Journal of Machine Learning Research 21 (2020) 1-63 Submitted 12/19; Published 7/20

Convergence of Sparse Variational Inference in Gaussian Processes Regression

David R. Burt DRB62@CAM.AC.UK
Carl Edward Rasmussen CER54@CAM.AC.UK
Department of Engineering, University of Cambridge, UK

Mark van der Wilk M.VDWILK@IMPERIAL.AC.UK
Department of Computing, Imperial College London, UK
*Prowler.io,*Cambridge, UK*

Editor: Kilian Weinberger

Abstract

Gaussian processes are distributions over functions that are versatile and mathematically convenient priors in Bayesian modelling. However, their use is often impeded for data with large numbers of observations, N, due to the cubic (in N) cost of matrix operations used in exact inference. Many solutions have been proposed that rely on $M \ll N$ *inducing variables* to form an approximation at a cost of $\mathcal{O}(NM^2)$. While the computational cost appears linear in N, the true complexity depends on how M must scale with N to ensure a certain quality of the approximation. In this work, we investigate upper and lower bounds on how M needs to grow with N to ensure high quality approximations. We show that we can make the KL-divergence between the approximate model and the exact posterior arbitrarily small for a Gaussian-noise regression model with $M \ll N$. Specifically, for the popular squared exponential kernel and D-dimensional Gaussian distributed covariates, $M = \mathcal{O}((\log N)^D)$ suffice and a method with an overall computational cost of $\mathcal{O}\big(N(\log N)^{2D}(\log\log N)^2\big)$ can be used to perform inference.

Keywords: Gaussian processes, approximate inference, variational methods, Bayesian non-parameterics, kernel methods

1. Introduction

Gaussian process (GP) priors are commonly used in Bayesian modelling due to their mathematical convenience and empirical success. The resulting models give flexible mean predictions, as well as useful estimates of uncertainty. GP priors are often used with a Gaussian likelihood for regression tasks, as the Bayesian posterior can be computed in closed form in this case. Additionally, in many instances, the kernel is differentiable with respect to hyperparameters, in which case hyperparameters can be efficiently learned using gradient-based optimization by maximizing the marginal likelihood, which can be computed analytically (also known as empirical Bayes, or type-II maximum likelihood). However, standard implementations of exact inference in Gaussian process regression models require storing and inverting a kernel matrix, imposing an $\mathcal{O}(N^2)$ memory cost and an $\mathcal{O}(N^3)$ computational cost, where N is the number of training examples. These computational constraints have

∗. Previous affiliation where significant portion of work was completed.

pushed researchers to adopt approximate methods in order to allow Gaussian process models to scale to large data sets.

Sparse methods (e.g. Seeger et al., 2003; Snelson and Ghahramani, 2006; Titsias, 2009b) rely on a set of *inducing variables* to represent the posterior distribution. While these methods have been widely adopted in research and application areas, there is a limited theoretical understanding of the effects of these approximations on the quality of posterior predictions, as well as what biases are introduced into hyperparameter selection when using approximations to the marginal likelihood. In this work, we aim to characterize the accuracy of sparse approximations. If all of the key properties of the exact model, i.e. the predictive mean and uncertainties and the marginal likelihood, are maintained by very sparse models, then a great deal of computation can be saved through these approximations.

We focus on the case of sparse inference in the variational framework of Titsias (2009b). We analyze the relationship between the level of sparsity used in performing inference, which dictates the computational cost, and the quality of the approximate posterior distribution. In particular, we analyze how many inducing variables should be used in order for the KL-divergence between the approximate posterior and the Bayesian posterior to be small. This offers theoretical insight into the trade-off between computation and quality of inference within the variational framework. From a practical perspective, our work suggests new methods for choosing which inducing variables to use to construct the approximation and provides theoretically grounded insight into the types of problems to which the sparse variational approach is particularly well-suited.

1.1. Our Contributions

- We derive bounds on the quality of variational inference in Gaussian process models. When our bounds are applied in the case of the squared exponential (SE) kernel and Gaussian or compactly supported inputs, we prove that the variational approximation can be made arbitrarily close to the true posterior with arbitrarily high probability using $\mathcal{O}((\log N)^D)$ inducing variables, where D is the dimensionality of the training inputs, leading to an overall computational cost of $\mathcal{O}(N(\log N)^{2D}(\log\log N)^2)$. Note that we consider D fixed throughout, implying a scaling in N that is nearly linear, i.e. $\mathcal{O}(N^{1+\epsilon})$, $\forall\epsilon > 0$.

- Our bounds measure the discrepancy to the true posterior using the KL-divergence between the approximate and exact posteriors. We also show that this implies convergence of the point-wise predictive means and variances.

- We show that theoretical guarantees on the quality of matrix approximation for existing methods for selecting regressors in sparse kernel ridge regression can be directly translated into guarantees on variational sparse GP regression. We demonstrate this for ridge leverage scores.

- We derive lower bounds on the number of inducing variables needed to ensure that the KL-divergence remains small. For the SE kernel and Gaussian covariate distribution, these lower bounds have the same dependence on the size of the data set as the upper bounds.

- Based on the theoretical results, we provide recommendations on how to select inducing variables in practice, and demonstrate empirical improvements.

This paper is an extension of the work Burt et al. (2019) presented at ICML 2019.

1.2. Overview of this Paper

In Section 2 we introduce notation and review the Gaussian process regression model, as well as sparse variational inference for Gaussian process models. In Section 3, we discuss practical considerations regarding assessing the quality of sparse variational inference using upper bounds on the log marginal likelihood that can be computed after observing a data set. In Section 4, we prove our main results, which bound the quality of the sparse approximate posterior, as measured by the KL-divergence. In order to do this, we consider methods for selecting inducing inputs inspired by methods used to obtain theoretical guarantees on sparse kernel ridge regression. Section 5 considers specific, commonly studied kernels and covariate distributions and investigates the implications of our results in these instances. We provide concrete computational complexities for finding arbitrarily accurate approximations to GPs. In Section 6, we consider the inverse problem, and show that in certain instances the KL-divergence will be large unless the number of inducing variables increases sufficiently quickly as a function of the size of the data set. Section 7 discusses practical insights and limitations of the theory as applied to real-world problems.

2. Background and Notation

In this section, we review exact inference in Gaussian process models, as well as sparse methods for approximate inference in these models. We particularly focus on the formulation of sparse methods based on variational inference (Titsias, 2009b). Throughout the paper, we use boldface letters to denote random variables, and the same letter in non-bold to denote a realization of this random variable. We follow the standard shorthand notation adopted in many Bayesian machine learning papers and denote probability densities by lower case letters p and q, with the distribution to which they are associated inferred by the name of the argument; e.g. $p(X, y)$ is the density of a joint distribution over random variables $\mathbf{X}$ and $\mathbf{y}$ evaluated at $\mathbf{X} = X$ and $\mathbf{y} = y$.

2.1. Gaussian Processes

A Gaussian process is a collection of real-valued random variables indexed by a set $\mathcal{X}$, such that any finite collection of these random variables is jointly Gaussian distributed. While most commonly $\mathcal{X}$ is a subset of $\mathbb{R}^D$, Gaussian processes can be indexed by other sets. Such a process can be viewed as defining a distribution over functions $\mathbf{f} : \mathcal{X} \to \mathbb{R}$, for which the distribution of function values for a finite set of inputs is Gaussian.

A procedure for specifying the first two moments of any finite marginal distribution in a consistent manner defines a GP. This can be done by selecting a *mean function* $\mu : \mathcal{X} \to \mathbb{R}$ and a symmetric, positive semi-definite *covariance function* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The finite dimensional marginal indexed by $X = (x_1, \ldots, x_N)^\intercal \subset \mathcal{X}$, denoted $\mathbf{f}_X$, is distributed as

$$\mathbf{f}_X \sim \mathcal{N}(\mu_X, \mathrm{K}_{\mathrm{ff}}), \tag{1}$$

with $\mu_X = (\mu(x_1), \ldots, \mu(x_N))^\intercal$ and $\mathrm{K_{ff}}$ an $N \times N$ matrix with $[\mathrm{K_{ff}}]_{n,n'} = k(x_n, x_{n'})$. Properties such as smoothness, variance and characteristic lengthscale of functions that are sampled from the GP are determined by the covariance function. The covariance function is often parameterized in such a way that these properties can be adjusted based on properties of the observed data.

2.2. Gaussian Process Regression

In this work we perform Bayesian regression using a Gaussian process as the prior distribution over the function we want to learn. We observe a data set of N training examples, $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$ with $x_n \in \mathcal{X}$ and $y_n \in \mathbb{R}$ and want to infer a posterior distribution over functions $\mathbf{f} : \mathcal{X} \to \mathbb{R}$ that relate the inputs to the outputs. We define $X = (x_1, \ldots, x_N)^\intercal$, $y = (y_1, \ldots, y_N)^\intercal$ and $\mathbf{f}_X = (\mathbf{f}(x_1), \cdots \mathbf{f}(x_N))^\intercal$. More generally for any finite, set $X' \subset \mathcal{X}, |X'| = S < \infty$, which we assume has a fixed ordering, we will use $\mathbf{f}_{X'}$ to denote the (random) vector in $\mathbb{R}^S$, formed by considering the Gaussian process at indices $x \in X'$.

We specify our Bayesian model through a prior and likelihood. We place a GP prior, which for notational convenience we assume has zero mean function i.e. $\mu \equiv 0$, on the function $\mathbf{f}$ so that

$$\mathbf{f} \sim \mathcal{GP}(0, k). \tag{2}$$

To allow for deviations from $\mathbf{f}$ in the observations, we model the data y as a noisy observation of this process through the likelihood

$$\mathbf{y} \,|\, \mathbf{f}_X \sim \mathcal{N}(\mathbf{f}_X, \sigma^2 \mathrm{I}), \tag{3}$$

where the noise variance σ^2, is a model hyperparameter and I is an $N \times N$ identity matrix.

Since the likelihood and the prior are conjugate in this model, Bayesian inference can be performed in closed form. The posterior density over the latent function values at any finite collection of T new data points $X^\star = (x_1^\star, \cdots, x_T^\star)^\intercal$ is given by

$$\begin{aligned} p(f_{X^\star} \,|\, \mathcal{D}) &= \int_{f_X \in \mathbb{R}^N} p(f_{X^\star}, f_X \,|\, \mathcal{D}) \,\mathrm{d}f_X \\ &= \int_{f_X \in \mathbb{R}^N} p(f_{X^\star} \,|\, f_X) p(f_X \,|\, \mathcal{D}) \,\mathrm{d}f_X. \end{aligned} \tag{4}$$

Both $p(f_{X^\star} \,|\, f_X)$ and $p(f_X \,|\, \mathcal{D})$ are Gaussian densities and the marginal distribution of a Gaussian is Gaussian, so $p(f_{X^\star} \,|\, \mathcal{D})$ is also a Gaussian density. The posterior predictive distribution over the inputs $\mathbf{f}_{X^\star}$ has mean vector and covariance matrix

$$\widehat{\mu}_\star = \mathrm{K}_{\star\mathrm{f}}(\mathrm{K_{ff}} + \sigma^2 \mathrm{I})^{-1} y \quad \text{and} \quad \widehat{\Sigma}_{\star\star} = \mathrm{K}_{\star\star} - \mathrm{K}_{\star\mathrm{f}}(\mathrm{K_{ff}} + \sigma^2 \mathrm{I})^{-1} \mathrm{K}_{\star\mathrm{f}}^\intercal, \tag{5}$$

where $\mathrm{K}_{\star\mathrm{f}}$ is $T \times N$ matrix with $[\mathrm{K}_{\star\mathrm{f}}]_{t,n} = k(x_t^\star, x_n)$ and $\mathrm{K}_{\star\star}$ is a $T \times T$ matrix with $[\mathrm{K}_{\star\star}]_{t,t'} = k(x_t^\star, x_{t'}^\star)$.

The *marginal likelihood* is of interest in Bayesian models for selecting the properties of the model, which are determined by hyperparameters. Point estimates of model hyperparameters are commonly obtained by maximizing the marginal likelihood with respect to the

noise variance σ^2, and any parameters of the prior covariance function k. In the case of conjugate regression described above, the log marginal likelihood takes the form

$$\log p(y) = -\frac{1}{2}\log\det(\mathrm{K_{ff}} + \sigma^2\mathrm{I}) - \frac{1}{2}y^\intercal(\mathrm{K_{ff}} + \sigma^2\mathrm{I})^{-1}y - \frac{N}{2}\log 2\pi. \tag{6}$$

The quadratic term measures how well the data y lines up with degrees of variation that are allowed under the prior. The log-determinant term measures how much variation there is in the prior, and penalizes priors which are widely spread. The combination of these terms in the log marginal likelihood balances the ability of the model to fit the data with model complexity, which allows a suitable model to be chosen; see Rasmussen and Williams (2006) for more discussion of the marginal likelihood as a tool for model selection as well as an introduction to Gaussian processes.

Despite closed-form expressions for both the predictive posterior (Eq. 5) and marginal likelihood (Eq. 6), exact inference in Gaussian process regression models is impractical for large data sets due to the cost of storing and inverting the kernel matrix $\mathrm{K_{ff}}$, leading to $\mathcal{O}(N^2)$ memory and $\mathcal{O}(N^3)$ time complexities. Sparse approximations have been widely adopted to address this issue.

2.3. Approximate Inference for Gaussian Processes

Approximate inference in Gaussian process regression is performed for a different reason than in most Bayesian models. Approximate inference is usually applied when the exact posterior is *analytically* intractable. In our case, we can analytically write down the posterior, but the cost of computation is often prohibitive. The methods we discuss here all approximate the posterior with a different Gaussian process which has more favorable computational properties. As this approximate posterior has a similar form to the exact posterior, and we can control the trade-off between accuracy and computation, it is plausible that our approximation may be very accurate.

2.3.1. Inducing Variable Methods

The large cost of computing the posterior GP comes from needing to infer a Gaussian distribution for the function values at all N input locations. Sparse approximations (Seeger et al., 2003; Snelson and Ghahramani, 2006; Titsias, 2009b) avoid this cost by instead computing an approximate posterior that only depends on the data through the process at $M \ll N$ locations.

The aim of these methods is to compress the combined effect of a large number of input and output pairs into a distribution over function values at a small set of inputs. In regions where data is dense, there is often redundant information about what the function is actually doing, so little is lost in performing this approximation. The selected input locations and their corresponding function values are named inducing inputs and outputs respectively, and together are named *inducing points*. Later, it was suggested that more general linear transformations of the process could also be used to compress knowledge into (Lázaro-Gredilla and Figueiras-Vidal, 2009). We generally refer to these approaches as *inducing variable* methods. In all of these methods, a low-rank matrix appears in place of $\mathrm{K_{ff}}$ in the computation of the posterior predictive and log marginal likelihood. This matrix can be manipulated with a much lower computational cost than working with $\mathrm{K_{ff}}$ directly.

The success of inducing variable methods depends heavily on *which* M random variables are chosen to represent the knowledge about the function f. Because in this work we are concerned with characterizing how large M should be, we need a good method for choosing the inducing variables, as well as a meaningful criterion for judging the quality of the resulting approximation. The variational formulation of Titsias (2009b) is of particular interest, as it uses a well-defined divergence for characterizing the quality of the posterior, which can also be used as a guide for selecting the inducing variables.

2.3.2. The Variational Formulation

Variational inference proceeds by defining a family of candidate distributions $\mathcal{Q}$, and then selecting the distribution $Q \in \mathcal{Q}$ that minimizes the KL-divergence between the approximation and the posterior. In practice, elements of $\mathcal{Q}$ are parameterized and the approximate posterior is selected by choosing an initial approximation which is then refined by finding a local minimum of the KL-divergence as a function of the variational parameters. In variational GP methods (Titsias, 2009b; Hensman et al., 2013) $\mathcal{Q}$ consists of GPs with finite dimensional marginal densities of the form

$$q(f_{X'}, U) = q(U)p(f_{X'} \,|\, U)\,, \tag{7}$$

for any $X' \subset \mathcal{X}, |X'| < \infty$, where q is the density of the approximate posterior at this collection of points, and $p(f_{X'} \,|\, U)$ is the density of the prior distribution of $\mathbf{f}_{X'}$ at $f_{X'}$ conditioned on the random variables $\mathbf{U}$ evaluated at $\mathbf{U} = U$. In inducing point approximations, we take the inducing variables to be point evaluations of f, i.e. $\mathbf{U} = \mathbf{f}_Z$, with inducing inputs $Z \subset \mathcal{X}$ and $|Z| = M$.

As discussed in the previous section, we can also define inducing variables as linear transformations of the prior process of the form

$$\mathbf{u}_m = \int_{\mathcal{X}} g_m(x)\mathbf{f}(x)\,\mathrm{d}\rho(x),$$

where we assume ρ is a measure on $\mathcal{X}$ defined with respect to an appropriate σ-algebra and $g_m \in L^1(\mathcal{X}, \rho)$. If ρ is taken to be a discrete measure, then these features correspond to (weighted) sums of inducing points; while other forms of these inducing variables of this form have been explored (Lázaro-Gredilla and Figueiras-Vidal, 2009; Hensman et al., 2018).

The density $q(U)$ is chosen to be an M-dimensional Gaussian density. This choice of variational family induces a Gaussian process approximate posterior with mean and covariance functions

$$\mu_Q(x) = k_{f(x)\mathrm{u}}\mathrm{K}_{\mathrm{uu}}^{-1}\mu_{\mathrm{U}}\,, \quad \text{and} \quad k_Q(x, x') = k(x, x') + k_{f(x)\mathrm{u}}\mathrm{K}_{\mathrm{uu}}^{-1}(\Sigma_{\mathrm{U}} - \mathrm{K}_{\mathrm{uu}})\mathrm{K}_{\mathrm{uu}}^{-1}k_{\mathrm{u}f(x')}\,,$$

where $\mu_{\mathrm{U}}, \Sigma_{\mathrm{U}}$ are the mean and covariance of $q(U)$, K_{uu} is the $M \times M$ matrix with entries $[\mathrm{K}_{\mathrm{uu}}]_{m,m'} = \mathrm{cov}(\mathbf{u}_m, \mathbf{u}_{m'})$, $k_{f(x)\mathrm{u}}$ is the row vector with entries $[k_{f(x')\mathrm{u}}]_m = \mathrm{cov}(\mathbf{f}(x), \mathbf{u}_m)$ and $k_{\mathrm{u}f(x)}$ is a column vector defined similarly. The variational parameters consist of Z, which determines the random variables that are included in $\mathbf{U}$, and μ_{U} and Σ_{U}, which determine the distribution over $\mathbf{U}$.

As is usually the case in variational inference, minimizing the KL-divergence is done indirectly by maximizing a lower bound to the marginal likelihood, $\mathcal{L}$ (also known as the

evidence lower bound, or ELBO), which has $\mathrm{KL}[Q||P]$ as its slack:

$$\mathcal{L} + \mathrm{KL}[Q||P] = \log p(y) \implies \operatorname*{argmin}_{Q \in \mathcal{Q}} \mathrm{KL}[Q||P] = \operatorname*{argmax}_{Q \in \mathcal{Q}} \mathcal{L}(Q)\,. \tag{8}$$

where P denotes the (exact) posterior process (Matthews et al., 2016).

When the likelihood is isotropic Gaussian, the unique optimum for the parameters $\{\mu_{\mathrm{U}}, \Sigma_{\mathrm{U}}\}$ can be computed in closed form. Using these optimal values, we obtain the ELBO as it was introduced by Titsias (2009b),

$$\mathcal{L} = -\frac{1}{2}\log\det(\mathrm{Q_{ff}} + \sigma^2\mathrm{I}) - \frac{1}{2}y^{\intercal}(\mathrm{Q_{ff}} + \sigma^2\mathrm{I})^{-1}y - \frac{N}{2}\log 2\pi - \frac{1}{2\sigma^2}\mathrm{tr}(\mathrm{K_{ff}} - \mathrm{Q_{ff}})\,, \tag{9}$$

where $\mathrm{Q_{ff}} = \mathrm{K_{uf}^{\intercal}}\mathrm{K_{uu}^{-1}}\mathrm{K_{uf}}$ and $\mathrm{K_{uf}}$ is the $M \times N$ matrix with entries $[\mathrm{K_{uf}}]_{i,j} = \mathrm{cov}(\mathbf{u}_i, \mathbf{f}(x_j))$.

