Supplementary Material
"Sparse Gaussian Processes Revisited: Bayesian Approaches to Inducing-Variable Approximations"

Simone Rossi, Markus Heinonen, Edwin V. Bonilla, Zheyang Shen and Maurizio Filippone

1 A Primer on Inference in Sparse GP Models

A Gaussian process (GP) defines a distribution over functions $f(\mathbf{x}) : \mathbb{R}^D \to \mathbb{R}$ such that for any subset of points $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ the function values $\{f(\mathbf{x}_1), \dots, f(\mathbf{x}_N)\}$ follow a Gaussian distribution (Rasmussen and Williams, 2005). A GP is fully described by a mean function $m(\mathbf{x})$ and a covariance function $\kappa(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta})$ with hyper-parameters $\boldsymbol{\theta}$. Given a supervised learning problem with N pairs of inputs $\mathbf{x}_i$ and labels y_i, $\mathcal{D} = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbb{R}^D, y_i \in \mathbb{R}\}_{i=1,\dots,N}$, we consider a GP prior over functions which are fed to a suitable likelihood function to model the observed labels.

Denoting by $\mathbf{f} \in \mathbb{R}^N$ the realizations of the GP random variables at the N inputs $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ and assuming a zero-mean GP prior, we have that $p(\mathbf{f}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{\mathrm{xx}|\boldsymbol{\theta}})$, where $\mathbf{K}_{\mathrm{xx}|\boldsymbol{\theta}}$ is the covariance matrix obtained by evaluating $\kappa(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta})$ over all input pairs $\mathbf{x}_i, \mathbf{x}_j$ (we will drop the explicit parameterization on $\boldsymbol{\theta}$ to keep the notation uncluttered). In the Bayesian setting, given a suitable likelihood function $p(\mathbf{y}|\mathbf{f})$, the objective is to infer the posterior $p(\mathbf{f}|\mathbf{y})$ given N pairs of inputs and labels. This inference problem is analytically tractable for few cases, e.g. using a Gaussian likelihood $\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma^2\mathbf{I})$, but it involves the costly $\mathcal{O}(N^3)$ inversion of the covariance matrix $\mathbf{K}_{\mathrm{xx}}$.

Sparse GPs are a family of approximate models that address the scalability issue by introducing a set of M inducing variables $\mathbf{u} = (u_1, \dots, u_M)$ at corresponding inducing inputs $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_M\}$ such that $u = f(\mathbf{z})$ (Snelson and Ghahramani, 2005). These inducing variables are assumed to be drawn from the same GP as the original process, yielding the joint prior $p(\mathbf{f}, \mathbf{u}) = p(\mathbf{u})p(\mathbf{f}|\mathbf{u})$ with

$$\begin{aligned} p(\mathbf{u}) &= \mathcal{N}(0, \mathbf{K}_{\mathrm{zz}}) \\ p(\mathbf{f}|\mathbf{u}) &= \mathcal{N}\left(\mathbf{K}_{\mathrm{xz}}\mathbf{K}_{\mathrm{zz}}^{-1}\mathbf{u}, \mathbf{K}_{\mathrm{xx}} - \mathbf{K}_{\mathrm{xz}}\mathbf{K}_{\mathrm{zz}}^{-1}\mathbf{K}_{\mathrm{zx}}\right), \end{aligned} \tag{1}$$

where $\mathbf{K}_{\mathrm{zz}} \equiv k(\mathbf{Z}, \mathbf{Z})$, $\mathbf{K}_{\mathrm{xz}} \equiv k(\mathbf{X}, \mathbf{Z})$ and $\mathbf{K}_{\mathrm{xz}} = \mathbf{K}_{\mathrm{zx}}^T$. After introducing the inducing variables, the interest is in obtaining a posterior distribution over $\mathbf{f}$ by relying on the set of inducing variables $\mathbf{u}$ so as to avoid costly algebraic operations with $\mathbf{K}_{\mathrm{xx}} \in \mathbb{R}^{N \times N}$. A general framework to do this for any likelihood and at scale (using mini-batches) can be obtained using variational inference techniques Titsias (2009a); Hensman et al. (2013); Bonilla et al. (2019). The main innovation in Titsias (2009a) is the formulation of an approximate posterior $q(\mathbf{f}, \mathbf{u})$ within variational inference (Jordan et al., 1999) so as to develop such a framework. This variational distribution formulation has come to be known as Titsias' trick and has the form:

$$q(\mathbf{f}, \mathbf{u}) = q(\mathbf{u})p(\mathbf{f}|\mathbf{u}). \tag{2}$$

Following the variational inference approach, and using the above approximate posterior, we introduce the evidence lower bound (ELBO),

$$\log p(\mathbf{y}) \geq -\mathrm{KL}\left[q(\mathbf{u}) \parallel p(\mathbf{u})\right] + \mathbb{E}_{q(\mathbf{f},\mathbf{u})} \log p(\mathbf{y} \mid \mathbf{f}), \tag{3}$$

where the Kullback-Leibler divergence (KL) term only involves M-dimensional distributions, as the conditional prior (Eq. 1) is also used in the approximate posterior (Eq. 2), which results in the KL involving N-dimensional distributions vanish. The second term in the expression above is usually referred to as the expected log likelihood (ELL) and, for factorized conditional likelihoods, it can be computed efficiently using quadrature or Monte Carlo

(MC) sampling (see, e.g., Hensman et al., 2015a). Thus, posterior estimation under this framework involves constraining $q(\mathbf{u})$ to have a parametric form (usually a Gaussian) and finding its parameters so as to optimize the ELBO above. This optimization can be carried out using stochastic-gradient methods operating on mini-batches yielding a time complexity of $O(M^3)$.

