arXiv:1903.03571v1 [stat.ML] 8 Mar 2019

Rates of Convergence for Sparse Variational Gaussian Process Regression

David R. Burt [1] **Carl E. Rasmussen** [1] **Mark van der Wilk** [2]

Abstract

Excellent variational approximations to Gaussian process posteriors have been developed which avoid the $\mathcal{O}\left(N^3\right)$ scaling with dataset size N. They reduce the computational cost to $\mathcal{O}\left(NM^2\right)$, with $M \ll N$ being the number of *inducing variables*, which summarise the process. While the computational cost seems to be linear in N, the true complexity of the algorithm depends on how M must increase to ensure a certain quality of approximation. We address this by characterising the behavior of an upper bound on the KL divergence to the posterior. We show that with high probability the KL divergence can be made arbitrarily small by growing M more slowly than N. A particular case of interest is that for regression with normally distributed inputs in D-dimensions with the popular Squared Exponential kernel, $M = \mathcal{O}(\log^D N)$ is sufficient. Our results show that as datasets grow, Gaussian process posteriors can truly be approximated cheaply, and provide a concrete rule for how to increase M in continual learning scenarios.

1. Introduction

Gaussian processes (GPs) [Rasmussen & Williams, 2006] are distributions over functions that are convenient priors in Bayesian models. They can be seen as infinitely wide neural networks [Neal, 1996], and are particularly popular in regression models, as they produce good uncertainty estimates, and have closed-form expressions for the posterior and marginal likelihood. The computational cost of exact computation of these quantities is their most well-known practical drawback, as it scales as $\mathcal{O}\left(N^3\right)$ in time and $\mathcal{O}\left(N^2\right)$ in memory where N is the number of training examples. Low-rank approximations [Quiñonero Candela & Rasmussen, 2005] choose a small set of M *inducing variables* which summarise the entire posterior, reducing the cost to $\mathcal{O}\left(NM^2 + M^3\right)$ time and $\mathcal{O}\left(NM + M^2\right)$ memory.

While the computational cost of adding inducing variables is well understood, results on how many are needed to achieve a good approximation are lacking. As the dataset size increases, we cannot expect to keep the capacity of the approximation constant without the quality deteriorating. This dependence of M on N is hidden in the computational scaling bounds above. Taking into account the rate at which M needs to increase with N to achieve a particular approximation accuracy determines a more realistic sense of the costs of scaling Gaussian processes. If M is required to scale linearly with N, low-rank approximation yields a constant factor improvement, while a slower rate ensures asymptotically better scaling.

The KL divergence to the true posterior is a common metric for assessing the quality of an approximate posterior, and is minimized by variational methods. Approximate GPs are commonly trained using variational inference [Titsias, 2009], which minimize the KL divergence between the approximate and full process posteriors [Matthews et al., 2016]. In this work, we choose the KL divergence as our metric for the approximate posterior's quality. We show that under intuitive assumptions the number of inducing variables only needs to grow at a sublinear rate to make the evidence lower bound (ELBO) tight to the marginal likelihood, and for the KL between approximation and posterior to go to zero. This shows that very sparse approximations can likely be found for large datasets, without introducing much bias into the hyperparameters selected using the ELBO, and with an approximate posterior that accurately reflects the predictions and uncertainties in the true posterior.

The core idea of our proof is to use upper bounds on the KL divergence that depend on the quality of a Nyström approximation to the data covariance matrix. Using existing results, we show this error can be understood in terms of the spectrum of an infinite-dimensional integral operator. Specialized to the case of stationary prior kernels, our main result proves that the greater the smoothness of functions in the prior and the greater the concentration of observations in input space, the sparser an approximation can be made.

[1] University of Cambridge, Cambridge, UK [2] PROWLER.io, Cambridge, UK. Correspondence to: David R. Burt <drb62@cam.ac.uk>.

Main Results Our main results assume that the training inputs are drawn i.i.d from some fixed distribution. We prove bounds of the form,

$$KL(Q\|\hat{P}) \leq \mathcal{O}\left(\frac{g(M,N)}{\delta\sigma_n^4} f(M)\right)$$

with probability at least $1-\delta$. The function $f(M)$ is rapidly decaying, and depends on both the kernel and input distribution. The function g is either quadratic or linear in N depending on our assumption on training outputs and is either linear or constant in M depending on the inducing variables used for inference. Bounds of this form are proven for a certain set of inducing variables using spectral knowledge of the prior in Section 3 and for inducing points in Section 4.

2. Background and Notation

2.1. Gaussian Process Regression

We are concerned with the problem of Gaussian process regression. Namely, we have observed *training data*, $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ with $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathbb{R}$. Our goal is to predict outputs y^* for new inputs $\mathbf{x}^*$ while taking into account the uncertainty we have about $f(\cdot)$ due to the limited size of the training set. We follow a Bayesian approach by placing a prior over f, and a likelihood to relate f to the observed data through some observation noise. Our model is

$$f \sim \mathcal{GP}(\nu(\cdot), k(\cdot,\cdot)), \quad y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_n^2),$$

where $\nu : \mathcal{X} \to \mathbb{R}$ is the *mean function* and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the *covariance function*. We take $\nu \equiv 0$; the general case can be derived similarly after first centering the process. We use the posterior for making predictions, and the marginal likelihood for selecting hyperparameters, both of which have closed-form expressions for this model [Rasmussen & Williams, 2006]. The marginal likelihood is of particular interest to us, as the quality of its approximation and our posterior approximation is linked. Its form is

$$\mathcal{L} = -\frac{1}{2}\mathbf{y}^\mathsf{T}\mathbf{K}_n^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K}_n| - c, \tag{1}$$

where $c = \frac{N}{2}\log(2\pi)$, $\mathbf{K}_n = \mathbf{K_{ff}} + \sigma_n^2\mathbf{I}$, and $\mathbf{K_{ff}}$ denotes the data covariance matrix with entries $[\mathbf{K_{ff}}]_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$.

2.2. Sparse Variational Gaussian Process Regression

While all quantities of interest have analytic expressions, their computation is intractable for large datasets due to the $\mathcal{O}(N^3)$ time complexity of the determinant and inverse. Numerous approaches have been proposed (e.g. Quiñonero Candela & Rasmussen [2005] or Rahimi & Recht [2008]) to avoid this cost, which rely on a low-rank approximation to $\mathbf{K_{ff}}$ that allows the necessary matrix inverse to be computed in $\mathcal{O}(NM^2)$ where M is the rank of the approximating matrix.

We consider the variational framework developed by Titsias [2009], which minimizes a KL divergence [Matthews et al., 2016] to the true posterior process from an approximate GP of the form

$$\mathcal{GP}\left(\mathbf{k}_{\cdot\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\boldsymbol{\mu}, k_{\cdot\cdot} + \mathbf{k}_{\cdot\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}(\boldsymbol{\Sigma} - \mathbf{K}_{\mathbf{uu}})\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{k}_{\mathbf{u}\cdot}\right), \tag{2}$$

where $[\mathbf{k}_{\mathbf{u}\cdot}]_i = k(\cdot, \mathbf{z}_i)$, $[\mathbf{K_{uf}}]_{m,i} := k(\mathbf{z}_m, \mathbf{x}_i)$ and $[\mathbf{K_{uu}}]_{m,n} := k(\mathbf{z}_m, \mathbf{z}_n)$. The properties of this variational distribution are determined by specifying the density of the function values $\mathbf{u} \in \mathbb{R}^M$ at *inducing points* $Z = \{\mathbf{z}_m\}_{m=1}^M$ to be $q(\mathbf{u}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The variational parameters consist of Z, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$.

Hensman et al. [2013] suggest optimizing all the variational parameters as free parameters, so that minibatches over the data can be used. The formulation originally developed by Titsias [2009] found the minimum of the convex optimization problem for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ explicitly, at the cost of requiring a full sweep over the training data. This resulted in the following evidence lower bound (ELBO):

$$\mathcal{L}_{\text{lower}} = -\frac{1}{2}\mathbf{y}^\mathsf{T}\mathbf{Q}_n^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{Q}_n| - c - \frac{t}{2\sigma_n^2} \tag{3}$$

where $\mathbf{Q}_n = \mathbf{Q_{ff}} + \sigma_n^2\mathbf{I}$, $\mathbf{Q_{ff}} = \mathbf{K}_{\mathbf{uf}}^\mathsf{T}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K_{uf}}$ and $t = \mathrm{Tr}(\mathbf{K_{ff}} - \mathbf{Q_{ff}})$. Matthews et al. [2016] showed that the KL divergence between the approximate GP posterior (eq. 2) and the true posterior process is equal to $\mathcal{L} - \mathcal{L}_{\text{lower}}$:

$$\mathrm{KL}\left(Q\|\hat{P}\right) = \mathcal{L} - \mathcal{L}_{\text{lower}}\,. \tag{4}$$

Instead of maximizing the intractable marginal likelihood (eq. 1), this framework suggests to jointly maximize the ELBO w.r.t. the variational and hyperparameters. This comes at the cost of introducing some bias in the hyperparameter estimation [Turner & Sahani, 2011], notably the overestimation of the σ_n^2 [Bauer et al., 2016]. Adding extra inducing points always reduces the KL gap [Titsias, 2009; Matthews, 2016; Bauer et al., 2016], which allows the bias to be practically eliminated when enough inducing variables are used.