2.3.3. Measuring the Quality of a Variational Approximation

In order to assess whether variational inference leads to an accurate approximation to the posterior, we need to choose a definition of what it means for an approximation to be accurate. We choose to measure the quality of an approximation in terms of the KL-divergence, $\mathrm{KL}[Q||P]$. This KL-divergence is 0 if and only if the approximate posterior is equal to the exact posterior. Under this measure, an approximation is considered good if this KL-divergence is small.

Variational approximations using this KL-divergence have been criticized for failing to provide guarantees on important quantities such as posterior estimates of the mean and variance. Huggins et al. (2019) observed that there exist Gaussian distributions such that the (normalized) difference between the means of the distributions is exponentially large as a function of the KL-divergence between the two distributions, as is the ratio of the variances. This has been used to motivate variational approaches based on other notions of divergence, as well as a more careful assessment of the quality of the approximations obtained via variational inference (Huggins et al., 2020).

However, in our case of sparse Gaussian process regression, a sufficiently small KL-divergence between the approximate and true posterior implies bounds on the approximation quality of the marginal posterior mean and variance function. Proposition 1 states one such bound:

Proposition 1 *Suppose* $2\mathrm{KL}[Q||P] \leq \gamma \leq \frac{1}{5}$. *For any* $x^\star \in \mathcal{X}$, *let* μ_1 *denote the posterior mean of the variational approximation at* $x^\star$ *and* μ_2 *denote the mean of the exact posterior at* $x^\star$. *Similarly, let* σ_1^2, σ_2^2 *denote the variances of the approximate and exact posteriors at* $x^\star$. *Then,*

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

The proof (Appendix A) uses that the KL-divergence between any pair of joint distributions upper bounds the KL-divergence between marginals of these distributions. It then suffices to bound the difference between the mean and variance of univariate Gaussian distributions with a small KL-divergence between them.

Proposition 1 implies that in cases where we can prove the KL-divergence between the approximate posterior and the exact posterior is very small, we are guaranteed to obtain similar marginal predictions with the variational approximation to those we would obtain with the exact model. We note that direct approaches to bounding marginal moments may lead to tighter bounds on these quantities (e.g. Calandriello et al., 2019), but we prefer to consider the KL-divergence due to its connection to the variational objective function.

The consequences of a small KL-divergence for hyperparameter selection using the evidence lower bound are more subtle, as both the approximate posterior and exact posterior depend on model hyperparameters, and it is generally difficult to ensure that the KL-divergence is uniformly small. We will be discuss these issues in more detail in Section 7.

2.3.4. Computation and Accuracy Trade-Offs

The ELBO (Eq. 9) as well as the corresponding choices for μ_{U} and Σ_{U} (needed for making predictions) can be computed in $\mathcal{O}(NM^2)$ time, and with $\mathcal{O}(NM)$ space. If a good approximation can be found with $M \ll N$, the savings in computational cost are large compared to exact inference. From Eq. (9) we see that the approximation is perfect when choosing $Z = X$, as this leads to $\mathrm{Q}_{\mathrm{ff}} = \mathrm{K}_{\mathrm{ff}}$. However, no computation is saved in this setting. We seek a more complete understanding of the trade-off between accuracy and computational cost when $M < N$ by understanding how M needs to grow with N to ensure an approximation of a certain quality. We derive probabilistic upper and lower bounds on this rate that depend on kernel properties that can be analyzed *before* observing any data.

2.4. Spectrum of Kernels and Mercer's Theorem

In the previous section, we noted that sparse methods imply a low-rank approximation Q_{ff} to the kernel matrix K_{ff}. In order to understand the impact of sparsity on the variational posterior, it is necessary to understand how well K_{ff} can be approximated by a rank-M matrix. This depends on the behavior of the eigenvalues of K_{ff}.

For small data sets, an eigendecomposition of K_{ff} allows direct empirical analysis. However, for problems where sparse approximations are actually of interest, eigendecompositions are not available within our computational constraints. However, even without access to a specific data set, we can reason that properties of the training inputs have a large impact on the properties of the eigendecomposition of the kernel matrix. For example, consider the case of a squared exponential kernel given by

$$k_{\mathrm{SE}}(x, x') = v \exp\left(-\frac{\|x - x'\|_2^2}{2\ell^2}\right),$$

where $v > 0$ is the signal variance, which controls the variance of marginal distributions of the prior and $\ell > 0$ is the lengthscale, which controls how quickly the correlation between function values decreases as a function of the distance between their inputs. If each covariate in our training data set is sufficiently far apart relative to the lengthscale, then $\mathrm{K}_{\mathrm{ff}} \approx v\mathrm{I}$ and any approximation by low-rank matrix will be of poor quality. Alternatively, if each covariate takes the same value, K_{ff} is a rank-one matrix and $\mathrm{Q}_{\mathrm{ff}} = \mathrm{K}_{\mathrm{ff}}$ if a single inducing point is placed at the location of the covariates. Therefore, in order to make statements

about the eigenvalues of $\mathrm{K_{ff}}$, we will need to make assumptions about the locations of training covariates X.

For the remainder of the paper, we assume $\mathcal{X} = \mathbb{R}^D$ (the generalization of most of our results is straightforward). One method for understanding the eigenvalues of $\mathrm{K_{ff}}$ is to suppose the x_i are realizations of random variables $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mu_x$, where μ_x is a probability measure on $\mathbb{R}^D$ with density $p(x)$. Under this assumption, the limiting properties of the kernel matrix are captured by the *kernel operator*, $\mathcal{K} : L^2(\mathbb{R}^D, \mu_x) \to L^2(\mathbb{R}^D, \mu_x)$ defined by

$$(\mathcal{K}g)(x) = \int g(x')k(x, x')p(x') \, \mathrm{d}x'. \tag{10}$$

If the kernel is continuous and bounded, then $\mathcal{K}$ has countably many eigenvalues. We denote these eigenvalues in non-increasing order, so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. Corresponding to each non-zero eigenvalue λ_m there is an eigenfunction ϕ_m which can be chosen to be continuous.

Moreover, the collection $\{\phi_m\}_{m=1}^{\infty}$ can be chosen so that the eigenfunctions all have unit norm and are mutually orthogonal as elements of $L^2(\mathbb{R}^D, \mu_x)$. In this case, Mercer's theorem (Mercer, 1909; König, 1986) states that for sufficiently nice k, for all $x \in \mathbb{R}^D$ such that $p(x) > 0$,

$$k(x, x') = \sum_{m=1}^{\infty} \lambda_m \phi_m(x) \phi_m(x') \quad \text{and} \quad \sum_{m=1}^{\infty} \lambda_m < \infty, \tag{11}$$

where the sum on the left converges absolutely and uniformly.[1]

The bounds we derive in the remainder of this work will depend on how rapidly the eigenvalues $\{\lambda_m\}_{m=1}^{\infty}$ decay. As they are absolutely summable, they must decay faster than $1/m$. The decay of these eigenvalues is closely related to the complexity of the non-parametric model as well as the generalization properties of the posterior (Micchelli and Wahba, 1979; Plaskota, 1996). Generally, these eigenvalues decay faster for covariate distributions that are concentrated in a small volume, and for kernels that give smooth mean predictors (Widom, 1963, 1964). Therefore, the bounds we prove in Section 4 can be seen as verifying the intuition that sparse variational approximations can be successfully applied to models with smooth prior kernels, as well as data sets with densely clustered covariates.

2.5. Inducing Variable Selection and Related Bounds

While the kernel eigenvalues determine how well a kernel matrix *can* be approximated, the quality of an actual approximation depends on how the inducing variables are chosen. Inducing point selection has been widely studied for many methods that require constructing a Nystrm approximation, like sparse Gaussian processes and kernel ridge regression (KRR). In the simplest case, a subset can be uniformly sampled from the training inputs. Bounds on the quality of the resulting matrix approximation, and downstream Kernel Ridge Regression predictor have been found for this case (Bach, 2013; Gittens and Mahoney, 2016) and depend heavily on assumptions about the covariate distribution and resulting kernel matrix. In the Gaussian process literature, some specific low-rank parametric approximations based on spectral information about the kernel operator or matrix have been proposed (Zhu et al.,

1. See Rasmussen and Williams (2006), section 4.3 for more discussion of Mercer's theorem.

1997; Ferrari-Trecate et al., 1999; Solin and Srkk, 2020) together with analysis on the rate of decrease in error with additional features. However, these methods generally are either limited in the types of kernels they can be applied to or have higher computational complexity than inducing point methods.

Heuristic inducing point selection methods have also been proposed in the hope of improving performance, for instance approximately minimizing $\mathrm{tr}(\mathrm{K}_{\mathrm{ff}} - \mathrm{Q}_{\mathrm{ff}})$ (Smola and Schölkopf, 2000), approximating the information gain of including a data point in the posterior (Seeger et al., 2003), or using the k-means centres of the input distribution (Hensman et al., 2013, 2015).

Two methods from the KRR literature are of particular interest: sampling from a Determinantal Point Process (DPP) (Li et al., 2016), and ridge leverage scores (Alaoui and Mahoney, 2015; Rudi et al., 2015; Calandriello et al., 2017). Theoretical guarantees exist in the literature for these methods applied to KRR, as well as empirical evidence of their efficacy compared to uniform sampling. The initial version of this work (Burt et al., 2019) analyzed convergence of the sparse variational GP posterior and marginal likelihood using the DPP initialization. Concurrently, Calandriello et al. (2019) used ridge leverage scores to show the DTC approximation (Seeger et al., 2003; Quiñonero-Candela and Rasmussen, 2005) can be made similar to the true posterior, in terms of pointwise predictive means and variances. Given the similarity between the DTC and variational posteriors, we include an analysis of ridge leverage sampling in this extended work to also provide results of convergence of the ELBO, and of the posterior in terms of the KL, which also implies pointwise convergence of the predictive means and variances.

3. Assessing Variational Inference: a Posteriori Bounds on the KL-divergence

We begin our investigation by considering how to choose the number of inducing variables for a specific data set. The simplest approach to assessing whether sufficiently many inducing points are used is to gradually increase the number of inducing points, and assess how the evidence lower bound changes with each additional point. If the ELBO increases only slightly or not at all when an additional inducing point is added, it is tempting to conclude that the approximate posterior is very close to the exact posterior. However, this is not a sufficient condition for the approximation to have converged. It could be the case that the last inducing point placed was not placed effectively, or that increasing from M to $M+1$ inducing points has little impact, but increasing to $M+c$, for some $c > 1$, inducing points would lead to significantly better performance if these points are well-placed.

A more refined mechanism for assessing the quality of the variational posterior would be to consider an upper bound on the KL-divergence that can be computed in similar computational time to the ELBO. Such a bound was proposed by Titsias (2014) and discussed as a method for assessing convergence in Kim and Teh (2018). In order to state this bound, we first need to introduce some notation. Let $t \coloneqq \mathrm{tr}(\mathrm{K}_{\mathrm{ff}} - \mathrm{Q}_{\mathrm{ff}})$ denote the trace of $\mathrm{K}_{\mathrm{ff}} - \mathrm{Q}_{\mathrm{ff}}$ and $\|\mathrm{K}_{\mathrm{ff}} - \mathrm{Q}_{\mathrm{ff}}\|_{\mathrm{op}}$ denote the operator norm of $\mathrm{K}_{\mathrm{ff}} - \mathrm{Q}_{\mathrm{ff}}$, which in this case is equal to the largest eigenvalue of this matrix as it is symmetric positive semidefinite.

Lemma 2 (Titsias, 2014) *For any $y \in \mathbb{R}^N, X \in \mathcal{X}^N$, and set of M inducing variables, U,*

$$\log p(y) \leq \mathcal{U}_1 \leq \mathcal{U}_2$$

where

$$\mathcal{U}_1 \coloneqq -\frac{1}{2}\log\det(\mathrm{Q_{ff}} + \sigma^2\mathrm{I}) - \frac{1}{2}y^T(\mathrm{Q_{ff}} + \|\mathrm{K_{ff}} - \mathrm{Q_{ff}}\|_{\mathrm{op}}\mathrm{I} + \sigma^2\mathrm{I})^{-1}y - \frac{N}{2}\log 2\pi,$$

and

$$\mathcal{U}_2 \coloneqq -\frac{1}{2}\log\det(\mathrm{Q_{ff}} + \sigma^2\mathrm{I}) - \frac{1}{2}y^T(\mathrm{Q_{ff}} + t\mathrm{I} + \sigma^2\mathrm{I})^{-1}y - \frac{N}{2}\log 2\pi. \tag{12}$$

For completeness we give a brief derivation of Lemma 2 in Appendix B, which essentially follows the derivation of Titsias (2014).

In problems where sparse GP regression is applied, computing the largest eigenvalue of $\mathrm{K_{ff}} - \mathrm{Q_{ff}}$ in order to compute $\mathcal{U}_1$ is computationally prohibitive. However, $\mathrm{tr}(\mathrm{K_{ff}} - \mathrm{Q_{ff}})$ can be computed in $\mathcal{O}(NM^2)$, so that $\mathcal{U}_2$ can be computed efficiently.

As $\mathcal{L} = \log p(y) - \mathrm{KL}[Q||P]$, we have

$$\mathrm{KL}[Q||P] = \log p(y) - \mathcal{L} \leq \mathcal{U}_1 - \mathcal{L} \leq \mathcal{U}_2 - \mathcal{L}. \tag{13}$$

If the difference between the upper and lower bounds is small, we can therefore be sure that sufficiently many inducing points are being used for the KL-divergence to be small. This suggests a refinement of the method for selecting the number of inducing points discussed earlier: continue to place more inducing points until the difference between the upper and lower bounds is small.

This raises the question: how many inducing variables do we need for the KL-divergence to be small in a typical problem? The upper bounds discussed above assess the approximation *a posteriori*, i.e. for a given data set and a given approximation. We would like to characterize the required number of inducing variables for a whole class of problems, *before* observing any data. This allows us to understand *a priori* how much computation is needed to solve a particular problem. For example, if the number of inducing variables M needs to grow linearly with the number of observations N, then the $\mathcal{O}(NM^2)$ cost of the approximation effectively scales cubically in N, i.e. in the same way as the exact implementation. In Section 4, we show that under intuitive assumptions, the number of inducing points can be taken to be much smaller than the size of the data set, while still giving approximations with small KL-divergences.

4. Convergence of Sparse Variational Inference in Gaussian Processes

In this section, we prove upper bounds on the KL-divergence between the approximate posterior and the exact posterior that depend on the number of inducing points used in inference, properties of the prior and distributional assumptions on the training covariates. The proof proceeds in three parts:

1. Derive an upper bound on the KL-divergence for a fixed data set and fixed set of inducing points that only depends on the quality of the approximation of $\mathrm{K_{ff}}$ by $\mathrm{Q_{ff}}$. In order to do this we make assumptions about the data generating process for y.

2. Suggest a method for selecting inducing inputs that obtains a high quality low-rank approximation to $\mathrm{K_{ff}}$. This yields an upper bound on the KL-divergence depending only on the eigenvalues of $\mathrm{K_{ff}}$. We consider using a k-determinantal point process or ridge leverage scores as the initialization method.

3. Relate eigenvalues of the kernel matrix back to those of the corresponding kernel operator, Eq. (10), through assumptions on the distribution of the covariates.

The second step has precedent in the literature on sparse kernel ridge regression. For example, Li et al. (2016) consider using a k-DPP to select the sparse regressors. Meanwhile ridge leverage scores have been studied in the setting of sparse kernel ridge regression and Gaussian process regression (Alaoui and Mahoney, 2015; Rudi et al., 2015; Calandriello et al., 2017, 2019), and have been shown to lead to strong statistical guarantees.

The third step in our analysis is similar to the analysis carried out when studying generalization and approximation bounds for Gaussian processes and other kernel methods. We use a generalization of a lemma proven in Shawe-Taylor et al. (2005) for this step.

In order to carry out our analysis, especially steps 2 and 3, we will treat X, y and Z as realizations of random variables $\mathbf{X}$, $\mathbf{y}$ and $\mathbf{Z}$ and make distributional assumptions about these random variables. This will allow us to make statements about bounds that hold in expectation or with fixed probability.

4.1. A-Posteriori Upper Bounds on the KL-divergence Revisited

In Section 3, we considered bounds on the KL-divergence that can be computed for a specific data set. In this section, we first derive an upper bound on the KL-divergence that only depends on the squared norm of $\mathbf{y}$, with no additional assumptions on the distribution of the $\mathbf{y}$ (Lemma 3). We then derive a second bound, given in Lemma 4, that improves on Lemma 3 in expectation, under the stronger assumption that $\mathbf{y}|\mathbf{Z},\mathbf{X} \sim \mathcal{N}(0, \mathrm{K_{ff}}+\sigma^2\mathrm{I})$. This assumption is satisfied if $\mathbf{y}$ is distributed according to the prior model and the distributions of $\mathbf{Z}$ and $\mathbf{y}$ are independent, i.e. the inducing inputs are chosen without reference to y. While our results are stated in terms of inducing points, the proofs generalize without modification to other inducing variables of the form discussed in Section 2.3.2.

4.1.1. Upper bounds on the KL-divergence

We first consider the case where we make few assumptions on the distribution of $\mathbf{y}$.

Lemma 3 *For any $y \in \mathbb{R}^N, X \in \mathcal{X}^N$, and any $Z \in \mathcal{X}^M$*

$$\mathrm{KL}[Q||P] \leq \mathcal{U}_1 - \mathcal{L} \leq \frac{1}{2\sigma^2}\left(t + \frac{\zeta\|y\|_2^2}{\zeta+\sigma^2}\right) \leq \frac{1}{2\sigma^2}\left(t + \frac{t\|y\|_2^2}{t+\sigma^2}\right),$$

with $t = \mathrm{tr}(\mathrm{K_{ff}} - \mathrm{Q_{ff}})$ and $\zeta = \|\mathrm{K_{ff}} - \mathrm{Q_{ff}}\|_{\mathrm{op}}$.

The first inequality has already been established (Eq. 13). The remainder of the proof, given in Appendix C relies on properties of symmetric positive semi-definite (SPSD) matrices.

In most applications, it is reasonable to assume that the data generating process for $\mathbf{y}$ is such that $\|\mathbf{y}\|_2^2 \leq RN$ almost surely, or at least $\mathbb{E}\left[\|\mathbf{y}\|_2^2 \,\middle|\, \mathbf{X}, \mathbf{Z}\right] \leq RN$, for some constant $R > 0$. For example, if $\mathbf{y}$ is formed by evaluating a bounded function corrupted by Gaussian or bounded noise and the location of the inducing inputs is independent of $\mathbf{y}$ then Lemma 3 allows us to bound the conditional expectation $\mathbb{E}[\mathrm{KL}[Q||P] \,|\, \mathbf{X}, \mathbf{Z}]$.

4.1.2. Average Case Analysis for the Prior Model

In Lemma 3, we did not make any assumption on the distribution of $\mathbf{y}$. From the Bayesian perspective, it is natural to make stronger distributional assumptions on $\mathbf{y}|\mathbf{X}$. We will see that in some instances stronger assumptions can lead to a much tighter upper bound than Lemma 3 that holds in expectation.

The natural candidate distribution for $\mathbf{y}$ is the prior distribution, that is $\mathbf{y}|\mathbf{X} \sim \mathcal{N}(0, \mathrm{K_{ff}} + \sigma^2\mathrm{I})$; if we additionally assume that the distributions of $\mathbf{y}|\mathbf{X}$ and $\mathbf{Z}|\mathbf{X}$ are independent, then this implies $\mathbf{y}|\mathbf{X}, \mathbf{Z} \sim \mathcal{N}(0, \mathrm{K_{ff}} + \sigma^2\mathrm{I})$. In this case we can derive upper and lower bounds on the conditional expectation of the KL-divergence conditioned on $\mathbf{X}$ and $\mathbf{Z}$.

