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

Simone Rossi
EURECOM (France)

Markus Heinonen
Aalto University (Finland)

Edwin V. Bonilla
CSIRO's Data61 (Australia)

Zheyang Shen
Aalto University (Finland)

Maurizio Filippone
EURECOM (France)

Abstract

Variational inference techniques based on inducing variables provide an elegant framework for scalable posterior estimation in Gaussian process (GP) models. Besides enabling scalability, one of their main advantages over sparse approximations using direct marginal likelihood maximization is that they provide a robust alternative for point estimation of the inducing inputs, i.e. the location of the inducing variables. In this work we challenge the common wisdom that optimizing the inducing inputs in the variational framework yields optimal performance. We show that, by revisiting old model approximations such as the fully-independent training conditionals endowed with powerful sampling-based inference methods, treating both inducing locations and GP hyper-parameters in a Bayesian way can improve performance significantly. Based on stochastic gradient Hamiltonian Monte Carlo, we develop a fully Bayesian approach to scalable GP and deep GP models, and demonstrate its state-of-the-art performance through an extensive experimental campaign across several regression and classification problems.

1 INTRODUCTION

Bayesian kernel machines based on Gaussian processes (GPs) combine the modeling flexibility of kernel methods with the ability to carry out sound quantification of uncertainty (Rasmussen and Williams, 2006). Modeling and inference with GPs have evolved considerably over the last few years with key contributions in the direction of scalability to virtually any number of datapoints and generality within automatic differentiation frameworks (Matthews et al., 2017; Krauth et al., 2017). This has been possible thanks to the combination of stochastic variational inference techniques with representations based on inducing variables (Titsias, 2009a; Lázaro-Gredilla and Figueiras-Vidal, 2009; Hensman et al., 2013), random features (Rahimi and Recht, 2008; Cutajar et al., 2017; Gal and Ghahramani, 2016), and structured approximations (Wilson and Nickisch, 2015; Wilson et al., 2016b). These advancements have now made GPs attractive to a variety of applications and likelihoods (Matthews et al., 2017; van der Wilk et al., 2017; Bonilla et al., 2019).

In this work, we focus on the variationally sparse GP framework originally formulated by Titsias (2009a) and later developed by Hensman et al. (2013, 2015a) to scale up to large datasets via stochastic optimization. In these formulations, the GP prior is augmented with inducing variables (drawn from the same prior) and their posterior is approximated and estimated via variational inference. In contrast, the location of the inducing variables, which we refer to as the inducing inputs, are simply optimized along with covariance hyper-parameters. In line with earlier evidence that Bayesian treatments of GPs are beneficial (Neal, 1997; Barber and Williams, 1997; Murray and Adams, 2010; Filippone and Girolami, 2014), posterior inference of the inducing variables *jointly* with covariance hyper-parameters has been shown to improve performance (Hensman et al., 2015b).

Despite these significant insights with regards to the benefits of full Bayesian inference over latent variables in GP models, the common practice is to optimize the inducing inputs, even in very recent GP developments (Havasi et al., 2018; Shi et al., 2020; Giraldo and Álvarez, 2019). In fact, the original work of Tit-

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

Table 1: A summary of previous works on inference methods for GPs. $\boldsymbol{\theta}, \mathbf{u}, \mathbf{Z}$ refer to the GP hyper-parameters, inducing variables and inducing inputs, respectively. (✗) indicates that variables are optimized.

	Model	Inference $\boldsymbol{\theta}$	Inference $\mathbf{u}$	Inference $\mathbf{Z}$	Reference
	MCMC-GP	MCMC	-	-	Neal (1997); Barber and Williams (1997)
	SVGP	✗	VB	✗	Hensman et al. (2015a)
	FITC-SVGP	✗	VB (heterosk.)	✗	Titsias (2009a)
	SGHMC-DGP	✗	MCMC	✗	Havasi et al. (2018)
	IPVI-DGP	✗	IP	✗	Yu et al. (2019)
	MCMC-SVGP	MCMC	MCMC	✗	Hensman et al. (2015b)
	BSGP	MCMC	MCMC	MCMC	This work

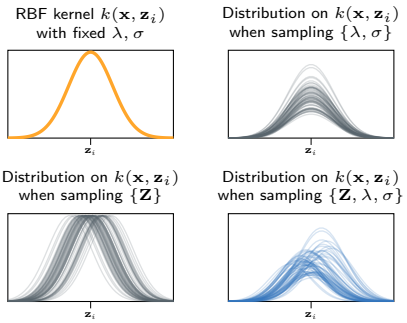

Figure 1: Representation of the induced posterior distribution on the covariance function at location $\mathbf{x}$.

sias (2009a) advocates for a treatment of the inducing inputs as variational parameters to avoid overfitting. Furthermore, later work concludes that point estimation of the inducing inputs through optimization of the variational objective is an 'optimal' treatment (Hensman et al., 2015b, §3). As we will see in § 2.1, the justification for inducing-input optimization in Hensman et al. (2015b) relies on being able to optimize both the prior and the posterior, and therefore, contradicts the fundamental principles of Bayesian inference. We summarize previous works on inference methods for GPs in Table 1, which we will use for comparison in our experiments.

Thus, we revisit the role of the inducing inputs in GP models and their treatment as variational parameters or even hyper-parameters. Given their potential high dimensionality and that the typical number of inducing variables goes beyond hundreds/thousands (Shi et al., 2020), we argue that they should be treated simply as model variables and, therefore, having priors and carrying out efficient posterior inference over them is an important—although challenging—problem. An illustration of the richer modeling capabilities offered by treating inducing inputs in a Bayesian fashion is given in Fig. 1.

Contributions. Firstly, we challenge the common wisdom that optimizing the inducing inputs in the variational framework yields optimal performance. We show that, by revisiting old model approximations such as the fully independent training conditionals (FITC; see Quiñonero-Candela and Rasmussen, 2005) endowed with powerful sampling-based inference methods, treating both inducing locations and GP hyper-parameters in a Bayesian way can improve performance significantly. We describe the conceptual justification and the mathematical details of our general formulation in § 2 and § 3. We then demonstrate that our approach yields state-of-the-art performance across a wide range of competitive benchmark methods, large-scale datasets and a variety of GP and deep GP models (§ 4).

2 BAYESIAN SPARSE GAUSSIAN PROCESSES

We are interested in supervised learning problems with N input-label training pairs $\{\mathbf{X}, \mathbf{y}\} \overset{\text{def}}{=} \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where we consider a conditional likelihood $p(\mathbf{y} \,|\, \mathbf{f})$ and $\mathbf{f}$ is drawn from a zero-mean GP prior with covariance function $k(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta})$ with hyper-parameters $\boldsymbol{\theta}$. Thus, 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 $N \times N$ covariance matrix obtained by evaluating $k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta})$ over all input pairs $\{\mathbf{x}_i, \mathbf{x}_j\}$. Inference in these types of models generally involves the costly $\mathcal{O}(N^3)$ operations to compute the inverse and log-determinant of the covariance matrix $\mathbf{K}_{\mathrm{xx}|\boldsymbol{\theta}}$.

