Aalto University

OPEN ACCESS

This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Li, Rui; John, ST; Solin, Arno

Improving Hyperparameter Learning under Approximate Inference in Gaussian Process Models

Published in:
Proceedings of the 40th International Conference on Machine Learning

Published: 01/07/2023

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY

Please cite the original version:
Li, R., John, ST., & Solin, A. (2023). Improving Hyperparameter Learning under Approximate Inference in Gaussian Process Models. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (pp. 19595-19615). (Proceedings of Machine Learning Research; Vol. 202). JMLR. https://proceedings.mlr.press/v202/li23m.html

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

Improving Hyperparameter Learning under Approximate Inference in Gaussian Process Models

Rui Li [1] ST John [1] Arno Solin [1]

Abstract

Approximate *inference* in Gaussian process (GP) models with non-conjugate likelihoods gets entangled with the *learning* of the model hyperparameters. We improve hyperparameter learning in GP models and focus on the interplay between variational inference (VI) and the learning target. While VI's lower bound to the marginal likelihood is a suitable objective for inferring the approximate posterior, we show that a direct approximation of the marginal likelihood as in Expectation Propagation (EP) is a better learning objective for hyperparameter optimization. We design a hybrid training procedure to bring the best of both worlds: it leverages conjugate-computation VI for inference and uses an EP-like marginal likelihood approximation for hyperparameter learning. We compare VI, EP, Laplace approximation, and our proposed training procedure and empirically demonstrate the effectiveness of our proposal across a wide range of data sets.

1. Introduction

Gaussian processes (GPs, Rasmussen & Williams, 2006) provide a plug-and-play approach for inference and learning, with principled ways of incorporating prior knowledge over functions and quantifying uncertainty. While GP regression under a conjugate (Gaussian) likelihood can be carried out elegantly in closed form, we focus on the non-conjugate case, where exact inference is intractable. Training of the GP consists of *inferring* the approximate posterior and *learning* the hyperparameters of the model. For clarity, by *training* we refer to the combination of inference and learning.

In a supervised learning setting, GPs are typically trained to optimize performance on the training samples (as in empirical risk minimization, Vapnick, 1998). Under the GP paradigm, the go-to solution to learning is finding $\boldsymbol{\theta}^\star$ that maximizes the marginal likelihood. The marginal likelihood summarizes the probability that we would generate the observations $\mathbf{y}$ with the model parameters $\boldsymbol{\theta}$ if we would sample over the prior. It is formed by *marginalizing* over the latent functions from the GP prior, thus also known as the *evidence*. Even if this does not capture all aspects of *generalization* (see discussion in Vehtari et al., 2016; Lotfi et al., 2022), it is still used as a practical proxy for performance on unseen test points (see Fig. 1 for the proxy and test performance on the IONOSPHERE benchmark data set).

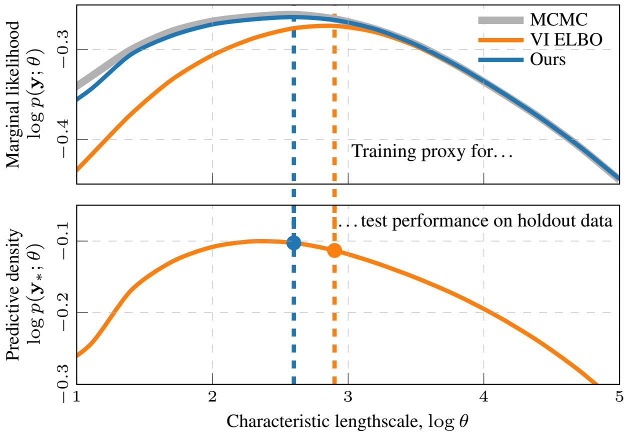

Figure 1. **Practical benefits** on IONOSPHERE: Marginal likelihood (top) acts as a training proxy for predictive density (bottom) of unseen future data. Our training objective produces a better point for prediction that also matches the MCMC baseline.

Under approximate inference, the marginalization step entangles the representation of the posterior with the learning target evaluation. The common approach is to assume an approximative Gaussian form for the posterior, so that the inference problem turns into finding a 'good' parameterization for the Gaussian (see Wilkinson et al., 2023, for a recent discussion on linearization/Gaussianization approaches). The simplest approach is the so-called Laplace's approximation (LA, Williams & Barber, 1998), which uses a second-order Taylor approximation. It is efficient but not very accurate. Variational inference (VI, Opper & Archam-

---

[1] Department of Computer Science, Aalto University, Finland, and Finnish Center for Artificial Intelligence (FCAI). Correspondence to: Rui Li <rui.li@aalto.fi>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

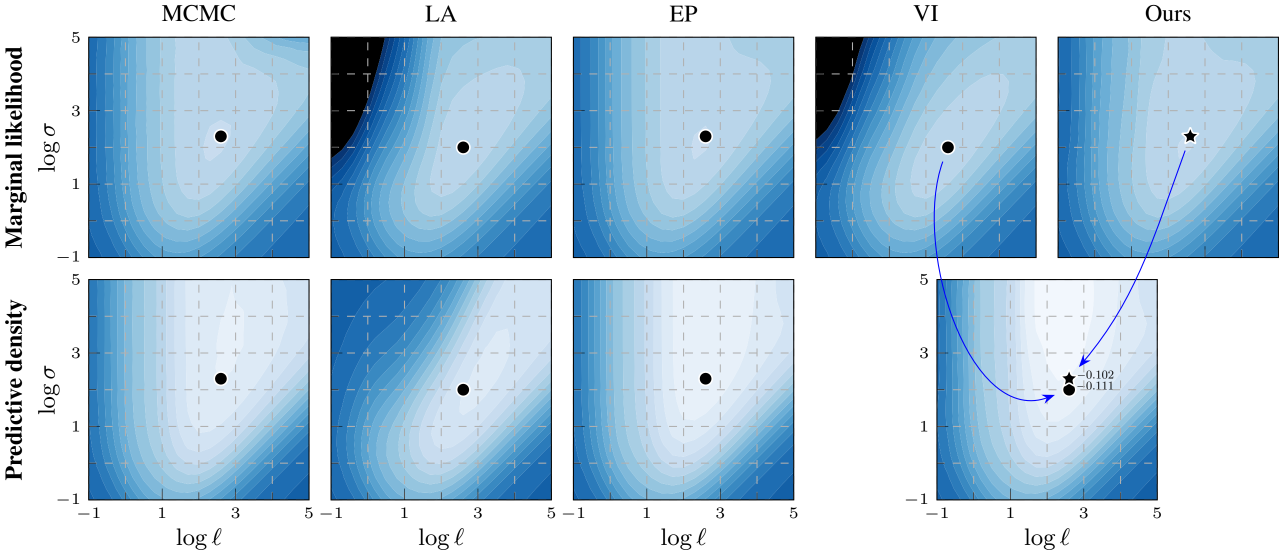

Figure 2. **Log marginal likelihood / predictive density surfaces** for the IONOSPHERE data set by varying kernel magnitude σ and lengthscale ℓ. The colour scale is the same in all plots: -0.8 ▨ 0 (normalized by n). Optimal hyperparameters are shown by a black marker. EP and our EP-like marginal likelihood estimation match the MCMC baseline better than VI or LA, thus providing a learning proxy. For prediction, our method still leverages the same variational representation as VI.

beau, 2009) and expectation propagation (EP, Minka, 2001) are two commonly used approximate inference methods for non-conjugate GP models, which have complementary advantages: VI optimizes a lower bound of the marginal likelihood, is easy to implement, straightforward to use, and the convex optimization problem is guaranteed to converge. However, it is known to underestimate variance (Bui et al., 2017). EP on the other hand requires implementation-wise tuning per likelihood and is not guaranteed to converge (Vehtari et al., 2020). However, it does provide a good approximation for the marginal likelihood (Kuss & Rasmussen, 2005; Nickisch & Rasmussen, 2008).

For model performance on unseen test data, the learning of hyperparameters plays a crucial role. Thus we strongly advocate against the common practice of jointly optimizing variational and hyperparameters using the ELBO, as the training target is only representative for the variational parameters. We build on work by Khan & Lin (2017) and Adam et al. (2021) that separate the learning of hyperparameters from inferring the variational parameters, and capture a link between VI and EP: the approximate posterior obtained through VI has exactly the same structure as the approximate posterior of EP. We obtain an EP-like marginal likelihood estimate from the VI approximate posterior for full and sparse GPs with no added computational cost. We propose a hybrid training procedure that combines the complementary advantages of natural-gradient VI and EP.

The contributions of this paper are as follows. *(i)* We improve generalizability in non-conjugate GP models with no extra computational cost by augmenting VI with an EP-like learning target for hyperparameter learning. *(ii)* We demonstrate our EP-like learning target is closer to an MCMC baseline and thus provides a better learning objective. We empirically compare the quality of the approximate marginal likelihood in LA, EP, and VI, and our proposed learning target. *(iii)* We show our method improves generalizability via experiments in binary classification for full and sparse GP models and robust regression.

2. Approximate Inference

In this section, we review common approximate inference methods in Gaussian process (GP) models. GP models put a GP prior over functions:

$$\text{GP prior:} \quad f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}')), \tag{1}$$

where $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$ is an input vector, $\mu(\mathbf{x})$ is the mean function, and $\kappa(\mathbf{x}, \mathbf{x}')$ is the covariance (kernel) function. This GP prior is linked to the data set $\mathcal{D} = (\mathbf{X}, \mathbf{y}) = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ of input–output pairs through a likelihood function that maps the latent function value $f(\mathbf{x})$ to the observations. We assume the likelihood factorizes over observations:

$$\text{Likelihood:} \quad \mathbf{y} \mid \mathbf{f} \sim \textstyle\prod_{i=1}^n p(y_i \mid f(\mathbf{x}_i)). \tag{2}$$

The posterior is given by $p(\mathbf{f} \mid \mathbf{y}; \boldsymbol{\theta}) \propto p(\mathbf{y} \mid \mathbf{f}; \boldsymbol{\theta})\, p(\mathbf{f}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the model (hyper)parameters of the likelihood, mean function, and kernel, and $\mathbf{f}$ is the vector of function values evaluated at the inputs. Prediction at a new test input $\mathbf{x}_*$ is obtained by computing the predictive distribution $p(f(\mathbf{x}_*) \mid \mathcal{D}, \mathbf{x}_*)$.

Probabilistic inference For (conjugate) Gaussian likelihoods, $p(y_i \mid f_i) = \mathrm{N}(y_i \mid f(\mathbf{x}_i), \sigma_\mathrm{n}^2)$, the posterior is available in closed form as a Gaussian distribution. For non-Gaussian likelihood models the inference problem needs to be approached with *approximative* inference methods. Sampling schemes (see Sec. 5.1 for our baseline solution) can tackle this, but for efficient inference one typically employs an approximative Gaussian posterior of the form

$$\text{Approximate posterior:} \quad q(\mathbf{f}) = \mathrm{N}(\mathbf{m}, \mathbf{S}). \tag{3}$$

Its 'optimal' parameterization (Opper & Archambeau, 2009) is given in terms of $2n$ parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ such that $\mathbf{m} = \mathbf{K}\boldsymbol{\alpha}$ and $\mathbf{S} = (\mathbf{K}^{-1} + \mathrm{diag}(\boldsymbol{\beta}))^{-1}$, where $\mathbf{K}$ is an $n \times n$ matrix with $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ as the ij^th entry. The inference problem thus turns into (efficiently) finding a (good) representation of the posterior in terms of Eq. (3) by minimizing some measure of error. Typical approaches for this are the Laplace approximation (local linearisation of the problem), expectation propagation (approximately minimizing $\mathrm{D_{KL}}\left[p(\mathbf{f} \mid \mathbf{y}) \,\middle\|\, q(\mathbf{f})\right]$ from approximate to true posterior), or variational inference (minimizing $\mathrm{D_{KL}}\left[q(\mathbf{f}) \,\middle\|\, p(\mathbf{f} \mid \mathbf{y})\right]$). EP is expected to be the most accurate method (see discussion in Vehtari et al., 2016) and Laplace to have the smallest computational overhead.

Learning under the GP paradigm In probabilistic machine learning, 'learning' typically amounts to finding point estimates for the hyperparameters $\boldsymbol{\theta}$ in the likelihood, mean function, and kernel by optimizing w.r.t. the log marginal likelihood:

$$\text{Learning target:} \quad \boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} \log p(\mathbf{y}; \boldsymbol{\theta}). \tag{4}$$

For Gaussian likelihoods, the marginal likelihood is available in closed form. For non-conjugate models, we can only optimize a proxy to the marginal likelihood $p(\mathbf{y}; \boldsymbol{\theta}) = \int p(\mathbf{y} \mid \mathbf{f}; \boldsymbol{\theta}) \, p(\mathbf{f}; \boldsymbol{\theta}) \, \mathrm{d}\mathbf{f}$, which depends on the approximate inference scheme and how it represents the posterior.

2.1. Laplace Approximation (LA)

A local Taylor expansion of the log posterior gives the Laplace approximation (LA, Williams & Barber, 1998). By defining $\Psi(\mathbf{f}) = \log(p(\mathbf{y} \mid \mathbf{f}) \, p(\mathbf{f}; \boldsymbol{\theta}))$, the approximate posterior $q(\mathbf{f})$ is obtained through a second-order Taylor expansion of $\Psi(\mathbf{f})$ around its maximum at $\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} \Psi(\mathbf{f})$ (the posterior mode): $p(\mathbf{f} \mid \mathbf{y}; \boldsymbol{\theta}) \propto \exp(\Psi(\mathbf{f})) \approx \exp\left(\Psi(\hat{\mathbf{f}}) + \frac{1}{2}(\mathbf{f} - \hat{\mathbf{f}})^\top \nabla^2 \Psi(\mathbf{f})|_{\mathbf{f}=\hat{\mathbf{f}}} (\mathbf{f} - \hat{\mathbf{f}})\right)$. This is proportional to $\mathrm{N}(\mathbf{f} \mid \hat{\mathbf{f}}, \mathbf{A}^{-1}) = q(\mathbf{f})$, where $\mathbf{A} = -\nabla^2 \Psi(\mathbf{f})|_{\mathbf{f}=\hat{\mathbf{f}}}$ is the Hessian of the negative log posterior at $\hat{\mathbf{f}}$. The log marginal likelihood is approximated as

$$\begin{aligned} \log p(\mathbf{y}; \boldsymbol{\theta}) &= \log \textstyle\int \exp(\Psi(\mathbf{f})) \, \mathrm{d}\mathbf{f} \\ &\approx \log \textstyle\int \exp\left(\Psi(\hat{\mathbf{f}}) - \frac{1}{2}(\mathbf{f} - \hat{\mathbf{f}})^\top \mathbf{A} (\mathbf{f} - \hat{\mathbf{f}})\right) \mathrm{d}\mathbf{f}. \end{aligned} \tag{5}$$

2.2. Expectation Propagation (EP)

Expectation Propagation (EP, Minka, 2001) is based on an approximation $q(\mathbf{f})$ that factorizes in the same way as the target posterior $p(\mathbf{f} \mid \mathbf{y}; \boldsymbol{\theta}) \propto p(\mathbf{f}; \boldsymbol{\theta}) \prod_{i=1}^n p(y_i \mid f_i; \boldsymbol{\theta})$: each likelihood term is approximated with a *site* function $t_i(f_i; \boldsymbol{\zeta}_i)$, and

$$q(\mathbf{f}; \boldsymbol{\theta}, \boldsymbol{\zeta}) \propto p(\mathbf{f}; \boldsymbol{\theta}) \textstyle\prod_{i=1}^n t_i(f_i; \boldsymbol{\zeta}_i). \tag{6}$$

For GP models, the sites $t_i(f_i; \boldsymbol{\zeta}_i)$ are chosen to be (unnormalized) Gaussians, and hence the global approximation $q(\mathbf{f})$ is also Gaussian. EP aims to minimize $\mathrm{D_{KL}}\left[p(\mathbf{f} \mid \mathbf{y}; \boldsymbol{\theta}) \,\middle\|\, q(\mathbf{f}; \boldsymbol{\theta}, \boldsymbol{\zeta})\right]$ w.r.t. $\boldsymbol{\zeta}$. This KL cannot be computed directly. Instead, EP updates the sites in an iterative fashion; the parameters of one site $\boldsymbol{\zeta}_i$ are tuned by minimizing the local Kullback–Leibler divergence

$$\begin{aligned} \mathrm{D_{KL}}\Big[p(y_i \mid f_i; \boldsymbol{\theta}) \, p(\mathbf{f}; \boldsymbol{\theta}) \textstyle\prod_{j \neq i} t_j(f_j; \boldsymbol{\zeta}_j) \\ \Big\| \, t_i(f_i; \boldsymbol{\zeta}_i) \, p(\mathbf{f}; \boldsymbol{\theta}) \textstyle\prod_{j \neq i} t_j(f_j; \boldsymbol{\zeta}_j)\Big], \end{aligned} \tag{7}$$

where in the first argument the $n - 1$ other likelihood terms have been replaced by their current site approximation. The optimal values of $\boldsymbol{\zeta}_i$ in this step can be determined by matching the first two moments. This iterative process often works well in practice, but can be numerically unstable (*e.g.*, for Student-t likelihood) and is not guaranteed to converge in the general case (see Vehtari et al., 2020).

