Rates of Convergence for Sparse Variational Gaussian Process Regression

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

Abstract

Excellent variational approximations to Gaussian process posteriors have been developed which avoid the $\mathcal{O}(N^3)$ scaling with dataset size N. They reduce the computational cost to $\mathcal{O}(NM^2)$, with $M \ll N$ 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 show that with high probability the KL divergence can be made arbitrarily small by growing M more slowly than N. A particular case is that for regression with normally distributed inputs in D-dimensions with the Squared Exponential kernel, $M = \mathcal{O}(\log^D N)$ suffices. Our results show that as datasets grow, Gaussian process posteriors can 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 most well known drawback of GP regression is the computational cost of the exact calculation of these quantities, which scales as $\mathcal{O}(N^3)$ in time and $\mathcal{O}(N^2)$ in memory where N is the number of training examples. Low-rank approximations [Quiñonero Candela & Rasmussen, 2005] choose M *inducing variables* which summarise the entire posterior, reducing the cost to $\mathcal{O}(NM^2 + M^3)$ time and $\mathcal{O}(NM + M^2)$ memory.

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

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

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. Taking into account the rate at which M must 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.

Approximate GPs are commonly trained using variational inference [Titsias, 2009], which minimizes the KL divergence between the approximate and full posterior processes [Matthews et al., 2016]. We use 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 for the KL from the approximation to the posterior to go to zero. This shows that very sparse approximations can be used for GP models on large datasets, without introducing much bias into the hyperparameters selected using the evidence lower bound (ELBO), and with approximate posteriors that accurately reflect the predictions and uncertainties in the models' posteriors.

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. In the case of stationary kernels, our main result proves that priors with smoother sample functions, and datasets with more concentrated inputs admit sparser approximations.

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

$$\mathrm{KL}\left(Q \| \hat{P}\right) \leq \mathcal{O}\left(\frac{g(M,N)}{\sigma_n^2 \delta}\left(1 + \frac{c\|\mathbf{y}\|_2^2}{\sigma_n^2}\right)\right)$$

with probability at least $1-\delta$, where $\hat{P}$ is the posterior Gaussian process, and Q is a variational approximation and $\mathbf{y}$ are the training targets. The constant c is either 0 or 1; $g(M,N)$ depends on both the kernel and input distribution,

grows linearly in N and generally decays rapidly in M. Theorems 1 to 4 are all of this form; the first two give results of this form for a collection of inducing variables defined using spectral knowledge of an operator depending on the prior, while the last two theorems prove the bounds for standard inducing points.

2. Background and Notation

2.1. Gaussian Process Regression

We are concerned with the problem of Gaussian process regression. We observe *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 [Rasmussen & Williams, 2006]. The log marginal likelihood is of particular interest to us, as the quality of its approximation and our posterior approximation is linked. Its form is

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

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

2.2. Sparse Variational Gaussian Process Regression

While all quantities of interest have analytic expressions, their computation is prohibitively expensive for large datasets due to the $\mathcal{O}(N^3)$ time complexity of the log-determinant and matrix inverse. Numerous approaches which rely on a low-rank approximation to $\mathbf{K_{ff}}$ have been proposed (e.g. Quiñonero Candela & Rasmussen [2005] or Rahimi & Recht [2008]), which allow the log-determinant and 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 the KL divergence [Matthews et al., 2016] to the posterior process from an approximate GP of the form

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

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 $q(\mathbf{u}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Z, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ are variational parameters.