Full joint distribution of sparse approximations. Sparse GPs are a family of approximate models that address the scalability issue by introducing a set of M inducing variables $\mathbf{u} = (u_1, \ldots, u_M)$ at corresponding inducing inputs $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_M\}$ such that $u_i = f(\mathbf{z}_i)$ (see, e.g., Quiñonero-Candela and Rasmussen, 2005). These inducing variables are assumed to be drawn from the same GP as the original process, yielding the joint prior $p(\mathbf{f}, \mathbf{u}) = p(\mathbf{u})p(\mathbf{f} \,|\, \mathbf{u})$. In the spirit of Bayesian modeling, any uncertainty in the covariance should be accounted for. Thinking of GP hyper-parameters and inducing inputs as parameters of the covariance function, a distribution over these induces a distribution over the covariance function, which enriches the modeling capabilities of these models (see, e.g., Jang et al. (2017)). We consider a general formulation where we place priors $p_{\boldsymbol{\psi}}(\boldsymbol{\theta})$ over covariance hyper-parameters and $p_{\boldsymbol{\xi}}(\mathbf{Z})$ over inducing inputs with hyper-parameters $\boldsymbol{\psi}, \boldsymbol{\xi}$,

$$p(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{u}, \mathbf{f}, \mathbf{y} \,|\, \mathbf{X}) = \tag{1}$$
$$= p_{\boldsymbol{\psi}}(\boldsymbol{\theta}) p_{\boldsymbol{\xi}}(\mathbf{Z}) p(\mathbf{u} \,|\, \mathbf{Z}, \boldsymbol{\theta}) p(\mathbf{f} \,|\, \mathbf{u}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) p(\mathbf{y} \,|\, \mathbf{f}),$$

where $p(\mathbf{u} \,|\, \mathbf{Z}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{\mathrm{zz}|\boldsymbol{\theta}})$, $p(\mathbf{f} \,|\, \mathbf{u}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{K}_{\mathrm{xz}|\boldsymbol{\theta}} \mathbf{K}_{\mathrm{zz}|\boldsymbol{\theta}}^{-1} \mathbf{u}, \mathbf{K}_{\mathrm{xx}|\boldsymbol{\theta}} - \mathbf{K}_{\mathrm{xz}|\boldsymbol{\theta}} \mathbf{K}_{\mathrm{zz}|\boldsymbol{\theta}}^{-1} \mathbf{K}_{\mathrm{xz}|\boldsymbol{\theta}}^{\top})$. The matrices $\mathbf{K}_{\mathrm{zz}|\boldsymbol{\theta}}, \mathbf{K}_{\mathrm{xz}|\boldsymbol{\theta}}$ denote the covariance matrices

computed between points in $\mathbf{Z}$ and $\{\mathbf{X}, \mathbf{Z}\}$, respectively. We assume a factorized likelihood $p(\mathbf{y} \mid \mathbf{f}) = \prod_{n=1}^{N} p(y_n \mid f_n)$ and make no assumptions about the other distributions. In this general formulation, approaches that do not consider priors over covariance hyper-parameters or inducing inputs correspond to improper uniform priors in Eq. 1.

2.1 Scalable inference frameworks for GPs

Let $\boldsymbol{\Psi} \stackrel{\text{def}}{=} \{\mathbf{u}, \mathbf{Z}, \boldsymbol{\theta}\}$ be the variables whose posterior we wish to infer. Our main object of interest is the log joint marginal obtained by integrating out the latent variables $\mathbf{f}$ in Eq. 1, i.e., $\log p(\mathbf{y}, \boldsymbol{\Psi} \mid \mathbf{X}) = \log \int_{\mathbf{f}} p(\mathbf{y} \mid \mathbf{f}) p(\mathbf{f} \mid \boldsymbol{\Psi}, \mathbf{X}) \mathrm{d}\mathbf{f} + \log p(\boldsymbol{\Psi})$. In particular, we are interested in discussing approximations to this that decompose over observations, allowing the use of stochastic optimization techniques to scale up to large datasets. In the literature of sparse GPs (see, e.g., Bauer et al., 2016; Bui et al., 2017), two of the most influential methods for carrying out inference on such models are based on the variational free energy (VFE) framework (Titsias, 2009a) and the fully independent training conditional (FITC) framework (Snelson and Ghahramani, 2006).

VFE approximations. The key innovation in Titsias (2009a) is the definition of the approximate posterior $q(\mathbf{f}, \mathbf{u}) \stackrel{\text{def}}{=} q(\mathbf{u}) p(\mathbf{f} \mid \boldsymbol{\Psi}, \mathbf{X})$, where $q(\mathbf{u})$ is the variational posterior, which yields the evidence lower bound (ELBO)

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) \geq \tag{2}$$
$$-\operatorname{KL}\left[q(\mathbf{u}) \parallel p(\mathbf{u} \mid \mathbf{Z}, \boldsymbol{\theta})\right] + \mathbb{E}_{q(\mathbf{f}, \mathbf{u})} \log p(\mathbf{y} \mid \mathbf{f}) \stackrel{\text{def}}{=} \mathcal{L}_{\text{ELBO}} .$$

We note that this approach does not incorporate priors over inducing inputs or hyper-parameters. Inference involves constraining $q(\mathbf{u})$ to a parametric form and finding its parameters to optimize the ELBO. Titsias (2009a) correctly argues that in the regression setting the variational approach to inducing variable approximations should be more robust to overfitting than a direct marginal likelihood maximization approach of traditional approximate models such as those described in Quiñonero-Candela and Rasmussen (2005). Indeed, if inducing inputs $\mathbf{Z}$ are optimized then the resulting ELBO provides an additional regularization term (see Titsias, 2009a, §3 for details). However, as we shall see later, the benefits of being Bayesian about the inducing inputs and estimating their posterior distribution can be superior to those obtained by this regularization.

