Deep Gaussian Processes for Regression using Approximate Expectation Propagation

Thang D. Bui
University of Cambridge
tdb40@cam.ac.uk

Daniel Hernández-Lobato
Universidad Autónoma de Madrid
daniel.hernandez@uam.es

Yingzhen Li
University of Cambridge
yl494@cam.ac.uk

José Miguel Hernández-Lobato
Harvard University
jmhl@seas.harvard.edu

Richard E. Turner
University of Cambridge
ret26@cam.ac.uk

February 15, 2016

Abstract

Deep Gaussian processes (DGPs) are multi-layer hierarchical generalisations of Gaussian processes (GPs) and are formally equivalent to neural networks with multiple, infinitely wide hidden layers. DGPs are nonparametric probabilistic models and as such are arguably more flexible, have a greater capacity to generalise, and provide better calibrated uncertainty estimates than alternative deep models. This paper develops a new approximate Bayesian learning scheme that enables DGPs to be applied to a range of medium to large scale regression problems for the first time. The new method uses an approximate Expectation Propagation procedure and a novel and efficient extension of the probabilistic backpropagation algorithm for learning. We evaluate the new method for non-linear regression on eleven real-world datasets, showing that it always outperforms GP regression and is almost always better than state-of-the-art deterministic and sampling-based approximate inference methods for Bayesian neural networks. As a by-product, this work provides a comprehensive analysis of six approximate Bayesian methods for training neural networks.

1 Introduction

Gaussian Processes (GPs) are powerful nonparametric distributions over continuous functions that can be used for both supervised and unsupervised learning problems (Rasmussen & Williams, 2005). In this article, we study a multi-layer hierarchical generalisation of GPs or deep Gaussian Processes (DGPs) (Damianou & Lawrence, 2013) for supervised learning tasks. A GP is equivalent to an infinitely wide neural network with single hidden layer and similarly a DGP is a multi-layer neural network with multiple infinitely wide hidden layers (Neal, 1995). The mapping between layers in this type of network is parameterised by a GP, and, as a result, DGPs retain useful properties of GPs such as nonparametric modelling power and well-calibrated predictive uncertainty estimates. In addition, DGPs employ a hierarchical structure of GP mappings and therefore are arguably more flexible, have

a greater capacity to generalise, and are able to provide better predictive performance (Damianou, 2015). This family of models is attractive as it can also potentially discover layers of increasingly abstract data representations, in much the same way as their deep parametric counterparts, but it can also handle and propagate uncertainty in the hierarchy.

The addition of non-linear hidden layers can also potentially overcome practical limitations of *shallow* GPs. First, modelling real-world complex datasets often requires rich, hand-designed covariance functions. DGPs can perform input warping or dimensionality compression or expansion, and automatically learn to construct a kernel that works well for the data at hand. As a result, learning in this model provides a flexible form of Bayesian kernel design. Second, the functional mapping from inputs to outputs specified by a DGP is non-Gaussian which is a more general and flexible modelling choice. Third, DGPs can repair damage done by sparse approximations to the representational power of each GP layer. For example, pseudo datapoint based approximation methods for DGPs trade model complexity for a lower computational complexity of $\mathcal{O}(NLM^2)$ where N is the number of datapoints, L is the number of layers, and M is the number of pseudo datapoints. This complexity scales quadratically in M whereas the dependence on the number of layers L is only linear. Therefore, it can be cheaper to increase the representation power of the model by adding extra layers rather than by adding more pseudo datapoints.

The focus of this paper is Bayesian learning of DGPs, which involves inferring the posterior over the layer mappings and hyperparameter optimisation via the marginal likelihood. Unfortunately, exact Bayesian learning in this model is analytically intractable and as such approximate inference is needed. Current proposals in the literature do not scale well and have not been compared to alternative deep Bayesian models. We will first review the model and past work in Section 2, and then make the following contributions:

- We propose a new approximate inference scheme for DGPs for regression, using a sparse GP approximation, a novel approximate Expectation Propagation scheme and the probabilistic backpropagation algorithm, resulting in a computationally efficient, scalable and easy to implement algorithm (Sections 3, 4 and 5).
- We demonstrate the validity of our method in supervised learning tasks on various medium to large scale datasets and show that the proposed method is always better than GP regression and is almost always better than state-of-the-art approximate inference techniques for multi-layer neural networks (Section 8).

2 Deep Gaussian processes

We first review DGPs and existing literature on approximate inference and learning for DGPs. Suppose we have a training set comprising of N D-dimensional input and observation pairs $(\mathbf{x}_n, y_n)$. For ease of presentation, the outputs are assumed to be real-valued scalars, but other types of data can be easily accommodated[1]. The probabilistic representation of a DGP comprising of L layers can be written as follows,

$$p(f_l|\theta_l) = \mathcal{GP}(f_l; \mathbf{0}, \mathbf{K}_l), \quad l = 1, \cdots, L$$
$$p(\mathbf{h}_l|f_l, \mathbf{h}_{l-1}, \sigma_l^2) = \prod_n \mathcal{N}(h_{l,n}; f_l(h_{l-1,n}), \sigma_l^2), \quad h_{1,n} = \mathbf{x}_n$$

[1]We also discuss how to handle non-Gaussian likelihoods in the supplementary material.

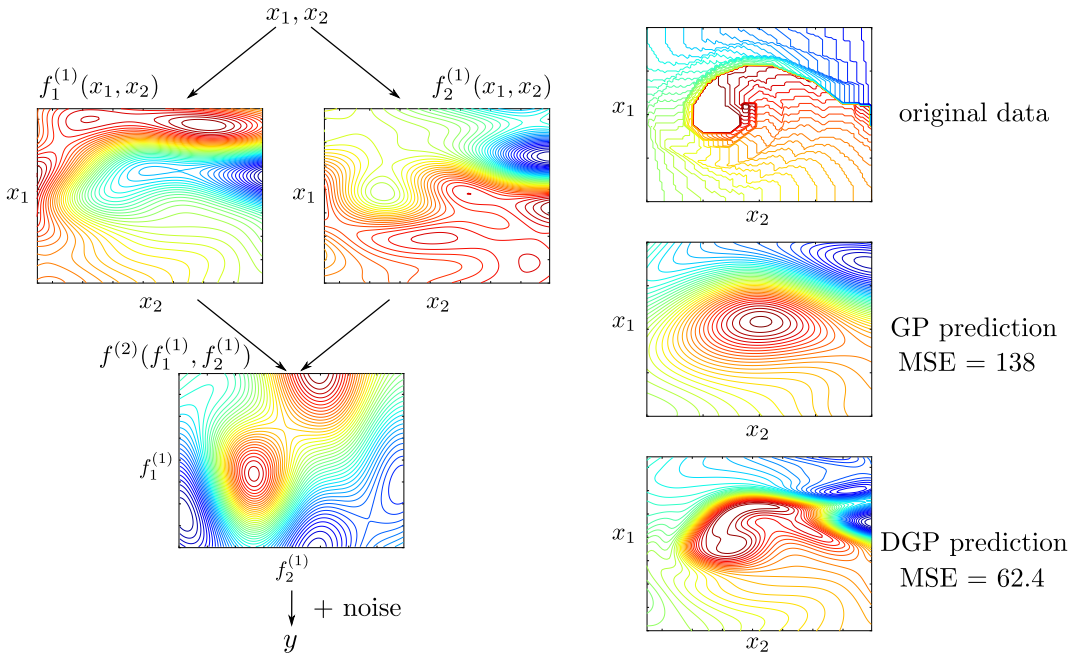

Figure 1: A deep GP example that has two GP layers and one 2-D hidden layer. The training output is the state values of the mountain car problem. The left graphs show latent functions in each layer, two functions in the first layer and one in the second layer, learnt by using the proposed approach. The right graph shows the training data [top] and the predictions of the overall function mapping from inputs to outputs made by a GP [middle] and the DGP on the left [bottom].

$$p(\mathbf{y}|f_L, \mathbf{h}_{L-1}, \sigma_L^2) = \prod_n \mathcal{N}(y_n; f_L(h_{L-1,n}), \sigma_L^2)$$

where hidden layers[2] are denoted $h_{l,n}$ and the functions in each layer, f_l. More formally, we place a zero mean GP prior over the mapping f_l, that is, given the inputs to f_l any finite set of function values are distributed under the prior according to a multivariate Gaussian $p(\mathbf{f}_l) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K_{ff}})$. Note that these function values and consequently the hidden variables are not marginally normally distributed, as the inputs are random variables. When $L = 1$, the model described above collapses back to GP regression or classification. When the inputs $\{\mathbf{x}_n\}$ are unknown and random, the model becomes a DGP latent variable model, which has been studied in Damianou & Lawrence (2013).

An example of DGPs when $L = 2$ and $\dim(h_1) = 2$ is shown in Figure 1. We use this network with the proposed approximation and training algorithm to fit a value function of the mountain car problem (Sutton & Barto, 1998) from a small number of noisy evaluations. This function is particularly difficult for models such as GP regression with a standard exponentiated quadratic kernel due to a *steep value function cliff*, but is well handled by a DGP with only two GP layers. Interestingly the functions in the first layer are fairly simple and learn to cover or explain different parts of the input space.

[2]Hidden variables in the intermediate layers can and will generally have multiple dimensions but we have omitted this here to lighten the notation.

We are interested in inferring the posterior distribution over the latent function mappings and the intermediate hidden variables, as well as obtaining a marginal likelihood estimate for hyperparameter tuning and model comparison. Due to the nonlinearity in the hierarchy, these quantities are analytically intractable. As such, approximate inference is required. The simplest approach is to obtain the *maximum a posteriori* estimate of the hidden variables (Lawrence & Moore, 2007). However, this procedure is prone to over-fitting and does not provide uncertainty estimates. An alternative existing approach is based on a variational-free-energy method proposed by Damianou & Lawrence (2013), extending the seminal work on variational sparse GPs by Titsias (2009). In this scheme, a variational approximation over both latent functions and hidden variables is chosen such that a free energy is both computationally and analytically tractable. Critically, as a variational distribution over the hidden variables is used in this approach, in addition to one over the inducing outputs, the number of variational parameters increases linearly with the number of training datapoints which hinders the use of this method for large scale datasets. Furthermore, initialisation for this scheme is a known issue, even for a modest number of datapoints (Turner & Sahani, 2011). An extension of Damianou & Lawrence (2013) that has skip links from the inputs to every hidden layer in the network was proposed in Dai et al. (2015), based on suggestions provided in Duvenaud et al. (2014). Recent work by Hensman & Lawrence (2014) introduces a nested variational scheme that only requires a variational distribution over the inducing outputs, removing the parameter scaling problem of Damianou & Lawrence (2013). However, both approaches of Dai et al. (2015) and Hensman & Lawrence (2014) have not been fully evaluated on benchmark supervised learning tasks or on medium to large scale datasets, nor compared to alternative deep models.

A special case of DGPs when $L = 2$ and the sole hidden layer h_1 is only one dimensional is warped GPs (Snelson et al., 2004; Lázaro-Gredilla, 2012). In Lázaro-Gredilla (2012) a variational approach, in a similar spirit to Titsias (2009) and Damianou & Lawrence (2013), was used to jointly learn the latent functions. In contrast, the latent function in the second layer is assumed to be deterministic and parameterised by a small set of parameters in Snelson et al. (2004), which can be learnt by maximising the analytically tractable marginal likelihood. However, the performance of warped GPs is often similar to a standard GP, most likely due to the narrow bottleneck in the hidden layer.

Our work differs substantially from the above and introduces an alternative approximate inference scheme for DGPs based on three approximations. First, in order to sidestep the cubic computational cost of GPs we leverage a well-known pseudo point sparse approximation (Snelson & Ghahramani, 2006; Quiñonero-Candela & Rasmussen, 2005). Second, an approximation to the Expectation Propagation (EP) energy function (Seeger, 2007), a marginal likelihood estimate, is optimised directly to find an approximate posterior over the inducing outputs. Third, the optimisation demands analytically intractable moments that are approximated by nesting Assumed Density Filtering (Hernández-Lobato & Adams, 2015). The proposed algorithm is not restricted to the warped GP case and is applicable to non-Gaussian observation models.

The complexity of our method is similar to that of the variational approach proposed in Damianou & Lawrence (2013), $\mathcal{O}(NLM^2)$, but is much less memory intensive, $\mathcal{O}(LM^2)$ vs. $\mathcal{O}(NL + LM^2)$. These costs are competitive to those of the nested variational approach in Hensman & Lawrence (2014).

