Scalable Gaussian Process Inference with Finite-data Mean and Variance Guarantees

Jonathan H. Huggins
Harvard University

Trevor Campbell
University of British Columbia

Mikołaj Kasprzak
University of Luxembourg

Tamara Broderick
MIT

Abstract

Gaussian processes (GPs) offer a flexible class of priors for nonparametric Bayesian regression, but popular GP posterior inference methods are typically prohibitively slow or lack desirable finite-data guarantees on quality. We develop a scalable approach to approximate GP regression, with finite-data guarantees on the accuracy of our pointwise posterior mean and variance estimates. Our main contribution is a novel objective for approximate inference in the nonparametric setting: the *preconditioned Fisher (pF) divergence.* We show that unlike the Kullback–Leibler divergence (used in variational inference), the pF divergence bounds the 2-Wasserstein distance, which in turn provides tight bounds on the pointwise error of mean and variance estimates. We demonstrate that, for sparse GP likelihood approximations, we can minimize the pF divergence efficiently. Our experiments show that optimizing the pF divergence has the same computational requirements as variational sparse GPs while providing comparable empirical performance—in addition to our novel finite-data quality guarantees.

1 Introduction

Gaussian processes (GPs) offer a versatile class of models for functions. In particular, GPs are able to capture complex, highly non-linear relationships in data. Unfortunately, even in the setting of GP regression with Gaussian noise, exact GP inference—that is, computing the posterior mean and covariance functions of the GP given N observations—incurs a prohibitive $O(N^3)$

running time and $O(N^2)$ memory cost (Rasmussen and Williams, 2006). By contrast, modern, large-scale data sets require algorithms with time and memory requirements at most linear in N. A natural question arises: can a reliably good GP *approximation* be found with at most linear cost? To understand what counts as a "good" approximation, we observe that practitioners tend to report pointwise estimates and uncertainties for the GP regression curve, especially pointwise posterior means and standard deviations (Sacks et al., 1989; Kaufman and Sain, 2010; Kaufman et al., 2011; Gramacy and Lee, 2012; Rasmussen and Ghahramani, 2002; Rasmussen and Williams, 2006; Snoek et al., 2012; Osborne et al., 2012). To the best of our knowledge, it is an open problem whether non-trivial, non-asymptotic theoretical bounds exist on the quality of these estimates when running linear-time approximations to GP regression. We believe we provide the first approach that simultaneously satisfies the following desiderata: (1) linear running time and memory cost in N, (2) theoretical bounds on the quality of pointwise estimates of the posterior mean, standard deviation, and variance for finite data, and (3) a practical mechanism for provably decreasing these bounds to zero.

Previous approaches to scaling GP inference primarily fall into two categories: sparse GP methods (Smola and Bartlett, 2000; Seeger et al., 2003; Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2005; Titsias, 2009; Bauer et al., 2016; Hensman et al., 2013; de G Matthews et al., 2016; Bui et al., 2017) and structured kernel matrix approximations (Wilson and Nickisch, 2015; Ding et al., 2017; Katzfuss, 2017; Gardner et al., 2018; Izmailov et al., 2018). The sparse GP approach introduces a set of M *inducing points*, which are chosen either greedily or at random and can then be optimized. The most widely used and top-performing sparse GP methods are variational and originate with the work of Titsias (2009). These methods optimize the inducing point locations using a variational evidence lower bound (Titsias, 2009; Bauer et al., 2016; Hensman et al., 2013; de G Matthews et al., 2016; Bui et al., 2017) and have $O(M^2N)$ time and $O(MN)$ space complexity. Structured kernel approximations replace the

$N \times N$ kernel matrix with an approximation that can be represented more compactly and operated on more efficiently. The time and space requirements of these methods vary and usually depend on the dimensionality of the input space. Neither approach is guaranteed to provide accurate estimates of the posterior mean or variance. Typically, these methods are validated empirically on small datasets (Bauer et al., 2016).

In fact, after introducing GP regression in more detail in Section 2, we note in Section 3 that a posterior approximation can be close in Kullback-Leibler (KL) divergence to the exact posterior and still exhibit bad posterior mean and variance approximations. Thus, rather than pursue variational methods with KL divergence, we instead observe that closeness in *2-Wasserstein distance* implies closeness of means and covariances—as well as many other posterior functionals (e.g. expectations of any function with a bounded gradient). For GPs, it is not possible to efficiently compute the 2-Wasserstein distance to the exact posterior, so instead we develop a theoretically-motivated proxy.

In particular, we build from a related metric called the *Fisher distance*. The squared Fisher distance between two distributions is defined as the expected squared norm of the difference between their score functions. This distance has been successfully applied in the finite-dimensional setting both to design practical algorithms (Campbell and Broderick, 2019) and for theoretical analysis (Huggins and Zou, 2017; Ogden, 2017). In Section 4, we adapt the Fisher distance to the infinite-dimensional GP setting and call our new measure of fit the *preconditioned Fisher (pF) divergence*. We carefully design the pF divergence to work with likelihoods that are parameterized by functions—as opposed to, e.g., the finite vector parameters that are more familiar from the parametric setting. We show that the pointwise mean and variance approximation errors are bounded by a constant times the pF divergence, and thus minimizing the pF divergence results in small mean and variance approximation error. In Section 5 we demonstrate that the pF divergence can be computed and efficiently minimized for a sparse GP likelihood approximation. Our experiments in Section 6 on simulated and real data show that our method (1) yields comparable empirical performance with the existing state-of-the-art variational methods while (2) using as much or less computation. Thus, using the pF divergence as an objective for approximate GP inference is a promising alternative to the variational approach, offering state-of-the-art real-world performance while crucially providing finite-data guarantees on the accuracy of the posterior mean and variance estimates.

2 Sparse Gaussian process regression

A Gaussian process on covariate space $\mathcal{X}$ is determined by a mean function $\mu : \mathcal{X} \to \mathbb{R}$ and a positive-definite covariance, or kernel, function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Take any two sets of locations $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \in \mathcal{X}^N$ and $\boldsymbol{X}' = (\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{N'}) \in \mathcal{X}^{N'}$ and any function $f : \mathcal{X} \to \mathbb{R}$. Let $f_{\boldsymbol{X}} := (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N))$, and let $k_{\boldsymbol{X}\boldsymbol{X}'}$ be the $N \times N'$ matrix where $[k_{\boldsymbol{X}\boldsymbol{X}'}]_{nn'} = k(\boldsymbol{x}_n, \boldsymbol{x}'_{n'})$. If $f \sim \text{GP}(\mu, k)$ is a GP-distributed function, f is marginally Gaussian at any set of locations: $f_{\boldsymbol{X}} \sim \mathcal{N}(\mu_{\boldsymbol{X}}, k_{\boldsymbol{X}\boldsymbol{X}})$.