2.3. Interdomain Inducing Features

The posterior parameterized in Equation 2 does not necessarily need to be parameterized by specifying the density $q(\mathbf{u})$ of function values of the GP. Lázaro-Gredilla & Figueiras-Vidal [2009] showed that one can specify $q(\mathbf{u})$ on integral transformations of $f(\cdot)$. Using these *interdomain* inducing variables can lead to sparser representations, or computational benefits [Hensman et al., 2018]. Interdomain inducing variables are defined by

$$u_m = \int_{\mathcal{X}} f(\mathbf{x}) g(\mathbf{x}; \mathbf{z}_m) dx\,.$$

When $g(\mathbf{x}; \mathbf{z}_m) = \delta(\mathbf{x} - \mathbf{z}_m)$ the u_m are inducing points. Interdomain features are practical, as they only require replacing $\mathbf{k}_{\mathbf{u}\cdot}$ and $\mathbf{K}_{\mathbf{uu}}$ in Equation 2 with transforms of the original kernel k. We will investigate particular interdomain transformations with interesting convergence properties.

2.4. Upper Bounds on the Marginal Likelihood

Combined with Equation 3, an upper bound on Equation 1 can show the KL divergence is small. This indicates inference has been successful and hyperparameter estimates are likely to have little bias. Titsias [2014] introduced an upper bound that can be computed in $\mathcal{O}(NM^2)$:

$$\mathcal{L}_{\text{upper}} := -c - \frac{1}{2}\log(|\mathbf{Q}_n|) - \frac{1}{2}\mathbf{y}^{\mathsf{T}}(\mathbf{Q}_n + t\mathbf{I})^{-1}\mathbf{y}. \quad (5)$$

This gives a *data-dependent* upper bound, that can be computed after seeing the data.

2.5. Spectral Properties of the Covariance Matrix

While for specific, small datasets the properties of the covariance matrix can be analyzed numerically, in order to understand these quantities for a typical dataset and for large datasets, we need another approach. The *covariance operator,* $\mathcal{K}$ is an operator on function spaces that captures the limiting properties of $\mathbf{K}_{\mathbf{ff}}$ for large N. It is defined by

$$\mathcal{K}g(\mathbf{x}') = \int_{\mathcal{X}} g(\mathbf{x})k(\mathbf{x}, \mathbf{x}')p(\mathbf{x})d\mathbf{x}, \quad (6)$$

where $p(\mathbf{x})$ is some (unknown) probability density from which the inputs are assumed to be drawn. We assume that $\mathcal{K}$ is compact, which is the case if p is a probability density and $k(\mathbf{x}, \mathbf{x}')$ is bounded. Under this assumption, the spectral theorem tells us that $\mathcal{K}$ has only discrete eigenvalues. The finite sequence of eigenvalues of $\frac{1}{N}\mathbf{K}_{\mathbf{ff}}$ approach the infinite sequence of eigenvalues of $\mathcal{K}$ [Koltchinskii et al., 2000], and in a certain sense the eigenspaces of $\frac{1}{N}\mathbf{K}_{\mathbf{ff}}$ approach those of $\mathcal{K}$, [Koltchinskii, 1998]. Mercer [1909] tells us that we can write our kernel as,

$$k(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^{\infty} \lambda_m \phi_m(\mathbf{x})\phi_m(\mathbf{x}'), \quad (7)$$

where the $(\lambda_m, \phi_m)_{i=1}^{\infty}$ are eigenvalue-eigenfunction pairs of the operator $\mathcal{K}$. Additionally,

$$\sum_{m=1}^{\infty} \lambda_m < \infty.$$

We assume without loss of generality the eigenfunctions are orthonormal with respect to $L^2(\mathcal{X})_p$.

2.6. Selecting the Number of Inducing Variables

Sections 2.2 and 2.4 gave lower and upper bounds to the marginal likelihood for a specific dataset. Their difference upper bounds the KL divergence (eq. 4). These results imply procedures for selecting the number of inducing points to balance computational cost and approximation accuracy. Based on the lower bound alone, common advice is to stop increasing M when the lower bound no longer improves, which is necessary but not sufficient for the bound to be tight. If the upper bound is taken into consideration, a good approximation is guaranteed when the difference between the bounds converges to zero. When performing hyperparameter selection, we also need to guarantee that there are no other settings of the hyperparameters which have higher marginal likelihood than our current best estimate. In this situation, we also require the upper bound for candidate hyperparameters to be below the current lower bound. In practice eliminating all choices of hyperparameters using this approach is not feasible without additionally taking a prior on values of hyperparameters, as the upper bound is very loose for certain hyperparameter choices, notably small choices of likelihood noise.

These procedures rely on bounds computed for a given dataset. While practically useful, they do not make predictions for a wide variety of tasks. In the following, we focus on *a priori* bounds, and asymptotic behavior as $N \to \infty$ and M grows as a function of N. These bounds provide guarantees of how the variational method scales computationally for *any* dataset satisfying intuitive conditions. This is particularly important for continual learning scenarios, where we incrementally observe more data. With our a priori results we can guarantee that the growth in required computation will not exceed a certain rate.

3. Bounds on the KL Divergence for Eigenfunction Inducing Features

In this section, we prove a priori and asymptotic bounds on the KL divergence for regression using a certain type of inducing variables. These inducing variables rely on spectral knowledge of the covariance matrix or the associated operator, and have certain near optimal properties in terms of minimizing the KL divergence in expectation over training outputs generated according to the prior generative model. The lemmas and proofs in this section form the basis for bounds on the KL divergence for inducing points (Section 4).

3.1. A posteriori Bounds on The KL divergence

We first consider *a posteriori* bounds on the KL divergence that hold for any $\mathbf{y}$. These are derived by looking at the difference between $\mathcal{L}_{\text{upper}}$ and $\mathcal{L}_{\text{lower}}$, and are primarily useful

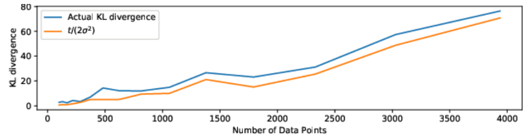

Figure 1. As N increases for fixed M the expected KL divergence increases. $t/2\sigma_n^2$ is a lower bound for the expected value over the KL divergence when $\mathbf{y}$ is generated according to our prior model.

in that they only depend on $\mathbf{y}$ through its norm. This makes them amenable to use in asymptotic statements derived in later sections.

Lemma 1. *Let $\widetilde{\mathbf{K}}_{\mathbf{ff}} = \mathbf{K}_{\mathbf{ff}} - \mathbf{Q}_{\mathbf{ff}}$. We use t to denote the trace of this matrix and $\widetilde{\lambda}_{max}$ to denote its largest eigenvalue. Then,*

$$\begin{aligned}\mathrm{KL}\left(Q\|\hat{P}\right) &\leq \frac{1}{2\sigma_n^2}\left(t + \frac{\widetilde{\lambda}_{max}\|\mathbf{y}\|_2^2}{\sigma_n^2 + \widetilde{\lambda}_{max}}\right) \\ &\leq \frac{t}{2\sigma_n^2}\left(1 + \frac{\|\mathbf{y}\|_2^2}{\sigma_n^2 + t}\right).\end{aligned}$$

The proof bounds the difference between a refinement of $\mathcal{L}_{\text{upper}}$ also proven by Titsias [2014] and $\mathcal{L}_{\text{lower}}$ through an algebraic manipulation and is given in Appendix A. The second inequality is a consequence of $t \geq \widetilde{\lambda}_{max}$.

We typically expect $\|\mathbf{y}\|_2^2 = \mathcal{O}(N)$, which is the case when the variance of the observed ys is bounded independent of N, so if $t \ll 1/N$ the KL divergence will be small.

3.2. A priori Bounds: Averaging over y

Lemma 1 is typically overly pessimistic. It assumes $\mathbf{y}$ is in the span of the largest eigenvector of $\widetilde{\mathbf{K}}_{\mathbf{ff}}$. In this section, we consider a bound that holds *a priori* over the training outputs. This allows us to bound the KL divergence for a 'typical' dataset. To formalize this, we assume $\mathbf{y}$ is a sample from our prior generative model.

Lemma 2. *For any set of $\{\mathbf{x}_i\}_{i=1}^N$, if the training outputs $\{y_i\}_{i=1}^N$ are generated according to our prior generative model, then*

$$\frac{t}{2\sigma_n^2} \leq \mathbb{E}_y\left[KL\left(Q\|\hat{P}\right)\right] \leq \frac{t}{\sigma_n^2} \tag{8}$$

The lower bound tells us that *even if the training data is contained in an interval of fixed length, we need to use more inducing points for problems with large N if we want to ensure the sparse approximation has converged.* This is shown in Figure 1 for data uniformly sampled on the interval $[0, 5]$ with 15 inducing points.