Lemma 4 *Suppose* $\mathbf{y}|\mathbf{X}, \mathbf{Z} \sim \mathcal{N}(0, \mathrm{K_{ff}} + \sigma^2\mathrm{I})$. *For any* $X \in \mathcal{X}^N$ *and* $Z \in \mathcal{X}^M$,

$$t/(2\sigma^2) \leq \mathbb{E}[\mathrm{KL}[Q||P] \,|\, \mathbf{Z} = Z, \mathbf{X} = X] \leq t/\sigma^2$$

where $t = \mathrm{tr}(\mathrm{K_{ff}} - \mathrm{Q_{ff}})$ *and* $\mathrm{K_{ff}}$ *and* $\mathrm{Q_{ff}}$ *are defined with respect to this* X, Z *as in Section 2.*

Remark 5 *Note that if* $\mathbf{y} \sim \mathcal{N}(0, \mathrm{K_{ff}} + \sigma^2\mathrm{I})$,

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{y}\|^2 \,\middle|\, \mathbf{X} = X, \mathbf{Z} = Z\right] &= \mathbb{E}[\mathrm{tr}(\mathbf{y}^T\mathbf{y}) \,|\, \mathbf{X} = X, \mathbf{Z} = Z] \\ &= \mathrm{tr}(\mathbb{E}[\mathbf{y}\mathbf{y}^T \,|\, \mathbf{X} = X, \mathbf{Z} = Z]) = \mathrm{tr}(\mathrm{K_{ff}} + \sigma^2\mathrm{I}).\end{aligned}$$

Therefore, under the strong assumption that $\mathbf{y}$ *is sampled from the prior model, Lemma 4 gives a significantly stronger bound on the expected KL-divergence as compared to Lemma 3.*

Proof Sketch of Lemma 4 Recall that $\mathrm{KL}[Q||P] = \log p(\mathbf{y}) - \mathcal{L}$. Taking conditional expectations on both sides,

$$\mathbb{E}[\mathrm{KL}[Q||P] \,|\, \mathbf{X} = X, \mathbf{Z} = Z] = \mathbb{E}[\log p(\mathbf{y}) - \mathcal{L} \,|\, \mathbf{X} = X, \mathbf{Z} = Z]. \tag{14}$$

Let $n(y; m, S)$ denote the density of a (multivariate) Gaussian random variable with mean m and covariance matrix S evaluated at y. Then,

$$\log p(\mathbf{y}) = \log n(\mathbf{y}; 0, \mathrm{K_{ff}} + \sigma^2\mathrm{I})$$

and

$$\mathcal{L} = \log n(\mathbf{y}; 0, \mathrm{Q_{ff}} + \sigma^2\mathrm{I}) - \frac{1}{2\sigma^2}\mathrm{tr}(\mathrm{K_{ff}} - \mathrm{Q_{ff}}).$$

Using this in Eq. (14),

$$\begin{aligned}\mathbb{E}[\mathrm{KL}[Q||P] \,|\, \mathbf{X} = X, \mathbf{Z} = Z] &= \mathbb{E}\left[\log \frac{n(\mathbf{y}; 0, \mathrm{K_{ff}} + \sigma^2\mathrm{I})}{n(\mathbf{y}; 0, \mathrm{Q_{ff}} + \sigma^2\mathrm{I})} \,\middle|\, \mathbf{X} = X, \mathbf{Z} = Z\right] + \frac{t}{2\sigma^2} \\ &= \mathrm{KL}\left[\mathcal{N}(0, \mathrm{K_{ff}} + \sigma^2\mathrm{I})||\mathcal{N}(0, \mathrm{Q_{ff}} + \sigma^2\mathrm{I})\right] + \frac{t}{2\sigma^2}. \end{aligned} \tag{15}$$

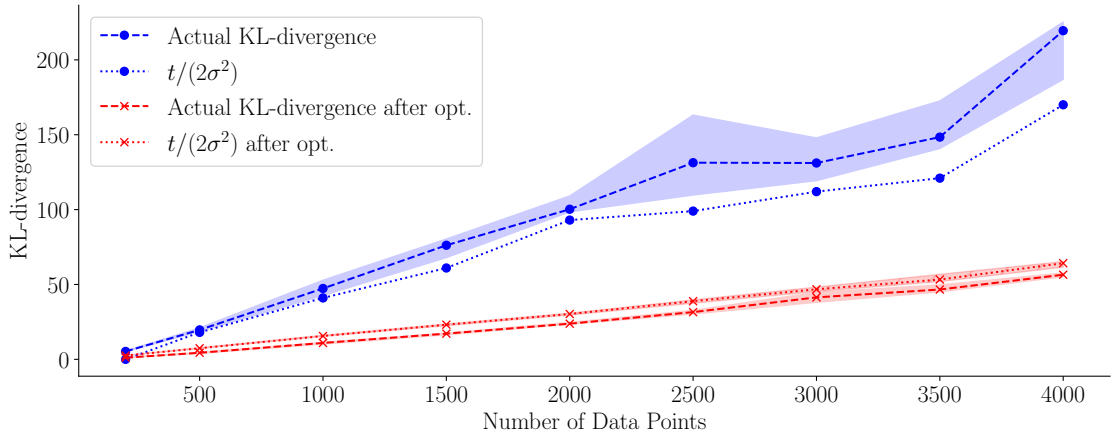

Figure 1: A comparison of the KL-divergence and the lower bound in Lemma 4. For each value of N we first fix a set of covariates and $M = 30$ inducing points and then compute both the trace and KL-divergence (shown in blue), with $y \sim \mathcal{N}(0, \mathrm{K_{ff}} + \sigma^2\mathrm{I})$. The dashed line shows the median of 20 random y's, while the shaded region represents 20-80 percentile regions. We then optimize the locations of the inducing points via gradient descent on the ELBO. As N increases, both the trace and KL-divergence increase. When the inducing points are selected via optimizing the ELBO, the KL-divergence is typically somewhat lower than the lower bound in Lemma 4, while if Z is chosen without reference to y the lower bound on the expected value of the KL-divergence holds.

The lower bound follows from the non-negativity of KL-divergence. The proof of the upper bound (Appendix C) relies on the formula for the KL-divergence between multivariate Gaussian distributions as well as the identity $|\mathrm{tr}(AB)| \leq \mathrm{tr}(A)\|B\|_{\mathrm{op}}$ for SPSD matrices A and B (Tao, 2012, Exercise 1.3.26). ■

Remark 6 *The lower bound in Lemma 4 holds in expectation* conditioned *on* $\mathbf{X} = X$ *and* $\mathbf{Z} = Z$*, with* $\mathbf{y}$ *distributed according to our prior. Common practice is to optimize the inducing inputs with respect to the ELBO, which depends on* y*. We may therefore expect that the KL-divergence will be somewhat smaller than predicted by the average case lower bound in Lemma 4 after this optimization. This is shown in Fig. 1, where we generate a data set satisfying the conditions of the lemma, and look at the trace and KL-divergence before and after gradient based optimization of the ELBO with respect to the inducing inputs. In Section 6, we establish lower bounds that hold for any* $y \in \mathbb{R}^N$ *conditioned on* $\mathbf{X} = X$ *and* $\mathbf{Z} = Z$ *and are therefore applicable to the case when inducing points are selected via gradient-based methods.*

4.2. Initialization of Inducing Points

In the previous section, we derived upper bounds on $\mathbb{E}[\mathrm{KL}[Q||P]|\mathbf{X},\mathbf{Z}]$ that depend on assumptions about the distribution of $\mathbf{y}$. These bounds depend on either the trace or the largest eigenvalue of $\mathrm{K_{ff}}-\mathrm{Q_{ff}}$, and will therefore be small if $\mathrm{K_{ff}}\approx\mathrm{Q_{ff}}$. We begin this section with a brief overview of inducing variable selection methods, of which we will analyze two in the context of sparse variational inference. Using known results on the quality of the resulting matrix approximations, we can then obtain bounds on the KL-divergence for a fixed set of training inputs $\mathbf{X}=X$.

4.2.1. Minimizing the upper bounds

We take a brief detour from discussing initializations of inducing inputs to discuss the set of inducing variables that minimize the upper bounds in Lemmas 3 and 4.

Let $\mathrm{K_{ff}}=\mathrm{V}\Lambda\mathrm{V}^\intercal$, where $\mathrm{V}=[v_1,v_2,\cdots v_N]$ is an $N\times N$ orthogonal matrix and $\Lambda=\mathrm{diag}(\tilde{\lambda}_1,\ldots,\tilde{\lambda}_N)$ is a diagonal matrix of eigenvalues of $\mathrm{K_{ff}}$ ordered such that $\tilde{\lambda}_1\geq\tilde{\lambda}_2\geq\cdots\geq\tilde{\lambda}_N\geq 0$. As $\mathrm{K_{ff}}$ is SPSD, such a decomposition exists. Define $\mathrm{K}_M=\mathrm{V}_M\Lambda_M\mathrm{V}_M^\intercal$, where V_M is an $N\times M$ matrix containing the first M columns of V and Λ_M is an $M\times M$ diagonal matrix with entries, $\tilde{\lambda}_1,\ldots,\tilde{\lambda}_M$, in other words K_M is the rank-M truncated singular value decomposition of $\mathrm{K_{ff}}$.

Both the trace and the operator norm are unitarily invariant, so K_M is the optimal rank-M approximation to $\mathrm{K_{ff}}$ according to either of these norms.[2] In particular, for any rank M $N\times N$ SPSD matrix A satisfying $\mathrm{A}\prec\mathrm{K_{ff}}$ (i.e. $\mathrm{K_{ff}}-A$ is SPSD), $\mathrm{tr}(\mathrm{K_{ff}}-\mathrm{K}_M)\leq\mathrm{tr}(\mathrm{K_{ff}}-\mathrm{A})$ and $\|\mathrm{K_{ff}}-\mathrm{K}_M\|_{\mathrm{op}}\leq\|\mathrm{K_{ff}}-\mathrm{A}\|_{\mathrm{op}}$ (see Horn and Johnson, 1990, Theorem 7.4.9.1).

As any subset of M inducing variables will lead to a rank-M matrix $\mathrm{Q_{ff}}\prec\mathrm{K_{ff}}$ this implies

$$\|\mathrm{K_{ff}}-\mathrm{Q_{ff}}\|_{\mathrm{op}}\geq\|\mathrm{K_{ff}}-\mathrm{K}_M\|_{\mathrm{op}}=\tilde{\lambda}_{M+1}\text{ and }\mathrm{tr}(\mathrm{K_{ff}}-\mathrm{Q_{ff}})\geq\mathrm{tr}(\mathrm{K_{ff}}-\mathrm{K}_M)=\sum_{m=M+1}^{N}\tilde{\lambda}_m.$$

Consider the inducing features defined as linear combinations of the random variables associated to evaluating the latent function at each observed input location, with weights coming from the eigenvectors of $\mathrm{K_{ff}}$, i.e.

$$\mathbf{u}_m=\frac{1}{\tilde{\lambda}_m}\sum_{i=1}^{N}v_{i,m}\mathbf{f}(x_i).$$

Then,

$$\mathrm{cov}(\mathbf{u}_m,\mathbf{u}_{m'})=\frac{1}{\tilde{\lambda}_m\tilde{\lambda}_{m'}}\mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}v_{m,j}v_{m',i}\mathbf{f}(x_i)\mathbf{f}(x_j)\right]=\frac{1}{\tilde{\lambda}_m\tilde{\lambda}_{m'}}\sum_{i=1}^{N}\sum_{j=1}^{N}v_{m,i}v_{m',j}k(x_i,x_j).$$

The final two sums are the quadratic form, $v_m^\intercal\mathrm{K_{ff}}v_{m'}$. As v_m is an eigenvector of $\mathrm{K_{ff}}$ and v_m is orthogonal to v'_m unless $m=m'$, this simplifies to $\mathrm{cov}(u_m,u_{m'})=\frac{\delta_{m,m'}}{\tilde{\lambda}_m}$. Similarly,

$$\mathrm{cov}(\mathbf{u}_m,\mathbf{f}(x_n))=\frac{1}{\tilde{\lambda}_m}\mathbb{E}\left[\sum_{i=1}^{N}v_{m,n}\mathbf{f}(x_i)\mathbf{f}(x_n)\right]=\frac{1}{\tilde{\lambda}_m}\sum_{i=1}^{N}v_{m,n}k(x_i,x_n)=\frac{1}{\tilde{\lambda}_m}[\mathrm{K_{ff}}v_m]_n=v_{m,n}.$$

2. While the trace is not generally a matrix norm, it agrees with the norm $\|\cdot\|_1$ as $\mathrm{K_{ff}}-\mathrm{Q_{ff}}$ is SPSD.

From these expressions, it follows that for these features $\mathrm{K}_{\mathrm{uf}} = \mathrm{V}_M$ and $\mathrm{K}_{\mathrm{uu}}^{-1} = \Lambda_M$. Therefore, $\mathrm{Q}_{\mathrm{ff}} = \mathrm{K}_M$, and these features minimize our upper bounds among any set of M inducing variables. Unfortunately, computing the matrices K_{uf} and K_{uu} in this case involves computing the first M eigenvalues and eigenvectors of K_{ff}, which lies outside of our desired computational budget of $\mathcal{O}(N\mathrm{poly}(M)\mathrm{polylog}(N))$.

4.2.2. M-Determinantal point processes

We now return to the more practical case of using inducing points for sparse variational inference. In order to derive non-trivial upper bounds on $\mathrm{tr}(\mathrm{K}_{\mathrm{ff}} - \mathrm{Q}_{\mathrm{ff}})$ and $\|\mathrm{K}_{\mathrm{ff}} - \mathrm{Q}_{\mathrm{ff}}\|_{\mathrm{op}}$, we need a sufficiently good method for placing inducing points. When using differentiable kernel functions, many practitioners select the locations of the inducing points with gradient-based methods by maximizing the ELBO. As this is a high-dimensional, non-convex optimization algorithm, directly analyzing the result of this procedure is beyond our analysis.

In this section, we assume M inducing points are subsampled from data according to an approximate *M-determinantal point process* (M-DPP) (Kulesza and Taskar, 2011) and use known bounds on the expected value of $\mathrm{tr}(\mathrm{K}_{\mathrm{ff}} - \mathrm{Q}_{\mathrm{ff}})$.[3] We note that if this scheme is used as an initialization prior to a gradient-based optimization of the evidence lower bound with respect to the inducing inputs, the resulting KL-divergence will be at least as small, so our bounds still apply after optimization of variational parameters.

Given an SPSD matrix L, an M-determinantal point process (Kulesza and Taskar, 2011) with kernel matrix L defines a discrete probability distribution over subsets of the N columns of L, with positive probability only assigned to subsets of cardinality M. The probability of any subset of cardinality M is proportional to the determinant of the principal submatrix formed by selecting those columns and the corresponding rows, that is for any set Z of M columns of L

$$\Pr(\mathbf{Z} = Z) = \frac{\det(\mathrm{L}_{Z,Z})}{\sum_{|Z'|=M} \det(\mathrm{L}_{Z',Z'})}.$$

where $\mathrm{L}_{Z,Z}$ is the principal submatrix of L with columns in Z. For a thorough introduction to determinantal point processes, as well as an implementation of many sampling methods, see Gautier et al. (2019).

As the determinant of $\mathrm{L}_{Z,Z}$ corresponds to the volume of the parallelepiped in $\mathbb{R}^M$ formed by the columns in Z, M-determinantal point processes introduce strong negative correlations between points sampled. This leads to samples that are more dispersed than subsets selected uniformly (Fig. 2). This intuition, as well as the following result due to Belabbas and Wolfe (2009) serves as motivation for using an M-DPP in order to select the location of inducing points.

Lemma 7 (Belabbas and Wolfe, 2009, Theorem 1) *Let* L *be a SPSD* $N \times N$ *matrix with eigenvalues* $\eta_1 \geq \cdots \geq \eta_N \geq 0$. *Suppose a set of points are sampled according to an* M*-determinantal point process with kernel matrix* L. *Define the (random) matrix* $\mathrm{L}_{\mathbf{Z}} = \mathrm{L}_{\mathbf{Z},N}^{T}\mathrm{L}_{\mathbf{Z},\mathbf{Z}}^{-1}\mathrm{L}_{\mathbf{Z},N}$ *where* $\mathrm{L}_{\mathbf{Z},\mathbf{Z}}$ *is the* $M \times M$ *principal submatrix of* L *with columns in* $\mathbf{Z}$ *and*

3. The standard terminology is k-DPP. We use M as this determines the number of inducing points and to avoid confusion with the kernel function.

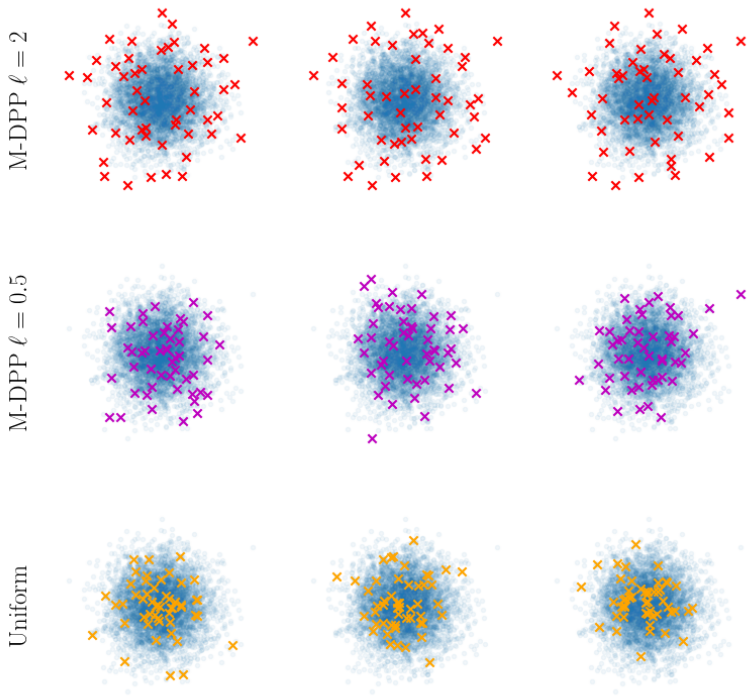

Figure 2: M-Determinantal point processes introduce a strong form of negative correlation between points, leading to samples that appear over-dispersed compared to uniform sampling. In the top two rows, we show (approximate) samples from a M-DPP with kernel matrix determined by a SE-kernel with two different length scales and Gaussian distributed covariates, with 50 points drawn for each sample. Sampling is performed via MCMC after a greedy initialization. In the bottom row, we show subsets of size 50 selected uniformly from the covariates.

$\mathrm{L}_{N,\mathbf{Z}}$ *is the* $N \times M$ *matrix with columns* $\mathbf{Z}$*. Then,*

$$\mathbb{E}[\mathrm{tr}(\mathrm{L} - \mathrm{L}_{\mathbf{Z}})] \leq (M+1) \sum_{m=M+1}^{N} \eta_m. \tag{16}$$

If $\mathrm{L} = \mathrm{K}_{\mathrm{ff}}$ and the inducing points are selected as a subset of data points corresponding to the columns selected by the M-DPP, then $\mathrm{L}_{\mathbf{Z}} = \mathbf{Q}_{\mathbf{ff}}$. This tells us that using an M-DPP to choose inducing inputs will make $\mathbb{E}[\mathrm{tr}(\mathrm{K}_{\mathrm{ff}} - \mathbf{Q}_{\mathbf{ff}}) \,|\, \mathbf{X} = X]$ relatively close to its optimal value of $\sum_{m=M+1}^{N} \tilde{\lambda}_m$.

