing algorithm template: Iterative minimization of local divergences to a tractable family of distributions yields a small global divergence. From that viewpoint, FV and KL are considered as special cases with divergence measure $\mathrm{KL}(\mathbb{Q}||\mathbb{P})$ combined with factorial and Gaussian distributions.

There is also a link between local and global methods, namely from the KL to the LA method. The necessary conditions for the LA method do hold *on average* for the KL method (Opper and Archambeau, 2008).

Finally, LR neither qualifies as local nor global—it is just a heuristic way of setting $\mathbf{m}$ and $\mathbf{W}$.

10. Markov Chain Monte Carlo (MCMC)

The only way of getting a handle on the ground truth for the moments Z, $\mathbf{m}$ and $\mathbf{V}$ is by applying sampling techniques. In the limit of long runs, one is guaranteed to get the right answer. But in practice, these methods can be very slow, compared to analytic approximations discussed previously. MCMC runs are rather supposed to provide a gold standard for the comparison of the other methods.

It turns out to be most challenging to obtain reliable marginal likelihood estimates as it is equivalent to solving the free energy problem in physics. We employ Annealed Importance Sampling (AIS) and thermodynamic integration to yield the desired marginal likelihoods. Instead of starting annealing from the prior distribution, we propose to directly start from an approximate posterior in order to speed up the sampling process.

Accurate estimates of the first and second moments can be obtained by sampling directly from the (unnormalized) posterior using Hybrid Monte Carlo methods (Neal, 1993).

10.1 Thermodynamic Integration

The goal is to calculate the marginal likelihood $Z = \int \mathbb{P}(\mathbf{y}|\mathbf{f})\,\mathbb{P}(\mathbf{f}|\mathbf{X})\,\mathrm{d}\mathbf{f}$. AIS (Neal, 1993, 2001) works with intermediate quantities $Z_t := \int \mathbb{P}(\mathbf{y}|\mathbf{f})^{\tau(t)}\,\mathbb{P}(\mathbf{f}|\mathbf{X})\,\mathrm{d}\mathbf{f}$. Here, $\tau : \mathbb{N} \supset [0,T] \to [0,1] \subset \mathbb{R}$ denotes an inverse temperature schedule with the properties $\tau(0) = 0$, $\tau(T) = 1$ and $\tau(t+1) \geq \tau(t)$ leading to $Z_0 = \int \mathbb{P}(\mathbf{f}|\mathbf{X})\,\mathrm{d}\mathbf{f} = 1$ and $Z_T = Z$.

On the other hand, we have $Z = Z_T/Z_0 = \prod_{t=1}^{T} Z_t/Z_{t-1}$—an expanded fraction. Each factor Z_t/Z_{t-1} can be approximated by importance sampling with samples $\mathbf{f}_s$ from the "intermediate posterior" $\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},t-1) := \mathbb{P}(\mathbf{y}|\mathbf{f})^{\tau(t-1)}\,\mathbb{P}(\mathbf{f}|\mathbf{X})\,/Z_{t-1}$ at time t.

$$\begin{aligned}
\frac{Z_t}{Z_{t-1}} &= \frac{\int \mathbb{P}(\mathbf{y}|\mathbf{f})^{\tau(t)}\,\mathbb{P}(\mathbf{f}|\mathbf{X})\,\mathrm{d}\mathbf{f}}{Z_{t-1}} = \int \frac{\mathbb{P}(\mathbf{y}|\mathbf{f})^{\tau(t)}}{\mathbb{P}(\mathbf{y}|\mathbf{f})^{\tau(t-1)}} \frac{\mathbb{P}(\mathbf{y}|\mathbf{f})^{\tau(t-1)}\,\mathbb{P}(\mathbf{f}|\mathbf{X})}{Z_{t-1}}\mathrm{d}\mathbf{f} \\
&= \int \mathbb{P}(\mathbf{y}|\mathbf{f})^{\Delta\tau(t)}\,\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},t-1)\,\mathrm{d}\mathbf{f} \\
&\approx \frac{1}{S}\sum_{s=1}^{S} \mathbb{P}(\mathbf{y}|\mathbf{f}_s)^{\Delta\tau(t)}, \qquad \mathbf{f}_s \sim \mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},t-1)\,.
\end{aligned}$$

This works fine for small temperature changes $\Delta\tau(t) := \tau(t) - \tau(t-1)$. In the limit, we smoothly interpolate between $\mathbb{P}(\mathbf{y}|\mathbf{f})^0\,\mathbb{P}(\mathbf{f}|\mathbf{X})$ and $\mathbb{P}(\mathbf{y}|\mathbf{f})^1\,\mathbb{P}(\mathbf{f}|\mathbf{X})$, that is, we start by sampling from the prior and finally approach the posterior. Note that sampling is algorithmically possible even though the distribution is only known up to a constant factor.

10.2 Amelioration Using an Approximation to the Posterior

In practice, the posterior can be quite different from the prior. That means that individual fractions Z_t/Z_{t-1} may be difficult to estimate. One can make these fractions more similar by increasing the number of steps T or by "starting" from a distribution close to the posterior rather than from the prior. Let $\mathbb{Q}(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V}) \approx \mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},T) = \mathbb{P}(\mathbf{y}|\mathbf{f})\,\mathbb{P}(\mathbf{f}|\mathbf{X})\,/Z_T$ denote an approximation to the posterior. Setting $\mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V}) = \mathbb{Q}(\mathbf{y}|\mathbf{f})\,\mathbb{P}(\mathbf{f}|\mathbf{X})$, one can calculate the effective likelihood $\mathbb{Q}(\mathbf{y}|\mathbf{f})$ by division (see Appendix B.2).

For the integration we use $Z_t = \int \mathbb{P}(\mathbf{y}|\mathbf{f})^{\tau(t)}\,\mathbb{Q}(\mathbf{y}|\mathbf{f})^{1-\tau(t)}\mathbb{P}(\mathbf{f}|\mathbf{X})\,\mathrm{d}\mathbf{f}$ where $Z_0 = \int \mathbb{Q}(\mathbf{y}|\mathbf{f})\,\mathbb{P}(\mathbf{f}|\mathbf{X})\,\mathrm{d}\mathbf{f}$ can be computed analytically. Again, each factor $\frac{Z_t}{Z_{t-1}}$ of the expanded fraction can be approximated

by importance sampling from the modified intermediate posterior:

$$\begin{aligned}\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},t-1) &= \mathbb{P}(\mathbf{y}|\mathbf{f})^{\tau(t-1)}\mathbb{Q}(\mathbf{y}|\mathbf{f})^{1-\tau(t-1)}\mathbb{P}(\mathbf{f}|\mathbf{X})/Z_{t-1} \\ &= \left[\frac{\mathbb{P}(\mathbf{y}|\mathbf{f})}{\mathbb{Q}(\mathbf{y}|\mathbf{f})}\right]^{\tau(t-1)}\mathbb{Q}(\mathbf{y}|\mathbf{f})\,\mathbb{P}(\mathbf{f}|\mathbf{X})/Z_{t-1},\end{aligned}$$

$$\begin{aligned}\frac{Z_t}{Z_{t-1}} &= \frac{\int \mathbb{P}(\mathbf{y}|\mathbf{f})^{\tau(t)}\mathbb{Q}(\mathbf{y}|\mathbf{f})^{1-\tau(t)}\mathbb{P}(\mathbf{f}|\mathbf{X})\,\mathrm{d}\mathbf{f}}{Z_{t-1}} \\ &= \int \frac{\mathbb{P}(\mathbf{y}|\mathbf{f})^{\tau(t)}\mathbb{Q}(\mathbf{y}|\mathbf{f})^{1-\tau(t)}}{\mathbb{P}(\mathbf{y}|\mathbf{f})^{\tau(t-1)}\mathbb{Q}(\mathbf{y}|\mathbf{f})^{1-\tau(t-1)}}\frac{\mathbb{P}(\mathbf{y}|\mathbf{f})^{\tau(t-1)}\mathbb{Q}(\mathbf{y}|\mathbf{f})^{1-\tau(t-1)}\mathbb{P}(\mathbf{f}|\mathbf{X})}{Z_{t-1}}\mathrm{d}\mathbf{f} \\ &= \int \left[\frac{\mathbb{P}(\mathbf{y}|\mathbf{f})}{\mathbb{Q}(\mathbf{y}|\mathbf{f})}\right]^{\Delta\tau(t)}\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},t-1)\,\mathrm{d}\mathbf{f} \\ &\approx \frac{1}{S}\sum_{s=1}^{S}\left[\frac{\mathbb{P}(\mathbf{y}|\mathbf{f}_s)}{\mathbb{Q}(\mathbf{y}|\mathbf{f}_s)}\right]^{\Delta\tau(t)}, \qquad \mathbf{f}_s \sim \mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},t-1).\end{aligned}$$

The choice of $\mathbb{Q}(\mathbf{f})$ to be a good approximation to the true posterior makes the fraction $\mathbb{P}(\mathbf{y}|\mathbf{f})/\mathbb{Q}(\mathbf{y}|\mathbf{f})$ as constant as possible, which in turn reduces the error due to the finite step size in thermodynamical integration.

10.3 Algorithm

If only one sample $\mathbf{f}_t$ is used per temperature $\tau(t)$, the value of the entire fraction is obtained as

$$\ln\frac{Z_t}{Z_{t-1}} = \Delta\tau(t)\left[\ln\mathbb{P}(\mathbf{y}|\mathbf{f}_t) - \ln\mathbb{Q}(\mathbf{y}|\mathbf{f}_t)\right]$$

which gives rise to the full estimate

$$\ln Z \approx \sum_{t=1}^{T}\ln\frac{Z_t}{Z_{t-1}} = \ln Z_{\mathbb{Q}} + \sum_{t=1}^{T}\Delta\tau(t)\left[\ln\mathbb{P}(\mathbf{y}|\mathbf{f}_t) + \frac{1}{2}(\mathbf{f}_t-\tilde{\mathbf{m}})^{\top}\mathbf{W}(\mathbf{f}_t-\tilde{\mathbf{m}})\right]$$

for a single run r. The finite temperature change bias can be removed by combining results Z_r from R different runs by their arithmetic mean $\frac{1}{R}\sum_r Z_r$ (Neal, 2001)

$$\ln Z = \ln\int \mathbb{P}(\mathbf{y}|\mathbf{f})\,\mathbb{P}(\mathbf{f}|\mathbf{X})\,\mathrm{d}\mathbf{f} \approx \ln\left(\frac{1}{R}\sum_{r=1}^{R}Z_r\right).$$

Finally, the only primitive needed to obtain MCMC estimates of Z, $\mathbf{m}$ and $\mathbf{V}$ is an efficient sampler for the "intermediate" posterior $\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},t-1)$. We use Hybrid Monte Carlo sampling (Neal, 1993).

10.4 Results

If the posterior is very close to the prior (as in regimes 7-9 of Figure 3), it does not make a difference, which we start from. However, if the posterior can be well approximated by a Gaussian

(regimes 4-6), but is sufficiently different from the prior, then the method decreases variance and consequently improves runtimes of AIS. Different approximation methods lead also to differences in the improvement. Namely, the Laplace approximation performs worse than the approximation found by Expectation Propagation because Laplace's method approximates around the mode which can be far away from the mean.

For our evaluations of the approximations to the marginal likelihood, however we started the algorithm from the prior. Otherwise, one might be worried of biasing the MCMC simulation towards the initial distribution in cases where the chain fails to mix properly.

11. Implementation

Implementations of all methods discussed are provided at `http://www.kyb.mpg.de/~hn/approxXX.tar.gz`. The code is designed as an extension to the Gaussian Processes for Machine Learning (GPML) (Rasmussen and Williams, 2006) Matlab Code.[3] Approximate inference for Gaussian processes is done by the `binaryGP.m` function, which takes as arguments the covariance function, the likelihood function and the approximation method. The existing GPML package provides `approxLA.m` for Laplace's method and `approxEP.m` for Expectation Propagation. These implementations are generic to the likelihood function. We provide `cumGauss.m` and `logistic.m` that were designed to avoid numerical problems. In the extension, `approxKL.m`, `approxVB.m`, `approxFV.m` and `approxTAP.m` are included, among others not discussed here, for example sparse and online methods outside the scope of the current investigation. The implementations are straight-forward, although special care has been taken to avoid numerical problems e.g., situations where $\mathbf{K}$ is close to singular. More concretely, we use the well-conditioned matrix[4] $\mathbf{B} = \mathbf{W}^{\frac{1}{2}}\mathbf{K}\mathbf{W}^{\frac{1}{2}} + \mathbf{I} = \mathbf{L}\mathbf{L}^{\top}$ and its Cholesky decomposition to calculate $\mathbf{V} = \left(\mathbf{K}^{-1} + \mathbf{W}\right)^{-1}$ or $\mathbf{k}_*^{\top}\left(\mathbf{K} + \mathbf{W}^{-1}\right)^{-1}\mathbf{k}_*$. The posterior mean is represented in terms of $\boldsymbol{\alpha}$ to avoid multiplications with $\mathbf{K}^{-1}$ and facilitate predictions.

Especially LA and EP show a high level of robustness along the full spectrum of possible hyperparameters. KL uses Gauss-Hermite quadrature; we did not notice problems stemming therefrom. The FV and TAP methods work very reliably, although, we had to add a small (10^{-6}) ridge for FV to regularize $\mathbf{K}$. As a general statement, we did not observe any numerical problems for a wide range of hyperparameters reaching from reasonable values to very extreme scales.

In addition to the code for the algorithms, we provide also a tarball containing all necessary scripts to reproduce the figures of the paper. We offer two versions: The first version contains only the code for running the experiments and drawing the figures.[5] The second version additionally includes the results of the experiments.[6]

12. Experiments

The purpose of the experiments is to illustrate the strengths and weaknesses of the different approximation methods. First of all, the quality of the approximation itself in terms of posterior moments Z,

3. The package is available at `http://www.gaussianprocess.org/gpml/code`.
4. All eigenvalues λ of $\mathbf{B}$ satisfy $1 \leq \lambda \leq 1 + \frac{n}{4}\max_{ij}\mathbf{K}_{ij}$, thus $\mathbf{B}^{-1}$ and $|\mathbf{B}|$ can be safely computed.
5. The code base ($\sim$ 9Mb) can be obtained from `http://www.kyb.mpg.de/~hn/supplement_code.tar.gz`.
6. The complete code base ($\sim$ 400Mb) including all simulation results and scripts to generate figures is stored at `http://www.kyb.mpg.de/~hn/supplement_all.tar.gz`.

m and **V** is studied. At a second level, building on the "low-level" features, we compare predictive performance in terms of the predictive probability p_* given by (Equations 4 and 6):

$$p_* := \mathbb{P}(y_* = 1 | \mathbf{x}_*, \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}) \quad \approx \quad \int \mathrm{sig}(f_*) \, \mathcal{N}\left(f_* | \mu_*, \sigma_*^2\right) \mathrm{d}f_*. \tag{16}$$

On a third level, we assess higher order properties such as the information score, describing how much information the model managed to extract about the target labels, and the error rate—a binary measure of whether a test input is assigned the right class. Uncertainty predictions provided by the model are not captured by the error rate.