Titsias [2009] found the minimum of the convex optimization problem for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ explicitly, resulting in the following evidence lower bound (ELBO):

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

where $\mathbf{Q}_n = \mathbf{Q_{ff}} + \sigma_n^2\mathbf{I}$, $\mathbf{Q_{ff}} = \mathbf{K_{uf}}^\intercal\mathbf{K_{uu}}^{-1}\mathbf{K_{uf}}$ and $t = \operatorname{Tr}(\mathbf{K_{ff}} - \mathbf{Q_{ff}})$. Matthews et al. [2016] showed that the KL divergence from the approximate GP posterior (eq. 2) to the 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 marginal likelihood (eq. 1), this framework jointly maximizes the ELBO (eq. 3) w.r.t. the variational and hyperparameters. This comes at the cost of introducing bias in hyperparameter estimation [Turner & Sahani, 2011], notably the overestimation of the σ_n^2 [Bauer et al., 2016]. Adding inducing points reduces the KL gap [Titsias, 2009], and the bias is practically eliminated when enough inducing variables are used.

2.3. Interdomain Inducing Features

Lázaro-Gredilla & Figueiras-Vidal [2009] showed that one can specify the distribution $q(\mathbf{u})$, 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

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

When $g(\mathbf{x}; \mathbf{z}_m) = \delta(\mathbf{x} - \mathbf{z}_m)$ the u_m are inducing points. Interdomain features require replacing $\mathbf{k}_{\mathbf{u}\cdot}$ and $\mathbf{K_{uu}}$ in Equation 2 with integral transforms of the kernel. In later sections, we 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 shows 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 $\mathcal{O}(NM^2)$:

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

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,* $\mathcal{K}$, captures the limiting properties of $\mathbf{K_{ff}}$ for large N. It is defined by

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

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

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

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 require the upper bound for candidate hyperparameters to be below the current lower bound. Eliminating all choices of hyperparameters using this approach is not feasible without 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 this work, 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 using inducing features that rely on spectral knowledge of the covariance matrix or the associated operator. These features 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 $\mathbf{y}$, derived by looking at the difference between $\mathcal{L}_{\text{upper}}$ and $\mathcal{L}_{\text{lower}}$. We will use these bounds in later sections to analyze asymptotic convergence properties.

Lemma 1. *Let* $\widetilde{\mathbf{K}}_{\mathbf{ff}} = \mathbf{K_{ff}} - \mathbf{Q_{ff}}$, $t = \operatorname{Tr}(\widetilde{\mathbf{K}}_{\mathbf{ff}})$ *and* $\widetilde{\lambda}_{max}$ *denote the largest eigenvalue of* $\widetilde{\mathbf{K}}_{\mathbf{ff}}$. *Then,*

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

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 $\mathbf{y}$

Lemma 1 is typically overly pessimistic. It assumes $\mathbf{y}$ is parallel to 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*

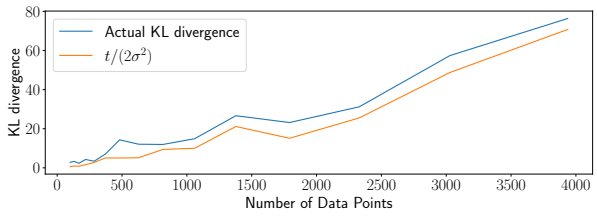

Figure 1. As N increases for fixed M the expected KL divergence increases. $t/2\sigma_n^2$ is a lower bound for the expected value over the KL divergence when $\mathbf{y}$ is generated according to our prior model.

model, then

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

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.

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

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 in 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 $\mathbf{K_{ff}}$. 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 $\mathbf{K_{ff}}$. We show in Appendix C,

$$\mathrm{cov}(u_m, u_k) = \mathbf{w}^{(m)\mathsf{T}}\mathbf{K_{ff}}\mathbf{w}^{(k)} = \lambda_k(\mathbf{K_{ff}})\delta_{m,k},$$

and

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

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 $\mathbf{K_{uf}}$ and $\mathbf{K_{uu}}$. $\mathbf{K_{uu}}$ contains the first M eigenvalues of $\mathbf{K_{ff}}$, $\mathbf{K_{uf}}$ 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]. With these features $\mathbf{Q_{ff}}$ is the *optimal* rank-M approximation to $\mathbf{K_{ff}}$ and leads to