Sketch of Proof.

$$\begin{aligned}\mathbb{E}_y\left[\mathrm{KL}\left(Q\|\hat{P}\right)\right] &= \frac{t}{2\sigma_n^2} + \int \mathcal{N}(\mathbf{y}; 0, \mathbf{K}_n) \\ &\times \log\left(\frac{\mathcal{N}(\mathbf{y}; 0, \mathbf{K}_n)}{\mathcal{N}(\mathbf{y}; 0, \mathbf{Q}_n)}\right)\mathrm{d}\mathbf{y}\end{aligned}$$

The second term on the right is a KL divergence between centered Gaussian distributions. The lower bound follows from Jensen's inequality. The proof of the upper bound (Appendix B), bounds this KL divergence above by $t/(2\sigma_n^2)$.

3.3. Minimizing the Upper Bound: An Idealized Case

We now consider the set of M interdomain inducing features that minimize the upper bounds of both Lemmas 1 and 2. Taking into account the lower bound in Lemma 2, they must be within a factor of two of the optimal features defined without reference to training outputs under the assumption of Lemma 2. Consider

$$u_m := \sum_{i=1}^{N} w_i^{(m)} f(\mathbf{x}_i)$$

where $w_i^{(m)}$ is the i^{th} entry in the m^{th} eigenvector of $\mathbf{K}_{\mathbf{ff}}$. That is, u_m is a linear combination of inducing points placed at each data point, with weights coming from the entries of the m^{th} eigenvector of $\mathbf{K}_{\mathbf{ff}}$. We show in Appendix C,

$$\mathrm{cov}(u_m, u_k) = \mathbf{w}^{(m)\intercal}\mathbf{K}_{\mathbf{ff}}\mathbf{w}^{(k)} = \lambda_k(\mathbf{K}_{\mathbf{ff}})\delta_{m,k},$$

and

$$\mathrm{cov}(u_m, f(\mathbf{x}_i)) = \left[\mathbf{K}_{\mathbf{ff}}\mathbf{w}^{(m)}\right]_i = \lambda_m(\mathbf{K}_{\mathbf{ff}})w_i^{(m)}.$$

Inference with these features can be seen as the variational equivalent of the optimal parametric projection of the model derived by Ferrari-Trecate et al. [1999].

Computation with these features requires computing the matrices $\mathbf{K}_{\mathbf{uf}}$ and $\mathbf{K}_{\mathbf{uu}}$. $\mathbf{K}_{\mathbf{uu}}$ contains the first M eigenvalues of $\mathbf{K}_{\mathbf{ff}}$, $\mathbf{K}_{\mathbf{uf}}$ contains the corresponding eigenvectors. Computing the first M eigenvalues and vectors can be done in $\mathcal{O}(N^2M)$ using, for example, Lanczos iteration [Lanczos, 1950]. The $\mathbf{Q}_{\mathbf{ff}}$ matrix for these features is the *optimal* rank-M approximation to $\mathbf{K}_{\mathbf{ff}}$ which leads to

$$\tilde{\lambda}_{max} = \lambda_{M+1}(\mathbf{K}_{\mathbf{ff}}) \quad \text{and} \quad t = \sum_{m=M+1}^{N} \lambda_m(\mathbf{K}_{\mathbf{ff}}).$$

3.4. Eigenfunction Inducing Features

We now modify the construction given in Section 3.3 to no longer depend on the specific $\mathbf{K}_{\mathbf{ff}}$ matrix (which in turn depends of the specific training inputs) and instead depend on assumptions about the training data. This construction is the *a priori* counterpart of the eigenvector inducing features, as it is defined prior to observing a specific set of training inputs.

Consider the limit as we have observed a large amount of data, so that $\frac{1}{N}\mathbf{K}_{\mathbf{ff}} \to \mathcal{K}$. This leads us to replace the eigenvalues, $\lambda(\mathbf{K}_{\mathbf{ff}})$, with the operator eigenvalues, λ, and the eigenvectors, $\mathbf{w}$, with the eigenfunctions, ϕ, yielding

$$u_m = \int_{\mathcal{X}} f(\mathbf{x})\phi_m(\mathbf{x})p(\mathbf{x})d\mathbf{x}, \tag{9}$$

Now $p(\mathbf{x})$ is a parameter of the inducing features that can be treated variationally. We note that changing p also changes the eigenvalues and eigenvectors. In Appendix C, we show

$\text{cov}(u_m, u_k) = \lambda_m\delta_{m,k}$ and $\text{cov}(u_m, f(\mathbf{x}_i)) = \lambda_m\phi_m(\mathbf{x}_i)$.

These features can be seen as the variational equivalent of methods utilizing truncated priors proposed in Zhu et al. [1997], which are the optimal linear M dimensional parametric GP approximation defined *a priori*, in terms of minimizing expected mean square error.

In the case of the SE-Kernel and Gaussian inputs, inference with these features can be performed in $\mathcal{O}(NM^2)$ using the closed form expressions for eigenfunctions and values [Zhu et al., 1997]. For Matérn kernels with inputs uniform on $[a, b]$, expressions for the eigenfunctions and eigenvalues needed for these computations can be found in Youla [1957]. However, the formulas involve solving systems of transcendental equations limiting the practical applicability of inference with these features for the Matérn class.

3.5. A priori Bounds on the KL divergence for Eigenfunction Features

Having developed the necessary preliminary results, we turn our attention to the proof of bounds on the KL divergence for inference with the eigenfunction features.

Theorem 1. *Suppose N training inputs are drawn i.i.d according to input density $p(\mathbf{x})$. For inference with M eigenfunction inducing variables defined with respect to the prior kernel and covariance, With probability at least $1-\delta$,*

$$\text{KL}\Big(Q\|\hat{P}\Big) \leq \frac{C}{2\sigma_n^2\delta}\left(1 + \frac{\|\mathbf{y}\|_2^2}{\sigma_n^2}\right) \tag{10}$$

where we have defined $C = N\sum_{i=M+1}^{\infty}\lambda_i$, and the λ_i are the eigenvalues of the integral operator $\mathcal{K}$ associated to the prior kernel and $p(\mathbf{x})$.

Theorem 2. *With the assumptions and notation of Theorem 1 and the additional assumption that $\mathbf{y}$ is distributed according to a sample from the prior generative model, with probability at least $1-\delta$,*

$$\text{KL}\Big(Q\|\hat{P}\Big) \leq \frac{C}{\delta\sigma_n^2}, \tag{11}$$

Sketch of Proof of Theorems 1 and 2. We first prove a bound on t that holds in expectation over input data matrices of size N with entries drawn i.i.d from $p(\mathbf{x})$. A direct computation of $\mathbf{Q}_{\mathbf{ff}}$ shows that $[\mathbf{Q}_{\mathbf{ff}}]_{i,j} = \sum_{m=1}^{M}\lambda_m\phi_m(\mathbf{x}_i)\phi_m(\mathbf{x}_j)$. Using the Mercer expansion of the kernel matrix and subtracting,

$$\left[\widetilde{\mathbf{K}}_{\mathbf{ff}}\right]_{i,i} = \sum_{m=M+1}^{\infty}\lambda_m\phi_m^2(\mathbf{x}_i).$$

Summing this and taking the expectation,

$$\mathbb{E}_{\mathbf{X}}[t] = N\sum_{m=M+1}^{\infty}\lambda_m\mathbb{E}_{\mathbf{x}}\left[\phi_m^2(\mathbf{x})\right] = N\sum_{m=M+1}^{\infty}\lambda_m.$$

The second equality follows from the eigenfunction having unit norm in $L^2(\mathcal{X})_p$. Using this with Lemmas 1 and 2, as well as Markov's Inequality leads to Theorems 1 and 2 respectively. □

Remark 1. *Theorem 1 can be turned into a bound that holds with high probability under additional assumptions on the eigenfunctions (e.g. finite fourth moment) using the assumption that the $\mathbf{x}_i$ are i.i.d and a concentration inequality in place of Markov's inequality.*

3.6. Square Exponential Kernel and Gaussian Inputs

For the SE-kernel in one-dimension with hyperparameters (v, ℓ^2) and $p(x) \sim \mathcal{N}(0, \sigma^2)$,

$$\lambda_m = v\sqrt{2a/A}B^{m-1}$$

where $a = 1/(4\sigma^2)$, $b = 1/(2\ell^2)$, $c = \sqrt{a^2 + 2ab}$, $A = a+b+c$ and $B = b/A$ [Zhu et al., 1997]. In this case, using the geometric series formula,

$$\sum_{m=M+1}^{\infty}\lambda_m = \frac{v\sqrt{2a}}{(1-B)\sqrt{A}}B^M.$$

Using this bound with Theorems 1 and 2, we see that by choosing $M = \mathcal{O}(\log N)$, under the assumptions of either theorem, as N tends to infinity, we can obtain a bound on the KL divergence that tends to 0 as N tends to infinity.