Restricting the form of $q(\mathbf{u})$ is suboptimal, and Hensman et al. (2015b) proposes to sample from the optimal posterior approximation instead. By applying Jensen's inequality to bound the log joint marginal we obtain the following formulation,

$$\log p(\mathbf{y}, \boldsymbol{\Psi} \mid \mathbf{X}) \geq \tag{3}$$
$$\mathbb{E}_{p(\mathbf{f} \mid \boldsymbol{\Psi}, \mathbf{X})} \log p(\mathbf{y} \mid \mathbf{f}) + \log p(\boldsymbol{\Psi}) \stackrel{\text{def}}{=} \log \widetilde{p}_{\text{VFE}}(\mathbf{y}, \boldsymbol{\Psi} \mid \mathbf{X}).$$

This is the same expression derived in Hensman et al. (2015b), although following a different derivation showing that $\widetilde{p}_{\text{VFE}}$ indeed yields the optimal distribution under the VFE framework of Eq. 2. However, Hensman et al. (2015b) argues that a Bayesian treatment of inducing inputs is unnecessary and concludes that the optimal prior is $p(\mathbf{Z}) = q(\mathbf{Z}) = \delta(\mathbf{Z} - \hat{\mathbf{Z}})$, where $\delta(\cdot)$ is Dirac's delta function and $\hat{\mathbf{Z}}$ is the set of inducing inputs that maximizes the ELBO (Hensman et al., 2015b, §3). We find such a justification flawed as it contradicts the fundamental principles of Bayesian inference. Indeed, the derivation by Hensman et al. (2015b) relies on minimizing both sides of the KL term in Eq. 2, allowing for a 'free-form' optimization of the prior, which ultimately negates the necessity of all prior choices and defeats the purpose of a Bayesian treatment.

FITC approximations. As an alternative, we can approximate the log joint of Eq. 1 by imposing independence in the conditional distribution (see Quiñonero-Candela and Rasmussen, 2005, for details), i.e., parameterizing $p(\mathbf{f} \mid \boldsymbol{\Psi}, \mathbf{X}) = \mathcal{N}\left(\mathbf{K}_{\text{xz}|\boldsymbol{\theta}} \mathbf{K}_{\text{zz}|\boldsymbol{\theta}}^{-1} \mathbf{u}, \operatorname{diag}\left[\mathbf{K}_{\text{xx}|\boldsymbol{\theta}} - \mathbf{K}_{\text{xz}|\boldsymbol{\theta}} \mathbf{K}_{\text{zz}|\boldsymbol{\theta}}^{-1} \mathbf{K}_{\text{xz}|\boldsymbol{\theta}}^{\top}\right]\right)$,

$$\log p(\mathbf{y}, \boldsymbol{\Psi} \mid \mathbf{X}) \approx$$
$$\underbrace{\sum_{n=1}^{N} \log \mathbb{E}_{p(f_n \mid \boldsymbol{\Psi}, \mathbf{X})}\left[p(y_n \mid f_n)\right] + \log p(\boldsymbol{\Psi})}_{\stackrel{\text{def}}{=} \log \tilde{p}_{\text{FITC}}(\mathbf{y}, \boldsymbol{\Psi} \mid \mathbf{X})} . \tag{4}$$

This same formulation of the FITC objective can be also been obtain by modifying the likelihood or the prior rather then the conditional distribution (Snelson and Ghahramani, 2006; Titsias, 2009a; Bauer et al., 2016).

We now see that, when considering i.i.d. conditional likelihoods, both approximations, $\log \tilde{p}_{\text{VFE}}$ and $\log \tilde{p}_{\text{FITC}}$, yield objectives that decompose on the observations, enabling scalable inference methods. In particular, we aim to sample from the posterior over all the latent variables using scalable approaches such as stochastic gradient Hamiltonian Monte Carlo (SGHMC; Chen et al., 2014). The main question is what approach should be preferred and how they relate to their optimization counterparts.

2.2 Sampling with VFE or FITC?

We will show in § 4 that our proposal that samples from the posterior according to Eq. 4 consistently outperforms that in Eq. 3. To understand why the 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). Indeed, Titsias (2009b) shows that, in the standard regression case with homoskedastic observation noise, VFE 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. 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. 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). Effectively, as described by Quiñonero-Candela and Rasmussen (2005), the 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. a conditional prior with zero variance.

Consequently, the main reason for the superior performance of VFE in earlier approaches, despite providing a less accurate predictive posterior than FITC's, was that inducing input estimation was less prone to overfitting due to the use of the variational objective, which provided an extra regularization term. However, by placing priors over the inducing inputs $\mathbf{Z}$ as well as over covariance hyper-parameters (as we propose in this work), regularization over these parameters becomes unnecessary. For this reason, we expect the *log of the expectation* in Eq. 4 to provide more accurate results than the *expectation of the log* in Eq. 3. Finally, it is important to point out that a variational formulation equivalent to FITC has also been proposed (see Titsias, 2009b, App. C). Our experimental evaluation, also assesses the benefits of our method with respect to that approach. Full details of this analysis can be found in the supplement.

3 PRACTICAL CONSIDERATIONS AND EXTENSION TO DEEP GPs

In this section we describe practical considerations in our Bayesian sparse Gaussian process (BSGP) framework, including inference techniques, prior choices and extensions to deep Gaussian processes. Recalling that $\boldsymbol{\Psi} = \{\boldsymbol{\theta}, \mathbf{u}, \mathbf{Z}\}$ represents the set of variables to infer and, using Eq. 4, their posterior can be obtained as

$$\log p(\boldsymbol{\Psi} \,|\, \mathbf{y}, \mathbf{X}) = \log \mathbb{E}_{p(\mathbf{f} \,|\, \boldsymbol{\Psi}, \mathbf{X})} p(\mathbf{y} \,|\, \mathbf{f}) + \log p(\mathbf{u} \,|\, \boldsymbol{\theta}, \mathbf{Z}) + \log p_{\boldsymbol{\xi}}(\mathbf{Z}) + \log p_{\psi}(\boldsymbol{\theta}) - \log C. \quad (5)$$

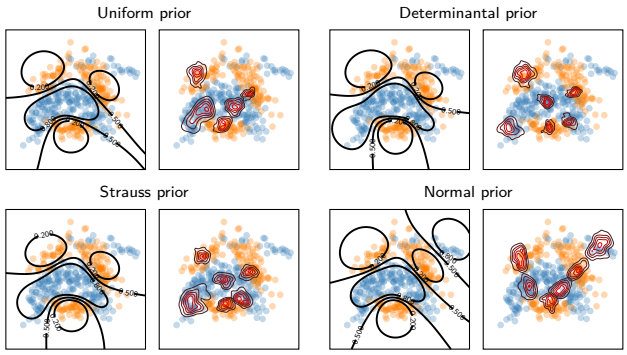

Figure 2: Illustration of a binary classification task on the BANANA dataset. *Left*: the decision bounds of the average classifier. *Right*: the posterior marginals of the inducing inputs.