3 The Fully Independent Training Conditional approximation

The computational complexity of full GP models scales cubically with the number of training instances, making it intractable in practice. Sparse approximation techniques are therefore often necessary. They can be coarsely put into two classes: ones that explicitly sparsify and create a semi-parametric representation that approximates the original model, and ones that retain the original non-parametric properties and perform sparse approximation to the exact posterior. The method used here, Fully Independent Training Conditional (FITC), falls into the first category. The FITC approximation is formed by considering a small set of M function values $\mathbf{u}$ in the infinite dimensional vector f and assuming conditional independence between the remaining values given the set $\mathbf{u}$ (Snelson & Ghahramani, 2006; Quiñonero-Candela & Rasmussen, 2005). This set is often called inducing outputs or pseudo targets and their input locations $\mathbf{z}$ can be chosen by optimising the approximate marginal likelihood. The resulting model can be written as follows,

$$
\begin{aligned}
p(\mathbf{u}_l|\theta_l) &= \mathcal{N}(\mathbf{u}_l; \mathbf{0}, \mathbf{K}_{\mathbf{u}_{l-1},\mathbf{u}_{l-1}}), \quad l = 1, \cdots, L \\
p(\mathbf{h}_l|\mathbf{u}_l, \mathbf{h}_{l-1}, \sigma_l^2) &= \prod_n \mathcal{N}(h_{l,n}; \mathbf{C}_{n,l}\mathbf{u}_l, \mathbf{R}_{n,l}), \\
p(\mathbf{y}|\mathbf{u}_L, \mathbf{H}_{L-1}, \sigma_L^2) &= \prod_n \mathcal{N}(y_n; \mathbf{C}_{n,L}\mathbf{u}_L, \mathbf{R}_{n,L}).
\end{aligned}
$$

where $\mathbf{C}_{n,l} = \mathbf{K}_{\mathbf{h}_{l,n},\mathbf{u}_l}\mathbf{K}^{-1}_{\mathbf{u}_l,\mathbf{u}_l}$ and $\mathbf{R}_{n,l} = \mathbf{K}_{\mathbf{h}_{l,n},\mathbf{h}_{l,n}} - \mathbf{K}_{\mathbf{h}_{l,n},\mathbf{u}_l}\mathbf{K}^{-1}_{\mathbf{u}_l,\mathbf{u}_l}\mathbf{K}_{\mathbf{u}_l,\mathbf{h}_{l,n}} + \sigma_l^2\mathrm{I}$. Note that the function outputs index the covariance matrices, for example $\mathbf{K}_{\mathbf{h}_{l,n},\mathbf{u}_l}$ denotes the covariance between $\mathbf{h}_{l,n}$ and $\mathbf{u}_l$, and takes $\mathbf{h}_{l-1,n}$ and $\mathbf{z}_l$ as inputs respectively. This is important when propagating uncertainty through the network. The FITC approximation creates a semi-parametric model, but one which is cleverly structured so that the induced non-stationary noise captures the uncertainty introduced from the sparsification. The computational complexity of inference and hyperparameter tuning in this approximate model is $\mathcal{O}(NM^2)$ which means M needs to be smaller than N to provide any computational gain (i.e. the approximation should be sparse). The quality of the approximation largely depends on the number of inducing outputs M and the complexity of the underlying function, i.e. if the function's characteristic lengthscale is small, M needs to be large and vice versa. As M tends to N and $\mathbf{z} = \mathbf{X}$, i.e. the inducing inputs and training inputs are shared, the approximate model reverts back to the original GP model. The graphical model is shown in Figure 2 [left].

4 Approximate Bayesian inference via EP

Having specified a probabilistic model for data using a deep sparse Gaussian processes we now consider inference for the inducing outputs $\mathbf{u} = \{\mathbf{u}_l\}_{l=1}^L$ and learning of the model parameters $\alpha = \{\mathbf{z}_l, \theta_l\}_{l=1}^L$. The posterior distribution over the inducing outputs can be written as $p(\mathbf{u}|\mathbf{X}, \mathbf{y}) \propto p(\mathbf{u}) \prod_n p(y_n|\mathbf{u}, \mathbf{X}_n)$. This quantity can then be used for output prediction given a test input, $p(y^*|\mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int \mathrm{d}\mathbf{u}\, p(\mathbf{u}|\mathbf{X}, \mathbf{y})\, p(y^*|\mathbf{u}, \mathbf{x}^*)$. The model hyperparameters can be tuned by maximising the marginal likelihood $p(\mathbf{y}|\alpha) = \int \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{h}\, p(\mathbf{u}, \mathbf{h})\, p(\mathbf{y}|\mathbf{u}, \mathbf{h}, \alpha)$. However, both the posterior of $\mathbf{u}$ and the marginal likelihood are not analytically tractable when there is more than one GP layer in the model. As such, approximate inference is needed; here we make use of the EP energy function with a tied factor constraint similar to that proposed in the Stochastic Expectation Propagation (SEP) algorithm (Li et al., 2015) to produce a scalable, convergent approximate inference method.

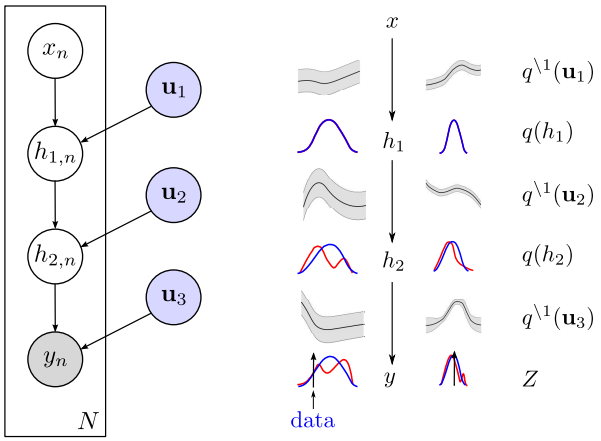

Figure 2: Left: The graphical model of our DGP-FITC model where the inducing outputs $\{\mathbf{u}_l\}$ play a role of global parameters. Right: A Gaussian moment-matching procedure to compute $\log \mathcal{Z}$. The bottom arrows denote the value of the observation and the left and right graphs [before and after an update respectively] show how the algorithm makes the final propagated Gaussian fit to the data, i.e. the model is trained so that training points are more probable after each update. The red curves show the distribution over hidden variables before being approximated by a Gaussian in blue. Best viewed in colour.

4.1 EP, Stochastic EP and the EP approximate energy

In EP (Minka, 2001), the approximate posterior is assumed to be $q(\mathbf{u}) \propto p(\mathbf{u}) \prod_n \tilde{t}_n(\mathbf{u})$ where $\{\tilde{t}_n(\mathbf{u})\}_{n=1}^N$ are the approximate data factors. Each factor approximately captures the contribution of datapoint n makes to the posterior and, in this work, they take an unnormalised Gaussian form. The factors can be found by running an iterative procedure which often requires several passes through the training set for convergence[3]. The EP algorithm also provides an approximation to the marginal likelihood,

$$\log p(\mathbf{y}|\alpha) \approx \mathcal{F}(\alpha) = \phi(\theta) - \phi(\theta_{\text{prior}}) + \sum_{n=1}^{N} \log \tilde{\mathcal{Z}}_n$$
$$\text{where} \quad \log \tilde{\mathcal{Z}}_n = \log \mathcal{Z}_n + \phi(\theta^{\backslash n}) - \phi(\theta),$$

where $\theta, \theta^{\backslash n}$ and θ_{prior} are the natural parameters of $q(\mathbf{u})$, the cavity $q^{\backslash n}(\mathbf{u})$ [$q^{\backslash n}(\mathbf{u}) \propto q(\mathbf{u})/\tilde{t}_n(\mathbf{u})$] and $p(\mathbf{u})$ respectively, $\phi(\theta)$ is the log normaliser of a Gaussian distribution with natural parameters θ, and $\log \mathcal{Z}_n = \log \int \mathrm{d}\mathbf{u}\, q^{\backslash n}(\mathbf{u})\, p(y_n|\mathbf{u}, \mathbf{X}_n)$ (Seeger, 2007). Unfortunately, EP is not guaranteed to converge, but if it does, the fixed points lie at the stationary points of the EP energy, which is given by $-\mathcal{F}(\alpha)$. Furthermore, EP requires the approximate factors to be stored in memory, which has a cost of $\mathcal{O}(NLM^2)$ in this application as we need to store the mean and the covariance matrix for each factor.

[3]We summarise the EP steps in the supplementary material.

4.2 Direct EP energy minimisation with a tied factor constraint

In order to reduce the expensive memory footprint of EP, the data-factors are tied. That is the posterior $p(\mathbf{u}|\mathbf{X}, \mathbf{y})$ is approximated by $q(\mathbf{u}) \propto p(\mathbf{u})g(\mathbf{u})^N$, where the factor $g(\mathbf{u})$ could be thought of as an *average* data factor that captures the average effect of a likelihood term on the posterior. Approximations of this form were recently used in the SEP algorithm (Li et al., 2015) and although seemingly limited, in practice were found to perform almost as well as full EP while significantly reducing EP's memory requirement, from $\mathcal{O}(NLM^2)$ to $\mathcal{O}(LM^2)$ in our case.

The original SEP work devised modified versions of the EP updates appropriate for the new form of the approximate posterior. Originally we applied this method to DGPs (details of this approach including hyperparameter optimisation are included in the supplementary material). However, an alternative approach was found to have superior performance, which is to optimise the EP energy function directly (for both the approximating factors and the hyperparameters). Normally, optimisation of the EP energy requires a double-loop algorithm, which is computationally inefficient, however the use of tied factors simplifies the approximate marginal likelihood and allows direct optimisation. The energy becomes,

$$\begin{aligned}\mathcal{F}(\alpha) &= \phi(\theta) - \phi(\theta_{\text{prior}}) + \sum_{n=1}^{N}\Big[\log \mathcal{Z}_n + \phi(\theta^{\backslash 1}) - \phi(\theta)\Big] \\ &= (1-N)\phi(\theta) + N\phi(\theta^{\backslash 1}) - \phi(\theta_{\text{prior}}) + \sum_{n=1}^{N}\log \mathcal{Z}_n\end{aligned}$$

since the cavity distribution $q^{\backslash n}(\mathbf{u}) \propto q(\mathbf{u})/\tilde{t}_n(\mathbf{u}) = q(\mathbf{u})/g(\mathbf{u}) = q^{\backslash 1}(\mathbf{u})$ is the same for all training points. This elegantly removes the need for a double-loop algorithm, since we can posit a form for the approximate posterior and optimise the above approximate marginal likelihood directly. However, it is important to note that, in general, optimising this objective will not give the same solution as optimising the full EP energy. The new energy produces an approximation formed by averaging the moments of $q^{\backslash 1}(\mathbf{u})\, p(y_n|\mathbf{u}, \mathbf{x}_n)$ over datapoints, whereas EP averages natural parameters, which is arguably more sensible but less tractable.

In detail, we assume the tied factor takes a Gaussian form with natural parameters θ_1. As a result, the approximate posterior and the cavity are also Gaussian with natural parameters $\theta = \theta_{\text{prior}} + N\theta_1$ and $\theta^{\backslash 1} = \theta_{\text{prior}} + (N-1)\theta_1$ respectively. This means that we can compute the first three terms in the energy function exactly. However, it remains to compute $\log \mathcal{Z}_n = \log \int \mathrm{d}\mathbf{u}\, q^{\backslash 1}(\mathbf{u})\, p(y_n|\mathbf{u}, \mathbf{x}_n)$ which we will discuss next.

5 Probabilistic backpropagation for deep Gaussian processes

Computing $\log \mathcal{Z}_n$ in the objective function above is analytically intractable for $L \geq 1$ since the likelihood given the inducing outputs $\mathbf{u}$ is nonlinear and the propagation of the Gaussian cavity through each layer results in a complex distribution. However, for certain choices of covariance functions $\{\mathbf{K}_l\}_{l=1}^{L}$, it is possible to use an efficient and accurate approximation which propagates a Gaussian through the first layer of the network and projects this non-Gaussian distribution back to a moment matched Gaussian before propagating through the next layer and repeating the same steps. This scheme is an algorithmic identical to Assumed Density Filtering and a central part of the probabilistic backpropagation algorithm that has been applied to standard neural networks (Hernández-Lobato & Adams, 2015).

The aim is to compute $\log \mathcal{Z}$ and its gradients with respect to the parameters such as θ_1 or the hyperparameters of the model[4]. By reintroducing the hidden variables in the middle layers, we perform a Gaussian approximation to $\mathcal{Z}$ in a sequential fashion, as illustrated in Figure 2 [right]. We take a two layer case as a running example:

$$\begin{aligned}\mathcal{Z} &= \int \mathrm{d}\mathbf{u}\, p(y|\mathbf{x},\mathbf{u})\, q^{\backslash 1}(\mathbf{u}) \\ &= \int \mathrm{d}h_1\, \mathrm{d}\mathbf{u}_2\, p(y|h_1,\mathbf{u}_2)\, q^{\backslash 1}(\mathbf{u}_2) \int \mathrm{d}\mathbf{u}_1\, p(h_1|\mathbf{x},\mathbf{u}_1)\, q^{\backslash 1}(\mathbf{u}_1)\end{aligned}$$

One key difference between our approach and the variational free energy method of Damianou & Lawrence (2013) is that our algorithm does not retain an explicit approximate distribution over the hidden variables. Instead, we approximately integrate them out when computing $\log \mathcal{Z}$ as follows.