The log marginal likelihood is directly approximated as

$$\log p(\mathbf{y}; \boldsymbol{\theta}) \approx \mathcal{L}_\text{EP}(\boldsymbol{\zeta}, \boldsymbol{\theta}) = \log \int p(\mathbf{f}; \boldsymbol{\theta}) \prod_{i=1}^n t_i(f_i; \boldsymbol{\zeta}_i) \, \mathrm{d}\mathbf{f}, \tag{8}$$

which is known to lead to a good objective for learning hyperparameters (see Jylänki et al., 2011).

2.3. Variational Inference (VI)

Variational Inference (VI, Opper & Archambeau, 2009) approximates the GP posterior $p(\mathbf{f} \mid \mathbf{y}; \boldsymbol{\theta})$ with a Gaussian distribution $q(\mathbf{f}; \boldsymbol{\xi})$ parameterized by $\boldsymbol{\xi}$. VI minimizes the reverse KL $\mathrm{D_{KL}}\left[q(\mathbf{f}; \boldsymbol{\xi}) \,\middle\|\, p(\mathbf{f} \mid \mathbf{y}; \boldsymbol{\theta})\right]$ by maximizing the following evidence lower bound (ELBO):

$$\begin{aligned} \log p(\mathbf{y}; \boldsymbol{\theta}) \geq \mathcal{L}_\text{VI}(\boldsymbol{\xi}, \boldsymbol{\theta}) = \sum_{i=1}^n \mathbb{E}_{q(f_i; \boldsymbol{\xi}_i)}\left[\log p(y_i \mid f_i; \boldsymbol{\theta})\right] \\ - \mathrm{D_{KL}}\left[q(\mathbf{f}; \boldsymbol{\xi}) \,\middle\|\, p(\mathbf{f}; \boldsymbol{\theta})\right], \end{aligned} \tag{9}$$

w.r.t. variational parameters $\boldsymbol{\xi}$. VI optimizes a lower bound on the marginal likelihood, so is guaranteed to converge, which is a strength over EP. As known from Hensman et al. (2013) and motivated by Khan et al. (2013), it has been desirable to *not* use the optimal parameterization in terms of $2n$ parameters, as the resulting optimization problem is non-convex. Instead, it is common to declare a variational distribution over the full posterior, $q(\mathbf{f};\boldsymbol{\xi}) = \mathrm{N}(\mathbf{m},\mathbf{S})$, and optimize the ELBO w.r.t. this mean–covariance parameterization[1] $\boldsymbol{\xi} = (\mathbf{m},\mathbf{S})$ using a general-purpose optimizer (*e.g.* Adam, Kingma & Ba, 2015).

In practice, the same lower bound $\mathcal{L}_{\mathrm{VI}}(\boldsymbol{\xi},\boldsymbol{\theta})$ is used to optimize variational parameters as well as hyperparameters, *i.e.*, inference and learning are coupled into a single optimization. This approach is commonplace, even though it is well-known to result in biased hyperparameters (Kuss & Rasmussen, 2005; Nickisch & Rasmussen, 2008; Bui et al., 2017).

3. Learning in the Dual Parameterization

We design a hybrid training procedure that augments VI with an EP-like learning target for hyperparameter learning. Our work builds upon the dual parameterization (Khan & Lin, 2017). Because the Gaussian distribution is part of the exponential family, we can write the approximate posterior as $q(\mathbf{f}) = \mathrm{N}(\mathbf{m},\mathbf{S}) = \exp\left(\boldsymbol{\eta}^\top \mathbf{T}(\mathbf{f}) - a(\boldsymbol{\eta})\right)$, where $\boldsymbol{\eta} = (\mathbf{S}^{-1}\mathbf{m}, -\frac{1}{2}\mathbf{S}^{-1})$, $\mathbf{T}(\mathbf{f}) = (\mathbf{f},\mathbf{f}\mathbf{f}^\top)$ are the sufficient statistics, and $\exp(-a(\boldsymbol{\eta}))$ is a normalization term. This leads to two additional parameterizations of $q(\mathbf{f})$: using the natural parameters $\boldsymbol{\eta}$, or using the expectation parameters $\boldsymbol{\mu} = \mathbb{E}_{q(\mathbf{f})}[\mathbf{T}(\mathbf{f})] = (\mathbf{m}, \mathbf{S}+\mathbf{m}\mathbf{m}^\top)$.

Khan & Lin (2017) showed that in the natural parameterization of the approximate posterior, *natural gradient descent* (NGD, Amari, 1998) (in the natural parameters space $\boldsymbol{\eta}$) will have the same computational cost as ordinary gradient descent on $\boldsymbol{\xi} = (\mathbf{m},\mathbf{S})$. The approximate posterior under this parameterization is

$$q(\mathbf{f};\boldsymbol{\lambda},\boldsymbol{\theta}) \propto p(\mathbf{f};\boldsymbol{\theta}) \prod\nolimits_{i=1}^{n} \underbrace{\exp\langle\boldsymbol{\lambda}_i, \mathbf{T}(f_i)\rangle}_{\triangleq\, t_i(f_i;\boldsymbol{\lambda}_i)}, \tag{10}$$

where $\boldsymbol{\lambda}_i = \nabla_{\boldsymbol{\mu}_i}\mathbb{E}_{q(f_i;\boldsymbol{\lambda}_i,\boldsymbol{\theta})}[\log p(y_i \mid f_i;\boldsymbol{\theta})]$. The natural parameters of the approximate posterior $q(\mathbf{f})$ are $\boldsymbol{\eta} = \boldsymbol{\lambda}_0 + \boldsymbol{\lambda}$, where $\boldsymbol{\lambda}_0 = (\mathbf{0}, -\frac{1}{2}\mathbf{K}^{-1})$ are the natural parameters of the prior $p(\mathbf{f};\boldsymbol{\theta})$ and $\boldsymbol{\lambda}$ are the parameters of the likelihood approximation term $t(\mathbf{f})$. Then, we could also parameterize the approximate posterior with $\boldsymbol{\lambda}$, to which we refer as the 'dual' $\boldsymbol{\lambda}$ parameterization.

Crucially, the approximate posterior (10) has the same form as its EP counterpart (6). This links EP with VI, which is the starting point for our proposed learning objective. The similarity *per se* has been visible in, *e.g.*, Chang et al. (2020); Adam et al. (2021), but it had not been explored further.

Algorithm 1 Training procedure for improved hyperparameter learning by a VEM-style iteration.

```
Initialize var. parameters $\boldsymbol{\lambda}^{(0)}$ and hyperparameters $\boldsymbol{\theta}^{(0)}$
Specify total training iteration $K$, the number of E-steps and M-steps per iteration $J_{\mathrm{E}}$, $J_{\mathrm{M}}$
for $k = 1, 2, \ldots, K$ do
    With $J_E$ nat. grad. steps and learning rate $\rho_E$, optimize
        $\boldsymbol{\lambda}^{(k)} \leftarrow \arg\max_{\boldsymbol{\lambda}} \mathcal{L}_{\mathrm{VI}}\left(\boldsymbol{\lambda}, \boldsymbol{\theta}^{(k-1)}\right)$
    With $J_M$ grad. steps and learning rate $\rho_M$, optimize
        $\boldsymbol{\theta}^{(k)} \leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{EP}}\left(\boldsymbol{\lambda}^{(k)}, \boldsymbol{\theta}\right)$
    if $\mathcal{L}_{\mathrm{EP}}\left(\boldsymbol{\lambda}^{(k)}, \boldsymbol{\theta}^{(k)}\right) < \mathcal{L}_{\mathrm{EP}}\left(\boldsymbol{\lambda}^{(k-1)}, \boldsymbol{\theta}^{(k-1)}\right)$ then
        return $\boldsymbol{\lambda}^{(k-1)}, \boldsymbol{\theta}^{(k-1)}$
    end if
end for
return $\boldsymbol{\lambda}^{(k)}, \boldsymbol{\theta}^{(k)}$
```

3.1. Our Proposed Objective for Learning

Natural gradient descent can efficiently optimize the variational parameters $\boldsymbol{\lambda}$, and we can combine it with another optimizer for the hyperparameters, leading to a natural separation of the inference and learning steps (Salimbeni et al., 2018). As discussed by Adam et al. (2021), this can be seen as a Variational Expectation–Maximization (VEM) procedure. Under this setup, inference/learning is performed by alternating between optimizing the variational distribution in the $\boldsymbol{\lambda}$ parameterization and taking gradient steps for finding $\boldsymbol{\theta}$ by iterating the following steps at the k^{th} iteration:

$$\text{E-step (inference):}\quad \boldsymbol{\lambda}^{(k+1)} \leftarrow \arg\max_{\boldsymbol{\lambda}} \mathcal{L}_{\mathrm{E}}(\boldsymbol{\lambda},\boldsymbol{\theta}^{(k)}),$$

$$\text{M-step (learning):}\quad \boldsymbol{\theta}^{(k+1)} \leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{M}}(\boldsymbol{\lambda}^{(k+1)},\boldsymbol{\theta}),$$

where the objective for both the inference and learning steps is the ELBO in Eq. (9) under the $\boldsymbol{\lambda}$ parameterization: $\mathcal{L}_{\mathrm{E}} \equiv \mathcal{L}_{\mathrm{M}} \equiv \mathcal{L}_{\mathrm{VI}}$. Note: Even if the parameterization and optimization procedure are different, the inference and learning objective are the same as in Sec. 2.3 and typically expected to converge at the same optima.

VEM deals with the variational inference problem by casting inference into an optimization problem that is solved by NGD, which appears both principled and efficient. We conjecture that the ELBO in Eq. (9) is not the best objective. Conveniently, the dual parameterization in the VI posterior Eq. (10) is formed as a product of the prior and Gaussian sites $t_i(f_i;\boldsymbol{\lambda}_i)$ just as in EP (*cf.* Eq. (6)). This provides a representation of the posterior that is directly EP-like and allows us to estimate the log marginal likelihood by plugging $\boldsymbol{\lambda}_i$ from Eq. (10) into $\boldsymbol{\zeta}_i$ in Eq. (8):

$$\mathcal{L}_{\mathrm{M}} \equiv \mathcal{L}_{\mathrm{EP}}(\boldsymbol{\lambda},\boldsymbol{\theta}) = \log\int p(\mathbf{f};\boldsymbol{\theta}) \prod\nolimits_{i=1}^{n} t_i(f_i;\boldsymbol{\lambda}_i)\,\mathrm{d}\mathbf{f}, \tag{11}$$

giving the target for the M-step.

---

[1] In practice it may be beneficial to optimize in the *whitened* (or *non-centered*) parameterization $\boldsymbol{\xi} = (\mathbf{m}', \mathbf{S}')$ s.t. $\mathbf{m} = \mathbf{L}\mathbf{m}'$ and $\mathbf{S} = \mathbf{L}\mathbf{S}'\mathbf{L}^\top$, where $\mathbf{L} = \mathrm{Cholesky}(\mathbf{K})$ (Gorinova et al., 2020).

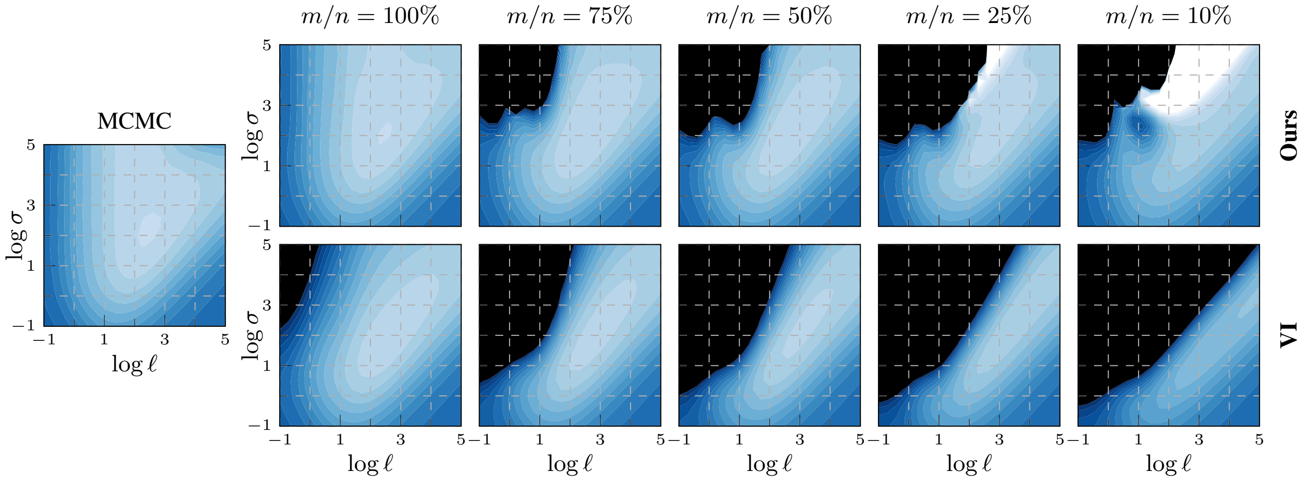

Figure 3. **Sparse approximation:** log marginal likelihood surfaces for the IONOSPHERE data set, changing the fraction m/n of the number of inducing points m vs. $n = 351$ data points. The colour scale is the same in all plots: -0.8 ▬ 0; values below the predefined range are plotted as black. For moderate sparsification, our EP-like marginal likelihood estimation (top) matches the full MCMC baseline better than VI (bottom). For extreme sparsification (10%: $m = 35$), neither approximation resembles the full surface.

Our hybrid training procedure uses the variational objective in the E-step to ensure a good representation of the posterior. Then in the M-step, we use an EP-like (and thus closer to marginal likelihood) objective for hyperparameter learning, at no additional computational cost. The algorithm is sketched out in Alg. 1. Although our procedure requires implementing two training objectives, this is not likelihood-specific and has minimal implementation overhead.

4. Sparse Approximation for Large Data Sets

Regardless of conjugacy, inference in GP models for large-scale data sets is challenging due to an $\mathcal{O}(n^3)$ computational bottleneck. In this section, we extend our hybrid training procedure to the sparse case.

A common approach to tackle this scalability issue is to summarize the information contained in the original data set into a smaller but more effective *pseudo*-data set, making the computational complexity tractable (see Fig. 3 and Quiñonero-Candela & Rasmussen, 2005). The pseudo-inputs are referred to as inducing points and denoted as $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^m$, where $m \ll n$ (Seeger, 2003; Csató, 2002; Quiñonero-Candela & Rasmussen, 2005; Williams & Seeger, 2001). The pseudo-outputs are referred to as inducing variables and denoted as $\mathbf{u} = f(\mathbf{Z})$.

One common choice for the form of approximate posterior, as first introduced in Titsias (2009), is $q(\mathbf{f}, \mathbf{u}; \boldsymbol{\xi}_{\mathbf{u}}, \boldsymbol{\theta}) = p(\mathbf{f} \,|\, \mathbf{u}; \boldsymbol{\theta})\, q(\mathbf{u}; \boldsymbol{\xi}_{\mathbf{u}})$, where $p(\mathbf{f} \,|\, \mathbf{u})$ is the GP conditional and $q(\mathbf{u}; \boldsymbol{\xi}_{\mathbf{u}}) = \mathrm{N}(\mathbf{m}_{\mathbf{u}}, \mathbf{S}_{\mathbf{u}})$ the approximate posterior in $\mathbf{u}$. In this form, $\mathbf{y}$ can only affect $\mathbf{f}$ through $\mathbf{u}$, which means the information in the original data set is summarized in $\mathbf{u}$.