3.7. Matérn Kernels and Uniform Measure

For the Matérn kernel $k + 1/2$, we have $\lambda_m \asymp m^{-2k-2}$ [Ritter et al., 1995], so $\sum_{m=M+1}^{\infty}\lambda_m = \mathcal{O}(M^{-2k-1})$. In order for the bound in Theorem 2 to converge to 0, we need $\lim_{N\to\infty} N/M^{2k+1} \to 0$. This holds if $M = N^\alpha$ for $\alpha > 2k+1$. For $k > 0$, this bound tells us that the number of inducing features can grow sublinearly with the amount of data.

4. Bounds for Inducing Points

We have shown that by exploiting spectral knowledge of either $\mathbf{K_{ff}}$ or $\mathcal{K}$ we can obtain bounds on the KL divergence that indicate that the number of inducing features can be taken much smaller than the number of data points. While a mathematically elegant approach, the practical applicability of the interdomain features defined is limited by computational considerations in the case of the eigenvector features and by the lack of analytic expressions for $\mathbf{K_{uf}}$ in most cases for the eigenfunction features, as well as the need to define a parametric family containing the generally unknown input density.

In contrast, inducing points can be efficiently applied to any kernel. In this section, we show that with a good initialization based on the empirical input data distribution, inducing points lead to bounds, that are only slightly weaker than the interdomain approaches suggested so far.

Proving this amounts to obtaining bounds on the trace of the error of a *Nyström approximation* to $\mathbf{K_{ff}}$. The Nyström approximation, popularized for kernel methods by [Williams & Seeger, 2001], approximates a positive semi-definite symmetric matrix by subsampling columns. If M columns, $\{\mathbf{c_i}\}_{i=1}^{M}$, are selected from $\mathbf{K_{ff}}$, the approximation used is $\mathbf{K_{ff}} \approx \mathbf{C}\overline{\mathbf{C}}^{-1}\mathbf{C}^{\mathsf{T}}$, where $\mathbf{C} = [\mathbf{c}_1, \mathbf{c_2}, \ldots, \mathbf{c_M}]$ and $\overline{\mathbf{C}}$ is the $M \times M$ principal submatrix associated to the $\{\mathbf{c_i}\}_{i=1}^{M}$. Note that if inducing points are placed at the points associated to each column in the data matrix, then $\mathbf{K_{uu}} = \overline{\mathbf{C}}$ and $\mathbf{K}_{\mathbf{uf}}^{\mathsf{T}} = \mathbf{C}$, so $\mathbf{C}\overline{\mathbf{C}}^{-1}\mathbf{C}^{\mathsf{T}} = \mathbf{Q_{ff}}$.

Lemma 3. *[Belabbas & Wolfe, 2009] Given a symmetric positive semidefinite matrix, $\mathbf{K_{ff}}$, if M columns are selected to form a Nyström approximation such that the probability of selecting a subset of columns, Z is proportional to the determinant of the principal submatrix formed by these columns and the matching rows, then,*

$$\mathbb{E}_Z[\mathrm{Tr}(\mathbf{K_{ff}} - \mathbf{Q_{ff}})] \leq (M+1) \sum_{m=M+1}^{N} \lambda_m(\mathbf{K_{ff}}). \quad (12)$$

This lemma (along with the lower bound in Lemma 2) tells us that on average well-initialized inducing points perform within a multiplicative factor of $(M+1)$ of the eigenvector inducing features, without the need to perform a spectral decomposition of $\mathbf{K_{ff}}$.

The selection scheme described introduces negative correlations between inducing points locations, leading the $\mathbf{z}_i$ to be well-dispersed amongst the training data, as shown in Figure 2. The strength of these negative correlations is tailored to the prior kernel used in inference.

The proposed initialization scheme is equivalent to sampling Z according to a discrete k-Determinantal Point Process (k-DPP), defined over $\mathbf{X}$. Belabbas & Wolfe [2009] suggested that sampling from this distribution, which has support over $\binom{N}{M}$ subsets of columns, may be computationally infeasible. However, as observed in Hennig & Garnett [2016], exact sampling from a k-DPP can be performed sequentially. In the discrete setting with a k-DPP defined over N points, this algorithm has computational cost $\mathcal{O}(NM^2)$. The algorithm subsamples the next inducing point from the data with probability proportional to variance of a noiseless GP fit on the already selected inducing points at candidate points. Details of the algorithm are given in Appendix D.

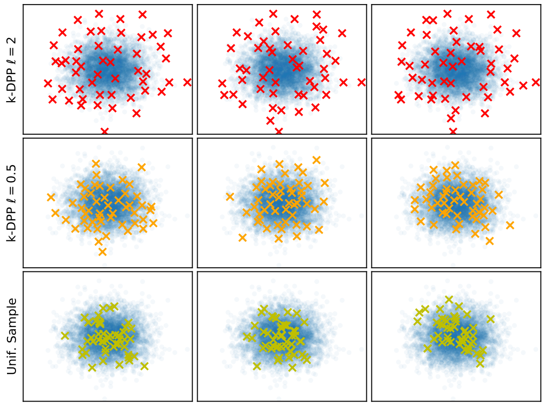

Figure 2. Uniform subsampling of data may lead to inducing points clustered in one area (bottom), while determinant based sampling, shown with a SE kernel with $\ell = 2$ (top) and with $\ell = .5$ (middle) leads to better spacing. The bounds proven hold when the kernel used for initializing points is the same as the kernel used in inference.

4.1. A Priori Bounds On the KL Divergence for Inducing Points

We now state the analogues of Theorems 1 and 2 for inducing points.

Theorem 3. *Suppose N training inputs are drawn i.i.d according to some input density $p(\mathbf{x})$. Sample M inducing points from the training with the probability assigned to any set of size M equal to the probability assigned to the corresponding subset by a $k-DPP$ with $k = M$. With probability at least $1 - \delta$,*

$$\mathrm{KL}\Big(Q\|\hat{P}\Big) \leq \frac{T}{2\sigma_n^2\delta}\left(1 + \frac{\|\mathbf{y}\|_2^2}{\sigma_n^2}\right) \quad (13)$$

where we have defined $T = N(M+1)\sum_{i=M+1}^{\infty}\lambda_i$, and the λ_i are the eigenvalues of the integral operator $\mathcal{K}$ associated to the prior kernel and $p(\mathbf{x})$.

Theorem 4. *With the assumptions and notation of Theorem 3 and the additional assumption that $\mathbf{y}$ is distributed according to a sample from the prior generative model, with* *probability at least* $1-\delta$,

$$\mathrm{KL}\Big(Q\|\hat{P}\Big) \le \frac{T}{\delta\sigma_n^2}, \tag{14}$$

Proof. We prove Theorem 4. Theorem 3 follows the same argument replacing the expectation over $\mathbf{y}$ with the bound given by Lemma 1.

$$\begin{aligned}\mathbb{E}_{\mathbf{X}}\Big[\mathbb{E}_Z\Big[\mathbb{E}_{\mathbf{y}}\Big[\mathrm{KL}\Big(Q\|\hat{P}\Big)\Big]\Big]\Big] &\le \sigma_n^{-2}\mathbb{E}_{\mathbf{X}}[\mathbb{E}_Z[t]] \\ &\le \sigma_n^{-2}(M+1)\mathbb{E}_{\mathbf{X}}\left[\sum_{m=M+1}^{N}\lambda_m(\mathbf{K_{ff}})\right] \\ &\le (M+1)N\sigma_n^{-2}\sum_{m=M+1}^{\infty}\lambda_m.\end{aligned}$$

The first two inequalities use Lemmas 2 and 3. The third follows from noting that the sum inside the expectation is the error in trace norm of the optimal rank M approximation to the covariance matrix for any given $\mathbf{X}$, and is therefore bounded above by the error from the rank M approximation due to eigenfunction features. We showed that this error is in expectation equal to $N\sum_{m=M+1}^{\infty}\lambda_m$ so this must be an upper bound on the expectation in the second to last line.

As the KL divergence is non-negative, Markov's inequality can be applied, leading to for any $\delta \in (0,1)$ with probability at least $1-\delta$,

$$\mathrm{KL}\Big(Q\|\hat{P}\Big) \le (M+1)N\delta^{-1}\sigma_n^{-2}\sum_{m=M+1}^{\infty}\lambda_m. \qquad \square$$

Figure 3 compares the actual KL divergence, the *a posteriori* bound derived by $\mathcal{L}_{\text{upper}} - \mathcal{L}_{\text{lower}}$, and the bounds proven in Theorems 3 and 4 on a dataset with normally distributed training inputs and $\mathbf{y}$ drawn from the generative model.