The next important question to address is whether a M-DPP can be sampled with sufficiently low computational complexity for this to be a practical method for selecting inducing inputs. Naively computing the probability distribution over all $\binom{N}{M}$ subsets of size M is prohibitively expensive. Kulesza and Taskar (2011) gave an algorithm that runs in

Algorithm 1 MCMC algorithm for approximately sampling from an M-DPP (Anari et al., 2016)

Input: Training inputs $X = \{x_i\}_{i=1}^N$, number of points to choose, M, kernel k, T number of steps of MCMC to run.
Returns: An (approximate) sample from a M-DPP with kernel matrix $\mathrm{K_{ff}}$ formed by evaluating k at X.
Initialize M columns by greedily selecting columns to maximize the determinant of the resulting submatrix. Call this set of indices of these columns Z_0.
for $\tau \leq T$ **do**
  Sample i uniformly from Z_τ and j uniformly from $X \setminus Z_\tau$. Define $Z' = Z_\tau \setminus \{i\} \cup \{j\}$,
  Compute $p_{i \to j} := \frac{1}{2} \min\{1, \det(\mathrm{K}_{Z'})/\det(\mathrm{K}_{Z_\tau})\}$
  With probability $p_{i\to j}$, $Z_{\tau+1} = Z'$ otherwise, $Z_{\tau+1} = Z_\tau$
end for
Return: Z_T

polynomial time and yields exact samples from an M-DPP. Unfortunately, this algorithm involves computing an eigendecomposition of the $N \times N$ kernel matrix ($\mathrm{K_{ff}}$ in our case), which is computationally prohibitive.

Recently, Dereziński et al. (2019) gave an algorithm for obtaining an exact sample from an M-DPP in time that is polynomial in M and nearly-linear in N. However, the polynomial in M is high. We instead consider an approximate algorithm and therefore derive the following simple corollary of Lemma 7.

Corollary 8 *Let ρ denote an M-DPP with kernel matrix L, satisfying $L_{i,i} \leq v$. Let ρ' denote a measure on subsets of columns of L with cardinality M such that $\mathrm{TV}(\rho, \rho') \leq \epsilon$ for some $\epsilon > 0$, where $\mathrm{TV}(\rho, \rho') := \frac{1}{2} \sum_{|Z|=M} \rho(Z) - \rho'(Z)$. Then*

$$\mathbb{E}_{\rho'}[\mathrm{tr}(\mathrm{L} - \mathrm{L}_{\mathbf{Z}})] \leq 2Nv\epsilon + (M+1) \sum_{m=M+1}^{N} \eta_m,$$

where η_m is the m^{th} largest eigenvalue of L.

Proof

$$\begin{aligned}
\mathbb{E}_{\rho'}[\mathrm{tr}(\mathrm{L} - \mathrm{L}_{\mathbf{Z}})] &= \sum_{|Z|=M} \mathrm{tr}(\mathrm{L} - \mathrm{L}_Z)(\rho'(Z) + \rho(Z) - \rho(Z)) \\
&= \mathbb{E}_{\rho}[\mathrm{tr}(\mathrm{L} - \mathrm{L}_{\mathbf{Z}})] + \sum_{|Z|=M} \mathrm{tr}(\mathrm{L} - \mathrm{L}_Z)(\rho'(Z) - \rho(Z)) \\
&\leq \mathbb{E}_{\rho}[\mathrm{tr}(\mathrm{L} - \mathrm{L}_{\mathbf{Z}})] + 2\mathrm{TV}(\rho, \rho') \max_{Z:|Z|=M} (\mathrm{tr}(\mathrm{L} - \mathrm{L}_Z)).
\end{aligned}$$

The corollary is completed by noting that for all Z, $\mathrm{tr}(\mathrm{L} - \mathrm{L}_Z) \leq \mathrm{tr}(\mathrm{L}) \leq Nv$. ■

Corollary 8 shows that sufficiently accurate approximate sampling from an M-DPP only has a small effect on the quality of the resulting $\mathrm{Q_{ff}}$. High quality approximate samples can

be drawn using a simple Markov Chain algorithm described in Anari et al. (2016), given as Algorithm 1. This MCMC algorithm is well-studied in the context of M-DPPs and their generalizations, and is known to be rapidly mixing (Anari et al., 2016; Hermon and Salez, 2019).

Lemma 9 (Hermon and Salez, 2019, Corollary 1) *Let ρ be an M-DPP with $N \times N$ kernel matrix L. Fix $\epsilon \in (0,1)$. Then Algorithm 1 produces a sample from a distribution ρ' satisfying*

$$\mathrm{TV}(\rho, \rho') \leq \epsilon$$

in not more than $T(\epsilon) = 2MN\Big(\log\log\Big(\frac{1}{\rho(Z_0)}\Big) + \log\frac{2}{\epsilon^2}\Big)$ iterations, where Z_0 is the subset of columns at which the Markov chain is initialized.

Since the determinant of a matrix is equal to the product of the determinant of a principal submatrix times the determinant of the Schur complement of this submatrix, the greedy initialization used in Algorithm 1 is equivalent to starting with $U = \emptyset$ and iteratively adding $\operatorname{argmax}_{x \in X} k(x,x) - \mathrm{k}_{f(x)u}\mathrm{K}_{\mathrm{uu}}^{-1}\mathrm{k}_{uf(x)}$ to U. This can be performed in time $\mathcal{O}(NM^2)$, for example by computing the pivot rules of a rank-M incomplete Cholesky decomposition of K_{ff} (Chen et al., 2018, Algorithm 1).

The per iteration cost of Algorithm 1 is dominated by computing the acceptance ratio, which can be performed in $\mathcal{O}(M^2)$, by iteratively updating a Cholesky or QR factorization of the matrix associated to the current set of columns. This makes the total cost of obtaining an ϵ-approximate sample $\mathcal{O}\big(NM^3 \log\log(1/\rho(Z_{\mathrm{greedy}})) + NM^3 \log 2/\epsilon^2\big)$, where Z_{greedy} denotes the set of columns selected by greedily maximizing the determinant of the submatrix. Moreover, the subset selected by the algorithm is known to have a probability at least $1/(M!)^2$ of the maximum probability subset (Çivril and Magdon-Ismail, 2009; Anari et al., 2016). By using the fact that the the maximum probability subset is more probable than the uniformly distributed probability, we obtain

$$\rho(Z_{\mathrm{greedy}}) \geq \left(M!^2 \binom{N}{M}\right)^{-1} \geq (MN)^{-M},$$

giving an overall complexity of not more than $\mathcal{O}\big(NM^3(\log\log N + \log M + \log 1/\epsilon^2)\big)$. This gives us a method for initializing inducing points conditioned on input locations, such that we can relate $\mathbb{E}[\mathrm{tr}(\mathbf{K_{ff}} - \mathbf{Q_{ff}}) \,|\, \mathbf{X} = X]$ to $\mathrm{tr}(\mathrm{K}_{\mathrm{ff}} - \mathrm{K}_M)$.

We now take a brief detour to consider a different approach to initializing inducing inputs before completing the proof of a priori bounds on the KL-divergence.

4.2.3. Ridge Leverage Scores

While using an M-DPP to select inducing inputs allows us to bound $\mathbb{E}[\mathrm{tr}(\mathrm{K}_{\mathrm{ff}} - \mathbf{Q_{ff}}) \,|\, \mathbf{X} = X]$, this method has a significant drawback as opposed to other methods of initialization: the computational cost of running the MCMC algorithm to obtain approximate samples dominates the cost of sparse inference. Ridge leverage score (RLS) sampling offers an alternative that runs in $\mathcal{O}(NM^2)$, while retaining strong theoretical guarantees on the quality of the resulting approximation. In this section, we give a brief discussion of ridge leverage scores as well an algorithm of Musco and Musco (2017) that allows for efficient approximations to ridge leverage scores.

The ω-ridge leverage score of a point $x_n \in X$ of a Gaussian process regressor, which we denote by $\ell^\omega(x_n)$ is defined as $1/\omega$ times the posterior variance at x_n of the process with noise variance ω, i.e.

$$\ell^\omega(x_n) = \frac{1}{\omega}\big(k(x_n, x_n) - \mathrm{k}_{\mathrm{nf}}^\intercal(\mathrm{K}_{\mathrm{ff}} + \omega\mathrm{I})^{-1}\mathrm{k}_{\mathrm{nf}}\big),$$

where $\mathrm{k}_{\mathrm{nf}}^\intercal = [k(x_1, x_n), k(x_2, x_n), \ldots, k(x_n, x_n)]$. RLS sampling uses these values as an importance distribution for selecting which points to include in sparse kernel methods. Intuitively, points at which there is high posterior uncertainty must be 'far' from other points, and therefore informative.

Computing the ridge leverage scores exactly is too computationally expensive, as it involves inverting the kernel matrix. However, practical approximate versions of leverage sampling algorithms that retain strong theoretical guarantees have been developed.

Ridge leverage based sampling algorithms select a subset of training data to use as inducing points. Each point is sampled independently into the subset with probability proportional to its leverage score. Approximate versions of this algorithm generally rely on overestimating the ridge leverage scores, which lead to equally strong accuracy guarantees compared to using the exact ridge leverage scores, at the cost of sampling more points in the approximation.

We consider the application of Algorithm 3 in Musco and Musco (2017) to the problem of selecting inducing inputs for sparse variational inference in GP models. This algorithm comes with the following bounds on the quality of the resulting Nyström approximation.

Lemma 10 (Musco and Musco, 2017, Theorem 14, Appendix D) *Given $X \in \mathcal{X}^N$ and a kernel k, let K_{ff} denote the $N \times N$ covariance matrix associated to X and k. Fix $\delta \in (0, \frac{1}{32})$ and $S \in \mathbb{N}$. There exists a universal constant c and algorithm with run time $\mathcal{O}(N\mathbf{M}^2)$ and memory complexity $\mathcal{O}(N\mathbf{M})$ that with probability $1 - 3\delta$ returns $\mathbf{M} \leq cS\log(S/\delta)$ columns of K_{ff} such that the resulting Nyström approximation, $\mathbf{Q_{ff}}$, satisfies*

$$\|\mathrm{K}_{\mathrm{ff}} - \mathbf{Q_{ff}}\|_{\mathrm{op}} \leq \frac{1}{S}\sum_{m=S+1}^{N} \tilde{\lambda}_m.$$

While in Section 4.2.2 M was fixed and the quality of the resulting approximation was random, in the algorithm discussed above $\mathbf{M}$ is additionally random.

An alternative approach to sampling using ridge leverage scores specifies a desired level of accuracy of the resulting approximation, and the number of points selected is chosen to obtain this approximation quality with fixed probability. This has the advantage of not requiring the user to manually select the number of inducing points, but may lead to a number of inducing points being used that exceeds a practical computational budget. We discuss the application of this approach to variational Gaussian process regression in Appendix G.

4.3. A-Priori Bounds on the KL-divergence

In the previous sections, the results on the quality of approximation depended on the eigenvalues of K_{ff}. As these eigenvalues depend on the covariates X, and we would like to

make statements that apply to a wide-range of data sets, we assume X is a realization of a random variable $\mathbf{X}$, and make assumptions about the distribution of $\mathbf{X}$.

If each $\mathbf{x} \in \mathbf{X}$ is i.i.d. distributed, according to some measure with continuous density $p(x)$, in the limit as the amount of data tends to infinity, the matrix $\frac{1}{N}\mathbf{K_{ff}}$ behaves like the operator $\mathcal{K}$ (Koltchinskii and Giné, 2000) defined with respect to this p. For finite sample sizes, the large eigenvalues of $\frac{1}{N}\mathbf{K_{ff}}$ tend to overestimate the corresponding eigenvalues of $\mathcal{K}$ and the small eigenvalues of $\frac{1}{N}\mathbf{K_{ff}}$ tend to underestimate the small eigenvalues of $\mathcal{K}$. We make this precise through a minor generalization of a lemma of Shawe-Taylor et al. (2005).

Lemma 11 *Suppose that N covariates are distributed in $\mathbb{R}^D$ such that the marginal distribution, μ_{x_n}, of each covariate, $\mathbf{x}_n$ has a continuous density $p_n(x)$, and there exists a distribution with continuous density $q(x)$ satisfying $p_n(x) < c_n q(x)$ for some $c_n > 0$ for all n. Let $\tilde{\boldsymbol{\lambda}}_m$ denote the m^{th} largest eigenvalue of the random matrix $\mathbf{K_{ff}}$ formed by a continuous, bounded kernel and these covariates. Let λ_m denote the m^{th} largest eigenvalue of the integral operator corresponding to the distribution with density q, $\mathcal{K}_q$. Then, for any $M \geq 1$*

$$\mathbb{E}\left[\frac{1}{N}\sum_{m=M+1}^{N}\tilde{\boldsymbol{\lambda}}_m\right] \leq \bar{c}\sum_{m=M+1}^{\infty}\lambda_m,$$

where $\bar{c} = \frac{1}{N}\sum_{n=1}^{N} c_n$.

Proof For $\mathbf{X}$ taking values in $(\mathbb{R}^D)^N$, and any rank-M matrix SPSD $\mathbf{\Phi} \prec \mathbf{K_{ff}}$, we have

$$\frac{1}{N}\sum_{m=M+1}^{N}\tilde{\boldsymbol{\lambda}}_m = \frac{1}{N}\mathrm{tr}(\mathbf{K_{ff}} - \mathbf{K}_M) \leq \frac{1}{N}\mathrm{tr}(\mathbf{K_{ff}} - \mathbf{\Phi}),$$

with $\mathbf{K}_M$ defined as in Section 4.2.1. The inequality follows form the optimality of $\mathbf{K}_M$ as a rank-M approximation to $\mathbf{K_{ff}}$ in the Schatten-1 norm (sum of absolute value of singular values).

As k is a continuous bounded kernel we can apply Mercer's theorem to represent $k(x, x')$ with respect to the eigenfunctions of the operator $\mathcal{K}_q$ giving $[\mathbf{K_{ff}}]_{i,j} = \sum_{m=1}^{\infty}\lambda_m\phi_m(\mathbf{x}_i)\phi_m(\mathbf{x}_j)$, and $\mathbb{E}_q\left[\phi(\mathbf{x}_n)^2\right] = 1$.

Consider the rank-M approximation to $\mathbf{K_{ff}}$ given by truncating this Mercer expansion, $[\mathbf{\Phi}]_{i,j} = \sum_{m=1}^{M}\lambda_m\phi_m(\mathbf{x}_i)\phi_m(\mathbf{x}_j)$. Then $[\mathbf{K_{ff}} - \mathbf{\Phi}]_{i,j} = \sum_{m=M+1}^{\infty}\lambda_m\phi_m(\mathbf{x}_i)\phi_m(\mathbf{x}_j)$, so $\mathbf{K_{ff}} - \mathbf{\Phi} \succ 0$.

For any covariates $\{\mathbf{x}_n\}_{n=1}^{N}$ satisfying the conditions of the lemma,

$$\frac{1}{N}\sum_{m=M+1}^{N}\tilde{\boldsymbol{\lambda}}_m \leq \frac{1}{N}\mathrm{tr}(\mathbf{K_{ff}} - \mathbf{\Phi}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{m=M+1}^{\infty}\lambda_m\phi_m(\mathbf{x}_n)^2.$$

Taking expectations on both sides with respect to the covariate distribution,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{N}\sum_{m=M+1}^{N}\tilde{\boldsymbol{\lambda}}_m\right] &\leq \frac{1}{N}\sum_{n=1}^{N}\sum_{m=M+1}^{\infty}\lambda_m\int\phi_m(x)^2 p_n(x)\,\mathrm{d}x \\ &\leq \frac{1}{N}\sum_{n=1}^{N}c_n\sum_{m=M+1}^{\infty}\lambda_m\int\phi_m(x)^2 q(x)\,\mathrm{d}x = \bar{c}\sum_{m=M+1}^{\infty}\lambda_m\ ,\end{aligned}$$

The interchanging of integral and sum is justified by Fubini's theorem as each ϕ_m is square integrable, each eigenvalue is non-negative, and the sum converges by Mercer's theorem. We used the non-negativity of $\phi_m(x)^2$ in the second inequality to bound the expectation of $\phi_m(x)^2$ under p_n in terms of its expectation under q. ■

Corollary 12 *Suppose the covariate distribution has identically distributed marginals, each with density $p(x)$, then*

$$\mathbb{E}\left[\frac{1}{N}\sum_{m=M+1}^{N}\tilde{\boldsymbol{\lambda}}_m\right] \leq \sum_{m=M+1}^{\infty}\lambda_m,$$

where λ_m is the m^{th} largest eigenvalue of the operator associated to the kernel and the distribution with continuous density $p(x)$.

This corollary follows from Lemma 11 by taking $q = p$ and $c_n = 1$ for all n. For simplicity, we will state our main results using the assumptions of this corollary, though the generalization to cases with non-identical marginals satisfying the conditions of Lemma 11 is immediate. We have now accumulated the necessary preliminaries to prove our main theorems.

4.3.1. Bounds on the KL-divergence for M-DPP Sampling

Theorem 13 *Suppose N training inputs are drawn according to a distribution on $\mathbb{R}^D$ with identical marginal distributions, each with density $p(x)$. Let k be a continuous kernel such that $k(x,x) < v$ for all $x \in \mathbb{R}^D$. Suppose $\mathbf{y}$ is distributed such that $\mathbb{E}\left[\|\mathbf{y}\|_2^2 \,\middle|\, \mathbf{X}\right] \leq RN$ almost surely for some $R \geq 0$. Sample M inducing points from the training data according to an ϵ-approximation to a M-DPP with kernel matrix $\mathrm{K_{ff}}$. Then,*

$$\mathbb{E}[\mathrm{KL}[Q||P]] \leq \frac{1}{2}\left(1 + \frac{RN}{\sigma^2}\right)\frac{(M+1)N\sum_{m=M+1}^{\infty}\lambda_m + 2Nv\epsilon}{\sigma^2}, \tag{17}$$

where the expectation is taken over the covariates, the mechanism for initializing inducing points and the observations.

Proof of Theorem 13 We use Lemma 3, Corollary 8 and take expectations with respect to $\mathbf{Z}$, noting that $\mathbf{Z}|\mathbf{X}$ is independent of $\mathbf{y}|\mathbf{X}$ so that $\mathbb{E}[\mathrm{tr}(\mathbf{K_{ff}} - \mathbf{Q_{ff}}) \,|\, \mathbf{X}] = \mathbb{E}[\mathrm{tr}(\mathbf{K_{ff}} - \mathbf{Q_{ff}}) \,|\, \mathbf{y}, \mathbf{X}]$,

$$\mathbb{E}[\mathrm{KL}[Q||P] \,|\, \mathbf{y}, \mathbf{X}] \leq \frac{N}{2\sigma^2}\left(1 + \frac{\|\mathbf{y}\|_2^2}{\sigma^2}\right)\left(\frac{M+1}{N}\sum_{m=M+1}^{N}\tilde{\boldsymbol{\lambda}}_m + 2v\epsilon\right). \tag{18}$$

Now using the assumption that $\mathbb{E}\left[\|\mathbf{y}\|_2^2 \,\middle|\, \mathbf{X}\right] \leq RN$ almost surely, and taking expectation over $\mathbf{y}$,

$$\mathbb{E}[\mathrm{KL}[Q||P] \,|\, \mathbf{X}] \leq \frac{N}{2\sigma^2}\left(1 + \frac{RN}{\sigma^2}\right)\left(\frac{M+1}{N}\sum_{m=M+1}^{N}\tilde{\boldsymbol{\lambda}}_m + 2v\epsilon\right). \tag{19}$$

Finally, taking expectation with respect to the covariate distribution over the covariate distribution and applying Lemma 11,

$$\mathbb{E}[\mathrm{KL}[Q||P]] \leq \frac{1}{2}\left(1 + \frac{RN}{\sigma^2}\right)\left(\frac{N(M+1)\sum_{m=M+1}^{\infty}\lambda_m + 2Nv\epsilon}{\sigma^2}\right). \tag{20}$$

■

Theorem 14 *With the same assumptions on the covariates and inducing point distributions as in Theorem 13, but with the assumption that* $\mathbf{y}|\mathbf{X}$ *is conditionally Gaussian distributed with mean zero and covariance matrix* $\mathrm{K_{ff}} + \sigma^2\mathrm{I}$*,*

$$\mathbb{E}[\mathrm{KL}[Q||P]] \leq \frac{(M+1)N\sum_{m=M+1}^{\infty}\lambda_m + 2Nv\epsilon}{\sigma^2} \tag{21}$$

where the expectation is taken over the covariate distribution, the observation distribution and the initialization mechanism.