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

3.4. Eigenfunction Inducing Features

We now modify the construction given in Section 3.3 to no longer depend on $\mathbf{K_{ff}}$ explicitly (which depends on 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 $\frac{1}{N}\mathbf{K_{ff}} \to \mathcal{K}$. This leads us to replace the eigenvalues, $\{\lambda_m(\mathbf{K_{ff}})\}_{m=1}^M$, with the operator eigenvalues, $\{\lambda_m\}_{m=1}^M$, and the eigenvectors, $\{\mathbf{w^{(m)}}\}_{m=1}^M$, with the eigenfunctions, $\{\phi_m\}_{m=1}^M$, yielding

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

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

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

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 the closed form expressions for eigenfunctions and values are known [Zhu et al., 1997]. For Matérn kernels with inputs uniform on $[a, b]$, expressions for the eigenfunctions and eigenvalues needed in order to compute $\mathbf{K_{uf}}$ and $\mathbf{K_{uu}}$ can be found in Youla [1957]. However, the formulas involve solving systems of transcendental equations limiting the practical applicability of these features for Matérn kernels.

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

Having developed the necessary preliminary results, we turn our attention to 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 $p(\mathbf{x})$, with probability at least $1-\delta$,*

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

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 if $\mathbf{y}$ is distributed according to a sample from the prior generative model, with probability at least $1-\delta$,*

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

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 $\mathbf{Q_{ff}}$ shows that $[\mathbf{Q_{ff}}]_{i,j} = \sum_{m=1}^{M}\lambda_m\phi_m(\mathbf{x}_i)\phi_m(\mathbf{x}_j)$. 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,

$$\mathbb{E}_{\mathbf{X}}[t] = N\sum_{m=M+1}^{\infty}\lambda_m\mathbb{E}_{\mathbf{x}}\left[\phi_m^2(\mathbf{x})\right] = N\sum_{m=M+1}^{\infty}\lambda_m. \tag{12}$$

The second equality follows from the eigenfunctions having unit norm. Combining Equation 12 with Lemmas 1 and 2, and Markov's Inequality yields Theorems 1 and 2 respectively. □

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, 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 $k+1/2$ kernel, $\lambda_m \asymp m^{-2k-2}$ [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 $\lim_{N\to\infty}\frac{N}{M^{2k+1}}\to 0$. This holds if $M = N^{\alpha}$ for $\alpha > \frac{1}{2k+1}$. For $k > 0$, this bound indicates the number of inducing features can grow sublinearly with the amount of data.

4. Bounds for Inducing Points

We have shown that using spectral knowledge of either $\mathbf{K_{ff}}$ or $\mathcal{K}$ we obtain bounds on the KL divergence indicating that the number of inducing features can be much smaller than the number of data points. While mathematically convenient, the practical applicability of the interdomain features used 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 difficulties in parameterizing the 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 $\mathbf{K_{ff}} \approx \mathbf{C}\overline{\mathbf{C}}^{-1}\mathbf{C}^{\intercal}$, 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 $\mathbf{K_{uf}^{\intercal}} = \mathbf{C}$, so $\mathbf{Q_{ff}} = \mathbf{C}\overline{\mathbf{C}}^{-1}\mathbf{C}^{\intercal}$.

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,*

$$\mathbb{E}_Z[\mathrm{Tr}(\mathbf{K_{ff}}-\mathbf{Q_{ff}})] \leq (M+1)\sum_{m=M+1}^{N}\lambda_m(\mathbf{K_{ff}}). \tag{13}$$