5. Consequences of Theorem 3 and Theorem 4

Having established our main results, we investigate implications for sparse GP regression. Our first two corollaries consider Gaussian inputs and the squared exponential kernel, and show that in D dimensions, choosing $M = \mathcal{O}(\log^D(N))$ is sufficient in order for the KL divergence to converge with high probability. We then briefly summarize convergence rates for other stationary kernels. Finally we point out consequences of our definition of convergence for the quality of the pointwise posterior mean and uncertainty.

5.1. Comparison of Consequences of Theorems

Using the explicit formula for the eigenvalues given in Section 3.6, we arrive at the following corollary:

Corollary 1. *In the setting of regression with an SE-kernel and Gaussian distributed inputs, take the assumptions of Theorem 3 and the additional assumption* $\|\mathbf{y}\|_2^2 \le RN, R \ge 0$ *with probability at least* $0 < 1-\delta' \le 1$. *Then for any* $\epsilon > 0$, *with probability at least* $1-\delta-\delta'$,

$$\mathrm{KL}\Big(Q\|\hat{P}\Big) \le N^{-\epsilon}\left(\frac{R}{\sigma_n^2} + \frac{1}{N}\right). \tag{15}$$

when inference is performed with $M = \frac{(3+\epsilon)\log(N)+\log D}{\log(B^{-1})}$, *where* $D = \frac{v\sqrt{2a}}{2\sqrt{A}\sigma_n^2\delta(1-B)}$.

Remark 2. *The assumption* $\|\mathbf{y}\|_2^2 \le RN$ *with probability at least* $1-\delta' > 0$ *is very weak. For example, if* $\mathbf{y}$ *is actually a noisy realization of some integrable function with homoscedastic noise,*

$$\sum_{i=1}^{N} y_i^2 \le \sum_{i=1}^{N} f(\mathbf{x}_i)^2 + \sum_{i=1}^{N}\epsilon_i^2 + \textit{small}$$

The first sum is asymptotically $N\int f(\mathbf{x})p(\mathbf{x})dx$ *and the second is* $N\sigma_n^2$.

Remark 3. *A slightly sharper, in terms of constants, version of Corollary 1 can be obtained if we instead take the assumptions of Theorem 4.*

Remark 4. *By choosing* δ *to be* $\mathcal{O}(N^{-\epsilon/2})$, *and* C *to grow slowly with* N *so that* δ', *tends to zero we obtain a bound that converges to zero with high probability still using only* $\mathcal{O}(\log(N))$ *features.*

The consequence of Corollary 1 is shown in Figure 4, in which we gradually increase N, choosing $M = C\log(N) + C_0$, and see the KL divergence converges as an inverse power of N. The training outputs are generated from a sample from the prior generative model.

We see that for the SE-kernel and Gaussian inputs, the difference in what we prove about the scaling of M with N between inducing points and the eigenfunction features differs by only a constant. For the Matérn kernel $k+1/2$, this is not the case. In particular, in order to prove the KL divergence is small with our bound in Theorem 4, we need to choose $M = N^{\alpha}$ with $\alpha > 1/(2k)$ instead of $\alpha > 1/(2k+1)$. This difference is particularly stark in the case of the popular Matérn $3/2$ kernel, for which our bounds tell us that inference with inducing points requires $\mathcal{O}(\sqrt{N})$ as opposed to $\mathcal{O}(N^{1/3})$ for the eigenfunction features. Whether this is an artifact of the proof, caused by the initialization scheme, or an inherent limitation for inducing points is an interesting area for future work.

5.2. Multidimensional Data, Effect of Input Density and other Kernels

If $\mathcal{X} = \mathbb{R}^D$, it is common to choose a *separable kernel*, meaning the kernel can be written as a product of kernels

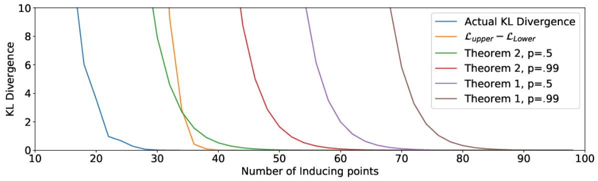

Figure 3. Rates of convergence as M increases on fixed dataset of size $N = 1000$, with a squared exponential kernel with $\ell = .3, v = 1, \sigma_n = 1$ and $x \sim \mathcal{N}(0,1)$ and $\mathbf{y}$ sampled from the prior.

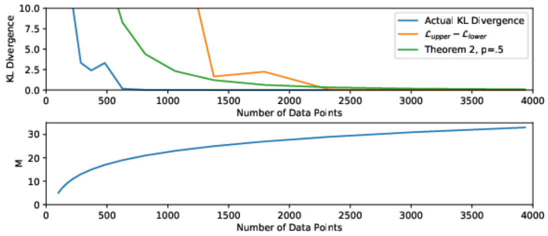

Figure 4. Corollary 1 tells us that increasing M like $\log(N)$ gives an upper bound on the KL divergence that decays like an inverse power of N.

along each dimension. If this choice of prior is made, and input densities factor over the dimensions, then the eigenvalues of $\mathcal{K}$ are the product of the eigenvalues along each dimension. In the case of the SE-ARD kernel and Gaussian input distributions, this means $\lambda_{m^D} = \mathcal{O}(B^M)$, where B corresponds to the largest B value coming from any of the one dimensional SE-kernels inolved in the product and the associated marginal input densities. Omitting constants, we obtain an analogous statement to Corollary 1 in D-dimensions.

Corollary 2. *For any fixed $\epsilon, \delta > 0$ under the assumptions of Corollary 1, but with the SE-ARD kernel in D dimensions and $p(\mathbf{x})$ a multivariate Gaussian, we can choose $M = \mathcal{O}(\log^D(N))$ inducing points so that with probability at least $1 - \delta$,*

$$\mathrm{KL}\left(Q \| \hat{P}\right) \leq \delta^{-1}\epsilon.$$

While for the SE-kernel and Gaussian input density we saw that M can grow polylogarithmically in N, and the KL divergence still converges, this is not the case for other regression with other kernels or input distribution.

In general, closed form expressions for the eigenvalues of arbitrary kernels with respect to various distributions are not known. However, for stationary kernels and compactly supported input distributions the asymptotic rate of decay of the eigenvalues of $\mathcal{K}$ are well-understood, thanks to the work of Widom [1963; 1964] and Ritter et al. [1995]. The intuitive explanation of these results is that smooth kernels, with concentrated input distributions have rapidly decaying eigenvalues. In contrast, kernels such as the Matérn-1/2 that define processes that are not smooth have slowly decaying eigenvalues. Table 1 summarizes these results and their implications for the number of inducing points needed for our bounds to converge with popular stationary kernels.

5.3. Pointwise Approximate Posterior

In practice, we are frequently concerned with pointwise estimates of the posterior mean and variance. It is therefore desirable to show that the approximate variational posterior gives similar estimates of the quatities as the true posterior.

Huggins et al. [2018] derived an approximation method for sparse GP inference with provable guarantees about point-wise mean and variance estimates of the posterior process. They additionally show that moderately large KL divergence may result in large deviations in posterior mean and variance estimates. In this section, we show that if M is sufficiently large that the KL divergence converges to zero, our variational estimates of mean and variance also converge to the posterior values.

The chain rule of KL divergence [Matthews et al., 2016], tells us that

$$\begin{aligned}&\mathrm{KL}(\mu_{\mathcal{X}} \| \nu_{\mathcal{X}}) = \mathrm{KL}(\mu_{\mathbf{x}_*} \| \nu_{\mathbf{x}_*})\\&\quad + \mathbb{E}_{\mu_{\mathbf{x}_*}}\left[\mathrm{KL}\big(\mu_{\mathcal{X}\setminus\mathbf{x}_*|\mathbf{x}_*} \| \nu_{\mathcal{X}\setminus\mathbf{x}_*|\mathbf{x}_*}\big)\right] \geq \mathrm{KL}(\mu_{\mathbf{x}_*} \| \nu_{\mathbf{x}_*}).\end{aligned}$$

In other words, the KL divergence between posterior processes upper bounds the KL divergence between any of the posterior marginals. Therefore, to provide pointwise guarantees about posterior inference, we need only consider bounds on the mean and variance of a one-dimensional Gaussian with a small KL divergence.

In Appendix B, we prove:

Proposition 1. *Suppose q and p are one dimensional Gaussian distributions with means μ_1 and μ_2 and variances σ_1 and σ_2, such that $2KL(q\|p) = \epsilon \leq \frac{1}{5}$, then*

$$|\mu_1 - \mu_2| \leq \sigma_2\sqrt{\epsilon} \leq \frac{\sigma_1\sqrt{\epsilon}}{\sqrt{1 - \sqrt{3\epsilon}}}$$

and

$$(1 - \sqrt{3\epsilon}) < \frac{\sigma_1^2}{\sigma_2^2} < (1 + \sqrt{3\epsilon}).$$

If $\epsilon \to 0$, proposition 1 implies $\mu_1 \to \mu_2$ and $\sigma_1 \to \sigma_2$. Using this and Theorems 3 and 4, *the posterior mean and variance converge pointwise to those of the full model using $M \ll N$ inducing features.*

6. Related Work

Statistical guarantees for convergence of parametric GP approximations [Zhu et al., 1997; Ferrari-Trecate et al., 1999],

Table 1. The number of features needed for our bounds to converge for several kernels assuming D is fixed, these hold for any $\epsilon_D > 0$.