First, we can exactly marginalise out the inducing outputs for each GP layer, leading to $\mathcal{Z} = \int \mathrm{d}h_1\, q(y|h_1)\, q(h_1)$ where $q(h_1) = \mathcal{N}(h_1; m_1, v_1)$, $q(y|h_1) = \mathcal{N}(y|h_1; m_{2|h_1}, v_{2|h_1})$ and

$$\begin{aligned}m_1 &= \mathbf{K}_{h_1,\mathbf{u}_1}\mathbf{K}^{-1}_{\mathbf{u}_1,\mathbf{u}_1}\mathbf{m}_1^{\backslash 1}, \\ v_1 &= \sigma_1^2 + K_{h_1,h_1} - \mathbf{K}_{h_1,\mathbf{u}_1}\mathbf{K}^{-1}_{\mathbf{u}_1,\mathbf{u}_1}\mathbf{K}_{\mathbf{u}_1,h_1} + \mathbf{K}_{h_1,\mathbf{u}_1}\mathbf{K}^{-1}_{\mathbf{u}_1,\mathbf{u}_1}\mathbf{V}_1^{\backslash 1}\mathbf{K}^{-1}_{\mathbf{u}_1,\mathbf{u}_1}\mathbf{K}_{\mathbf{u}_1,h_1}, \\ m_{2|h_1} &= \mathbf{K}_{h_2,\mathbf{u}_2}\mathbf{K}^{-1}_{\mathbf{u}_2,\mathbf{u}_2}\mathbf{m}_2^{\backslash 1}, \\ v_{2|h_1} &= \sigma_2^2 + K_{h_2,h_2} - \mathbf{K}_{h_2,\mathbf{u}_2}\mathbf{K}^{-1}_{\mathbf{u}_2,\mathbf{u}_2}\mathbf{K}_{\mathbf{u}_2,h_2} + \mathbf{K}_{h_2,\mathbf{u}_2}\mathbf{K}^{-1}_{\mathbf{u}_2,\mathbf{u}_2}\mathbf{V}_1^{\backslash 1}\mathbf{K}^{-1}_{\mathbf{u}_2,\mathbf{u}_2}\mathbf{K}_{\mathbf{u}_2,h_2}.\end{aligned}$$

Following (Girard et al., 2003; Barber & Schottky, 1998; Deisenroth & Mohamed, 2012), we can use the law of iterated conditionals to approximate the difficult integral in the equation above by a Gaussian $\mathcal{Z} \approx \mathcal{N}(y|m_2, v_2)$ where the mean and variance take the following form,

$$\begin{aligned}m_2 &= \mathrm{E}_{q(h_1)}[m_{2|h_1}] \\ v_2 &= \mathrm{E}_{q(h_1)}[v_{2|h_1}] + \mathrm{var}_{q(h_1)}[m_{2|h_1}]\end{aligned}$$

which results in

$$\begin{aligned}m_2 &= \mathrm{E}_{q(h_1)}[\mathbf{K}_{h_2,\mathbf{u}_2}]\mathbf{A} \\ v_2 &= \sigma_2^2 + \mathrm{E}_{q(h_1)}[K_{h_2,h_2}] + \mathrm{tr}\left(\mathbf{B}\mathrm{E}_{q(h_1)}[\mathbf{K}_{\mathbf{u}_2,h_2}\mathbf{K}_{h_2,\mathbf{u}_2}]\right) - m_2^2\end{aligned}$$

where $\mathbf{A} = \mathbf{K}^{-1}_{\mathbf{u}_2,\mathbf{u}_2}\mathbf{m}_2^{\backslash 1}$ and $\mathbf{B} = \mathbf{K}^{-1}_{\mathbf{u}_2,\mathbf{u}_2}(\mathbf{V}_2^{\backslash 1} + \mathbf{m}_2^{\backslash 1}\mathbf{m}_2^{\backslash 1,\mathrm{T}})\mathbf{K}^{-1}_{\mathbf{u}_2,\mathbf{u}_2} - \mathbf{K}^{-1}_{\mathbf{u}_2,\mathbf{u}_2}$. The equations above require the expectations of the kernel matrix under a Gaussian distribution over the inputs, which are analytically tractable for widely used kernels such as exponentiated quadratic, linear or a more general class of spectral mixture kernels (Titsias & Lawrence, 2010; Wilson & Adams, 2013). In addition, the approximation above can be improved for networks that have multidimensional intermediate variables, by using a Gaussian with a non-diagonal covariance matrix. We discuss this in the supplementary material.

As the mean and variance of the Gaussian approximation in each intermediate layer can be computed analytically, their gradients with respect to the mean and variance of the input distribution, as well as the parameters of the current layers are also available. Since we require the gradients of the approximation to $\log \mathcal{Z}$, we need to store these results in the forward propagation step, compute

[4]We ignore the data index here to lighten the notation

the approximate $\log \mathcal{Z}$ and its gradients at the output layer and use the chain rule in the backward step to differentiate through the ADF procedure. This procedure is reminiscent of the backpropagation algorithm in standard parametric neural networks, hence the name *probabilistic backpropagation* (Hernández-Lobato & Adams, 2015).

6 Stochastic optimisation for scalable training

The propagation and moment-matching as described above costs $\mathcal{O}(LM^2)$ and needs to be repeated for all datapoints in the training set in batch mode, resulting in an overall complexity of $\mathcal{O}(NLM^2)$. Fortunately, the last term of the objective in Section 4.2 is a sum of independent terms, i.e. its computation can be distributed, resulting in a substantial decrease in computational cost. Furthermore, the objective is suitable for stochastic optimisation. In particular, an unbiased noisy estimate of the objective and its gradients can be obtained using a minibatch of training datapoints,

$$\mathcal{F} \approx -(N-1)\phi(\theta) + N\phi(\theta^{\backslash 1}) - \phi(\theta_{\text{prior}}) + \frac{N}{|B|}\sum_{b=1}^{|B|} \log \mathcal{Z}_b,$$

where $|B|$ denotes the minibatch size.

7 Approximate predictive distribution

Given the approximate posterior and a new test input x^*, we wish to make a prediction about the test output y^*. That is to find $p(y^*|x^*, \mathbf{X}, \mathbf{Y}) \approx \int \mathrm{d}\mathbf{u}\, p(y^*|x^*, \mathbf{u})\, q(\mathbf{u}|\mathbf{X}, \mathbf{Y})$. This predictive distribution is not analytically tractable, but fortunately, we can approximate it by a Gaussian in a similar fashion to the method described in Section 5. That is, a single forward pass is performed, in which each layer takes in a Gaussian distribution over the input, incorporates the approximate posterior of the inducing outputs and approximates the output distribution by a Gaussian. An alternative to obtain the prediction is to forward sample from the model, but we do not use this approach in the experiments.

8 Experiments

We implement and compare the proposed approximation scheme to state-of-the-art methods for Bayesian neural networks. We first detail our implementation in Section 8.1 and then discuss the experimental results in Sections 8.2 and 8.3.

8.1 Experimental details

In all the experiments reported here, we use Adam with the default learning rate (Kingma & Ba, 2015) for optimising our objective function. We use an exponentiated quadratic kernel with ARD lengthscales for each layer. The hyperparameters and pseudo point locations are different between functions in each layer. The lengthscales and inducing inputs of the first GP layer are sensibly initialised based on the median distance between datapoints in the input space and the k-means cluster centers respectively. We use long lengthscales and initial inducing inputs between $[-1, 1]$ for the higher layers to force them to start with an identity mapping. We parameterise the natural parameters of the average factor and initialise them with small random values. We evaluate the predictive performance on the test set using two popular metrics: root mean squared error (RMSE) and mean log likelihood (MLL).

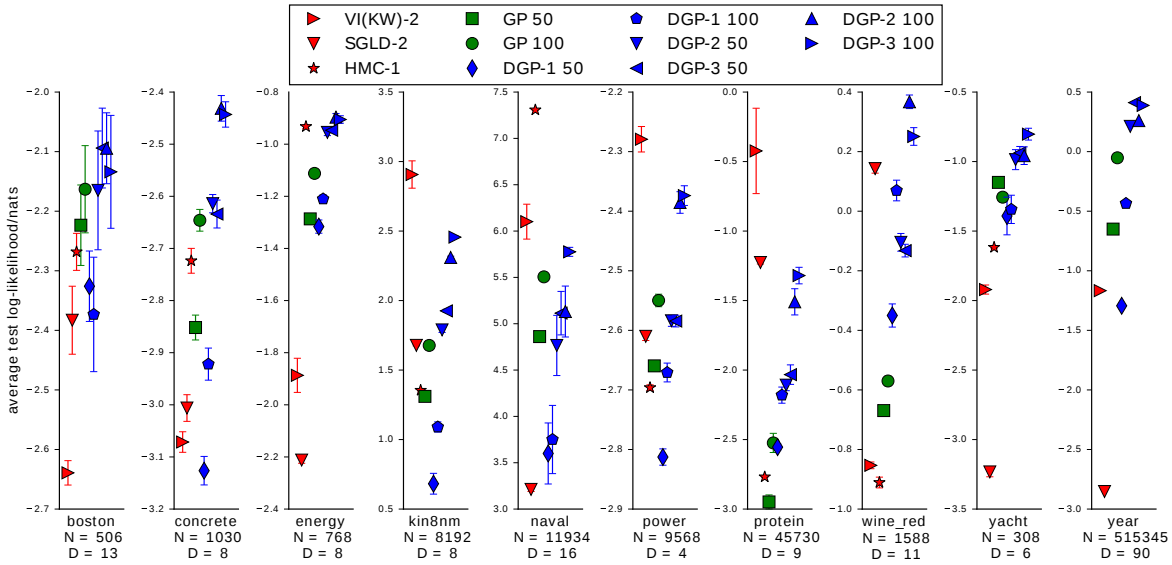

Figure 3: Average predictive log likelihood of existing approaches for BNNs and GPs, and the proposed method for DGPs, across 10 datasets. The higher the better, and best viewed in colour. Full results are included in the supplementary material.

8.2 Regression on UCI datasets

We validate the proposed approach for training DGPs in regression experiments using several datasets from the UCI repository. In particular, we use the ten datasets and train/test splits used in Hernández-Lobato & Adams (2015) and Gal & Ghahramani (2015): 1 split for the *year* dataset [$N \approx 500000, D = 90$], 5 splits for the *protein* dataset [$N \approx 46000, D = 9$], and 20 for the others.

We compare our method (FITC-DGP) against sparse GP regression using FITC (FITC-GP) and Bayesian neural network (BNN) regression using several state-of-the-art deterministic and sampling-based approximate inference techniques. As baselines, we include the results for BNNs reported in Hernández-Lobato & Adams (2015), BNN-VI(G)-1 and BNN-PBP-1, and in Gal & Ghahramani (2015), BNN-Dropout-1. The results reported for these methods are for networks with one hidden layer of 50 units (100 units for *protein* and *year*). Specifically, BNN-VI(G) uses a mean-field Gaussian approximation for the weights in the network, and obtains the stochastic estimates of the bound and its gradient using a Monte Carlo approach (Graves, 2011). BNN-PBP employs Assumed Density Filtering and the probabilistic backpropagation algorithm to obtain a Gaussian approximation for the weights (Hernández-Lobato & Adams, 2015). BNN-Dropout is a recently proposed technique that employs *dropout* during training as well as at prediction time, that is to average over several predictions, each made by the entire network with a random proportion of the weights set to zero (Gal & Ghahramani, 2015). We implement other methods as follows,

- DGP: we evaluate three different architectures of DGPs, each with two GP layers and one hidden layer of one, two and three dimensions respectively (DGP-1, DGP-2 and DGP-3). We include the results for two settings of the number of inducing outputs, $M = 50$ and $M = 100$ respectively. Note that for the bigger datasets *protein* and *year*, we use $M = 100$ and $M = 200$ but do not annotate this in Figure 3. We choose these settings to ensure the run time for our method is smaller or comparable to that of other methods for BNNs.

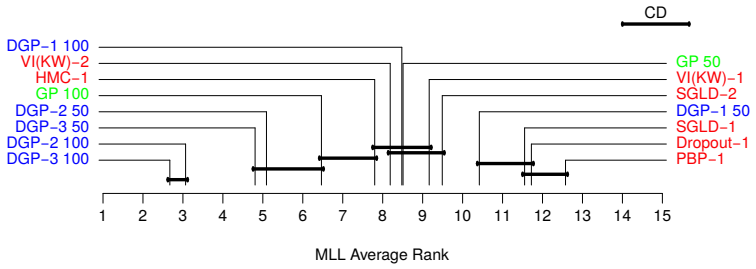

Figure 4: The average rank of all methods across the datasets and their train/test splits, generated based on Demšar (2006). See the text for more details.

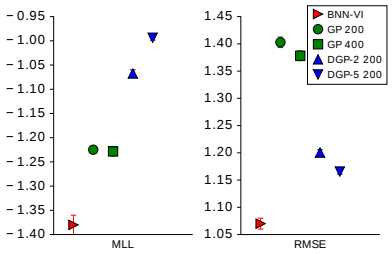

Figure 5: MLL and RMSE results for the photovoltaic molecule regression experiment.

- GP: we use the same number of pseudo datapoints as in DGP (GP 50 and GP 100).
- BNN-VI(KW): this method, similar to Graves (2011), employs a mean-field Gaussian variational approximation but evaluates the variational free energy using the *reparameterisation trick* proposed in Kingma & Welling (2014). We use a diagonal Gaussian prior for the weights and fix the prior variance to 1. The noise variance of the Gaussian noise model is optimised together with the means and variances of the variational approximation using the variational free energy. We test two different network architectures with the rectified linear activation function, and one and two hidden layers, each of 50 units (100 for the two big datasets), denoted by VI(KW)-1 and VI(KW)-2 respectively.
- BNN-SGLD: we reuse the same networks with one and two hidden layers as with VI(KW) and approximately sample from the posterior over the weights using Stochastic Gradient Langevin Dynamics (SGLD) (Welling & Teh, 2011). We place a diagonal Gaussian prior over the weights, and parameterise the observation noise variance as $\sigma^2 = \log(1 + \exp(\kappa))$, a broad Gaussian prior over κ and sample κ using the same SGLD procedure. Two step sizes, one for the weights and one for κ, were manually tuned for each dataset. We use Autograd for the implementation of BNN-SGLD and BNN-VI(KW) (github.com/HIPS/autograd).
- BNN-HMC: We run Hybrid Monte Carlo (HMC) (Neal, 1993) using the MCMCstuff toolbox (Vanhatalo & Vehtari, 2006) for networks with one hidden layer. We place a Gaussian prior over the network weights and a broad inverse Gamma hyper-prior for the prior variance. We also assume an inverse Gamma prior over the observation noise variance. The number of leapfrog steps and step size are first tuned using Bayesian optimisation using the pybo package (github.com/mwhoffman/pybo). Note that this procedure takes a long time (e.g. 3 days for protein) and the *year* dataset is too large to be handled in this way.

Figure 3 shows the average test log likelihood (MLL) for a subset of methods with their standard errors. We exclude methods that perform consistently poorly to improve readability. Full results and many more comparisons are included in the supplementary material. We also evaluate the average rank of the MLL performance of all methods across the datasets and their train/test splits and include the results in Figure 4. This figure is generated using the comparison scheme provided by Demšar (2006), and shows statistical differences in the performance of the methods. More precisely, if the gap between the average ranks of any two methods is above the critical distance (shown on the top right), the two methods' performances are statistically significantly different. Methods that are not

significantly different from each other are linked by a solid line. The rank result shows that DGPs with our inference scheme are the best performing methods overall. Specifically, the DGP-3-100 architecture obtains the best performance on 6 out of 10 datasets and are competitive on the remaining four datasets. The performance of other DGP variants follow closely with the exception for DGP-1 which is a standard warped GP, the network with one dimensional hidden layer. DGP-1 performs poorly compared to GP regression, but is still competitive with several methods for BNNs. The results also strongly indicate that the predictive performance is almost always improved by adding extra hidden layers or extra hidden dimensions or extra inducing outputs.