In GP regression (Rasmussen and Williams, 2006), we model noisy observed output y_n at input location $\boldsymbol{x}_n$:

$$y_n \,|\, f, \boldsymbol{x}_n \overset{\text{indep}}{\sim} \mathcal{N}(f(\boldsymbol{x}_n), \sigma^2), \quad n = 1, \ldots, N$$
$$f \sim \text{GP}(0, k),$$

where $\sigma^2 > 0$ is the observation noise variance and we make a standard assumption that the mean function is identically zero. We collect the response variables as $\boldsymbol{y} := (y_1 \ldots, y_N) \in \mathbb{R}^N$. The posterior distribution given the data $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{y})$ is the Gaussian process $\pi = \text{GP}(\mu_{\boldsymbol{D}}, k_{\boldsymbol{D}})$, where $\mu_{\boldsymbol{D}}(\boldsymbol{x}) := k_{\boldsymbol{x}\boldsymbol{X}}[k_{\boldsymbol{X}\boldsymbol{X}} + \sigma^2 I]^{-1}\boldsymbol{y}$ and $k_{\boldsymbol{D}}(\boldsymbol{x}, \boldsymbol{x}') := k(\boldsymbol{x}, \boldsymbol{x}') - k_{\boldsymbol{x}\boldsymbol{X}}[k_{\boldsymbol{X}\boldsymbol{X}} + \sigma^2 I]^{-1}k_{\boldsymbol{X}\boldsymbol{x}'}$ are, respectively, the posterior mean and kernel functions. The posterior mean and kernel functions are usually the quantities of interest since they fully define the GP posterior distribution. They can then be used for prediction, in Bayesian optimization, in the design of computer experiments, and for other tasks (Sacks et al., 1989; Kaufman and Sain, 2010; Kaufman et al., 2011; Gramacy and Lee, 2012; Rasmussen and Ghahramani, 2002; Rasmussen and Williams, 2006; Snoek et al., 2012; Osborne et al., 2012). Computing the posterior exactly requires $O(N^3)$ time due to the cost of inversion of the $N \times N$ matrix $k_{\boldsymbol{X}\boldsymbol{X}} + \sigma^2 I$. Sparse Gaussian process approximations make GP inference more scalable by introducing a set of M *inducing points* $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_m)$ and evaluating the function f only at those locations (Seeger et al., 2003; Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2005; Titsias, 2009). The key idea is to replace the exact log-likelihood $\mathcal{L}(f) := \log p_{\mathcal{N}}(\boldsymbol{y} \,|\, f_{\boldsymbol{X}}, \sigma^2 I)$ with an approximate log-likelihood that makes use of the observation that if $f \sim \text{GP}(0, k)$, then $\mathbb{E}[f_{\boldsymbol{X}} \,|\, f_{\tilde{\boldsymbol{X}}}] = k_{\boldsymbol{X}\tilde{\boldsymbol{X}}}k_{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}}^{-1}f_{\tilde{\boldsymbol{X}}}$. For example, the deterministic training conditional (DTC) approximation (Seeger et al., 2003; Quiñonero-Candela and Rasmussen, 2005) is $\tilde{\mathcal{L}}_{\text{DTC}}(f) := \log p_{\mathcal{N}}(\boldsymbol{y} \,|\, k_{\boldsymbol{X}\tilde{\boldsymbol{X}}}k_{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}}^{-1}f_{\tilde{\boldsymbol{X}}}, \sigma^2 I)$.

3 Wasserstein and Fisher distances

Scalable GP inference methods based on optimizing a variational lower bound for the model evidence have so far provided the best empirical performance when

compared to alternative methods (Titsias, 2009; Bauer et al., 2016; Hensman et al., 2013, 2015). While variational inference is an elegant approach, there are no finite-data guarantees on the accuracy of the approximate mean and covariance functions produced by variational methods. The issue is that variational methods minimize the KL divergence between the sparse posterior approximation and the exact posterior (Titsias, 2009; Bauer et al., 2016; Hensman et al., 2013), but a small KL divergence does not necessarily imply small error in the mean and covariance estimates. For clarity, we consider the case of finite-dimensional Gaussians.

Proposition 3.1. $\forall \delta > 0$ *there exist distributions* $\eta = \mathcal{N}(\mu, s^2)$ *and* $\tilde{\eta} = \mathcal{N}(\tilde{\mu}, \tilde{s}^2)$ *on* $\mathbb{R}$ *such that* $\mathrm{KL}(\tilde{\eta}||\eta) = \delta$, $(\mu - \tilde{\mu})^2 = \tilde{s}^2\{\exp(2\delta) - 1\}$, *and* $\tilde{s}^2 = \exp(-2\delta)s^2$.

See proof in Appendix B. Proposition 3.1 shows that, for example, if $\mathrm{KL}(\tilde{\eta}||\eta) = 5$ then the mean estimate may be off by more than $148\tilde{s}$. Since $\tilde{s}$ provides a natural unit of uncertainty about the parameter value, we see that a moderate Kullback–Leibler divergence can correspond to a very large error in the mean estimate.

An alternative to KL divergences that does imply closeness of means and covariances is the 2-Wasserstein distance. For the remainder of the paper let η, $\tilde{\eta}$, and ν denote probability measures on a measurable space $(\mathbb{H}, \mathcal{B})$, where $(\mathbb{H}, \langle\cdot, \cdot\rangle_{\mathbb{H}})$ is a Hilbert space. We assume these distributions are absolutely continuous with respect to a base measure λ. We will typically take η to be the GP regression posterior π, $\tilde{\eta}$ to be an approximation $\tilde{\pi}$ to the posterior, and λ to be the prior π_0. Let $\Gamma(\eta, \tilde{\eta})$ denote the collection of distributions γ on $\mathbb{H} \times \mathbb{H}$ such that γ has marginal distributions η and $\tilde{\eta}$: that is, $\eta = \gamma(\cdot, \mathbb{H})$ and $\tilde{\eta} = \gamma(\mathbb{H}, \cdot)$. The 2-Wasserstein distance between η and $\tilde{\eta}$ is given by

$$\mathcal{W}_2(\eta, \tilde{\eta}) := \inf_{\gamma \in \Gamma(\eta, \tilde{\eta})} \left\{ \int \|\theta - \tilde{\theta}\|_{\mathbb{H}}^2 \gamma(\mathrm{d}\theta, \mathrm{d}\tilde{\theta}) \right\}^{1/2}.$$

Closeness in 2-Wasserstein implies closeness of means, standard deviations, and many other expectations:

Proposition 3.2. *Assume that* $\mathbb{H} = \mathbb{R}$, *that* η *has mean* μ *and variance* s^2, *and that* $\tilde{\eta}$ *has mean* $\tilde{\mu}$ *and variance* $\tilde{s}^2$. *If* $\mathcal{W}_2(\eta, \tilde{\eta}) \leq \varepsilon$, *then* $|\mu - \tilde{\mu}| \leq \varepsilon$, $|s - \tilde{s}| \leq 2\varepsilon$, *and, for any function* $\phi : \mathbb{R} \to \mathbb{R}$ *such that* $|\mathrm{d}\phi/\mathrm{d}\theta| \leq L < \infty$, $|\mathbb{E}_{\theta \sim \eta}[\phi(\theta)] - \mathbb{E}_{\tilde{\theta} \sim v}[\phi(\tilde{\theta})]| \leq L\epsilon$.

Similar results hold for distributions in $\mathbb{R}^d$ and for Gaussian processes (detailed later in Theorem 4.3 and Appendix D). Unfortunately it is not feasible to directly minimize the 2-Wasserstein distance between the exact posterior π and some approximation $\tilde{\pi}$ because it would require computing the posterior, including its normalizing constant. Therefore, we introduce an alternative divergence measure we call the ν-*Fisher distance*. Crucially, the ν-Fisher distance satisfies two properties: (1)

it does not depend on the normalizing constant of the target distribution, so it can be computed in practice, and (2) it provides an upper bound on the 2-Wasserstein distance, so it provides accuracy guarantees on means and standard deviations. Let $L^2(\nu)$ denote the space of functions that are square-integrable with respect to ν: $\phi \in L^2(\nu) \implies \|\phi\|_{L^2(\nu)} := (\int \phi(\theta)^2 \nu(\mathrm{d}\theta))^{1/2} < \infty$.

Definition 3.3. When $\mathbb{H} = \mathbb{R}^d$, the ν-**Fisher distance** between two distributions is defined as the square root of the expected squared Euclidean distance between the score functions of the two densities:

$$d_{\mathrm{F},\nu}(\eta, \tilde{\eta}) := \left\| \left\| \nabla \log \frac{\mathrm{d}\tilde{\eta}}{\mathrm{d}\theta} - \nabla \log \frac{\mathrm{d}\eta}{\mathrm{d}\theta} \right\|_2 \right\|_{L^2(\nu)}$$

The ν-Fisher distance implies the following 2-Wasserstein bound, which for simplicity we state for the most relevant case of when η and $\tilde{\eta}$ are (finite-dimensional) Gaussians. For a matrix A, let $\sigma_{\min}(A)$ denote the smallest absolute eigenvalue of A and for a function ϕ, let $\|\phi\|_\infty$ denote the usual sup-norm.