Accurate marginal likelihood estimates Z are a key to hyperparameter learning. In that respect, Z can be seen as a high-level feature and as the "zeroth" posterior moment at the same time.

A summary of the whole section is provided in Table 1.

12.1 Data Sets

One main idea of the paper is to study the general behavior of approximate GP classification. Our results for the different approximation methods are not specific to a particular data set but apply to a wide range of application domains. This is reflected by the choice of our reference data sets, widely used in the machine learning literature. Due to limited space, we don't include the full experiments on all data sets in this paper. However, we have verified that the same qualitative conclusions hold for all the data sets considered. The full results are available via the web.[7]

Data set	n_{train}	n_{test}	d	Brief description of problem domain
Breast	300	383	9	Breast cancer[8]
Crabs	100	100	6	Sex of Leptograpsus crabs[9]
Ionosphere	200	151	34	Classification of radar returns from the ionosphere[10]
Pima	350	418	8	Diabetes in Pima Indians[11]
Sonar	108	100	60	Sonar signals bounced by a metal or rock cylinder[12]
USPS 3 vs. 5	767	773	256	Binary sub-problem of the USPS handwritten digit data set[13]

12.2 Results

In the following, we report our experimental results covering posterior moments and predictive performance. Findings for all 5 methods are provided to make the methods as comparable as possible.

7. See links in Footnotes 5 and 6.
8. Data set at `http://mlearn.ics.uci.edu/databases/breast-cancer-wisconsin/`.
9. Data set at `http://www.stats.ox.ac.uk/pub/PRNN/`.
10. Data set at `http://mlearn.ics.uci.edu/databases/ionosphere/`.
11. Data set at `http://mlearn.ics.uci.edu/databases/pima-indians-diabetes/`.
12. Data set at `ftp://ftp.ics.uci.edu/pub/machine-learning-databases/undocumented/connectionist-bench/sonar/`.
13. Data set at `http://www.gaussianprocess.org/gpml/data/`.

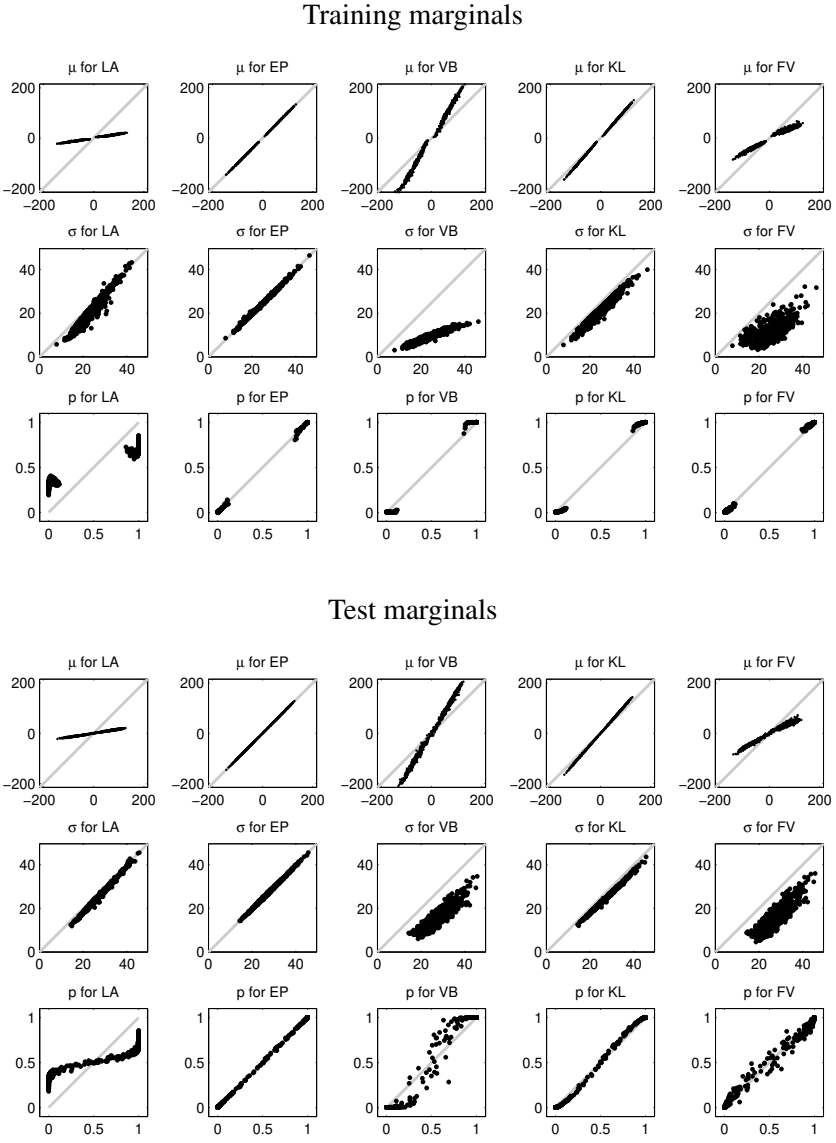


Figure 6: Marginals of USPS 3 vs. 5 for a highly non-Gaussian posterior: Each row consists of five plots showing MCMC ground truth on the x-axis and LA, EP, VB, KL and FV on the y-axis. Based on the logistic likelihood function and the squared exponential covariance function with parameters $\ln \ell = 2.25$ and $\ln \sigma_f = 4.25$ we plot the marginal means, standard deviations and resulting predictive probabilities in rows 1-3. We are working in regime 2 of Figure 3 that means the posterior is highly non-Gaussian. The upper part shows marginals of training points and the lower part shows test point marginals.

	LA	EP*	VB logit|probit	KL	FV	MCMC
idea	quadratic expansion around the mode	marginal moment matching	lower bound on indiv. likelihoods	KL minim., average w.r.t. wrong $\mathbb{Q}(\mathbf{f})$	best free-form factorial	sampling, thermo-dynamic integration
algorithm	Newton steps	iterative matching	Newton steps	Newton steps	fixed-point iteration	Hybrid MC, AIS
complexity	$O(n^3)$	$O(n^3)$	$O(n^3)$	$O(8n^3)$	$O(n^3)$	$O(n^3)$
speed	very fast	fast	fast	slow	very fast	very slow
running time	1	10	8	150	4	>500
likelihood properties	1st-3rd log. derivative	$\mathcal{N}$-integrals	lower bound	simple evaluation	$\mathcal{N}$-integrals	1st log derivative
evidence Z	–	$\approx$	– –	–	– – –	=
mean $\mathbf{m}$	– –	$\approx$	++| – –	+	–	=
covariance $\mathbf{V}$	–	$\approx$	– –	–	– –	=
information I	–	$\approx$	$\approx$| –	$\approx$	–	=
PRO	speed	practical accuracy		principled method	speed	theoretical accuracy
CON	mean$\neq$mode, low info I	speed	strong over-confidence	overconfidence	factorizing approxima-tion	very slow

Table 1: Feature summary of the considered algorithms: For each of the six algorithms under consideration, the major properties are listed in the above table. The basic idea of the method along with its computational algorithm and complexity is summarized, the requirements to the likelihood functions are given, the accuracy of evidence and moment estimates as well as information is outlined and some striking advantages and drawbacks are compared. Six relations characterize accuracy: – – – extreme underestimation, – – heavy underestimation, – underestimation, = ground truth, $\approx$ good approximation, + overestimation and ++ heavy overestimation. Running times were calculated by running each algorithm for 9 different hyperparameter regimes and both likelihoods on all data sets. An average running time per data set was calculated for each method and scaled to yield 1 for LA. In the table, the average of these numbers are shown. We are well aware of the fact, that these numbers also depend on our Matlab implementations and choices of convergence thresholds.

12.2.1 MEAN $\mathbf{m}$ AND (CO)VARIANCE $\mathbf{V}$

The posterior process, or equivalently the posterior distribution over the latent values $\mathbf{f}$, is determined by its location parameter $\mathbf{m}$ and its width parameter $\mathbf{V}$. In that respect, these two low-level quantities are the basis for all further calculations. In general, one can say that the methods show

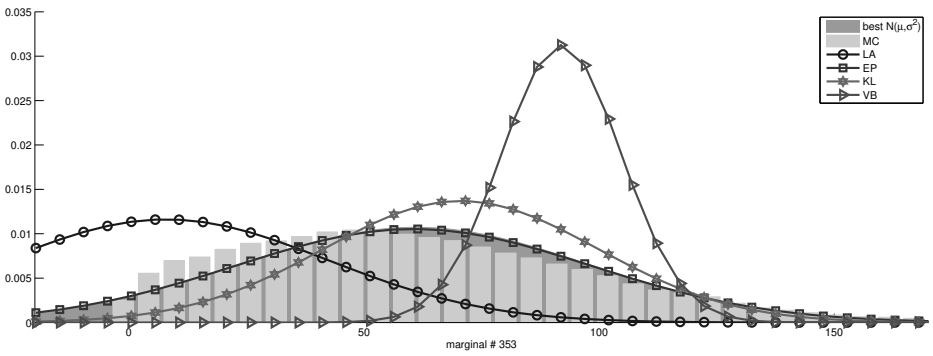


Figure 7: Marginals USPS 3 vs. 5 for digit #353 ≡ : Posterior marginals for one special training point from Figure 6 is shown. Ground truth in terms of true marginal and best Gaussian marginal (matching the moments of the true marginal) are plotted in gray, Gaussian approximations are visualized as lines. For multivariate Gaussians $\mathcal{N}(\mathbf{m},\mathbf{V})$, the i-th marginal is given by $\mathcal{N}([\mathbf{m}]_i,[\mathbf{V}]_{ii})$. Thus, the mode m_i of marginal i coincides with the i-th coordinate of the mode of the joint $[\mathbf{m}]_i$. This relation does not hold for general skewed distribution. Therefore, the marginal given by the Laplace approximation is not centered at the mode of the true marginal.

significant differences in the case of highly non-Gaussian posteriors (regimes 1-5 of Figure 3). Even in the two-dimensional toy example of Figures 4 and 5, significant differences are apparent. The means are inaccurate for LA and VB; whereas the variances are somewhat underestimated by LA and KL and severely so by VB. Marginal means $\mathbf{m}$ and variances $\mathrm{dg}(\mathbf{V})$ for USPS 3 vs. 5 are shown in Figure 6; an exemplary marginal is pictured in Figure 7 for all approximate methods and the MCMC estimate. Along the same lines, a close-to-Gaussian posterior is illustrated in Figure 8. We chose the hyperparameters for the non Gaussian case of Figure 6 to maximize the EP marginal likelihood (see Figure 9), whereas the hyperparameters of Figure 8 were selected to yield a posterior that is almost Gaussian but still has reasonable predictive performance.

The LA method has the principled weakness of expanding around the mode. In high-dimensional spaces, the mode can be very far away from the mean (Kuss and Rasmussen, 2005). The absolute value of the mean is strongly underestimated. Furthermore, the posterior is highly curved at its mode which leads to an underestimated variance, too. These effects can be seen in the first column of Figures 6 and 7, although in the close-to-Gaussian regime LA works well, Figure 8. For large latent function scales σ_f^2, in the limit $\sigma_f^2 \to \infty$, the likelihood becomes a step function, the mode approaches the origin and the curvature at the mode becomes larger. Thus the approximate posterior as found by LA becomes a zero-mean Gaussian which is much too narrow.

The EP method almost perfectly agrees with the MCMC estimates, second column of Figure 6. That means, iterative matching of approximate marginal moments leads to accurate marginal moments of the posterior.

The KL method minimizes the KL-divergence $\mathrm{KL}(\mathbb{Q}(\mathbf{f}) \parallel \mathbb{P}(\mathbf{f})) = \int \mathbb{Q}(\mathbf{f}) \ln \frac{\mathbb{Q}(\mathbf{f})}{\mathbb{P}(\mathbf{f})} \mathrm{d}\mathbf{f}$ with the average taken to the approximate distribution $\mathbb{Q}(\mathbf{f})$. The method is *zero-forcing* i.e., in regions where $\mathbb{P}(\mathbf{f})$ is very small, $\mathbb{Q}(\mathbf{f})$ has to be very small as well. In the limit that means $\mathbb{P}(\mathbf{f}) = 0 \Rightarrow \mathbb{Q}(\mathbf{f}) = 0$.

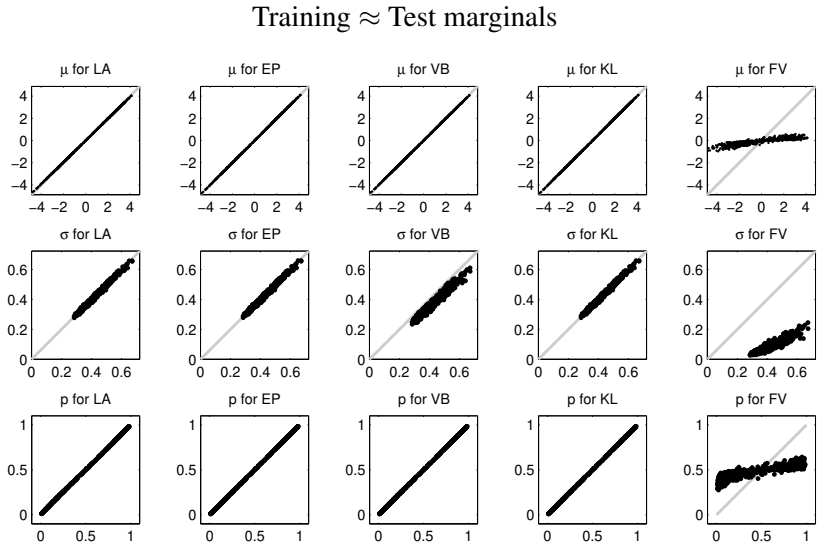


Figure 8: Marginals of USPS 3 vs. 5 for a close-to-Gaussian posterior: Using the squared exponential covariance and the logistic likelihood function with parameters $\ln \ell = 3$ and $\ln \sigma_f = 0.5$, we plot the marginal means, standard deviations and resulting predictive probabilities in rows 1-3. Only the quantities for the trainings set are shown, because the test set results are very similar. We are working in regime 8 of Figure 3 that means the posterior is of rather Gaussian shape. Each row consists of five plots showing MCMC ground truth on the x-axis and LA, EP, VB, KL and FV on the y-axis.

Thus, the support of $\mathbb{Q}(\mathbf{f})$ is smaller than the support of $\mathbb{P}(\mathbf{f})$ and hence the variance is underestimated. Typically, the posterior has a long tail away from zero as seen in Figure 3 regimes 1-5. The zero forcing property shifts the mean of the approximation away from the origin, which results in a slightly overestimated mean, fourth column of Figure 6.