The best non-GP method is BNN-VI(KW)-2 which obtains the best performance on three datasets. However, this method performs poorly on 6 out of 7 remaining datasets, pushing down the corresponding average rank. Despite this, VI(KW) is the best method among all deterministic approximations for BNNs with one or two hidden layers. Overall, the VI approach without the *reparameterisation trick* of Graves, Dropout and PBP perform poorly in comparison and give inaccurate predictive uncertainty.

Sampling based methods such as SGLD and HMC obtain good predictive performance overall, but often require more tuning compared to other methods. In particular, HMC appears superior on one dataset, and competitively close to DGP's performance on three other datasets; however, this method does not scale to large datasets.

The results for the RMSE metric follow the same trends with DGP-2 and DGP-3 performing as well or better compared to other methods. Interestingly, BNN-SGLD, despite being ranked relatively low according to the MLL metric, often provides good RMSE results. Full results are included in the supplementary material.

8.3 Predicting the efficiency of organic photovoltaic molecules

Having demonstrated the performance of our inference scheme for DGPs, we carry out an additional regression experiment on a challenging dataset. We obtain a subset of 60,000 organic molecules and their power conversion efficiency from the Harvard Clean Energy Project (HCEP) (available at http://www.molecularspace.org) (Hachmann et al., 2011). We use 50,000 molecules for training and 10,000 for testing. The molecules are represented using 512-dimensional binary feature vectors, which were generated using the RDKit package, based on the molecular structures in the canonical SMILES format and a bond radius of 2. The power conversion efficiency of these molecules was estimated using density functional theory, determining whether a molecule could be potentially used as solar cell. The overall aim of the HCEP is to find *organic* molecules that are as efficient as their *silicon* counterparts. Our aim here is to show DGPs are effective predictive models that provide good uncertainty estimates, which can be used for tasks such as Bayesian Optimisation.

We test the method on two DGPs with one hidden layer of 2 and 5 dimensions, denoted by DGP-2 and DGP-5 respectively and each GP is sparsified using 200 inducing outputs. We compare these against two FITC-GP architectures with 200 and 400 pseudo datapoints respectively. We also repeat the experiment using a Bayesian neural network with two hidden layers, each of 400 hidden units. We use the variational approach with the *reparameterisation trick* of Kingma & Welling (2014) to perform inference in this model. The noise variance was fixed to 0.16 based on a suggestion in Pyzer-Knapp et al. (2015). Figure 5 shows the predictive performance by five architectures. The DGP with a five dimensional hidden layer significantly outperforms others in terms of test MLL, including the shallow structure with considerably more pseudo datapoints. This result demonstrates the efficacy of DGPs in providing good predictive uncertainty estimates, even when the kernel used is a *simple* exponentiated quadratic kernel and the input features are binary. Surprisingly, VI(KW), although performing poorly

as measured by the MLL, makes good predictions for the mean.

9 Summary

This paper has introduced a new and powerful deterministic approximation scheme for DGPs based upon an approximate EP algorithm and the FITC approximation to sidestep the computational and analytical intractability. A novel extension of the probabilistic backpropagation algorithm was developed to address a difficult marginalisation problem in the approximate EP algorithm used. The new method was evaluated on eleven datasets and compared against a number of state-of-the-art algorithms for Bayesian neural networks. The results show that the new method for training DGPs is superior on 7 out of 11 datasets considered, and performs comparably on the remainder, demonstrating that DGPs are a competitive alternative to multi-layer Bayesian neural networks for supervised learning tasks.

The proposed method, in principle, can be applied to classification and unsupervised learning. However, initial work on classification using DGPs, as included in the supplementary, does not show a substantial gain over a GP. This issue is potentially related to the diagonal Gaussian approximation currently used for the hidden layers from the second layer onwards. A non-diagonal approximation is feasible but more expensive. This can be easily addressed because the computation of our training method can be distributed on GPUs for example, making it even more scalable. We will investigate both problems in future work.

Acknowledgements

The authors would like to thank Nilesh Tripuraneni, Alex Matthews, Jes Frellsen and Carl Rasmussen for insightful comments and discussion. TDB thanks Google for funding his European Doctoral Fellowship. JMHL acknowledges support from the Rafael del Pino Foundation. DHL and JMHL acknowledge support from Plan Nacional I+D+i, Grant TIN2013-42351-P, and from CAM, Grant S2013/ICE-2845 CASI-CAM-CM. YL thanks the Schlumberger Foundation for her Faculty for the Future PhD fellowship. RET thanks EPSRC grants EP/G050821/1 and EP/L000776/1.

References

Barber, D. and Schottky, B. Radial basis functions: a Bayesian treatment. In *Advances in Neural Information Processing Systems 10*, 1998.

Dai, Zhenwen, Damianou, Andreas, González, Javier, and Lawrence, Neil. Variational auto-encoded deep Gaussian processes. *arXiv preprint arXiv:1511.06455*, 2015.

Damianou, Andreas. *Deep Gaussian processes and variational propagation of uncertainty*. PhD thesis, University of Sheffield, 2015.

Damianou, Andreas C and Lawrence, Neil D. Deep Gaussian processes. In *16th International Conference on Artificial Intelligence and Statistics*, pp. 207–215, 2013.

Deisenroth, Marc and Mohamed, Shakir. Expectation propagation in Gaussian process dynamical systems. In *Advances in Neural Information Processing Systems 25*, pp. 2609–2617, 2012.

Demšar, Janez. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

Duvenaud, David, Rippel, Oren, Adams, Ryan P., and Ghahramani, Zoubin. Avoiding pathologies in very deep networks. In *17th International Conference on Artificial Intelligence and Statistics*, 2014.

Gal, Yarin and Ghahramani, Zoubin. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *arXiv preprint arXiv:1506.02142*, 2015.

Girard, Agathe, Rasmussen, Carl Edward, Quiñonero-Candela, Joaquin, and Murray-Smith, Roderick. Gaussian process priors with uncertain inputs — application to multiple-step ahead time series forecasting. In *Advances in Neural Information Processing Systems 15*, pp. 529–536, 2003.

Graves, Alex. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems 25*, pp. 2348–2356, 2011.

Hachmann, Johannes, Olivares-Amaya, Roberto, Atahan-Evrenk, Sule, Amador-Bedolla, Carlos, Sánchez-Carrera, Roel S, Gold-Parker, Aryeh, Vogt, Leslie, Brockway, Anna M, and Aspuru-Guzik, Alán. The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *The Journal of Physical Chemistry Letters*, 2(17): 2241–2251, 2011.

Hensman, James and Lawrence, Neil D. Nested variational compression in deep Gaussian processes. *arXiv preprint arXiv:1412.1370*, 2014.

Hernández-Lobato, José Miguel and Adams, Ryan P. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *32nd International Conference on Machine Learning*, 2015.

Kingma, D. P. and Ba, J. Adam: a method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015.

Kingma, Diederik P. and Welling, Max. Stochastic gradient VB and the variational auto-encoder. In *The International Conference on Learning Representations*, 2014.

Lawrence, Neil D. and Moore, Andrew J. Hierarchical Gaussian process latent variable models. In *24th International Conference on Machine Learning*, ICML '07, pp. 481–488, New York, NY, USA, 2007.

Lázaro-Gredilla, Miguel. Bayesian warped Gaussian processes. In *Advances in Neural Information Processing Systems 25*, pp. 1619–1627, 2012.

Li, Yingzhen, Hernández-Lobato, José Miguel, and Turner, Richard E. Stochastic expectation propagation. In *Advances in Neural Information Processing Systems 29*, 2015.

Minka, Thomas P. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

Neal, Radford M. Bayesian learning via stochastic dynamics. In *Advances in Neural Information Processing Systems 6*, pp. 475–482, 1993.

Neal, Radford M. *Bayesian learning for neural networks*. PhD thesis, University of Toronto, 1995.

Pyzer-Knapp, Edward O, Li, Kewei, and Aspuru-Guzik, Alan. Learning from the Harvard clean energy project: The use of neural networks to accelerate materials discovery. *Advanced Functional Materials*, 25(41):6495–6502, 2015.

Quiñonero-Candela, Joaquin and Rasmussen, Carl Edward. A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

Rasmussen, Carl Edward and Williams, Christopher K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

Seeger, Matthias. Expectation propagation for exponential families. Technical report, Department of EECS, University of California at Berkeley, 2007.

Snelson, Edward and Ghahramani, Zoubin. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems 19*, pp. 1257–1264, 2006.

Snelson, Edward, Rasmussen, Carl Edward, and Ghahramani, Zoubin. Warped Gaussian processes. In *Advances in Neural Information Processing Systems 17*, pp. 337–344, Cambridge, MA, USA, 2004.

Sutton, Richard S. and Barto, Andrew G. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.

Titsias, Michalis K. Variational learning of inducing variables in sparse Gaussian processes. In *12th International Conference on Artificial Intelligence and Statistics*, pp. 567–574, 2009.

Titsias, Michalis K and Lawrence, Neil D. Bayesian Gaussian process latent variable model. In *13th International Conference on Artificial Intelligence and Statistics*, pp. 844–851, 2010.

Turner, R. E. and Sahani, M. Two problems with variational expectation maximisation for time-series models. In Barber, D., Cemgil, T., and Chiappa, S. (eds.), *Bayesian Time series models*, chapter 5, pp. 109–130. Cambridge University Press, 2011.

Vanhatalo, Jarno and Vehtari, Aki. MCMC methods for MLP-network and Gaussian process and stuff–a documentation for Matlab toolbox MCMCstuff. 2006. Laboratory of computational engineering, Helsinki university of technology.

Welling, Max and Teh, Yee W. Bayesian learning via stochastic gradient Langevin dynamics. In *28th International Conference on Machine Learning*, pp. 681–688, 2011.

Wilson, Andrew and Adams, Ryan. Gaussian process kernels for pattern discovery and extrapolation. In *30th International Conference on Machine Learning*, pp. 1067–1075, 2013.

Appendices

A Extra experimental results

A.1 Regression

Due to the page limitation of the main text, we include here several figures and tables showing the full experimental results and analyses from the regression experiments on 10 UCI datasets. Note that the results for DGPs reported here could be improved further by increasing the number of pseudo datapoints. We choose 50 and 100 pseudo datapoints (or 100 and 200 for the big datasets) so that the training time and prediction time are comparable across all methods. Next we show the full results for the implemented methods and the their average rank across all train/test splits.

- Figures 6 and 7 show the full MLL results for all methods and all datasets. Part of these results have been included in the main text. These figures show that DGPs with our approximation scheme is superior as measured by the MLL metric, obtaining the top spot in the average ranking table.

- Figures 8 and 9 show the full RMSE results for all methods. Surprisingly, though not doing well on the MLL metric, i.e. providing inaccurate predictive uncertainty, BNN-SGLD with one and two layers are very good at predicting the mean of the test set. DGPs, on average, rival or perform better than this approximate sampling scheme and other methods.

- Figures 10 and 11 show the subset of the MLL results above, for GP architectures, and their average ranking. This again demonstrate that DGPs are more flexible than GPs, hence always obtain better predictive performance. The only exception is the network with a one dimensional hidden layer or a warped GP which performs poorly relative to other architectures.

- Similarly, Figures 12 and 13 show evidence that increasing the number of layers and hidden dimensions helps improving the accuracy of the predictions.

- We include a similar analysis for approximate inference methods for BNNs in Figures 14, 15, 16 and 17. This set of results demonstrates that VI(KW) and SGLD with two hidden layers provide good performance on the test sets, outperforming other methods in shallower architectures. HMC with one hidden layer performs well overall, but its running time is much larger compared to other methods. Other deterministic approximations [VI(G), PBP and Dropout] perform poorly overall.

Tables 3 and 4 show the average test log-likelihood and error respectively for all datasets. The best deterministic method for each dataset is bolded, the best method overall (deterministic and sampling) is underlined and emphasised in italic. The average ranks of the methods across the 10 datasets are also included.