Proposition 3.4. *Assume* $\mathbb{H} = \mathbb{R}^d$, $\eta = \mathcal{N}(\mu, \Sigma)$, $\tilde{\eta} = \mathcal{N}(\tilde{\mu}, \tilde{\Sigma})$. *Then for any probability measure* ν *such that* $\eta \ll \nu$, $\mathcal{W}_2(\eta, \tilde{\eta}) \leq \sigma_{\min}(\tilde{\Sigma})^{-1} \|\mathrm{d}\eta/\mathrm{d}\nu\|_\infty^{1/2} d_{\mathrm{F},\nu}(\eta, \tilde{\eta})$.

The special case when $\nu = \eta$ or $\tilde{\eta}$ is known as (e.g.) the *Fisher divergence* and has appeared in many applications. It serves as the objective in score matching, a model estimation technique (Sriperumbudur et al., 2017; Hyvarinen, 2005). Johnson and Barron (2004) and Ley and Swan (2013) bounded certain integral probability measures (e.g. the total variation and Kolmogorov distances) in terms of the Fisher divergence while Huggins and Zou (2017) bounded the 1-Wasserstein distance in terms of the Fisher divergence.[1] In the Bayesian setting, Campbell and Broderick (2019, 2018) minimized the ν-Fisher distance to construct a high-quality posterior approximation based on a coreset. Campbell and Broderick (2019, 2018) found their coreset method performed very well empirically, including accurate posterior mean and variance estimates. Ogden (2017) and Dalalyan and Karagulyan (2017) developed related guarantees—the former using a score function approach similar in spirit to the ν-Fisher distance. Finally, Huggins et al. (2017) used the results of Huggins and Zou (2017) to prove finite-data Wasserstein guarantees for a scalable approximate inference algorithm for generalized linear models.

4 Infinite-dimensional spaces

Together, Propositions 3.2 and 3.4 show that small ν-Fisher distance between two finite-dimensional Gaus-

---

[1]Regularity conditions were required in both cases.

sians implies small differences in their means and standard deviations. Based on these theoretical foundations and the empirical successes of Campbell and Broderick (2019, 2018), the ν-Fisher distance appears to be an attractive objective for optimizing the sparse GP likelihood approximation. However, there are a number of subtleties that must be addressed before using this distance, or a similar distance, for GP regression—in particular, due to the infinite-dimensional parameter space in this case. In the GP setting the Hilbert space $\mathbb{H}$ is a function space, so instead of gradients of the log-likelihood, we must use a functional derivative $\mathcal{D}$, taken with respect to the Hilbert norm $\|\cdot\|_{\mathbb{H}}$. That is, for a function $F : \mathbb{H} \to \mathbb{R}$, $\mathcal{D}F : \mathbb{H} \to \mathbb{H}$ is the linear operator that satisfies $\lim_{h\to 0} \|F(f + h) - F(f) - (\mathcal{D}F)(h)\|_{\mathbb{H}} / \|h\|_{\mathbb{H}} = 0$ for all $f \in \mathbb{H}$. The major obstacle to overcome is that the naïve extension of the ν-Fisher distance can differ from the 2-Wasserstein distance by an arbitrarily large multiplicative factor. If we tried to directly translate the ν-Fisher distance to the GP setting we would get

$$d_{\mathrm{F},\nu}(\eta, \tilde{\eta}) := \left\| \left\| \mathcal{D} \log \frac{\mathrm{d}\tilde{\eta}}{\mathrm{d}\lambda} - \mathcal{D} \log \frac{\mathrm{d}\eta}{\mathrm{d}\lambda} \right\|_{\mathbb{H}} \right\|_{L^2(\nu)}.$$

Notice that the derivatives are now functional derivatives and that the Radon-Nikodym derivatives are no longer with respect to Lebesgue measure, but rather with respect to the base measure λ.

To gain intuition for what can go wrong in the infinite-dimensional setting, recall from Proposition 3.4 that in the finite-dimensional Gaussian case, the constant in the bound depends on $\sigma_{\min}(\tilde{\Sigma})^{-1}$, the reciprocal of smallest eigenvalue of the covariance of $\tilde{\eta}$. In infinite dimensions, we replace $\tilde{\Sigma}$ with the *covariance operator* $\mathcal{C}_{\tilde{\eta}}$ associated with $\tilde{\eta}$ (Ibragimov and Rozanov, 1978, Ch. 1.4), the formal definition of which we defer. The eigenvalues of $\mathcal{C}_{\tilde{\eta}}$ get arbitrarily close to zero, so $\sigma_{\min}(\mathcal{C}_{\tilde{\eta}})^{-1} = \infty$ and the bound is vacuous. The next example illustrates the issue for GP regression.

Example 4.1. Consider a 1-dimensional Gaussian process on $\mathcal{X} = \mathbb{R}$ with squared exponential kernel $k(x, x') = e^{-(x-x')^2/2}$. Let η (respectively $\tilde{\eta}$) denote the GP regression posterior distribution with $N = 1$, $x_1 = 0$, and $y_1 = t$ (resp. $y_1 = \tilde{t}$). Then $d_{\mathrm{F},\nu}(\eta, \tilde{\eta}) = c(1 + \sigma^{-2})\mathcal{W}_2(\eta, \tilde{\eta})$, where c is a constant that depends on the choice of $\mathbb{H}$ and the factor of $(1 + \sigma^{-2})$ can be made arbitrarily large by taking $\sigma^2 \to 0$. See Appendix C for a detailed derivation.

Example 4.1 shows that the Fisher divergence bound on the Wasserstein distance could be arbitrarily large even when the Wasserstein distance itself is finite. To understand how to fix the problem, we can take inspiration from the finite-dimensional setting; suppose $\tilde{\Sigma}$ had a very small but non-zero eigenvalue. We could

perform a change of basis: $\theta \mapsto \tilde{\Sigma}^{1/2}\theta$. In the new basis, $\tilde{\eta}$ would have covariance equal to the identity matrix I, and hence the minimum eigenvalue would be one. Working with the score function, this change of variable corresponds to replacing $\nabla \log \frac{\mathrm{d}\tilde{\eta}}{\mathrm{d}\theta}(\theta) = -\tilde{\Sigma}^{-1}(\theta - \tilde{\mu})$ with $\tilde{\Sigma}\nabla \log \pi(\theta) = -(\theta - \tilde{\mu})$, where the final expression is equal to the gradient of the log density of a Gaussian with covariance I. Formally, this procedure carries over to the infinite-dimensional case: we replace the functional derivative $\mathcal{D}$ with $\mathcal{C}_{\tilde{\eta}}\mathcal{D}$.

To precisely define the covariance operator, we must describe the Hilbert space $\mathbb{H}$ in more detail. Specifically, we will take $\mathbb{H}$ to be a *reproducing kernel Hilbert space* (RKHS) (Steinwart and Christmann, 2008, Ch. 4). For a positive definite *reproducing kernel* $r : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the corresponding RKHS $\mathbb{H}_r$ comes equipped with an inner product $\langle \cdot, \cdot \rangle_r$. The function $r_{\boldsymbol{x}} := r(\boldsymbol{x}, \cdot)$ is the evaluation function at $\boldsymbol{x}$ (that is, the inner product and reproducing kernel satisfy the reproducing property): for any $f \in \mathbb{H}_r$ and $\boldsymbol{x} \in \mathcal{X}$, $\langle r_{\boldsymbol{x}}, f \rangle_r = f(\boldsymbol{x})$. We take r to be fixed but for clarity make the r-dependence explicit. Now we can define the covariance operator $\mathcal{C}_{\tilde{\eta}}$ associated with a distribution $\tilde{\eta}$ as the self-adjoint operator that satisfies the identity $\langle r_{\boldsymbol{x}}, \mathcal{C}_{\tilde{\eta}}r_{\boldsymbol{x}'} \rangle_r = \mathrm{Cov}(f(\boldsymbol{x}), f(\boldsymbol{x}'))$, where $f \sim \tilde{\eta}$.