The marginal posterior over the function $f(\cdot)$ is

$$q_{\mathbf{u}}(f(\cdot); \boldsymbol{\xi}_{\mathbf{u}}, \boldsymbol{\theta}) = \int p(f(\cdot) \,|\, \mathbf{u}; \boldsymbol{\theta})\, q(\mathbf{u}; \boldsymbol{\xi}_{\mathbf{u}})\, \mathrm{d}\mathbf{u}, \tag{12}$$

where $p(f(\cdot) \,|\, \mathbf{u}; \boldsymbol{\theta})$ is the distribution of the GP prior conditioned on $f(\mathbf{Z}) = \mathbf{u}$. Variational parameters $\boldsymbol{\xi}_{\mathbf{u}} = (\mathbf{m}_{\mathbf{u}}, \mathbf{S}_{\mathbf{u}})$ can be inferred by optimizing the following ELBO:

$$\begin{aligned}\log p(\mathbf{y}; \boldsymbol{\theta}) &\geq \mathcal{L}_{\mathrm{VI}}^{\mathrm{sparse}}(\boldsymbol{\xi}_{\mathbf{u}}, \boldsymbol{\theta}) \\ &= \textstyle\sum_{i=1}^n \mathbb{E}_{q_{\mathbf{u}}(f_i; \boldsymbol{\xi}_{\mathbf{u}}, \boldsymbol{\theta})}[\log p(y_i \,|\, f_i; \boldsymbol{\theta})] \\ &\quad - \mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{u}; \boldsymbol{\xi}_{\mathbf{u}}) \,\middle\|\, p(\mathbf{u}; \boldsymbol{\theta})\right],\end{aligned} \tag{13}$$

where $q_{\mathbf{u}}(f_i; \boldsymbol{\xi}_{\mathbf{u}}, \boldsymbol{\theta}) = \mathrm{N}(f_i \,|\, \mathbf{a}_i^\top \mathbf{m}_{\mathbf{u}}, \kappa_{ii} - \mathbf{a}_i^\top (\mathbf{K}_{\mathbf{uu}} - \mathbf{S}_{\mathbf{u}})\mathbf{a}_i)$, $\mathbf{a}_i = \mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{k}_{\mathbf{u},i}$, $\mathbf{K}_{\mathbf{uu}} = \kappa(\mathbf{Z}, \mathbf{Z})$, and $\mathbf{k}_{\mathbf{u},i} = \kappa(\mathbf{Z}, \mathbf{x}_i)$ for the i^{th} data sample. Now under the dual parameterization, we denote the converged dual parameters of the posterior marginal $q_{\mathbf{u}}^*(f_i)$ as $\boldsymbol{\lambda}_i^*$. Adam et al. (2021) suggest designing a similar VEM procedure that exploits the structure of the $q(\mathbf{u})$ in terms of the $2n$ dual parameters. The natural parameters of $q^*(\mathbf{u})$ are

$$(\mathbf{S}_{\mathbf{u}}^*)^{-1}\mathbf{m}_{\mathbf{u}}^* = \mathbf{K}_{\mathbf{uu}}^{-1} \underbrace{\left(\textstyle\sum_{i=1}^n \mathbf{k}_{\mathbf{u}i}\lambda_{1,i}^*\right)}_{=\bar{\boldsymbol{\lambda}}_1^*} \tag{14}$$

$$(\mathbf{S}_{\mathbf{u}}^*)^{-1} = \mathbf{K}_{\mathbf{uu}}^{-1} + \mathbf{K}_{\mathbf{uu},}^{-1} \underbrace{\left(\textstyle\sum_{i=1}^n \mathbf{k}_{\mathbf{u}i}\lambda_{2,i}^*\mathbf{k}_{\mathbf{u},i}^\top\right)}_{\bar{\boldsymbol{\Lambda}}_2^*} \mathbf{K}_{\mathbf{uu}}^{-1}, \tag{15}$$

where the quantities $\mathbf{K}_{\mathbf{uu}}$ and $\mathbf{k}_{\mathbf{u},i}$ directly depend on $\boldsymbol{\theta}$ and we can express the ELBO as the partition function of a Gaussian distribution. Adam et al. (2021) use a *tied* parameterization (motivated by site-tying in EP, see Bui et al., 2017; Li et al., 2015) that relaxes the need of storing all the $\{\boldsymbol{\lambda}_i^*\}_{i=1}^n$ and instead stores only $\bar{\boldsymbol{\lambda}}_1^*$ (length m) and $\bar{\boldsymbol{\Lambda}}_2^*$ (size $m \times m$), which avoids the storage issue for large data sets. This extends the results of Khan & Lin to the sparse case where the resulting approximate posterior is

$$q(\mathbf{f}, \mathbf{u}; \boldsymbol{\lambda}_\mathbf{u}, \boldsymbol{\theta}) \propto p(\mathbf{f} \mid \mathbf{u}; \boldsymbol{\theta})\, p(\mathbf{u}; \boldsymbol{\theta}) \times \prod_{i=1}^n \underbrace{\exp \langle \boldsymbol{\lambda}_{\mathbf{u},i}, \mathbf{T}(\mathbf{a}_i^\top \mathbf{u}) \rangle}_{\triangleq\, t_i(\mathbf{u}; \boldsymbol{\lambda}_{\mathbf{u},i})}, \tag{16}$$

where $\boldsymbol{\lambda}_{\mathbf{u},i} = \nabla_{\boldsymbol{\mu}_{\mathbf{u},i}} \mathbb{E}_{q_\mathbf{u}(f_i; \boldsymbol{\lambda}_{\mathbf{u},i}, \boldsymbol{\theta})}[\log p(y_i \mid f_i; \boldsymbol{\theta})]$. This gives rise to the sparse E-step for inference under the sparse VEM scheme, where $\mathcal{L}_\text{E}^\text{sparse} \equiv \mathcal{L}_\text{VI}^\text{sparse}$.

Our proposed sparse objective for learning In EP, the tied representation for constraining the problem to a summary $\boldsymbol{\zeta}_\mathbf{u}$ that scales in m rather than n gives rise to a sparse expectation propagation approach (Bui et al., 2017), where the log marginal likelihood is approximated as

$$\log p(\mathbf{y}; \boldsymbol{\theta}) \approx \mathcal{L}_\text{EP}^\text{sparse}(\boldsymbol{\zeta}_\mathbf{u}, \boldsymbol{\theta}) = \log \int q(\mathbf{f}, \mathbf{u}; \boldsymbol{\zeta}_\mathbf{u}, \boldsymbol{\theta})\, \mathrm{d}\mathbf{f}\, \mathrm{d}\mathbf{u}$$

$$= \log \int p(\mathbf{f} \mid \mathbf{u}; \boldsymbol{\theta})\, p(\mathbf{u}; \boldsymbol{\theta}) \prod_{i=1}^n t_i(\mathbf{u}; \boldsymbol{\zeta}_{\mathbf{u},i})\, \mathrm{d}\mathbf{f}\, \mathrm{d}\mathbf{u}. \tag{17}$$

Under dual parameterization VI, the approximate posterior Eq. (16) has the same structure as the EP approximate posterior in Eq. (17). An EP-like estimate of the log marginal likelihood can thus be calculated by injecting $\boldsymbol{\lambda}_{\mathbf{u},i}$ from Eq. (16) into $\boldsymbol{\zeta}_{\mathbf{u},i}$ in Eq. (17), thus giving $\mathcal{L}_\text{M}^\text{sparse} \equiv \mathcal{L}_\text{EP}^\text{sparse}(\boldsymbol{\lambda}_\mathbf{u}, \boldsymbol{\theta})$ which is a sparse EP-like learning objective under sparse variational inference. Note: $\mathcal{L}_\text{EP}^\text{sparse}(\boldsymbol{\lambda}_\mathbf{u}, \boldsymbol{\theta})$ has the same computational cost as VI.

5. Experiments

We provide a range of experiments, in which we demonstrate *effectiveness* and *practicality* of the proposed approach, and highlight similarities and differences between learning under the three most common approximative inference methods (LA, EP, and VI).

As the log marginal likelihood is a surrogate for the generalization ability of the model to unseen data, we evaluate the marginal likelihood estimations of different methods (Sec. 5.1) to see whether our EP-like marginal likelihood provides a better learning target. We then evaluate our hybrid training on non-conjugate tasks in binary classification and Student-t regression on small and mid-sized data sets (Secs. 5.2 and 5.3).

How the sparsity affects the learning target is not obvious, therefore in Sec. 5.4 we first investigate the influence of sparse approximation on the learning target. We then evaluate our hybrid training procedure on binary classification tasks with sparse approximation. For all experiments in the main paper, we use an isotropic Matérn-$^5/_2$ kernel. We also provide results under automatic relevance determination (ARD) with the same kernel (in App. D.4), where we only include results for data sets that could be confirmed to have converged.

We implement the variational methods in GPflow (Matthews et al., 2017), use reference implementations of LA and EP from GPy (GPy, since 2012), and base our MCMC implementation on the GPML toolbox (Rasmussen & Nickisch, 2010). Additionally, we use the GPstuff toolbox (Vanhatalo et al., 2013) for the custom LA and EP implementation for the Student-t likelihood. We implement EP and VI convergence checks; details in App. D.4.

5.1. Quality of Marginal Likelihood Approximations

We compare the quality of marginal likelihood approximations of LA, VI, EP, and our EP-like VI with gold-standard MCMC. We demonstrate this on a binary classification task on the IONOSPHERE data set, with SONAR, USPS, PARKINSONS, and MONKS-2 in App. C.2. We estimate the log marginal likelihood on a 21×21 grid of values for the log hyperparameters $\log \boldsymbol{\theta} = (\log \ell, \log \sigma)$ and plot the contour on the grid. For each hyperparameter setting, we fix the hyperparameters and evaluate the approximate log marginal likelihood based on the inferred approximate posterior.

Markov Chain Monte Carlo baseline MCMC is exact in the limit of long runs and thus provides a gold standard for log marginal likelihood estimation. Kuss & Rasmussen (2005) and Nickisch & Rasmussen (2008) proposed a sampling scheme based on annealed importance sampling (AIS, Neal, 2001) for obtaining a good estimate of the marginal likelihood (see App. C.1 for details). The baseline was computed by running $21 \times 21 = 441$ jobs in parallel on a cluster.

Experiment results As shown in the top row of Fig. 2 on the IONOSPHERE benchmark data set, the marginal likelihood estimation of EP closely matches the MCMC baseline, whereas that of VI looks clearly different. Notably, when we estimate the marginal likelihood by $\mathcal{L}_\text{EP}(\boldsymbol{\lambda}, \boldsymbol{\theta})$ using the 'site' parameters of dual VI (Ours), the contour shapes become much closer to the MCMC result, demonstrating the improvement of using this EP-like marginal likelihood estimation. To investigate whether the improved marginal likelihood estimation also leads to better generalization, we select the optimal hyperparameter location across the grid values and compare the log predictive density on the test set (bottom row of Fig. 2). The optimal hyperparameter location of EP-like VI (Ours) is closer to MCMC than VI and generalizes well. We show the same analysis for different data sets covering different types of classification tasks (from general classification to small images) in App. C.2, with the same conclusion. In Figs. 6 to 9, VI and EP conform to their

Table 1. **Binary classification:** log predictive density (higher is better) on different data sets from the *Bayesian benchmarks* over 5-fold cross-validation with 10 different seeds. Best results and those not statistically significantly different from them under a paired t-test are **bolded**. We provide MCMC results for reference and exclude it from bolding. MCMC gives the best results on all data sets except Balloons. All inference methods perform well overall, while our training objective delivers the most reliable performance.

	(n, d)	LA	EP	VI	Ours	MCMC
TRAINS	(10, 30)	-0.702 ± 0.025	$\mathbf{-0.698\pm0.033}$	-0.702 ± 0.037	$\mathbf{-0.691\pm0.046}$	-0.692 ± 0.025
BALLOONS	(16, 5)	-0.660 ± 0.125	-0.650 ± 0.128	-0.649 ± 0.185	$\mathbf{-0.607\pm0.227}$	-0.684 ± 0.076
FERTILITY	(100, 10)	$\mathbf{-0.388\pm0.122}$	$\mathbf{-0.384\pm0.149}$	$\mathbf{-0.393\pm0.136}$	$\mathbf{-0.397\pm0.139}$	-0.382 ± 0.126
PITTSBURG-BRIDGES-T-OR-D	(102, 8)	$\mathbf{-0.299\pm0.081}$	-0.321 ± 0.108	$\mathbf{-0.290\pm0.110}$	$\mathbf{-0.293\pm0.116}$	-0.306 ± 0.115
ACUTE-NEPHRITIS	(120, 7)	-0.203 ± 0.012	-0.046 ± 0.007	-0.007 ± 0.002	$\mathbf{-0.005\pm0.002}$	-0.005 ± 0.002
ACUTE-INFLAMMATION	(120, 7)	-0.184 ± 0.018	-0.052 ± 0.007	-0.007 ± 0.002	$\mathbf{-0.007\pm0.002}$	-0.007 ± 0.003
ECHOCARDIOGRAM	(131, 11)	-0.424 ± 0.093	$\mathbf{-0.418\pm0.095}$	$\mathbf{-0.425\pm0.110}$	-0.428 ± 0.112	-0.437 ± 0.127
HEPATITIS	(155, 20)	$\mathbf{-0.370\pm0.071}$	-0.372 ± 0.072	$\mathbf{-0.364\pm0.090}$	-0.367 ± 0.094	-0.369 ± 0.091
PARKINSONS	(195, 23)	-0.260 ± 0.031	-0.295 ± 0.056	-0.160 ± 0.050	$\mathbf{-0.141\pm0.046}$	-0.145 ± 0.044
BREAST-CANCER-WISC-PROG	(198, 34)	$\mathbf{-0.458\pm0.075}$	-0.473 ± 0.091	$\mathbf{-0.457\pm0.085}$	$\mathbf{-0.460\pm0.088}$	-0.464 ± 0.085
SPECT	(265, 23)	$\mathbf{-0.593\pm0.049}$	$\mathbf{-0.590\pm0.055}$	-0.594 ± 0.054	-0.595 ± 0.054	-0.596 ± 0.051
STATLOG-HEART	(270, 14)	-0.395 ± 0.064	$\mathbf{-0.389\pm0.061}$	-0.396 ± 0.071	-0.397 ± 0.071	-0.397 ± 0.070
HABERMAN-SURVIVAL	(306, 4)	$\mathbf{-0.530\pm0.053}$	-0.532 ± 0.059	-0.531 ± 0.055	-0.531 ± 0.055	-0.520 ± 0.063
IONOSPHERE	(351, 34)	-0.224 ± 0.042	-0.230 ± 0.042	$\mathbf{-0.170\pm0.048}$	$\mathbf{-0.170\pm0.055}$	-0.179 ± 0.058
HORSE-COLIC	(368, 26)	-0.463 ± 0.059	$\mathbf{-0.452\pm0.057}$	-0.467 ± 0.072	-0.473 ± 0.082	-0.469 ± 0.079
CONGRESSIONAL-VOTING	(435, 17)	-0.640 ± 0.028	$\mathbf{-0.639\pm0.030}$	-0.641 ± 0.030	-0.642 ± 0.029	-0.644 ± 0.027
CYLINDER-BANDS	(512, 36)	-0.488 ± 0.038	-0.500 ± 0.041	-0.465 ± 0.049	$\mathbf{-0.451\pm0.052}$	-0.451 ± 0.049
BREAST-CANCER-WISC-DIAG	(569, 31)	-0.085 ± 0.026	-0.140 ± 0.020	-0.077 ± 0.044	$\mathbf{-0.075\pm0.045}$	-0.076 ± 0.043
ILPD-INDIAN-LIVER	(583, 10)	$\mathbf{-0.513\pm0.040}$	-0.520 ± 0.041	$\mathbf{-0.512\pm0.043}$	$\mathbf{-0.512\pm0.043}$	-0.512 ± 0.042
MONKS-2	(601, 7)	-0.491 ± 0.025	-0.512 ± 0.028	-0.464 ± 0.031	$\mathbf{-0.442\pm0.033}$	-0.437 ± 0.032
STATLOG-AUSTRALIAN-CREDIT	(690, 15)	$\mathbf{-0.630\pm0.026}$	-0.639 ± 0.036	$\mathbf{-0.630\pm0.026}$	$\mathbf{-0.630\pm0.026}$	-0.630 ± 0.025
CREDIT-APPROVAL	(690, 16)	$\mathbf{-0.342\pm0.047}$	$\mathbf{-0.342\pm0.050}$	$\mathbf{-0.341\pm0.052}$	-0.342 ± 0.052	-0.341 ± 0.052
BREAST-CANCER-WISC	(699, 10)	$\mathbf{-0.094\pm0.025}$	$\mathbf{-0.093\pm0.023}$	$\mathbf{-0.093\pm0.029}$	-0.093 ± 0.029	-0.093 ± 0.029
BLOOD	(748, 5)	$\mathbf{-0.478\pm0.039}$	$\mathbf{-0.479\pm0.040}$	-0.478 ± 0.039	-0.478 ± 0.039	-0.478 ± 0.039
PIMA	(768, 9)	$\mathbf{-0.474\pm0.033}$	$\mathbf{-0.476\pm0.038}$	$\mathbf{-0.474\pm0.035}$	$\mathbf{-0.474\pm0.035}$	-0.474 ± 0.035
MAMMOGRAPHIC	(961, 6)	$\mathbf{-0.407\pm0.038}$	$\mathbf{-0.407\pm0.040}$	$\mathbf{-0.408\pm0.040}$	-0.408 ± 0.040	-0.408 ± 0.040
STATLOG-GERMAN-CREDIT	(1000, 25)	$\mathbf{-0.491\pm0.030}$	$\mathbf{-0.491\pm0.032}$	$\mathbf{-0.492\pm0.032}$	$\mathbf{-0.492\pm0.032}$	-0.492 ± 0.032
Bold Count		14	13	13	16	/

stereotypes of being over- and under-confident, respectively, while Ours tends to have slightly better calibration.

For completeness, and motivated by the seminal work of Kuss & Rasmussen (2005), we provide back-to-back comparisons of both marginal log likelihood and predictive density surfaces also for LA. In Fig. 2, the marginal likelihood surface of LA resembles that of VI, while for the predictive density surface VI more closely resembles MCMC compared to LA.

5.2. Non-Conjugate Tasks in Bayesian Benchmarks

The log marginal likelihood is a surrogate for the generalization ability of the model to unseen data. To explicitly evaluate the generalization ability of our hybrid training procedure, we compare it against LA, EP, and VI on commonly used benchmark classification tasks. We use common small and mid-sized data sets ($n \leq 1000$) and do full GP inference with 5-fold cross-validation. We use the *Bayesian Benchmarks* suite (github.com/secondmind-labs/bayesian_benchmarks) for evaluating the methods.