A.2 Binary and multiclass classification

We test our approximate inference scheme for DGPs with non-Gaussian noise models. However, as shown in Tables 1 and 2, DGPs often obtain a marginal gain over GPs, as compared to some substantial improvement in the regression experiments above. We speculate that this is due to our

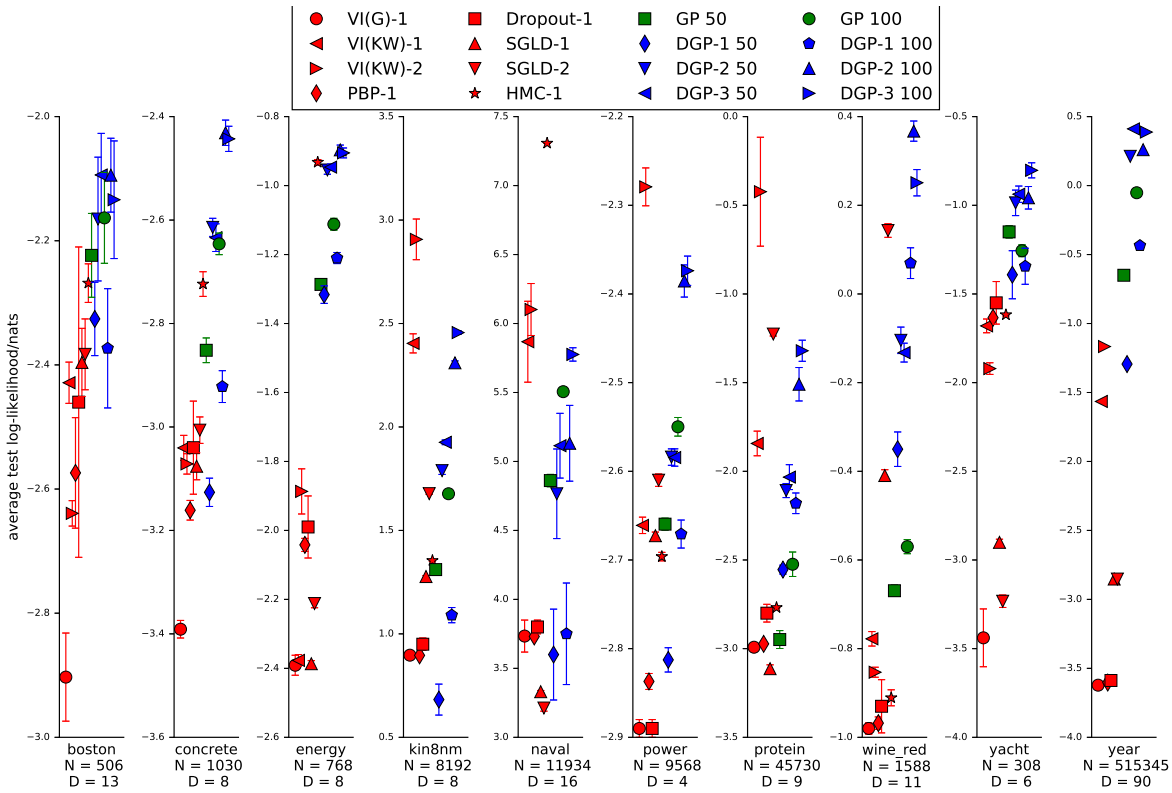

Figure 6: Average test log likelihood for all methods

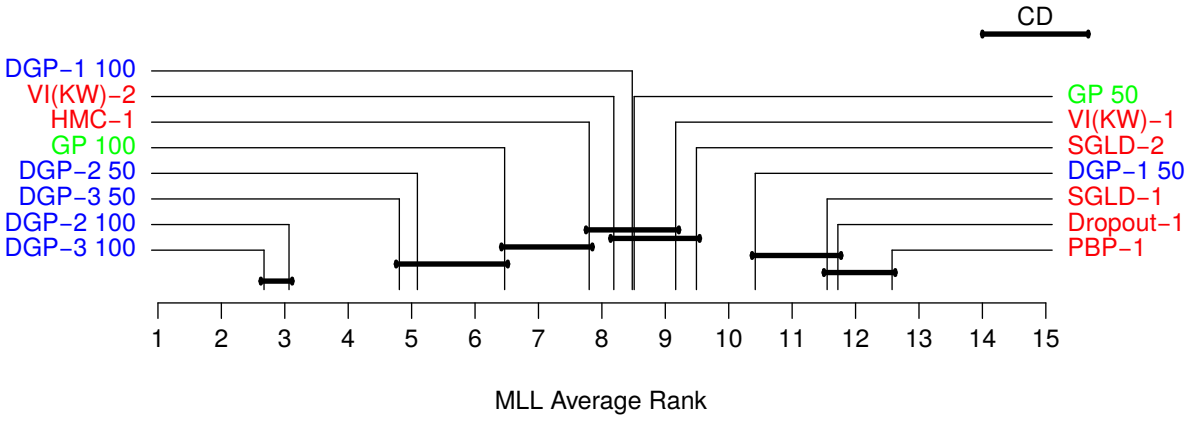

Figure 7: The average rank based on the test MLL of all methods across the datasets and their train/test splits, generated based on Demšar (2006). See the main text for more details.

current initialisation strategy and our diagonal Gaussian approximation at last layer for multiclass classification. We will follow this up in future work.

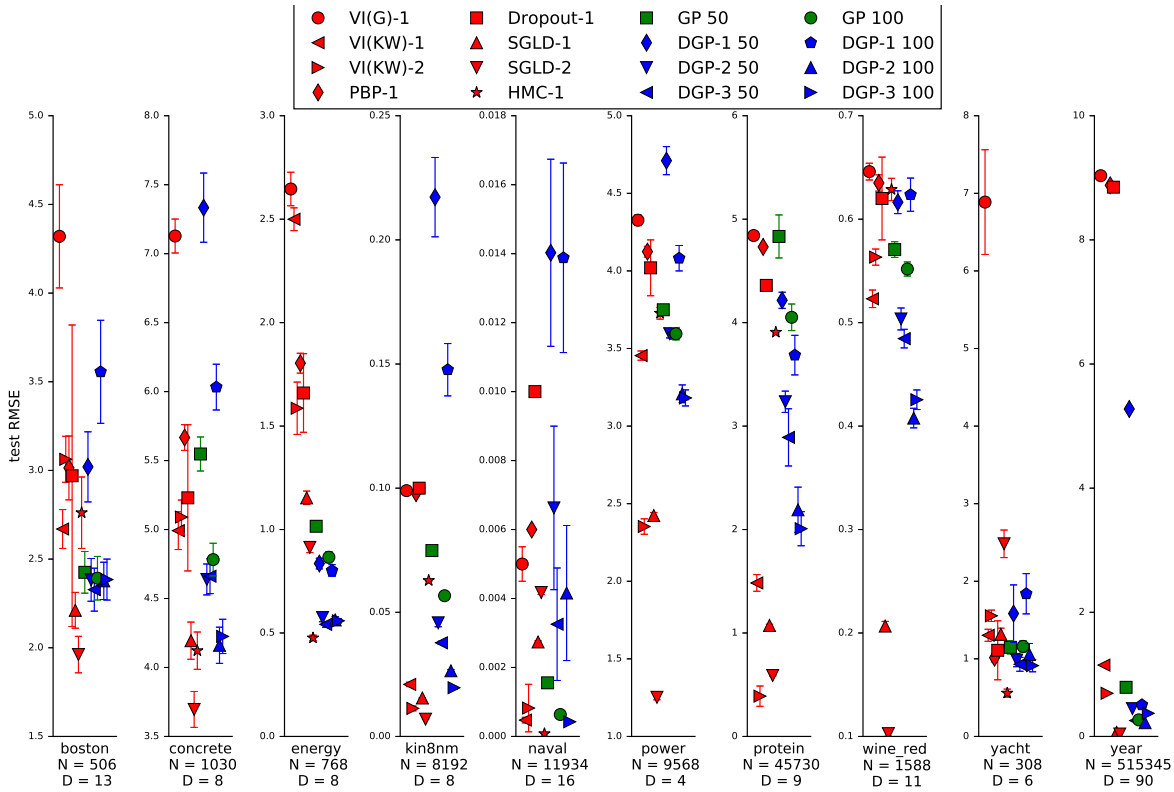

Figure 8: Average test RMSE for all methods

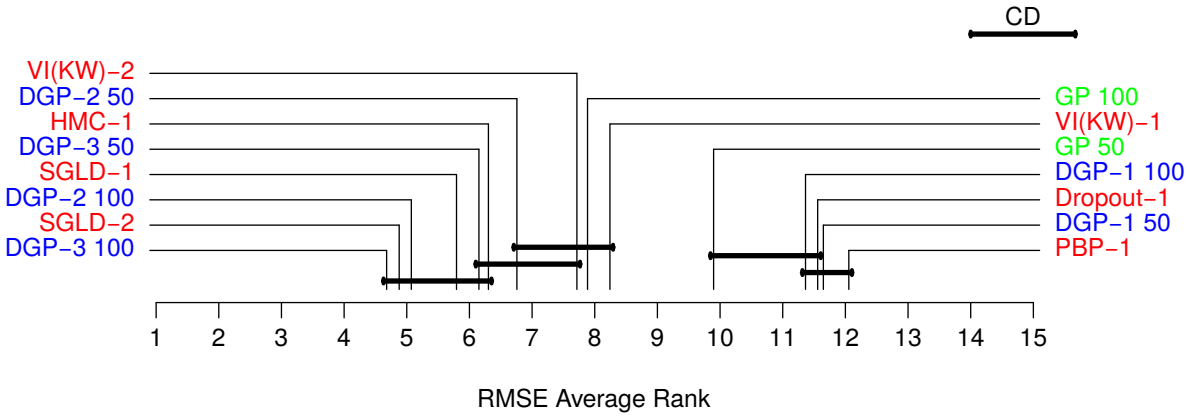

Figure 9: The average rank based on the test RMSE of all methods across the datasets and their train/test splits, generated based on Demšar (2006). See the main text for more details.

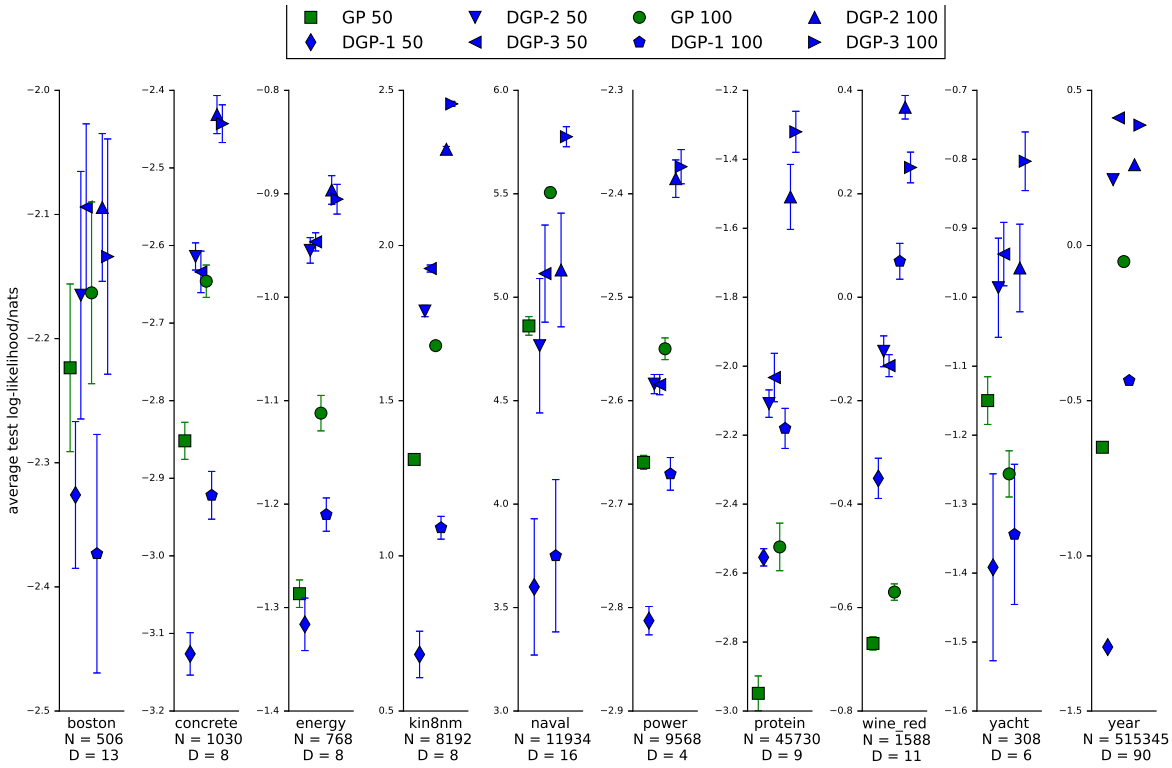

Figure 10: Average test log likelihood for GP methods

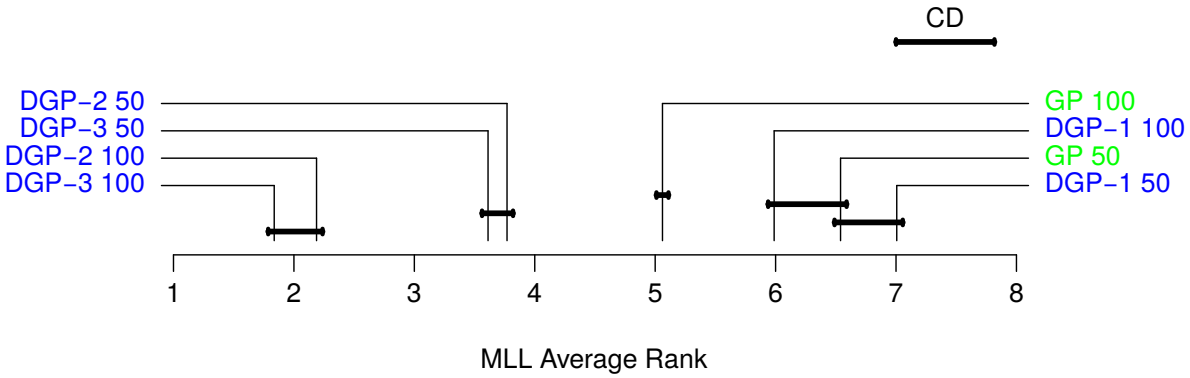

Figure 11: The average rank based on the test MLL for GP/DGP models across the datasets and their train/test splits, generated based on Demšar (2006). See the main text for more details.