1.1 MCMC for Variationally Sparse GPs

An alternative treatment of the inducing variables under the variational framework described above is to avoid constraining $q(\mathbf{u})$ to having any parametric form or admitting simplistic factorizing assumptions. As shown by Hensman et al. (2015b), this can be, in fact, achieved by finding the optimal (unconstrained) distribution $q(\mathbf{u})$ that maximizes the ELBO in Eq. 3 and sampling from it using techniques such as Markov chain Monte Carlo (MCMC). This optimal distribution can be shown to have the form

$$\log q(\mathbf{u}) = \mathbb{E}_{p(\mathbf{f}|\mathbf{u})} \log p(\mathbf{y}|\mathbf{f}) + \log p(\mathbf{u}) + C, \tag{4}$$

where C is an unknown normalizing constant. This expression makes it apparent that sparse variational GPs can be seen as GP models with a Gaussian prior over the inducing variables and a likelihood which has a complicated form due to the expectation under the conditional $p(\mathbf{f}|\mathbf{u})$. This observation makes it possible to derive MCMC samplers for the posterior over $\mathbf{u}$, thus relaxing the constraint of having to deal with a fixed form approximation. The only difficulty is that the likelihood requires the computation of an expectation; however, as mentioned above, for most modeling problems where the likelihood factorizes, this expectation can be calculated as a sum of univariate integrals, for which it is easy to employ numerical quadrature. Hensman et al. (2015b) also include the sampling of the hyper-parameters $\boldsymbol{\theta}$ jointly with $\mathbf{u}$; however, in order to do this efficiently, a whitening representation is employed, whereby the inducing variables are reparameterized as $\mathbf{u} = \mathbf{L}_{zz}\boldsymbol{\nu}$, with $\mathbf{K}_{zz} = \mathbf{L}_{zz}\mathbf{L}_{zz}^\top$. The sampling scheme then amounts to sampling from the joint posterior over $\boldsymbol{\nu}, \boldsymbol{\theta}$.

The sampling scheme used by Hensman et al. (2015b) employs a more efficient method based on Hamiltonian Monte Carlo (HMC, Duane et al., 1987; Neal, 2010). Given a potential energy function defined as $U(\mathbf{u}) = -\log p(\mathbf{u}, \mathbf{y}) = -\log p(\mathbf{u}|\mathbf{y}) + C$, Hamiltonian Monte Carlo (HMC) introduces auxiliary momentum variables $\mathbf{r}$ and it generates samples from the joint distribution $p(\mathbf{u}, \mathbf{r})$ by simulation of the Hamiltonian dynamics

$$\begin{aligned} \mathrm{d}\mathbf{u} &= \mathbf{M}^{-1}\mathbf{r}\mathrm{d}t\,, \\ \mathrm{d}\mathbf{r} &= -\nabla U(\mathbf{u})\mathrm{d}t\,, \end{aligned}$$

where $\mathbf{M}$ is the so called mass matrix, followed by a Metropolis accept/reject step.

1.2 Stochastic gradient HMC for Deep models

Different from classic HMC where it is required to compute the full gradients $\nabla U(\mathbf{u}) = -\nabla \log p(\mathbf{u}|\mathbf{y})$, stochastic gradient Hamiltonian Monte Carlo (SGHMC) (Chen et al., 2014) allows to sample from the true intractable posterior by means of stochastic gradients, and without the need of Metropolis accept/reject steps, which would require access to the whole data set. By modeling the stochastic gradient noise as normally distributed $\mathcal{N}(\mathbf{0}, \mathbf{V})$, the (discretized) Hamiltonian dynamics are updated as follows

$$\begin{aligned} \mathbf{u}_{t+1} &= \mathbf{u}_{t+1} + \varepsilon\mathbf{M}^{-1}\mathbf{r}_t\,, \\ \mathbf{r}_{t+1} &= \mathbf{r}_t - \varepsilon\widetilde{\nabla U}(\mathbf{u}) - \varepsilon\mathbf{C}\mathbf{M}^{-1}\mathbf{r}_t + \mathcal{N}(0, 2\varepsilon(\mathbf{C} - \tilde{\mathbf{B}}))\,, \end{aligned}$$

where ε is the step size, $\mathbf{C}$ is a user defined friction term and $\tilde{\mathbf{B}}$ is the estimated diffusion matrix of the gradient noise; see e.g., Springenberg et al. (2016) for ideas on how to estimate these parameters.

SGHMC is the primary inference method used by Havasi et al. (2018) for obtaining samples from the posterior distribution over the latent variables. Recently, this has been approached using adversarial inference methods (Yu et al., 2019).

1.3 Other Approaches to Scalable and Bayesian GPs

It is worth mentioning that other approaches to scalable inference in GPs have been proposed, which feature the possibility to operate using mini-batches. For example, looking at the feature-space view of kernel machines,

Rahimi and Recht (2008) show how random features can be obtained for shift invariant covariance functions, like the commonly used squared exponential. These approximations are also useful for addressing the scalability of GPs and deep Gaussian processes (DGPs), as showed by Lázaro-Gredilla et al. (2010) and Cutajar et al. (2017). Similarly, the work on structured approximations of GPs Saatçi (2011) has found applications to develop a scalable framework for GPs, later developed to include the possibility to learn deep learning-based representations for the input Wilson et al. (2016).

The Gaussian process latent variable model (GPLVM) proposed by Lawrence (2005) is a popular approach to Bayesian nonlinear dimensionality reduction and its Bayesian extensions such as those develped by Titsias and Lawrence (2010) consider a prior over the inputs of a GP. Although these methods can be used for training GPs with missing or uncertain inputs, we are not aware of previous work adopting such methodologies for inducing inputs within scalable sparse GP models.

2 Discussion on the objectives: VFE vs FITC

To understand why the fully independent training conditional (FITC) objective makes sense we need to go back to the original work of Titsias (2009a,b) and the seminal work of Quiñonero-Candela and Rasmussen (2005). For this, we will consider the regression case and then we can easily generalize our reasoning to the classification case. Titsias (2009b) shows that, in the standard regression case with isotropic observation noise, his variational free energy (VFE) optimization framework yields exactly the same predictive posterior as the projected process (PP) approximation (Seeger et al., 2003), which is referred to as the deterministic training conditional (DTC) approximation in Quiñonero-Candela and Rasmussen (2005). The optimal variational posterior distribution is given by:

$$\begin{aligned} q^*(\mathbf{u} \,|\, \boldsymbol{\theta}) &= \mathcal{N}(\mathbf{u}; \mathbf{m}, \mathbf{S}), \\ \mathbf{m} &= \sigma^{-2} \mathbf{K}_{\text{zz}} \boldsymbol{\Sigma} \mathbf{K}_{\text{zx}} \mathbf{y} \\ \mathbf{S} &= \mathbf{K}_{\text{zz}} \boldsymbol{\Sigma} \mathbf{K}_{\text{zz}}, \text{ where} \\ \boldsymbol{\Sigma} &= \left(\mathbf{K}_{\text{zz}} + \sigma^{-2} \mathbf{K}_{\text{zx}} \mathbf{K}_{\text{xz}} \right)^{-1}, \end{aligned} \tag{5}$$