Kernel	Input Distribution	λ_m	M, Theorem 3	M, Theorem 4
SE-Kernel	Compact Support	$\mathcal{O}\big(\exp(-A\frac{m}{d}\log\frac{m}{d})\big), A > 0$	$\mathcal{O}(\log^D(N))$	$\mathcal{O}(\log^D(N))$
SE-Kernel	Gaussian	$\mathcal{O}\big(\exp(-\frac{m}{d}\big)$	$\mathcal{O}(\log^D(N))$	$\mathcal{O}(\log^D(N))$
Matérn k+1/2	Uniform	$\mathcal{O}\Big(M^{-2k-2}\log(M)^{2(d-1)(k+1)}\Big)$	$\mathcal{O}\Big(N^{1/k+\epsilon_D}\Big)$	$\mathcal{O}\Big(N^{1/(2k)+\epsilon_D}\Big)$

lead to similar conclusions about the choice of approximating rank. Ferrari-Trecate et al. [1999] showed that given N data points, using a truncated SVD of the prior covariance matrix with rank M such that $\lambda_M \ll \sigma_n^2/N$ results in almost no change in the model, at least in terms of expected mean squared error. Our results can be considered the equivalent for variational inference, showing that theoretical guarantees can be established for *non-parametric* approximate inference. Our average case analysis leads to choosing M such that $\sum_{m=M+1}^{\infty}\lambda_m \ll \sigma_n^2/N$ with the interdomain features or $(M+1)\sum_{m=M+1}^{\infty}\lambda_m \ll \sigma_n^2/N$, using inducing points in order to ensure the KL divergence is small.

Connections between high-quality Nyström approximation and computationally efficient models have also been utilized in kernel ridge regression. Alaoui & Mahoney [2015] showed that replacing the full covariance matrix with its Nyström approximation, sampled according to the 'leverage scores' converges in mean square error when the number of columns scales with the effective dimensionality of the problem. The effective dimensionality is roughly the same as the number of eigenvalues of $\mathbf{K_{ff}}$ larger than the parameter of ridge regression. This essentially coincides with the choice proposed in Ferrari-Trecate et al. [1999] for GPs when parameter of ridge regression is set so that the posterior mean coincides with the mean of a GP.

7. Conclusion

We proved bounds on the KL divergence from the variational approximation of sparse GP regression to the posterior that depend only on the decay of the eigenvalues of the covariance operator for the prior kernel. These bounds prove the intuitive result, *smooth kernels with training data concentrated in a small region admit high quality, very sparse approximations*. These bounds prove that *truly sparse non-parametric inference with $M \ll N$* can still provide reliable estimates of the marginal likelihood and pointwise posterior.

Extensions to models with non-conjugate likelihoods, especially bounding the additional error introduced by sparsity in the framework of Hensman et al. [2015], pose a promising direction for future research.

References

Alaoui, A. and Mahoney, M. W. Fast randomized kernel ridge regression with statistical guarantees. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 775–783. Curran Associates, Inc., 2015.

Bauer, M., van der Wilk, M., and Rasmussen, C. E. Understanding probabilistic sparse Gaussian process approximations. In *Advances in neural information processing systems*, pp. 1533–1541, 2016.

Belabbas, M.-A. and Wolfe, P. J. Spectral methods in machine learning and new strategies for very large datasets. *Proceedings of the National Academy of Sciences*, 106 (2):369–374, January 2009. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0810600105.

Ferrari-Trecate, G., Williams, C. K., and Opper, M. Finite-dimensional approximation of Gaussian processes. In *Advances in neural information processing systems*, pp. 218–224, 1999.

Hennig, P. and Garnett, R. Exact sampling from determinantal point processes. *arXiv preprint arXiv:1609.06840*, 2016.

Hensman, J., Fusi, N., and Lawrence, N. D. Gaussian processes for big data. In *Uncertainty in Artificial Intelligence*, pp. 282. Citeseer, 2013.

Hensman, J., Matthews, A. G. d. G., and Ghahramani, Z. Scalable variational Gaussian process classification. 2015.

Hensman, J., Durrande, N., and Solin, A. Variational fourier features for gaussian processes. *Journal of Machine Learning Research*, 18(151):1–52, 2018.

Huggins, J. H., Campbell, T., Kasprzak, M., and Broderick, T. Scalable Gaussian process inference with finite-data mean and variance guarantees. *CoRR*, abs/1806.10234, 2018.

Koltchinskii, V., Giné, E., et al. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1): 113–167, 2000.

Koltchinskii, V. I. Asymptotics of spectral projections of some random matrices approximating integral operators. In *High dimensional probability*, pp. 191–227. Springer, 1998.

Lanczos, C. *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*. 1950.

Lázaro-Gredilla, M. and Figueiras-Vidal, A. Inter-domain Gaussian Processes for Sparse Inference using Inducing Features. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 22*, pp. 1087–1095. Curran Associates, Inc., 2009.

Matthews, A. G. d. G. *Scalable Gaussian process inference using variational methods*. PhD Thesis, 2016.

Matthews, A. G. d. G., Hensman, J., Turner, R., and Ghahramani, Z. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. *Journal of Machine Learning Research*, 51:231–239, 2016.

Mercer, J. Functions of positive and negative type, and their connection the theory of integral equations. *Phil. Trans. R. Soc. Lond. A*, 209(441-458):415–446, 1909.

Neal, R. M. *Bayesian learning for neural networks*, volume 118. Springer, 1996.

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

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pp. 1177–1184, 2008.

Rasmussen, C. E. and Williams, C. K. *Gaussian processes for machine learning*. MIT Press, 2006.

Ritter, K., Wasilkowski, G. W., and Woniakowski, H. Multivariate integration and approximation for random fields satisfying Sacks-Ylvisaker conditions. *The Annals of Applied Probability*, 5(2):518–540, 1995. ISSN 10505164.

Titsias, M. K. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, pp. 567–574, 2009.

Titsias, M. K. *Variational Inference for Gaussian and Determinantal Point Processes*. December 2014. Published: Workshop on Advances in Variational Inference (NIPS 2014).

Turner, R. E. and Sahani, M. *Two problems with variational expectation maximisation for time series models*, pp. 104–124. Cambridge University Press, 2011.

Widom, H. Asymptotic behavior of the eigenvalues of certain integral equations. *Transactions of the American Mathematical Society*, 109(2):278–295, 1963. ISSN 00029947.

Widom, H. Asymptotic behavior of the eigenvalues of certain integral equations. ii. *Archive for Rational Mechanics and Analysis*, 17(3):215–229, 1964.

Williams, C. K. I. and Seeger, M. Using the Nyström Method to Speed Up Kernel Machines. In Leen, T. K., Dietterich, T. G., and Tresp, V. (eds.), *Advances in Neural Information Processing Systems 13*, pp. 682–688. MIT Press, 2001.

Youla, D. The solution of a homogeneous Wiener-Hopf integral equation occurring in the expansion of second-order stationary random functions. *IRE Transactions on Information Theory*, 3(3):187–193, 1957.

Zhu, H., Williams, C. K. I., Rohwer, R., and Morciniec, M. Gaussian Regression and Optimal Finite Dimensional Linear Models. In *Neural Networks and Machine Learning*, pp. 167–184. Springer-Verlag, 1997.

A. Proof Of Lemma 1

Titsias [2014] actually proves the tighter upper bound,

$$\mathcal{L} \leq \mathcal{L}'_{\text{upper}} := -\frac{N}{2}\log(2\pi) \\ -\frac{1}{2}\log(|\mathbf{Q}_n|) - \frac{1}{2}\mathbf{y}^\intercal\Big(\mathbf{Q}_n + \widetilde{\lambda}_{max}\mathbf{I}\Big)^{-1}\mathbf{y}$$

Subtracting,

$$\mathcal{L}'_{\text{upper}} - \mathcal{L}_{\text{lower}} \\ = \frac{t}{2\sigma_n^2} + \frac{1}{2}\Big(\mathbf{y}^\intercal\Big(\mathbf{Q}_n^{-1} - (\mathbf{Q}_n + \widetilde{\lambda}_{max}\mathbf{I})^{-1}\Big)\mathbf{y}\Big). \quad (16)$$

Since $\mathbf{Q_{ff}}$ is symmetric positive semidefinite, $\mathbf{Q}_n$ is positive definite with eigenvalues bounded below by σ_n^2. Write, $\mathbf{Q}_n = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\intercal$, where $\mathbf{U}$ is unitary and $\mathbf{\Lambda}$ is a diagonal matrix with non-increasing diagonal entries $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_N \geq \sigma_n^2$.