Table 1: Binary cla. experiment: Average test log-likelihood/nats

Dataset	**GP D-1**	**DGP D-1-1**	**DGP D-2-1**	**DGP D-3-1**
australian	**-0.51±0.01**	-0.51±0.02	-0.51±0.02	-0.53±0.02
breast	-0.05±0.01	-0.04±0.01	**-0.04±0.01**	-0.04±0.01
crabs	**-0.03±0.01**	-0.10±0.05	-0.03±0.01	-0.03±0.01
ionoshere	-0.17±0.02	-0.17±0.03	-0.16±0.03	**-0.16±0.02**
pima	-0.40±0.01	-0.39±0.01	-0.40±0.02	**-0.39±0.01**
sonar	-0.32±0.03	**-0.29±0.03**	-0.30±0.03	-0.31±0.03

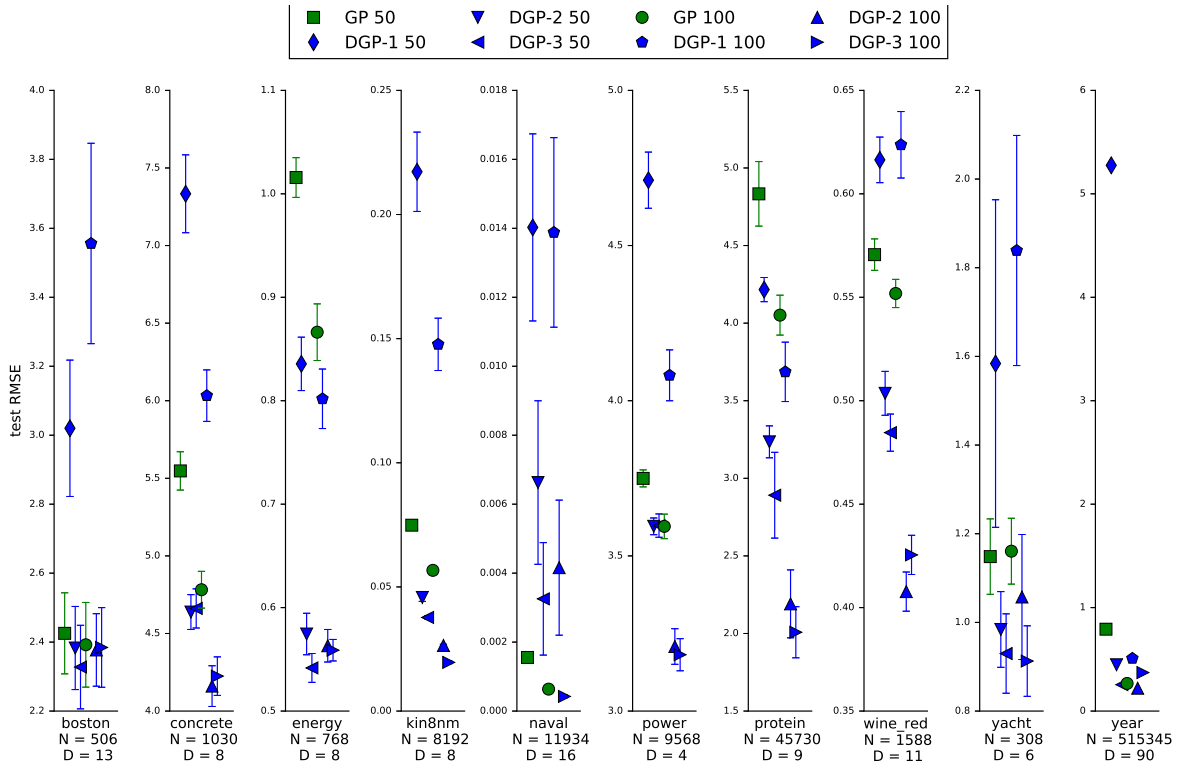

Figure 12: Average test RMSE for GP methods

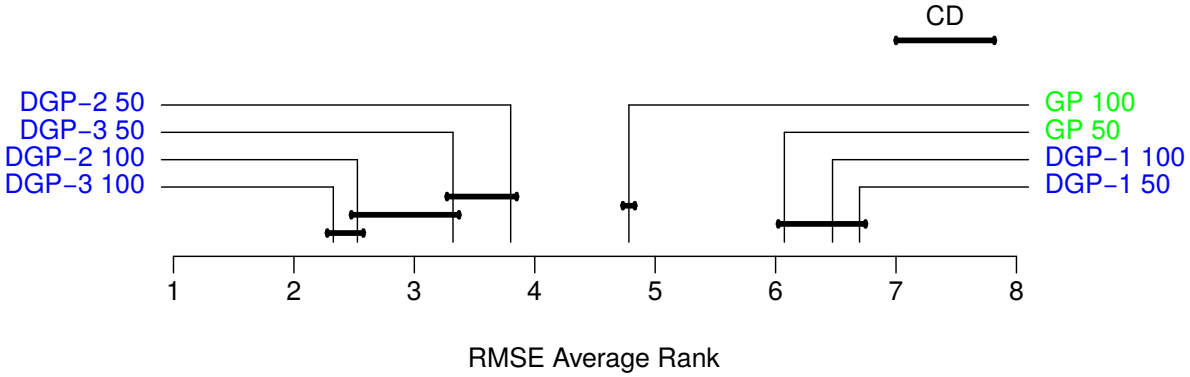

Figure 13: The average rank based on the test RMSE for GP/DGP models across the datasets and their train/test splits, generated based on Demšar (2006). See the main text for more details.

Table 2: Multiclass experiment: Average test log-likelihood/nats

Dataset	N	D	K	**GP D-K**	**DGP D-1-K**	**GP D-2-K**	**DGP D-3-K**
glass	214	9	6	-0.79±0.02	**-0.71±0.02**	-0.72±0.02	-0.71±0.02
new-thyroid	215	5	3	-0.05±0.01	-0.05±0.01	-0.05±0.02	**-0.04±0.01**
svmguide2	319	20	3	-0.54±0.02	-0.53±0.02	-0.52±0.02	**-0.51±0.02**
wine	178	13	3	-0.10±0.01	**-0.07±0.01**	-0.07±0.01	-0.07±0.01

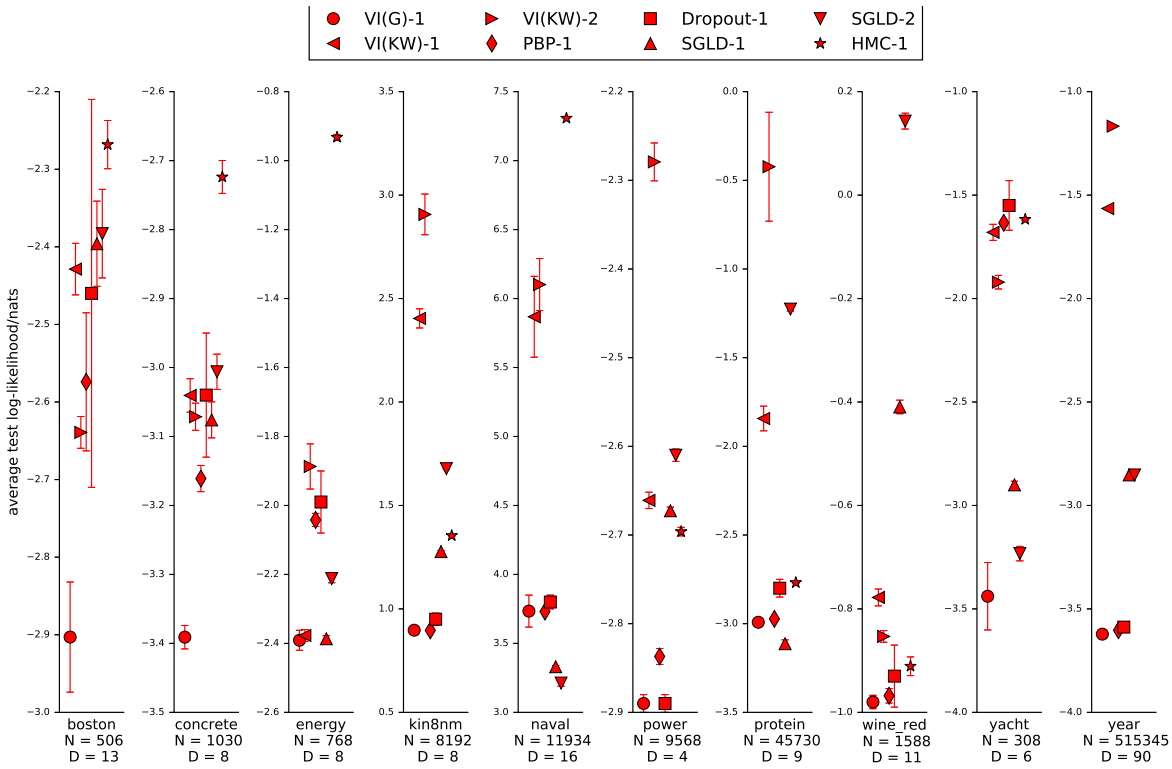

Figure 14: Average test log likelihood for methods with BNNs

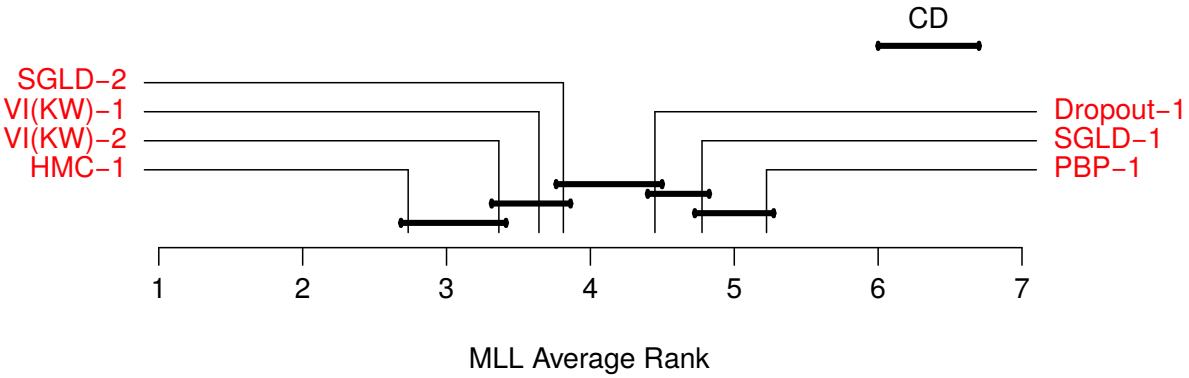

Figure 15: The average rank based on the test MLL for methods on BNNs across the datasets and their train/test splits, generated based on Demšar (2006). See the main text for more details.

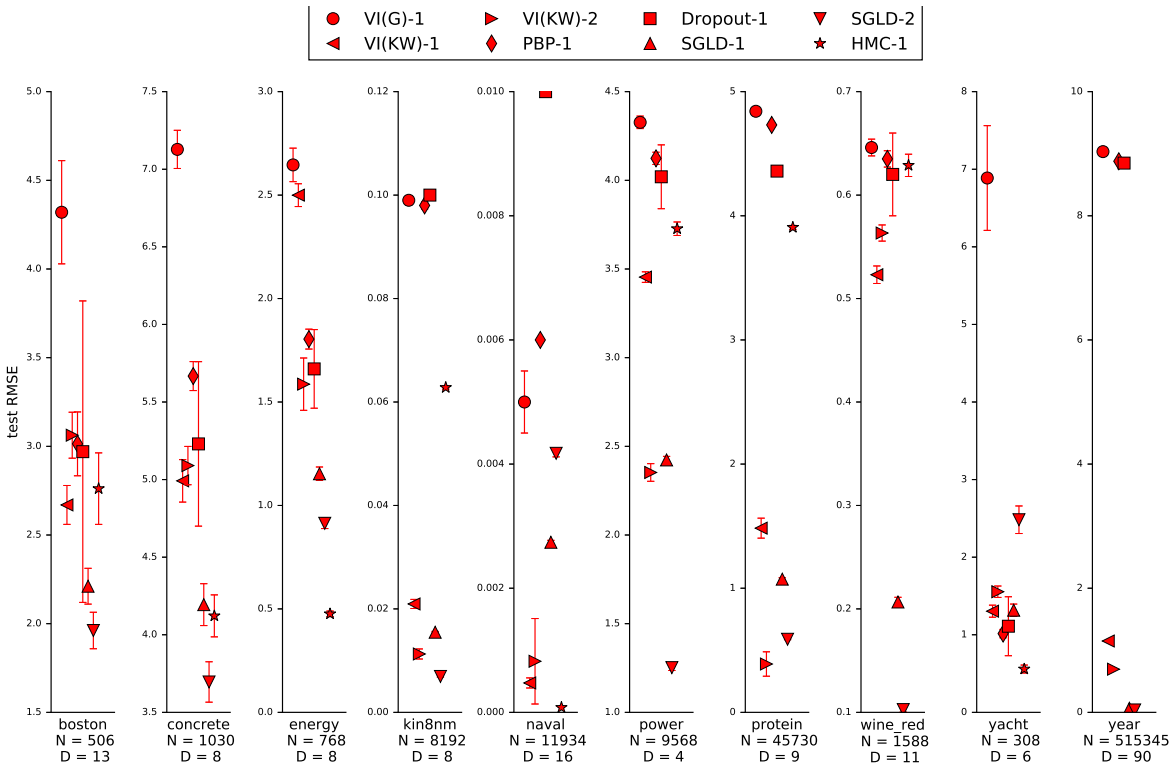

Figure 16: Average test RMSE for methods with BNNs

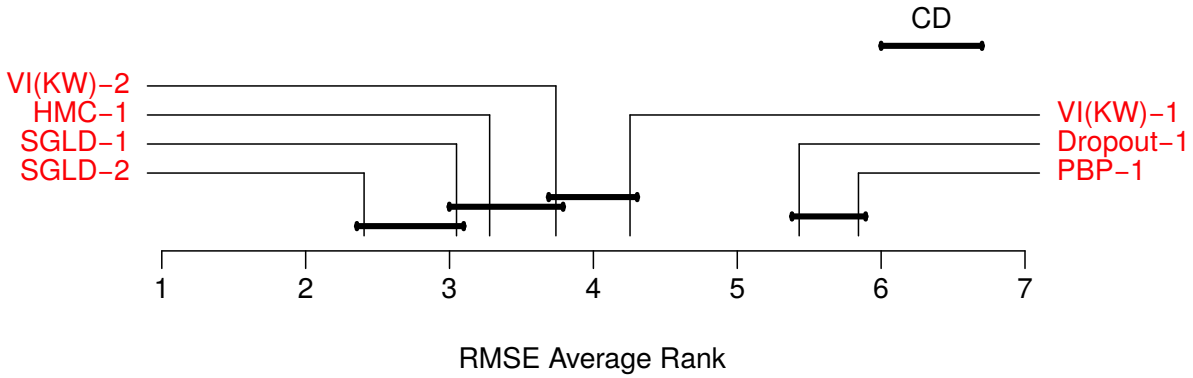

Figure 17: The average rank based on the test RMSE for methods on BNNs across the datasets and their train/test splits, generated based on Demšar (2006). See the main text for more details.