We use Markov chain Monte Carlo (MCMC) techniques, in particular stochastic gradient Hamiltonian Monte Carlo (SGHMC) (Chen et al., 2014; Havasi et al., 2018), to obtain samples from the intractable $p(\boldsymbol{\Psi} \,|\, \mathbf{y}, \mathbf{X})$. Unlike Hamiltonian Monte Carlo (HMC), which requires computing the exact gradient $\nabla \log p(\boldsymbol{\Psi} \,|\, \mathbf{y}, \mathbf{X})$ and the exact unnormalized posterior to evaluate the acceptance (Neal, 2010), SGHMC obtains samples from the posterior with stochastic gradients and without evaluating the Metropolis ratio (see supplement for details). With a factorized likelihood $p(\mathbf{y}|\mathbf{f})$ and an energy function $U(\boldsymbol{\Psi}) = -\log p(\boldsymbol{\Psi} \,|\, \mathbf{y}, \mathbf{X}) + \log C$, we sample Eq. 5 over minibatches of data.

3.1 Choosing priors

Next, we discuss prior choices for the inducing inputs and covariance hyper-parameters. The inducing inputs $\mathbf{Z}$ support the sparse Gaussian process interpolation, which motivates matching the inducing prior to the data distribution $p(\mathbf{X})$. We begin by proposing a simple Normal (N) prior $p_N(\mathbf{Z}) = \prod_{j=1}^{M} \mathcal{N}(\mathbf{z}_j|\mathbf{0}, \mathbf{I})$, which matches the mean and variance of the normalized data distribution, and favors inducing inputs toward the baricenter of the data inputs.

We also explore two priors based on point processes, which consider distributions over point sets (González et al., 2016). Point processes can induce repulsive effects penalizing configurations where inducing points are clumped together. The determinantal point process (DPP), defined through $p_D(\mathbf{Z}) \propto \det \mathbf{K}_{\mathbf{Z}\mathbf{Z}|\boldsymbol{\theta}}$, relates the probability of inducing inputs to the volume of space spanned by the covariance (Lavancier et al., 2015). DPP is a repulsive point process, which gives higher probabilities to input diversity, controlled by the hyper-parameters $\boldsymbol{\xi} \equiv \boldsymbol{\theta}$. We then consider the Strauss process (see e.g. Daley and Vere-Jones, 2003; Strauss, 1975), $p_S(\mathbf{Z}) \propto \lambda^M \gamma^{\sum_{\mathbf{z},\mathbf{z}' \in \mathbf{Z}} \delta(|\mathbf{z}-\mathbf{z}'|<r)}$, where $\lambda > 0$ is

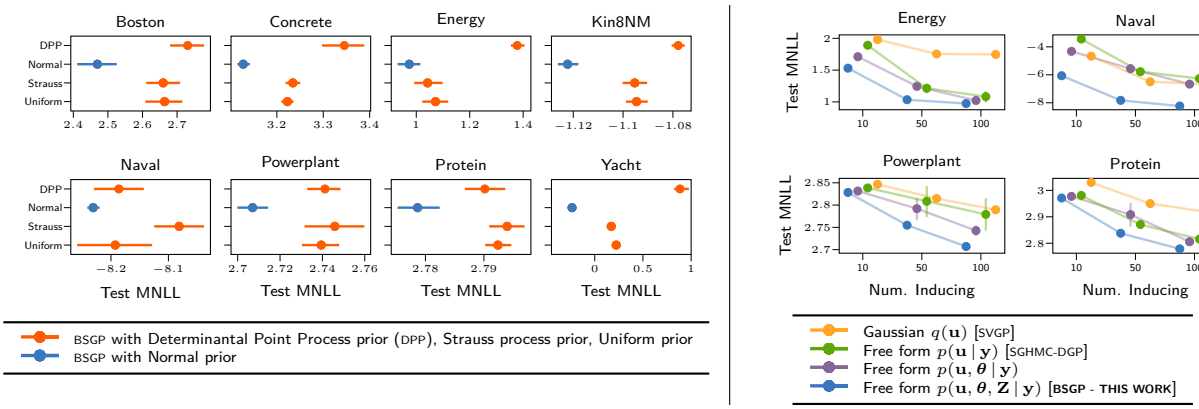

Figure 3: *Left:* analysis of different priors on inducing locations for BSGP on the UCI benchmarks: determinantal point process (DPP), Strauss process, uniform. *Right:* ablation study on the effect of performing posterior inference on different sets of variables. From SVGP, where the posterior is constrained to be Gaussian and the remaining parameters are point-estimated, to our proposal BSGP, where we infer a free-form posterior for all $\boldsymbol{\Psi} = \{\mathbf{u}, \boldsymbol{\theta}, \mathbf{Z}\}$. We refer the reader to Table 1 for details on the methods (colors are matched).

the intensity, and $0 < \gamma \leq 1$ is the repulsion coefficient which decays the prior as a function of the number of input pairs that are within distance r. The Strauss prior (S) tends to maintain the minimum distance between inducing inputs, parameterized by $\boldsymbol{\xi} = (\lambda, \gamma, r)$. We finally consider an uninformative uniform prior (U), $\log p_U(\mathbf{Z}) = 0$, which effectively provides no contribution to the evaluation of the posterior. To gain insights on the choice of these priors, we set up a comparative analysis on the BANANA dataset (Fig. 2). We observe that the posterior densities on the inducing inputs are multimodal and highly non-Gaussian, further confirming the necessity of free-form inference. Both Strauss and DPP-based priors encourage configurations where the inducing inputs are evenly spread. The Normal and Uniform priors, instead, focus exclusively on aligning the inducing inputs in a way that is sensible to accurately model the intricate classification boundary between the classes. This insight is confirmed by our the extensive experimental validation in § 4.

Prior on covariance hyper-parameters. Choosing priors on the hyper-parameters has been discussed in previous works on Bayesian inference for GPs (see e.g. Filippone and Girolami, 2014). Throughout this paper, we use the RBF covariance with automatic relevance determination (ARD), marginal variance σ and independent lengthscales λ_i per feature (Mackay, 1994). On these two hyper-parameters we place a lognormal prior with unit variance and means equal to 1 and 0.05 for $\boldsymbol{\lambda}$ and σ, respectively.