Finally, the VB method can be seen as a more constrained version of the KL method with deteriorated approximation properties. The variance underestimation and mean overestimation is magnified, third column of Figure 6. Due to the required lower bounding property of each individual likelihood term, the approximate posterior has to obey severe restrictions. Especially, the lower bound to the cumulative Gaussian cannot adjust its width since the asymptotic behavior does not depend on the variational parameter (Equation 14).

The FV method has a special rôle because it does not lead to a Gaussian approximation to the posterior but to the closest (in terms of KL-divergence) factorial distribution. If the prior is quite isotropic (regimes 1,4 and 7 of Figure 3), the factorial approximation provides a reasonable approximation. If the latent function values are correlated, the approximation fails. Because of the zero forcing property, mentioned in the discussion of the KL method, both the means and the variances are underestimated. Since a factorial distribution cannot capture correlations, the effect can be severe. It is worth mentioning that there is no difference whether the posterior is close to a

Gaussian or not. In that respect, the FV method complements the LA method, which has difficulties in regimes 1, 2 and 4 of Figure 3.

12.2.2 PREDICTIVE PROBABILITY p_* AND INFORMATION SCORE I

Low-level features like posterior moments are not a goal per se, they are only needed for the purpose of calculating predictive probabilities. Figures 4 and 6 show predictive probabilities in the last row.

In principle, a bad approximation in terms of posterior moments can still provide reasonable predictions. Consider the predictive probability from Equation 16 using a cumulative Gaussian likelihood

$$p_* \quad = \quad \int \mathrm{sig}_{\mathrm{probit}}(f_*)\mathcal{N}(f_*|\mu_*,\sigma_*^2)\mathrm{d}f_* = \mathrm{sig}_{\mathrm{probit}}(\mu_*/\sqrt{1+\sigma_*^2}).$$

It is easy to see that the predictive probability p_* is constant if $\mu_*/\sqrt{1+\sigma_*^2}$ is constant. That means, moving mean μ_* and standard deviation σ_* along the hyperbolic curve $\mu_*^2/C^2-\sigma_*^2=1$, while keeping the sign of μ_* fixed, does not affect the probabilistic prediction. In the limit of large μ_* and large σ_*, rescaling does not change the prediction.

Summarizing all predictive probabilities p_i we consider the scaled information score I. As a baseline model we use the best model ignoring the inputs $\mathbf{x}_i$. This model simply returns predictions matching the class frequencies of the training set

$$B \quad = \quad -\sum_{y=\{+1,-1\}} \frac{n_{\mathrm{test}}^y}{n_{\mathrm{test}}^{+1}+n_{\mathrm{test}}^{-1}} \log_2 \frac{n_{\mathrm{train}}^y}{n_{\mathrm{train}}^{+1}+n_{\mathrm{train}}^{-1}} \leq 1[\mathrm{bit}].$$

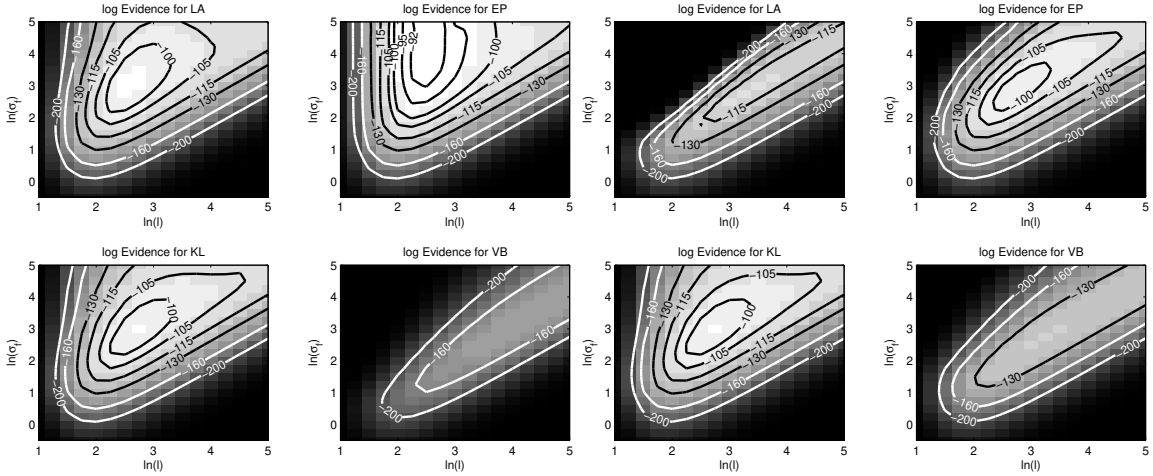
We take the difference between the baseline B (entropy) and the average negative log predictive probabilities $\log_2 \mathbb{P}(y_*|\mathbf{x}_*,\mathbf{y},\mathbf{X})$ to obtain the information score

$$I \quad = \quad B+\frac{1}{2n_{\mathrm{test}}}\sum_{i=1}^{n_{\mathrm{test}}}(1+y_i)\log_2(p_i)+(1-y_i)\log_2(1-p_i),$$

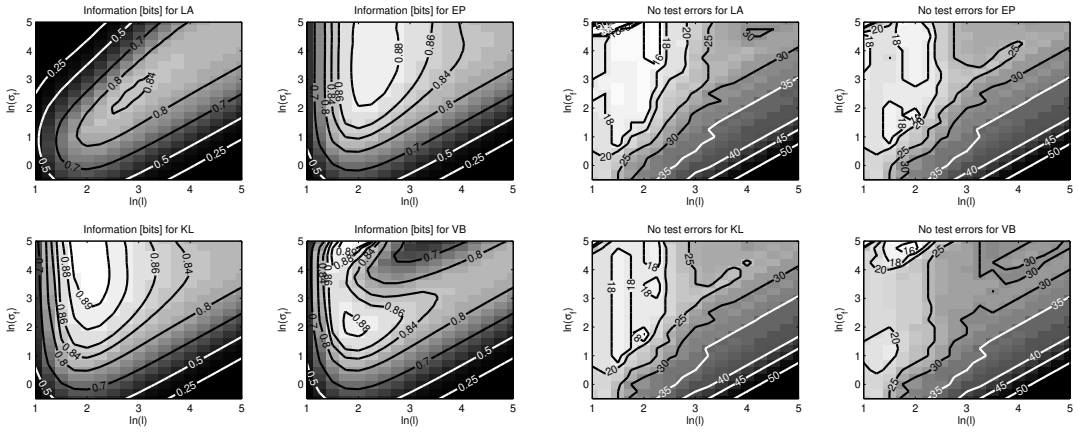
which is $1[\mathrm{bit}]$ for perfect (and confident) prediction and $0[\mathrm{bits}]$ for random guessing (for equiprobable classes). Figures 9(c), 10(middle) and 11(c) contain information scores for 5 different approximation methods on two different data sets as a function of the hyperparameters of the covariance function. According to the EP and KL plots (most prominently in Figure 11(c)), there are two strategies for a model to achieve good predictive performance:

- Find a good length scale ℓ (e.g., $\ln\ell\approx 2$) and choose a latent function scale σ_f above some threshold (e.g., $\ln\sigma_f > 3$).
- Start from a good set of hyperparameters (e.g., $\ln\ell\approx 2$, $\ln\sigma_f\approx 2$) and compensate a harder cutting likelihood ($\sigma_f^2\uparrow$) by making the data points more similar to each other ($\ell^2\uparrow$).

The LA method heavily underestimates the marginal means in the non-Gaussian regime (regimes 1-5 of Figure 3). As a consequence, the predictive probabilities are strongly under-confident in the non-Gaussian regime, first column of Figure 6. The information score's value is too small in the non-Gaussian regime, Figures 9(c) and 11(c).



(a) Evidence (b) Lower bound on evidence



(c) Information in bits (d) Number of errors

Figure 9: Evidence and classification performance for LA, EP, KL & VB on USPS 3 vs. 5: The length scale ℓ and the latent scale σ_f determine the working regime (1-9) of the Gaussian Process as drafted in Figure 3. We use the logistic likelihood and the squared exponential covariance function to classify handwritten digits. The four panels illustrate the model performance in terms of evidence, information and classification errors over the space of hyperparameters (ℓ, σ_f). For better visibility we choose a logarithmic scale of the axes. Panel (a) shows the inherent evidence approximation of the four methods and panel (b) contains the Jensen lower bound (Equation 9) on the evidence used in KL method. Both panels share the same contour levels for all four methods. Note that for the VB method, the general lower bound is a better evidence estimate than the bound provided by the method itself. Panel (c) and (d) show the information score and the number of misclassifications. One can read-off the divergence between posterior and approximation by recalling $\mathrm{KL}(\mathbb{Q}||\mathbb{P}) = \ln Z - \ln Z_B$ from Equation 10 and assuming $\ln Z_{\mathrm{EP}} \approx \ln Z$. In the figure this corresponds to subtracting Subplots (b, LA-VB) from Subplots (a, EP). Obviously, the divergence vanishes for close-to-Gaussian posteriors (regimes 3,5-6,7-9).

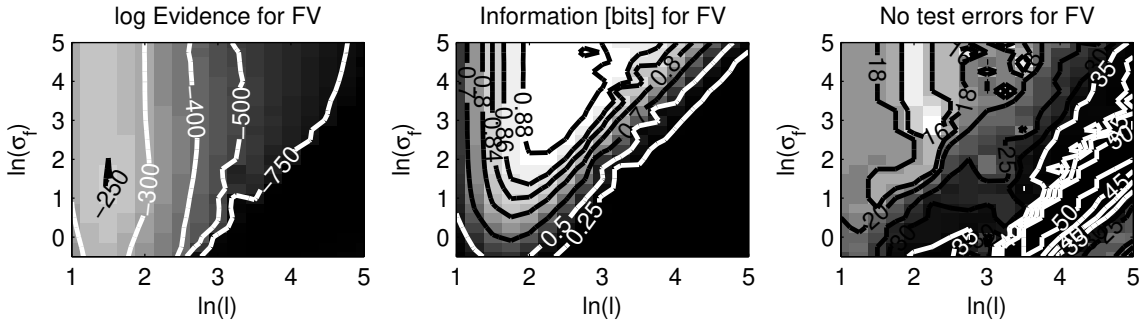


Figure 10: Evidence and classification performance for FV on USPS 3 vs. 5: The plots are a supplement to Figure 9 in that they make the factorial variational method comparable, even though we use the cumulative Gaussian likelihood. The levels of the contour lines for the information score and the number of misclassifications are the same as in Figure 9. For the marginal likelihood other contours are shown, since it has significantly different values.

Since the EP algorithm yields marginal moments very close to the MCMC estimates (second column of Figure 6), its predictive probabilities and information score is consequently also very accurate, Figures 9(c) and 11(c). The plots corresponding to EP can be seen as the quasi gold standard (Kuss and Rasmussen, 2005, Figures 4 and 5).

The KL method slightly underestimates the variance and slightly overestimates the mean which leads to slightly overconfident predictions, fourth column of Figure 6. Overconfidence, in general, leads to a degradation of the information score, however in this example, the information score is very close to the EP values and at the peak it is even slightly (0.01[bits]) higher, Figures 9(c) and 11(c).

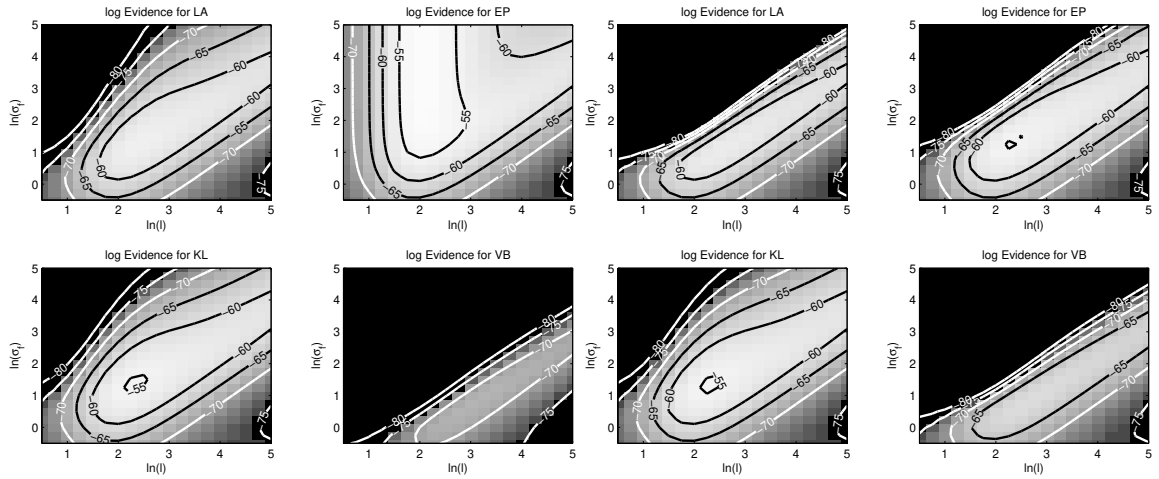
The VB method, again, has the same problems as the KL method only amplified. The predictions are overconfident, third column of Figure 6. Consequently, the information measured score in the non-Gaussian regime is too small. The logistic likelihood function (Figure 9(c)) yields much better results than the cumulative Gaussian likelihood function (Figure 11(c)).

Finally, as the FV method is accurate if the prior is isotropic, predictive probabilities and information scores are very high in regimes 1, 4 and 7 of Figure 3. For correlated priors, the FV method achieves only low information scores, Figure 10(middle). The method seems to benefit from the "hyperbolic scaling invariance" of the predictive probabilities mentioned earlier in that section because both the mean and the variance are strongly underestimated.

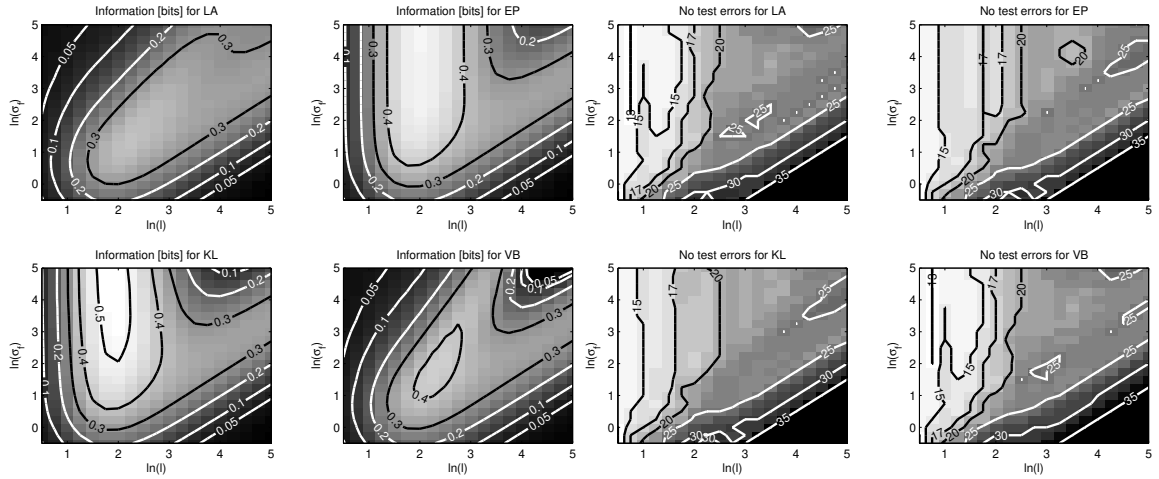
12.2.3 NUMBER OF ERRORS *E*

If one is only interested in the actual class and not in the associated confidence level, one can simply measure the number of misclassifications. Results for 5 approximation methods and 2 data sets are shown in Figures 9(d), 10(right) and 11(d).