Remark 4.1. We emphasize the very different roles played by the kernels k and r. Whereas k determines the prior covariance of the Gaussian process used for regression, r induces a reproducing kernel Hilbert space $\mathbb{H}_r$. Recall that η and $\tilde{\eta}$ should be thought of as exact and approximating GP regression posterior distributions, which depend on the choice of k. On the other hand, for a sample $f \sim \eta$ or $\tilde{\eta}$, by construction $f \in \mathbb{H}_r$. But, as we discuss further below, $f \notin \mathbb{H}_k$ almost surely.

We can give our definition of the modified ν-Fisher distance, which we call the ν-*preconditioned Fisher* (ν-*pF*) *divergence*.[2]

Definition 4.2. The ν-**preconditioned Fisher divergence** is defined by

$$d_{\mathrm{pF},\nu}(\tilde{\eta}\|\eta) := \left\| \left\| \mathcal{C}_{\tilde{\eta}}\mathcal{D} \log \frac{\mathrm{d}\tilde{\eta}}{\mathrm{d}\lambda} - \mathcal{C}_{\tilde{\eta}}\mathcal{D} \log \frac{\mathrm{d}\eta}{\mathrm{d}\lambda} \right\|_r \right\|_{L^2(\nu)}.$$

In the setting of Example 4.1, we can show that $d_{\mathrm{pF},\nu}(\tilde{\eta}\|\eta) = \mathcal{W}_2(\eta, \tilde{\eta})$ (see Appendix C). For a distribution η, let $\mu_\eta(\boldsymbol{x}) := \mathbb{E}_{f\sim\eta}[f(\boldsymbol{x})]$ and $k_\eta(\boldsymbol{x}, \boldsymbol{x}') := \mathbb{E}_{f\sim\eta}[(f(\boldsymbol{x}) - \mu_\eta(\boldsymbol{x}))(f(\boldsymbol{x}') - \mu_\eta(\boldsymbol{x}'))]$ denote the mean and covariance functions associated with η. More generally, we have the following powerful result:

Theorem 4.3. *Let η, $\tilde{\eta}$, and ν denote probability measures on the measurable space $(\mathbb{H}_r, \mathcal{B})$, all abso-*

---

[2] We use "divergence" instead of "distance" because the presence of the covariance operator $\mathcal{C}_{\tilde{\eta}}$ means that the pF divergence is not symmetric in its arguments.

lutely continuous with respect to a base measure λ. Assume η, $\tilde{\eta}$, and λ are Gaussian measures. Let $\varepsilon := \|\mathrm{d}\eta/\mathrm{d}\nu\|_\infty^{1/2}\, d_{\mathrm{pF},\nu}(\tilde{\eta}\|\eta)$ and $\underline{k}(\boldsymbol{x},\boldsymbol{x}) := k_{\tilde{\eta}}(\boldsymbol{x},\boldsymbol{x}) \wedge k_\eta(\boldsymbol{x},\boldsymbol{x})$. If $\varepsilon < \infty$, then $\mathcal{W}_2(\eta,\tilde{\eta}) \leq \varepsilon$ and $\forall \boldsymbol{x} \in \mathcal{X}$,

$$|\mu_{\tilde{\eta}}(\boldsymbol{x}) - \mu_\eta(\boldsymbol{x})| \leq r(\boldsymbol{x},\boldsymbol{x})^{1/2}\varepsilon$$

$$|k_{\tilde{\eta}}(\boldsymbol{x},\boldsymbol{x})^{1/2} - k_\eta(\boldsymbol{x},\boldsymbol{x})^{1/2}| \leq \sqrt{6}\,r(\boldsymbol{x},\boldsymbol{x})^{1/2}\varepsilon$$

$$|k_{\tilde{\eta}}(\boldsymbol{x},\boldsymbol{x}) - k_\eta(\boldsymbol{x},\boldsymbol{x})| \leq 3\,r(\boldsymbol{x},\boldsymbol{x})^{1/2}\underline{k}(\boldsymbol{x},\boldsymbol{x})^{1/2}\varepsilon \\ + 6\,r(\boldsymbol{x},\boldsymbol{x})\varepsilon^2,$$

and, for any $\phi : \mathbb{H} \to \mathbb{R}$ such that $|\mathcal{D}\phi| \leq L < \infty$,

$$|\mathbb{E}_{\tilde{f}\sim\tilde{\eta}}[\phi(\tilde{f})] - \mathbb{E}_{f\sim\eta}[\phi(f)]| \leq L\epsilon.$$

The proof of Theorem 4.3 (found in Appendix D) does not require η and $\tilde{\eta}$ to be Gaussian measures, but for simplicity we have stated it for that case. As long as ν is an over-approximation of η then we can use the ν-pF divergence to control the 2-Wasserstein distance. By "over-approximation" we mean that ν has heavier tails than η so $\|\mathrm{d}\eta/\mathrm{d}\nu\|_\infty$ is bounded. A bound on the 2-Wasserstein distance between η and $\tilde{\eta}$ implies control on the difference in their mean and covariance functions. The factors of $r(\boldsymbol{x},\boldsymbol{x})^{1/2}$ account for the structure of the functions sampled from the GP.

In our application of Theorem 4.3 we will take $\eta = \pi$, the exact posterior; $\tilde{\eta} = \tilde{\pi}$, the approximate posterior; and $\lambda = \pi_0$, the prior. Therefore $\log\frac{\mathrm{d}\pi}{\mathrm{d}\pi_0}(f) = \mathcal{L}(f)$, the exact log-likelihood, and $\log\frac{\mathrm{d}\tilde{\pi}}{\mathrm{d}\pi_0}(f) = \tilde{\mathcal{L}}(f)$, the approximate log-likelihood. In order to make use of the pF divergence algorithmically, we must choose three free parameters: (1) the reproducing kernel r, (2) the auxiliary distribution ν, and (3) a family of log-likelihood approximations $\tilde{\mathcal{L}}$. We address these choices in Section 4.1, Section 4.2, and Section 5, respectively.

Remark 4.4. A direct computation of the constant ε in Theorem 4.3 is difficult in practice. However, it could be approximated by taking $\nu = \eta$ and using an importance sampling approach to estimate $d_{\mathrm{pF},\eta}(\tilde{\eta}\|\eta)$.