3.2 Extension to deep Gaussian processes

BSGP can be easily extended to deep Gaussian process (DGP) models (Damianou and Lawrence, 2013), where we compose L sparse GP layers. Each layer is associated with a set of inducing inputs $\mathbf{Z}_\ell$, inducing variables $\mathbf{u}_\ell$ and hyper-parameters $\boldsymbol{\theta}_\ell$ (Salimbeni and Deisenroth, 2017). In our notation $\boldsymbol{\Psi} = \{\boldsymbol{\Psi}_\ell\}_{\ell=1}^L = \{\mathbf{Z}_\ell, \mathbf{u}_\ell, \boldsymbol{\theta}_\ell\}_{\ell=\ell}^L$. The joint distribution is

$$p(\mathbf{y}, \boldsymbol{\Psi}) = p(\mathbf{y} \mid \mathbf{f}_L) \prod_{\ell=1}^L p(\mathbf{f}_\ell \mid \mathbf{f}_{\ell-1}, \boldsymbol{\Psi}_\ell) p(\boldsymbol{\Psi}_\ell), \quad (6)$$

where we omit dependency on $\mathbf{X}$. In contrast to the 'shallow' joint distribution in Eq. 1 and posterior in Eq. 5, the 'hidden' layers $\mathbf{f}_\ell$ are marginalized with sampling and propagated up to the final layer L (Salimbeni and Deisenroth, 2017), which can be marginalized exactly if the likelihood is Gaussian or by quadrature (Hensman et al., 2015a). Full details and derivations for this more general case can be found in the supplement.

4 EXPERIMENTS

In this section, we provide empirical evidence that our BSGP outperforms previous inference/optimization approaches on shallow and deep GPs. We use eight of the classic UCI benchmark datasets with standardized features and split into eight folds with 0.8/0.2 train/test ratio. We train the competing models for 10,000 iterations with ADAM (Kingma and Ba, 2015), step size of 0.01 and a minibatch of 1,000 samples. The sampling methods are evaluated based on 256 samples collected after optimization. Following previous works (e.g. Rasmussen and Williams, 2006; Havasi et al., 2018; Yu et al., 2019), in order to evaluate and compare the full predictive posteriors we compute the mean negative loglikelihood (MNLL) on the test set (RMSEs are reported in the supplement for reference).

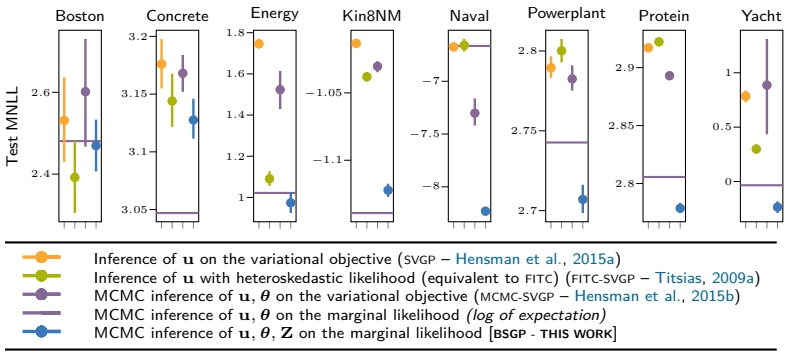

Figure 4: Analysis of different choices of objectives when used for optimization and sampling. We refer the reader to Table 1 for a description of the methods.

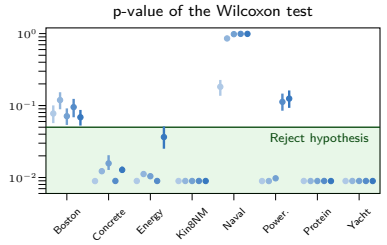

Figure 5: p-values of the hypothesis test that BSGP with VFE objective is better than BSGP with FITC objective; depth of the DGP from 1 (●) to 5 (●). For models with p-values < 0.05, we reject the hypothesis.

4.1 Prior analysis and ablation study

We start our empirical analysis with a comparative evaluation of the priors on inducing inputs described in § 3.1: DPP, Normal, Strauss and Uniform. We run our inference procedure on a shallow GP with 100 inducing points and we report the results in Fig. 3 (left). The results show that the Normal prior consistently outperforms the others. The uniform and Strauss priors behave similarly, while the DPP prior is consistently among the worst. We argue that the repulsive nature of the point process priors (DPP, particularly), although grounded on the intuition of covering the input space more evenly, constrains the smoothness of the functions up to the point that they become too simple to accurately model the data. With this, we select the Gaussian prior for the remaining experiments.

We now study the benefits of a Bayesian treatment of the inducing variables, inducing inputs, and hyperparameters with an ablation study. Using the same setup as before, we start with the baseline of SVGP (Hensman et al., 2013, 2015a), where the posterior on $\mathbf{u}$ is approximated using a Gaussian and $\mathbf{Z}, \boldsymbol{\theta}$ are optimized. We then incrementally add parameters to the list of variables that are sampled rather than optimized: only $\mathbf{u}$ (equivalent to SGHMC-DGP, (Havasi et al., 2018)), then $\{\mathbf{u}, \boldsymbol{\theta}\}$ and finally, our proposal, $\{\mathbf{u}, \boldsymbol{\theta}, \mathbf{Z}\}$. This experiment is repeated for different number of inducing points (10, 50 and 100). Fig. 3 (right) reports a summary of these results (full comparison in the supplement). This plot shows that each time we carry out free-form posterior inference on a bigger set of parameters rather than optimization, performance is enhanced, and our proposal outperforms previous approaches.

4.2 Choosing the objective: VFE vs FITC

In § 2 we discussed the role of the marginal and the variational free energy (VFE) objective when used for optimization and for sampling. In Fig. 4 we support the discussion with empirical results. The baseline is SVGP, for which the inference is approximate (Gaussian) and performed on the variational objective. Titsias (2009b, App. C) also considers a VFE formulation of FITC which corresponds to a GP regression with heteroskedastic noise variance. The likelihood needs to be augmented to handle heteroskedasticity, but inference can be carried out exactly on the variational objective. For these two methods, $\{\boldsymbol{\theta}, \mathbf{Z}\}$ are optimized. We also test MCMC-SVGP, the model proposed by Hensman et al. (2015b), implemented in GPflow (Matthews et al., 2017) with the same suggested experimental setup. This experiment indicates that having a free-form posterior on $\mathbf{u}, \boldsymbol{\theta}$ sampled from the variational objective does not dramatically improve on the exact Gaussian approximation of the FITC model, with both of them delivering superior performance with respect to SVGP. In the same setup of Hensman et al. (2015b) ($\mathbf{u}, \boldsymbol{\theta}$ sampled and $\mathbf{Z}$ optimized), we look at the effect of swapping the *expectation of log* with the *log of expectation* (which effectively means moving from the VFE objective to FITC); on the contrary, here we observe a significant increase in performance when using the latter, further confirming the discussion of the objectives in § 2.2. We finally conclude this section with an experiment where we try both objectives on our proposed BSGP

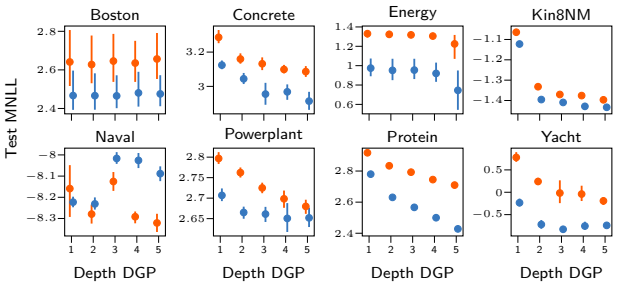

Figure 6: BSGP with the two different objectives for the sampler: FITC (●) and VFE (●).