where σ^2 is the observation-noise variance. It is easy to show that, given a Gaussian posterior over the inducing variables with mean and covariance $\mathbf{m}$ and $\mathbf{S}$, the posterior predictive distribution at test point $\mathbf{x}_\star$ is a Gaussian with mean and variance

$$\begin{aligned} \mu_y(\mathbf{x}_\star) &= \kappa(\mathbf{x}_\star, \mathbf{Z}) \mathbf{K}_{\text{zz}}^{-1} \mathbf{m} \\ \sigma_y^2(\mathbf{x}_\star) &= \kappa(\mathbf{x}_\star, \mathbf{x}_\star) - \kappa(\mathbf{x}_\star, \mathbf{Z}) \mathbf{K}_{\text{zz}}^{-1} \kappa(\mathbf{Z}, \mathbf{x}_\star) + \kappa(\mathbf{x}_\star, \mathbf{Z}) \mathbf{K}_{\text{zz}}^{-1} \mathbf{S} \mathbf{K}_{\text{zz}}^{-1} \kappa(\mathbf{Z}, \mathbf{x}_\star). \end{aligned} \tag{6}$$

Thus, replacing Eq. 5 in Eq. 6 we obtain:

$$\begin{aligned} \mu_y(\mathbf{x}_\star) &= \sigma^{-2} \kappa(\mathbf{x}_\star, \mathbf{Z}) \boldsymbol{\Sigma} \mathbf{K}_{\text{zx}} \mathbf{y} \\ \sigma_y^2(\mathbf{x}_\star) &= \kappa(\mathbf{x}_\star, \mathbf{x}_\star) - \kappa(\mathbf{x}_\star, \mathbf{Z}) \mathbf{K}_{\text{zz}}^{-1} \kappa(\mathbf{Z}, \mathbf{x}_\star) + \kappa(\mathbf{x}_\star, \mathbf{Z}) \boldsymbol{\Sigma} \kappa(\mathbf{Z}, \mathbf{x}_\star), \end{aligned} \tag{7}$$

which indeed corresponds to the predictive distribution of the DTC/PP approximation. Despite this equivalence, as highlighted in Titsias (2009a), the main difference is that the VFE framework provides a more robust approach to hyper-parameter estimation as the resulting ELBO corresponds to a regularized marginal likelihood of the DTC approach and hence should be more robust to overfitting. Nevertheless, the DTC/PP, and consequently the VFE, predictive distribution has been shown to be less accurate than the FITC approximation (Titsias, 2009a; Quiñonero-Candela and Rasmussen, 2005; Snelson, 2007).

2.1 The FITC Approximation

The FITC approximation considers the following approximate conditional prior:

$$
\begin{aligned}
p(\mathbf{f} \,|\, \mathbf{u}) &\approx \mathcal{N}\left(\mathbf{f}; \mathbf{K}_{\text{xz}}\mathbf{K}_{\text{zz}}^{-1}\mathbf{u}, \operatorname{diag}\left(\mathbf{K}_{\text{xx}} - \mathbf{K}_{\text{xz}}\mathbf{K}_{\text{zz}}^{-1}\mathbf{K}_{\text{zx}}\right)\right) \\
&= \prod_{n=1}^{N} p(f_n \,|\, \mathbf{u}) = \prod_{n=1}^{N} \mathcal{N}(f_n; \tilde{\mu}_n, \tilde{\sigma}_n^2), \text{ with} \\
\tilde{\mu}_n &= \kappa(\mathbf{x}_n, \mathbf{Z})\mathbf{K}_{\text{zz}}^{-1}\mathbf{u} \qquad (8) \\
\tilde{\sigma}_n^2 &= \kappa(\mathbf{x}_n, \mathbf{x}_n) - \kappa(\mathbf{x}_n, \mathbf{Z})\mathbf{K}_{\text{zz}}^{-1}\kappa(\mathbf{Z}, \mathbf{x}_n). \qquad (9)
\end{aligned}
$$

As we shall see later, is this factorization assumption in the conditional prior that will yield a decomposable objective amenable to stochastic gradient techniques. For now, consider the posterior predictive distribution under the FITC approximation[1]

$$
\begin{aligned}
\mu_{\text{FITC}}(\mathbf{x}_\star) &= \kappa(\mathbf{x}_\star, \mathbf{Z})\boldsymbol{\Sigma}_{\text{FITC}}\mathbf{K}_{\text{zx}}\boldsymbol{\Lambda}^{-1}\mathbf{y} \\
\sigma^2_{\text{FITC}}(\mathbf{x}_\star) &= \kappa(\mathbf{x}_\star, \mathbf{x}_\star) - \kappa(\mathbf{x}_\star, \mathbf{Z})\mathbf{K}_{\text{zz}}^{-1}\kappa(\mathbf{Z}, \mathbf{x}_\star) + \kappa(\mathbf{x}_\star, \mathbf{Z})\boldsymbol{\Sigma}_{\text{FITC}}\kappa(\mathbf{Z}, \mathbf{x}_\star), \text{ where} \\
\boldsymbol{\Lambda} &= \operatorname{diag}(\mathbf{K}_{\text{xx}} - \mathbf{K}_{\text{xz}}\mathbf{K}_{\text{zz}}^{-1}\mathbf{K}_{\text{zx}} + \sigma^2\mathbf{I}) \text{ and} \\
\boldsymbol{\Sigma}_{\text{FITC}} &= (\mathbf{K}_{\text{zz}} + \mathbf{K}_{\text{zx}}\boldsymbol{\Lambda}^{-1}\mathbf{K}_{\text{xz}})^{-1}.
\end{aligned} \qquad (10)
$$

We now see why FITC's predictive distribution above is more accurate than VFE's in Eq. 7, as we can obtain FITC's by replacing $\sigma^2\mathbf{I}$ in VFE's solution with $\operatorname{diag}(\mathbf{K}_{\text{xx}} - \mathbf{K}_{\text{xz}}\mathbf{K}_{\text{zz}}^{-1}\mathbf{K}_{\text{zx}}) + \sigma^2\mathbf{I}$. Effectively, as described in Quiñonero-Candela and Rasmussen (2005), VFE's solution (which is the same as DTC's) can be understood as considering a deterministic conditional prior $p(\mathbf{f} \,|\, \mathbf{u})$, i.e. with zero variance.