4.1 Choosing the reproducing kernel

The seemingly natural choice for the reproducing kernel r is k. However, $\mathbb{H}_k$ is not suitable because if $\mathbb{H}_k$ is infinite-dimensional (as is almost always the case in practice), then for $f \sim \mathrm{GP}(0,k)$, $\mathbb{P}\{f \in \mathbb{H}_k\} = 0$ (see Lukic and Beder (2001) and Rasmussen and Williams (2006, Ch. 6.1)). Therefore we must choose some $r \neq k$ for which $\mathbb{P}\{f \in \mathbb{H}_r\} = 1$. The following result (proved in Appendix E) greatly simplifies the issue:

Proposition 4.5. *Suppose $\mathbb{H}_k$ has an orthonormal basis $(e_j)_{j\geq 1}$ such that for any $\boldsymbol{x} \in \mathcal{X}$, $|e_j(\boldsymbol{x})| = o(j^{-1})$ as $j \to \infty$. Then for any finite set of points $\boldsymbol{S}$ and $\epsilon > 0$, there exists a kernel r that satisfies $\mathbb{P}\{f \in \mathbb{H}_r\} = 1$ and $\max_{\boldsymbol{x},\boldsymbol{x}'\in\boldsymbol{S}} |k(\boldsymbol{x},\boldsymbol{x}') - r(\boldsymbol{x},\boldsymbol{x}')| < \epsilon$.*

The decay condition $|e_j(\boldsymbol{x})| = o(j^{-1})$ is satisfied for Gaussian kernels (Steinwart and Christmann, 2008) and we suspect for Matérn kernels as well. Since we can take ϵ arbitrarily small (e.g. smaller than floating point error), as a practical matter we can choose $r = k$. Furthermore, since in the GP regression setting we only need to evaluate the posterior approximation at a finite number of locations (say, the training points $\boldsymbol{X}$ and some additional test locations $\boldsymbol{X}^*$), the restriction that $\boldsymbol{S}$ be a finite set is not problematic. Hence, we use $r = k$ in our experiments. In the kernel regression literature, our current approach is most closely related to fixed-design error bounds (Cortes et al., 2010; El Alaoui and Mahoney, 2015); but choosing $r \neq k$ might allow more powerful generalization guarantees (Caponnetto and De Vito, 2007; Rudi et al., 2015).

4.2 Choosing the auxiliary distribution

To apply Theorem 4.3, the auxiliary distribution ν must satisfy $\|\mathrm{d}\pi/\mathrm{d}\nu\|_\infty < \infty$. This finiteness condition is easy to achieve by letting $\nu = \mathrm{GP}(\hat{\mu},\hat{k})$ and constructing $\hat{\mu}$ and $\hat{k}$ via a simple subset of datapoints approximation (Rasmussen and Williams, 2006, Ch. 8.3.3). Let $\hat{\boldsymbol{X}} \subset \boldsymbol{X}$ be a random subset of the input locations of size M', and take ν to be the GP posterior conditional on observing $\hat{\boldsymbol{y}}$, the subset of $\boldsymbol{y}$ corresponding to $\hat{\boldsymbol{X}}$. Therefore the auxiliary mean and covariance functions are $\hat{\mu}(\boldsymbol{x}) = k_{\boldsymbol{x}\hat{\boldsymbol{X}}}[k_{\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}} + \sigma^2 I]^{-1}\hat{\boldsymbol{y}}$ and $\hat{k}(\boldsymbol{x},\boldsymbol{x}') = k(\boldsymbol{x},\boldsymbol{x}') - k_{\boldsymbol{x}\hat{\boldsymbol{X}}}[k_{\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}} + \sigma^2 I]^{-1}k_{\hat{\boldsymbol{X}}\boldsymbol{x}'}$. Let $\hat{\boldsymbol{D}} = (\hat{\boldsymbol{X}},\hat{\boldsymbol{y}})$ denote the auxiliary data, and let $\mathcal{Z}_{\boldsymbol{D}}$ and $\mathcal{L}_{\boldsymbol{D}}$ denote the marginal likelihood and log-likelihood of the data $\boldsymbol{D}$ (with $\mathcal{Z}_{\hat{\boldsymbol{D}}}$ and $\mathcal{L}_{\hat{\boldsymbol{D}}}$ defined analogously). We can conclude that

$$\begin{aligned}\frac{\mathrm{d}\pi}{\mathrm{d}\nu}(f) &= \frac{\mathcal{Z}_{\hat{\boldsymbol{D}}}e^{\mathcal{L}_{\boldsymbol{D}}(f)-\mathcal{L}_{\hat{\boldsymbol{D}}}(f)}}{\mathcal{Z}_{\boldsymbol{D}}} \\ &= \frac{\mathcal{Z}_{\hat{\boldsymbol{D}}}e^{-\sum_{\boldsymbol{x}_n\in\boldsymbol{X}\setminus\hat{\boldsymbol{X}}}(f(\boldsymbol{x}_n)-y_n)^2/(2\sigma^2)}}{(2\pi\sigma^2)^{\frac{N-M'}{2}}\mathcal{Z}_{\boldsymbol{D}}} \\ &\leq \frac{\mathcal{Z}_{\hat{\boldsymbol{D}}}}{(2\pi\sigma^2)^{\frac{N-M'}{2}}\mathcal{Z}_{\boldsymbol{D}}} < \infty.\end{aligned}$$

Hence Theorem 4.3 applies when the auxiliary distribution is a subset of datapoints approximation.

5 Preconditioned Fisher DTC

In this section we describe how to optimize the DTC inducing point approximation described in Section 2 using the pF divergence. We call the resulting inference algorithm *preconditioned Fisher DTC* (pF-DTC). Recall the DTC log-likelihood for the nth observation:

$$\tilde{\mathcal{L}}_n(f) := -\frac{1}{2\sigma^2}(k_{\boldsymbol{x}_n\tilde{\boldsymbol{X}}}k_{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}}^{-1}f_{\tilde{\boldsymbol{X}}} - y_n)^2 - \frac{1}{2}\log(2\pi\sigma^2).$$
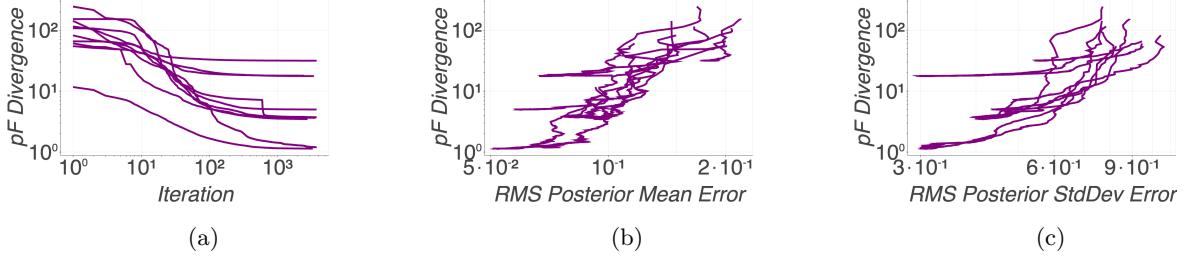
Figure 1: **(a)** Value of the pF-DTC objective on 10 runs for the airfoil data. **(b,c)** Relationship between the pF-DTC objective and the mean/standard deviation error on 10 runs for the airfoil dataset.

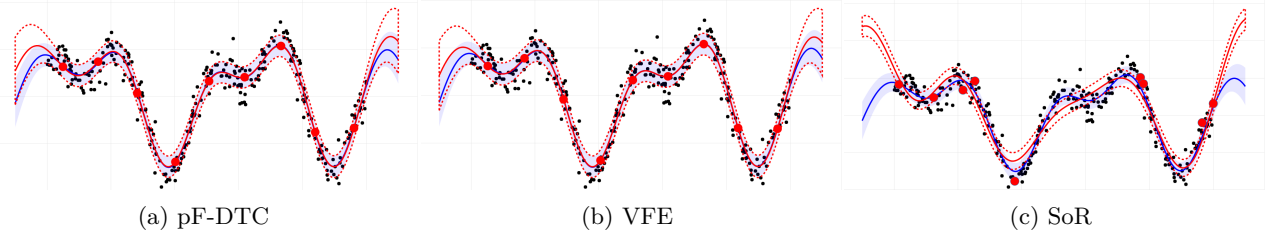(a) pF-DTC   (b) VFE   (c) SoR

Figure 2: Comparison of pF-DTC, VFE, and SoR on the synthetic example with $M = 9$ inducing points. Data is black, exact posterior mean and standard deviation are blue, and approximate fits and inducing points are red.