Interestingly, all four Gaussian approximation have very similar error rates. The reason is mainly due to the fact that all methods manage to compute the right sign of the marginal mean. Only the FV method with cumulative Gaussian likelihood seems a bit problematic, even though the



(a) Evidence (b) Lower bound on evidence



(c) Information in bits (d) Number of errors

Figure 11: Evidence and classification performance for LA, EP, KL & VB on Sonar: We show the same quantities as in Figure 9, only for the Sonar Mines versus Rocks data set and using the cumulative Gaussian likelihood function.

difference is only very small. Small error rates do not imply high information scores, it is rather the other way round. In Figure 9(d) at $\ln \ell = 2$ and $\ln \sigma_f = 4$ only 16 errors are made by the LA method while the information score (Figure 9(c)) is only of 0.25[bits].

Even the FV method yields very accurate classes, having only small error rates.

12.2.4 MARGINAL LIKELIHOOD Z

Agreement of model and data is typically measured by the marginal likelihood Z. Hyperparameters can conveniently be optimized using Z not least because the gradient $\frac{\partial \ln Z}{\partial \boldsymbol{\theta}}$ can be analytically and efficiently computed for all methods. Formally, the marginal likelihood is the volume of the product of prior and likelihood. In classification, the likelihood is a product of sigmoid functions (Figure 3), so that only the orthant $\{\mathbf{f}|\mathbf{f}\odot\mathbf{y}\geq\mathbf{0}\in\mathbb{R}^n\}$ contains values $\mathbb{P}(\mathbf{f}|\mathbf{y})\geq\frac{1}{2}$. In principle, evidences are bounded by $\ln Z\leq 0$ where $\ln Z=0$ corresponds to a perfect model. As pointed out in Section 2.1.1, the marginal likelihood for a model ignoring the data and having equiprobable targets has the value $\ln Z=-n\ln 2$, which serves as a baseline.

Evidences provided by LA, EP and VB for two data sets are shown in Figures 9(a), 10(left) and 11(a). As the Jensen bound can be applied to any Gaussian approximation of the posterior, we also report it in Figures 9(b) and 11(b).

The LA method strongly underestimates the evidence in the non-Gaussian regime, because it is forced to center its approximation at the mode, Figures 9(a) and 11(a). Nevertheless, there is a good agreement between the value of the marginal likelihood and the corresponding information score. The Jensen lower bound is not tight for the LA approximation, Figures 9(b) and 11(b).

The EP method yields the highest values among all other methods. As described in Section 2.1.2, for high latent function scales σ_f^2, the model becomes effectively independent of σ_f^2. This behavior is only to be seen for the EP method, Figures 9(a) and 11(a). Again, the Jensen bound is not tight for the EP method, Figures 9(b) and 11(b). The difference between EP and MCMC marginal likelihood estimate is vanishingly small (Kuss and Rasmussen, 2005, Figures 4 and 5).

The KL method directly uses the Jensen bound (Equation 8) which can only be tight for Gaussian posterior distributions. If the posterior is very skew, the bound inherently underestimates the marginal likelihood. Therefore, Figures 9(a) and 9(b) and Figures 11(a) and 11(b) show the same values. The disagreement between information score and marginal likelihood makes hyperparameter selection based on the KL method problematic.

The VB method's lower bound on the evidence turns out to be very loose, Figures 9(a) and 11(a). Theoretically, it cannot be better than the more general Jensen bound due to the additional constraints imposed by the individual bound on each likelihood factor, Figures 9(b) and 11(b). In practice, one uses the Jensen bound for hyperparameter selection. Again, the maximum of the bound to the evidence is not very helpful for finding regions of high information score.

Finally, the FV method only yields a poor approximation to the marginal likelihood due to the factorial approximation, Figure 10. The more isotropic the model becomes (small ℓ), the tighter is the bound. For strongly correlated priors (large ℓ) the evidence drops even below the baseline $\ln Z=-n\ln 2$. Thus, the bound is not adequate to do hyperparameter selection as its maximum does not lie in regions with high information score.

12.2.5 CHOICE OF LIKELIHOOD

In the experiments, we worked with two different likelihood functions, namely the logistic and the cumulative Gaussian likelihood. The two functions differ in their slope at the origin and their asymptotic behavior. We did not find empirical evidence supporting the use of either likelihood. Theoretically, the cumulative Gaussian likelihood should be less robust against outliers due to the quadratic asymptotics. Practically, the different slopes result in a shift of the latent function length scale in the order of $\ln\frac{1}{4}-\ln\frac{1}{\sqrt{2\pi}}\approx 0.46$ on a log scale in that the logistic likelihood prefers a

bigger latent scale. Only for the VB method, differences were significant because the logistic bound is more concise. Numerically, however the cumulative Gaussian is preferable.

12.3 Results Across Data Sets

We conclude with a quantitative summary of experiments conducted on 6 data sets (breast, crabs, ionosphere, diabetes, sonar, USPS 3 vs. 5), two different likelihoods (cumulative Gaussian, logistic) and 8 covariance functions (linear, polynomial of degree 1-3, Matérn $\nu \in \{\frac{3}{2}, \frac{5}{2}\}$, squared exponential and neural network) resulting in 96 trials. All 7 approximate classification methods were trained on a 16×16 grid of hyperparameters to compare their behavior under a wide range of conditions. We calculated the maximum (over the hyperparameter grid) amount of information, every algorithm managed to extract from the data in each of the 96 trials. The table shows the number of trials, where the respective algorithm had a maximum information score that was above the mean/median (over the 7 methods).

Test \ Method	**LA**	**EP**	**KL**	**VB**	**FV**	**LR**	**TAPnaive**
# trials, information below **mean**	31	0	0	6	34	92	31
# trials, information below **median**	54	0	0	15	48	96	51

13. Conclusions

In the present paper we provide a comprehensive overview of methods for approximate Gaussian process classification. We present an exhaustive analysis of the considered algorithms using theoretical arguments. We deliver thorough empirical evidence supporting our insights revealing the strengths and weaknesses of the algorithms. Finally, we make a unified and modular implementation of all methods available to the research community.

We are able to conclude that the Expectation Propagation algorithm is, in terms of accuracy, always the method of choice, except when you cannot afford the slightly longer running time compared to the Laplace approximation.

Our comparisons include the Laplace approximation and the Expectation Propagation algorithm (Kuss and Rasmussen, 2005). We extend the latter to the logistic likelihood. We apply Kullback-Leibler divergence minimization to Gaussian process classification and derive an efficient Newton algorithm. Although the principles behind this method have been known for some time, we are unaware that this method has been previously implemented for GPs in practise. The existing variational method (Gibbs and MacKay, 2000) is extended by a lower bound on the cumulative Gaussian likelihood and we provide an implementation based on Newton's method. Furthermore, we give a detailed analysis of the Factorial Variational method (Csató et al., 2000).

All methods are considered in a common framework, approximation quality is assessed, predictive performance is measured and model selection is benchmarked.

In practice, an approximation method has to satisfy a wide range of requirements. If **runtime** is the major concern or one is interested in **error rate** only, the Laplace approximation or label regression should be considered. Only Expectation Propagation and—although a lot slower—the KL-method deliver accurate **marginals** as well as reliable **class probabilities** and allow for faithful **model selection**.

If an application demands a **non-standard likelihood** function, this also affects the choice of the algorithm: The Laplace approximation requires derivatives, Expectation Propagation and the

Factorial Variational method need integrability with respect to Gaussian measures. However, the KL-method simply needs to evaluate the likelihood and known lower bounds naturally lead to the VB algorithm.

Finally, if the classification problem contains a lot of **label noise** (σ_f is small), the exact posterior distribution is effectively close to Gaussian. In that case, the choice of the approximation method is not crucial since in the Gaussian regime, they will give the same answer. For weakly coupled training data, the Factorial Variational method can lead to quite reasonable approximations.

As a future goal remains an in-depth understanding of the properties of sparse and online approximations to the posterior and a coverage of a broader range of covariance functions. Also, the approximation techniques discussed can be applied to other non-Gaussian inference problems besides the narrow applications to binary GP classification discussed here, and there is hope that some of the insights presented may be useful more generally.

Acknowledgments

Thanks to Manfred Opper for pointing us initially to the practical possibility of the KL method and the three anonymous reviewers.

Appendix A. Derivatives

In the following, we provide the expressions for the derivatives needed to implement the VB and the KL method.

A.1 Derivatives for VB

Some notational remarks. Partial derivatives w.r.t. one single parameter such as $\frac{\partial \mathbf{A}_\varsigma}{\partial \varsigma_i}$ or $\frac{\partial \mathbf{b}_\varsigma}{\partial \varsigma_i}$ stay matrices or vectors, respectively. Lowercase letters $\{\mathbf{a},\mathbf{b},\mathbf{c}\}_\varsigma$ indicate vectors, upper case letters $\{\mathbf{A},\mathbf{B},\mathbf{C}\}_\varsigma$ stand for the corresponding diagonal matrices with the vector as diagonal. The dot notation applies to both lower and uppercase letters and denote derivatives w.r.t. the variational parameter vector ς

$$\begin{aligned}
\dot{\mathbf{a}}_\varsigma &:= \left[\frac{\partial a_{\varsigma_i}}{\partial \varsigma_i}\right]_i = \frac{\partial \mathbf{a}_\varsigma}{\partial \varsigma}, \text{ vector,} \\
\ddot{\mathbf{a}}_\varsigma &:= \left[\frac{\partial^2 a_{\varsigma_i}}{\partial \varsigma_i^2}\right]_i = \frac{\partial^2 \mathbf{a}_\varsigma}{\partial \varsigma^2}, \text{ vector,} \\
\dot{\mathbf{A}}_\varsigma &:= \mathrm{Dg}\left(\dot{\mathbf{a}}_\varsigma\right).
\end{aligned}$$

The operators $\mathrm{Dg} : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ and $\mathrm{dg} : \mathbb{R}^{n\times n} \to \mathbb{R}^n$ manipulate matrix diagonals. The result of $\mathrm{Dg}(\mathbf{x})$ is a diagonal matrix $\mathbf{X}$ containing $\mathbf{x}$ as diagonal, whereas $\mathrm{dg}(\mathbf{X})$ returns the diagonal of $\mathbf{X}$ as a vector. Hence, we have $\mathrm{Dg}\left(\mathrm{dg}(\mathbf{x})\right) = \mathbf{x}$, but in general $\mathrm{dg}\left(\mathrm{Dg}(\mathbf{X})\right) = \mathbf{X}$ does only hold true for diagonal matrices.

A.1.1 SOME SHORTCUTS USED LATER ONWARDS

$$
\begin{aligned}
\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}} &:= \left(\mathbf{K}^{-1}-2\mathbf{A}_{\boldsymbol{\varsigma}}\right)^{-1} \overset{\mathbf{K} \text{ small}}{=} \mathbf{K}-\mathbf{K}\left(\mathbf{K}-\frac{1}{2}\mathbf{A}_{\boldsymbol{\varsigma}}^{-1}\right)^{-1}\mathbf{K},\\
\tilde{\mathbf{b}}_{\boldsymbol{\varsigma}} &:= \mathrm{Dg}(\mathbf{y})\mathbf{b}_{\boldsymbol{\varsigma}} = \mathbf{y}\odot\mathbf{b}_{\boldsymbol{\varsigma}},\\
\mathbf{l}_{\boldsymbol{\varsigma}} &:= \tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\tilde{\mathbf{b}}_{\boldsymbol{\varsigma}} = \left(\mathbf{K}^{-1}-2\mathbf{A}_{\boldsymbol{\varsigma}}\right)^{-1}\left(\mathbf{y}\odot\mathbf{b}_{\boldsymbol{\varsigma}}\right),\\
\frac{\partial \mathbf{l}_{\boldsymbol{\varsigma}}}{\partial \varsigma_j} &= \tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\left(2\frac{\partial \mathbf{A}_{\boldsymbol{\varsigma}}}{\partial \varsigma_j}\mathbf{l}_{\boldsymbol{\varsigma}}+\mathbf{y}\odot\frac{\partial \mathbf{b}_{\boldsymbol{\varsigma}}}{\partial \varsigma_j}\right),\\
\frac{\partial \mathbf{l}_{\boldsymbol{\varsigma}}}{\partial \theta_i} &= \tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\mathbf{K}^{-1}\frac{\partial \mathbf{K}}{\partial \theta_i}\mathbf{K}^{-1}\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\left(\mathbf{y}\odot\mathbf{b}_{\boldsymbol{\varsigma}}\right),\\
\dot{\mathbf{L}}_{\boldsymbol{\varsigma}} &:= \frac{\partial \mathbf{l}_{\boldsymbol{\varsigma}}}{\partial \boldsymbol{\varsigma}^{\top}} = \tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\left(2\mathrm{Dg}(\mathbf{l}_{\boldsymbol{\varsigma}})\dot{\mathbf{A}}_{\boldsymbol{\varsigma}}+\mathrm{Dg}(\mathbf{y})\dot{\mathbf{B}}_{\boldsymbol{\varsigma}}\right),\\
\mathbf{r}_{\boldsymbol{\varsigma}} &:= \dot{\mathbf{b}}_{\boldsymbol{\varsigma}}\odot\mathbf{y}\odot\mathbf{l}_{\boldsymbol{\varsigma}}+\mathrm{dg}\left(\mathbf{l}_{\boldsymbol{\varsigma}}\mathbf{l}_{\boldsymbol{\varsigma}}^{\top}\dot{\mathbf{A}}_{\boldsymbol{\varsigma}}\right)\\
&= \dot{\mathbf{b}}_{\boldsymbol{\varsigma}}\odot\mathbf{y}\odot\mathbf{l}_{\boldsymbol{\varsigma}}+\mathbf{l}_{\boldsymbol{\varsigma}}\odot\mathbf{l}_{\boldsymbol{\varsigma}}\odot\dot{\mathbf{a}}_{\boldsymbol{\varsigma}},\\
\frac{\partial \mathbf{r}_{\boldsymbol{\varsigma}}}{\partial \varsigma_j} &= \mathbf{y}\odot\mathbf{l}_{\boldsymbol{\varsigma}}\odot\frac{\partial \dot{\mathbf{b}}_{\boldsymbol{\varsigma}}}{\partial \varsigma_j}+\dot{\mathbf{b}}_{\boldsymbol{\varsigma}}\odot\mathbf{y}\odot\frac{\partial \mathbf{l}_{\boldsymbol{\varsigma}}}{\partial \varsigma_j}+2\mathbf{l}_{\boldsymbol{\varsigma}}\odot\dot{\mathbf{a}}_{\boldsymbol{\varsigma}}\odot\frac{\partial \mathbf{l}_{\boldsymbol{\varsigma}}}{\partial \varsigma_j}+\mathbf{l}_{\boldsymbol{\varsigma}}\odot\mathbf{l}_{\boldsymbol{\varsigma}}\odot\frac{\partial \dot{\mathbf{a}}_{\boldsymbol{\varsigma}}}{\partial \varsigma_j},\\
\dot{\mathbf{R}}_{\boldsymbol{\varsigma}} &:= \frac{\partial \mathbf{r}_{\boldsymbol{\varsigma}}}{\partial \boldsymbol{\varsigma}^{\top}} = \mathrm{Dg}\left(\mathbf{y}\odot\dot{\mathbf{b}}_{\boldsymbol{\varsigma}}+2\mathbf{l}_{\boldsymbol{\varsigma}}\odot\dot{\mathbf{a}}_{\boldsymbol{\varsigma}}\right)\dot{\mathbf{L}}_{\boldsymbol{\varsigma}}+\mathrm{Dg}\left(\mathbf{l}_{\boldsymbol{\varsigma}}\odot\left(\mathbf{y}\odot\ddot{\mathbf{b}}_{\boldsymbol{\varsigma}}+\mathbf{l}_{\boldsymbol{\varsigma}}\odot\ddot{\mathbf{a}}_{\boldsymbol{\varsigma}}\right)\right)\\
&= \mathrm{Dg}\left(\mathbf{y}\odot\dot{\mathbf{b}}_{\boldsymbol{\varsigma}}+2\mathbf{l}_{\boldsymbol{\varsigma}}\odot\dot{\mathbf{a}}_{\boldsymbol{\varsigma}}\right)\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\mathrm{Dg}\left(\mathbf{y}\odot\dot{\mathbf{b}}_{\boldsymbol{\varsigma}}+2\mathbf{l}_{\boldsymbol{\varsigma}}\odot\dot{\mathbf{a}}_{\boldsymbol{\varsigma}}\right)+\mathrm{Dg}\left(\mathbf{l}_{\boldsymbol{\varsigma}}\odot\left(\mathbf{y}\odot\ddot{\mathbf{b}}_{\boldsymbol{\varsigma}}+\mathbf{l}_{\boldsymbol{\varsigma}}\odot\ddot{\mathbf{a}}_{\boldsymbol{\varsigma}}\right)\right).
\end{aligned}
$$