2.2 Stochastic Updates Using the FITC Approximation

Now we can understand why the log of the expectation can provide more accurate results than the expectation of the log. Basically in the former we are using the FITC approximation while in the later we are using the VFE/DTC/PP approximation. It is easy to show that when using the FITC approximation, one can obtain a decomposable objective function that can be implemented at large scale using stochastic gradient techniques. Here we focus only on the expectation of the conditional likelihood (which is the crucial term) and in the regression setting for simplicity but the extension to the classification case (e.g. using quadrature) is straightforward.

$$
\begin{aligned}
\log p(\mathbf{y}, \mathbf{u} \,|\, \boldsymbol{\theta}) &= \log \mathbb{E}_{p(\mathbf{f} \,|\, \mathbf{u}, \boldsymbol{\theta})}\left[p(\mathbf{y} \,|\, \mathbf{f})\right] \\
&= \log \int_{\mathbf{f}} p(\mathbf{f} \,|\, \mathbf{u}, \boldsymbol{\theta}) p(\mathbf{y} \,|\, \mathbf{f}) \, \mathrm{d}\mathbf{f} \quad \text{will drop } \boldsymbol{\theta} \text{ for simplicity from now on} \\
&= \log \int_{f_1, \ldots, f_N} \prod_{n=1}^{N} p(y_n \,|\, f_n) p(f_n \,|\, \mathbf{u}) \, \mathrm{d}\mathbf{f} \\
&= \log \prod_{n=1}^{N} \int_{f_n} \mathcal{N}(y_n; f_n, \sigma^2) \mathcal{N}(f_n; \tilde{\mu}_n, \tilde{\sigma}_n^2) \, \mathrm{d}f_n \\
&= \log \prod_{n=1}^{N} p(y_n \,|\, \mathbf{u}) \\
&= \sum_{n=1}^{N} \log \mathcal{N}(y_n; \tilde{\mu}_n, \tilde{\sigma}_n^2 + \sigma^2), \qquad (11)
\end{aligned}
$$

where $\tilde{\mu}_n, \tilde{\sigma}_n^2$ are given by Eq. 8 and Eq. 9.

[1] Which is, in fact, the same as in the sparse Gaussian process (SPGP) framework of Snelson and Ghahramani (2007).

Binary classification. Similar results can be derived for binary classification with Bernoulli likelihood and response function $\lambda(f)$:

$$\log p(\mathbf{y}, \mathbf{u} \,|\, \boldsymbol{\theta}) = \log \mathbb{E}_{p(\mathbf{f} \,|\, \mathbf{u}, \boldsymbol{\theta})} \left[p(\mathbf{y} \,|\, \mathbf{f}) \right] = \log \prod_{n=1}^{N} \int_{f_n} \mathcal{N}(f_n; \tilde{\mu}_n, \tilde{\sigma}_n^2) \mathrm{Bern}(y_n; \lambda(f_n)) \, \mathrm{d}f_n. \tag{12}$$

When the response function is the cdf of a standard Normal distribution, i.e., $\lambda(f_n) = \Phi(f_n) \stackrel{\text{def}}{=} \int_{-\infty}^{f_n} \mathcal{N}(f_n; 0, 1) df_n$, which is also known as the probit regression model, the expectation above can be computed analytically to obtain:

$$\log p(\mathbf{y}, \mathbf{u} \,|\, \boldsymbol{\theta}) = \sum_{n=1}^{N} \log \mathrm{Bern}(y_n; \Phi(\tilde{\mu}_n / \sqrt{1 + \tilde{\sigma}_n^2})). \tag{13}$$

For other response functions the expectation in Eq. 12 can be estimated using quadrature.

2.3 An heteroskedastic version of the Gaussian likelihood

As Titsias (2009b) discussed in Appendix C, the FITC approximation corresponds to a GP regression with heteroskedastic noise variance

$$p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2 \mathbf{I} + \mathrm{diag}[\mathbf{K_{xx}} - \mathbf{K_{xz}} \mathbf{K_{zz}} \mathbf{K_{zx}}]). \tag{14}$$

If we apply this augmented likelihood to the variational expectations term, we get

$$\mathbb{E}_{q(\mathbf{f})} \log p(\mathbf{y}|\mathbf{f}, \sigma^2, \boldsymbol{\theta}) = -\frac{1}{2} \sum_{j=1}^{n} \left(\log 2\pi(\sigma^2 + \tilde{\sigma}_j^2) + \frac{(y_j - \tilde{\mu}_j)^2 + \tilde{\sigma}_j^2}{\sigma^2 + \tilde{\sigma}_j^2} \right). \tag{15}$$

Since Titsias (2009b) considers this VFE formulation, we also compare with it.

2.4 Concluding remarks

The main reasoning in Titsias (2009a)'s work behind the better performance of VFE, despite providing a less accurate predictive posterior than FITC's, was that hyper-parameter estimation was more robust due to the use of the variational objective (which provided an extra regularization term). Now, we have a better way to do inference on hyper-parameters and inducing inputs by placing priors on those and by carrying out free-form inference upon them with SGHMC.