We consider the case where $r = k$ at the locations of interest, which will simplify formulas (cf. Section 4.1). Let $Q_{\boldsymbol{XX}} := k_{\boldsymbol{X}\tilde{\boldsymbol{X}}} k_{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}}^{-1} k_{\tilde{\boldsymbol{X}}\boldsymbol{X}}$, $\bar{Q}_{\boldsymbol{X}\tilde{\boldsymbol{X}}} := k_{\boldsymbol{X}\tilde{\boldsymbol{X}}} k_{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}}^{-1}$, and $S_{\boldsymbol{XX}} := Q_{\boldsymbol{XX}}(I - (Q_{\boldsymbol{XX}} + \sigma^2 I)^{-1} Q_{\boldsymbol{XX}})^2$. Our main result in this section provides a formula for computing the pF divergence in the sparse GP setting and guarantees that, up to a constant independent of the inducing points, $d_{\mathrm{pF},\nu}(\tilde{\pi}||\pi)$ can be computed efficiently.

Proposition 5.1. *For the DTC log-likelihood approximation, if $\nu = \mathrm{GP}(\hat{\mu}, \hat{k})$, then when $r = k$,*

$$d_{\mathrm{pF},\nu}(\tilde{\pi}||\pi)$$
$$= \underbrace{\mathrm{Tr}((\hat{k}_{\boldsymbol{XX}} + (\hat{\mu}_{\boldsymbol{X}} - \boldsymbol{y})(\hat{\mu}_{\boldsymbol{X}} - \boldsymbol{y})^{\top})(k_{\boldsymbol{XX}} - Q_{\boldsymbol{XX}}))}_{\mathrm{(I)}}$$
$$+ \underbrace{\mathrm{Tr}(\hat{k}_{\boldsymbol{XX}} S_{\boldsymbol{XX}}) + \mathrm{Tr}(\hat{k}_{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}} \bar{Q}_{\boldsymbol{X}\tilde{\boldsymbol{X}}}^{\top} S_{\boldsymbol{XX}} \bar{Q}_{\boldsymbol{X}\tilde{\boldsymbol{X}}})}_{\mathrm{(IIa)}}$$
$$- \underbrace{2\,\mathrm{Tr}(\hat{k}_{\tilde{\boldsymbol{X}}\boldsymbol{X}} S_{\boldsymbol{XX}} \bar{Q}_{\boldsymbol{X}\tilde{\boldsymbol{X}}})}_{\mathrm{(IIb)}}$$
$$+ \underbrace{(\hat{\mu}_{\boldsymbol{X}} - \bar{Q}_{\boldsymbol{X}\tilde{\boldsymbol{X}}} \hat{\mu}_{\tilde{\boldsymbol{X}}})^{\top} S_{\boldsymbol{XX}} (\hat{\mu}_{\boldsymbol{X}} - \bar{Q}_{\boldsymbol{X}\tilde{\boldsymbol{X}}} \hat{\mu}_{\tilde{\boldsymbol{X}}})}_{\mathrm{(III)}} .$$

Furthermore, as long as vector multiplication by $\hat{k}_{\boldsymbol{XX}}$ takes $O(NM)$ time, $d_{\mathrm{pF},\nu}(\tilde{\pi}||\pi) - C(\boldsymbol{X})$ can be computed in $O(NM^2)$ time and $O(NM)$ space, where $C(\boldsymbol{X})$ is a function that does not depend on $\tilde{\boldsymbol{X}}$.

We can interpret the three terms in the expression for $d_{\mathrm{pF},\nu}(\tilde{\pi}||\pi)$ quite naturally. Term (I) measures how well the kernel matrix $k_{\boldsymbol{XX}}$ is approximated by the

Nyström approximation $Q_{\boldsymbol{XX}}$. To interpret term (II) = (IIa) + (IIb), recall that the exact posterior kernel is $k_{\boldsymbol{D}}(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}, \boldsymbol{x}') - k_{\boldsymbol{x}\boldsymbol{X}}[k_{\boldsymbol{XX}} + \sigma^2 I]^{-1} k_{\boldsymbol{X}\boldsymbol{x}'}$. Observing the $(Q_{\boldsymbol{XX}} + \sigma^2 I)^{-1}$ terms in $S_{\boldsymbol{XX}}$, we can see that term (II) measures how well the correction term $k_{\boldsymbol{x}\boldsymbol{X}}[k_{\boldsymbol{XX}} + \sigma^2 I]^{-1} k_{\boldsymbol{X}\boldsymbol{x}'}$ is estimated. Finally, term (III) measures the quality of $\bar{Q}_{\boldsymbol{X}\tilde{\boldsymbol{X}}} \hat{\mu}_{\tilde{\boldsymbol{X}}}$, which is the inducing point approximation to the auxiliary mean function $\hat{\mu}_{\boldsymbol{X}}$. If we had to compute the terms in $d_{\mathrm{pF},\nu}(\tilde{\pi}||\pi)$ involving $k_{\boldsymbol{XX}}$, then the time and space complexity would be, respectively, $O(MN^2)$ and $O(N^2)$. However, the terms that include $k_{\boldsymbol{XX}}$ do not depend on the inducing points, so we absorb them into $C(\boldsymbol{X})$.

6 Experiments

Our theory shows that pF-DTC can provide guarantees on the quality of the approximate GP posterior means and variances that practitioners typically report (Section 1). We now check empirically that our method yields competitive estimates of exact GP means and variances in practice. We compare pF-DTC to to variational DTC (VFE) (Titsias, 2009), subset of regressors (SoR) (Smola and Bartlett, 2000), and exact GP inference with a random subsample of the data (subsample). To compare to the exact GP as ground truth, all experiments use the same kernel (squared exponential kernel with separate length scales for each dimension) and identical hyperparameters. To that end, unless noted otherwise, we fix the kernel hyperparameters and the observation noise σ^2, all of which we learn with

a pilot run of VFE using 200 fixed randomly chosen inducing points. We found that if we optimized the hyperparameters for pF-DTC and VFE after inducing point optimization, they remained relatively stable. For pF-DTC we take $r = k$, as justified by Proposition 4.5. Although constructing ν from a subset of data is currently the most justified choice theoretically (see Section 4.2), preliminary experiments showed better empirical performance using SoR with a small random subset of the data.

We used one synthetic and five real datasets with dimensionality ranging from 1 to 11. In order to run the exact GP on all datasets we subsampled the original airline delays dataset used by Hensman et al. (2013) down to 10,000 observations. The remaining non-synthetic datasets were obtained from the UCI Machine Learning Repository. We held out the maximum of 1,000 observations or 20% of each dataset for testing and used the remainder for training (training set sizes were between 1,000 and 8,000). See Table 1 in Appendix A for full details. We repeated all experiments 10 times.

Behavior of the pF divergence. The pF-DTC objective is non-convex, so we first check that we can effectively optimize it. As seen in Fig. 1a, for the airfoil dataset, optimizing the locations of the 200 inducing points substantially reduces the size of the pF divergence, though clearly we are finding only local optima. The bounds in Theorem 4.3 suggest a linear relationship between the pF divergence and the errors in estimates of the posterior mean $\mu_{\boldsymbol{D}}(\boldsymbol{x})$ and standard deviation $k_{\boldsymbol{D}}(\boldsymbol{x}, \boldsymbol{x})^{1/2}$. Figs. 1b and 1c show an approximately linear relationship does in fact hold in both cases.

Mean and Variance Estimates. We first consider a simple one-dimensional synthetic example to compare pF-DTC to VFE and SoR. Fig. 2 shows that, with 9 inducing points, pF-DTC and VFE produce excellent, almost identical fits, while SoR performs substantially worse at estimating both the posterior mean and standard deviation. Bauer et al. (2016) found similarly poor performance for the fully independence training conditional (FITC) method of Snelson and Ghahramani (2005).