A.1.2 FIRST DERIVATIVES W.R.T. VARIATIONAL PARAMETERS ς_i YIELDING THE GRADIENT

$$
\begin{aligned}
\ln Z_B &= \mathbf{c}_{\boldsymbol{\varsigma}}^{\top}\mathbb{1}+\frac{1}{2}\tilde{\mathbf{b}}_{\boldsymbol{\varsigma}}^{\top}\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\tilde{\mathbf{b}}_{\boldsymbol{\varsigma}}-\frac{1}{2}\ln\left|\mathbf{I}-2\mathbf{A}_{\boldsymbol{\varsigma}}\mathbf{K}\right|,\\
\frac{\partial \ln Z_B}{\partial \varsigma_i} &= \frac{\partial c_i}{\partial \varsigma_i}+\tilde{\mathbf{b}}_{\boldsymbol{\varsigma}}^{\top}\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\left[\mathbf{y}\odot\frac{\partial \mathbf{b}_{\boldsymbol{\varsigma}}}{\partial \varsigma_i}+\frac{\partial \mathbf{A}_{\boldsymbol{\varsigma}}}{\partial \varsigma_i}\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\tilde{\mathbf{b}}_{\boldsymbol{\varsigma}}\right]+\mathrm{tr}\left(\left(\mathbf{I}-2\mathbf{A}_{\boldsymbol{\varsigma}}\mathbf{K}\right)^{-\top}\mathbf{K}\frac{\partial \mathbf{A}_{\boldsymbol{\varsigma}}}{\partial \varsigma_i}\right)\\
&\overset{\mathbf{l}_{\boldsymbol{\varsigma}},\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}}{=} \frac{\partial c_i}{\partial \varsigma_i}+\mathbf{l}_{\boldsymbol{\varsigma}}^{\top}\left[\mathbf{y}\odot\frac{\partial \mathbf{b}_{\boldsymbol{\varsigma}}}{\partial \varsigma_i}+\frac{\partial \mathbf{A}_{\boldsymbol{\varsigma}}}{\partial \varsigma_i}\mathbf{l}_{\boldsymbol{\varsigma}}\right]+\mathrm{tr}\left(\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\frac{\partial \mathbf{A}_{\boldsymbol{\varsigma}}}{\partial \varsigma_i}\right),\\
\frac{\partial \ln Z_B}{\partial \boldsymbol{\varsigma}} &= \left[\frac{\partial c_i}{\partial \varsigma_i}\right]_i+\dot{\mathbf{b}}_{\boldsymbol{\varsigma}}\odot\mathbf{y}\odot\left(\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\tilde{\mathbf{b}}_{\boldsymbol{\varsigma}}\right)+\mathrm{dg}\left(\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\tilde{\mathbf{b}}_{\boldsymbol{\varsigma}}\tilde{\mathbf{b}}_{\boldsymbol{\varsigma}}^{\top}\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\dot{\mathbf{A}}_{\boldsymbol{\varsigma}}\right)+\mathrm{dg}\left(\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\dot{\mathbf{A}}_{\boldsymbol{\varsigma}}\right)\\
&\overset{\mathbf{l}_{\boldsymbol{\varsigma}}}{=} \left[\frac{\partial c_i}{\partial \varsigma_i}\right]_i+\dot{\mathbf{b}}_{\boldsymbol{\varsigma}}\odot\mathbf{y}\odot\mathbf{l}_{\boldsymbol{\varsigma}}+\mathrm{dg}\left(\mathbf{l}_{\boldsymbol{\varsigma}}\mathbf{l}_{\boldsymbol{\varsigma}}^{\top}\dot{\mathbf{A}}_{\boldsymbol{\varsigma}}\right)+\mathrm{dg}\left(\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\dot{\mathbf{A}}_{\boldsymbol{\varsigma}}\right)\\
&\overset{\mathbf{r}_{\boldsymbol{\varsigma}}}{=} \left[\frac{\partial c_i}{\partial \varsigma_i}\right]_i+\mathbf{r}_{\boldsymbol{\varsigma}}+\mathrm{dg}\left(\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\dot{\mathbf{A}}_{\boldsymbol{\varsigma}}\right)\\
&= \dot{\mathbf{c}}_{\boldsymbol{\varsigma}}+\mathbf{l}_{\boldsymbol{\varsigma}}\odot\left(\dot{\mathbf{b}}_{\boldsymbol{\varsigma}}\odot\mathbf{y}+\mathbf{l}_{\boldsymbol{\varsigma}}\odot\dot{\mathbf{a}}_{\boldsymbol{\varsigma}}\right)+\mathrm{dg}\left(\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\right)\odot\dot{\mathbf{a}}_{\boldsymbol{\varsigma}}.
\end{aligned}
$$

A.1.3 SECOND DERIVATIVES W.R.T. VARIATIONAL PARAMETERS ς_i YIELDING THE HESSIAN

$$\begin{aligned}
\frac{\partial^2 \ln Z_B}{\partial \varsigma_j \partial \varsigma_i} &= \frac{\partial^2 c_i}{\partial \varsigma_j \partial \varsigma_i} + \frac{\partial \mathbf{r}_{\boldsymbol{\varsigma},i}}{\partial \varsigma_j} + \operatorname{tr}\left(2\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}} \frac{\partial \mathbf{A}_{\boldsymbol{\varsigma}}}{\partial \varsigma_j} \tilde{\mathbf{K}}_{\boldsymbol{\varsigma}} \frac{\partial \mathbf{A}_{\boldsymbol{\varsigma}}}{\partial \varsigma_i} + \tilde{\mathbf{K}}_{\boldsymbol{\varsigma}} \frac{\partial^2 \mathbf{A}_{\boldsymbol{\varsigma}}}{\partial \varsigma_j \partial \varsigma_i}\right), \\
\frac{\partial^2 \ln Z_B}{\partial \boldsymbol{\varsigma} \partial \boldsymbol{\varsigma}^\top} &= \left[\frac{\partial^2 c_i}{\partial \varsigma_i^2}\right]_{ii} + \frac{\partial \mathbf{r}_{\boldsymbol{\varsigma}}}{\partial \boldsymbol{\varsigma}^\top} + 2\left(\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}} \dot{\mathbf{A}}_{\boldsymbol{\varsigma}}\right) \odot \left(\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}} \dot{\mathbf{A}}_{\boldsymbol{\varsigma}}\right)^\top + \operatorname{Dg}\left(\operatorname{dg}(\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}) \odot \ddot{\mathbf{a}}_{\boldsymbol{\varsigma}}\right) \\
&= \ddot{\mathbf{C}}_{\boldsymbol{\varsigma}} + \ddot{\mathbf{R}}_{\boldsymbol{\varsigma}} + 2\left(\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}} \dot{\mathbf{A}}_{\boldsymbol{\varsigma}}\right) \odot \left(\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}} \dot{\mathbf{A}}_{\boldsymbol{\varsigma}}\right)^\top + \operatorname{Dg}\left(\operatorname{dg}(\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}) \odot \ddot{\mathbf{a}}_{\boldsymbol{\varsigma}}\right).
\end{aligned}$$

A.1.4 MIXED DERIVATIVES W.R.T. HYPER- θ_i AND VARIATIONAL PARAMETERS ς_i

$$\begin{aligned}
\frac{\partial^2 \ln Z_B}{\partial \theta_i \partial \boldsymbol{\varsigma}} &= \dot{\mathbf{a}}_{\boldsymbol{\varsigma}} \odot \frac{\partial}{\partial \theta_i}\left(\mathbf{l}_{\boldsymbol{\varsigma}} \odot \mathbf{l}_{\boldsymbol{\varsigma}} + \operatorname{dg}\left(\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\right)\right) + \dot{\mathbf{b}}_{\boldsymbol{\varsigma}} \odot \mathbf{y} \odot \frac{\partial \mathbf{l}_{\boldsymbol{\varsigma}}}{\partial \theta_i} \\
&= \dot{\mathbf{a}}_{\boldsymbol{\varsigma}} \odot \left(2\mathbf{l}_{\boldsymbol{\varsigma}} \odot \frac{\partial \mathbf{l}_{\boldsymbol{\varsigma}}}{\partial \theta_i} + \operatorname{dg}\left(\tilde{\mathbf{K}}_{\boldsymbol{\varsigma}} \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \theta_i} \mathbf{K}^{-1} \tilde{\mathbf{K}}_{\boldsymbol{\varsigma}}\right)\right) + \dot{\mathbf{b}}_{\boldsymbol{\varsigma}} \odot \mathbf{y} \odot \frac{\partial \mathbf{l}_{\boldsymbol{\varsigma}}}{\partial \theta_i}.
\end{aligned}$$

A.1.5 FIRST DERIVATIVES W.R.T. HYPERPARAMETERS θ_i:

For a gradient optimization with respect to $\boldsymbol{\theta}$, we need the gradient of the objective $\partial \ln Z_B / \partial \boldsymbol{\theta}$. Naïvely, the gradient is given by:

$$\begin{aligned}
\frac{\partial \ln Z_B}{\partial \theta_i} &= \frac{1}{2} \tilde{\mathbf{b}}_{\boldsymbol{\varsigma}}^\top \tilde{\mathbf{K}}_{\boldsymbol{\varsigma}} \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \theta_i} \mathbf{K}^{-1} \tilde{\mathbf{K}}_{\boldsymbol{\varsigma}} \tilde{\mathbf{b}}_{\boldsymbol{\varsigma}} + \operatorname{tr}\left(\left(\mathbf{I} - 2\mathbf{A}_{\boldsymbol{\varsigma}} \mathbf{K}\right)^{-\top} \mathbf{A}_{\boldsymbol{\varsigma}} \frac{\partial \mathbf{K}}{\partial \theta_i}\right) \\
&\overset{\mathbf{l}_{\boldsymbol{\varsigma}}}{=} \frac{1}{2} \mathbf{l}_{\boldsymbol{\varsigma}}^\top \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \theta_i} \mathbf{K}^{-1} \mathbf{l}_{\boldsymbol{\varsigma}} + \operatorname{tr}\left(\left(\mathbf{I} - 2\mathbf{A}_{\boldsymbol{\varsigma}} \mathbf{K}\right)^{-\top} \mathbf{A}_{\boldsymbol{\varsigma}} \frac{\partial \mathbf{K}}{\partial \theta_i}\right).
\end{aligned}$$

However, the optimal variational parameter $\boldsymbol{\varsigma}^*$ depends implicitly on the actual choice of $\boldsymbol{\theta}$ and one has to account for that in the derivative by adding an extra "implicit" term

$$\left.\frac{\partial \ln Z_B(\boldsymbol{\theta}, \boldsymbol{\varsigma})}{\partial \theta_i}\right|_{\boldsymbol{\varsigma}=\boldsymbol{\varsigma}^*} = \frac{\partial \ln Z_B(\boldsymbol{\theta}, \boldsymbol{\varsigma}^*)}{\partial \theta_i} + \sum_{j=1}^{n} \frac{\partial \ln Z_B(\boldsymbol{\theta}, \boldsymbol{\varsigma}^*)}{\partial \varsigma_j^*} \frac{\partial \varsigma_j^*}{\partial \theta_i}.$$

The question of how to find an expression for $\frac{\partial \boldsymbol{\varsigma}^*}{\partial \boldsymbol{\theta}}$ can be solved by means of the implicit function theorem for continuous and differentiable functions $\mathbf{F}$:

$$\mathbf{F} : \mathbb{R}^p \times \mathbb{R}^n \to \mathbb{R}^n, \quad \mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0} \quad \Rightarrow \quad \frac{\partial \mathbf{y}}{\partial \mathbf{x}}(\mathbf{x}) = -\left(\frac{\partial \mathbf{F}}{\partial \mathbf{y}}(\mathbf{x}, \mathbf{y}(\mathbf{x}))\right)^{-1} \frac{\partial \mathbf{F}}{\partial \mathbf{x}}(\mathbf{x}, \mathbf{y}(\mathbf{x})) \text{ if } \mathbf{F}(\mathbf{x}, \mathbf{y}(\mathbf{x})) = \mathbf{0}.$$

Setting $\mathbf{F}(\mathbf{x}, \mathbf{y}) \equiv \frac{\partial \ln Z_B}{\partial \boldsymbol{\varsigma}}(\boldsymbol{\theta}, \boldsymbol{\varsigma})$ leads to

$$\frac{\partial \boldsymbol{\varsigma}_\theta^*}{\partial \boldsymbol{\theta}^\top} = -\left(\frac{\partial^2 \ln Z_B(\boldsymbol{\theta}, \boldsymbol{\varsigma}_\theta^*)}{\partial \boldsymbol{\varsigma} \partial \boldsymbol{\varsigma}^\top}\right)^{-1} \frac{\partial^2 \ln Z_B(\boldsymbol{\theta}, \boldsymbol{\varsigma}_\theta^*)}{\partial \boldsymbol{\theta}^\top \partial \boldsymbol{\varsigma}}$$

and in turn combines to

$$\left.\frac{\partial \ln Z_B}{\partial \theta_i}\right|_{\boldsymbol{\varsigma}=\boldsymbol{\varsigma}^*} = \frac{\partial \ln Z_B}{\partial \theta_i} - \left(\frac{\partial \ln Z_B}{\partial \boldsymbol{\varsigma}}\right)^\top \left(\frac{\partial^2 \ln Z_B}{\partial \boldsymbol{\varsigma} \partial \boldsymbol{\varsigma}^\top}\right)^{-1} \frac{\partial^2 \ln Z_B}{\partial \theta_i \partial \boldsymbol{\varsigma}}$$

where all terms are known.