We can rewrite the second term (ignoring the factor of one half) in Equation 16 as,

$$(\mathbf{U}^\intercal\mathbf{y})^\intercal\Big(\mathbf{\Lambda}^{-1} - (\mathbf{\Lambda} + \widetilde{\lambda}_{max}\mathbf{I})^{-1}\Big)(\mathbf{U}^\intercal\mathbf{y}).$$

Define, $\mathbf{z} = (\mathbf{U}^\intercal \mathbf{y})$. Since $\mathbf{U}$ is unitary, $\|\mathbf{z}\| = \|\mathbf{y}\|$.

$$\begin{aligned}(\mathbf{U}^\intercal\mathbf{y})^\intercal\big(\boldsymbol{\Lambda}^{-1}- (\boldsymbol{\Lambda}+t\mathbf{I})^{-1}\big)(\mathbf{U}^\intercal\mathbf{y}) & \\ &= \mathbf{z}^\intercal\Big(\boldsymbol{\Lambda}^{-1}-(\boldsymbol{\Lambda}+\widetilde{\lambda}_{max}\mathbf{I})^{-1}\Big)\mathbf{z}\\ &=\sum_i z_i^2\frac{\widetilde{\lambda}_{max}}{\gamma_i^2+\gamma_i\widetilde{\lambda}_{max}}\\ &\le \|\mathbf{y}\|^2\frac{\widetilde{\lambda}_{max}}{\gamma_N^2+\gamma_N\widetilde{\lambda}_{max}}.\end{aligned}$$

The last inequality comes from noting that the fraction in the sum attains a maximum when γ_i is minimized. Since σ_n^2 is a lower bound on the smallest eigenvalue of $\mathbf{Q}_n$, we have,

$$\mathbf{y}^T\Big(\mathbf{Q}_n^{-1}-(\mathbf{Q}_n+\widetilde{\lambda}_{max}\mathbf{I})^{-1}\Big)\mathbf{y}\le\frac{\widetilde{\lambda}_{max}\|\mathbf{y}\|^2}{\sigma_n^4+\sigma_n^2\widetilde{\lambda}_{max}}.$$

Lemma 1 follows.

B. KL Divergence Gaussian Distributions

B.1. KL divergence between multivariate Gaussian distributions

We make use of the formula for KL divergences between multivariate Gaussian distributions in our proof of Lemma 2, and the univariate case in Proposition 1.

Recall the KL divergence from $p_1 \sim \mathcal{N}(\mathbf{m_1}, \mathbf{S_1})$ to $p_2 \sim \mathcal{N}(\mathbf{m_2}, \mathbf{S_2})$ both of dimension N is given by

$$\begin{aligned}\mathrm{KL}(p_1\|p_2) = \frac{1}{2}\Bigg(&\mathrm{Tr}\big(\mathbf{S_2}^{-1}\mathbf{S_1}\big)+\log\bigg(\frac{|\mathbf{S_2}|}{|\mathbf{S_1}|}\bigg)\\ &-(\mathbf{m_1}-\mathbf{m_2})^\intercal\mathbf{S_2}^{-1}(\mathbf{m_1}-\mathbf{m_2})\Bigg)\ge 0. \end{aligned}\tag{17}$$

The inequality is a special case of Jensen's inequality.

B.2. Proof of Upper Bound in Lemma 2

In the main text we showed,

$$\begin{aligned}\mathbb{E}_y\Big[\mathrm{KL}\Big(Q\|\hat{P}\Big)\Big] = \frac{t}{2\sigma_n^2}+\int\mathcal{N}(\mathbf{y};0,\mathbf{K}_n)&\\ \times\log\bigg(\frac{\mathcal{N}(\mathbf{y};0,\mathbf{K}_n)}{\mathcal{N}(\mathbf{y};0,\mathbf{Q}_n)}\bigg)\mathrm{d}\mathbf{y}&\end{aligned}$$

In order to complete the proof, we need to show that the second term on the right hand side is bounded above by $t/(2\sigma_n^2)$. Using Equation 17:

$$\begin{aligned}\mathbb{E}_y\Big[KL\Big(Q\|\hat{P}\Big)\Big] &= \frac{t}{2\sigma_n^2}-\frac{N}{2}+\frac{1}{2}\log\bigg(\frac{|\mathbf{Q}_n|}{|\mathbf{K}_n|}\bigg)\\ &\quad+\frac{1}{2}\mathrm{Tr}\big(\mathbf{Q}_n^{-1}(\mathbf{K}_n)\big)\\ &\le\frac{t}{2\sigma_n^2}-\frac{N}{2}+\frac{1}{2}\mathrm{Tr}\Big(\mathbf{Q}_n^{-1}(\mathbf{Q}_n+\widetilde{\mathbf{K}}_{\mathbf{ff}})\Big).\end{aligned}\tag{18}$$

The inequality follows from noting the log determinant term is negative, as $\mathbf{K}_n \succ \mathbf{Q}_n$. Simplifying the last term,

$$\begin{aligned}\frac{1}{2}\mathrm{Tr}(\mathbf{I})+\frac{1}{2}\mathrm{Tr}\Big(\mathbf{Q}_n^{-1}\widetilde{\mathbf{K}}_{\mathbf{ff}}\Big)&\le N/2+t\lambda_1\big(\mathbf{Q}_n^{-1}\big)/2\\ &=N/2+t/(2\sigma_n^2).\end{aligned}$$

The final inequality uses that for positive semi-definite symmetric matrices $\mathrm{Tr}(AB)\le\mathrm{Tr}(A)\lambda_1(B)$ which is a special case of Hölder's inequality. The final line uses that, when $M < N$, the largest eigenvalue of $\mathbf{Q}_n^{-1}$ is σ_n^{-2} (and more generally it is bounded above by this quantity). Using this in Equation 18 finishes the proof.

B.3. Proof of Proposition 1

Defining $\epsilon = 2KL(q\|p)$,

$$\begin{aligned}\epsilon &= \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{\sigma_2^2}-\log\bigg(\frac{\sigma_1^2}{\sigma_2^2}\bigg)-1 \qquad (19)\\ &\ge\frac{1}{2}(x-\log(x)-1)\end{aligned}$$

where we have defined $x=\frac{\sigma_1^2}{\sigma_2^2}$.

Applying the lower bound $x-\log(x)-1\ge(x-1)^2/2-(x-1)^3/3$,

$$\epsilon\ge(x-1)^2/2-(x-1)^3/3.$$

A bound on $|x-1|$ that holds for all ϵ can then be found with the cubic formula. Under the assumption that $\epsilon<\frac{1}{5}$, we have $x-\log(x)<1.2$ which implies $x\in[0.49,1.77]$. For x in this range, we have

$$x-\log(x)-1\ge(x-1)^2/3$$

So,

$$|x-1|\le\sqrt{\epsilon}$$

This proves that,

$$1-\sqrt{3\epsilon}<\frac{\sigma_1^2}{\sigma_2^2}<1+\sqrt{3\epsilon}.$$

From Equation 19 and $x-\log x>1$,

$$|\mu_1-\mu_2|\le\sigma_2\sqrt{\epsilon}.$$

Using our bound on the ratio of the variances completes the proof of Proposition 1.

C. Covariances for Interdomain Features

We compute the covariances for eigenvector and eigenfunction inducing features.

C.1. Eigenvector inducing features

Recall we have defined eigenvector inducing features by,

$$u_m = \sum_{i=1}^{N} w_i^{(m)} f(\mathbf{x}_i).$$

Then,

$$\begin{aligned}\mathrm{cov}(u_m, u_k) &= \mathbb{E}\left[\sum_{i=1}^{N} w_i^{(m)} f(\mathbf{x}_i) \sum_{j=1}^{N} w_j^{(k)} f(\mathbf{x}_j)\right] \\ &= \sum_{i=1}^{N} w_i^{(m)} \sum_{j=1}^{N} w_j^{(k)} \mathbb{E}[f(\mathbf{x}_i) f(\mathbf{x}_j)] \\ &= \sum_{i=1}^{N} w_i^{(m)} \sum_{j=1}^{N} w_j^{(k)} k(\mathbf{x}_i, \mathbf{x}_j).\end{aligned}$$

We know recognize this expression as $\mathbf{w}^{(m)\intercal}\mathbf{K}_{\mathbf{ff}}\mathbf{w}^{(k)}$. Using the defining property of eigenvectors as well as orthonormality,

$$\mathrm{cov}(u_m, u_k) = \lambda_k(\mathbf{K}_{\mathbf{ff}})\delta_{m,k}.$$

Similarly,

$$\begin{aligned}\mathrm{cov}(u_m, f(\mathbf{x}_i)) &= \mathbb{E}\left[\sum_{j=1}^{N} w_j^{(m)} f(\mathbf{x}_j) f(\mathbf{x}_i)\right] \\ &= \sum_{j=1}^{N} w_j^{(m)} \mathbb{E}[f(\mathbf{x}_j) f(\mathbf{x}_i)] \\ &= \sum_{j=1}^{N} w_j^{(m)} k(\mathbf{x}_j, \mathbf{x}_i)\end{aligned}$$

This is the i^{th} entry of the matrix vector product $\mathbf{K}_{\mathbf{ff}}\mathbf{w}^{(m)} = \lambda_m(\mathbf{K}_{\mathbf{ff}})\mathbf{w}_i^{(m)}$.