Evaluation on binary classification We consider binary classification with a Bernoulli likelihood on 27 data sets from the UCI repository (Dua & Graff, 2017). For all approximate inference methods, we set the same number of maximum training iterations and use the relative changes in the parameters of the model as convergence criteria. For our hybrid training procedure due to the conflicting objectives in E- and M-steps discussed in Sec. 5.5, we use a decrease in the EP learning objective as an additional convergence criterion. As a gold-standard baseline, we include MCMC results. For MCMC, we use a log-uniform hyperpriors to ensure a close match to the model setup in the other models.

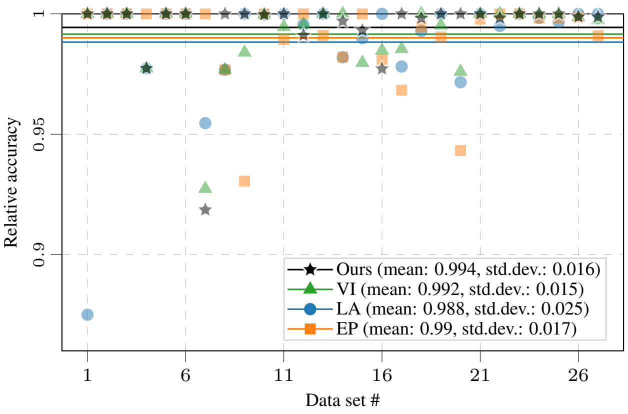

Figure 4. Mean relative accuracy compared to best method on each data set of Table 1 (over 5-fold CV repeated with 10 seeds). The horizontal lines indicate mean across all data sets; see legend for mean and standard deviation. Our approach yields reliable training, with the highest average relative accuracy and the least outliers.

To reduce the variance introduced by the training–test set split, we repeat the 5-fold CV with ten different seeds. The

Table 2. **Robust regression tasks:** log predictive density (higher is better) with a Student-t likelihood on different data sets (mean ± standard deviation over 5-fold cross-validation). The best results and those not statistically significantly different from them under a paired t-test are **bolded**. Our objective performs well overall.

	(n, d)	LA	EP	VI	Ours
NEAL	(100, 1)	$\mathbf{.317}{\pm}.440$	$\mathbf{.303}{\pm}.432$	$\mathbf{.295}{\pm}.426$	$\mathbf{.301}{\pm}.436$
BOSTON	(506, 13)	$-\mathbf{.210}{\pm}.069$	$-\mathbf{.190}{\pm}.053$	$-.206{\pm}.056$	$-\mathbf{.195}{\pm}.061$
STOCK	(1000, 1)	$1.910{\pm}.079$	$1.389{\pm}.242$	$1.917{\pm}.082$	$\mathbf{1.921}{\pm}.083$
Bold Count		2	2	1	3

Table 3. **Sparse approximation** for classification on different data sets: log predictive density (mean ± standard deviation). Higher is better. Results that are statistically significantly different under a paired t-test are **bolded**.

	LA	EP	VI	Ours
TITANIC	$-.217{\pm}.037$	$-.014{\pm}.004$	$-\mathbf{.011}{\pm}\mathbf{.003}$	$-.037{\pm}.005$
BANK	$-\mathbf{.247}{\pm}\mathbf{.006}$	$-\mathbf{.246}{\pm}\mathbf{.007}$	$-.249{\pm}.006$	$-.247{\pm}.007$
TWONORM	$-\mathbf{.060}{\pm}\mathbf{.007}$	$-\mathbf{.061}{\pm}\mathbf{.008}$	$-\mathbf{.060}{\pm}\mathbf{.008}$	$-.524{\pm}.208$
MUSHROOM	$-.129{\pm}.003$	$-.002{\pm}.000$	$-\mathbf{.001}{\pm}\mathbf{.000}$	$-.028{\pm}.001$
MAGIC	$-\mathbf{.285}{\pm}\mathbf{.333}$	$-.693{\pm}.000$	$-\mathbf{.008}{\pm}\mathbf{.001}$	$-.070{\pm}.002$
Bold Count	3	2	4	0

performance on the test set is given in Table 1 and Fig. 4. As shown in Table 1 for log predictive density, LA and EP have very similar performance to VI-based methods (VI and Ours) on most data sets. This empirically demonstrates that for binary classification on small and mid-sized data sets EP and LA generalize well. Our hybrid training procedure achieves the same test performance as VI on most data sets and outperforms VI on three data sets. It empirically demonstrates that when no sparse approximation is required, by using an improved estimation of the marginal likelihood for hyperparameter learning, we could potentially have better generalization ability at no additional computational cost. As the gold standard, MCMC gives the best results; notably, the gap between MCMC and approximate inference methods is relatively small on small data sets. In practice we often favour methods with stable performance over different data sets, *i.e.*, they might not always give the best performance but we can expect consistently good performances. To investigate the reliability of different methods, in Fig. 4 we plot the relative accuracy (on each data set we divide the results of each method by the highest accuracy on that data set) for individual data sets and the mean relative accuracy of each method across all data sets. Our approach achieves the most consistent performance on all data sets and thus yields reliable training. We include additional result tables with the same conclusion in App. D.3, including experiments with an ARD kernel and checks for initializing other methods with our optimal hyperparameters.

5.3. Robust (Student-t) Regression

We further test our hybrid training procedure on a more challenging robust regression task with a Student-t likelihood, a model which is not log-concave. In the likelihood, we fix the degrees of freedom, $\nu = 3$, and only train the noise scale together with hyperparameters. For LA and EP we follow the methods designed by Jylänki et al. (2011). For VI and our EP-like VI, to make the training procedure numerically stable we crop the gradient w.r.t. the second element of the natural parameters to prevent the approximate posterior covariance from becoming negative. We test on three benchmark data sets previously used for robust regression: the simulated data from Neal (2022), the Boston housing regression task, and the stock data from Solin & Särkkä (2015). Jylänki et al. (2011) point out that in Student-t regression EP provides a good approximation for marginal likelihood and, as shown in Table 2, by using an EP-like marginal likelihood estimation for hyperparameters learning our hybrid training procedure generalizes better than vanilla VI. The MCMC gold standard results for NEAL, BOSTON and STOCK are $0.309{\pm}0.454$, $-0.191{\pm}0.051$, and $1.586{\pm}0.034$, respectively.

5.4. Evaluation under Sparse Approximation

It is not obvious how a sparse approximation would affect the quality of marginal likelihood approximation. To be able to compare with the MCMC baseline on the full data set, we first analyse the influence of sparsification on IONOSPHERE. We choose 75%, 50%, 25%, and 10% random subsets of training data as inducing points. We estimate the log marginal likelihood as in Sec. 5.1 on a grid of values for the log hyperparameters. Fig. 3 shows the resulting contour surfaces. Unsurprisingly, as we reduce the number of inducing points, the estimation of the log marginal likelihood becomes less accurate. For moderate sparsification (75%, 50%), our EP-like marginal likelihood estimation matches the full MCMC baseline better than VI. For more extreme sparsification, both approximations show significant biases. This is because, with very few inducing points, only larger lengthscales make sense. To match the low-complexity approximation, large lengthscales are required and the ground truth marginal likelihood provided by MCMC becomes irrelevant.

Evaluation on large-scale binary classification We compare LA, EP, VI, and our proposed training procedure on five data sets from 2k to 19k data points (for details, see Table 11). We use k-means to select 500 inducing points from the input data and keep them fixed. The test set performance is given in Table 3. Here we are in the regime of m/n in the range of 25% to 2.5%, where the log marginal likelihood surface approximations of both VI and our EP-like approximation are likely to be biased away from the true marginal likelihood surface, and the approximate sparse model can no longer be considered a surrogate of the true model.

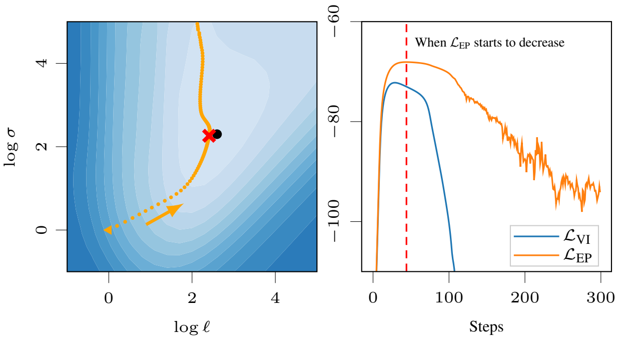

Figure 5. Interplay between the two optimization targets can result in overshooting the optimum (●). The optimizer starts at ◂ and once it passes ✖ (dashed red; stopping point), $\mathcal{L}_{\mathrm{EP}}$ starts to decrease. After that, the optimization becomes increasingly unstable.

5.5. Practical Considerations

VEM with the same objective for E- and M-step is analogous to coordinate ascent and hence guaranteed to always improve the objective. With $\mathcal{L}_{\mathrm{VI}}$ for E-step and $\mathcal{L}_{\mathrm{EP}}$ for M-step, this guarantee no longer holds. This can introduce interplay between the two targets, and in our experiments, we observed that (with certain data-splits/model setups) the optimization can overshoot past the optimum and then becomes increasingly unstable, see Fig. 5. In our experiments this conflict between $\mathcal{L}_{\mathrm{E}} = \mathcal{L}_{\mathrm{VI}}$ and $\mathcal{L}_{\mathrm{M}} = \mathcal{L}_{\mathrm{EP}}$ occurred in about 54% of cases. We address this issue by ending optimization once the $\mathcal{L}_{\mathrm{EP}}$ objective starts to decrease (see Alg. 1). Note that this is solely based on the training data and does not require any additional validation data or tuning per data set.

6. Discussion and Conclusions

In GP models, the training separates into inference and learning which are typically both cast into an optimization problem. In this paper, we improved hyperparameter learning in non-conjugate GP models by augmenting VI with an EP-like learning target. Our hybrid training procedure builds upon the dual variational GP formulation, introduced by Khan & Lin (2017) and extended to sparse GPs in Adam et al. (2021), which provides a link between VI and EP. We used the representation of the posterior from VI to obtain an EP-like approximation of the marginal likelihood for hyperparameter optimization—without any added computational complexity or computation time.

In the experiments, we evaluated our hybrid training procedure on binary classification tasks and robust regression. For full (non-sparse) models, the extensive results (over 1350 runs per method) show clear benefits in stability, reliability, and performance for our method. This shows the benefits of decoupling inference and learning. When more hyperparameters are present, as shown in Table 5 and Table 6 in the appendix, our method has the same performance on log predictive density as VI. However, as shown in Fig. 10, our method is still more reliable than VI. For sparse problems, similar empirical benefits could not be demonstrated.

We provide a reference implementation of the methods and code to reproduce the experiments at https://github.com/AaltoML/improved-hyperparameter-learning.

Acknowledgements

This work was supported by funding from the Academy of Finland (grant id 339730) and the Finnish Center for Artificial Intelligence (FCAI). We acknowledge the computational resources provided by the Aalto Science-IT project and CSC – IT Center for Science, Finland. We thank Aidan Scannell for his constructive comments on the manuscript.

References

Adam, V., Chang, P. E., Khan, M. E., and Solin, A. Dual parameterization of sparse variational Gaussian processes. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, volume 34, pp. 11474–11486. Curran Associates, Inc., 2021.

Amari, S.-I. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Bui, T. D., Yan, J., and Turner, R. E. A unifying framework for Gaussian process pseudo-point approximations using power expectation propagation. *Journal of Machine Learning Research (JMLR)*, 18:3649–3720, 2017.

Chang, P. E., Wilkinson, W. J., Khan, M. E., and Solin, A. Fast variational learning in state-space Gaussian process models. In *International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2020.

Csató, L. *Gaussian Processes: Iterative Sparse Approximations*. PhD thesis, Aston University, Birmingham, UK, 2002.

Dua, D. and Graff, C. UCI machine learning repository, 2017. http://archive.ics.uci.edu/ml.

Gorinova, M., Moore, D., and Hoffman, M. Automatic reparameterisation of probabilistic programs. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3648–3657. PMLR, 2020.

GPy. GPy: A Gaussian process framework in python. http://github.com/SheffieldML/GPy, since 2012.

Hensman, J., Fusi, N., and Lawrence, N. D. Gaussian processes for big data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 282–290. AUAI Press, 2013.

Jylänki, P., Vanhatalo, J., and Vehtari, A. Robust Gaussian process regression with a Student-t likelihood. *Journal of Machine Learning Research (JMLR)*, 12:3227–3257, 2011.

Khan, M. E. and Lin, W. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *Proceedings of Machine Learning Research*, pp. 878–887. PMRL, 2017.

Khan, M. E., Aravkin, A., Friedlander, M., and Seeger, M. Fast dual variational inference for non-conjugate latent Gaussian models. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 28 of *Proceedings of Machine Learning Research*, pp. 951–959. PMLR, 2013.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Kuss, M. and Rasmussen, C. E. Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research (JMLR)*, 6:1679–1704, 2005.

Li, Y., Hernández-Lobato, J. M., and Turner, R. E. Stochastic expectation propagation. In *Advances in Neural Information Processing Systems 28*, pp. 2323–2331. Curran Associates, Inc., 2015.

Lotfi, S., Izmailov, P., Benton, G., Goldblum, M., and Wilson, A. G. Bayesian model selection, the marginal likelihood, and generalization. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14223–14247. PMLR, 2022.

Matthews, A. G. d. G., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., Ghahramani, Z., and Hensman, J. GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research (JMLR)*, 18(40):1–6, 2017.

Minka, T. P. Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, Proceedings of Machine Learning Research, pp. 362–369, 2001.

Murray, I., Adams, R., and MacKay, D. Elliptical slice sampling. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 541–548. JMLR Workshop and Conference Proceedings, 2010.

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

Neal, R. M. Software for flexible Bayesian modeling and Markov Chain sampling, 2022. http://www.cs.toronto.edu/~radford/fbm.software.html.

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

Opper, M. and Archambeau, C. The variational Gaussian approximation revisited. *Neural Computation*, 21(3):786–92, 2009.

Quiñonero-Candela, J. and Rasmussen, C. E. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research (JMLR)*, 6:1939–1959, 2005.

Rasmussen, C. E. and Nickisch, H. Gaussian processes for machine learning (GPML) toolbox. *Journal of Machine Learning Research (JMLR)*, 11:3011–3015, 2010.

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

Salimbeni, H., Eleftheriadis, S., and Hensman, J. Natural gradients in practice: Non-conjugate variational inference in Gaussian process models. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84 of *Proceedings of Machine Learning Research*, pp. 689–697. PMLR, 2018.

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

Solin, A. and Särkkä, S. State space methods for efficient inference in Student-t process regression. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38 of *Proceedings of Machine Learning Research*, pp. 885–893. PMLR, 2015.

Titsias, M. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pp. 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR.

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., and Vehtari, A. GPstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research (JMLR)*, 14(1):1175–1179, 2013.

Vapnick, V. N. *Statistical Learning Theory*. Wiley, New York, 1998.

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., and Winther, O. Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *Journal of Machine Learning Research (JMLR)*, 17(1):3581–3618, 2016.

Vehtari, A., Gelman, A., Sivula, T., Jylänki, P., Tran, D., Sahai, S., Blomstedt, P., Cunningham, J. P., Schiminovich, D., and Robert, C. P. Expectation propagation as a way of life. *Journal of Machine Learning Research (JMLR)*, 21:1–53, 2020.

Wilkinson, W. J., Särkkä, S., and Solin, A. Bayes–Newton methods for approximate Bayesian inference with PSD guarantees. *Journal of Machine Learning Research (JMLR)*, 24(83):1–50, 2023.

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

Williams, C. K. and Seeger, M. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13 (NIPS)*, pp. 682–688. MIT Press, 2001.

Appendix

In the supplementary material, we include technical details of the methods that were omitted for brevity in the main paper (App. A). Additionally, we provide details on the experiments and evaluation setup (App. B–F) for reproducing the results in the main paper, and include further result tables and figures that extend the evaluation. The codes for the methods proposed in this paper are included as a separate supplement.