3 Extension to Deep Gaussian Processes

In this section, we derive the mathematical basis for of Bayesian treatment of inducing inputs in a DGP setting (Damianou and Lawrence, 2013). We assume a deep Gaussian process prior $f^L \circ f^{L-1} \circ \cdots f^1$, where each f^l is a GP. For notational brevity, we use $\boldsymbol{\theta}^l$ as both kernel hyper-parameters and inducing inputs of the l-th layer, and $\mathbf{f}^0$ as the input vector $\mathbf{X}$. Then we can write down the joint distribution over visible and latent variables (omitting the dependency on $\mathbf{X}$ for clarity) as

$$p\left(\mathbf{y}, \left\{\mathbf{f}^l, \mathbf{u}^l, \boldsymbol{\theta}^l\right\}_{l=1}^{L}\right) = p\left(\mathbf{y} \,\middle|\, \mathbf{f}^L\right) \prod_{l=1}^{L} p\left(\mathbf{f}^l \,\middle|\, \mathbf{u}^l, \mathbf{f}^{l-1}, \boldsymbol{\theta}^l\right) p\left(\mathbf{u}^l \,\middle|\, \boldsymbol{\theta}^l\right) p\left(\boldsymbol{\theta}^l\right). \tag{16}$$

Our goal is to estimate the posterior,

$$\begin{aligned} &\log \widetilde{p}\left(\left\{\mathbf{u}^l, \boldsymbol{\theta}^l\right\}_{l=1}^{L} \,\middle|\, \mathbf{y}\right) = \\ &= \log \mathbb{E}_{p(\{\mathbf{f}^l\} \,|\, \{\mathbf{u}^l, \boldsymbol{\theta}^l\})} p\left(\mathbf{y} \,\middle|\, \mathbf{f}^L\right) + \sum_{l=1}^{L} \left(\log p\left(\mathbf{u}^l \,\middle|\, \boldsymbol{\theta}^l\right) + \log p\left(\boldsymbol{\theta}^l\right) \right) - \log C. \end{aligned} \tag{17}$$

Here C is a normalizing constant, after integrating out $\{\mathbf{f}^l, \mathbf{u}^l, \boldsymbol{\theta}^l\}_{l=1}^L$ from the joint. While the distribution $\widetilde{p}$ is intractable, we have obtained the form of its (un-normalized) log joint, from which we can sample using HMC methods. However, Eq. 17 is not immediately computable owing to the intractable expectation term. More calculations reveal that we can, nevertheless, obtain estimates of this expectation term with Monte Carlo sampling

$$
\begin{aligned}
&\log \mathbb{E}_{p(\{\mathbf{f}^l\} \mid \{\mathbf{u}^l, \boldsymbol{\theta}^l\})} p\left(\mathbf{y} \,\middle|\, \mathbf{f}^L\right) \approx \\
&\quad \approx \log \mathbb{E}_{p\left(\{\mathbf{f}^l\}_{l=2}^L \,\middle|\, \widetilde{\mathbf{f}}^1, \{\mathbf{u}^l, \boldsymbol{\theta}^l\}_{l=2}^L\right)} p\left(\mathbf{y} \,\middle|\, \mathbf{f}^L\right), \quad \widetilde{\mathbf{f}}^1 \sim p\left(\mathbf{f}^1 \mid \mathbf{u}^1, \boldsymbol{\theta}^1, \mathbf{f}^0\right), \\
&\quad \approx \log \mathbb{E}_{p\left(\{\mathbf{f}^l\}_{l=3}^L \,\middle|\, \widetilde{\mathbf{f}}^2, \{\mathbf{u}^l, \boldsymbol{\theta}^l\}_{l=3}^L\right)} p\left(\mathbf{y} \,\middle|\, \mathbf{f}^L\right), \quad \widetilde{\mathbf{f}}^2 \sim p\left(\mathbf{f}^2 \,\middle|\, \mathbf{u}^2, \boldsymbol{\theta}^2, \widetilde{\mathbf{f}}^1\right), \\
&\quad \approx \dots \\
&\quad \approx \log \mathbb{E}_{p\left(\mathbf{f}^L \,\middle|\, \widetilde{\mathbf{f}}^{L-1}, \mathbf{u}^L, \boldsymbol{\theta}^L\right)} p\left(\mathbf{y} \,\middle|\, \mathbf{f}^L\right), \quad \widetilde{\mathbf{f}}^{L-1} \sim p\left(\mathbf{f}^{L-1} \,\middle|\, \mathbf{u}^{L-1}, \boldsymbol{\theta}^{L-1}, \widetilde{\mathbf{f}}^{L-2}\right), \\
&\quad = \sum_{n=1}^N \log \mathbb{E}_{p\left(f_n^L \mid \widetilde{f}_n^{L-1}, \mathbf{u}^L, \boldsymbol{\theta}^L\right)} p\left(y_n \mid f_n^L\right)
\end{aligned} \tag{18}
$$

Because of the layer-wise factorization of the joint likelihood (Eq. 16), each step of the approximation is unbiased. While it is possible to approximate the last-layer expectation with a Monte Carlo sample $\widetilde{f}_j^L$, the expectation is tractable when $y_j \mid f_j^L$ is a Gaussian or a Bernoulli distribution with a probit regression model, or is computable with one-dimensional quadrature (Hensman et al., 2015b).

4 Additional Results

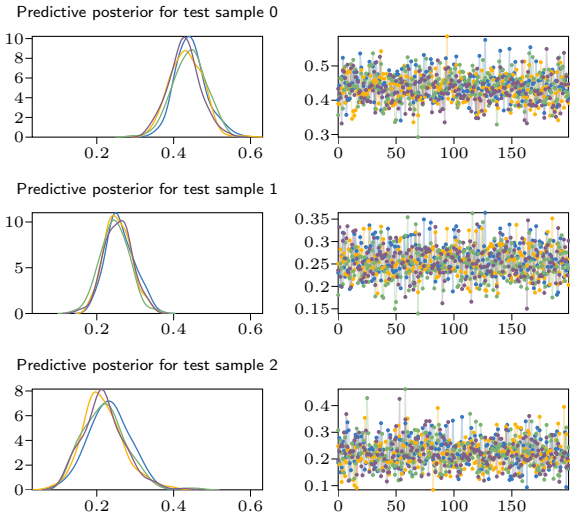

Figure 1: Traces for three test points on the Airline dataset (4 chains/200 samples).