This lemma tells us that we expect well-initialized inducing

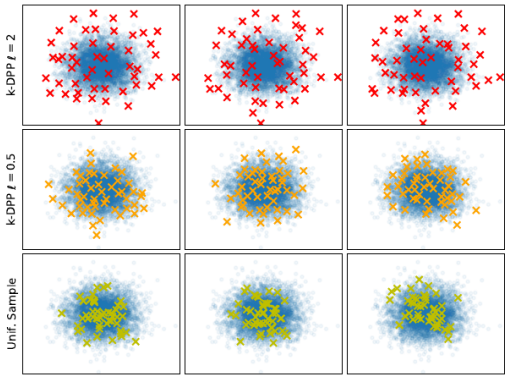

Figure 2. Determinant based sampling, with a SE kernel with $\ell = 2$ (top) and with $\ell = .5$ (middle) leads to better spacing than uniform sampling (bottom). The bounds proven hold when the kernel used for initializing points is the same as the kernel used in inference.

points to perform at within a factor of $M+1$ the eigenvector inducing features.

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 the variance of a noiseless GP fit on the already selected inducing points at candidate points. Details of the algorithm are given in Appendix D.

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

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

Theorem 3. *Suppose N training inputs are drawn i.i.d according to input density $p(\mathbf{x})$. Sample M inducing points from the training data 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 $1-\delta$,*

$$\mathrm{KL}\left(Q\|\hat{P}\right) \leq \frac{C(M+1)}{2\sigma_n^2\delta}\left(1+\frac{\|\mathbf{y}\|_2^2}{\sigma_n^2}\right) \tag{14}$$

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

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

$$\mathrm{KL}\left(Q\|\hat{P}\right) \leq C(M+1)(\delta\sigma_n^2)^{-1}. \tag{15}$$

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

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

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.

We apply Markov's inequality, yielding for any $\delta \in (0,1)$ with probability at least $1-\delta$,

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

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

We now investigate implications of our main results 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. *Suppose that* $\|\mathbf{y}\|_2^2 \leq RN$. *Fix* $\epsilon > 0$. *Assume the input data is normally distributed and regression in performed with a SE-kernel. Under the assumptions of Theorem 3, with probability* $1 - \delta$,

$$\mathrm{KL}\left(Q\|\hat{P}\right) \leq N^{-\epsilon}\left(\frac{R}{\sigma_n^2} + \frac{1}{N}\right). \tag{16}$$

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

The proof is given in Appendix E. If the lengthscale is much shorter than the standard deviation of the data then B will be near 1, implying that M will need to be large in order for the bound to converge.

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

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

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

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.

For the SE-kernel and Gaussian inputs, the rate that we prove M must increase for inducing points and eigenfunction features differs by a constant factor. For the Matérn $k + 1/2$ kernel, 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 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, 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*, i.e a kernel than can be written as a product of kernels along each dimension. If this choice of prior is made, and input densities factor over the dimensions, 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, we obtain an analogous statement to Corollary 1 in D-dimensions.

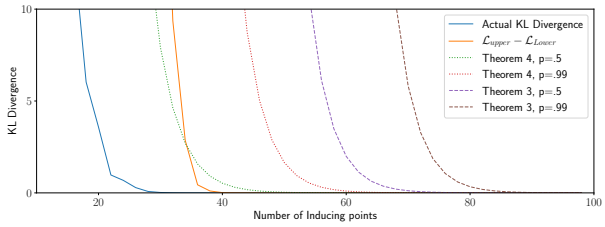

Figure 3. Rates of convergence as M increases on fixed dataset of size $N = 1000$, with a squared exponential kernel with $\ell = .3, v = 1, \sigma_n = 1$ and $x \sim \mathcal{N}(0,1)$ and $\mathbf{y}$ sampled from the prior.

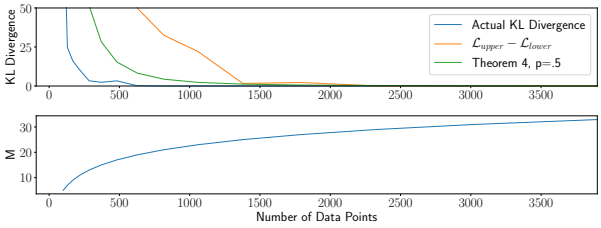

Figure 4. Corollary 1 states that for $M = C\log(N)$ the KL divergence that decays at least as fast as an inverse power of N.