A. Method Details

Sparse energy To extend the presentation in Sec. 4, we derive how to obtain the sparse EP marginal likelihood estimation from the VI approximate posterior. Following Eq. (77) in Bui et al. (2017), the sites $t_i(\mathbf{u};\boldsymbol{\zeta}_{\mathbf{u},i})$ in Eq. (17) are

$$t_i(\mathbf{u};\boldsymbol{\zeta}_{\mathbf{u},i}) \propto \exp\langle \mathbf{u}^\top \mathbf{a}_i \boldsymbol{\zeta}_{\mathbf{u},i,2}\boldsymbol{\zeta}_{\mathbf{u},i,1} - \frac{1}{2}\mathbf{u}^\top \mathbf{a}_i \boldsymbol{\zeta}_{\mathbf{u},i,2}\mathbf{a}_i^\top \mathbf{u}\rangle, \tag{18}$$

where $\mathbf{a}_i = \mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{k}_{\mathbf{u},i}$, $\mathbf{K}_{\mathbf{uu}} = \kappa(\mathbf{Z},\mathbf{Z})$, and $\mathbf{k}_{\mathbf{u},i} = \kappa(\mathbf{Z},\mathbf{x}_i)$ for the i^{th} data sample. Note that $\mathbf{u}^\top\mathbf{a}_i = \mathbf{a}_i^\top\mathbf{u}$ is a scalar. As $t_i(\mathbf{u};\boldsymbol{\lambda}_{\mathbf{u},i})$ in Eq. (16) is given by

$$\begin{aligned} t_i(\mathbf{u};\boldsymbol{\lambda}_{\mathbf{u},i}) &= \exp\langle \boldsymbol{\lambda}_{\mathbf{u},i}, \mathbf{T}(\mathbf{a}_i^\top\mathbf{u})\rangle \\ &= \exp\langle \boldsymbol{\lambda}_{\mathbf{u},i,1}\mathbf{a}_i^\top\mathbf{u} + \boldsymbol{\lambda}_{\mathbf{u},i,2}(\mathbf{a}_i^\top\mathbf{u})^2\rangle, \end{aligned} \tag{19}$$

we have the following correspondence between $\boldsymbol{\zeta}_{\mathbf{u}}$ and $\boldsymbol{\lambda}_{\mathbf{u}}$:

$$\boldsymbol{\zeta}_{\mathbf{u},i,2}\boldsymbol{\zeta}_{\mathbf{u},i,1} \Leftrightarrow \boldsymbol{\lambda}_{\mathbf{u},i,1} \qquad \text{and} \qquad -\frac{1}{2}\boldsymbol{\zeta}_{\mathbf{u},i,2} \Leftrightarrow \boldsymbol{\lambda}_{\mathbf{u},i,2}. \tag{20}$$

Following Eq. (126) in Bui et al. (2017), the sparse EP energy is (we omit $\boldsymbol{\theta}$ to make notation simpler)

$$\begin{aligned} \mathcal{L}_{\text{EP}}^{\text{sparse}}(\boldsymbol{\zeta}_{\mathbf{u}},\boldsymbol{\theta}) = \frac{1}{2}\log|\mathbf{S}_{\mathbf{u}}| + \frac{1}{2}\mathbf{m}_{\mathbf{u}}^\top\mathbf{S}_{\mathbf{u}}^{-1}\mathbf{m}_{\mathbf{u}} - \frac{1}{2}\log|\mathbf{K}_{\mathbf{uu}}| + \frac{1}{\alpha}\sum_n \log\mathcal{Z}_{\text{tilted},i} \\ + \sum_n \Big[-\frac{1}{2\alpha}\log(1-\mathbf{a}_i^\top\alpha\boldsymbol{\zeta}_{\mathbf{u},i,2}\mathbf{S}_{\mathbf{u}}\mathbf{a}_i) + \frac{1}{2}\frac{\mathbf{m}_{\mathbf{u}}^\top\mathbf{a}_i\boldsymbol{\zeta}_{\mathbf{u},i,2}\mathbf{a}_i^\top\mathbf{m}_{\mathbf{u}}}{1-\mathbf{a}_i^\top\alpha\boldsymbol{\zeta}_{\mathbf{u},i,2}\mathbf{S}_{\mathbf{u}}\mathbf{a}_i} \\ + \frac{1}{2}\boldsymbol{\zeta}_{\mathbf{u},i,1}\boldsymbol{\zeta}_{\mathbf{u},i,2}\mathbf{a}_i^\top\mathbf{V}_{\text{cav},i}\mathbf{a}_i\alpha\boldsymbol{\zeta}_{\mathbf{u},i,2}\boldsymbol{\zeta}_{\mathbf{u},i,1} - \boldsymbol{\zeta}_{\mathbf{u},i,1}\boldsymbol{\zeta}_{\mathbf{u},i,2}\mathbf{a}_i^\top\mathbf{V}_{\text{cav},i}\mathbf{S}_{\mathbf{u}}^{-1}\mathbf{m}_{\mathbf{u}}\Big], \end{aligned} \tag{21}$$

where the different terms are defined by

$$\mathbf{S}_{\text{cav},i} = \mathbf{S}_{\mathbf{u}} + \frac{\mathbf{S}_{\mathbf{u}}\mathbf{a}_i\alpha\boldsymbol{\zeta}_{\mathbf{u},i,2}\mathbf{a}_i^\top\mathbf{S}_{\mathbf{u}}}{1-\mathbf{a}_i^\top\,\alpha\,\boldsymbol{\zeta}_{\mathbf{u},i,2}\mathbf{S}_{\mathbf{u}}\mathbf{a}_i}, \tag{22}$$

$$\mathbf{S}_{\text{cav},i}^{-1}\mathbf{m}_{\text{cav},i} = \mathbf{S}_{\mathbf{u}}^{-1}\mathbf{m}_{\mathbf{u}} - \alpha\,\mathbf{a}_i\boldsymbol{\zeta}_{\mathbf{u},i,2}\boldsymbol{\zeta}_{\mathbf{u},i,1}, \tag{23}$$

$$\log\mathcal{Z}_{\text{tilted},i} = \log\int q_{\text{cav},i}(f_i)\,p^\alpha\,(y_i\,|\,f_i)\;\mathrm{d}f_i, \tag{24}$$

$$q_{\text{cav},i}(f_i) = \int p\,(f_i\,|\,\mathbf{u})\,\mathrm{N}(\mathbf{m}_{\text{cav},i},\mathbf{S}_{\text{cav},i})\,\mathrm{d}\mathbf{u}. \tag{25}$$

By substituting $\boldsymbol{\lambda}_{\mathbf{u},i}$ into $\boldsymbol{\zeta}_{\mathbf{u},i}$ in Eq. (21) using Eq. (20), we obtain the sparse EP marginal likelihood approximation with the VI approximate posterior. When $\alpha = 1$, Power-EP reduces to normal EP, which we use in our experiments.

B. Computational Details

All experiments ran on a cluster, which allowed us to parallelize jobs. This played a central role especially for the MCMC baseline results for the marginal likelihood surfaces, where we split into 441 separate jobs (per hyperparameter value combination), each of which were allocated 1–3 CPU cores and 1 Gb memory and ran 8–40 h depending on data set size.

C. Quality of Marginal Likelihood Approximation

Inspired by the work by Kuss & Rasmussen (2005) and Nickisch & Rasmussen (2008), we compare the quality of marginal likelihood approximations of LA, VI, EP, and our EP-like VI with a 'ground' truth obtained by annealed importance sampling (AIS, Neal, 2001). We demonstrate this on binary classification tasks, where we estimate the log marginal likelihood on a 21×21 grid of values for the log hyperparameters $\log \boldsymbol{\theta} = (\log \ell, \log \sigma)$ and plot the contour on the grid. For each hyperparameter setting, we fix the hyperparameters and evaluate the approximate log marginal likelihood based on the inferred approximate posterior. We then also visualize the log predictive density on hold-out test data as a similar contour plot, showing what the performance of the model would have been if the hyperparameters would have been chosen based on the log marginal likelihood surface in question under the specific inference scheme.

C.1. Markov Chain Monte Carlo Baseline

As in previous work, we use an MCMC approach as the gold-standard baseline. We use the annealed importance sampling (Neal, 2001) approach from Kuss & Rasmussen (2005) and Nickisch & Rasmussen (2008) that defines a sequence of $t = 0, 1, \ldots, T$ steps $Z_t = \int p(\mathbf{y} \,|\, \mathbf{f}; \boldsymbol{\theta})^{\tau(t)} p(\mathbf{f}; \boldsymbol{\theta}) \,\mathrm{d}\mathbf{f}$, where $\tau(t) = (t/T)^4$ (such that $\tau(0) = 0$ and $\tau(T) = 1$). The marginal likelihood can be rewritten as

$$p(\mathbf{y}; \boldsymbol{\theta}) = \frac{Z_T}{Z_0} = \frac{Z_T}{Z_{T-1}} \frac{Z_{T-1}}{Z_{T-2}} \cdots \frac{Z_1}{Z_0}, \tag{26}$$

where ${}^{Z_t}/_{Z_{t-1}}$ is approximated by importance sampling using samples from $q_t(\mathbf{f}) \propto p(\mathbf{y} \,|\, \mathbf{f}; \boldsymbol{\theta})^{\tau(t-1)} \, p(\mathbf{f}; \boldsymbol{\theta})$:

$$\begin{aligned}
\frac{Z_t}{Z_{t-1}} &= \frac{\int p(\mathbf{y} \,|\, \mathbf{f}; \boldsymbol{\theta})^{\tau(t)} p(\mathbf{f}; \boldsymbol{\theta}) \,\mathrm{d}\mathbf{f}}{Z_{t-1}} \\
&= \int \frac{p(\mathbf{y} \,|\, \mathbf{f}; \boldsymbol{\theta})^{\tau(t)}}{p(\mathbf{y} \,|\, \mathbf{f}; \boldsymbol{\theta})^{\tau(t-1)}} \frac{p(\mathbf{y} \,|\, \mathbf{f}; \boldsymbol{\theta})^{\tau(t-1)} p(\mathbf{f}; \boldsymbol{\theta})}{Z_{t-1}} \,\mathrm{d}\mathbf{f} \\
&\approx \frac{1}{S} \sum_{s=1}^{S} p(\mathbf{y} \,|\, \mathbf{f}_t^{(s)}; \boldsymbol{\theta})^{\tau(t) - \tau(t-1)}, \quad \text{where} \quad \mathbf{f}_t^{(s)} \sim \frac{p(\mathbf{y} \,|\, \mathbf{f}; \boldsymbol{\theta})^{\tau(t-1)} p(\mathbf{f}; \boldsymbol{\theta})}{Z_{t-1}}.
\end{aligned} \tag{27}$$

In practice, instead of sampling $\mathbf{f}_t$ from $\frac{p(\mathbf{y} \,|\, \mathbf{f})^{\tau(t-1)} \mathrm{N}(\mathbf{f} \,|\, \mathbf{m}, \mathbf{K})}{Z_{t-1}}$ directly, we use a parameterisation in terms of $\boldsymbol{\alpha} = \mathbf{K}^{-1}(\mathbf{f}_t - \mathbf{m})$ and sample from $\boldsymbol{\alpha} \sim P(\boldsymbol{\alpha}) = \frac{p(\mathbf{y} \,|\, \mathbf{K}\boldsymbol{\alpha} + \mathbf{m})^{\tau(t-1)} \mathrm{N}(\boldsymbol{\alpha} \,|\, \mathbf{0}, \mathbf{K}^{-1})}{Z_a}$ to increase numerical stability since $\log P(\boldsymbol{\alpha})$ and its gradient can be computed safely. We use elliptical slice sampling (Murray et al., 2010). Now, by using a single sample $S = 1$ and a large number of steps T, the estimation of log marginal likelihood can be written as

$$\log p(\mathbf{y}; \boldsymbol{\theta}) = \sum_{t=1}^{T} \log \frac{Z_t}{Z_{t-1}} \approx \sum_{t=1}^{T} (\tau(t) - \tau(t-1)) \log p(\mathbf{y} \,|\, \mathbf{f}_t; \boldsymbol{\theta}). \tag{28}$$

Following Kuss & Rasmussen (2005), we set $T = 8000$ and combine three estimates of log marginal likelihood by their geometric mean.

C.2. Experiment Results

In addition to the figure for the IONOSPHERE data set (Fig. 2) in the main paper, we include surface plots for four additional data sets in the appendix for more comprehensive comparisons. The marginal likelihood estimation on SONAR, USPS, PARKINSONS and MONKS-2 are given in Fig. 6, Fig. 7, Fig. 8, and Fig. 9, respectively. The IONOSPHERE and SONAR were included to make it easy to compare to previous work, and the other three chosen as an interesting subset covering different type of classification tasks (from general classification to small images).

Similarly to the results on IONOSPHERE in the main paper, when using the EP-like marginal likelihood estimation from the VI approximate posterior (Ours), the contour shapes become closer to the MCMC result, and the optimal hyperparameter location of EP-like VI (Ours) is closer to MCMC than VI. These effects help explain the quantitative results in the tables in the main paper and the supplement.

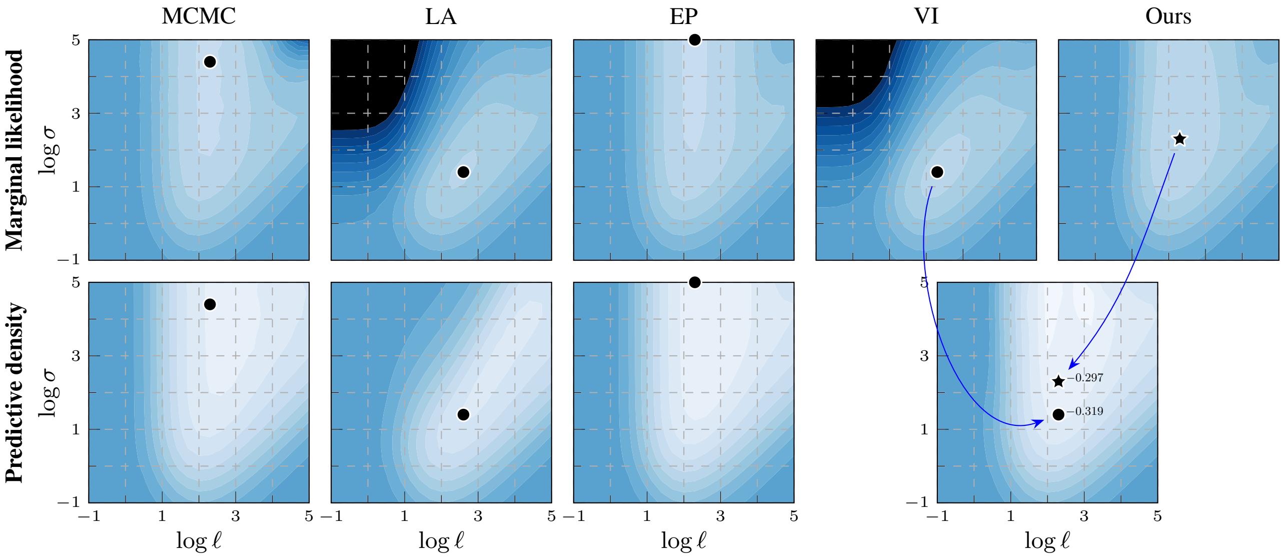

Figure 6. **Log marginal likelihood / predictive density surfaces** for the SONAR data set by varying kernel magnitude σ and lengthscale ℓ. The colour scale is the same in all plots: -1.0 ▬ -0.2 (normalized by n). Optimal hyperparameters of each method shown by a black marker. EP and our EP-like marginal likelihood estimation match the MCMC baseline better than VI or LA, thus providing a learning proxy. For prediction, our method still leverages the same variational representation as VI.

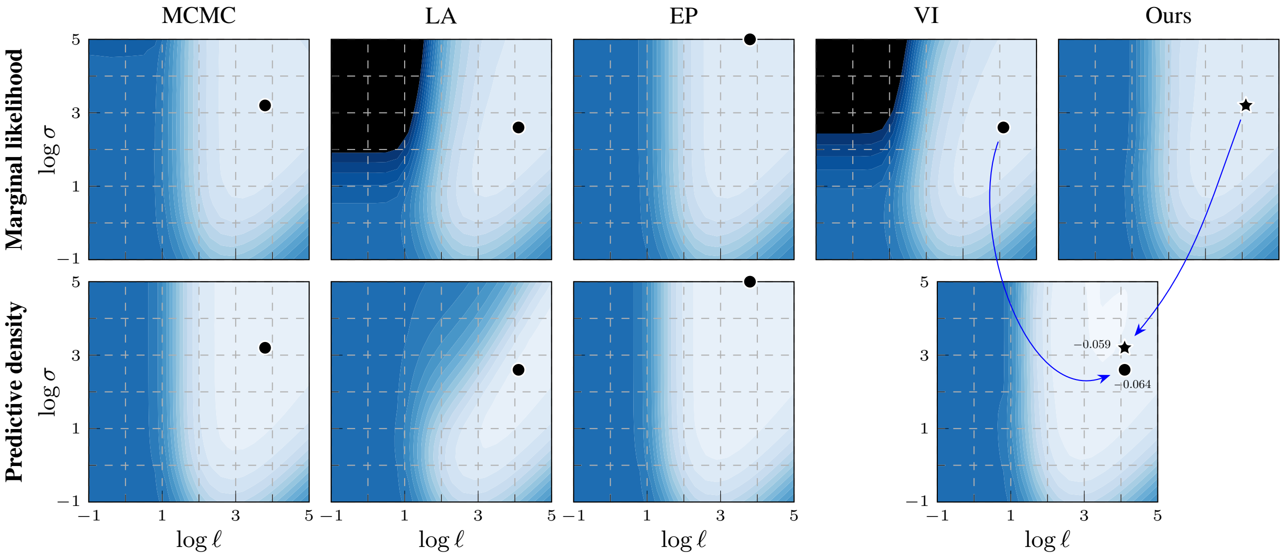

Figure 7. **Log marginal likelihood / predictive density surfaces** for the USPS data set (MNIST-like digits image) by varying kernel magnitude σ and lengthscale ℓ. The colour scale is the same in all plots: -1.0 ▬ 0 (normalized by n). Optimal hyperparameters shown by a black marker. EP and our EP-like marginal likelihood estimation match the MCMC baseline better than VI or LA, thus providing a learning proxy. For prediction, our method still leverages the same variational representation as VI.