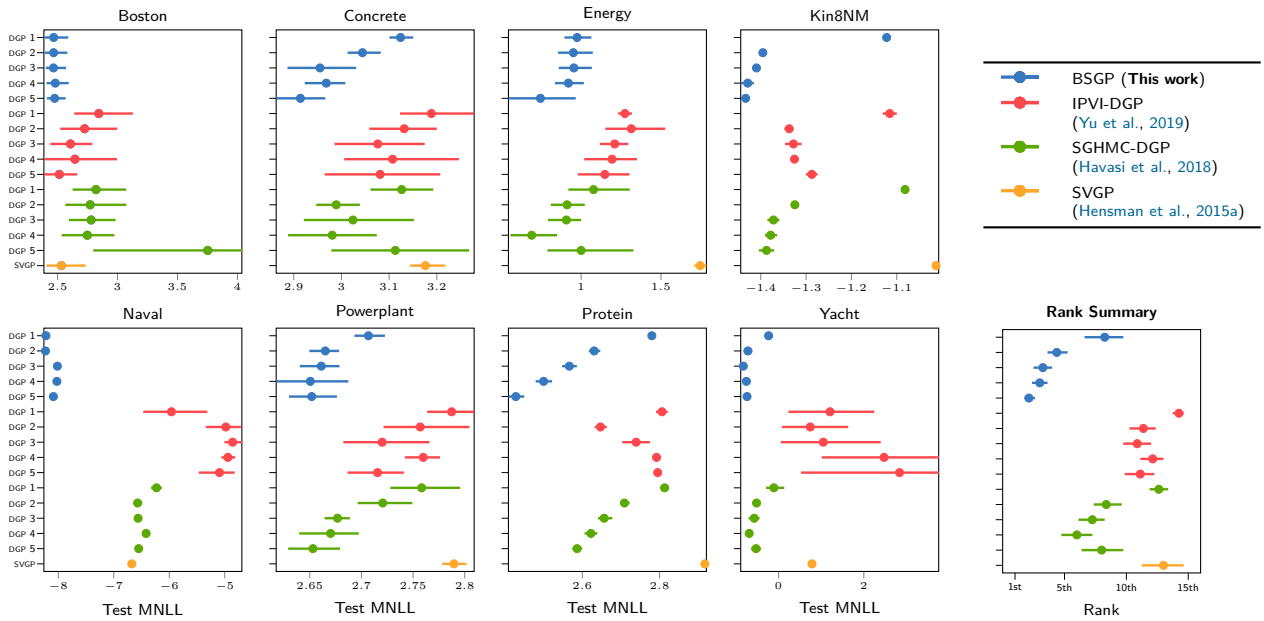

Figure 7: Test MNLL on UCI regression benchmarks (the error bars represent the 95%CI). The lower MNLL (i.e. to the left), the better. The number on the right of the method's name refers to the depth of the DGP. *Bottom right*: Rank summary of all methods.

and also different depths of the DGP (Fig. 6). Using the Wilcoxon signed-rank test (Wilcoxon, 1945), we test the null hypothesis of VFE objective being better than the proposed FITC. Fig. 5 shows that, for the majority of the cases, this can be rejected ($p < 0.05$).

4.3 Deep Gaussian processes on UCI benchmarks

We now report results on DGPs. We compare against two current state-of-the-art deep GP methods, SGHMC-DGP (Havasi et al., 2018) and IPVI-DGP (Yu et al., 2019), and against the shallow SVGP baseline (Hensman et al., 2015a). For a faithful comparison with IPVI-DGP we follow the recommended parameter configurations[1]. Using a standard setup, all models share $M = 100$ inducing points, the same RBF covariance with ARD and, for DGP, the same hidden dimensions (equal to the input dimension D). Fig. 7 shows the predictive test MNLL mean and 95% CI over the different folds over the UCI datasets, and also includes rank summaries. The proposed method clearly outperforms competing deep and shallow GPs. The improvements are particularly evident on NAVAL, a dataset known to be challenging to improve upon. Furthermore, the deeper models perform consistently better or on par with the shallow version, without incurring in any measurable overfitting even on small or medium sized datasets (see BOSTON and YACHT, for example).

[1]We use the IPVI-DGP implementation available at github.com/HeroKillerEver/ipvi-dgp

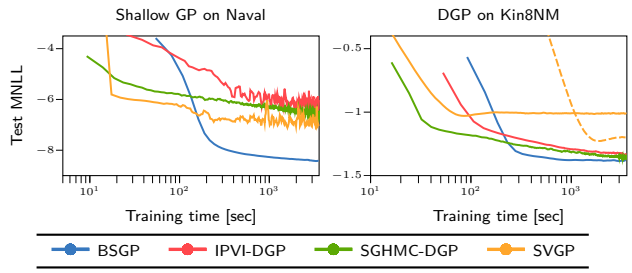

Figure 8: Comparison of test MNLL as function of training time. The dashed line on the right hand side plot corresponds to SVGP with $M = 1000$ inducing points.

Computational efficiency. Similarly to the baseline algorithms, each training iteration of BSGP involves the computation of the inverse covariance with complexity $\mathcal{O}(M^3)$. In Fig. 8 we compare the three main competitors with BSGP trained for a fixed training time budget of one hour for a shallow GP and a 2-layer DGP. The experiment is repeated four times on the same fold and the results are then averaged. Each run is performed on an isolated instance in a cloud computing platform with 8 CPU cores and 8 GB of reserved memory (Pace et al., 2017). Inference on the test set is performed every 250 iterations. This shows that BSGP converges considerably faster in wall-clock time, even though a single gradient step requires slightly more time.

Computing the predictive distribution, on the other hand, is more challenging as it requires recomputing the covariance matrices $\mathbf{K}_{xz}, \mathbf{K}_{zz}$ for each posterior

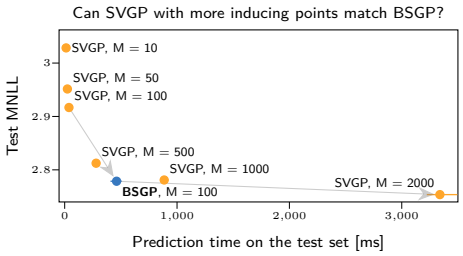

Figure 9: Comparison of test MNLL as a function of prediction time on the largest dataset (Protein).