On more complex problems, we consider root mean squared error (RMSE) of posterior mean and standard deviation estimates at all held-out test locations, as we vary the number of inducing points. Fig. 5 confirms that pF-DTC is competitive with VFE, validating the practicality of the pF divergence as an objective for approximate inference. VFE shows better performance for larger numbers of inducing points once the error dips below 10^{-3}, which we suspect is related to numerical issues. SoR shows surprisingly good performance on standard deviation error but worse mean estimation—

particularly on abalone. As expected, subsampling performed poorly.

KL divergence and comparison of objective functions. In Section 3 we argued that KL divergence is a suboptimal way to measure posterior approximation quality when the goal is good posterior mean and standard deviation estimates. We here show that this issue arises in practice. Figs. 4, A.1 and A.2 show the KL divergences for all four approximations. We see that VFE sometimes yields small KL divergence but worse posterior mean and standard deviation estimates. Similarly, we see that pF-DTC can exhibit larger KL divergence values but very good posterior mean and standard deviation estimates.

Predictive Performance. While our focus in this paper is on posterior mean and variance estimates, we also used the held-out test data to check RMSE predictive performance. As shown in Fig. 3, pF-DTC, SoR, and VFE performed quite similarly, with subsampling providing worse predictive accuracy.

Computation. In over 80% of our 60 experiments (10 experiments per dataset), pF-DTC used less computation than VFE. On average, VFE was 4.5 times slower than pF-DTC. Because they did not involve any optimization, SoR and subsample were orders of magnitude faster.

7 Discussion

In this paper we have developed an approach to scalable Gaussian process regression using a novel objective, the preconditioned Fisher divergence, which we show bounds the 2-Wasserstein distance. We were motivated by the need to guarantee the finite-data accuracy of posterior mean and covariance function estimates. Empirically we showed that using the pF divergence with a DTC likelihood approximation provides competitive mean and variance estimates with state-of-the-art approaches.

The current work demonstrates the feasibility of using the pF divergence, but there are many important directions for future work. The first is to scale to larger datasets, e.g. using stochastic optimization techniques similar in spirit to SVI for GPs (Hensman et al., 2013). A second direction is to use the pF divergence objective with other likelihood approximations. Structured kernel approximations are of particular interest since these are both fast to compute and can often better capture multi-scale structure (Wilson and Nickisch, 2015; Gardner et al., 2018). Finally, it remains to efficiently utilize the preconditioned Fisher divergence for other likelihoods, such as in GP classification.
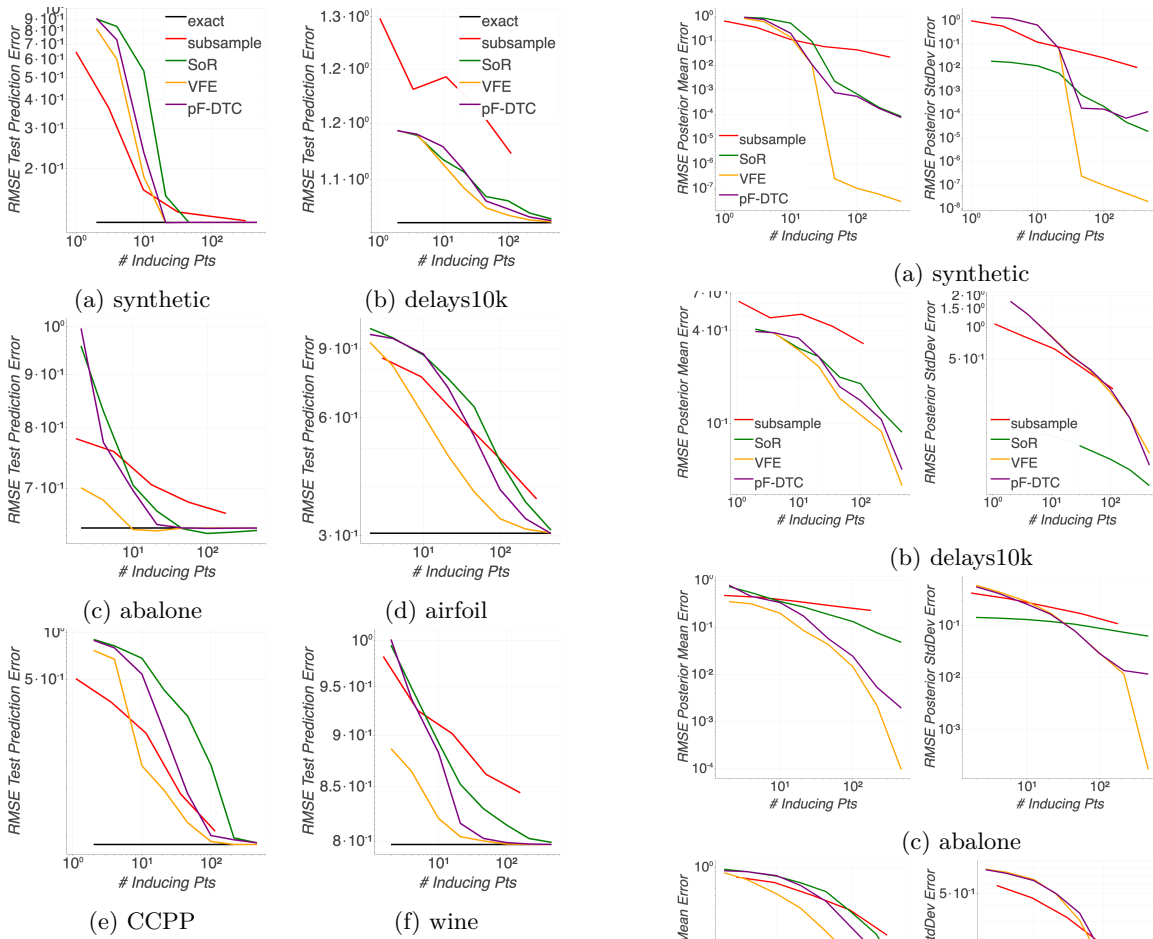
(a) synthetic

(b) delays10k

(c) abalone

(d) airfoil

(e) CCPP

(f) wine

Figure 3: Predictive performance as measured by root mean squared error.
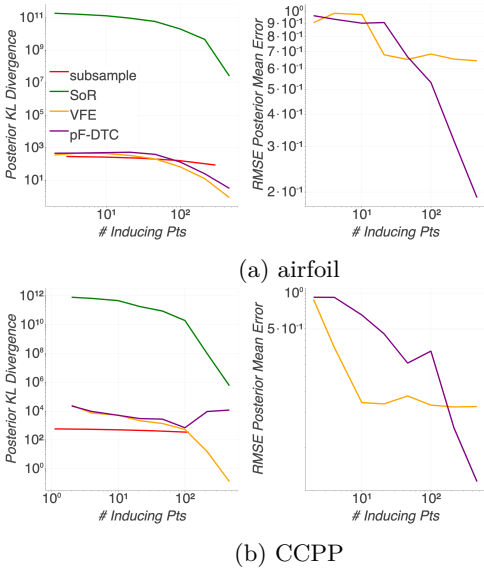