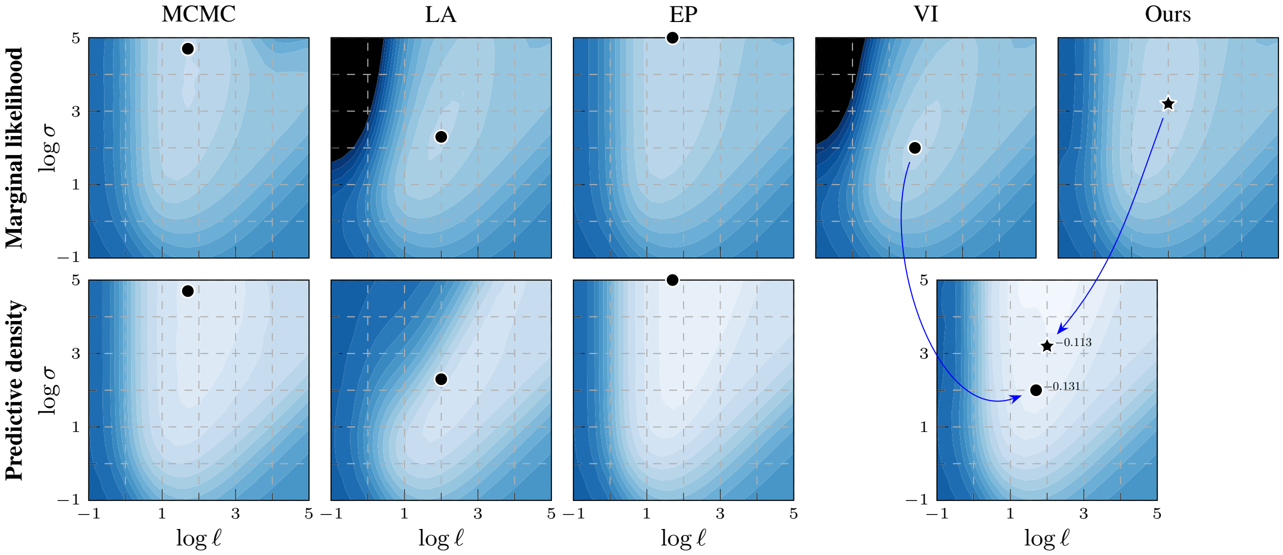

Figure 8. **Log marginal likelihood / predictive density surfaces** for the PARKINSONS data set by varying kernel magnitude σ and lengthscale ℓ. The colour scale is the same in all plots: -0.8 ▨ 0 (normalized by n). Optimal hyperparameters shown by a black marker. EP and our EP-like marginal likelihood estimation match the MCMC baseline better than VI or LA, thus providing a learning proxy. For prediction, our method still leverages the same variational representation as VI.

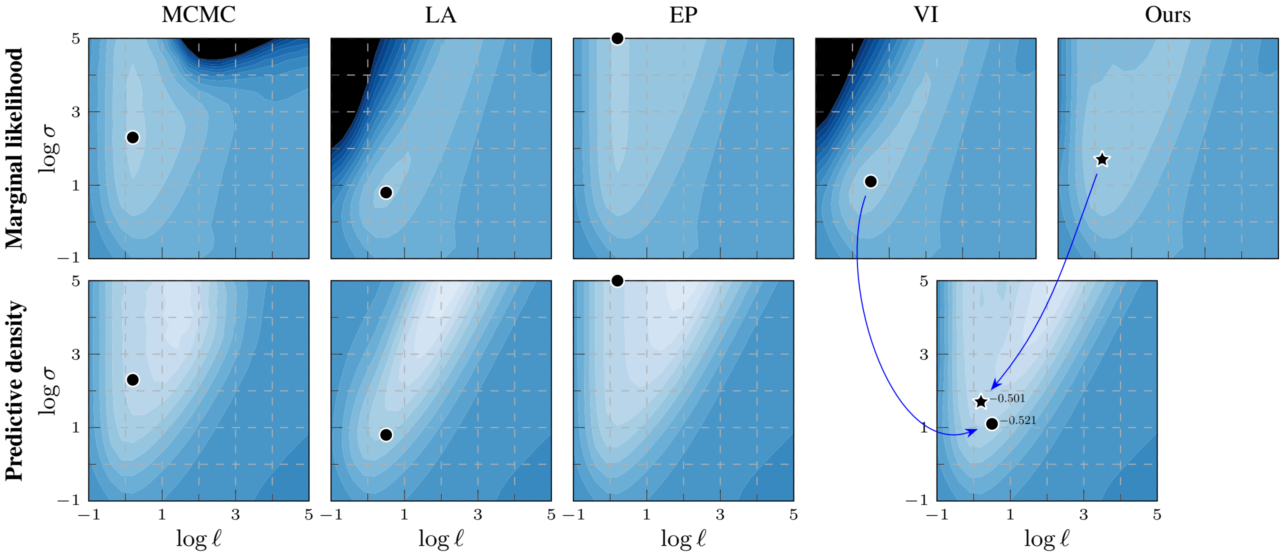

Figure 9. **Log marginal likelihood / predictive density surfaces** for the MONKS-2 data set by varying kernel magnitude σ and lengthscale ℓ. The colour scale is the same in all plots: -0.9 ▨ -0.3 (normalized by n). Optimal hyperparameters shown by a black marker. EP and our EP-like marginal likelihood estimation match the MCMC baseline better than VI or LA, thus providing a learning proxy. For prediction, our method still leverages the same variational representation as VI.

D. Non-Conjugate Tasks in Bayesian Benchmarks

We use the *Bayesian Benchmarks* suite (github.com/secondmind-labs/bayesian_benchmarks; originally by Salimbeni *et al.*) for evaluating the methods in binary classification. Bayesian benchmarks includes common evaluation data sets (typically from UCI) and makes it possible to run a large number of comparisons under a fixed evaluation setup. In the first part, we only include binary classification tasks (Bernoulli likelihood) with $n \leq 1000$. We follow the standard setup of input point normalization and splits in the evaluation suite.

D.1. Evaluation Metrics

We conduct 5-fold cross-validation and use test set accuracy and log predictive density to evaluate the test performance of each method (higher is better in both). To compare different methods, we use the paired t-test (with $p = 0.05$) that compares whether the best-performing method performs statistically significantly better than the others.

D.2. Experiment Setup

We initialize the hyperparameters with unit lengthscale and magnitude for all methods. For LA and EP, the hyperparameters are optimized by the default optimizer L-BFGS-B in GPy. For VI and our hybrid training procedure, each E step and M step consists of 20 iterations. In the E-step we set the learning rate of natural gradient descent to be 0.1. In the M-step we use the Adam optimizer (Kingma & Ba, 2015) with learning rate 0.01. We use the convergence criterion described in the main text, with a maximum number of at most $10\,000$ steps.

For MCMC, we use log uniform prior to ensure it is the same model as the approximate inference methods. We set the burn-in step to be 200, the number of samples to be 10000, and thin the sample with 2.

D.3. Experiment Results

The test set accuracies are given in Table 4. For test accuracies all methods achieve similar performance, which is to be expected as accuracy alone only captures where the decision boundary has been draw, completely disregarding second-order information. The log predictive density results are included in the main paper (Table 1). Our hybrid training procedure gives the most consistent results and thus achieves the most reliable training.

D.4. Ablation Studies

Automatic Relevance Determination Kernel We run experiments with automatic relevance determination (ARD) kernel to see whether our method would perform well when there are multiple hyperparameters. To ensure fair comparison we only include results on data sets where all methods have converged. The log predictive density and test set accuracy are given in Table 5 and Table 6 respectively. The mean relative accuracy is plotted in Fig. 10. Our training objective performs well overall and achieves reliable training.

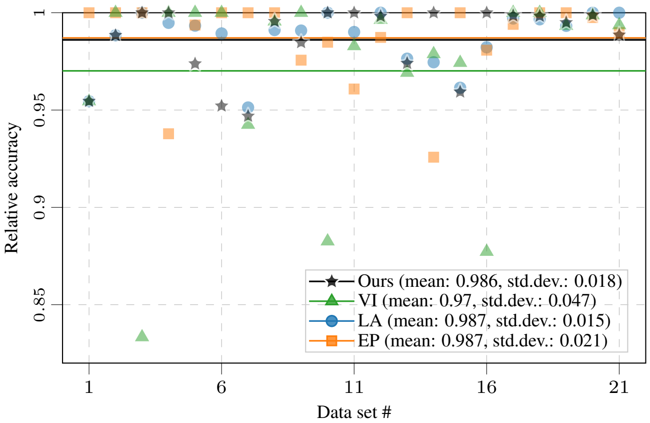

Figure 10. Mean relative accuracy compared to best method on each data set of Table 6. The horizontal lines indicate mean across all data sets; see legend for mean and standard deviation. Our approach yields reliable training.

Table 4. **Binary classification:** test set accuracy (higher is better) on different data sets from the *Bayesian benchmarks* over 5-fold cross-validation with 10 different seeds. Best results and those not statistically significantly different from them under a paired t-test are **bolded**. We provide MCMC results for reference (excluded from bolding). All inference methods perform well overall, while our training objective delivers the most reliable performance.

	(n, d)	LA	EP	VI	Ours	MCMC
TRAINS	(10, 30)	0.620 ± 0.394	$\mathbf{0.680\pm0.371}$	0.660 ± 0.353	$\mathbf{0.710\pm0.348}$	0.730 ± 0.349
BALLOONS	(16, 5)	$\mathbf{0.618\pm0.275}$	0.607 ± 0.282	$\mathbf{0.615\pm0.285}$	$\mathbf{0.650\pm0.291}$	0.625 ± 0.272
FERTILITY	(100, 10)	$\mathbf{0.879\pm0.062}$	$\mathbf{0.880\pm0.062}$	$\mathbf{0.879\pm0.062}$	$\mathbf{0.877\pm0.063}$	0.877 ± 0.061
PITTSBURG-BRIDGES-T-OR-D	(102, 8)	$\mathbf{0.874\pm0.070}$	$\mathbf{0.868\pm0.075}$	$\mathbf{0.875\pm0.072}$	$\mathbf{0.877\pm0.068}$	0.868 ± 0.069
ACUTE-NEPHRITIS	(120, 7)	$\mathbf{1.000\pm0.000}$	$\mathbf{1.000\pm0.000}$	$\mathbf{1.000\pm0.000}$	$\mathbf{1.000\pm0.000}$	1.000 ± 0.000
ACUTE-INFLAMMATION	(120, 7)	$\mathbf{1.000\pm0.000}$	$\mathbf{1.000\pm0.000}$	$\mathbf{1.000\pm0.000}$	$\mathbf{1.000\pm0.000}$	1.000 ± 0.000
ECHOCARDIOGRAM	(131, 11)	0.820 ± 0.064	$\mathbf{0.840\pm0.056}$	0.808 ± 0.069	0.808 ± 0.068	0.797 ± 0.073
HEPATITIS	(155, 20)	$\mathbf{0.830\pm0.059}$	0.819 ± 0.061	$\mathbf{0.833\pm0.059}$	$\mathbf{0.832\pm0.060}$	0.830 ± 0.057
PARKINSONS	(195, 23)	$\mathbf{0.951\pm0.031}$	0.888 ± 0.050	0.942 ± 0.032	$\mathbf{0.949\pm0.034}$	0.950 ± 0.031
BREAST-CANCER-WISC-PROG	(198, 34)	0.808 ± 0.058	0.793 ± 0.070	$\mathbf{0.815\pm0.057}$	$\mathbf{0.814\pm0.057}$	0.808 ± 0.060
SPECT	(265, 23)	$\mathbf{0.706\pm0.055}$	$\mathbf{0.703\pm0.057}$	$\mathbf{0.703\pm0.056}$	$\mathbf{0.705\pm0.056}$	0.707 ± 0.058
STATLOG-HEART	(270, 14)	0.830 ± 0.049	$\mathbf{0.838\pm0.047}$	0.830 ± 0.049	0.828 ± 0.050	0.822 ± 0.050
HABERMAN-SURVIVAL	(306, 4)	$\mathbf{0.741\pm0.048}$	$\mathbf{0.740\pm0.050}$	$\mathbf{0.741\pm0.049}$	$\mathbf{0.741\pm0.048}$	0.746 ± 0.047
IONOSPHERE	(351, 34)	0.927 ± 0.029	0.930 ± 0.035	$\mathbf{0.942\pm0.030}$	$\mathbf{0.945\pm0.027}$	0.943 ± 0.026
HORSE-COLIC	(368, 26)	0.810 ± 0.042	$\mathbf{0.817\pm0.040}$	0.805 ± 0.044	0.806 ± 0.043	0.805 ± 0.045
CONGRESSIONAL-VOTING	(435, 17)	$\mathbf{0.599\pm0.046}$	$\mathbf{0.597\pm0.046}$	$\mathbf{0.598\pm0.047}$	$\mathbf{0.597\pm0.049}$	0.600 ± 0.046
CYLINDER-BANDS	(512, 36)	0.778 ± 0.041	0.763 ± 0.045	0.782 ± 0.041	$\mathbf{0.794\pm0.040}$	0.795 ± 0.040
BREAST-CANCER-WISC-DIAG	(569, 31)	$\mathbf{0.969\pm0.015}$	$\mathbf{0.974\pm0.015}$	$\mathbf{0.970\pm0.015}$	$\mathbf{0.972\pm0.014}$	0.973 ± 0.016
ILPD-INDIAN-LIVER	(583, 10)	$\mathbf{0.718\pm0.035}$	$\mathbf{0.715\pm0.039}$	$\mathbf{0.719\pm0.035}$	$\mathbf{0.719\pm0.035}$	0.719 ± 0.036
MONKS-2	(601, 7)	0.758 ± 0.036	0.738 ± 0.034	0.760 ± 0.035	$\mathbf{0.773\pm0.035}$	0.778 ± 0.034
STATLOG-AUSTRALIAN-CREDIT	(690, 15)	$\mathbf{0.677\pm0.035}$	0.667 ± 0.032	$\mathbf{0.677\pm0.035}$	$\mathbf{0.677\pm0.035}$	0.678 ± 0.035
CREDIT-APPROVAL	(690, 16)	$\mathbf{0.859\pm0.031}$	0.856 ± 0.029	$\mathbf{0.860\pm0.030}$	$\mathbf{0.860\pm0.030}$	0.859 ± 0.030
BREAST-CANCER-WISC	(699, 10)	0.967 ± 0.013	$\mathbf{0.969\pm0.012}$	0.967 ± 0.013	0.967 ± 0.013	0.967 ± 0.013
BLOOD	(748, 5)	$\mathbf{0.783\pm0.033}$	$\mathbf{0.784\pm0.032}$	$\mathbf{0.783\pm0.034}$	$\mathbf{0.783\pm0.034}$	0.779 ± 0.031
PIMA	(768, 9)	$\mathbf{0.764\pm0.030}$	$\mathbf{0.765\pm0.031}$	$\mathbf{0.764\pm0.030}$	$\mathbf{0.764\pm0.030}$	0.763 ± 0.030
MAMMOGRAPHIC	(961, 6)	$\mathbf{0.823\pm0.026}$	$\mathbf{0.823\pm0.026}$	$\mathbf{0.823\pm0.026}$	$\mathbf{0.823\pm0.026}$	0.823 ± 0.026
STATLOG-GERMAN-CREDIT	(1000, 25)	$\mathbf{0.769\pm0.027}$	0.766 ± 0.028	$\mathbf{0.768\pm0.027}$	$\mathbf{0.768\pm0.027}$	0.767 ± 0.027
Bold Count		18	17	19	23	/

Table 5. **Binary classification:** log predictive density (higher is better) on different data sets with ARD kernel from the *Bayesian benchmarks* over 5-fold cross-validation. To ensure fair comparison we only include results on data sets where all methods have converged. Best results and those not statistically significantly different from them under a paired t-test are **bolded**.