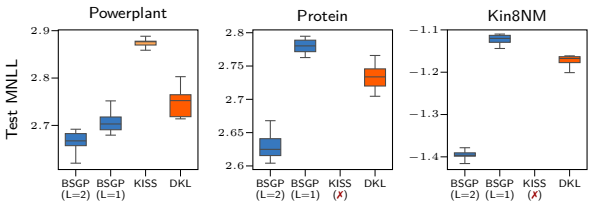

Figure 10: Comparison with structured inducing variables methods. KISS-GP could only run on the Powerplant dataset (hence the ✗ on Protein and Kin8NM).

sample $\{\mathbf{Z}, \boldsymbol{\theta}\}$, for an overall complexity linear in the number of posterior samples. This operation can be easily parallelized and implemented on GPUs but it could question the practicality of using a more involved inference method. In particular, it is relevant to study whether SVGP could deliver superior performance with a higher number of inducing points for less computational overhead. In Fig. 9 we study this trade-off on the biggest dataset considered (Protein): while it is evident that predictions with BSGP take more time (assuming a serial computation of the covariance matrices), it is also clear that the number of inducing points SVGP requires to (even marginally) improve upon BSGP is significantly larger (up to 20 times).

Structured inducing points. Finally, we run one last comparison with methods which exploit structure in the inputs. These models allow one to scale the number of inducing variables while maintaining computational tractability. Kernel Interpolation for Scalable Structured Gaussian Processes (KISS-GP) (Wilson and Nickisch, 2015) proposes to place the inducing inputs on a fixed and equally-spaced grid and to exploit Toeplitz/Kronecker structures with an iterative conjugate gradient method to further enhance scalability. Despite these benefits, KISS-GP is known to fall short with high-dimensional data ($D > 4$). This shortcoming was later addressed with Deep Kernel Learning (DKL) (Wilson et al., 2016a): using a deep neural network DKL projects the data in a lower dimensional manifold by learning an useful feature representation, which is then used as input to a KISS-GP. In Fig. 10 we have the comparison of BSGP with these two methods. KISS-GP could only run on Powerplant, with a 4-dimensional grid of size 10 (for a total of 10,000 inducing points). Here, BSGP delivers better performance despite having less inducing points. For DKL we followed the suggestion of Wilson et al. (2016a) to use a fully-connected neural network with a $[d-1000-1000-500-50-2]$ architecture as feature extractor and a grid size of 100 (for again a total of 10,000 inducing points). Training is performed by alternating optimization of the neural network weights and the KISS-GP parameters. Thanks to the flexibility of the feature extractor, this configuration is very competitive with our shallow BSGP, but it yields lower performance compared to a 2-layer DGP.

4.4 Large scale classification

The AIRLINE dataset is a classic benchmark for large scale classification. It collects delay information of all commercial flights in USA during 2008, counting more than 5 millions data points. The goal is to predict if a flight will be delayed based on 8 features, namely *month*, *day of month*, *day of week*, *airtime*, *distance*, *arrival time*, *departure time* and *age* of the plane. We preprocess the dataset following the guidelines provided in (Hensman et al., 2015b; Wilson et al., 2016a).

After a burn-in phase of 10,000 iterations, we draw 200 samples with 1000 simulation steps in between. We test on 100,000 randomly selected held-out points. We fit three models with $M = 100$ inducing points. Table 2 shows the predictive performance of three shallow GP models. The BSGP yields the best test error, MNLL, and test area under the curve (AUC). We assess the convergence of the predictive posterior by evaluating the $\hat{R}$-statistics (Gelman et al., 2004) over four independent SGHMC chains. This diagnostic yielded a $\hat{R} = 1.02 \pm 0.045$, which indicates good convergence. We report further convergence analysis in the supplement.

Table 2: AIRLINE dataset predictive test performance.

Model	Error (↓)	MNLL (↓)	AUC (↑)
SGHMC-GP	35.85%	0.646	0.671
SVGP	31.26%	0.595	0.730
BSGP	**30.46%**	**0.580**	**0.749**

As a further large scale example, we use the HIGGS dataset (Baldi et al., 2014), which has 11 millions data points with 28 features. This dataset was created by Monte Carlo simulations of particle dynamics in accelerators to detect the Higgs boson. We select 90% of the these points for training, while the rest is kept for testing. Table 3 reports the final test performance, showing that BSGP outperforms the competing meth-

Table 3: HIGGS dataset predictive test performance.

Model	Error (↓)	MNLL (↓)	AUC (↑)
SGHMC-GP	35.39%	0.628	0.698
SVGP	27.79%	0.544	0.796
BSGP	**26.97%**	**0.530**	**0.808**

ods. Interestingly, in both these large scale experiments, SGHMC-GP always falls back considerably w.r.t. BSGP and even SVGP. We argue that, with these large sized datasets, the continuous alternation of optimization of $\mathbf{Z}$ and $\boldsymbol{\theta}$ and sampling of $\mathbf{u}$ used by the authors (called Moving Window MCEM, see Havasi et al. (2018) for details) might have led to suboptimal solutions.

5 CONCLUSION & DISCUSSION

We have developed a fully Bayesian treatment of sparse Gaussian process models that considers the inducing inputs, along with the inducing variables and covariance hyper-parameters, as random variables, places suitable priors and carries out approximate inference over them. Our approach, based on SGHMC, investigated two conventional priors (Gaussian and uniform) for the inducing inputs as well as two point process based priors (the Determinantal and the Strauss processes).

By challenging the standard belief of most previous work on sparse GP inference that assumes the inducing inputs can be estimated point-wisely, we have developed a state-of-the-art inference method and have demonstrated its outstanding performance on both accuracy and running time on regression and classification problems. We hope this work can have an impact similar (or better) to other works in machine learning that have adopted more elaborate Bayesian machinery (e.g. Wallach et al., 2009) for long-standing inference problems in commonly used probabilistic models.

Finally, we believe it is worth investigating further more structured priors similar to those presented here (e.g. exploring different hyper-parameter settings), including a full joint treatment of inducing inputs and their number, i.e. $p(\mathbf{Z}, M)$. We leave this for future work. We are currently investigating ways to extend BSGP to convolutional Gaussian process (van der Wilk et al., 2017; Dutordoir et al., 2020; Blomqvist et al., 2020).

Acknowledgements. MF gratefully acknowledges support from the AXA Research Fund and the Agence Nationale de la Recherche (grant ANR-18-CE46-0002).

Bibliography

Baldi, P., Sadowski, P., and Whiteson, D. (2014). Searching for Exotic Particles in High-Energy Physics with Deep Learning. *Nature Communications*, 5(1):4308.

Barber, D. and Williams, C. K. I. (1997). Gaussian Processes for Bayesian Classification via Hybrid Monte Carlo. In *Advances in Neural Information Processing Systems 9*, pages 340–346. MIT Press.