(a) airfoil

(b) CCPP

Figure 4: KL divergences of the approximate posteriors and root mean squared error of the approximate posteriors for the VFE and pF-DTC trials with the smallest objective values.
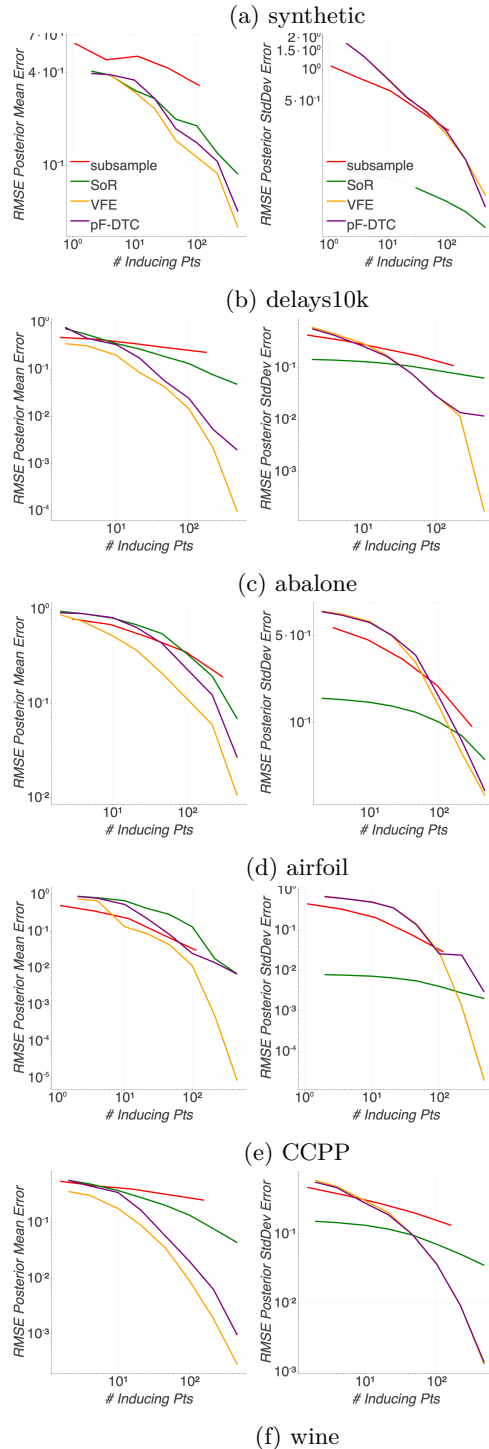

(a) synthetic

(b) delays10k

(c) abalone

(d) airfoil

(e) CCPP

(f) wine

Figure 5: Root mean squared error of the approximate posterior mean (left) and standard deviation (right) at the held-out test locations.

References

M. Bauer, M. van der Wilk, and C. E. Rasmussen. Understanding Probabilistic Sparse Gaussian Process Approximations. In *Advances in Neural Information Processing Systems*, 2016.

T. D. Bui, J. Yan, and R. E. Turner. A Unifying Framework for Gaussian Process Pseudo-Point Approximations using Power Expectation Propagation. *Journal of Machine Learning Research*, 18:1–72, Oct. 2017.

T. Campbell and T. Broderick. Bayesian Coreset Construction via Greedy Iterative Geodesic Ascent. In *International Conference on Machine Learning*, pages 1–13, 2018.

T. Campbell and T. Broderick. Automated Scalable Bayesian Inference via Hilbert Coresets. *Journal of Machine Learning Research*, 20:1–38, 2019.

A. Caponnetto and E. De Vito. Optimal Rates for the Regularized Least-Squares Algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

C. Cortes, M. Mohri, and A. Talwalkar. On the Impact of Kernel Approximation on Learning Accuracy. In *International Conference on Artificial Intelligence and Statistics*, 2010.

A. S. Dalalyan and A. G. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *arXiv.org*, math.ST, Sept. 2017.

A. G. de G Matthews, J. Hensman, R. E. Turner, and Z. Ghahramani. On Sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In *International Conference on Artificial Intelligence and Statistics*, 2016.

Y. Ding, R. Kondor, and J. Eskreis-Winkler. Multiresolution Kernel Approximation for Gaussian Process Regression. In *Advances in Neural Information Processing Systems*, 2017.

A. El Alaoui and M. W. Mahoney. Fast Randomized Kernel Methods With Statistical Guarantees. In *Advances in Neural Information Processing Systems*, 2015.

J. R. Gardner, G. Pleiss, R. Wu, K. Q. Weinberger, and A. G. Wilson. Product Kernel Interpolation for Scalable Gaussian Processes. In *International Conference on Artificial Intelligence and Statistics*, 2018.

R. B. Gramacy and H. K. H. Lee. Bayesian Treed Gaussian Process Models With an Application to Computer Modeling. *Journal of the American Statistical Association*, 103(483):1119–1130, Jan. 2012.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian Processes for Big Data. In *Uncertainty in Artificial Intelligence*, 2013.

J. Hensman, A. G. de G Matthews, and Z. Ghahramani. Scalable Variational Gaussian Process Classification. In *International Conference on Artificial Intelligence and Statistics*, 2015.

J. H. Huggins and J. Zou. Quantifying the accuracy of approximate diffusions and Markov chains. In *International Conference on Artificial Intelligence and Statistics*, 2017.

J. H. Huggins, R. P. Adams, and T. Broderick. PASS-GLM: polynomial approximate sufficient statistics for scalable Bayesian GLM inference. In *Advances in Neural Information Processing Systems*, 2017.

A. Hyvarinen. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

I. A. Ibragimov and Y. A. Rozanov. *Gaussian Random Processes*. Springer-Verlag, New York, NY, 1978.

P. Izmailov, A. Novikov, and D. Kropotov. Scalable Gaussian Processes with Billions of Inducing Inputs via Tensor Train Decomposition. In *International Conference on Artificial Intelligence and Statistics*, 2018.

O. Johnson and A. R. Barron. Fisher information inequalities and the central limit theorem. *Probability Theory and Related Fields*, 129(3):391–409, July 2004.

M. Katzfuss. A Multi-Resolution Approximation for Massive Spatial Datasets. *Journal of the American Statistical Association*, 112(517):201–214, Mar. 2017.

C. G. Kaufman and S. R. Sain. Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Analysis*, 5(1):123–149, Mar. 2010.

C. G. Kaufman, D. Bingham, S. Habib, K. Heitmann, and J. A. Frieman. Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology. *The Annals of Applied Statistics*, 5(4):2470–2492, Dec. 2011.

C. Ley and Y. Swan. Stein's density approach and information inequalities. *Electronic Communications in Probability*, 18(7):1–14, 2013.

M. N. Lukic and J. H. Beder. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society*, 353(10):3945–3969, 2001.

H. E. Ogden. On asymptotic validity of naive inference with an approximate likelihood. *Biometrika*, 104(1): 153–164, 2017.

M. A. Osborne, D. Duvenaud, R. Garnett, C. E. Rasmussen, S. J. Roberts, and Z. Ghahramani. Active learning of model evidence using Bayesian quadrature. *Advances in Neural Information Processing Systems*, 2012.

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

C. E. Rasmussen and Z. Ghahramani. Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 489–496, 2002.

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

A. Rudi, R. Camoriano, and L. Rosasco. Less is More: Nyström Computational Regularization. In *Advances in Neural Information Processing Systems*, 2015.

J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and Analysis of Computer Experiments. *Statistical Science*, 4(4):409–423, Nov. 1989.

M. W. Seeger, C. K. I. Williams, and N. D. Lawrence. Fast Forward Selection to Speed Up Sparse Gaussian Process Regression. In *International Conference on Artificial Intelligence and Statistics*, 2003.

A. J. Smola and P. L. Bartlett. Sparse Greedy Gaussian Process Regression. In *Advances in Neural Information Processing Systems*, 2000.

E. Snelson and Z. Ghahramani. Sparse Gaussian Processes using Pseudo-inputs. In *Advances in Neural Information Processing Systems*, 2005.

J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems*, 2012.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, A. Hyvarinen, and R. Kumar. Density Estimation in Infinite Dimensional Exponential Families. *Journal of Machine Learning Research*, 18:1–59, 2017.

I. Steinwart and A. Christmann. *Support Vector Machines.* Springer, New York, NY, 2008.

M. Titsias. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In *International Conference on Artificial Intelligence and Statistics*, 2009.

A. G. Wilson and H. Nickisch. Kernel Interpolation for Scalable Structured Gaussian Processes (KISS-GP). In *International Conference on Machine Learning*, 2015.