The proof of Theorem 14 is nearly identical to the proof of Theorem 13, applying Lemma 4 instead of Lemma 3 in the first line.

In certain instances, it may be desirable to have a bound that holds with fixed probability instead of in expectation. As $\mathrm{KL}[Q||P] \geq 0$, such a bound can be derived through applying Markov's inequality to Theorem 13 or Theorem 14 leading to the following corollaries:

Corollary 15 *Under the assumptions of Theorem 13, with probability at least* $1 - \delta$*,*

$$\mathrm{KL}[Q||P] \leq \frac{1}{2}\left(1 + \frac{RN}{\sigma^2}\right)\frac{(M+1)N\sum_{m=M+1}^{\infty}\lambda_m + 2Nv\epsilon}{\delta\sigma^2}.$$

Corollary 16 *Under the assumptions of Theorem 14, with probability at least* $1 - \delta$*,*

$$\mathrm{KL}[Q||P] \leq \frac{(M+1)N\sum_{m=M+1}^{\infty}\lambda_m + 2Nv\epsilon}{\delta\sigma^2}.$$

4.3.2. Bounds for Ridge Leverage Score Sampling

We now state and derive statements similar to Corollaries 15 and 16 for a ridge leverage score initialization utilizing Musco and Musco (2017, Algorithm 3). In order to this we us that for any SPSD A, $\mathrm{tr}(A) \leq N\|A\|_{\mathrm{op}}$, so that Lemma 3 implies

$$\mathrm{KL}[Q||P] \leq \frac{\|\mathbf{K_{ff}} - \mathbf{Q_{ff}}\|_{\mathrm{op}}}{2\sigma^2}\left(N + \frac{\|\mathbf{y}\|_2^2}{\sigma^2}\right), \tag{22}$$

and Lemma 4 implies,

$$\mathbb{E}[\mathrm{KL}[Q||P] \,|\, \mathbf{Z}, \mathbf{X}] \leq N\frac{\|\mathbf{K_{ff}} - \mathbf{Q_{ff}}\|_{\mathrm{op}}}{\sigma^2}. \tag{23}$$

Combining Lemma 10 and Corollary 12 and using Markov's inequality twice with Eq. (22) or Eq. (23) and a union bound respectively leads to the following bounds on the performance of sparse inference using ridge leverage scores:

Kernel	Covariate Distribution	M, Theorem 13	M, Theorem 17
SE-Kernel	Compact support	$\mathcal{O}((\log N)^D)$	$\mathcal{O}((\log N)^D \log\log N)$
SE-Kernel	Gaussian	$\mathcal{O}((\log N)^D)$	$\mathcal{O}((\log N)^D \log\log N)$
Matérn ν	Compact support	$\mathcal{O}(N^{\frac{2D}{2\nu-D}})$	$\mathcal{O}(N^{\frac{2D}{2\nu+D}} \log N)$

Table 1: The number of features needed for upper bounds to converge in $D-$dimensions. We assume the compactly supported distributions have bounded density functions.[5]

Theorem 17 *Take the same assumptions on* $\mathbf{X}$ *and* $\mathbf{y}|\mathbf{X}$ *as in Theorem 13. Fix* $\delta \in (0, 1/32)$ *and* $S \in \mathbb{N}$. *There exists a universal constant* c *such with probability* $1 - 5\delta$, *we have* $\mathbf{M} < cS\log(S/\delta)$ *and*

$$\mathrm{KL}[Q||P] \leq \frac{1}{2}\left(N + \frac{RN}{\sigma^2}\right)\frac{N\sum_{m=S+1}^{\infty}\lambda_m}{S\delta^2\sigma^2}$$

when inducing points are initialized using Musco and Musco (2017, Algorithm 3).

Theorem 18 *Take the same assumptions on* $\mathbf{X}$ *and* $\mathbf{y}|\mathbf{X}$ *as in Theorem 14. Fix* $\delta \in (0, 1/32)$ *and* $S \in \mathbb{N}$. *There exists a universal constant* c *such with probability* $1 - 5\delta$, *we have* $\mathbf{M} < cS\log(S/\delta)$ *and*

$$\mathrm{KL}[Q||P] \leq \frac{N^2\sum_{m=S+1}^{\infty}\lambda_m}{S\delta^2\sigma^2}$$

when inducing points are initialized using Musco and Musco (2017, Algorithm 3).

4.3.3. Are these bounds useful?

Having established probabilistic upper bounds on the KL-divergence resulting from sparse approximation, a simple question is whether these bounds offer any insight into the efficacy of sparse inference. If in order for the upper bounds to be small, we need to take $M = N$, then they would not be useful, as it is already known that by taking $Z = X$, exact inference is recovered. In the next section, we discuss bounds on the eigenvalues of $\mathcal{K}$ for common kernels and input distribution. These bounds show that for many inference problems, the upper bounds in Theorems 13, 14, 17 and 18 imply that the KL-divergence can be made small with $M \ll N$ inducing points.

5. Bounds for Specific Kernels and Covariate Distributions

In this section, we consider specific covariate distributions and commonly used kernels, and investigate the implications of the upper bounds derived in Section 4. These results are summarized in Table 1. We begin with the case of the popular squared exponential kernel and Gaussian covariates in one-dimension. This kernel and covariate distribution

5. Burt et al. (2019) stated bounds for a product of one-dimensional Matérn kernels, which differs from the commonly used multivariate Matérn kernel.

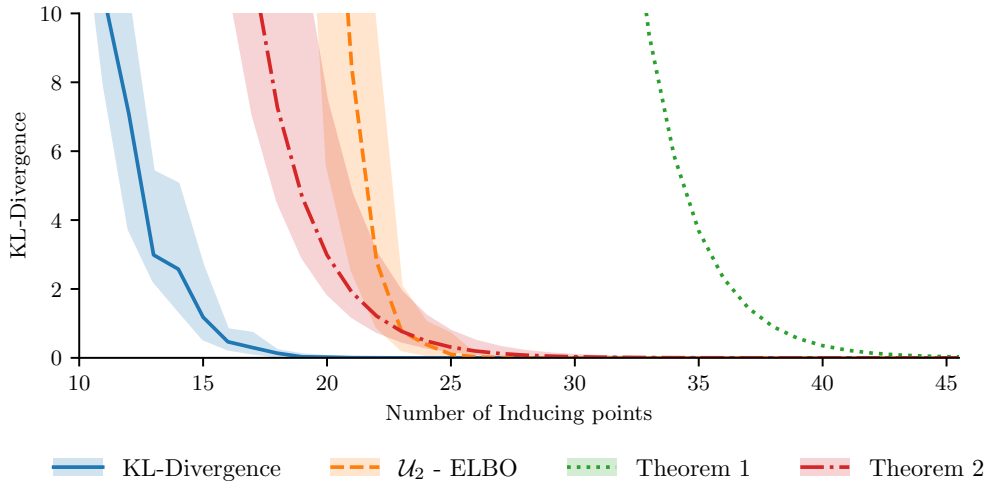

Figure 3: A comparison of the actual KL-divergence achieved, the bound given by Lemma 2 and the bounds derived in Corollaries 15 and 16 squared exponential kernel. For the actual KL-divegence and the bound given by Lemma 2, the dotted line shows the median of 20 independent trials and the shaded region shows the $20\% - 80\%$ regions, while for the bounds, the dotted line represents $\delta = .5$ and the shaded region $\delta \in [.2, .8]$.

are one of the few instances in which the eigenvalues of $\mathcal{K}$ have a simple analytic form. In Section 5.1.1, we consider the analogous multi-dimensional problem. In Section 5.2 we discuss implications for stationary kernels with compactly supported inputs, including the well-studied Matérn kernels.

5.1. Squared Exponential Kernel and Gaussian Covariate Distribution

In the case of the squared exponential kernel, with lengthscale parameter ℓ and variance v, that is

$$k_{\text{SE}}(x, x') = v \exp\left(-\frac{(x - x')^2}{2\ell^2}\right)$$

and one-dimensional covariates distributed according to $\mathcal{N}(0, \beta^2)$, the eigenvalues of $\mathcal{K}$ are (Zhu et al., 1997)

$$\lambda_m = \sqrt{\frac{2a}{A}} B^{m-1}, \tag{24}$$

where $a = (4\beta^2)^{-1}, b = (2\ell^2)^{-1}, A = a + b + \sqrt{a^2 + 4ab}$ and $B = b/A$. Note that $B < 1$ for any $\ell^2, \beta^2 > 0$, so the eigenvalues of this operator decay geometrically. However, the exact value of B depends on the lengthscale of the kernel and the variance of the covariate distribution. Short lengthscales and high standard deviations lead to values of B close to 1, which means that the eigenvalues decay more slowly. From a practical perspective, it is important to keep this in mind, as while the particular rates we obtain on how M

should grow as a function of N do not depend on the model hyperparameters, the implicit constants do.

Corollary 19 *Let k be a squared exponential kernel. Suppose that N real-valued (one-dimensional) covariates are observed, with identical Gaussian marginal distributions. Suppose the conditions of Theorem 13 are satisfied for some $R > 0$. Fix any $\gamma \in (0, 1]$. Then there exists an $M = \mathcal{O}(\log(N^3/\gamma))$ and an $\epsilon = \Theta(\gamma/N^2)$ such if inducing points are distributed according to an ϵ-approximate M-DPP with kernel matrix K_{ff},*

$$\mathbb{E}[\mathrm{KL}[Q||P]] \leq \gamma.$$

Similarly, for any $\delta \in (0, 1/32)$ using the ridge leverage algorithm of Musco and Musco (2017) and choosing S appropriately, with probability $1-5\delta$, $\mathbf{M} = \mathcal{O}\Big(\log \frac{N^2}{\delta^2\gamma} \log \frac{\log(N^2/\delta^2\gamma)}{\delta}\Big)$ and

$$\mathrm{KL}[Q||P] \leq \gamma.$$

The implicit constants depend on the kernel hyperparameters, the likelihood variance, the variance of the covariate distribution and R.

Remark 20 *If we consider γ and δ as fixed constants (independent of N), this implies that if inducing points are placed using an approximate M-DPP we can choose $M = \mathcal{O}(\log(N))$ inducing points leading to a computational cost of $\mathcal{O}(N(\log N)^4)$ while for approximate ridge leverage scores sampling $\mathcal{O}(\log N \log\log N)$ inducing points suffice leading to a cost at most $\mathcal{O}(N(\log N)^2(\log\log N)^2)$.*

The proof (Appendix D.1) consists of applying the geometric series formula to evaluate the sum of eigenvalues and choosing M, ϵ and S appropriately. All dependencies of the implicit constants on hyperparameters can be made explicit. Figure 3 illustrates the KL-divergence, the a posteriori bound given by $\mathcal{U}_2 -$ ELBO and the bounds from Theorems 13 and 14 in the case of a SE kernel and synthetic 1D distributed covariates.

Corollary 19 is illustrated in Fig. 4, in which we increase N and increase M logarithmically as a function of N in such a way that $\mathrm{KL}[Q||P]$ can be bounded above by a decreasing function in N.

5.1.1. The Multivariate Case

The generalization of Corollary 19 to the case of multi-dimensional input distributions is relatively straightforward. The multi-dimensional version of the squared exponential kernel can be written as a product of one dimensional kernels, i.e.

$$k_{\mathrm{SEARD}}(x, x') = v \exp\left(-\sum_{d=1}^{D} \frac{(x_d - x'_d)^2}{\ell_d^2}\right) = v \prod_{d=1}^{D} \exp\left(-\frac{(x_d - x'_d)^2}{\ell_d^2}\right),$$

where $\ell_d > 0$ for all d.

For any kernel that can be expressed as a product of one-dimensional kernels, and for any covariate distribution that is a product of one-dimensional covariate distributions,

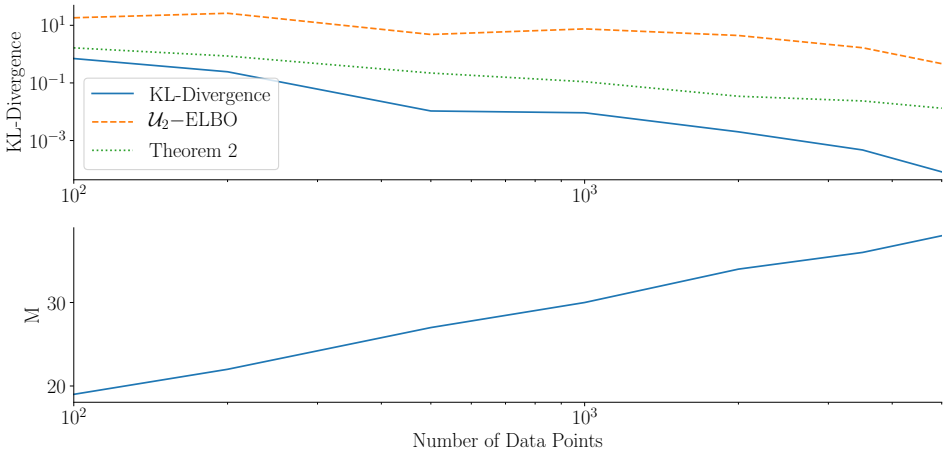

Figure 4: To illustrate Corollary 19, we incrementally increase the size of the data set, and set the number of inducing points to grow as the log of the data set size. The top plot shows the KL-divergence (blue) for each data set as N increases, plotted on a log-log scale. The bound in Theorem 14 (in green) tends to 0. We also compute the a posteriori upper bound, Lemma 2 (yellow). The bottom plot shows the number of inducing points, plotted against the number N with N on a log-scale.

the eigenvalues of the multi-dimensional covariance operator is the product of the one-dimensional analogues. When obtaining rates of convergence, we lose no generality in assuming that the kernel is isotropic as is the covariate distribution. Otherwise, consider the direction with the shortest lengthscale, and the covariate distribution with the largest standard deviation and the eigenvalues of this operator are larger than a constant multiple of the corresponding eigenvalues of the non-isotropic operator.

In the isotropic case, each eigenvalue is of the form,

$$\lambda_m = (2a/A)^{D/2} B^{m'},$$

for some integer m' with a, A and B defined as in the one-dimensional case. Note that m and m' are no longer equal. The number of times each eigenvalue with m' in the exponent is repeated is equal to the number of ways to write m' as a sum of D non-negative integers. By counting the multiplicity of each eigenvalue, Seeger et al. (2008) arrived at the bound

$$\lambda_{m+D-1} \leq (2a/A)^{D/2} B^{m^{1/D}}.$$

In order to prove a multi-dimensional analogue of Corollary 19 we need an upper bound on $\sum_{m=M+1}^{\infty} \lambda_m$. This can be derived with following an argument made by Seeger et al. (2008, Appendix II).

Proposition 21 *For a SE-kernel and Gaussian distributed covariates in $\mathbb{R}^D$, for $M \geq \frac{1}{\alpha} D^D + D - 1$, $\sum_{m=M+1}^{\infty} \lambda_m = \mathcal{O}(M \exp(-\alpha M^{1/D}))$, where $\alpha = -\log B > 0$ and the implicit constant depends on the dimension of the covariates, the kernel parameters and the covariance matrix of the covariate distribution.*

The proof of Proposition 21 is in Appendix D.1.

Corollary 22 *Let k be a SE-ARD kernel in D-dimensions. Suppose that N D-dimensional covariates are observed, so that each covariate has an identical multivariate Gaussian distribution, and that the distribution of training outputs satisfies $\mathbb{E}\left[\|\mathbf{y}\|^2 \,\middle|\, \mathbf{X}\right] \leq RN$. Fix any $\gamma \in (0, 1]$. Then there exists an $M = \mathcal{O}((\log N/\gamma)^D)$ and an $\epsilon = \mathcal{O}(N^2/\gamma)$ such if inducing inputs are distributed according to an ϵ-approximate M-DPP with kernel matrix K_{ff},*

$$\mathbb{E}[\mathrm{KL}[Q||P]] \leq \gamma.$$

The implicit constant depends on the kernel hyperparameters, the variance matrix of the covariate distribution, D and R. With the same assumptions but applying the RLS algorithm of Musco and Musco (2017) to selecting inducing inputs, for any $\delta \in (0, 1/32)$ there exists a choice of S such that with probability $1 - 5\delta$, $\mathbf{M} = \mathcal{O}\left(\left(\log \frac{N^2}{\delta\gamma}\right)^D (\log\log \frac{N^2}{\delta\gamma} + \log(1/\delta)\right)$ and

$$\mathrm{KL}[Q||P] \leq \gamma.$$

The proof follows from Proposition 21 and Theorem 13 or Theorem 17, by choosing parameters appropriately.

Remark 23 *If we allow the implicit constant to depend on γ and δ, this implies that for inducing ploints distributed accoding to an approximate M-DPP we can choose $M = \mathcal{O}((\log N)^D)$ inducing points leading to a computational cost of $\mathcal{O}(N(\log N)^{3D+1})$ while for approximate ridge leverage scores sampling $\mathcal{O}((\log N)^D \log\log N)$ inducing points suffice leading to a $\mathcal{O}(N(\log N)^{2D}(\log\log N)^2)$ computational cost.*

In order for the KL-divergence to be less than a fixed constant, the exponential scaling of the number of inducing points in the dimensions of the covariates is inevitable, as we will show in Section 6. However, practically the situation may not be quite so dire. First, many practioners use a SE-ARD kernel. If the data is essentially constant over many dimensions, then when training with empirical Bayes, the lengthscales of these dimensions tends to become large, effectively reducing the dimensionality of the inference problem. Additionally, in the case when covariates fall on a smooth, low-dimensional manifold, the decay of the eigenvalues only depends on the dimensionality and smoothness properties of this manifold, see Altschuler et al. (2019, Theorem 4). In addition, for a given problem, the dimensionality D is fixed, meaning that the dependence of the number of inducing points M depends polylogarithmically on N. This growth is slower than any polynomial, i.e. $(\log N)^D = o(N^\epsilon)$ for $\epsilon > 0$.

We also note that Corollary 22 can easily be adapted using Lemma 11 to show that if all of the $\mathbf{x}_n$ are drawn from any compactly supported distributions with continuous densities that are all bounded by some universal constant, the same asymptotic bound on the number of inducing points applies. This follows from noting that under these assumptions, $p_n(x)$, satisfies $p_n(x) < cq(x)$ where $q(x)$ is a Gaussian density for some $c > 0$, so we can apply Lemma 11 to bound the expectation of the sum of the matrix eigenvalues associated to p_n in terms of the eigenvalues associated to q.

5.2. Compactly Supported Inputs and Stationary Kernels

For most kernels and covariate distributions, solutions to the eigenfunction problem, $\mathcal{K}\phi = \lambda\phi$ cannot be found in closed form. In the case of stationary kernels defined on $\mathbb{R}^D$, that is kernels satisfying $k(x, x') = \kappa(x - x')$ for some $\kappa : \mathbb{R}^D \to \mathbb{R}$, the asymptotic properties of the eigenvalues are often understood (Widom, 1963, 1964).

Stationary, continuous kernels can be characterized through Bochner's theorem, which states that any such kernel is the Fourier transform of a positive measure, i.e.

$$\kappa(x - x') = \int_{\mathbb{R}^D} s(\omega) \exp(i\omega \cdot (x - x'))d\omega.$$

We will refer to $s(\omega)$ as the spectral density of k.[6] The decay of the spectral density conveys information about how smooth the kernel function is.