Table 1: Datasets used, including number of datapoints and their dimensionality.

name	n.	d-in	d-out
BOSTON	506	13	1
CONCRETE	1,030	8	1
ENERGY	768	8	2
KIN8NM	8,192	8	1
NAVAL	11,934	16	2
POWERPLANT	9,568	4	1
PROTEIN	45,730	9	1
YACHT	308	6	1
AIRLINE	5,934,530	8	2
HIGGS	11,000,000	28	2

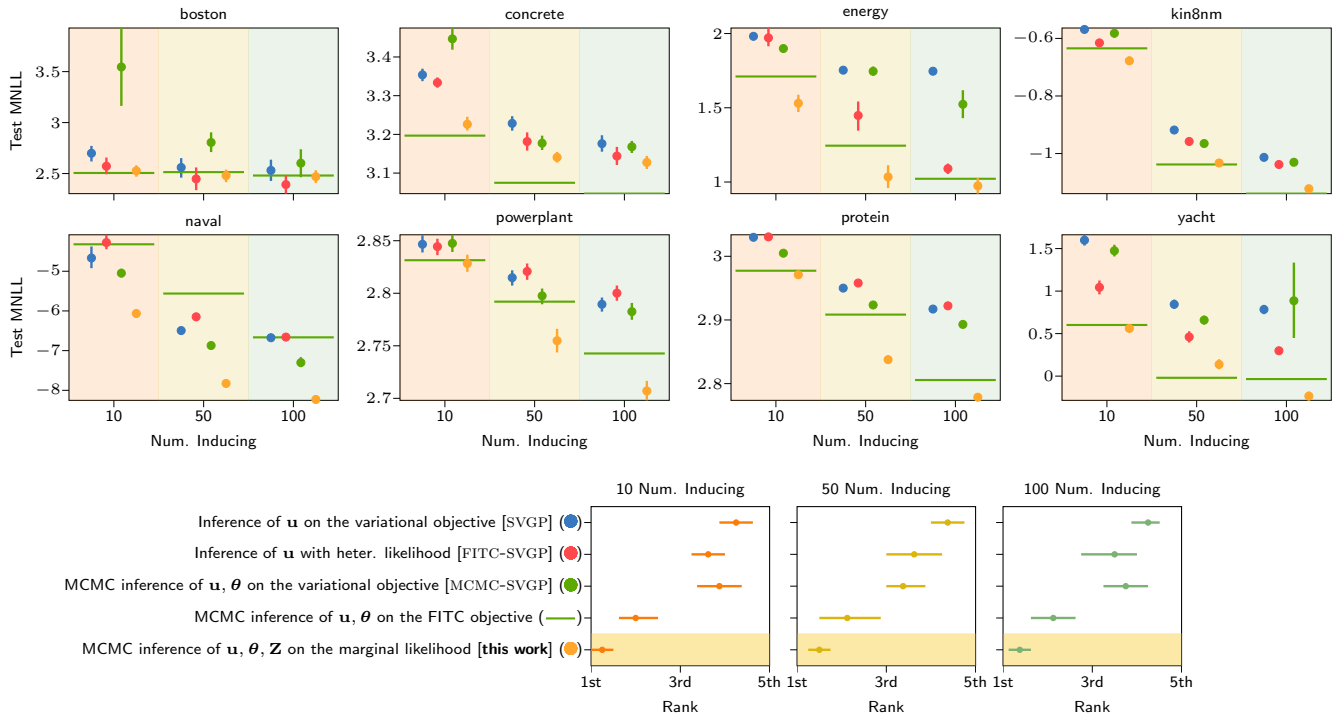

Figure 2: Empirical analysis of different choices of objectives for optimization and sampling.

Table 2: Tabular version of Figure 7 in the main paper.

DATASET NAME	TEST MNLL BOSTON	CONCRETE	ENERGY	KIN8NM	NAVAL	POWERPLANT	PROTEIN	YACHT
BSGP 1	2.47 (0.16)	3.12 (0.04)	0.97 (0.13)	−1.12 (0.01)	−8.22 (0.04)	2.71 (0.02)	2.78 (0.01)	−0.23 (0.13)
BSGP 2	2.47 (0.15)	3.04 (0.05)	0.95 (0.16)	−1.40 (0.01)	−8.23 (0.04)	2.67 (0.02)	2.63 (0.02)	−0.72 (0.15)
BSGP 3	2.47 (0.14)	2.96 (0.10)	0.95 (0.15)	−1.41 (0.01)	−8.02 (0.04)	2.66 (0.03)	2.57 (0.03)	−0.83 (0.10)
BSGP 4	2.48 (0.14)	2.97 (0.06)	0.92 (0.14)	−1.43 (0.02)	−8.03 (0.05)	2.65 (0.05)	2.50 (0.03)	−0.76 (0.13)
BSGP 5	2.48 (0.12)	2.91 (0.08)	0.75 (0.30)	−1.43 (0.01)	−8.09 (0.05)	2.65 (0.03)	2.43 (0.03)	−0.74 (0.08)
IPVI GP 1	2.84 (0.36)	3.19 (0.11)	1.27 (0.07)	−1.12 (0.02)	−5.96 (0.89)	2.79 (0.03)	2.81 (0.02)	1.21 (1.50)
IPVI GP 2	2.73 (0.35)	3.13 (0.11)	1.31 (0.28)	−1.34 (0.02)	−4.98 (0.48)	2.76 (0.07)	2.65 (0.02)	0.74 (1.13)
IPVI GP 3	2.61 (0.25)	3.08 (0.13)	1.21 (0.12)	−1.33 (0.03)	−4.86 (0.23)	2.72 (0.06)	2.74 (0.05)	1.05 (1.77)
IPVI GP 4	2.64 (0.44)	3.11 (0.18)	1.19 (0.25)	−1.33 (0.01)	−4.94 (0.20)	2.76 (0.02)	2.79 (0.01)	2.47 (2.34)
IPVI GP 5	2.51 (0.20)	3.08 (0.17)	1.15 (0.22)	−1.29 (0.02)	−5.09 (0.49)	2.72 (0.04)	2.80 (0.01)	2.84 (3.64)
SGHMC GP 1	2.82 (0.33)	3.13 (0.09)	1.08 (0.28)	−1.08 (0.01)	−6.23 (0.14)	2.76 (0.05)	2.81 (0.01)	−0.11 (0.28)
SGHMC GP 2	2.77 (0.37)	2.99 (0.07)	0.91 (0.15)	−1.32 (0.01)	−6.57 (0.11)	2.72 (0.04)	2.71 (0.02)	−0.52 (0.14)
SGHMC GP 3	2.78 (0.28)	3.02 (0.16)	0.91 (0.14)	−1.37 (0.02)	−6.56 (0.09)	2.68 (0.02)	2.66 (0.03)	−0.57 (0.19)
SGHMC GP 4	2.75 (0.34)	2.98 (0.13)	0.69 (0.22)	−1.38 (0.02)	−6.42 (0.08)	2.67 (0.04)	2.62 (0.02)	−0.69 (0.12)
SGHMC GP 5	3.75 (1.91)	3.11 (0.21)	1.00 (0.42)	−1.39 (0.02)	−6.55 (0.09)	2.65 (0.04)	2.59 (0.02)	−0.53 (0.18)
SVGP 1	2.53 (0.25)	3.18 (0.05)	1.75 (0.06)	−1.01 (0.01)	−6.67 (0.09)	2.79 (0.02)	2.92 (0.01)	0.78 (0.13)

Table 3: Normalized RMSE corresponding to results of Figure 7 in the main paper.