	(n, d)	LA	EP	VI	Ours
FERTILITY	(100, 10)	$\mathbf{-0.417\pm0.096}$	$\mathbf{-0.407\pm0.120}$	$\mathbf{-0.439\pm0.116}$	-0.456 ± 0.115
PITTSBURG-BRIDGES-T-OR-D	(102, 8)	$\mathbf{-0.320\pm0.057}$	$\mathbf{-0.325\pm0.085}$	$\mathbf{-0.325\pm0.055}$	$\mathbf{-0.320\pm0.061}$
ACUTE-INFLAMMATION	(120, 7)	-0.112 ± 0.005	-0.029 ± 0.004	$\mathbf{-1.668\pm2.754}$	$\mathbf{-0.005\pm0.001}$
PARKINSONS	(195, 23)	-0.260 ± 0.029	-0.214 ± 0.086	$\mathbf{-0.062\pm0.062}$	$\mathbf{-0.048\pm0.042}$
BREAST-CANCER-WISC-PROG	(198, 34)	$\mathbf{-0.498\pm0.053}$	$\mathbf{-0.476\pm0.068}$	$\mathbf{-0.516\pm0.080}$	$\mathbf{-0.644\pm0.092}$
SPECT	(265, 23)	$\mathbf{-0.608\pm0.059}$	$\mathbf{-0.614\pm0.064}$	$\mathbf{-0.643\pm0.056}$	-0.646 ± 0.068
STATLOG-HEART	(270, 14)	-0.433 ± 0.066	$\mathbf{-0.398\pm0.056}$	-0.450 ± 0.072	-0.472 ± 0.082
HABERMAN-SURVIVAL	(306, 4)	$\mathbf{-0.535\pm0.049}$	$\mathbf{-0.531\pm0.053}$	$\mathbf{-0.539\pm0.050}$	$\mathbf{-0.540\pm0.051}$
IONOSPHERE	(351, 34)	$\mathbf{-0.243\pm0.058}$	$\mathbf{-0.253\pm0.032}$	$\mathbf{-0.208\pm0.086}$	$\mathbf{-0.231\pm0.061}$
CONGRESSIONAL-VOTING	(435, 17)	$\mathbf{-0.642\pm0.030}$	$\mathbf{-0.639\pm0.033}$	-0.687 ± 0.043	$\mathbf{-0.644\pm0.081}$
CYLINDER-BANDS	(512, 36)	$\mathbf{-0.468\pm0.063}$	-0.501 ± 0.061	$\mathbf{-0.454\pm0.068}$	-0.478 ± 0.064
BREAST-CANCER-WISC-DIAG	(569, 31)	$\mathbf{-0.098\pm0.024}$	-0.122 ± 0.031	$\mathbf{-0.075\pm0.044}$	$\mathbf{-0.073\pm0.046}$
ILPD-INDIAN-LIVER	(583, 10)	-0.531 ± 0.030	$\mathbf{-0.518\pm0.035}$	$\mathbf{-0.519\pm0.036}$	$\mathbf{-0.518\pm0.037}$
MONKS-2	(601, 7)	-0.458 ± 0.029	-0.485 ± 0.038	-0.424 ± 0.037	$\mathbf{-0.399\pm0.036}$
STATLOG-AUSTRALIAN-CREDIT	(690, 15)	$\mathbf{-0.633\pm0.022}$	$\mathbf{-0.627\pm0.009}$	$\mathbf{-0.633\pm0.022}$	$\mathbf{-0.632\pm0.023}$
CREDIT-APPROVAL	(690, 16)	-0.294 ± 0.102	-0.148 ± 0.015	$\mathbf{-0.905\pm1.126}$	$\mathbf{-0.070\pm0.028}$
BREAST-CANCER-WISC	(699, 10)	$\mathbf{-0.102\pm0.026}$	$\mathbf{-0.101\pm0.028}$	$\mathbf{-0.102\pm0.034}$	$\mathbf{-0.104\pm0.035}$
BLOOD	(748, 5)	$\mathbf{-0.477\pm0.047}$	$\mathbf{-0.477\pm0.049}$	$\mathbf{-0.477\pm0.048}$	$\mathbf{-0.477\pm0.048}$
PIMA	(768, 9)	$\mathbf{-0.473\pm0.034}$	$\mathbf{-0.466\pm0.042}$	$\mathbf{-0.474\pm0.037}$	$\mathbf{-0.474\pm0.036}$
MAMMOGRAPHIC	(961, 6)	$\mathbf{-0.399\pm0.039}$	$\mathbf{-0.401\pm0.041}$	$\mathbf{-0.400\pm0.040}$	$\mathbf{-0.400\pm0.040}$
STATLOG-GERMAN-CREDIT	(1000, 25)	$\mathbf{-0.488\pm0.044}$	$\mathbf{-0.485\pm0.046}$	-0.501 ± 0.053	$\mathbf{-0.503\pm0.054}$
Bold Count		15	15	17	17

Convergence of EP To make sure EP fully converged, we initialized it with the trained hyperparameter values of our method and the log predictive density results and test set accuracies are given in Table 7 and Table 8. Compared with Table 9 and Table 10, regarding log predictive density EP has three more bolded results and regarding test set accuracies EP has two more bolded results. This underlines some of the issues associated with EP and speaks in favour of our method.

Table 6. **Binary classification:** test set accuracy (higher is better) on different data sets with ARD kernel from the *Bayesian benchmarks* over 5-fold cross-validation. To ensure fair comparison we only include results on data sets where all methods have converged. Best results and those not statistically significantly different from them under a paired t-test are **bolded**.

	(n, d)	LA	EP	VI	Ours
FERTILITY	(100, 10)	**0.840±0.049**	**0.880±0.060**	**0.840±0.049**	**0.840±0.049**
PITTSBURG-BRIDGES-T-OR-D	(102, 8)	**0.853±0.003**	**0.863±0.036**	**0.863±0.019**	**0.853±0.003**
ACUTE-INFLAMMATION	(120, 7)	**1.000±0.000**	**1.000±0.000**	**0.833±0.173**	**1.000±0.000**
PARKINSONS	(195, 23)	**0.985±0.021**	**0.928±0.071**	**0.990±0.021**	**0.990±0.021**
BREAST-CANCER-WISC-PROG	(198, 34)	**0.768±0.029**	**0.768±0.048**	**0.773±0.053**	**0.753±0.049**
SPECT	(265, 23)	**0.702±0.056**	**0.709±0.054**	**0.709±0.057**	**0.675±0.066**
STATLOG-HEART	(270, 14)	0.796±0.039	**0.837±0.027**	0.789±0.049	**0.793±0.054**
HABERMAN-SURVIVAL	(306, 4)	**0.732±0.043**	**0.738±0.038**	**0.735±0.044**	**0.735±0.044**
IONOSPHERE	(351, 34)	**0.923±0.026**	**0.909±0.026**	**0.932±0.016**	0.917±0.016
CONGRESSIONAL-VOTING	(435, 17)	**0.607±0.027**	**0.598±0.030**	**0.536±0.091**	**0.607±0.034**
CYLINDER-BANDS	(512, 36)	**0.791±0.050**	0.768±0.043	**0.785±0.078**	**0.799±0.048**
BREAST-CANCER-WISC-DIAG	(569, 31)	**0.975±0.010**	**0.963±0.022**	**0.972±0.013**	**0.974±0.013**
ILPD-INDIAN-LIVER	(583, 10)	**0.707±0.033**	**0.724±0.033**	0.702±0.023	**0.705±0.030**
MONKS-2	(601, 7)	0.765±0.040	0.727±0.044	0.769±0.038	**0.785±0.031**
STATLOG-AUSTRALIAN-CREDIT	(690, 15)	**0.651±0.032**	**0.677±0.018**	**0.659±0.035**	**0.649±0.033**
CREDIT-APPROVAL	(690, 16)	**0.962±0.020**	0.961±0.013	**0.859±0.201**	**0.980±0.011**
BREAST-CANCER-WISC	(699, 10)	**0.961±0.011**	**0.958±0.011**	**0.964±0.010**	**0.963±0.012**
BLOOD	(748, 5)	**0.786±0.038**	**0.789±0.036**	**0.789±0.036**	**0.787±0.037**
PIMA	(768, 9)	**0.766±0.025**	**0.771±0.028**	**0.766±0.025**	**0.767±0.023**
MAMMOGRAPHIC	(961, 6)	**0.831±0.012**	**0.829±0.016**	**0.830±0.013**	**0.830±0.013**
STATLOG-GERMAN-CREDIT	(1000, 25)	**0.787±0.028**	**0.779±0.027**	**0.782±0.028**	**0.778±0.028**
Bold Count		19	18	18	20

Table 7. **Binary classification:** log predictive density (higher is better) on different data sets from the *Bayesian benchmarks* over 5-fold cross-validation. We initialize EP with the trained hyperparameters values of our method. Best results and those not statistically significantly different from them under a paired t-test are **bolded**. The results are largely the same as Table 9, which means EP has converged.

	(n, d)	LA	EP (new init.)	VI	Ours
TRAINS	(10, 30)	**−0.695±0.011**	**−0.680±0.042**	**−0.692±0.023**	**−0.681±0.042**
BALLOONS	(16, 5)	**−0.707±0.146**	**−0.672±0.263**	**−0.711±0.239**	**−0.657±0.267**
FERTILITY	(100, 10)	**−0.379±0.099**	**−0.379±0.103**	**−0.378±0.103**	**−0.379±0.103**
PITTSBURG-BRIDGES-T-OR-D	(102, 8)	**−0.306±0.044**	**−0.295±0.058**	**−0.295±0.057**	**−0.296±0.059**
ACUTE-NEPHRITIS	(120, 7)	−0.203±0.010	−0.010±0.003	−0.006±0.002	**−0.005±0.001**
ACUTE-INFLAMMATION	(120, 7)	−0.172±0.033	−0.013±0.002	−0.008±0.001	**−0.007±0.001**
ECHOCARDIOGRAM	(131, 11)	**−0.420±0.073**	**−0.421±0.086**	**−0.420±0.085**	**−0.421±0.086**
HEPATITIS	(155, 20)	**−0.369±0.043**	**−0.362±0.052**	**−0.361±0.051**	**−0.361±0.052**
PARKINSONS	(195, 23)	−0.267±0.030	−0.153±0.050	−0.164±0.062	**−0.145±0.054**
BREAST-CANCER-WISC-PROG	(198, 34)	**−0.460±0.063**	**−0.458±0.070**	**−0.458±0.071**	**−0.458±0.070**
SPECT	(265, 23)	**−0.587±0.045**	**−0.590±0.051**	**−0.589±0.050**	**−0.590±0.051**
STATLOG-HEART	(270, 14)	**−0.392±0.071**	**−0.395±0.080**	**−0.394±0.080**	**−0.395±0.080**
HABERMAN-SURVIVAL	(306, 4)	**−0.535±0.050**	**−0.537±0.054**	**−0.537±0.054**	**−0.537±0.054**
IONOSPHERE	(351, 34)	−0.221±0.021	−0.173±0.031	**−0.169±0.031**	**−0.169±0.033**
HORSE-COLIC	(368, 26)	**−0.465±0.058**	**−0.469±0.068**	**−0.466±0.068**	**−0.469±0.068**
CONGRESSIONAL-VOTING	(435, 17)	**−0.636±0.026**	**−0.636±0.029**	**−0.636±0.029**	**−0.636±0.029**
CYLINDER-BANDS	(512, 36)	−0.482±0.030	**−0.442±0.038**	−0.457±0.038	**−0.441±0.040**
BREAST-CANCER-WISC-DIAG	(569, 31)	**−0.079±0.020**	**−0.069±0.040**	**−0.071±0.038**	**−0.071±0.045**
ILPD-INDIAN-LIVER	(583, 10)	**−0.516±0.033**	**−0.516±0.037**	**−0.516±0.037**	**−0.516±0.037**
MONKS-2	(601, 7)	−0.496±0.037	**−0.445±0.042**	−0.468±0.045	**−0.443±0.044**
STATLOG-AUSTRALIAN-CREDIT	(690, 15)	**−0.627±0.022**	**−0.627±0.023**	**−0.627±0.023**	**−0.627±0.023**
CREDIT-APPROVAL	(690, 16)	**−0.342±0.008**	**−0.341±0.009**	**−0.341±0.008**	**−0.341±0.009**
BREAST-CANCER-WISC	(699, 10)	**−0.091±0.030**	**−0.091±0.037**	**−0.091±0.037**	**−0.091±0.037**
BLOOD	(748, 5)	**−0.475±0.043**	**−0.476±0.044**	**−0.476±0.044**	**−0.476±0.044**
PIMA	(768, 9)	**−0.470±0.031**	**−0.471±0.033**	**−0.471±0.033**	**−0.471±0.033**
MAMMOGRAPHIC	(961, 6)	**−0.408±0.037**	**−0.408±0.039**	**−0.408±0.039**	**−0.408±0.039**
STATLOG-GERMAN-CREDIT	(1000, 25)	**−0.493±0.040**	**−0.494±0.043**	**−0.493±0.043**	**−0.494±0.043**
Bold Count		21	23	22	27

Convergence of VI To ensure that the comparison to VI is as fair as possible, we ran both the vanilla VI approach where we jointly optimize $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ using L-BFGS-B and the CVI approach by Khan & Lin (2017) that uses natural gradients for the E-step for faster convergence. The speed of convergence was not the main interest, and thus we verified that the VI results obtained by the two different approaches gave the same results—thus confirming that the VI results presented in the tables are all for converged optimization runs.

Table 8. **Binary classification:** test set accuracy (higher is better) on different data sets from the *Bayesian benchmarks* over 5-fold cross-validation. We initialize EP with the trained hyperparameters values of our method. Best results and those not statistically significantly different from them under a paired t-test are **bolded**. The results are largely the same as Table 10, which means EP has converged.

	(n, d)	LA	EP (new init.)	VI	Ours
TRAINS	(10, 30)	$\mathbf{0.700\pm0.400}$	$\mathbf{0.800\pm0.245}$	$\mathbf{0.800\pm0.245}$	$\mathbf{0.800\pm0.245}$
BALLOONS	(16, 5)	$\mathbf{0.667\pm0.298}$	$\mathbf{0.667\pm0.298}$	$\mathbf{0.667\pm0.298}$	$\mathbf{0.667\pm0.298}$
FERTILITY	(100, 10)	$\mathbf{0.880\pm0.060}$	$\mathbf{0.880\pm0.060}$	$\mathbf{0.880\pm0.060}$	$\mathbf{0.880\pm0.060}$
PITTSBURG-BRIDGES-T-OR-D	(102, 8)	$\mathbf{0.863\pm0.036}$	$\mathbf{0.863\pm0.036}$	$\mathbf{0.863\pm0.036}$	$\mathbf{0.863\pm0.036}$
ACUTE-NEPHRITIS	(120, 7)	$\mathbf{1.000\pm0.000}$	$\mathbf{1.000\pm0.000}$	$\mathbf{1.000\pm0.000}$	$\mathbf{1.000\pm0.000}$
ACUTE-INFLAMMATION	(120, 7)	$\mathbf{1.000\pm0.000}$	$\mathbf{1.000\pm0.000}$	$\mathbf{1.000\pm0.000}$	$\mathbf{1.000\pm0.000}$
ECHOCARDIOGRAM	(131, 11)	$\mathbf{0.808\pm0.061}$	$\mathbf{0.778\pm0.063}$	$\mathbf{0.785\pm0.076}$	$\mathbf{0.778\pm0.063}$
HEPATITIS	(155, 20)	$\mathbf{0.813\pm0.024}$	$\mathbf{0.832\pm0.024}$	$\mathbf{0.813\pm0.024}$	$\mathbf{0.832\pm0.024}$
PARKINSONS	(195, 23)	$\mathbf{0.959\pm0.048}$	$\mathbf{0.959\pm0.048}$	$\mathbf{0.944\pm0.050}$	$\mathbf{0.959\pm0.048}$
BREAST-CANCER-WISC-PROG	(198, 34)	$\mathbf{0.793\pm0.051}$	$\mathbf{0.793\pm0.051}$	$\mathbf{0.793\pm0.051}$	$\mathbf{0.793\pm0.051}$
SPECT	(265, 23)	$\mathbf{0.706\pm0.031}$	$\mathbf{0.706\pm0.023}$	$\mathbf{0.702\pm0.028}$	$\mathbf{0.706\pm0.023}$
STATLOG-HEART	(270, 14)	$\mathbf{0.830\pm0.036}$	$\mathbf{0.826\pm0.034}$	$\mathbf{0.830\pm0.036}$	$\mathbf{0.826\pm0.034}$
HABERMAN-SURVIVAL	(306, 4)	$\mathbf{0.725\pm0.039}$	$\mathbf{0.725\pm0.039}$	$\mathbf{0.725\pm0.039}$	$\mathbf{0.725\pm0.039}$
IONOSPHERE	(351, 34)	$\mathbf{0.932\pm0.025}$	$\mathbf{0.946\pm0.028}$	$\mathbf{0.949\pm0.026}$	$\mathbf{0.946\pm0.028}$
HORSE-COLIC	(368, 26)	$\mathbf{0.799\pm0.035}$	$\mathbf{0.802\pm0.039}$	0.791 ± 0.045	$\mathbf{0.802\pm0.039}$
CONGRESSIONAL-VOTING	(435, 17)	$\mathbf{0.605\pm0.024}$	$\mathbf{0.591\pm0.016}$	$\mathbf{0.595\pm0.013}$	$\mathbf{0.591\pm0.016}$
CYLINDER-BANDS	(512, 36)	0.785 ± 0.037	$\mathbf{0.803\pm0.030}$	$\mathbf{0.791\pm0.034}$	$\mathbf{0.803\pm0.031}$
BREAST-CANCER-WISC-DIAG	(569, 31)	0.972 ± 0.010	$\mathbf{0.975\pm0.009}$	$\mathbf{0.979\pm0.009}$	$\mathbf{0.977\pm0.011}$
ILPD-INDIAN-LIVER	(583, 10)	$\mathbf{0.719\pm0.025}$	$\mathbf{0.719\pm0.028}$	$\mathbf{0.715\pm0.027}$	$\mathbf{0.719\pm0.028}$
MONKS-2	(601, 7)	$\mathbf{0.740\pm0.049}$	$\mathbf{0.762\pm0.045}$	$\mathbf{0.744\pm0.046}$	$\mathbf{0.762\pm0.045}$
STATLOG-AUSTRALIAN-CREDIT	(690, 15)	$\mathbf{0.678\pm0.028}$	$\mathbf{0.678\pm0.028}$	$\mathbf{0.678\pm0.028}$	$\mathbf{0.678\pm0.028}$
CREDIT-APPROVAL	(690, 16)	$\mathbf{0.859\pm0.026}$	$\mathbf{0.862\pm0.026}$	$\mathbf{0.864\pm0.024}$	$\mathbf{0.862\pm0.026}$
BREAST-CANCER-WISC	(699, 10)	$\mathbf{0.969\pm0.016}$	$\mathbf{0.969\pm0.016}$	$\mathbf{0.969\pm0.016}$	$\mathbf{0.969\pm0.016}$
BLOOD	(748, 5)	$\mathbf{0.785\pm0.042}$	$\mathbf{0.786\pm0.044}$	$\mathbf{0.786\pm0.044}$	$\mathbf{0.786\pm0.044}$
PIMA	(768, 9)	$\mathbf{0.764\pm0.027}$	$\mathbf{0.767\pm0.026}$	$\mathbf{0.767\pm0.026}$	$\mathbf{0.767\pm0.026}$
MAMMOGRAPHIC	(961, 6)	$\mathbf{0.824\pm0.019}$	$\mathbf{0.823\pm0.019}$	$\mathbf{0.823\pm0.019}$	$\mathbf{0.823\pm0.019}$
STATLOG-GERMAN-CREDIT	(1000, 25)	$\mathbf{0.770\pm0.039}$	$\mathbf{0.769\pm0.037}$	$\mathbf{0.768\pm0.038}$	$\mathbf{0.769\pm0.037}$
Bold Count		25	27	26	27

Table 9. **Binary classification:** log predictive density (higher is better) on different data sets from the *Bayesian benchmarks* (mean ± standard deviation over 5-fold cross-validation). The best results and those not statistically significantly different from them under a paired t-test are **bolded**. For the classification accuracy all methods perform comparably, which is to be expected as accuracy alone only captures where the decision boundary has been draw, completely disregarding second-order information.