Table 3: Regression experiment: Average test log likelihood/nats

Dataset	N	D	VI(G)-1	VI(KW)-1	VI(KW)-2	PBP-1	Dropout-1	SGLD-1	SGLD-2	HMC-1	GP 50	DGP-1 50	DGP-2 50	DGP-3 50	GP 100	DGP-1 100	DGP-2 100	DGP-3 100
boston	506	13	-2.90±0.07	-2.43±0.03	-2.64±0.02	-2.57±0.09	-2.46±0.25	-2.40±0.05	-2.38±0.06	-2.27±0.03	-2.22±0.07	-2.33±0.06	-2.17±0.10	***-2.09±0.07***	-2.16±0.07	-2.37±0.10	-2.09±0.06	-2.13±0.09
concrete	1030	8	-3.39±0.02	-3.04±0.02	-3.07±0.02	-3.16±0.02	-3.04±0.09	-3.08±0.03	-3.01±0.03	-2.72±0.02	-2.85±0.02	-3.13±0.03	-2.61±0.02	-2.63±0.03	-2.65±0.02	-2.92±0.03	***-2.43±0.02***	-2.44±0.02
energy	768	8	-2.39±0.03	-2.38±0.02	-1.89±0.07	-2.04±0.02	-1.99±0.09	-2.39±0.01	-2.21±0.01	-0.93±0.01	-1.29±0.01	-1.32±0.03	-0.95±0.01	-0.95±0.01	-1.11±0.02	-1.21±0.02	***-0.90±0.01***	-0.91±0.01
kin8nm	8192	8	0.90±0.01	2.40±0.05	***2.91±0.10***	0.90±0.01	0.95±0.03	1.28±0.00	1.68±0.00	1.35±0.00	1.31±0.01	0.68±0.07	1.79±0.02	1.93±0.01	1.68±0.01	1.09±0.04	2.31±0.01	2.46±0.01
naval	11934	16	3.73±0.12	5.87±0.29	**6.10±0.19**	3.73±0.01	3.80±0.05	3.33±0.01	3.21±0.02	*7.31±0.00*	4.86±0.04	3.60±0.33	4.77±0.32	5.11±0.23	5.51±0.03	3.75±0.37	5.13±0.27	5.78±0.05
power	9568	4	-2.89±0.01	-2.66±0.01	***-2.28±0.02***	-2.84±0.01	-2.89±0.01	-2.67±0.00	-2.61±0.01	-2.70±0.00	-2.66±0.01	-2.81±0.01	-2.58±0.01	-2.58±0.01	-2.55±0.01	-2.67±0.02	-2.39±0.02	-2.37±0.02
protein	45730	9	-2.99±0.01	-1.84±0.07	***-0.42±0.31***	-2.97±0.00	-2.80±0.05	-3.11±0.02	-1.23±0.01	-2.77±0.00	-2.95±0.05	-2.55±0.03	-2.11±0.04	-2.03±0.07	-2.52±0.07	-2.18±0.06	-1.51±0.09	-1.32±0.06
red wine	1588	11	-0.98±0.01	-0.78±0.02	-0.85±0.01	-0.97±0.01	-0.93±0.06	-0.41±0.01	0.14±0.02	-0.91±0.02	-0.67±0.01	-0.35±0.04	-0.10±0.03	-0.13±0.02	-0.57±0.02	0.07±0.03	***0.37±0.02***	0.25±0.03
yacht	308	6	-3.44±0.16	-1.68±0.04	-1.92±0.03	-1.63±0.02	-1.55±0.12	-2.90±0.02	-3.23±0.03	-1.62±0.01	-1.15±0.03	-1.39±0.14	-0.99±0.07	-0.94±0.05	-1.26±0.03	-1.34±0.10	-0.96±0.06	***-0.80±0.04***
year	515345	90	-3.62±NA	-1.56±NA	-1.17±NA	-3.60±NA	-3.59±NA	-2.85±NA	-2.85±NA	NA±NA	-0.65±NA	-1.29±NA	0.21±NA	***0.41***±NA	-0.05±NA	-0.44±NA	0.26±NA	0.39±NA
Average Rank			15.10±0.39	9.00±1.18	7.50±1.70	13.70±0.40	12.10±0.64	12.50±0.75	9.40±1.42	8.80±1.38	8.20±0.69	10.80±0.95	5.30±0.51	4.20±0.66	6.10±0.57	8.20±0.72	2.80±0.49	**2.30±0.25**

Table 4: Regression experiment: Test root mean square error

Dataset	N	D	VI(G)-1	VI(KW)-1	VI(KW)-2	PBP-1	Dropout-1	SGLD-1	SGLD-2	HMC-1	GP 50	DGP-1 50	DGP-2 50	DGP-3 50	GP 100	DGP-1 100	DGP-2 100	DGP-3 100
boston	506	13	4.32±0.29	2.67±0.11	3.06±0.13	3.01±0.18	2.97±0.85	2.21±0.10	*1.96±0.10*	2.76±0.20	2.43±0.12	3.02±0.20	2.38±0.12	**2.33±0.12**	2.39±0.12	3.56±0.29	2.38±0.11	2.38±0.12
concrete	1030	8	7.13±0.12	4.99±0.14	5.09±0.12	5.67±0.09	5.23±0.53	4.19±0.13	*3.70±0.13*	4.12±0.14	5.55±0.12	7.33±0.25	4.64±0.11	4.66±0.13	4.78±0.12	6.03±0.17	**4.16±0.13**	4.23±0.12
energy	768	8	2.65±0.08	2.50±0.06	1.59±0.13	1.80±0.05	1.66±0.19	1.15±0.03	0.91±0.03	*0.48±0.01*	1.02±0.02	0.84±0.03	0.57±0.02	**0.54±0.01**	0.87±0.03	0.80±0.03	0.56±0.02	0.56±0.01
kin8nm	8192	8	0.10±0.00	0.02±0.00	**0.01±0.00**	0.10±0.00	0.10±0.00	0.02±0.00	*0.01±0.00*	0.06±0.00	0.07±0.00	0.22±0.02	0.05±0.00	0.04±0.00	0.06±0.00	0.15±0.01	0.03±0.00	0.02±0.00
naval	11934	16	0.01±0.00	0.00±0.00	0.00±0.00	0.01±0.00	0.01±0.00	0.00±0.00	0.00±0.00	*0.00±0.00*	0.00±0.00	0.01±0.00	0.01±0.00	0.00±0.00	0.00±0.00	0.01±0.00	0.00±0.00	**0.00±0.00**
power	9568	4	4.33±0.04	3.45±0.03	**2.35±0.05**	4.12±0.03	4.02±0.18	2.42±0.02	*1.25±0.02*	3.73±0.04	3.75±0.03	4.71±0.09	3.60±0.03	3.60±0.04	3.60±0.04	4.08±0.08	3.21±0.06	3.18±0.05
protein	45730	9	4.84±0.03	1.48±0.08	***0.39±0.10***	4.73±0.01	4.36±0.04	1.07±0.01	0.59±0.00	3.91±0.02	4.83±0.21	4.22±0.08	3.24±0.10	2.89±0.28	4.05±0.13	3.69±0.19	2.19±0.22	2.01±0.16
red wine	1588	11	0.65±0.01	0.52±0.01	0.56±0.01	0.64±0.01	0.62±0.04	0.21±0.00	*0.10±0.00*	0.63±0.01	0.57±0.01	0.62±0.01	0.50±0.01	0.48±0.01	0.55±0.01	0.62±0.02	**0.41±0.01**	0.43±0.01
yacht	308	6	6.89±0.67	1.30±0.08	1.55±0.07	1.01±0.05	1.11±0.38	1.32±0.08	2.48±0.18	*0.56±0.05*	1.15±0.09	1.58±0.37	0.98±0.09	0.93±0.09	1.16±0.07	1.84±0.26	1.06±0.14	**0.91±0.08**
year	515345	90	9.03±NA	1.15±NA	0.70±NA	8.88±NA	8.85±NA	0.07±NA	*0.04*±NA	NA±NA	0.79±NA	5.28±NA	0.45±NA	0.26±NA	0.27±NA	0.51±NA	**0.22**±NA	0.37±NA
Average Rank			14.90±0.50	7.90±1.09	7.60±1.42	12.50±0.85	12.00±0.62	4.80±1.08	4.20±1.55	7.50±1.72	10.10±0.74	13.20±0.88	7.00±0.76	5.50±0.72	7.60±0.60	12.20±0.99	4.90±0.57	**4.10±0.43**

B EP and SEP

In this section, we summarise the EP and SEP iterative procedures. The EP algorithm is often mistaken to be optimising $\mathrm{KL}(p(\mathbf{u}|\mathbf{X},\mathbf{y})||q(\mathbf{u}))$; however, this objective function is intractable. Instead, EP updates one approximate factor at a time by the following procedure: 1. remove the factor $\tilde{t}_n(\mathbf{u})$ to form the leave-one-out or cavity distribution $q^{\backslash n}(\mathbf{u}) \propto q(\mathbf{u})/\tilde{t}_n(\mathbf{u})$, 2. minimise $\mathrm{KL}(q^{\backslash n}(\mathbf{u})p(y_n|\mathbf{u},\mathbf{X}_n)||q(\mathbf{u}))$, resulting in a new approximate factor $\tilde{t}_n^{\mathrm{new}}(\mathbf{u})$ which can be 3. combined with the cavity to form the new approximate posterior. This procedure is iteratively performed for each datapoint, and often requires several passes through the training set for convergence. One disadvantage of the EP algorithm is the need to store the approximate factors in memory, which costs $\mathcal{O}(NM^2)$.

To sidestep this expensive memory requirement, the SEP algorithm proposes tying the approximate data factors, that is to make some or all factors the same. The simplest case is $q(\mathbf{u}) \propto p(\mathbf{u})g(\mathbf{u})^N$ where $g(\mathbf{u})$ is the *average* data factor. The SEP algorithm, similar to EP, involves iteratively finding the new approximate factor $g_{\mathrm{new}}(\mathbf{u})$, as follows: 1. remove the factor $\tilde{g}(\mathbf{u})$ to form the leave-one-out or cavity distribution $q^{\backslash 1}(\mathbf{u}) \propto q(\mathbf{u})/\tilde{g}(\mathbf{u})$, 2. minimise $\mathrm{KL}(q^{\backslash 1}(\mathbf{u})p(y_n|\mathbf{u},\mathbf{X}_n)||q(\mathbf{u}))$, resulting in a new approximate factor $\tilde{g}_{\mathrm{new}}(\mathbf{u})$ which can be 3. combined with the cavity to form the new approximate posterior, and in addition to EP, 4. perform an explicit update to the *average* factor $g(\mathbf{u})$: $g(\mathbf{u}) \leftarrow g^{1-\beta}(\mathbf{u})g^{\beta}_{\mathrm{new}}(\mathbf{u})$, where β is a small learning rate.

C EP/SEP moment matching step

We have proposed using the EP approximate marginal likelihood for direct optimisation of the approximate posterior over the pseudo datapoints and the hyperparameters. An alternative is to run SEP/EP to obtain the approximate posterior, and once this is done, obtain the approximate marginal likelihood for hyperparameter tuning and repeat.

As we use Gaussian EP/SEP, the deletion, the update step and the explicit update step in the case of SEP are straightforward. The moment matching step is equivalent to the following updates to the mean and covariance of the approximate posterior:

$$\begin{aligned}\mathbf{m} &= \mathbf{m}^{\backslash 1} + \mathbf{V}^{\backslash 1}\frac{\mathrm{d}\log\mathcal{Z}}{\mathrm{d}\mathbf{m}^{\backslash 1}} \\ \mathbf{V} &= \mathbf{V}^{\backslash 1} - \mathbf{V}^{\backslash 1}\left[\frac{\mathrm{d}\log\mathcal{Z}}{\mathrm{d}\mathbf{m}^{\backslash 1}}\left(\frac{\mathrm{d}\log\mathcal{Z}}{\mathrm{d}\mathbf{m}^{\backslash 1}}\right)^{\intercal} - 2\frac{\mathrm{d}\log\mathcal{Z}}{\mathrm{d}\mathbf{V}^{\backslash 1}}\right]\mathbf{V}^{\backslash 1},\end{aligned}$$

where $q^{\backslash 1}(\mathbf{u}) = \mathcal{N}(\mathbf{u};\mathbf{m}^{\backslash 1},\mathbf{V}^{\backslash 1})$ is the cavity distribution, obtained by the deletion step.

The inference scheme therefore reduces to evaluating the normalising constant $\mathcal{Z}$ and its gradient. Fortunately, we can approximately compute $\log\mathcal{Z}$ and its gradients using the probabilistic propagation algorithm, in exactly the same way as discussed in the main text.