DATASET NAME	TEST ERROR BOSTON	CONCRETE	ENERGY	KIN8NM	NAVAL	POWERPLANT	PROTEIN	YACHT
BSGP 1	0.36 (0.07)	0.40 (0.03)	0.13 (0.01)	0.31 (0.01)	0.02 (0.00)	0.23 (0.00)	0.72 (0.00)	0.04 (0.01)
BSGP 2	0.37 (0.07)	0.36 (0.02)	0.13 (0.01)	0.24 (0.00)	0.01 (0.00)	0.22 (0.00)	0.69 (0.00)	0.03 (0.01)
BSGP 3	0.37 (0.07)	0.32 (0.03)	0.13 (0.01)	0.24 (0.01)	0.01 (0.00)	0.22 (0.00)	0.68 (0.00)	0.03 (0.01)
BSGP 4	0.37 (0.07)	0.33 (0.02)	0.13 (0.01)	0.24 (0.01)	0.01 (0.00)	0.21 (0.00)	0.67 (0.01)	0.03 (0.01)
BSGP 5	0.37 (0.07)	0.32 (0.03)	0.12 (0.03)	0.23 (0.00)	0.01 (0.00)	0.21 (0.00)	0.65 (0.01)	0.03 (0.01)
IPVI GP 1	0.34 (0.04)	0.34 (0.02)	0.13 (0.01)	0.31 (0.01)	0.16 (0.17)	0.23 (0.00)	0.72 (0.01)	0.04 (0.02)
IPVI GP 2	0.35 (0.06)	0.32 (0.02)	0.13 (0.01)	0.25 (0.01)	0.62 (0.22)	0.22 (0.01)	0.68 (0.01)	0.03 (0.02)
IPVI GP 3	0.34 (0.05)	0.30 (0.03)	0.13 (0.01)	0.25 (0.01)	0.65 (0.09)	0.22 (0.01)	0.65 (0.01)	0.03 (0.02)
IPVI GP 4	0.33 (0.06)	0.31 (0.03)	0.12 (0.03)	0.25 (0.00)	0.70 (0.01)	0.22 (0.01)	0.65 (0.01)	0.04 (0.03)
IPVI GP 5	0.32 (0.04)	0.30 (0.04)	0.11 (0.04)	0.26 (0.01)	0.62 (0.22)	0.21 (0.01)	0.65 (0.01)	0.03 (0.01)
SGHMC GP 1	0.36 (0.08)	0.38 (0.03)	0.13 (0.01)	0.34 (0.01)	0.15 (0.13)	0.23 (0.00)	0.75 (0.01)	0.04 (0.01)
SGHMC GP 2	0.37 (0.07)	0.35 (0.03)	0.13 (0.01)	0.26 (0.01)	0.02 (0.01)	0.23 (0.00)	0.72 (0.01)	0.03 (0.01)
SGHMC GP 3	0.38 (0.08)	0.33 (0.03)	0.13 (0.01)	0.25 (0.01)	0.02 (0.00)	0.22 (0.00)	0.71 (0.01)	0.03 (0.01)
SGHMC GP 4	0.35 (0.09)	0.31 (0.02)	0.09 (0.04)	0.25 (0.01)	0.02 (0.00)	0.22 (0.00)	0.70 (0.01)	0.02 (0.01)
SGHMC GP 5	0.39 (0.07)	0.31 (0.02)	0.13 (0.01)	0.24 (0.01)	0.02 (0.00)	0.22 (0.00)	0.69 (0.00)	0.03 (0.01)
SVGP 1	0.33 (0.05)	0.35 (0.02)	0.14 (0.01)	0.32 (0.00)	0.03 (0.01)	0.23 (0.00)	0.73 (0.00)	0.04 (0.01)