A.2 Derivatives for KL

The lower bound $\ln Z_B$ to the log marginal likelihood $\ln Z$ is given by Equation 9 as

$$\ln Z \quad \geq \quad = \ln Z_B(\mathbf{m},\mathbf{V}) = a(\mathbf{y},\mathbf{m},\mathbf{V}) + \frac{1}{2}\ln\left|\mathbf{V}\mathbf{K}^{-1}\right| + \frac{n}{2} - \frac{1}{2}\mathbf{m}^\top\mathbf{K}^{-1}\mathbf{m} - \frac{1}{2}\mathrm{tr}\left(\mathbf{V}\mathbf{K}^{-1}\right)$$

where we used the shortcut $a(\mathbf{y},\mathbf{m},\mathbf{V}) = \sum_{i=1}^n \int \mathcal{N}(f_i|m_i,v_{ii}) \ln \mathrm{sig}(y_i f_i) \mathrm{d}f_i$. As a first step, we calculate the first derivatives of $\ln Z_B$ with respect to the posterior moments $\mathbf{m}$ and $\mathbf{V}$ to derive necessary conditions for the optimum by equating them with zero:

$$\begin{aligned}
\frac{\partial \ln Z_B}{\partial \mathbf{V}} = \frac{\partial a(\mathbf{y},\mathbf{m},\mathbf{V})}{\partial \mathbf{V}} + \frac{1}{2}\mathbf{V}^{-1} - \frac{1}{2}\mathbf{K}^{-1} \stackrel{!}{=} \mathbf{0} \quad &\Rightarrow \quad \mathbf{V} = \left(\mathbf{K}^{-1} - 2\mathrm{Dgdg}\frac{\partial a}{\partial \mathbf{V}}\right)^{-1}, \\
\frac{\partial \ln Z_B}{\partial \mathbf{m}} = \frac{\partial a(\mathbf{y},\mathbf{m},\mathbf{V})}{\partial \mathbf{m}} - \mathbf{K}^{-1}\mathbf{m} \stackrel{!}{=} \mathbf{0} \quad &\Rightarrow \quad \mathbf{m} = \mathbf{K}\frac{\partial a}{\partial \mathbf{m}}.
\end{aligned}$$

These two expressions are plugged in the original expression for $\ln Z_B$ using $\mathbf{A} = (\mathbf{I} - 2\mathbf{K}\mathbf{\Lambda})^{-1}$ and $\mathbf{\Lambda} = \mathrm{Dgdg}\frac{\partial a}{\partial \mathbf{V}}$ to yield:

$$\ln Z_B(\boldsymbol{\alpha},\mathbf{\Lambda}) \quad = \quad a\left(\mathbf{y},\mathbf{K}\boldsymbol{\alpha},(\mathbf{K}^{-1}-2\mathbf{\Lambda})^{-1}\right) + \frac{1}{2}\ln|\mathbf{A}| - \frac{1}{2}\mathrm{tr}\mathbf{A} + \frac{n}{2} - \frac{1}{2}\boldsymbol{\alpha}^\top\mathbf{K}\boldsymbol{\alpha}.$$

Our algorithm uses the parameters $\boldsymbol{\alpha}$, $\mathbf{\Lambda}$, so we calculate first and second derivatives to implement Newton's method.

A.2.1 FIRST DERIVATIVES W.R.T. PARAMETERS $\boldsymbol{\alpha}$, $\mathbf{\Lambda}$ YIELDING THE GRADIENT

$$\frac{\partial \ln Z_B}{\partial \boldsymbol{\lambda}} = \frac{\partial a}{\partial \boldsymbol{\lambda}} + \mathrm{dg}(\mathbf{V}) - \mathrm{dg}(\mathbf{V}\mathbf{A}^\top) \quad \text{and} \quad \frac{\partial \ln Z_B}{\partial \boldsymbol{\alpha}} = \frac{\partial a}{\partial \boldsymbol{\alpha}} - \mathbf{K}\boldsymbol{\alpha}.$$

Only the terms containing derivatives of a need further attention, namely

$$\frac{\partial a}{\partial \boldsymbol{\alpha}} = \mathbf{K}\frac{\partial a}{\partial \mathbf{m}} \qquad \text{and}$$

$$\begin{aligned}
\mathrm{d}(\mathrm{dg}\mathbf{V}) \quad &= \quad \mathrm{dg}\left[\mathrm{d}\left(\mathbf{K}^{-1}-2\mathbf{\Lambda}\right)^{-1}\right] = 2\mathrm{dg}\left[\mathbf{V}\,\mathrm{d}\mathbf{\Lambda}\,\mathbf{V}\right] = 2\mathrm{dg}\left[\sum_k \mathbf{v}_k\mathbf{v}_k^\top \mathrm{d}\lambda_k\right] = 2\sum_k (\mathbf{v}_k \odot \mathbf{v}_k)\,\mathrm{d}\lambda_k \\
&= \quad 2\,(\mathbf{V}\odot\mathbf{V})\,\mathrm{d}\boldsymbol{\lambda} \Rightarrow \frac{\partial \mathrm{dg}\mathbf{V}}{\partial \boldsymbol{\lambda}^\top} = 2\mathbf{V}\odot\mathbf{V}, \\
\frac{\partial a}{\partial \boldsymbol{\lambda}} \quad &= \quad 2(\mathbf{V}\odot\mathbf{V})\frac{\partial a(\mathbf{y},\mathbf{m},\mathbf{V})}{\partial \mathrm{dg}\mathbf{V}}.
\end{aligned}$$

As a last step, the derivatives w.r.t. $\mathbf{m}$ and the diagonal part of $\mathbf{V}$ yield

$$\begin{aligned}
\frac{\partial a}{\partial m_i} &= \int \frac{\partial \mathcal{N}(f|m_i,v_{ii})}{\partial m_i} \ln \mathrm{sig}(y_i f)\mathrm{d}f = \int \frac{f-m_i}{v_{ii}} \mathcal{N}(f|m_i,v_{ii}) \ln \mathrm{sig}(y_i f)\mathrm{d}f \\
&= \frac{1}{\sqrt{v_{ii}}} \int f \cdot \mathcal{N}(f) \ln \mathrm{sig}\left(\sqrt{v_{ii}} y_i f + m_i y_i\right) \mathrm{d}f, \\
\frac{\partial a}{\partial v_{ii}} &= \int \frac{\partial \mathcal{N}(f|m_i,v_{ii})}{\partial v_{ii}} \ln \mathrm{sig}(y_i f)\mathrm{d}f = \int \left(\frac{(f-m_i)^2}{v_{ii}^{\frac{3}{2}}} - \frac{1}{\sqrt{v_{ii}}} \right) \mathcal{N}(f|m_i,v_{ii}) \ln \mathrm{sig}(y_i f)\mathrm{d}f \\
&= \frac{1}{2v_{ii}} \int \left(f^2-1\right) \cdot \mathcal{N}(f) \ln \mathrm{sig}\left(\sqrt{v_{ii}} y_i f + m_i y_i\right) \mathrm{d}f.
\end{aligned}$$

A.2.2 SECOND DERIVATIVES W.R.T. PARAMETERS $\boldsymbol{\alpha}$, $\boldsymbol{\Lambda}$ YIELDING THE HESSIAN

Again, we proceed in two steps, calculating derivatives w.r.t. $\boldsymbol{\alpha}$ and $\boldsymbol{\Lambda}$ and by the chain rule compute those w.r.t. $\mathbf{m}$ and $\mathbf{V}$.

$$\begin{aligned}
\frac{\partial^2 \ln Z_B}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^\top} &= \frac{\partial^2 a}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^\top} + \mathbf{K} = \frac{\partial}{\partial \boldsymbol{\alpha}} \left[\frac{\partial a}{\partial \mathbf{m}^\top} \frac{\partial \mathbf{m}}{\partial \boldsymbol{\alpha}^\top} \right] + \mathbf{K} = \frac{\partial}{\partial \boldsymbol{\alpha}} \left[\frac{\partial a}{\partial \mathbf{m}^\top} \mathbf{K} \right] + \mathbf{K} \\
&= \frac{\partial}{\partial \boldsymbol{\alpha}} \left[\frac{\partial a}{\partial \mathbf{m}^\top} \right] \mathbf{K} + \mathbf{K} = \frac{\partial \mathbf{m}^\top}{\partial \boldsymbol{\alpha}} \frac{\partial}{\partial \mathbf{m}} \left[\frac{\partial a}{\partial \mathbf{m}^\top} \right] \mathbf{K} + \mathbf{K} \\
&= \mathbf{K} \frac{\partial^2 a}{\partial \mathbf{m} \partial \mathbf{m}^\top} \mathbf{K} + \mathbf{K}, \\
\frac{\partial^2 \ln Z_B}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\alpha}^\top} &= \frac{\partial^2 a}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\alpha}^\top} = \frac{\partial}{\partial \boldsymbol{\lambda}} \left[\frac{\partial a}{\partial \mathbf{m}^\top} \right] \mathbf{K} = \frac{\partial \left(\mathrm{dg}\mathbf{V}\right)^\top}{\partial \boldsymbol{\lambda}} \frac{\partial}{\partial \mathrm{dg}\mathbf{V}} \left[\frac{\partial a}{\partial \mathbf{m}^\top} \right] \mathbf{K} \\
&= 2\mathbf{V} \odot \mathbf{V} \frac{\partial^2 a}{\partial \mathrm{dg}\mathbf{V} \partial \mathbf{m}^\top} \mathbf{K}, \\
\frac{\partial^2 \ln Z_B}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}^\top} &= \frac{\partial^2 a}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}^\top} + \mathbf{R}, \quad \mathbf{R} := 2\mathbf{V} \odot \left(\mathbf{V} - \mathbf{A}\mathbf{V}^\top - \mathbf{V}\mathbf{A}^\top\right) \\
&= 2 \frac{\partial}{\partial \boldsymbol{\lambda}} \left[\frac{\partial a}{\partial \left(\mathrm{dg}\mathbf{V}\right)^\top} \mathbf{V} \odot \mathbf{V} \right] + \mathbf{R} \\
&= 2 \frac{\partial^2 a}{\partial \boldsymbol{\lambda} \partial \left(\mathrm{dg}\mathbf{V}\right)^\top} \mathbf{V} \odot \mathbf{V} + 2 \left[\frac{\partial a}{\partial \left(\mathrm{dg}\mathbf{V}\right)^\top} \frac{\partial \mathbf{V} \odot \mathbf{V}}{\partial \lambda_i} \right]_i + \mathbf{R} \\
&= 2 \frac{\partial \left(\mathrm{dg}\mathbf{V}\right)^\top}{\partial \boldsymbol{\lambda}} \frac{\partial^2 a}{\partial \mathrm{dg}\mathbf{V} \partial \left(\mathrm{dg}\mathbf{V}\right)^\top} \mathbf{V} \odot \mathbf{V} + 4 \left[\frac{\partial a}{\partial \left(\mathrm{dg}\mathbf{V}\right)^\top} \left(\mathbf{V} \odot \frac{\partial \mathbf{V}}{\partial \lambda_i} \right) \right]_i + \mathbf{R} \\
&= 4\mathbf{V} \odot \mathbf{V} \frac{\partial^2 a}{\partial \mathrm{dg}\mathbf{V} \partial \left(\mathrm{dg}\mathbf{V}\right)^\top} \mathbf{V} \odot \mathbf{V} + 8 \left[\frac{\partial a}{\partial \left(\mathrm{dg}\mathbf{V}\right)^\top} \left(\mathbf{V} \odot \left(\mathbf{v}_i \mathbf{v}_i^\top \right) \right) \right]_i + \mathbf{R}.
\end{aligned}$$

In the following, we abbreviate $\mathcal{N}(f|m_i,v_{ii})$ by $\mathcal{N}_i$.

$$\begin{aligned}
\frac{\partial^2 a}{\partial m_i^2} &= \int \frac{\partial^2 \mathcal{N}_i}{\partial m_i^2} \ln \operatorname{sig}(y_i f) \mathrm{d}f = \int \frac{(f-m_i)^2 - c_{ii}}{v_{ii}^2} \mathcal{N}_i \ln \operatorname{sig}(y_i f) \mathrm{d}f \\
&= \frac{1}{v_{ii}} \int \left(f^2 - 1\right) \cdot \mathcal{N}(f) \ln \operatorname{sig}(\sqrt{v_{ii}} y_i f + m_i y_i) \mathrm{d}f, \\
\frac{\partial^2 a}{\partial c_{ii} \partial m_i} &= \int \frac{\partial^2 \mathcal{N}_i}{\partial v_{ii} \partial m_i} \ln \operatorname{sig}(y_i f) \mathrm{d}f = \int \frac{(f-m_i)^3 - 3(f-m_i) v_{ii}}{2 v_{ii}^3} \mathcal{N}_i \ln \operatorname{sig}(y_i f) \mathrm{d}f \\
&= \frac{1}{2 v_{ii}^{\frac{3}{2}}} \int \left(f^3 - 3f\right) \cdot \mathcal{N}(f) \ln \operatorname{sig}\left(\sqrt{v_{ii}} y_i f + m_i y_i\right) \mathrm{d}f, \\
\frac{\partial^2 a}{\partial v_{ii}^2} &= \int \frac{\partial^2 \mathcal{N}_i}{\partial v_{ii}^2} \ln \operatorname{sig}(y_i f) \mathrm{d}f = \int \frac{(f-m_i)^4 - 6 v_{ii} (f-m_i)^2 + 3 v_{ii}^2}{4 v_{ii}^4} \mathcal{N}_i \ln \operatorname{sig}(y_i f) \mathrm{d}f \\
&= \frac{1}{4 v_{ii}^2} \int \left(f^4 - 6 f^2 + 3\right) \cdot \mathcal{N}(f) \ln \operatorname{sig}(\sqrt{v_{ii}} y_i f + m_i y_i) \mathrm{d}f.
\end{aligned}$$

A.2.3 FIRST DERIVATIVES W.R.T. HYPERPARAMETERS θ_i:

The direct gradient is given by the following equation where we have marked the dependency of the covariance $\mathbf{K}$ on θ_i by subscripts

$$\begin{aligned}
\frac{\partial \ln Z_B(\boldsymbol{\alpha}, \boldsymbol{\Lambda})}{\partial \theta_i} &= \boldsymbol{\alpha}^\top \frac{\partial \mathbf{K}_\theta}{\partial \theta_i} \frac{\partial a(\mathbf{y}, \mathbf{m}, \mathbf{V})}{\partial \mathbf{m}} + \mathrm{dg}\left(\mathbf{A} \frac{\partial \mathbf{K}_\theta}{\partial \theta_i} \mathbf{A}^\top\right)^\top \frac{\partial a(\mathbf{y}, \mathbf{m}, \mathbf{V})}{\partial \mathrm{dg}\mathbf{V}} \\
&\quad + \mathrm{tr}\left(\mathbf{A}^\top \boldsymbol{\Lambda} \frac{\partial \mathbf{K}_\theta}{\partial \theta_i}\right) - \mathrm{tr}\left(\mathbf{A} \frac{\partial \mathbf{K}_\theta}{\partial \theta_i} \boldsymbol{\Lambda} \mathbf{A}\right) - \frac{1}{2} \boldsymbol{\alpha}^\top \frac{\partial \mathbf{K}_\theta}{\partial \theta_i} \boldsymbol{\alpha}.
\end{aligned}$$

Again we have would have to add an implicit term to the gradient, but in our implementation, we forbore from doing so.