Bauer, M., van der Wilk, M., and Rasmussen, C. E. (2016). Understanding Probabilistic Sparse Gaussian Process Approximations. In *Advances in Neural Information Processing Systems 29*, pages 1533–1541. Curran Associates, Inc.

Blomqvist, K., Kaski, S., and Heinonen, M. (2020). Deep Convolutional Gaussian Processes. In *Machine Learning and Knowledge Discovery in Databases*, pages 582–597, Cham. Springer International Publishing.

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.

Bui, T. D., Yan, J., and Turner, R. E. (2017). A Unifying Framework for Gaussian Process Pseudo-Point Approximations using Power Expectation Propagation. *Journal of Machine Learning Research*, 18(1):3649–3720.

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

Cutajar, K., Bonilla, E. V., Michiardi, P., and Filippone, M. (2017). Random Feature Expansions for Deep Gaussian Processes. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 884–893. PMLR.

Daley, D. J. and Vere-Jones, D. (2003). *An introduction to the theory of point processes. Vol. I.* Probability and its Applications (New York). Springer-Verlag, second edition. Elementary theory and methods.

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*, volume 31 of *Proceedings of Machine Learning Research*, pages 207–215. JMLR.org.

Dutordoir, V., van der Wilk, M., Artemev, A., and Hensman, J. (2020). Bayesian Image Classification with Deep Convolutional Gaussian Processes. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, AISTATS 2020*, volume 108 of *Proceedings of Machine Learning Research*, pages 1529–1539. PMLR.

Filippone, M. and Girolami, M. (2014). Pseudo-marginal Bayesian inference for Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2214–2226.

Gal, Y. and Ghahramani, Z. (2016). Dropout As a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning, ICML 2016*, volume 48, pages 1050–1059. JMLR.org.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition.

Giraldo, J.-J. and Álvarez, M. A. (2019). A Fully Natural Gradient Scheme for Improving Inference of the Heterogeneous Multi-Output Gaussian Process Model. *arXiv preprint arXiv:1911.10225*.

González, J. A., Rodríguez-Cortés, F. J., Cronie, O., and Mateu, J. (2016). Spatio-temporal Point Process Statistics: A Review. *Spatial Statistics*, 18:505–544.

Havasi, M., Hernández-Lobato, J. M., and Murillo-Fuentes, J. J. (2018). Inference in Deep Gaussian Processes using Stochastic Gradient Hamiltonian Monte Carlo. In *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*, page 282–290. AUAI Press.

Hensman, J., Matthews, A., and Ghahramani, Z. (2015a). Scalable Variational Gaussian Process Classification. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015*, volume 38 of *Proceedings of Machine Learning Research*, pages 351–360. PMLR.

Hensman, J., Matthews, A. G., Filippone, M., and Ghahramani, Z. (2015b). MCMC for Variationally Sparse Gaussian Processes. In *Advances in Neural Information Processing Systems 28*, pages 1648–1656. Curran Associates, Inc.

Jang, P. A., Loeb, A., Davidow, M., and Wilson, A. G. (2017). Scalable Lévy Process Priors for Spectral Kernel Learning. In *Advances in Neural Information Processing Systems 30*, pages 3940–3949.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *Proceedings of the Third International Conference on Learning Representations, ICLR 2015*, San Diego, USA.

Krauth, K., Bonilla, E. V., Cutajar, K., and Filippone, M. (2017). AutoGP: Exploring the Capabilities and Limitations of Gaussian Process Models. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017*. AUAI Press.

Lavancier, F., Møller, J., and Rubak, E. (2015). Determinantal Point Process Models and Statistical Inference. *Royal Statistical Society B*, 77:853–877.

Lázaro-Gredilla, M. and Figueiras-Vidal, A. (2009). Inter-domain Gaussian Processes for Sparse Inference using Inducing Features. In *Advances in Neural Information Processing Systems 22*, pages 1087–1095. Curran Associates, Inc.

Mackay, D. J. C. (1994). Bayesian Methods for Backpropagation Networks. In *Models of Neural Networks III*, chapter 6, pages 211–254. Springer.

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

Murray, I. and Adams, R. P. (2010). Slice Sampling Covariance Hyperparameters of Latent Gaussian Models. In *Advances in Neural Information Processing Systems 23*, pages 1732–1740. Curran Associates, Inc.

Neal, R. M. (1997). Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification. *Technical Report*.

Neal, R. M. (2010). MCMC using Hamiltonian dynamics. In Handbook of Markov Chain Monte Carlo (eds S. Brooks, A. Gelman, G. Jones, XL Meng). Chapman and Hall/CRC Press.

Pace, F., Venzano, D., Carra, D., and Michiardi, P. (2017). Flexible Scheduling of Distributed Analytic Applications. In *Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, pages 100–109. IEEE Press.

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 *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc.

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

Salimbeni, H. and Deisenroth, M. (2017). Doubly Stochastic Variational Inference for Deep Gaussian Processes. In *Advances in Neural Information Processing Systems 30*, pages 4588–4599. Curran Associates, Inc.

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*.

Shi, J., Titsias, M. K., and Mnih, A. (2020). Sparse Orthogonal Variational Inference for Gaussian Processes. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*, volume 108 of *Proceedings of Machine Learning Research*, pages 1932–1942. PMLR.

Snelson, E. and Ghahramani, Z. (2006). Sparse Gaussian Processes using Pseudo-Inputs. In *Advances in Neural Information Processing Systems 18*, pages 1257–1264. MIT Press.

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

Strauss, D. J. (1975). A Model for Clustering. *Biometrika*, 62(2):467–475.

Titsias, M. K. (2009a). Variational Learning of Inducing Variables in Sparse Gaussian Processes. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 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*.

van der Wilk, M., Rasmussen, C. E., and Hensman, J. (2017). Convolutional Gaussian Processes. In *Advances in Neural Information Processing Systems 30*, pages 2849–2858. Curran Associates, Inc.

Wallach, H. M., Mimno, D. M., and McCallum, A. (2009). Rethinking LDA: Why Priors Matter. In *Advances in Neural Information Processing Systems 22*, pages 1973–1981. Curran Associates, Inc.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83.

Wilson, A. and Nickisch, H. (2015). Kernel Interpolation for Scalable Structured Gaussian Processes (KISS-GP). In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, pages 1775–1784.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. (2016a). Deep Kernel Learning. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016*, volume 51 of *Proceedings of Machine Learning Research*, pages 370–378, Cadiz, Spain. PMLR.

Wilson, A. G., Hu, Z., Salakhutdinov, R. R., and Xing, E. P. (2016b). Stochastic Variational Deep Kernel Learning. In *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 *Advances in Neural Information Processing Systems 32*, pages 14475–14486. Curran Associates, Inc.