Widom's theorem (Widom, 1963) relates the decay of the eigenvalues of $\mathcal{K}$ to the decay of s. Widom's theorem applies to input distributions with compact support and stationary kernels with spectral density satisfying several regularit conditions (stated in Appendix D). Seeger et al. (2008) give a corollary of Widom's theorem, which is sufficient in many instances to obtain bounds on the number of inducing points needed for Theorems 13 and 14 to converge.

Lemma 24 (Seeger et al., 2008, Theorem 2) *Let k be an isotropic kernel (i.e. $\kappa(\alpha) = \kappa(\alpha')$ if $\|\alpha\| = \|\alpha'\|$). Suppose k satisfies the criteria of Widom's theorem, the covariate distribution has density zero outside a ball of radius T around the origin, and is bounded above by τ, then*

$$\lambda_m \leq \tau(2\pi)^D s\left(\frac{2\Gamma(D/2+1)^{2/D}}{T} m^{1/D}\right)(1 + o(1)).$$

5.2.1. Matérn kernels and compactly supported input distribution

Matérn kernels are widely applied to problems where the data generating process is believed to lead to non-smooth functions, and are known to satisfy the conditions of Widom's theorem (Seeger et al., 2008). These kernels are defined as (Rasmussen and Williams, 2006),

$$k_{\text{Mat}}(x, x') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\|x - x'\|_2}{\ell}\right)^{\nu} K_\nu\left(\frac{\|x - x'\|_2}{\ell}\right), \tag{25}$$

where K_ν is a modified Bessel function. The spectral density of the Matérn kernel is

$$s(\omega) = \frac{\ell^D \Gamma(\nu + D/2)}{\pi^{D/2}\Gamma(\nu)}\left(1 + (\ell\omega)^2\right)^{-(\nu+\frac{D}{2})}$$

which is proportional to a Student's t-distribution with $2\nu + D$ degrees of freedom. This spectral density only decays polynomially, with the degree of the polynomial depending on ν and D. Here $\ell > 0$ is the lengthscale and $\nu > 0$ is a 'smoothness' parameter often chosen as $\nu \in \{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}\}$. The posterior mean is $\lfloor\nu\rfloor$-times differentiable, which relates to the slower decay of the spectral density.

6. We assume $\kappa(x - x')$ decays sufficiently rapidly so that such a continuous spectral density exists.

Lemma 24 tells us that for compactly supported covariates with bounded density and the Matérn kernel with smoothness paramater ν

$$\lambda_m = \mathcal{O}(m^{\frac{-2\nu-D}{D}}).$$

It follows that $\sum_{m=M+1}^{D} \lambda_m = \mathcal{O}(M^{\frac{-2\nu}{D}})$. From this, we can derive a result of the same form as Corollary 22 for Matérn kernels and compactly supported input distributions.

Corollary 25 *Suppose the conditions of Theorem 13 on* $\mathbf{X}$ *and* $\mathbf{y}$ *are satisfied for some* $R > 0$. *Let* k *be a Matérn kernel with smoothness parameter* ν. *Suppose that* N *covariates are observed, each with an identical distribution with bounded density and compact support on* $\mathbb{R}^D$. *Fix any* $\gamma \in (0, 1]$. *Then for* $\nu > D/2$ *if inducing points are initialized using an* ϵ*-approximate* M*-DPP with and* $\epsilon = \mathcal{O}(N^2/\gamma)$ *there exists an* $M = \mathcal{O}(N^{\frac{2D}{2\nu-D}}\gamma^{\frac{D}{2\nu-D}})$ *such that*

$$\mathbb{E}[\mathrm{KL}[Q||P]] \leq \gamma.$$

Under the same assumptions if inducing points are initialized using the RLS algorithm of Musco and Musco (2017) with $\delta \in (0, 1/32)$ *there exists an* $S = \mathcal{O}\left(N^{\frac{2D}{2\nu+D}}(\gamma\delta^2)^{\frac{-D}{2\nu+D}}\right)$ *such that with probability at least* $1 - 5\delta$,

$$\mathrm{KL}[Q||P] \leq \gamma,$$

and $\mathbf{M} \leq S \log \frac{S}{\delta}$.

Remark 26 *If we consider* γ *and* δ *as fixed constants (independent of* N*), this implies that for an initialization with* M*-DPP we can choose* $M = \mathcal{O}(N^{\frac{2D}{2\nu-D}})$ *inducing points leading to a computational cost of* $\mathcal{O}(N^{\frac{2\nu+5D}{2\nu-D}}\log(N))$ *while for approximate ridge leverage scores sampling* $\mathcal{O}(N^{\frac{2D}{2\nu+D}}\log N)$ *inducing points suffice leading to a cost at most* $\mathcal{O}(N^{\frac{2\nu+5D}{2\nu+D}}(\log N)^2)$.

The first part corollary follows from Theorem 13, noting that we need to choose M such that

$$CN^2M\sum_{m=M+1}^{\infty}\lambda_m \leq C'N^2M^{\frac{D-2\nu}{D}} \leq \gamma\delta/2$$

for some constants C, C'. The second part follows from similar considerations applied to Theorem 17.

These bounds on M are vacuous (i.e. are no smaller than $M = \mathcal{O}(N)$) for Matérn kernels in high dimensional spaces or with low smoothness parameters. Additionally, the cost of sampling the M-DPP using Algorithm 1 makes this inference scheme less expensive than exact GP inference only when $\nu > 2D$. If we instead make the stronger assumptions required by Theorem 14, we can choose $M = \mathcal{O}(N^{\frac{D}{2\nu-D}})$, which implies a computational complexity less than exact GP regression if $\nu > \frac{5}{4}D$.

The bounds for the RLS initialization are generally sharper, and are non-vacuous for all $\nu > D/2$ with the weaker assumptions on $\mathbf{y}$. Additionally, the computational complexity of choosing inducing points using the RLS algorithm is the same as the cost of inference up to logarithmic factors, so that for $\nu > D/2$ the cost of sparse inference with the RLS initialization is (asymptotically) smaller than the cubic cost of exact GP regression.

6. Lower Bounds on the Number of Inducing Points Needed

In Sections 4 and 5, we showed that for many problems the number of inducing points can grow sub-linearly with the number of data points, while maintaining a small KL-divergence between the approximate and exact posteriors. In this section we consider the inverse question, i.e. how many inducing points are necessary to avoid having the KL-divergence grow as the amount of data increases? In this section, we prove a-priori lower bounds on the KL-divergence under similar assumptions to those used in proving the upper bounds in Section 4.

Naively, it appears that the lower bound in Lemma 4 gives us a starting place for a lower bound on the KL-divergence. From this bound,

$$\mathbb{E}[\mathrm{KL}[Q||P] \mid \mathbf{Z}, \mathbf{X}] \geq \frac{\mathbf{t}}{2\sigma^2}. \tag{26}$$

While lower bounding this quantity can be done using the approach taken in this section, it is not the most interesting quantity to study, as we average over $\mathbf{y}$ *conditioned* on $\mathbf{X}$ and $\mathbf{Z}$. This would not give a valid lower bound if the locations of the inducing points depend on $\mathbf{y}$, as illustrated in Fig. 1. This approach would establish a lower bound for initialization schemes considered in the previous sections, as well as any initialization scheme that does not take the observed y into account, but not the common practice of performing gradient ascent on the ELBO with respect to inducing inputs.

In this section, we establish a lower bound on the number of inducing variables needed for the KL-divergence not to become large, which is valid regardless of the method for selecting inducing variables or the distribution of $\mathbf{y}$. These bounds assume that the covariates are independent and identically distributed (in contrast to the upper bounds, which do not require independence and require a slightly weaker condition than identical marginals). The independence assumption is necessary in order to lower bound the eigenvalues of the covariance matrix. For example, if all of the covariates were identically distributed and equal, the covariance matrix would be rank-1 and so a single inducing point could be used regardless of the size of the data set.

The proof of the lower bounds proceeds in two parts:

1. First, we derive a lower bound on $\mathrm{KL}[Q||P]$ that holds for any y and Z, but depends on the eigenvalues of $\mathrm{K_{ff}}$.

2. Second, we use a result on the concentration of eigenvalues of the kernel matrix to those of the corresponding operator due to Braun (2006) to derive a lower bound that holds with fixed probability under the assumption that the covariates are independent and identically distributed.

In the case of SE-kernel and Gaussian covariates, we establish a lower bound with the same dependence on N as our upper bounds, that is we need $M = \Omega((\log N)^D)$. In the case of Matérn kernels with uniform covariates and $\nu > 1$, we establish a lower bound that increases as a power of N. However, there is a large gap between our upper and lower bounds for Matérn kernels, indicating room for improvement. These results are summarized in Table 2. While our results are stated in terms of inducing points, they hold for more general inducing variables.

Kernel	Input Distribution	M
SE-Kernel	Gaussian	$\Omega((\log N)^D)$
Matérn ν	Uniform	$\Omega\left(N^{\frac{2\nu D}{(2\nu+5D)(2\nu+D)}-\epsilon}\right)$

Table 2: The lower bounds we establish on the number of features needed so that the KL-divergence does not increase as a function of N. Here ϵ is a positive constant that can be chosen arbitrarily close to 0. While the bound for the the SE-kernel matches our upper bounds up to terms that are constant in N, we expect the lower bound for the Matérn kernel can be raised significantly.

6.1. A Lower Bound

In this section, we derive a lower bound on the KL-divergence that holds for any y and Z and depends on $\mathbf{X}$.

Lemma 27 *Given a kernel k, likelihood model with variance σ^2 and random covariates $\mathbf{X}$. Then,*

$$\min_{Z\in\mathcal{X}^M}\min_{y\in\mathbb{R}^N}\mathrm{KL}[Q||P] \geq \frac{1}{2}\sum_{m=M+1}^{N}\frac{\tilde{\boldsymbol{\lambda}}_m}{\sigma^2} - \log\left(1+\frac{\tilde{\boldsymbol{\lambda}}_m}{\sigma^2}\right)$$

where $\tilde{\boldsymbol{\lambda}}_m$ denotes the m^{th} largest eigenvalue of the matrix $\mathbf{K_{ff}}$ determined by the covariates and kernel.

The proof (Appendix E) follows from noting that for any X, y, Z we have $\mathrm{KL}[Q||P] = \log p(y) - \mathcal{L}(y, Z) \geq \log p(0) - \mathcal{L}(0, Z)$, where we have defined $\mathcal{L}(y, Z)$ to be the evidence bound resulting from the triple X, y, Z (and suppressed dependence on X). The bound follows from writing both the trace and log determinant in terms of eigenvalues and using that $\mathbf{K_{ff}} \succ \mathbf{Q_{ff}}$, so that the m^{th} largest eigenvalue of $\mathbf{K_{ff}}$ is greater than the m^{th} largest eigenvalue of $\mathbf{Q_{ff}}$ for all $1 \leq m \leq N$.

In order to establish lower bounds on the number of inducing variables needed to ensure the KL-divergence does not grow as a function of N, it suffices to analyze the behavior of the lower bound in Lemma 27 for random covariates as a function of both M and N.

6.2. Structure of the Argument

In order to derive a lower bound on the KL-divergence Lemma 27, we can consider just the largest term in the sum appearing in Lemma 27, as all the terms are non-negative. For $a > 3$, $\log(1 + a) \leq a/2$. Therefore, if $\tilde{\boldsymbol{\lambda}}_{M+1}/\sigma^2 > 3$, we have $\mathrm{KL}[Q||P] \geq \frac{\tilde{\boldsymbol{\lambda}}_{M+1}}{4\sigma^2}$.

Under the supposition that $\tilde{\boldsymbol{\lambda}}_{M+1}/\sigma^2 > 3$, we can apply the triangle inequality to Lemma 27 to give,

$$\mathrm{KL}[Q||P] \geq \frac{N(\lambda_{M+1} - |\lambda_{M+1} - \frac{1}{N}\tilde{\boldsymbol{\lambda}}_{M+1}|)}{4\sigma^2}$$

Therefore, for any $M = M(N)$ such that:

1. We can show a relative error bound on the approximation of matrix eigenvalues with operator eigenvalues of the form $\frac{|\lambda_{M+1} - \frac{1}{N}\tilde{\boldsymbol{\lambda}}_{M+1}|}{\lambda_{M+1}} < 1 - \gamma_N$ for some $\gamma_N \in (0,1)$,
2. $N\gamma_N\lambda_{M+1}$ tends to infinity as N tends to infinity,

it must be the case that the KL-divergence tends to infinity as a function of N (at a rate $\Omega(N\gamma_N\lambda_{M+1})$).

6.3. Concentration of Eigenvalues

In order to complete the argument in the previous section, we need a more fine-grained understanding of the behavior of eigenvalues of $\mathbf{K_{ff}}$ than given in Lemma 11. For this, we rely on the following result:

Lemma 28 (Braun, 2006, Theorem 4) *Let k be a continuous kernel with $k(x,x) \leq v$ for all $x \in \mathbb{R}^D$. Fix $\delta \in (0,1)$. Suppose $\mathbf{x}_1, \cdots, \mathbf{x}_N$ are realizations of i.i.d. random variables sampled according to some measure on $\mathbb{R}^D$ with density $p(x)$. Then for all m and any $1 \leq r \leq N$, with probability at least $1-\delta$,*

$$\begin{aligned} |\lambda_m - \frac{1}{N}\tilde{\boldsymbol{\lambda}}_m| &\leq \lambda_m r\sqrt{\frac{r(r+1)v}{\lambda_r N\delta}} + \sum_{s=r}^{\infty}\lambda_s + \sqrt{\frac{2v\sum_{s=r+1}^{\infty}\lambda_s}{N\delta}} \\ &= \mathcal{O}\left(\lambda_m r^2 \lambda_r^{-1/2} N^{-1/2}\delta^{-1/2} + \sum_{s=r}^{\infty}\lambda_s + \sqrt{\frac{\sum_{s=r+1}^{\infty}\lambda_s}{N\delta}}\right). \end{aligned} \tag{27}$$

where λ_m is the m^{th} eigenvalue of the integral operator $\mathcal{K} : (\mathcal{K}g)(x') = \int g(x)k(x,x')p(x)dx$ and $\tilde{\boldsymbol{\lambda}}_m$ is the m^{th} eigenvalue of $\mathbf{K_{ff}}$.

For the remainder of this section, we consider specific cases of kernel and input distributions for which we know properties of the spectrum, and derive lower bounds on the number of features needed so that the KL-divergence is not an increasing function of N.

6.4. Squared Exponential Kernel and Gaussian Covariates

We begin with the one-dimensional SE kernel and Gaussian covariates. The multivariate case follows a similar, though more involved argument and will be discussed in Section 6.5. Recall, if the covariates have variance β^2, then for any $r \in \mathbb{N}$

$$\lambda_r = v\sqrt{\frac{2a}{A}}B^{r-1} \quad \text{and} \quad \sum_{s=r}^{\infty}\lambda_s = \frac{\lambda_r}{1-B},$$

where $a = (4\beta^2)^{-1}, b = (2\ell^2)^{-1}, A = a + b + \sqrt{a^2 + 4ab}$ and $B = b/A$. For some $\eta \in (0,1)$, choose $r = 1 + \lceil \log_B(1-B)\sqrt{A/(2av^2)}N^{-\eta}\rceil$, so that $\sum_{s=r}^{\infty}\lambda_r \leq N^{-\eta}$ and $\lambda_r \geq B(1-B)N^{-\eta}$. Hence Eq. (27) implies that for all m with probability at least $1-\delta$,

$$\frac{|\lambda_m - \frac{1}{N}\tilde{\boldsymbol{\lambda}}_m|}{\lambda_m} \leq r\sqrt{\frac{r(r+1)v}{B(1-B)N^{1-\eta}\delta}} + \lambda_m^{-1}N^{-\eta} + \sqrt{\frac{2v}{\lambda_m^2 N^{1+\eta}\delta}}. \tag{28}$$

For any fixed $\delta \in (0,1)$ the first term on the right hand side tends to zero with N since $\eta < 1$. For any $M \leq \log_B(\sqrt{A/(2av^2)}N^{-\eta}\sqrt{\delta})$,

$$\frac{|\lambda_{M+1} - \frac{1}{N}\tilde{\boldsymbol{\lambda}}_{M+1}|}{\lambda_{M+1}} \leq r\sqrt{\frac{r(r+1)v}{B(1-B)N^{1-\eta}\delta}} + \frac{\sqrt{\delta}}{2} + \sqrt{\frac{v}{2N^{1-\eta}}}. \tag{29}$$

The second term is less than $1/2$ and the last term tends to 0 for large N. We conclude that for any such M, the KL-divergence is bounded below by $c_N \frac{N\lambda_{M+1}}{8\sigma^2} = \Omega(N^{1-\eta})$, where $\lim_{N\to\infty} c_N = 1$.

Remark 29 *For univariate Gaussian kernels, if we choose $\eta = .01$ in the above argument, we get that for any $M \leq \frac{1}{\log(1/B)}\log(N^{.01}\sqrt{\frac{2av^2}{\delta A}}) = \Omega(\log N)$, the KL-divergence is $\Omega(N^{.99})$, and will therefore be large as N increases. We therefore need M to grow faster than this to avoid this if we want the KL-divergence to be small for large N.*

6.5. The Isotropic SE-kernel and Multidimensional Gaussian Covariates

In order to obtain lower bounds in the multivariate case, we first obtain a lower bound on the individual eigenvalues of the operator $\mathcal{K}$.

Proposition 30 *Suppose k is an isotropic SE-kernel in D dimensions with lengthscale ℓ and variance v. Suppose the training covariates are independently identically distributed according to an isotropic Gaussian measure, μ, on $\mathbb{R}^D$ with covariance matrix $\beta^2\mathrm{I}$. For any $r \in \mathbb{N}$, we have*

$$\lambda_r \geq \left(\frac{2a}{A}\right)^{D/2} B^{Dr^{1/D}}.$$

where λ_r denotes the r^{th} largest eigenvalue of the operator $\mathcal{K} : L^2(\mathbb{R}^D, \mu) \to L^2(\mathbb{R}^D, \mu)$ defined by $(\mathcal{K}g)(x') = \int g(x)k(x,x')p(x)dx$ with $p(x)$ the density of the multivariate Gaussian at x.

The proof (Appendix E) relies on a counting argument and standard bounds on binomial coefficients. We can now combine Lemma 28 and Propositions 21 and 30 in order to bound the multivariate SE-kernel with Gaussian inputs.

Proposition 31 *Let k be an isotropic SE-kernel. Suppose N covariates are sampled independent and identically from an isotropic Gaussian density with variance β^2 along each dimension. Define $M(N)$ to be any function of N such that $\lim_{N\to\infty} M(N)/(\log N)^D = 0$; i.e. $M(N) = o((\log N)^D)$. Suppose inference is performed using any set of inducing inputs, Z such that $|Z| = M(N)$. Then for any $y \in \mathbb{R}^N$, for any $\epsilon > 0$ and for any $\delta \in (0,1)$, with probability at least $1-\delta$, $\mathrm{KL}[Q||P] = \Omega(N^{1-\epsilon})$.*

Remark 32 *To illustrate the meaning of this lower bound, consider the case when $M(N) = (\log N)^{D-1}$. Then for N sufficiently large, we have $\mathrm{KL}[Q||P] > N^{.99}$ implying for large N the KL-divergence must be large. On the other hand from our upper bounds in Section 5, we know that if we fix any positive constant γ there exists a constant C (that does not depend on N) such that if inference is performed with $M(N) = c(\log N)^D$ inducing inputs placed according to an approximate M-DPP, the KL-divergence is less than γ in expectation.*

The proof follows from Lemma 27 by choosing r appropriately in Lemma 28 to bound the empirical eigenvalues and using Propositions 21 and 30 to bound eigenvalues in the appropriate directions to control the error term. Details are given in Appendix E.