C.2. Eigenfunction inducing features

Recall we have defined eigenfunction inducing features by,

$$u_m = \int \phi_m(\mathbf{x}) f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}.$$

Then,

$$\begin{aligned}&\mathrm{cov}(u_m, u_k) \\ &= \mathbb{E}\left[\int \phi_m(\mathbf{x}) f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \int \phi_k(\mathbf{x}') f(\mathbf{x}') p(\mathbf{x}') d\mathbf{x}'\right] \\ &= \int \phi_m(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \int \phi_k(\mathbf{x}') \mathbb{E}[f(\mathbf{x}) f(\mathbf{x}')] p(\mathbf{x}') d\mathbf{x}' \\ &= \int \phi_m(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \int \phi_k(\mathbf{x}') k(\mathbf{x}, \mathbf{x}') p(\mathbf{x}') d\mathbf{x}'.\end{aligned}$$

Algorithm 1 Initialization of Inducing Points

```
**Input:** Training inputs $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, number of points to choose, $M$, kernel $k$.
**Returns:** $Z$, a sample of $M$ inducing points drawn proportional to the determinant of $\mathbf{K}_{Z,Z}$
Initialize $Z = \{\}$
while $|Z| < M$ do
  for $\mathbf{x}_i \in \mathbf{X} \setminus Z$ do
    $[\mathbf{k}_{Z,i}]_m := \mathrm{cov}(\mathbf{z}_m, \mathbf{x}_i)$.
    $V_i = k(\mathbf{x}_i, \mathbf{x}_i) - \mathbf{k}_{i,Z}\mathbf{K}_{Z,Z}^{-1}\mathbf{k}_{Z,i}$,
  end for
  Sample $\mathbf{x}_i$ with probability proportional to $V_i$
  Add $\mathbf{x}_i$ to $Z$
end while
```

The expectation and integration may be interchanged by Fubini's theorem, as both integrals converge absolutely since $p(\mathbf{x})$ is a probability density, the $\phi_m(\mathbf{x})$ are in $L^2(\mathcal{X})_p \cap L^1(\mathcal{X})_p$ and k is bounded.

We may then apply the eigenfunction property to the inner integral and orthonormality of eigenfunctions to the result yielding,

$$\mathrm{cov}(u_m, u_k) = \lambda_k \int \phi_k(\mathbf{x}')\phi_m(\mathbf{x})p(\mathbf{x})d\mathbf{x} = \lambda_k \delta_{m,k}.$$

With similar considerations,

$$\begin{aligned}&\mathrm{cov}(u_m, f(\mathbf{x}_i)) \\ &= \mathbb{E}\left[\int \phi_m(\mathbf{x}) f(\mathbf{x}) f(\mathbf{x}_i) p(\mathbf{x}) d\mathbf{x}\right] \\ &= \int \phi_m(\mathbf{x}) \mathbb{E}[f(\mathbf{x}) f(\mathbf{x}_i)] p(\mathbf{x}) d\mathbf{x} \\ &= \lambda_m \phi_m(\mathbf{x}_i).\end{aligned}$$

D. Sampling from a Discrete k-DPP

In this section, we give the algorithm, Algorithm 1, described in Hennig & Garnett [2016] adapted to the discrete setting which is relevant to our application. We additionally show that it can be implemented with complexity $\mathcal{O}(NM^2)$.

D.1. Efficient Implementation of Algorithm 1

We will denote $\mathbf{K}_{\mathbf{Z}} = \mathbf{K}_{\mathbf{Z},\mathbf{Z}}$. We view $\mathbf{K}_{\mathbf{Z}}$ as a block matrix of the form:

$$\mathbf{K}_{\mathbf{Z}} = \begin{bmatrix} \mathbf{K}_{\mathbf{Z-1}} & \mathbf{k}_{\mathbf{m}} \\ \mathbf{k}_{\mathbf{m}}^{\intercal} & k(\mathbf{z}_m, \mathbf{z}_m) \end{bmatrix}$$

where $\mathbf{k}_{\mathbf{m}}$ is an $(m-1) \times 1$ column vector with $[\mathbf{k}_{\mathbf{m}}]_i = k(\mathbf{z}_i, \mathbf{z}_m)$. Using block matrix inversion,

$$\mathbf{K}_{\mathbf{Z}}^{-1} = \begin{bmatrix} \mathbf{K}_{\mathbf{Z-1}}^{-1} + \frac{1}{r}\mathbf{K}_{\mathbf{Z-1}}^{-1}\mathbf{k}_{\mathbf{m}}\mathbf{k}_{\mathbf{m}}{}^{\intercal}\mathbf{K}_{\mathbf{Z-1}}^{-1} & -\frac{1}{r}\mathbf{K}_{\mathbf{Z-1}}^{-1}\mathbf{k}_{\mathbf{m}} \\ -\frac{1}{r}\mathbf{k}_{\mathbf{m}}{}^{\intercal}\mathbf{K}_{\mathbf{Z-1}}^{-1} & \frac{1}{r} \end{bmatrix}$$

with $r = k(\mathbf{z}_m, \mathbf{z}_m) - \mathbf{k_m}^\mathsf{T}\mathbf{K}_{\mathbf{Z-1}}^{-1}\mathbf{k_m}$. Define,

$$\mathbf{t}_{\mathbf{Z},\mathbf{i}} = \mathbf{K}_{\mathbf{Z}}^{-1}\mathbf{k}_{\mathbf{Z},\mathbf{i}}.$$

With this definition,

$$\mathbf{K}_{\mathbf{Z}}^{-1} = \begin{bmatrix} \mathbf{K}_{\mathbf{Z-1}}^{-1} + \frac{1}{r}\mathbf{t}_{\mathbf{Z-1},\mathbf{m}}\mathbf{t}_{\mathbf{Z-1},\mathbf{m}}^\mathsf{T} & -\frac{1}{r}\mathbf{t}_{\mathbf{Z-1},\mathbf{m}} \\ -\frac{1}{r}\mathbf{t}_{\mathbf{Z-1},\mathbf{m}} & \frac{1}{r} \end{bmatrix}$$

and

$$r = k(\mathbf{z}_m, \mathbf{z}_m) - \mathbf{k_m}^\mathsf{T}\mathbf{t}_{\mathbf{Z-1},\mathbf{m}}.$$

where by an abuse of notation, we assumed $\mathbf{z}_m$ is also $\mathbf{x}_m$. Additionally,

$$V_i = k(\mathbf{x}_i, \mathbf{x}_j) - \mathbf{k}_{\mathbf{Z},\mathbf{i}}^\mathsf{T}\mathbf{t}_{\mathbf{Z},\mathbf{i}}.$$

We assume the kernel can be evaluated in constant time. The second term is an inner product between vectors of length m, and therefore has computational cost $\mathcal{O}(m)$.

We need to show that given $\mathbf{t}_{\mathbf{Z-1},\mathbf{j}}$ for all j, $\mathbf{t}_{\mathbf{Z},\mathbf{i}}$ can be updated in linear time. Using the formula for $\mathbf{K}_{\mathbf{Z}}^{-1}$ and writing

$$\mathbf{k}_{\mathbf{Z},\mathbf{i}} = \begin{bmatrix} \mathbf{k}_{\mathbf{Z},\mathbf{i}} \\ k(\mathbf{z}_m, \mathbf{x}_i) \end{bmatrix},$$

we arrive at:

$$\mathbf{t}_{\mathbf{Z},\mathbf{i}} = \begin{bmatrix} \mathbf{t}_{\mathbf{Z-1},\mathbf{i}} + \frac{1}{r}\left(\mathbf{t}_{\mathbf{Z-1},\mathbf{m}}\mathbf{t}_{\mathbf{Z-1},\mathbf{m}}^\mathsf{T}\mathbf{k}_{\mathbf{Z},\mathbf{i}} - k(\mathbf{z}_m, \mathbf{x}_i)\mathbf{t}_{\mathbf{Z-1},\mathbf{m}}\right) \\ \frac{1}{r}\left(-\mathbf{t}_{\mathbf{Z-1},\mathbf{m}}^\mathsf{T}\mathbf{k}_{\mathbf{Z},\mathbf{i}} + k(\mathbf{z}_m, \mathbf{x}_i)\right) \end{bmatrix}$$

By performing the matrix operations in the correct order, this also consists of only inner products of length $(m-1)$ vectors and can be computed in $\mathcal{O}(m)$.