D Computing the gradients of $\log\mathcal{Z}$

Let m_l and v_l be the mean and variance of the output Gaussian at the l-th layer in the forward propagation step, as we have shown in the main text,

$$m_l = \psi_{l,1}\mathbf{A}_l \tag{1}$$

$$v_l = \sigma_l^2 + \psi_{l,0} + \mathrm{tr}\left(\mathbf{B}_l\psi_{l,2}\right) - m_l^2 \tag{2}$$

where

$$\psi_{l,0} = \mathrm{E}_{q(h_1)}[K_{h_l,h_l}] \tag{3}$$
$$\psi_{l,1} = \mathrm{E}_{q(h_{l-1})}[\mathbf{K}_{h_l,\mathbf{u}_l}] \tag{4}$$
$$\psi_{l,1} = \mathrm{E}_{q(h_{l-1})}[\mathbf{K}_{\mathbf{u}_l,h_l}\mathbf{K}_{h_l,\mathbf{u}_l}] \tag{5}$$
$$\mathbf{A}_l = \mathbf{K}_{\mathbf{u}_l,\mathbf{u}_l}^{-1}\mathbf{m}_l^{\backslash 1} \tag{6}$$
$$\mathbf{B}_l = \mathbf{K}_{\mathbf{u}_l,\mathbf{u}_l}^{-1}(\mathbf{V}_l^{\backslash 1} + \mathbf{m}_l^{\backslash 1}\mathbf{m}_l^{\backslash 1,\mathrm{T}})\mathbf{K}_{\mathbf{u}_l,\mathbf{u}_l}^{-1} - \mathbf{K}_{\mathbf{u}_l,\mathbf{u}_l}^{-1} \tag{7}$$

In the forward propagation step, we need to compute the gradients of m_l and v_l w.r.t. α_l, the parameters of the model and m_{l-1} and v_{l-1}, the mean and variance of the distribution over the input. Let $\beta_l = \{\alpha_l, m_{l-1}, v_{l-1}\}$ As $\mathbf{A}_l$ and $\mathbf{B}_l$ are shared between datapoints, one trick to reduce the computation required for each datapoint is to compute the gradients w.r.t. $\mathbf{A}$ and $\mathbf{B}$ first, then combine them at the end of each minibatch. If we assume that $\mathbf{A}_l$ and $\mathbf{B}_l$ are fixed, the gradients of m_l and v_l are as follows

$$\frac{\mathrm{d}m_l}{\mathrm{d}\beta_l} = \frac{\mathrm{d}\psi_{l,1}}{\mathrm{d}\beta_l}\mathbf{A}_l \tag{8}$$
$$\frac{\mathrm{d}v_l}{\mathrm{d}\beta_l} = \frac{\mathrm{d}\sigma_l^2}{\mathrm{d}\beta_l} + \frac{\mathrm{d}\psi_{l,0}}{\mathrm{d}\beta_l} + \mathrm{tr}\left(\mathbf{B}_l\frac{\mathrm{d}\psi_{l,2}}{\mathrm{d}\beta_l}\right) - 2m_l\frac{\mathrm{d}m_l}{\mathrm{d}\beta_l} \tag{9}$$
$$\frac{\mathrm{d}m_l}{\mathrm{d}\mathbf{A}_l} = \psi_{l,1}^{\intercal} \tag{10}$$
$$\frac{\mathrm{d}m_l}{\mathrm{d}\mathbf{B}_l} = \mathbf{0} \tag{11}$$
$$\frac{\mathrm{d}v_l}{\mathrm{d}\mathbf{A}_l} = -2m_l\frac{\mathrm{d}m_l}{\mathrm{d}\mathbf{A}_l} \tag{12}$$
$$\frac{\mathrm{d}v_l}{\mathrm{d}\mathbf{B}_l} = \psi_{l,2}^{\intercal} \tag{13}$$

At the end of the forward step, we can obtain $Z = q(y) = \mathcal{N}(y; m_L, v_L)$, leading to,

$$\log \mathcal{Z} = -\frac{1}{2}\log(2\pi v_L) - \frac{1}{2}\frac{(y-m_L)^2}{v_L} \tag{14}$$
$$\frac{\mathrm{d}\log\mathcal{Z}}{\mathrm{d}m_L} = \frac{y-m_L}{v_L} \tag{15}$$
$$\frac{\mathrm{d}\log\mathcal{Z}}{\mathrm{d}v_L} = -\frac{1}{2v_L} + \frac{1}{2}\frac{(y-m_L)^2}{v_L^2}. \tag{16}$$

We are now ready to perform the backpropagation step, that is we compute the gradients of $\log\mathcal{Z}$ w.r.t. parameters at a layer α_l using the chain rule,

$$\frac{\mathrm{d}\log\mathcal{Z}}{\mathrm{d}\alpha_l} = \frac{\mathrm{d}\log\mathcal{Z}}{\mathrm{d}m_l}\frac{\mathrm{d}m_l}{\mathrm{d}\alpha_l} + \frac{\mathrm{d}\log\mathcal{Z}}{\mathrm{d}v_l}\frac{\mathrm{d}v_l}{\mathrm{d}\alpha_l}. \tag{17}$$

Similarly, we can compute the gradients w.r.t. the mean and variance of the input distribution, m_{l-1} and v_{l-1}, and $\mathbf{A}_l$ and $\mathbf{B}_l$.

E Computing the gradients of the approximate marginal likelihood

The approximate marginal likelihood as discussed in the main text is as follows,

$$\mathcal{F} = -(N-1)\phi(\theta) + N\phi(\theta^{\backslash 1}) - \phi(\theta_{\text{prior}}) + \sum_{n=1}^{N} \log \mathcal{Z}_n \tag{18}$$

where $\theta, \theta^{\backslash 1}$ and θ_{prior} are the natural parameters of $q(\mathbf{u})$, $q^{\backslash 1}(\mathbf{u})$ and $p(\mathbf{u})$ respectively, $\phi(\theta)$ is the log normaliser or log partition function of a Gaussian distribution with natural parameters θ or mean $\mathbf{m}$ and covariance $\mathbf{V}$,

$$\phi(\theta) = \frac{1}{2}\log|\mathbf{V}| + \frac{1}{2}\mathbf{m}^\intercal \mathbf{V}^{-1}\mathbf{m}, \tag{19}$$

α is the model hyperameters that we need to tune, and $\log \mathcal{Z}_n = \log \int q^{\backslash n}(\mathbf{u}) p(y_n|\mathbf{u}, \mathbf{X}_n) \mathrm{d}\mathbf{u}$. Consider the gradient of this objective function w.r.t. one parameter α_i,

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{F}}{\mathrm{d}\alpha_i} &= -(N-1)\frac{\mathrm{d}\phi(\theta)}{\mathrm{d}\alpha_i} + N\frac{\mathrm{d}\phi(\theta^{\backslash 1})}{\mathrm{d}\alpha_i} \\
&\quad - \frac{\mathrm{d}\phi(\theta_{\text{prior}})}{\mathrm{d}\alpha_i} + \sum_{n=1}^{N}\frac{\mathrm{d}\log \mathcal{Z}_n}{\mathrm{d}\alpha_i} \\
&= -(N-1)\frac{\mathrm{d}\phi(\theta)}{\mathrm{d}\theta}\frac{\mathrm{d}\theta}{\mathrm{d}\alpha_i} + N\frac{\mathrm{d}\phi(\theta^{\backslash 1})}{\mathrm{d}\theta^{\backslash 1}}\frac{\mathrm{d}\theta^{\backslash 1}}{\mathrm{d}\alpha_i} \\
&\quad - \frac{\mathrm{d}\phi(\theta_{\text{prior}})}{\mathrm{d}\theta_{\text{prior}}}\frac{\mathrm{d}\theta_{\text{prior}}}{\mathrm{d}\alpha_i} + \sum_{n=1}^{N}\frac{\mathrm{d}\log \mathcal{Z}_n}{\mathrm{d}\alpha_i} \\
&= -(N-1)\eta^\intercal \frac{\mathrm{d}\theta}{\mathrm{d}\alpha_i} + N\eta^{\backslash 1,\intercal}\frac{\mathrm{d}\theta^{\backslash 1}}{\mathrm{d}\alpha_i} \\
&\quad - \eta_{\text{prior}}^\intercal \frac{\mathrm{d}\theta_{\text{prior}}}{\mathrm{d}\alpha_i} + \sum_{n=1}^{N}\frac{\mathrm{d}\log \mathcal{Z}_n}{\mathrm{d}\alpha_i}
\end{aligned}$$

where η, $\eta_{\backslash 1}$ and η_{prior} are the expected sufficient statistics under the $q(\mathbf{u})$, $q^{\backslash 1}(\mathbf{u})$ and $p(\mathbf{u})$ respectively. Specifically, for Gaussian approximate EP as discussed in the main paper, the natural parameters are as follows,

$$\begin{aligned}
&q(\mathbf{u}) : \theta = \theta_{\text{prior}} + N\theta_g \\
&q^{\backslash 1}(\mathbf{u}) : \theta^{\backslash 1} = \theta_{\text{prior}} + (N-1)\theta_g \\
&p(\mathbf{u}) : \theta_{\text{prior}}
\end{aligned}$$

leading to

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{F}}{\mathrm{d}\alpha_i} &= \left[-(N-1)\eta^\intercal + N\eta^{\backslash 1,\intercal} - \eta_{\text{prior}}^\intercal\right]\frac{\mathrm{d}\theta_{\text{prior}}}{\mathrm{d}\alpha_i} \\
&\quad + N(N-1)\left[-\eta^\intercal + \eta^{\backslash 1,\intercal}\right]\frac{\mathrm{d}\theta_g}{\mathrm{d}\alpha_i} + \sum_{n=1}^{N}\frac{\mathrm{d}\log \mathcal{Z}_n}{\mathrm{d}\alpha_i}
\end{aligned}$$

F Dealing with non-Gaussian likelihoods

In this section, we discuss how to compute the log of $\mathcal{Z} = \int \mathrm{d}\mathbf{u}\, q^{\backslash 1}(\mathbf{u})\, p(y|\mathbf{u}, \mathbf{x})$ when we have a non-Gaussian likelihood $p(y|\mathbf{u}, \mathbf{x})$. For example, if the observations are binary, we can use the probit likelihood, that is $p(y|\mathrm{f}_L, h_{L-1}) = \phi(y\mathrm{f}_L)$ where ϕ is the Gaussian cdf. We now need to compute,

$$\begin{aligned}
\mathcal{Z} &= \int q^{\backslash 1}(\mathbf{u}) p(y|\mathbf{u}, \mathbf{x}) \mathrm{d}\mathbf{u} \\
&= \int q^{\backslash 1}(\mathbf{u}) p(\mathrm{f}_L|h_{L-1}, \mathbf{u}_L) p(y|\mathrm{f}_L) \mathrm{d}\mathbf{u} \mathrm{d}h_{L-1} \mathrm{d}\mathrm{f}_L \\
&\approx \int \mathcal{N}(\mathrm{f}_L; m_\mathrm{f}, v_\mathrm{f}) p(y|\mathrm{f}_L) \mathrm{d}\mathrm{f}_L
\end{aligned}$$

where we can find $q(\mathrm{f}_L) = \mathcal{N}(\mathrm{f}_L; m_\mathrm{f}, v_\mathrm{f})$ using the forward pass of the probabilistic backpropagation. The final integral above can be computed exactly, leading to,

$$\mathcal{Z} \approx \phi\left(\frac{y m_\mathrm{f}}{\sqrt{v_\mathrm{f} + 1}}\right)$$

If we have a different likelihood and there is no simple approximation available as above, we can evaluate $\mathcal{Z}$ by Monte Carlo averaging, that is to draw samples from $q(\mathrm{f}_L)$, evaluate the likelihood, then sum and normalise accordingly. However, as we are interested in $\log \mathcal{Z}$ and its gradients, the objective and gradients obtained by Monte Carlo will be slightly biased. This bias is, however, can be significantly reduced by using more samples.

G Improving the Gaussian approximation

In this section, we discuss how to obtain a non-diagonal Gaussian approximation for the hidden variables from the second layer and above, when computing $\log \mathcal{Z}$. Consider a DGP with two GP layer, a one dimensional hidden layer and two dimensional observations $\mathbf{y} = [y_1, y_2]$. Following the derivation in the main text, we can exactly marginalise out the inducing outputs for each GP layer:

$$\mathcal{Z} = \int \mathrm{d}h_1 q(\mathbf{y}|h_1) q(h_1) \tag{20}$$

where $q(h_1) = \mathcal{N}(h_1; m_1, v_1)$ and

$$\begin{aligned}
q(\mathbf{y}|h_1) &= \mathcal{N}(\mathbf{y}|h_1; \mathbf{m}_{\mathbf{y}|h_1}, \mathbf{V}_{\mathbf{y}|h_1}) \\
&= \mathcal{N}\left(\mathbf{y}|h_1; \begin{bmatrix} m_{y_1|h_1} \\ m_{y_2|h_1} \end{bmatrix}, \begin{bmatrix} v_{y_1|h_1} & 0 \\ 0 & v_{y_2|h_1} \end{bmatrix}\right)
\end{aligned}$$

since we assume that there are two independent GPs in the second layer, and the distribution above is a conditional given the input to the second layer, h_1. Importantly, we need to integrate out h_1 in eqn. (20). As such, the resulting distribution over $\mathbf{y}$ become a complicated distribution in which y_1 and y_2 are strongly correlated. Consequently, any approximation that breaks this dependency could be poor. We aim to approximate this distribution by a non-diagonal Gaussian with the same moments, that is in words, the approximating Gaussian will have the mean being the expected mean, and the new covariance being the expected covariance plus the covariance of the mean,

$$\mathbf{m}_\mathbf{y} = \mathrm{E}_{q(h_1)}[\mathbf{m}_{\mathbf{y}|h_1}] \tag{21}$$

$$\mathbf{V_y} = \mathrm{E}_{q(h_1)}[\mathbf{V_{y|h_1}}] + \mathrm{covar}_{q(h_1)}[\mathbf{m_{y|h_1}}] \tag{22}$$

Substitute the mean and covariance of the conditional $q(\mathbf{y}|h_1)$ into the above expressions gives us,

$$\mathbf{m_y} = \begin{bmatrix} \mathrm{E}_{q(h_1)}[m_{y_1|h_1}] \\ \mathrm{E}_{q(h_1)}[m_{y_2|h_1}] \end{bmatrix} \tag{23}$$

and

$$\begin{aligned} \mathbf{V_y} = & \begin{bmatrix} \mathrm{E}_{q(h_1)}[v_{y_1|h_1}] & 0 \\ 0 & \mathrm{E}_{q(h_1)}[v_{y_2|h_1}] \end{bmatrix} \\ & + \begin{bmatrix} \mathrm{E}_{q(h_1)}[m^2_{y_1|h_1}] & \mathrm{E}_{q(h_1)}[m_{y_1|h_1} m_{y_2|h_1}] \\ \mathrm{E}_{q(h_1)}[m_{y_1|h_1} m_{y_2|h_1}] & \mathrm{E}_{q(h_1)}[m^2_{y_2|h_1}] \end{bmatrix} \\ & - \mathbf{m_y}\mathbf{m_y}^\intercal \end{aligned} \tag{24}$$

Note that the diagonal elements of $\mathbf{V_y}$ are identical to the expression for the variance in the main text for the single dimensional case.