Appendix B. Auxiliary Calculations

In the following, we enumerate some calculations we removed from the main text in order to improve on readability.

B.1 Limits of the Covariance Matrix and Corresponding Marginal Likelihood

We investigate the behavior of the covariance matrix $\mathbf{K}$ for extreme lengthscales ℓ. The matrix is given by $[\mathbf{K}]_{ij} = \sigma_f^2 g(|\mathbf{x}_i - \mathbf{x}_j|/\ell)$ where $g : \mathbb{R} \to \mathbb{R}$ is monotonously decreasing and continuous with $g(0) = 1$ and $\lim_{t \to \infty} g(t) = 0$. >From this definition we have $[\mathbf{K}]_{ii} = \sigma_f^2$. We define $\Delta_{ij} := |\mathbf{x}_i - \mathbf{x}_j|/\ell > 0$ for $i \neq j$. From

$$\begin{aligned}
\lim_{\ell \to 0} [\mathbf{K}]_{ij} &\overset{i \neq j}{=} \lim_{\ell \to 0} \sigma_f^2 g(|\mathbf{x}_i - \mathbf{x}_j|/\ell) = \sigma_f^2 \lim_{\Delta_{ij} \to \infty} g(\Delta_{ij}) = 0, \\
\lim_{\ell \to \infty} [\mathbf{K}]_{ij} &\overset{i \neq j}{=} \lim_{\ell \to \infty} \sigma_f^2 g(|\mathbf{x}_i - \mathbf{x}_j|/\ell) = \sigma_f^2 \lim_{\Delta_{ij} \to 0} g(\Delta_{ij}) = 1
\end{aligned}$$

we conclude

$$\begin{aligned}\lim_{\ell\to 0}\mathbf{K} &= \sigma_f^2\mathbf{I},\\ \lim_{\ell\to\infty}\mathbf{K} &= \sigma_f^2\mathbb{1}\mathbb{1}^\top.\end{aligned}$$

The sigmoids are normalized $\mathrm{sig}(-f_i)+\mathrm{sig}(f_i)=1$ and the Gaussian is symmetric $\mathcal{N}(f_i)=\mathcal{N}(-f_i)$. Consequently, we have

$$\begin{aligned}\int \mathrm{sig}(y_i f_i)\,\mathcal{N}(f_i|0,\sigma_f^2)\mathrm{d}f_i &= \int \mathrm{sig}(f_i)\,\mathcal{N}(f_i|0,\sigma_f^2)\mathrm{d}f_i\\ &= \int_{-\infty}^{0}\mathrm{sig}(f_i)\,\mathcal{N}(f_i|0,\sigma_f^2)\mathrm{d}f_i+\int_0^{\infty}\mathrm{sig}(f_i)\,\mathcal{N}(f_i|0,\sigma_f^2)\mathrm{d}f_i\\ &= \int_0^{\infty}\mathrm{sig}(-f_i)\,\mathcal{N}(-f_i|0,\sigma_f^2)\mathrm{d}f_i+\int_0^{\infty}\mathrm{sig}(f_i)\,\mathcal{N}(f_i|0,\sigma_f^2)\mathrm{d}f_i\\ &= \int_0^{\infty}[\mathrm{sig}(-f_i)+\mathrm{sig}(f_i)]\,\mathcal{N}(f_i|0,\sigma_f^2)\mathrm{d}f_i\\ &= \int_0^{\infty}1\cdot\mathcal{N}(f_i|0,\sigma_f^2)\mathrm{d}f_i=\frac{1}{2}.\end{aligned}$$

The marginal likelihood is given by

$$\begin{aligned}Z &= \int \mathbb{P}(\mathbf{y}|\mathbf{f})\,\mathbb{P}(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})\,\mathrm{d}\mathbf{f}\\ &= \int\prod_{i=1}^{n}\mathrm{sig}(y_i f_i)\,|2\pi\mathbf{K}|^{-\frac{1}{2}}\exp(-\frac{1}{2}\mathbf{f}^\top\mathbf{K}^{-1}\mathbf{f})\mathrm{d}\mathbf{f}.\end{aligned}$$

B.1.1 LENGTHSCALE TO ZERO

For $\mathbf{K}=\sigma_f^2\mathbf{I}$ the prior factorizes and we get

$$\begin{aligned}Z_{\ell\to 0} &= \prod_{i=1}^{n}\int\mathrm{sig}(y_i f_i)\,\frac{1}{\sqrt{2\pi\sigma_f^2}}\exp(-\frac{f_i^2}{2\sigma_f^2})\mathrm{d}f_i\\ &\overset{(17)}{=} \prod_{i=1}^{n}\frac{1}{2}=2^{-n}.\end{aligned}$$

B.1.2 LENGTHSCALE TO INFINITY

To get $\mathbf{K} \to \sigma_f^2 \mathbb{1}\mathbb{1}^\top$ we write $\mathbf{K} = \sigma_f^2 \mathbf{1} + \varepsilon^2 \mathbf{I}$ with $\mathbf{1} = \mathbb{1}\mathbb{1}^\top$ and let $\varepsilon \to 0$. The eigenvalue decomposition of $\mathbf{K}$ is written as $\mathbf{K} = \sum_{i=1}^n \mathbf{u}_i \mathbf{u}_i^\top \lambda_i$ with $\mathbf{u}_1 = \frac{1}{\sqrt{n}}\mathbb{1}$, $\lambda_1 = \sigma_f^2 + \varepsilon^2$ and all other $\lambda_i = \varepsilon^2$

$$
\begin{aligned}
Z_{\frac{1}{\varepsilon}} &\overset{\mathbf{K}=\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top}{=} && \int \prod_{i=1}^n \operatorname{sig}(y_i f_i) \, |2\pi\mathbf{\Lambda}|^{-\frac{1}{2}} \exp(-\frac{1}{2}\mathbf{f}^\top \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^\top \mathbf{f}) \mathrm{d}\mathbf{f} \\
&\overset{\mathbf{t}=\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}^\top\mathbf{f}}{=} && \int \prod_{i=1}^n \operatorname{sig}\left(y_i \sqrt{\lambda_i} \cdot \mathbf{t}^\top \mathbf{u}_i\right) |2\pi\mathbf{\Lambda}|^{-\frac{1}{2}} \exp(-\frac{1}{2}\mathbf{t}^\top\mathbf{t}) \left|\mathbf{\Lambda}^{\frac{1}{2}}\right| \mathrm{d}\mathbf{t} \\
&= && \int \prod_{i=1}^n \operatorname{sig}\left(y_i \sqrt{\lambda_i} \cdot \mathbf{t}^\top \mathbf{u}_i\right) \mathcal{N}(t_i) \mathrm{d}\mathbf{t} \\
&= && \int \operatorname{sig}\left(\sqrt{\frac{\sigma_f^2 + \varepsilon^2}{n}} \cdot \mathbf{t}^\top \mathbb{1}\right) \mathcal{N}(t_1) \prod_{i=2}^n \left[\operatorname{sig}\left(\varepsilon \cdot \mathbf{t}^\top \mathbf{u}_i\right)\right] \mathcal{N}(t_i) \mathrm{d}\mathbf{t}, \\
Z_{\ell\to\infty} = \lim_{\varepsilon\to 0} Z \quad &= && \int \operatorname{sig}\left(\frac{\sigma_f}{\sqrt{n}} \cdot \mathbf{t}^\top \mathbb{1}\right) \mathcal{N}(t_1) \prod_{i=2}^n \left[\frac{1}{2}\right] \mathcal{N}(t_i) \mathrm{d}\mathbf{t} \\
&\overset{(17)}{=} && 2^{-n+1} \int \operatorname{sig}\left(\frac{\sigma_f}{\sqrt{n}} \cdot \mathbf{t}^\top \mathbb{1}\right) \mathcal{N}(\mathbf{t}) \mathrm{d}\mathbf{t} \\
&\overset{r=\mathbf{t}^\top\mathbb{1}}{=} && 2^{-n+1} \int \operatorname{sig}\left(\frac{\sigma_f}{\sqrt{n}} \cdot r\right) \mathcal{N}(r) \mathrm{d}r \\
&\overset{(17)}{=} && 2^{-n}.
\end{aligned}
$$

B.1.3 LATENT SCALE TO ZERO

We define $\sigma_f^2 \tilde{\mathbf{K}} = \mathbf{K}$ and $\sigma_f \tilde{\mathbf{f}} = \mathbf{f}$ and derive

$$
\begin{aligned}
Z_{\sigma_f} &= \int \prod_{i=1}^n \operatorname{sig}(y_i f_i) \, |2\pi\mathbf{K}|^{-\frac{1}{2}} \exp(-\frac{1}{2}\mathbf{f}^\top \mathbf{K}^{-1}\mathbf{f}) \mathrm{d}\mathbf{f} \\
&= \int \prod_{i=1}^n \operatorname{sig}\left(y_i \sigma_f \tilde{f}_i\right) |2\pi\mathbf{K}|^{-\frac{1}{2}} \exp(-\frac{\sigma_f^2}{2}\tilde{\mathbf{f}}^\top \mathbf{K}^{-1}\tilde{\mathbf{f}}) \sigma_f^n \mathrm{d}\tilde{\mathbf{f}} \\
&= \int \prod_{i=1}^n \operatorname{sig}\left(y_i \sigma_f \tilde{f}_i\right) \left|2\pi\sigma_f^2\tilde{\mathbf{K}}\right|^{-\frac{1}{2}} \exp(-\frac{\sigma_f^2}{2}\tilde{\mathbf{f}}^\top \sigma_f^{-2}\tilde{\mathbf{K}}^{-1}\tilde{\mathbf{f}}) \sigma_f^n \mathrm{d}\tilde{\mathbf{f}} \\
&= \int \prod_{i=1}^n \left[\operatorname{sig}\left(y_i \sigma_f \tilde{f}_i\right)\right] \mathcal{N}\left(\tilde{\mathbf{f}} | \mathbf{0}, \tilde{\mathbf{K}}\right) \mathrm{d}\tilde{\mathbf{f}}, \\
Z_{\sigma_f\to 0} = \lim_{\sigma_f\to 0} Z &= \int \prod_{i=1}^n \left[\frac{1}{2}\right] \mathcal{N}\left(\tilde{\mathbf{f}} | \mathbf{0}, \tilde{\mathbf{K}}\right) \mathrm{d}\tilde{\mathbf{f}} = 2^{-n}.
\end{aligned}
$$

Note that the functions, we are using are all well-behaved, such that the limits do exist.

B.2 Posterior Divided by Prior = Effective Likelihood

$$\begin{aligned}
\mathbb{Q}(\mathbf{y}|\mathbf{f}) &= \frac{\mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V})}{\mathbb{P}(\mathbf{f}|\mathbf{X})} = \frac{\mathcal{N}\left(\mathbf{f}|\mathbf{m},\left(\mathbf{K}^{-1}+\mathbf{W}\right)^{-1}\right)}{\mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{K})} \\
&= \frac{\mathcal{N}\left(\mathbf{f}|\tilde{\mathbf{m}},\mathbf{W}^{-1}\right)}{\mathcal{N}(\tilde{\mathbf{m}}|\mathbf{0},\mathbf{K}+\mathbf{W}^{-1})}, \quad \tilde{\mathbf{m}} = (\mathbf{K}\mathbf{W})^{-1}\mathbf{m}+\mathbf{m} \\
&= \frac{(2\pi)^{-\frac{n}{2}}\left|\mathbf{W}^{-1}\right|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{f}-\tilde{\mathbf{m}})^{\top}\mathbf{W}(\mathbf{f}-\tilde{\mathbf{m}})\right)}{(2\pi)^{-\frac{n}{2}}\left|\mathbf{K}+\mathbf{W}^{-1}\right|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\tilde{\mathbf{m}}^{\top}(\mathbf{K}+\mathbf{W}^{-1})^{-1}\tilde{\mathbf{m}}\right)} \\
&= \sqrt{|\mathbf{K}\mathbf{W}+\mathbf{I}|}\frac{\exp\left(-\frac{1}{2}(\mathbf{f}-\tilde{\mathbf{m}})^{\top}\mathbf{W}(\mathbf{f}-\tilde{\mathbf{m}})\right)}{\exp\left(-\frac{1}{2}\tilde{\mathbf{m}}^{\top}(\mathbf{K}+\mathbf{W}^{-1})^{-1}\tilde{\mathbf{m}}\right)} \\
&=: \frac{1}{Z_{\mathbb{Q}}}\exp\left(-\frac{1}{2}(\mathbf{f}-\tilde{\mathbf{m}})^{\top}\mathbf{W}(\mathbf{f}-\tilde{\mathbf{m}})\right), \\
\ln Z_{\mathbb{Q}} &= -\frac{1}{2}\tilde{\mathbf{m}}^{\top}\left(\mathbf{K}+\mathbf{W}^{-1}\right)^{-1}\tilde{\mathbf{m}}-\frac{1}{2}\ln|\mathbf{K}\mathbf{W}+\mathbf{I}|
\end{aligned}$$

B.3 Kullback-Leibler Divergence for KL method

We wish to calculate the divergence between the approximate posterior, a Gaussian, and the true posterior