6.6. Lower Bounds for Kernels with Polynomial Decay

As discussed in Section 5 some popular choices of kernels lead to eigenvalues that decay polynomially instead of exponentially. For example, Widom (1963, Theorem 2.1) implies that the eigenvalues of the operator associated to the Matérn kernel with smoothness parameter ν and covariates uniformly distributed in the unit cube has eigenvalues satisfying $C_1 m^{\frac{-2\nu+D}{D}} \leq \lambda_m \leq C_2 m^{\frac{-2\nu+D}{D}}$ for some constant C_1 and C_2 independent of m i.e. $\lambda_m = \Theta(m^{\frac{-2\nu+D}{D}})$.[7]

Proposition 33 *Let k be a continuous kernel, and μ a measure on $\mathbb{R}^D$ with density p such that the associated operator $\mathcal{K}$ has eigenvalue satisfying $C_1 m^{-\eta} \leq \lambda_m \leq C_2 m^{-\eta}$ for all $m \geq 1$, some $\eta > 1$ and constants $C_1, C_2 > 0$. Suppose inference is performed using any set of inducing inputs Z such that $|Z| = M(N)$ with $M(N)$ any function such that $\lim_{N\to\infty} \frac{M(N)}{N^\zeta} = c$ (i.e. $M = O(N^\zeta)$) for some $c < \infty$ and $\zeta \in (0, \frac{\eta-1}{\eta(4+\eta)})$. Then for any $\delta \in (0,1)$ with probability at least $1-\delta$, $\mathrm{KL}[Q||P] = \Omega(N^{1-\eta\zeta})$.*

Proof We have $\lambda_r = \Theta(r^{-\eta})$ so $\sum_{s=r}^{\infty} \lambda_s = \Theta(r^{1-\eta})$. Choose $r = N^\gamma$ for some $\gamma \in (0,1)$ and $M+1 = N^\zeta$. In this case the error term in Lemma 28 becomes:

$$\frac{|\lambda_{M+1} - \frac{1}{N}\tilde{\boldsymbol{\lambda}}_{M+1}|}{\lambda_{M+1}} = \mathcal{O}\Big(\delta^{-1/2}(N^{2\gamma+\gamma\eta/2-1/2} + N^{\zeta\eta+\gamma(1-\eta)} + N^{\zeta\eta+\gamma(1-\eta)/2-1/2})\Big).$$

Following the earlier proof sketch, we must show the RHS tends to a value less than 1. To ensure the first of the three summands is small, we choose $\gamma \in (0, \frac{1}{4+\eta})$. Given this choice, for large N the third summand in the error term is always smaller than the second, so that this entire term is $o(1)$ given the supposition that $\zeta \leq \gamma\frac{\eta-1}{\eta}$. We conclude that if $M = N^{-\zeta}$ for $\zeta \in (0, \gamma\frac{\eta-1}{\eta})$ with probability $1-\delta$, the lower bound in Lemma 27 is at least $N\lambda_{M+1} = \Omega(N^{1-\zeta\eta})$. ■

In the case of D-dimensional Matérn kernels and a uniform covariate distribution ($\eta = \frac{2\nu+D}{D}$), by choosing ζ as large as possible, this means that for an arbitrary $\epsilon > 0$, the KL-divergence is lower bounded by an increasing function of N if fewer than $\Omega\Big(N^{\frac{2\nu D}{(2\nu+5D)(2\nu+D)}-\epsilon}\Big)$ inducing variables are used. This lower bound on the number of inducing variables becomes vacuous (i.e. the exponent tends to 0) as $\eta \to 1$ from above, meaning it is not useful when applied to many kernels that we expect would be very difficult to approximate. There is a large gap between the upper and lower bound, particularly when η is near 1 (i.e. for non-smooth kernels). The gap between the bounds is in part introduced by needing to choose M so that the error term from Lemma 28 remains lower order. If we heuristically allow ourselves to replace matrix eigenvalues with the corresponding scaled operator eigenvalues and neglect the error term, we obtain a lower bound of $\Omega(N^{\frac{1}{\eta}})$, bringing the lower bound

7. See Seeger et al. (2008) for more details on the derivation of this from Widom's Theorem.

more closely in line with the upper bound of $\mathcal{O}(N^{\frac{1}{\eta-1}})$. The remaining gap between these bounds is essentially due to only bounding a single eigenvalue in the lower bound, while bounding the sum of eigenvalues in the upper bound. Improving the analysis to close the gap between the upper and lower bounds is important for better understanding the efficacy of sparse methods with non-smooth kernels.

7. Practical Considerations

Up to this point, we proved statements about the *asymptotic* scaling properties of variational sparse inference. Our results indicated which models could be well-approximated with relatively few inducing points for sufficiently large data sets. In this section, we investigate the limitations and practical implications of our results to real situations with finite amounts of data. We consider the applicability of our results to practical implementations, and perform empirical analyzes on how marginal likelihood bounds converge. Additionally, our proof suggests a specific procedure for choosing inducing points that differs from methods that are currently commonly applied. We empirically investigate this procedure, and provide recommendations on how to initialize inducing points.

7.1. Finite Precision in Practical Implementations

Any practical implementation of a Gaussian process method will be influenced by the finite precision with which floating-point numbers are represented in a computer. These issues are not explicitly addressed in our mathematical analysis, which assume calculations are in exact arithmetic. Here, we briefly discuss the effects of this finite precision on 1) the implementation, 2) the precision to which we can expect convergence in practice compared to our analysis, and 3) the way that this is quantified by marginal likelihood bounds.

7.1.1. Ill-conditioning & Cholesky Decomposition Failure

Finding various quantities for Gaussian process regression requires computing log determinants and matrix inverses. When the smallest and largest eigenvalues of the kernel matrix are many orders of magnitude apart, these computations become *ill-conditioned*, meaning that small changes on the input can lead to large changes to the output. For example, tiny changes in the elements of the vector $\mathbf{f}_X$ can lead to huge variations in the vector $\mathrm{K}_{\mathrm{ff}}^{-1}\mathbf{f}_X$ when K_{ff} has an eigenvalue close to zero (see Deisenroth et al., 2019, §6.2 for a visual illustration). This typically occurs when considering many highly-correlated inputs to the GP (e.g clusters of nearby points with similar input values). These points have a high probability of having very similar function outputs under the prior. This ill-conditioning arises naturally in GPs when considering e.g. evaluating the prior density on function values: small differences in the function values result in huge changes to the value of the probability density. If the sensitivity of the calculations becomes too large, then the finite precision with which numbers are represented can lead to considerable error.

In the variational methods we consider, determinants and inverses are found based on the Cholesky decomposition of the kernel matrix: $\mathrm{K}_{\mathrm{ff}} + \sigma^2\mathrm{I}$ for exact implementations, and

K_{uu} for sparse approximations.[8] When faced with a problem that is too ill-conditioned, most Cholesky implementations terminate with an exception. This can be seen as desirable from the point of view that a successful run usually indicates an accurate result.

Even in cases when the data set can be well-described by a GP model with hyperparameters that lead to reasonably well-conditioned matrices, conditioning problems frequently arise during training, when the log marginal likelihood or ELBO is values for other candidate hyperparameter values. For example, for stationary kernels, large lengthscales contribute to conditioning problems, as they increase the correlation between distant points. Hyperparameters are typically found by (approximately) maximizing the log marginal likelihood (Eqs. 6 and 9). Since these objective and their derivatives can be evaluated in closed-form, fast-converging quasi-Newton methods such as (L-)BFGS are commonly used. These methods often propose large steps, which lead to the evaluation of hyperparameter settings where the Cholesky decomposition raises an exception. Even though these hyperparameter settings are often of poor quality, (L-)BFGS still requires an evaluation of the objective function to continue the search. The Cholesky errors must therefore be avoided to successfully complete the entire optimization procedure.

7.1.2. Improving Matrix Conditioning

Increasing the smallest eigenvalue of the kernel matrix improves the conditioning. In exact implementations this can be done by increasing the likelihood noise variance, as we need to decompose $K_{ff} + \sigma^2 I$ which has eigenvalues that are lower-bounded by σ[9]. On the other hand, the sparse variational approximation requires inverting K_{uu} *without any noise*. However, it is important to note that the conditioning of K_{uu} is better than K_{ff} for two reasons. Firstly, it is a smaller matrix, and often issues of conditioning are less severe for smaller matrices. Secondly, if inducing points are selected using a method that introduces negative correlations clusters of highly-correlated points are unlikely to appear in K_{uu}. Nevertheless, it is still possible for the Cholesky decomposition to fail, particularly when trying different hyperparameter settings when maximizing the ELBO.

To improve robustness in the sparse approximation, a small diagonal "jitter" matrix ϵI, with ϵ commonly around 10^{-6}, is added to K_{uu}, introducing a lower bound the on its eigenvalues. This change is often enough to avoid decomposition errors during optimization. While this modification changes the problem that is solved, the effect is typically small. Some software packages (e.g. GPy, since 2012) increase jitter adaptively by catching exceptions inside the optimization loop to only introduce bias where it is necessary.

7.1.3. Quantifying the Effect of Jitter

Adding jitter to the covariance matrix K_{uu} corresponds to defining the inducing variables as noisy observations of the GP. While this still produces a valid approximation to the posterior and ELBO (Titsias, 2009a), the approximation obtained is generally of marginally lower quality and there is a small amount of corresponding slack in the ELBO (Matthews, 2016,

8. Conjugate gradient and Lanczos methods also give exact answers when they are run for sufficient iterations, and have been successfully applied in practice (Gibbs and Mackay, 1997; Davies, 2015; Gardner et al., 2018).
9. This can be done by reparameterizing the noise to have a lower bound.

Theorem 4). We summarize the effect on the ELBO and upper bound $\mathcal{U}_2$ in the following proposition. Define $\mathrm{Q}_{\mathrm{ff}}(\epsilon) \coloneqq \mathrm{K}_{\mathrm{uf}}^{\mathsf{T}}(\mathrm{K}_{\mathrm{uu}} + \epsilon\mathrm{I})^{-1}\mathrm{K}_{\mathrm{uf}}$, so that $\mathrm{Q}_{\mathrm{ff}}(0) = \mathrm{Q}_{\mathrm{ff}}$.

Proposition 34 *Let $\mathcal{L}_\epsilon$ denote the evidence lower bound computed with jitter $\epsilon \geq 0$ added to K_{uu}, that is*

$$\mathcal{L}_\epsilon = -\frac{1}{2}\log\det(\mathrm{Q}_{\mathrm{ff}}(\epsilon) + \sigma^2\mathrm{I}) - \frac{1}{2}y^{\mathsf{T}}(\mathrm{Q}_{\mathrm{ff}}(\epsilon) + \sigma^2\mathrm{I})^{-1}y - \frac{N}{2}\log 2\pi - \frac{1}{2\sigma^2}\mathrm{tr}(\mathrm{K}_{\mathrm{ff}} - \mathrm{Q}_{\mathrm{ff}}(\epsilon)).$$

Then $\mathcal{L}_\epsilon$ is monotonically decreasing in ϵ. Similarly if $\mathcal{U}_\epsilon$ denotes the upper bound Eq. (12) computed with added jitter to K_{uu}, that is

$$\mathcal{U}_\epsilon \coloneqq -\frac{1}{2}\log\det(\mathrm{Q}_{\mathrm{ff}}(\epsilon) + \sigma^2\mathrm{I}) - \frac{1}{2}y^{\mathsf{T}}(\mathrm{Q}_{\mathrm{ff}}(\epsilon) + \mathrm{tr}(\mathrm{K}_{\mathrm{ff}} - \mathrm{Q}_{\mathrm{ff}}(\epsilon))\mathrm{I} + \sigma^2\mathrm{I})^{-1}y - \frac{N}{2}\log 2\pi.$$

Then $\mathcal{U}_\epsilon$ is monotonically increasing in ϵ. In particular, adding jitter can only make the upper bound on the log marginal likelihood larger and the ELBO smaller.

The proof (Appendix F) is a consequence of $\mathrm{Q}_{\mathrm{ff}}(\epsilon) + \sigma^2\mathrm{I} \succ \mathrm{Q}_{\mathrm{ff}}(\epsilon') + \sigma^2\mathrm{I}$ for $\epsilon' > \epsilon \geq 0$. Proposition 34 shows that even with jitter the upper and lower bounds are still valid. However, they are not exactly equal, even when $M = N$, due to the additional gap caused by the jitter. From a practical point of view, the impact is typically very small, with a gap being introduced on the order of a few nats.

To summarize, we saw 1) that jitter was needed to stabilize the computation of the hyperparameter objective functions using standard implementations of Cholesky decomposition, and 2) that jitter and finite floating-point precision prevented the approximate posterior and bounds from converging to their exact values. Notably, we can quantify the effect of the finite precision calculations *using the same bounds* as what is used to determine the effect of using an approximate inducing point posterior (Proposition 34). The variational bounds we analyzed in this work therefore provide a unified way of measuring the effect of both exact arithmetic approximate posteriors and the impact of finite precision arithmetic on the quality of the approximation.

7.2. Placement of Inducing Inputs

When training a sparse Gaussian process regression (Titsias, 2009b) model, we need to select the kernel hyperparameters as well as the inducing inputs, with the hyperparameters determining the generalization characteristics of the model, and the inducing inputs the quality of the sparse approximation. In the time since Snelson and Ghahramani (2006) and Titsias (2009b) introduced joint objective functions for all parameters, it has become commonplace to find the final set of inducing variables by optimizing the objective function together with the hyperparameters. Because this makes the inducing input initialization procedure less critical for final performance, less attention has been placed on it in recent years than in e.g. the kernel ridge regression literature.[10] However, the number of optimization parameters added by the inducing inputs is often large, and convergence can be slow, which makes the optimization cumbersome.

10. Kernel ridge regression lacks a joint objective function for the approximation and hyperparameters. Hyperparameters are commonly selected through cross-validation.

Our results suggest that gradient-based optimization of the ELBO is not necessary for setting the inducing variables. Selecting enough of inducing points is sufficient for obtaining an arbitrarily good approximation. For the common squared exponential kernel if hyperparameters are fixed, our upper and lower bounds imply that optimization of inducing points leads to at most a constant factor fewer inducing points than a good initialization. In this section, we investigate the performance of various inducing point selection methods in practice. We consider commonly used methods (uniform sampling, K-means, and gradient-based optimization), and methods that our proofs are based on (M-DPP and RLS). In addition, we propose using the intialization used for approximately sampling the M-DPP (Algorithm 1) as an inducing point selection method. While our theoretical results do not prove anything for this method, it avoids the additional cost and complexity of running a Markov chain, while still being leading to negative correlations between inducing point locations. We refer to this method as *greedy variance selection*, since it greedily selects the next inducing point based on which has the highest marginal variance in the conditioned prior $p(\mathbf{f} \mid \mathbf{u})$, i.e. $\operatorname{argmax} \operatorname{diag}[\mathrm{K_{ff}} - \mathrm{K_{uf}}\mathrm{K_{uu}^{-1}}\mathrm{K_{uf}}]$.[11]

We set the free parameters for each of the methods as follows. For K-means, we run the Scipy implementation of K-means++ with M centres. Gradient-based optimization is initialized using greedy variance selection. This choice was made since it was found to perform better than uniform selection, and our goal is to quantify how much can be gained by doing gradient-based optimization, and whether it is worth the cost. We ran 10^4 steps of L-BFGS, at which point any improvement was negligible compared to adding more inducing variables. Approximate M-DPP sampling was done following Algorithm 1, using 10^4 iterations of MCMC. For RLS, we use an adaptation of the public implementation of Musco and Musco (2017, Algorithm 3), which omits many of the constants derived in the proofs, and therefore loses theoretical guarantees.[12] We additionally modify the algorithm to ensure that it selects exactly M inducing points.

We consider 3 data sets from the UCI repository that are commonly used in benchmarking regression algorithms, "Naval" ($N_{train} = 10740, N_{test} = 1194, D = 14$) , "Elevators" ($N_{train} = 14939, N_{test} = 1660, D = 18$) and "Energy" ($N_{train} = 691, N_{test} = 77, D = 8$). These data sets were chosen as near-exact sparse approximations could be found, so convergence could be illustrated.[13] Naval is the result of a physical simulation and the observations are essentially noiseless. To make statistical estimation more difficult, we add independent Gaussian noise with standard deviation 0.0068 to each observation. For all experiments, we use a squared exponential kernel with automatic relevance determination (ARD), i.e. a separate lengthscale per input dimension.

11. An equivalent approach, derived through different motivations, has been previously used for approximating the kernel matrix in SVMs (Fine and Scheinberg, 2001) and applied to sparse GP approximations (Foster et al., 2009).
12. Their implementation is available at: `https://github.com/cnmusco/recursive-nystrom`.
13. Not all data sets exhibit this property. For instance, the "kin40k" data set still isn't near convergence when $M = \frac{N}{2}$ due to very short optimal lengthscales. A step functions being present would cause this, and would indicate that squared exponential kernels are inappropriate.

7.2.1. Fixed Hyperparameters

We first consider regression with fixed hyperparameters to illustrate convergence in a situation that is directly comparable to our theoretical results. We investigate which of the inducing point selection methods recovers the exact model with the fewest inducing points. The hyperparameters are set to the optimal values for an exact GP model, or for "Naval" a sparse GP with 1000 inducing points. We find the hyperparmeters by maximizing the exact GP log marginal likelihood using L-BFGS. This setting is for illustrative purposes only, as computing exact log marginal likelihood is not feasible in practical situations where sparse methods are of actual interest. In the next section we consider hyperparameters that are learned using the ELBO (Eq. 9).

Figure 5 shows the performance of various methods of selecting inducing points as we vary M, as measured by the evidence lower bound, test root mean squared error and per data point test negative log predictive density. From the results, we can observe the following:

- For very sparse models where the ELBO is considerably lower than the true marginal likelihood, gradient-based tuning of the inducing inputs consistently performs best in all metrics.
- The benefit of gradient-based tuning is small when many inducing points are added, *provided they are added in the good locations*. Greedy variance selection and M-DPP find these good locations, as they consistently recover the true GP's performance with only a small number of additional inducing variables.
- K-means, uniform subsampling, and RLS tend to underperform, and require far more inducing variables to converge to the exact solution. In our experiment, they never converge quicker than greedy variance selection.
- In terms of the upper bound, greedy variance selection and M-DPP sampling both provide the best results.

Greedy variance selection seems to provide all the desirable properties in this case: convergence to the exact results with few inducing variables, simple to implement, and fast since it does not require as many expensive operations as optimization or sampling. The approximate ridge leverage score algorithm is also reasonably fast, and perhaps careful tuning of hyperparameters or different algorithms for approximate ridge leverage scores could lead to improved performance in practice.

7.2.2. Training procedure and hyperparameter optimization

In the previous section, the hyperparameters were fixed to values maximizing the log marginal likelihood. When sparse GP approximations are applied in practice, these optimal values are unknown, and they are instead found via maximizing the ELBO, as an approximation to maximizing the exact log marginal likelihood. This comes at the cost of introducing a bias in the hyperparameters towards models where $\mathrm{KL}[Q||P]$ is small (Turner and Sahani, 2011), as implied by Eq. (8). The most noticeable effect in sparse GP regression is the overestimation of σ^2 and a bias toward models with smoother sample functions (Bauer et al., 2016).