Corollary 2. *For any fixed* $\epsilon, \delta > 0$ *under the assumptions of Corollary 1, with a 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* $1 - \delta$, $\mathrm{KL}\left(Q\|\hat{P}\right) \leq \delta^{-1}\epsilon$.

The proof uses ideas from Seeger et al. [2008] and is given in Appendix E. While for the SE-kernel and Gaussian input density M can grow polylogarithmically in N, and the KL divergence still converges, this is not the case for regression with other kernels or input distribution.

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. For Lebesgue measure on $[a, b]$ the Sacks-Ylivasker conditions of order r (Appendix F), which can be roughly thought of as meaning that realizations of the process are r

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

KERNEL	INPUT DISTRIBUTION	DECAY OF λ_m	M, THEOREM 3	M, THEOREM 4
SE-KERNEL	COMPACT SUPPORT	$\mathcal{O}\big(\exp(-\alpha \frac{m}{d} \log \frac{m}{d})\big)$	$\mathcal{O}(\log^D(N))$	$\mathcal{O}(\log^D(N))$
SE-KERNEL	GAUSSIAN	$\mathcal{O}\big(\exp(-\alpha \frac{m}{d})\big)$	$\mathcal{O}(\log^D(N))$	$\mathcal{O}(\log^D(N))$
MATÉRN K+1/2	UNIFORM ON INTERVAL	$\mathcal{O}\Big(M^{-2k-2} \log(M)^{2(d-1)(k+1)}\Big)$	$\mathcal{O}\Big(N^{1/k+\epsilon_D}\Big)$	$\mathcal{O}\Big(N^{1/(2k)+\epsilon_D}\Big)$

times differentiable with probability 1 [Ritter et al., 1995], implies an eigendecay of $\lambda_m \asymp m^{-2r-2}$. Table 1 summarizes the spectral decay of several stationary kernels, as well as the implications for the number of inducing points needed for inference to provably converge with our bounds.

5.3. Pointwise Approximate Posterior

In applications, often pointwise estimates of the posterior mean and variance are of interest. It is therefore desirable to show that the approximate variational posterior gives similar estimates of these quantities as the true posterior.

Huggins et al. [2019] derived an approximation method for sparse GP inference with provable guarantees about pointwise mean and variance estimates of the posterior process. They show that approximations with moderately small KL divergences can still have large deviations in 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 converge to the posterior values.

By the chain rule of KL divergence [Matthews et al., 2016],

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

In words, the KL divergence between posterior processes upper bounds the KL divergence between any of the posterior marginals. Therefore, bounds on the mean and variance of a one-dimensional Gaussian with a small KL divergence imply pointwise guarantees about posterior inference when the KL divergence between processes is small.

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 $2\mathrm{KL}(q\|p) = \epsilon \leq \frac{1}{5}$, then*

$$|\mu_1 - \mu_2| \leq \sigma_2 \sqrt{\epsilon} \leq \frac{\sigma_1 \sqrt{\epsilon}}{\sqrt{1 - \sqrt{3\epsilon}}} \text{ and } |1 - \sigma_1^2/\sigma_2^2| < \sqrt{3\epsilon}.$$

The proof is in Appendix B. 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], lead to similar conclusions about the choice of approximating rank. Ferrari-Trecate et al. [1999] showed that given N data points, using a rank M truncated SVD of the prior covariance matrix, such that $\lambda_M \ll \sigma_n^2/N$ results in almost no change in the model, 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.

Alaoui & Mahoney [2015] showed using a Nyström approximation, sampled according to the 'leverage scores', to the covariance matrix of ridge regression leads to a model that converges in mean square error when the number of columns scales with the effective dimensionality of the problem. The effective dimensionality is determined by the decay of eigenvalues of $\mathbf{K_{ff}}$ and the parameter of ridge regression. This leads to similar conclusion as the results of Ferrari-Trecate et al. [1999] mentioned previously, when the parameter of ridge regression is such 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, bounding the additional error introduced by sparsity in the framework of Hensman et al. [2015], pose a promising direction for future research.