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{Q}(\mathbf{f}|\boldsymbol{\theta}) \parallel \mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})\right) &= \int \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V})\ln\frac{\mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V})}{\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})}\mathrm{d}\mathbf{f} \\
&\overset{(2)}{=} \int \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V})\ln\frac{Z\cdot\mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V})}{\mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V})\prod_{i=1}^{n}\mathbb{P}(y_i|f_i)}\mathrm{d}\mathbf{f} \\
&= \ln Z + \int \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V})\ln\mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V})\,\mathrm{d}\mathbf{f} \\
&\quad - \int \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V})\ln\prod_{i=1}^{n}\mathbb{P}(y_i|f_i)\mathrm{d}\mathbf{f} \\
&\quad - \int \mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V})\ln\mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{K})\mathrm{d}\mathbf{f}.
\end{aligned}$$

There are three Gaussian integrals to evaluate; the entropy of the approximate posterior and two other expectations

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{Q}(\mathbf{f}|\boldsymbol{\theta}) \parallel \mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta})\right) &= \ln Z - \frac{1}{2}\ln|\mathbf{V}| - \frac{n}{2} - \frac{n}{2}\ln 2\pi \\
&\quad - \int \mathcal{N}(f)\left[\sum_{i=1}^{n}\ln\mathrm{sig}\left(\sqrt{v_{ii}}y_i f + m_i y_i\right)\right]\mathrm{d}f \qquad (17) \\
&\quad + \frac{n}{2}\ln 2\pi + \frac{1}{2}\ln|\mathbf{K}| + \frac{1}{2}\mathbf{m}^{\top}\mathbf{K}^{-1}\mathbf{m} + \frac{1}{2}\mathrm{tr}\left(\mathbf{K}^{-1}\mathbf{V}\right).
\end{aligned}$$

Summing up and dropping the constant (w.r.t. $\mathbf{m}$ and $\mathbf{V}$) terms, we arrive at

$$\mathrm{KL}(\mathbf{m},\mathbf{V}) \stackrel{c}{=} -\int \mathcal{N}(f)\left[\sum_{i=1}^{n} \ln \mathrm{sig}\left(\sqrt{v_{ii}}y_i f + m_i y_i\right)\right] \mathrm{d}f - \frac{1}{2}\ln|\mathbf{V}| + \frac{1}{2}\mathbf{m}^\top \mathbf{K}^{-1}\mathbf{m} + \frac{1}{2}\mathrm{tr}\left(\mathbf{K}^{-1}\mathbf{V}\right).$$

B.4 Gaussian Integral for VB Lower Bound

$$\begin{aligned}
Z_B &= \int \mathbb{P}(\mathbf{f}|\mathbf{X})\,\mathbb{Q}(\mathbf{y}|\mathbf{f},\mathbf{A},\mathbf{b},\mathbf{c})\,\mathrm{d}\mathbf{f} = \int \mathcal{N}(\mathbf{f}|0,\mathbf{K}) \exp\left(\mathbf{f}^\top \mathbf{A}\mathbf{f} + (\mathbf{b}\odot\mathbf{y})^\top \mathbf{f} + \mathbf{c}^\top \mathbb{1}\right) \mathrm{d}\mathbf{f} \\
&= \frac{\exp\left(\mathbf{c}^\top \mathbb{1}\right)}{\sqrt{(2\pi)^n |\mathbf{K}|}} \int \exp\left(-\frac{1}{2}\mathbf{f}^\top \left(\mathbf{K}^{-1} - 2\mathbf{A}\right)\mathbf{f} + (\mathbf{b}\odot\mathbf{y})^\top \mathbf{f}\right) \mathrm{d}\mathbf{f} \\
&= \frac{\exp\left(\mathbf{c}^\top \mathbb{1}\right)}{\sqrt{(2\pi)^n |\mathbf{K}|}} \sqrt{\frac{(2\pi)^n}{|\mathbf{K}^{-1} - 2\mathbf{A}|}} \exp\left(\frac{1}{2}(\mathbf{b}\odot\mathbf{y})^\top \left(\mathbf{K}^{-1} - 2\mathbf{A}\right)^{-1} (\mathbf{b}\odot\mathbf{y})\right) \\
&= \frac{\exp\left(\mathbf{c}^\top \mathbb{1}\right)}{\sqrt{|\mathbf{I} - 2\mathbf{A}\mathbf{K}|}} \exp\left(\frac{1}{2}(\mathbf{b}\odot\mathbf{y})^\top \left(\mathbf{K}^{-1} - 2\mathbf{A}\right)^{-1} (\mathbf{b}\odot\mathbf{y})\right), \\
\ln Z_B &= \mathbf{c}^\top \mathbb{1} + \frac{1}{2}(\mathbf{b}\odot\mathbf{y})^\top \left(\mathbf{K}^{-1} - 2\mathbf{A}\right)^{-1} (\mathbf{b}\odot\mathbf{y}) - \frac{1}{2}\ln|\mathbf{I} - 2\mathbf{A}\mathbf{K}|.
\end{aligned}$$

B.5 Lower Bound for the Cumulative Gaussian Likelihood

A lower bound

$$\mathrm{sig}_{\mathrm{probit}}(y_i f_i) \;\geq\; \mathbb{Q}(y_i|f_i,\varsigma_i) = a_i f_i^2 + b_i f_i + c_i$$

for the cumulative Gaussian likelihood function is derived by matching the function at one point ς

$$\mathbb{Q}(y_i = +1|f_i,\varsigma_i) \;=\; \mathrm{sig}_{\mathrm{probit}}(\varsigma_i), \forall i$$

and by matching the first derivative

$$\left.\frac{\partial}{\partial f_i}\ln \mathbb{Q}(y_i = +1|f_i,\varsigma_i)\right|_{\varsigma_i} \;=\; \frac{\partial \ln \mathrm{sig}_{\mathrm{probit}}(y_i f_i)}{\partial f_i} = \frac{\mathcal{N}(\varsigma_i)}{\mathrm{sig}_{\mathrm{probit}}(\varsigma_i)}, \forall i$$

at this point for a tight approximation. Solving for these constraints leads to the coefficients

$$\begin{aligned}
\text{asymptotic behavior} \Rightarrow a_i &= -\frac{1}{2}, \\
\text{first derivative} \Rightarrow b_i &= \varsigma_i + \frac{\mathcal{N}(\varsigma_i)}{\mathrm{sig}_{\mathrm{probit}}(\varsigma_i)}, \\
\text{point matching} \Rightarrow c_i &= \left(\frac{\varsigma_i}{2} - b_i\right)\varsigma_i + \log \mathrm{sig}_{\mathrm{probit}}(\varsigma_i).
\end{aligned}$$

B.6 Free Form Optimization for FV

We make a factorial approximation $\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X}) \approx \mathbb{Q}(\mathbf{f}) := \prod_i \mathbb{Q}(f_i)$ to the posterior by minimizing

$$\begin{aligned}\mathrm{KL}[\mathbb{Q}(\mathbf{f})||\mathbb{P}(\mathbf{f})] &= \int \prod_{i=1}^n \mathbb{Q}(f_i) \ln \frac{Z \cdot \prod_{i=1}^n \mathbb{Q}(f_i)}{\mathcal{N}(\mathbf{f}|\mathbf{m},\mathbf{V}) \prod_{i=1}^n \mathbb{P}(y_i|f_i)} \mathrm{d}\mathbf{f} \\ &= \sum_i \int \mathbb{Q}(f_i) \ln \frac{\mathbb{Q}(f_i)}{\mathbb{P}(y_i|f_i)} \mathrm{d}f_i + \frac{1}{2} \int \prod_{i=1}^n \mathbb{Q}(f_i) \mathbf{f}^\top \mathbf{K}^{-1} \mathbf{f} \mathrm{d}\mathbf{f} + \mathrm{const}_{\mathbf{f}}.\end{aligned}$$

Free-form optimization proceeds by equating the functional derivative with zero

$$\frac{\delta \mathrm{KL}}{\delta \mathbb{Q}(f_i)} = \ln \mathbb{Q}(f_i) + 1 - \ln \mathbb{P}(y_i|f_i) + \frac{1}{2} \frac{\delta}{\delta \mathbb{Q}(f_i)} \int \prod_{i=1}^n \mathbb{Q}(f_i) \mathbf{f}^\top \mathbf{K}^{-1} \mathbf{f} \mathrm{d}\mathbf{f}. \tag{18}$$

We abbreviate the integral in the last term with ξ and rewrite it in terms of simple one-dimensional integrals $m_l = \int f_l \mathbb{Q}(f_l) \mathrm{d}f_l$ and $v_l = \int f_l^2 \mathbb{Q}(f_l) \mathrm{d}f_l - m_l^2$

$$\begin{aligned}\xi &= \int \prod_i \mathbb{Q}_i \sum_{j,k} f_j \left[\mathbf{K}^{-1}\right]_{jk} f_k \mathrm{d}\mathbf{f} \\ &= \int \prod_{i \neq l} \mathbb{Q}_i \left[\int \mathbb{Q}_l \left(f_l^2 \left[\mathbf{K}^{-1}\right]_{ll} + 2 f_l \sum_{j \neq l} f_j \left[\mathbf{K}^{-1}\right]_{jl} + \sum_{j \neq l, k \neq l} f_j \left[\mathbf{K}^{-1}\right]_{jk} f_k \right) \mathrm{d}f_l \right] \mathrm{d}\mathbf{f}_{\neg l} \\ &= \int \prod_{i \neq l} \mathbb{Q}_i \left[\left[\mathbf{K}^{-1}\right]_{ll} \underbrace{\int f_l^2 \mathbb{Q}_l \mathrm{d}f_l}_{v_l + m_l^2} + 2 \left(\sum_{j \neq l} f_j \left[\mathbf{K}^{-1}\right]_{jl}\right) \underbrace{\int f_l \mathbb{Q}_l \mathrm{d}f_l}_{m_l} + \sum_{j \neq l, k \neq l} f_j \left[\mathbf{K}^{-1}\right]_{jk} f_k \right] \mathrm{d}\mathbf{f}_{\neg l} \\ &= \left[\mathbf{K}^{-1}\right]_{ll} (v_l + m_l^2) + 2 \sum_{j \neq l} m_j \left[\mathbf{K}^{-1}\right]_{jl} m_l + \int \prod_{i \neq l} \mathbb{Q}_i \sum_{j \neq l, k \neq l} f_j \left[\mathbf{K}^{-1}\right]_{jk} f_k \mathrm{d}\mathbf{f}_{\neg l} \\ &= \text{induction over } l \\ &= \sum_l \left[\mathbf{K}^{-1}\right]_{ll} (v_l + m_l^2) + 2 \sum_{j<l} m_j \left[\mathbf{K}^{-1}\right]_{jl} m_l.\end{aligned}$$

Plugging this into Equation 18 and using $\frac{\delta \int f_l^p \mathbb{Q}(f_l) \mathrm{d}f_l}{\delta \mathbb{Q}(f_l)} = f_l^p$, we find

$$\begin{aligned}\frac{\delta \mathrm{KL}}{\delta \mathbb{Q}(f_i)} &= \ln \mathbb{Q}(f_i) + 1 - \ln \mathbb{P}(y_i|f_i) + \frac{1}{2} f_i \left[\mathbf{K}^{-1}\right]_{ii} f_i + f_i \sum_l \left[\mathbf{K}^{-1}\right]_{il} m_l \stackrel{!}{\equiv} 0 \\ \Rightarrow \mathbb{Q}(f_i) &\propto \exp\left(-\frac{1}{2} f_i \left[\mathbf{K}^{-1}\right]_{ii} f_i - f_i \sum_{l \neq i} \left[\mathbf{K}^{-1}\right]_{il} m_l \right) \mathbb{P}(y_i|f_i) \\ \Rightarrow \mathbb{Q}(f_i) &\propto \mathcal{N}\left(f_i \middle| m_i - \frac{\left[\mathbf{K}^{-1}\mathbf{m}\right]_i}{\left[\mathbf{K}^{-1}\right]_{ii}}, \left[\mathbf{K}^{-1}\right]_{ii}^{-1} \right) \mathbb{P}(y_i|f_i)\end{aligned}$$

as the functional form of the best possible factorial approximation, namely a product of the true likelihood times a Gaussian with the same precision as the prior marginal.

References

Yasemin Altun, Thomas Hofmann, and Alex Smola. Gaussian process classification for segmenting and annotating sequences. In *International Conference on Machine Learning*, 2004.

Wei Chu, Zoubin Ghahramani, Francesco Falciani, and David L. Wild. Biomarker discovery in microarray gene expression data with gaussian processes. *Bioinformatics*, 21:3385–3393, 2005.

Lehel Csató, Ernest Fokoué, Manfred Opper, and Bernhard Schottky. Efficient Approaches to Gaussian Process Classification. In *Neural Information Processing Systems 12*, pages 251–257. MIT Press, 2000.

Mark N. Gibbs and David J. C. MacKay. Variational Gaussian Process Classifiers. *IEEE Transactions on Neural Networks*, 11(6):1458–1464, 2000.

Mark Girolami and Simon Rogers. Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors. *Neural Computation*, 18:1790–1817, 2006.

Ashish Kapoor and Rosalind W. Picard. Multimodal affect recognition in learning environments. In *ACM international conference on Multimedia*, 2005.

Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Active learning with gaussian processes for object categorization. In *ICCV*, 2007.

Malte Kuss and Carl Edward Rasmussen. Assessing Approximate Inference for Binary Gaussian Process Classification. *Journal of Machine Learning Research*, 6:1679 – 1704, 10 2005.

David J. C. MacKay. Bayesian Interpolation. *Neural Computation*, 4(3):415–447, 1992.

Thomas P. Minka. Expectation Propagation for Approximate Bayesian Inference. In *UAI*, pages 362–369. Morgan Kaufmann, 2001a.

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

Tom Minka. Divergence Measures and Message Passing. Technical report, Microsoft Research, 2005.

Radford M. Neal. Annealed Importance Sampling. *Statistics and Computing*, 11:125–139, 2001.

Radford M. Neal. Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, September 1993.

Manfred Opper and Cédric Archambeau. The Variational Gaussian Approximation Revisited. *Neural Computation*, accepted, 2008.

Manfred Opper and Ole Winther. Gaussian Processes for Classification: Mean Field Algorithms. *Neural Computation*, 12(11):2655–2684, 2000.

Manfred Opper and Ole Winther. Expectation Consistent Approximate Inference. *Journal of Machine Learning Research*, 6:2177–2204, 2005.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C*. Cambridge University Press, 2nd edition, February 1993.

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

Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *JMLR*, 5:101–141, 2004.

Anton Schwaighofer, Volker Tresp, Peter Mayer, Alexander K. Scheel, and Gerhard Müller. The RA scanner: Prediction of rheumatoid joint inflammation based on laser imaging. In *NIPS*, 2002.

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

Matthias Seeger. Bayesian Methods for Support Vector Machines and Gaussian Processes. Master's thesis, Universität Karlsruhe, 1999.

S. Sundararajan and S. S. Keerthi. Predictive Approaches for Choosing Hyperparameters in Gaussian Processes. *Neural Computation*, 13:1103–1118, 2001.

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

Mingjun Zhong, Fabien Lotte, Mark Girolami, and Anatole Lécuyer. Classifying eeg for brain computer interfaces using gaussian processes. *Pattern Recognition Letters*, 29:354–359, 2008.