#		(n, d)	LA	EP	VI	Ours
1	TRAINS	(10, 30)	$\mathbf{-0.695\pm0.011}$	$\mathbf{-0.687\pm0.023}$	$\mathbf{-0.692\pm0.023}$	$\mathbf{-0.681\pm0.042}$
2	BALLOONS	(16, 5)	$\mathbf{-0.707\pm0.146}$	$\mathbf{-0.684\pm0.161}$	$\mathbf{-0.711\pm0.239}$	$\mathbf{-0.657\pm0.267}$
3	FERTILITY	(100, 10)	$\mathbf{-0.379\pm0.099}$	$\mathbf{-0.384\pm0.138}$	$\mathbf{-0.378\pm0.103}$	$\mathbf{-0.379\pm0.103}$
4	PITTSBURG-BRIDGES-T-OR-D	(102, 8)	$\mathbf{-0.306\pm0.044}$	$\mathbf{-0.316\pm0.060}$	$\mathbf{-0.295\pm0.057}$	$\mathbf{-0.296\pm0.059}$
5	ACUTE-NEPHRITIS	(120, 7)	-0.203 ± 0.010	-0.047 ± 0.009	-0.006 ± 0.002	$\mathbf{-0.005\pm0.001}$
6	ACUTE-INFLAMMATION	(120, 7)	-0.172 ± 0.033	-0.055 ± 0.005	-0.008 ± 0.001	$\mathbf{-0.007\pm0.001}$
7	ECHOCARDIOGRAM	(131, 11)	$\mathbf{-0.420\pm0.073}$	$\mathbf{-0.412\pm0.084}$	$\mathbf{-0.420\pm0.085}$	$\mathbf{-0.421\pm0.086}$
8	HEPATITIS	(155, 20)	$\mathbf{-0.369\pm0.043}$	$\mathbf{-0.376\pm0.046}$	$\mathbf{-0.361\pm0.051}$	$\mathbf{-0.361\pm0.052}$
9	PARKINSONS	(195, 23)	-0.267 ± 0.030	-0.302 ± 0.090	-0.164 ± 0.062	$\mathbf{-0.145\pm0.054}$
10	BREAST-CANCER-WISC-PROG	(198, 34)	$\mathbf{-0.460\pm0.063}$	$\mathbf{-0.478\pm0.083}$	$\mathbf{-0.458\pm0.071}$	$\mathbf{-0.458\pm0.070}$
11	SPECT	(265, 23)	$\mathbf{-0.587\pm0.045}$	$\mathbf{-0.587\pm0.052}$	$\mathbf{-0.589\pm0.050}$	$\mathbf{-0.590\pm0.051}$
12	STATLOG-HEART	(270, 14)	$\mathbf{-0.392\pm0.071}$	$\mathbf{-0.380\pm0.058}$	$\mathbf{-0.394\pm0.080}$	$\mathbf{-0.395\pm0.080}$
13	HABERMAN-SURVIVAL	(306, 4)	$\mathbf{-0.535\pm0.050}$	$\mathbf{-0.539\pm0.059}$	$\mathbf{-0.537\pm0.054}$	$\mathbf{-0.537\pm0.054}$
14	IONOSPHERE	(351, 34)	-0.221 ± 0.021	-0.227 ± 0.016	$\mathbf{-0.169\pm0.031}$	$\mathbf{-0.169\pm0.033}$
15	HORSE-COLIC	(368, 26)	-0.465 ± 0.058	$\mathbf{-0.455\pm0.060}$	$\mathbf{-0.466\pm0.068}$	$\mathbf{-0.469\pm0.068}$
16	CONGRESSIONAL-VOTING	(435, 17)	$\mathbf{-0.636\pm0.026}$	$\mathbf{-0.633\pm0.027}$	$\mathbf{-0.636\pm0.029}$	$\mathbf{-0.636\pm0.029}$
17	CYLINDER-BANDS	(512, 36)	-0.482 ± 0.030	-0.495 ± 0.030	-0.457 ± 0.038	$\mathbf{-0.441\pm0.040}$
18	BREAST-CANCER-WISC-DIAG	(569, 31)	$\mathbf{-0.079\pm0.020}$	-0.140 ± 0.015	$\mathbf{-0.071\pm0.038}$	$\mathbf{-0.071\pm0.045}$
19	ILPD-INDIAN-LIVER	(583, 10)	$\mathbf{-0.516\pm0.033}$	$\mathbf{-0.521\pm0.033}$	$\mathbf{-0.516\pm0.037}$	$\mathbf{-0.516\pm0.037}$
20	MONKS-2	(601, 7)	-0.496 ± 0.037	-0.518 ± 0.042	-0.468 ± 0.045	$\mathbf{-0.443\pm0.044}$
21	STATLOG-AUSTRALIAN-CREDIT	(690, 15)	$\mathbf{-0.627\pm0.022}$	$\mathbf{-0.636\pm0.030}$	$\mathbf{-0.627\pm0.023}$	$\mathbf{-0.627\pm0.023}$
22	CREDIT-APPROVAL	(690, 16)	$\mathbf{-0.342\pm0.008}$	$\mathbf{-0.343\pm0.012}$	$\mathbf{-0.341\pm0.008}$	$\mathbf{-0.341\pm0.009}$
23	BREAST-CANCER-WISC	(699, 10)	$\mathbf{-0.091\pm0.030}$	$\mathbf{-0.092\pm0.028}$	$\mathbf{-0.091\pm0.037}$	$\mathbf{-0.091\pm0.037}$
24	BLOOD	(748, 5)	$\mathbf{-0.475\pm0.043}$	$\mathbf{-0.476\pm0.045}$	$\mathbf{-0.476\pm0.044}$	$\mathbf{-0.476\pm0.044}$
25	PIMA	(768, 9)	$\mathbf{-0.470\pm0.031}$	$\mathbf{-0.472\pm0.035}$	$\mathbf{-0.471\pm0.033}$	$\mathbf{-0.471\pm0.033}$
26	MAMMOGRAPHIC	(961, 6)	$\mathbf{-0.408\pm0.037}$	$\mathbf{-0.408\pm0.041}$	$\mathbf{-0.408\pm0.039}$	$\mathbf{-0.408\pm0.039}$
27	STATLOG-GERMAN-CREDIT	(1000, 25)	$\mathbf{-0.493\pm0.040}$	$\mathbf{-0.493\pm0.041}$	$\mathbf{-0.493\pm0.043}$	$\mathbf{-0.494\pm0.043}$
	Bold Count		20	20	22	27

Table 10. **Binary classification:** test set accuracy (higher is better) on different data sets from the *Bayesian benchmarks* (mean ± standard deviation over 5-fold cross-validation). The best results and those not statistically significantly different from them under a paired t-test are **bolded**. For the classification accuracy all methods perform comparably, which is to be expected as accuracy alone only captures where the decision boundary has been draw, completely disregarding second-order information.

	(n, d)	LA	EP	VI	Ours
TRAINS	(10, 30)	$\mathbf{0.700{\pm}0.400}$	$\mathbf{0.800{\pm}0.245}$	$\mathbf{0.800{\pm}0.245}$	$\mathbf{0.800{\pm}0.245}$
BALLOONS	(16, 5)	$\mathbf{0.667{\pm}0.298}$	$\mathbf{0.667{\pm}0.298}$	$\mathbf{0.667{\pm}0.298}$	$\mathbf{0.667{\pm}0.298}$
FERTILITY	(100, 10)	$\mathbf{0.880{\pm}0.060}$	$\mathbf{0.880{\pm}0.060}$	$\mathbf{0.880{\pm}0.060}$	$\mathbf{0.880{\pm}0.060}$
PITTSBURG-BRIDGES-T-OR-D	(102, 8)	$\mathbf{0.863{\pm}0.036}$	$\mathbf{0.883{\pm}0.066}$	$\mathbf{0.863{\pm}0.036}$	$\mathbf{0.863{\pm}0.036}$
ACUTE-NEPHRITIS	(120, 7)	$\mathbf{1.000{\pm}0.000}$	$\mathbf{1.000{\pm}0.000}$	$\mathbf{1.000{\pm}0.000}$	$\mathbf{1.000{\pm}0.000}$
ACUTE-INFLAMMATION	(120, 7)	$\mathbf{1.000{\pm}0.000}$	$\mathbf{1.000{\pm}0.000}$	$\mathbf{1.000{\pm}0.000}$	$\mathbf{1.000{\pm}0.000}$
ECHOCARDIOGRAM	(131, 11)	$0.808{\pm}0.061$	$\mathbf{0.847{\pm}0.043}$	$\mathbf{0.785{\pm}0.076}$	$0.778{\pm}0.063$
HEPATITIS	(155, 20)	$\mathbf{0.813{\pm}0.024}$	$\mathbf{0.813{\pm}0.032}$	$\mathbf{0.813{\pm}0.024}$	$\mathbf{0.832{\pm}0.024}$
PARKINSONS	(195, 23)	$\mathbf{0.959{\pm}0.048}$	$0.892{\pm}0.075$	$\mathbf{0.944{\pm}0.050}$	$\mathbf{0.959{\pm}0.048}$
BREAST-CANCER-WISC-PROG	(198, 34)	$\mathbf{0.793{\pm}0.051}$	$\mathbf{0.793{\pm}0.069}$	$\mathbf{0.793{\pm}0.051}$	$\mathbf{0.793{\pm}0.051}$
SPECT	(265, 23)	$\mathbf{0.706{\pm}0.031}$	$\mathbf{0.698{\pm}0.027}$	$\mathbf{0.702{\pm}0.028}$	$\mathbf{0.706{\pm}0.023}$
STATLOG-HEART	(270, 14)	$\mathbf{0.830{\pm}0.036}$	$\mathbf{0.833{\pm}0.039}$	$\mathbf{0.830{\pm}0.036}$	$\mathbf{0.826{\pm}0.034}$
HABERMAN-SURVIVAL	(306, 4)	$\mathbf{0.725{\pm}0.039}$	$\mathbf{0.719{\pm}0.036}$	$\mathbf{0.725{\pm}0.039}$	$\mathbf{0.725{\pm}0.039}$
IONOSPHERE	(351, 34)	$\mathbf{0.932{\pm}0.025}$	$\mathbf{0.932{\pm}0.024}$	$\mathbf{0.946{\pm}0.024}$	$\mathbf{0.946{\pm}0.028}$
HORSE-COLIC	(368, 26)	$\mathbf{0.799{\pm}0.035}$	$\mathbf{0.807{\pm}0.032}$	$\mathbf{0.791{\pm}0.045}$	$\mathbf{0.802{\pm}0.039}$
CONGRESSIONAL-VOTING	(435, 17)	$\mathbf{0.605{\pm}0.024}$	$\mathbf{0.593{\pm}0.014}$	$\mathbf{0.595{\pm}0.013}$	$\mathbf{0.591{\pm}0.016}$
CYLINDER-BANDS	(512, 36)	$0.785{\pm}0.037$	$\mathbf{0.777{\pm}0.044}$	$\mathbf{0.791{\pm}0.034}$	$\mathbf{0.803{\pm}0.031}$
BREAST-CANCER-WISC-DIAG	(569, 31)	$\mathbf{0.972{\pm}0.010}$	$\mathbf{0.974{\pm}0.019}$	$\mathbf{0.977{\pm}0.012}$	$\mathbf{0.977{\pm}0.011}$
ILPD-INDIAN-LIVER	(583, 10)	$\mathbf{0.719{\pm}0.025}$	$\mathbf{0.712{\pm}0.020}$	$\mathbf{0.717{\pm}0.029}$	$\mathbf{0.719{\pm}0.028}$
MONKS-2	(601, 7)	$\mathbf{0.740{\pm}0.049}$	$0.719{\pm}0.039$	$\mathbf{0.744{\pm}0.046}$	$\mathbf{0.762{\pm}0.045}$
STATLOG-AUSTRALIAN-CREDIT	(690, 15)	$\mathbf{0.678{\pm}0.028}$	$\mathbf{0.677{\pm}0.021}$	$\mathbf{0.678{\pm}0.028}$	$\mathbf{0.678{\pm}0.028}$
CREDIT-APPROVAL	(690, 16)	$\mathbf{0.859{\pm}0.026}$	$\mathbf{0.864{\pm}0.020}$	$\mathbf{0.864{\pm}0.024}$	$\mathbf{0.862{\pm}0.026}$
BREAST-CANCER-WISC	(699, 10)	$\mathbf{0.969{\pm}0.016}$	$\mathbf{0.969{\pm}0.016}$	$\mathbf{0.969{\pm}0.016}$	$\mathbf{0.969{\pm}0.016}$
BLOOD	(748, 5)	$\mathbf{0.785{\pm}0.042}$	$\mathbf{0.785{\pm}0.041}$	$\mathbf{0.786{\pm}0.044}$	$\mathbf{0.786{\pm}0.044}$
PIMA	(768, 9)	$\mathbf{0.764{\pm}0.027}$	$\mathbf{0.766{\pm}0.025}$	$\mathbf{0.767{\pm}0.026}$	$\mathbf{0.767{\pm}0.026}$
MAMMOGRAPHIC	(961, 6)	$\mathbf{0.824{\pm}0.019}$	$\mathbf{0.823{\pm}0.017}$	$\mathbf{0.823{\pm}0.019}$	$\mathbf{0.823{\pm}0.019}$
STATLOG-GERMAN-CREDIT	(1000, 25)	$\mathbf{0.770{\pm}0.039}$	$\mathbf{0.763{\pm}0.038}$	$\mathbf{0.768{\pm}0.038}$	$\mathbf{0.769{\pm}0.037}$
Bold Count		25	25	27	26

E. Robust (Student-t) Regression

For additional insight, we include results for a Student-t likelihood model that allows heavy-tailed noise in the observations. This likelihood is not log-concave, which makes LA and EP more challenging. Previous works in the field do not list many standard data sets for benchmarking, but we have gathered three previously used benchmark problems for comparison.

E.1. Experiment Setup

We preprocessed the data and selected an interval of 1000 data points from the original data. We initialize all methods with unit lengthscale, kernel magnitude, and likelihood variance. We fix the degrees of freedom in the likelihood to $\nu = 3$. For LA and EP, we follow the methods designed in Jylänki et al. (2011). For VI and our hybrid training procedure, E and M steps each have 20 iterations. In the E-step we set the learning rate of natural gradient descent to be 0.001. In the M-step we use the Adam optimizer (Kingma & Ba, 2015) and set the learning rate to 0.001. For NEAL we set the maximal steps to be 2000 and for BOSTON and STOCK we set the maximal steps to be 5000.

F. Evaluation on Sparse Approximation

To further study the proposed approach, we include results for a couple of sparse GP classification problems as proof-of-concept.

Table 11. Data sets information for sparse binary classification.

Data set	TITANIC	BANK	TWONORM	MUSHROOM	MAGIC
(n, d)	(2201, 4)	(4521, 17)	(7400, 21)	(8124, 22)	(19020, 11)

F.1. Experiment Setup

We initialize all methods with unit lengthscale and magnitude. We use k-means to select a set of 500 inducing points and then fix the location of inducing points. Learning the inducing points would add an additional layer of optimization, and this is not in the scope of this paper. We use 20 iterations for both the E and M steps. In the E-step we set the learning rate of natural gradient descent to 0.1. In the M-step we use the Adam optimizer (Kingma & Ba, 2015) and set the learning rate to 0.01. The data sets information are given in Table 11.