Acknowledgements

We would like to thank the reviewers for their helpful feedback and suggestions.

References

Alaoui, A. and Mahoney, M. W. Fast Randomized Kernel Ridge Regression with Statistical Guarantees. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pp. 775–783. 2015.

Bauer, M., van der Wilk, M., and Rasmussen, C. E. Understanding probabilistic sparse Gaussian process approximations. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1533–1541, 2016.

Belabbas, M.-A. and Wolfe, P. J. Spectral Methods in Machine Learning and new Strategies for very Large Datasets. In *Proceedings of the National Academy of Sciences (PNAS)*, volume 106, pp. 369–374, 2009.

Ferrari-Trecate, G., Williams, C. K., and Opper, M. Finite-dimensional Approximation of Gaussian Processes. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 218–224, 1999.

Hennig, P. and Garnett, R. Exact Sampling from Determinantal Point Processes. *arXiv preprint arXiv:1609.06840*, 2016.

Hensman, J., Matthews, A., and Ghahramani, Z. Scalable Variational Gaussian Process Classification. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 351–360, 2015.

Hensman, J., Durrande, N., and Solin, A. Variational Fourier Features for Gaussian Processes. In *Journal of Machine Learning Research*, volume 18, pp. 1–52, 2018.

Huggins, J. H., Campbell, T., Kasprzak, M., and Broderick, T. Scalable Gaussian Process Inference with Finite-data Mean and Variance Guarantees. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Koltchinskii, V. and Giné, E. Random Matrix Approximation of Spectra of Integral Operators. In *Bernoulli*, volume 6, pp. 113–167, 2000.

Lanczos, C. An Iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integral Operators. In *Journal of Research of the National Bureau of Standards*, pp. 255–282, 1950.

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

Matthews, A. G. d. G., Hensman, J., Turner, R., and Ghahramani, Z. On Sparse Variational Methods and the Kullback-Leibler Divergence between Stochastic Processes. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 231–239, 2016.

Mercer, J. Functions of Positive and Negative Type, and their Connection the Theory of Integral Equations. In *Phil. Trans. R. Soc. Lond. A*, volume 209, pp. 415–446. The Royal Society, 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. In *Journal of Machine Learning Research*, volume 6, pp. 1939–1959, 2005.

Rahimi, A. and Recht, B. Random Features for Large-scale Kernel Machines. In *Advances in Neural Information Processing Systems (NIPS)*, 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 Woźniakowski, H. Multivariate Integration and Approximation for Random Fields Satisfying Sacks-Ylvisaker Conditions. In *The Annals of Applied Probability*, volume 5, pp. 518–540. Institute of Mathematical Statistics, 1995.

Seeger, M. W., Kakade, S. M., and Foster, D. P. Information Consistency of Nonparametric Gaussian Process Methods. In *IEEE Transactions on Information Theory*, volume 54, pp. 2376–2382. IEEE, 2008.

Titsias, M. K. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 567–574, 2009.

Titsias, M. K. Variational Inference for Gaussian and Determinantal Point Processes. In *Workshop on Advances in Variational Inference (NIPS)*, December 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. I. In *Transactions of the American Mathematical Society*, volume 109, pp. 278–295. American Mathematical Society, 1963.

Widom, H. Asymptotic behavior of the eigenvalues of certain integral equations. II. In *Archive for Rational Mechanics and Analysis*, volume 17, pp. 215–229. Springer, 1964.

Williams, C. K. I. and Seeger, M. Using the Nyström Method to Speed Up Kernel Machines. In *Advances in Neural Information Processing Systems (NIPS) 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. In *IRE Transactions on Information Theory*, volume 3, pp. 187–193. IEEE, 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.