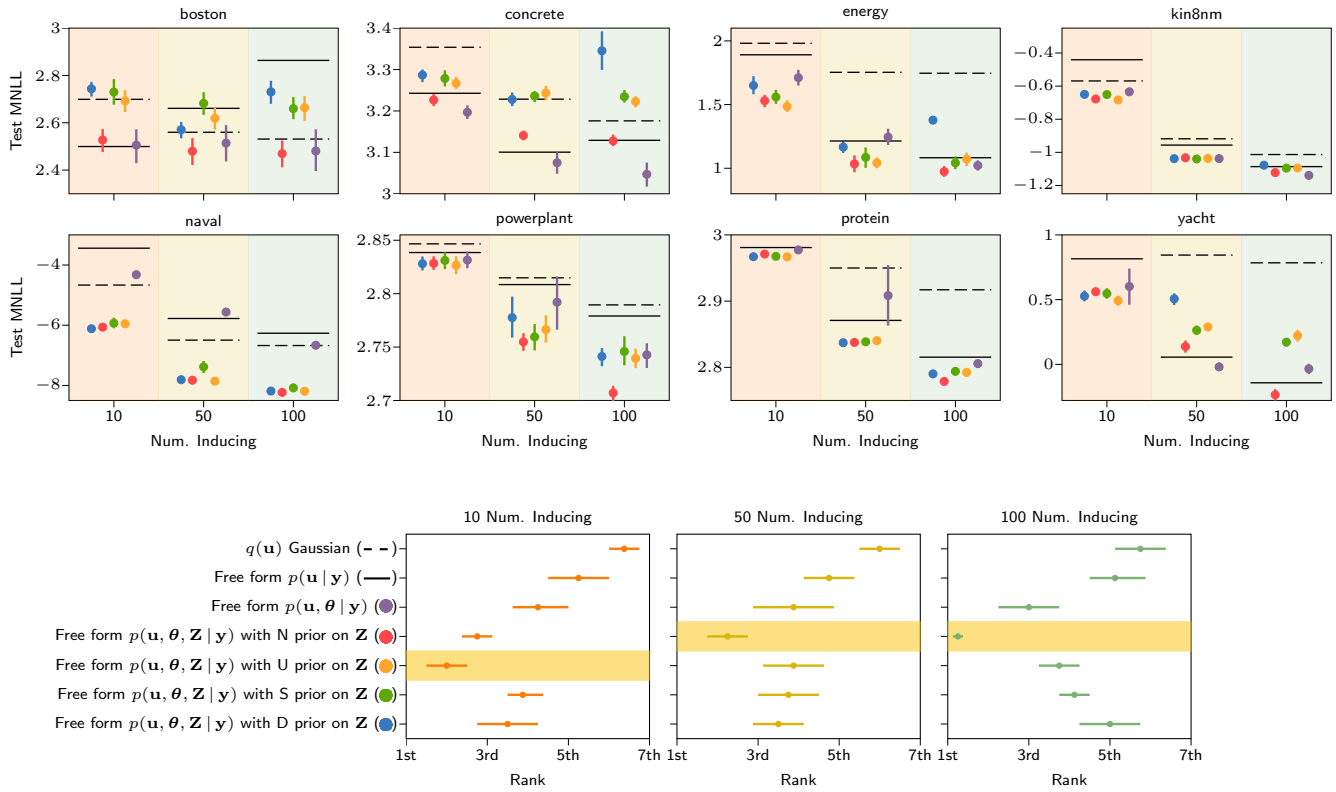

Figure 3: Ablation study on the Test MNLL based on the UCI benchmark for different number of inducing variables and for determinantal point process prior (●), normal prior (●), Strauss process prior (●) and uniform prior (●). These are compared with (●), corresponding to the case of inducing positions optimized and inducing variables and covariance hyper-parameters sampled, with (—) where only inducing variables are inferred, while the rest is optimized (similarly to SGHMC-DGP). Finally (- -) is the classic SVGP, where everything is optimized.

Bibliography

Bonilla, E. V., Krauth, K., and Dezfouli, A. (2019). Generic inference in latent Gaussian process models. *Journal of Machine Learning Research*, 20(117):1–63.

Chen, T., Fox, E., and Guestrin, C. (2014). Stochastic Gradient Hamiltonian Monte Carlo. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 1683–1691, Bejing, China. PMLR.

Cutajar, K., Bonilla, E. V., Michiardi, P., and Filippone, M. (2017). Random feature expansions for deep Gaussian processes. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 884–893, International Convention Centre, Sydney, Australia. PMLR.

Damianou, A. C. and Lawrence, N. D. (2013). Deep Gaussian Processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2013, Scottsdale, AZ, USA, April 29 - May 1, 2013*, volume 31 of *JMLR Proceedings*, pages 207–215. JMLR.org.

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

Havasi, M., Hernández-Lobato, J. M., and Murillo-Fuentes, J. J. (2018). Inference in Deep Gaussian Processes using Stochastic Gradient Hamiltonian Monte Carlo. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 7506–7516. Curran Associates, Inc.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, page 282–290, Arlington, Virginia, USA. AUAI Press.

Hensman, J., Matthews, A., and Ghahramani, Z. (2015a). Scalable Variational Gaussian Process Classification. In Lebanon, G. and Vishwanathan, S. V. N., editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 351–360, San Diego, California, USA. PMLR.

Hensman, J., Matthews, A. G., Filippone, M., and Ghahramani, Z. (2015b). MCMC for Variationally Sparse Gaussian Processes. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 1648–1656. Curran Associates, Inc.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233.

Lawrence, N. (2005). Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models. *Journal of Machine Learning Research*, 6:1783–1816.

Lázaro-Gredilla, M., Quinonero-Candela, J., Rasmussen, C. E., and Figueiras-Vidal, A. R. (2010). Sparse Spectrum Gaussian Process Regression. *Journal of Machine Learning Research*, 11:1865–1881.

Neal, R. M. (2010). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 54:113–162.

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

Rahimi, A. and Recht, B. (2008). Random Features for Large-Scale Kernel Machines. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc.

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

Saatçi, Y. (2011). *Scalable Inference for Structured Gaussian Process Models*. PhD thesis, University of Cambridge.

Seeger, M., Williams, C. K., and Lawrence, N. D. (2003). Fast forward selection to speed up sparse Gaussian process regression. In *Artificial Intelligence and Statistics*.

Snelson, E. and Ghahramani, Z. (2005). Sparse Gaussian Processes using Pseudo-inputs. In *NIPS*.

Snelson, E. and Ghahramani, Z. (2007). Local and global sparse Gaussian process approximations. In Meila, M. and Shen, X., editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, AISTATS 2007, San Juan, Puerto Rico, March 21-24, 2007*, volume 2 of *JMLR Proceedings*, pages 524–531. JMLR.org.

Snelson, E. L. (2007). *Flexible and efficient Gaussian process models for machine learning*. PhD thesis, UCL (University College London).

Springenberg, J. T., Klein, A., Falkner, S., and Hutter, F. (2016). Bayesian Optimization with Robust Bayesian Neural Networks. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 4134–4142. Curran Associates, Inc.

Titsias, M. and Lawrence, N. D. (2010). Bayesian gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 844–851.

Titsias, M. K. (2009a). Variational Learning of Inducing Variables in Sparse Gaussian Processes. In Dyk, D. A. and Welling, M., editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009*, volume 5 of *JMLR Proceedings*, pages 567–574. JMLR.org.

Titsias, M. K. (2009b). Variational model selection for sparse Gaussian process regression. *Report, University of Manchester, UK*.

Wilson, A. G., Hu, Z., Salakhutdinov, R. R., and Xing, E. P. (2016). Stochastic Variational Deep Kernel Learning. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 2586–2594. Curran Associates, Inc.

Yu, H., Chen, Y., Low, B. K. H., Jaillet, P., and Dai, Z. (2019). Implicit Posterior Variational Inference for Deep Gaussian Processes. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 14475–14486. Curran Associates, Inc.