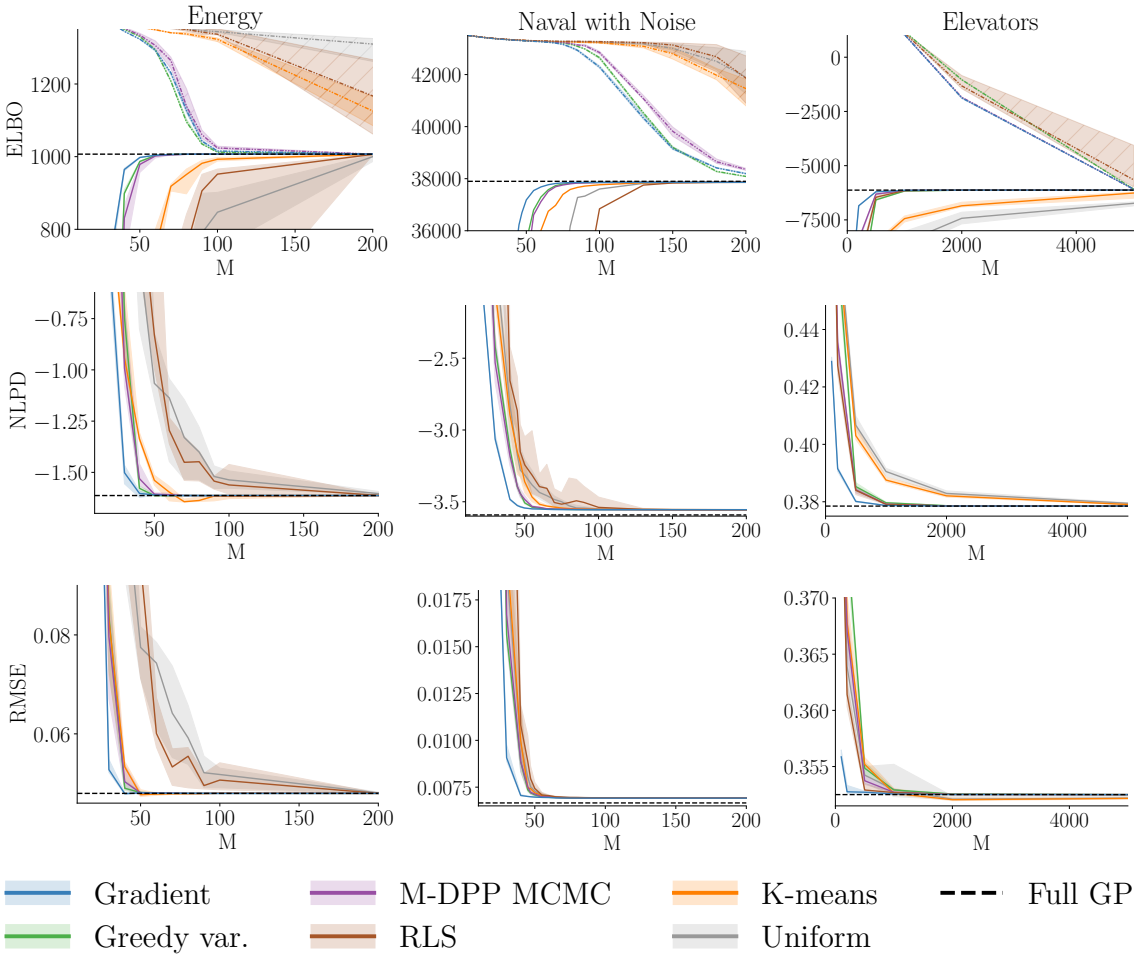

Figure 5: Performance of various methods for selecting inducing inputs on 3 data sets (each column corresponds to one data set) with fixed model hyperparameters. The top row shows the evidence lower bounds as well as an upper bound on the log marginal likelihood ($\mathcal{U}_2$ in Lemma 2), the middle row the per datapoint negative log predictive density on held-out test points and the bottom row the root mean square error on test data. The solid lines show the median of 10 initializations using the given method for selecting inducing inputs, while the shaded region represents the $20 - 80\%$. The dashed black line shows the performance of the exact GP regressor on the given data set.

Our results imply that for large enough data sets, an approximation with high sparsity can be found for the optimal hyperparameter setting that has a small KL-divergence to the posterior. This implies that the bias in the hyperparameter selection also is likely to be small. An impediment to finding the high-quality approximation for the optimal hyperparameters, is that our results depends on the inducing inputs being chosen based on

properties of the kernel with the same hyperparameters that are used for inference. Here, we investigate several procedures for jointly choosing the hyperparameters and inducing inputs, in the regime where enough inducing variables are used to recover a close to exact model.

We propose a new procedure based on the greedy variance selection discussed in the previous section. To account for the changing hyperparameters, we alternately optimize the hyperparameters, and reinitialize the inducing inputs with greedy variance selection *using the updated hyperparameters*. This avoids the high-dimensional non-convex optimization of the inducing inputs, while still being able to tailor the inducing inputs to the kernel. In effect, the method behaves a bit like variational Expectation-Maximization (EM) (Beal and Ghahramani, 2003), with the inducing input selection taking the place of finding the posterior. When enough inducing points are used, the reinitialization is good enough to make the ELBO almost tight for the current setting of hyperparameters. We terminate when the reinitialization does not improve the ELBO. We note that reinitialization would not benefit K-means or uniform initializations (beyond random chance), as the inducing points that are selected do not depend on the setting of the kernel hyperparameters.

We start our evaluation by running all methods from the previous section in addition to greedy variance selection with reinitialization (Fig. 6). The initial inducing inputs are set with the untrained initialized hyperparameters, after which the hyperparameters are maximized w.r.t. ELBO (Eq. 9) using L-BFGS, with the reinitialization being applied for "Greedy variance (reinit.)". We observe that the reinitialized greedy variance method provides consistent fast convergence to the exact model.

To evaluate the benefit of gradient-based optimization, we compare it to the reinitialized greedy variance method (the best from Fig. 6), as well as K-means. For the initial setting of the inducing inputs when optimizing inducing inputs, we use the greedy variance selection (denoted "gradient"). Since Fig. 6 shows that optimization of the inducing inputs is not needed to converge to the exact solution, the question becomes whether it is *faster* to perform gradient-based optimization. We choose M to be the smallest value for which the ELBO given by the gradient method converges to within a few nats of the exact marginal likelihood based on Fig. 5. We plot the optimization traces in Fig. 7 for several runs to account for random variation in the initializations.

In this constrained setting, we see different behaviours on the different data sets. One constant is that placing inducing points using K-means leads to sub-optimal performance compared to the best method. For the Energy data set, "greedy var (reinit)" suffers from convergence to local optima. This is caused by the low sparsity, and disappears if more inducing points are used (see Fig. 6). For the Naval data set, we see *very* slow convergence when using gradient-based optimization initialized with greedy variance selection. K-means underperforms and also suffers from local optima, with reinitialization reliably reaching the best ELBO. For elevators, reinitialization reaches the optimal ELBO fastest.

We note that in the reinitialization method the hyperparameter optimization step was terminated when L-BFGS had determined convergence according to the default Scipy settings. This leads to a characteristic "step" pattern in the optimization traces, where progress halts for many iterations towards the end of a hyperparameter optimization phase, followed by large gains after a reinitialization of the inducing inputs. By terminating the hyperparameter optimization earlier after signs of stagnation, the reinitialization method could be

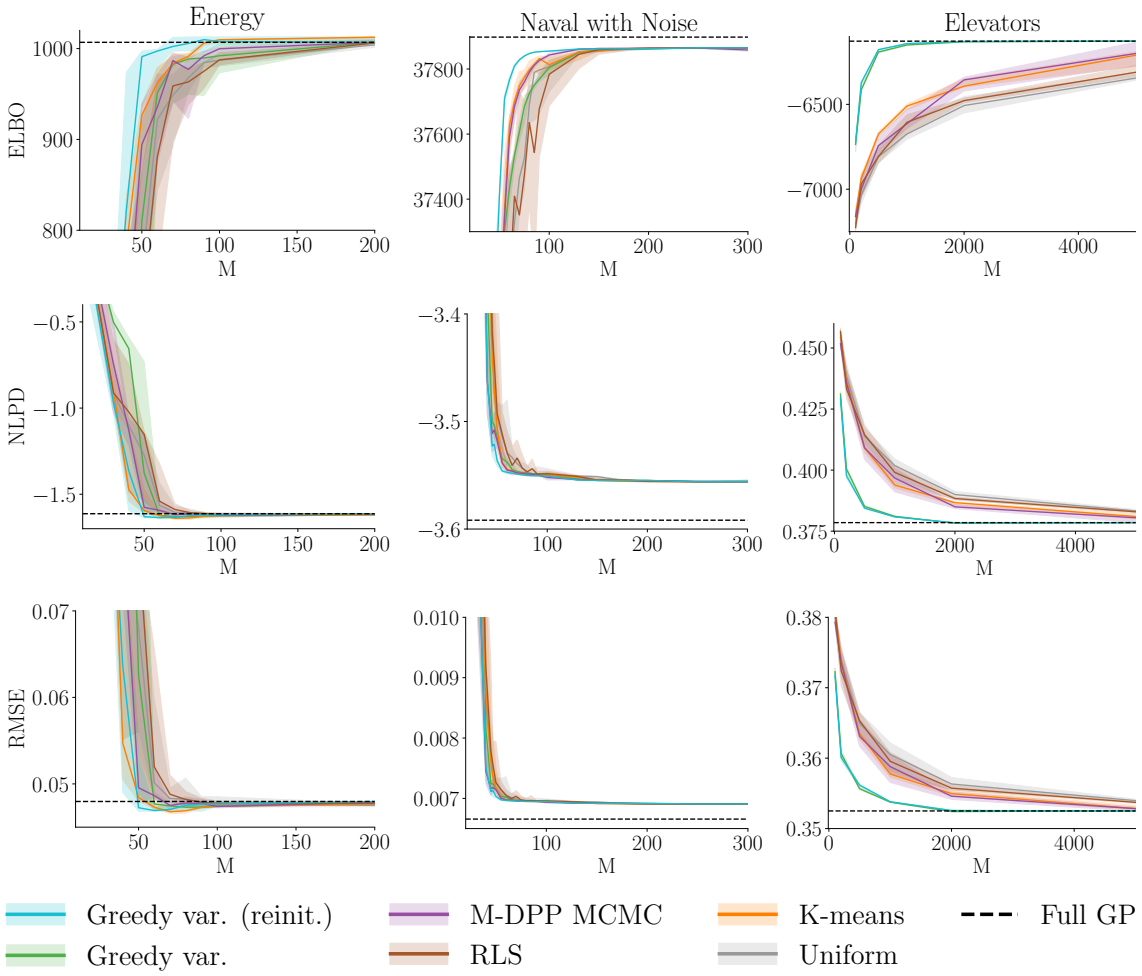

Figure 6: Performance of various methods for selecting inducing inputs on 3 data sets (each column corresponds to one data set) with model parameters learned via L-BFGS. The top row shows the evidence lower bounds, the middle row the per datapoint negative log predictive density on held-out test points and the bottom row the root mean square error on test data. The solid lines show the median of 10 initializations using the given method for selecting inducing inputs, while the shaded region represents the $20 - 80\%$. The dashed black line shows the performance of the exact GP regressor on the given data set.

significantly sped up. In addition, we measure computational cost through the number of function evaluations. This does not take into account the additional cost of computing the gradients for the inducing inputs, which make up the bulk of parameters that are to be optimized. As the amount of computation needed to reinitialization the inducing points is

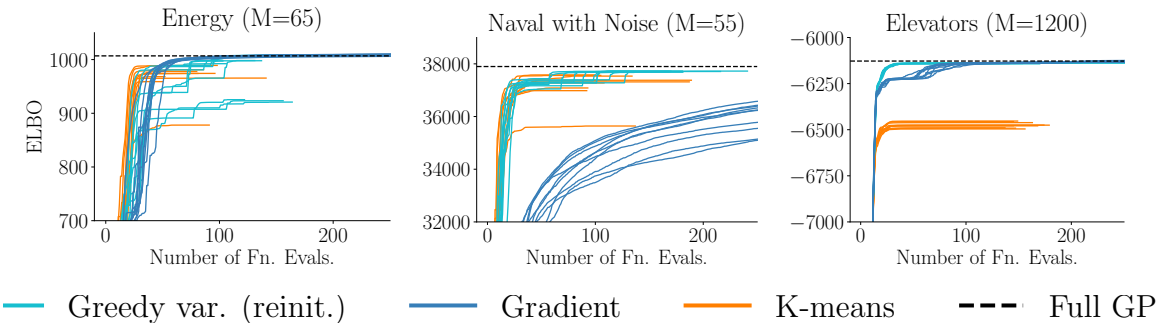

Figure 7: ELBO against the number of function evaluations called during a run of L-BFGS. "Gradient" uses gradient-based optimization for the inducing inputs as well as the hyperparameters, with the inducing inputs initialized using greedy variance selection.

comparable to the computation required in as single iteration of gradient descent, Fig. 7 likely understates the computational savings of the reinitialization method.

7.2.3. Recommendation for Inducing Input Selection

The main conclusion from our empirical results is that well-chosen inducing inputs that are reinitialized during hyperparameter optimization give highly accurate variational approximations to the results of exact GPs. While performing gradient-based optimization of the inducing inputs may lead to improved performance in settings that are constrained to be very sparse, in some instances it is not worth the additional effort. We provide a GPflow-based (Matthews et al., 2017) implementation of the initialization methods and experiments that builds on other open source software (Coelho, 2017; Virtanen et al., 2020), available at `https://github.com/markvdw/RobustGP`.

It is important to note that we only considered data sets where sparse approximations were practically possible. The "kin40k" UCI data set is a notable example where a squared exponential GP regression model with learned hyperparameters could not accurately be approximated, due to a lengthscale that continuously decreased with increasing M. Given the underfitting and significant hyperparameter bias (Bauer et al., 2016), one can question whether variational approximations are appropriate. In cases where the covariates are less heavily correlated under the prior, conjugate gradient approaches (Gibbs and Mackay, 1997; Davies, 2015; Gardner et al., 2018) may be better. We choose to not make a recommendation for how to choose inducing variables in cases where the variational approximation is poor.

8. Conclusions

We provide guarantees on the quality of variational sparse Gaussian process regression when many fewer inducing variables are used than data points. We also consider lower bounds on the number of inducing variables needed in order to ensure that the KL-divergence between

the approximate posterior and the full posterior is not large. These bounds provide insight into the number of inducing points that should be used for a variety of tasks, as well as suggest the sorts of problems to which sparse variational inference is well-suited. We also include an empirical results comparing the efficacy of different methods for selecting inducing inputs, which is of practical importance to the Gaussian process community. We believe that there is a great deal of interesting future research to be done on the role of sparsity in variational Gaussian process inference; both in refining the bounds given in this work and in better understanding non-conjugate inference schemes, such as those developed in Hensman et al. (2015).

Acknowledgments

We would particularly like to thank Guillaume Gautier for pointing out an error in the exact k-DPP sampling algorithm cited in an earlier version of this work, and for guiding us through recent work on sampling k-DPPs that led to an amended proof. MvdW would additionally like to thank James Hensman for his guidance, and PROWLER.io for providing an excellent research environment while this work was developed.

Appendix A. Proof of Bound on Mean and Variance of One-dimensional Marginal Distributions

Proposition 1 *Suppose* $2\mathrm{KL}[Q||P] \leq \gamma \leq \frac{1}{5}$. *For any* $x^\star \in \mathcal{X}$, *let* μ_1 *denote the posterior mean of the variational approximation at* $x^\star$ *and* μ_2 *denote the mean of the exact posterior at* $x^\star$. *Similarly, let* σ_1^2, σ_2^2 *denote the variances of the approximate and exact posteriors at* $x^\star$. *Then,*

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

Proof By the chain rule of KL-divergence, we have

$$2\mathrm{KL}[q(f(x^*))||p(f(x^*)|\mathcal{D})] \leq 2\mathrm{KL}[Q||P] \leq \gamma \tag{30}$$

for any $x^\star \in \mathcal{X}$.

For any $x^\star \in \mathcal{X}$, the KL-divergence on the right hand side of Eq. (30) is a KL-divergence between one-dimensional Gaussian distributions, and has the form,

$$\gamma \geq 2\mathrm{KL}[q(f(x^*))||p(f(x^*)|\mathcal{D})] = \frac{\sigma_1^2}{\sigma_2^2} - 1 - \log\frac{\sigma_1^2}{\sigma_2^2} + \frac{(\mu_1-\mu_2)^2}{\sigma_2^2} \geq \frac{\sigma_1^2}{\sigma_2^2} - 1 - \log\frac{\sigma_1^2}{\sigma_2^2}. \tag{31}$$

Define $r = \sigma_1^2/\sigma_2^2$, so Eq. (31) becomes $\gamma \geq r - 1 - \log r$. For $\gamma < \frac{1}{5}$, we have $r - \log(r) < 1.2$, so $r \in [.493, 1.78]$. For r in this range, we have, $\gamma \geq r - 1 - \log r \geq (r-1)^2/3$. Solving, for r, we obtain the bound,

$$\left|1 - \frac{\sigma_1^2}{\sigma_2^2}\right| \leq \sqrt{3\gamma}. \tag{32}$$

We now turn to the proof of the bound relating μ_1 and μ_2. From Eq. (31) and because $r - 1 - \log r > 0$ for $r > 0$, $\frac{(\mu_1-\mu_2)^2}{\sigma_2^2} \leq \gamma$. Rearranging, $|\mu_1 - \mu_2| \leq \sigma_2\sqrt{\gamma}$. The final bound on the mean follows from Eq. (32), which implies that,

$$\sigma_2 \leq \frac{\sigma_1}{\sqrt{1-\sqrt{3\gamma}}}.$$

■

Appendix B. Proofs of A-Posteriori Bounds

In this section, we restate and prove the upper bound on the marginal likelihood given in Titsias (2014).

Lemma 2 (Titsias, 2014) *For any $y \in \mathbb{R}^N, X \in \mathcal{X}^N$, and set of M inducing variables, U,*

$$\log p(y) \leq \mathcal{U}_1 \leq \mathcal{U}_2$$

where

$$\mathcal{U}_1 \coloneqq -\frac{1}{2}\log\det(\mathrm{Q_{ff}} + \sigma^2\mathrm{I}) - \frac{1}{2}y^\intercal(\mathrm{Q_{ff}} + \|\mathrm{K_{ff}} - \mathrm{Q_{ff}}\|_{\mathrm{op}}\mathrm{I} + \sigma^2\mathrm{I})^{-1}y - \frac{N}{2}\log 2\pi,$$

and

$$\mathcal{U}_2 \coloneqq -\frac{1}{2}\log\det(\mathrm{Q_{ff}} + \sigma^2\mathrm{I}) - \frac{1}{2}y^\intercal(\mathrm{Q_{ff}} + t\mathrm{I} + \sigma^2\mathrm{I})^{-1}y - \frac{N}{2}\log 2\pi. \tag{12}$$

This result relies on several properties of symmetric positive semi-definite (SPSD) matrices, which we state in Proposition 35.

Proposition 35 (Horn and Johnson (1990), Corollary 7.7.4) *Let $\succ$ denote the partial order on SPSD matrices induced by $A \succ B \iff A - B$ is SPSD. Then if $A \succ B$ are $N \times N$ SPSD matrices,*

1. $\det(A) \geq \det(B)$,
2. *If A^{-1}, B^{-1} exist, then $A^{-1} \prec B^{-1}$.*
3. *If $\lambda_1(A) \geq \ldots \geq \lambda_N(A)$, $\lambda_1(B) \geq \ldots \geq \lambda_N(B)$ denote the eigenvalues of A and B respectively, $\lambda_i(A) \geq \lambda_i(B)$ for all $1 \leq i \leq N$.*

We also use that $\mathrm{Q_{ff}} \prec \mathrm{K_{ff}}$, which follows from properties of Schur complements of PSD matrices (Gallier, 2010, Proposition 2.1).

Proof of Lemma 2 This proof follows that of Titsias (2014). Recall Eq. (6),

$$\begin{aligned}
\log p(y) &= -\frac{1}{2}\log\det(\mathrm{K_{ff}} + \sigma^2\mathrm{I}) - \frac{1}{2}y^\intercal(\mathrm{K_{ff}} + \sigma^2\mathrm{I})^{-1}y - \frac{N}{2}\log 2\pi \\
&\leq -\frac{1}{2}\log\det(\mathrm{Q_{ff}} + \sigma^2\mathrm{I}) - \frac{1}{2}y^\intercal(\mathrm{K_{ff}} + \sigma^2\mathrm{I})^{-1}y - \frac{N}{2}\log 2\pi \\
&\leq -\frac{1}{2}\log\det(\mathrm{Q_{ff}} + \sigma^2\mathrm{I}) - \frac{1}{2}y^\intercal(\mathrm{Q_{ff}} + \|\mathrm{K_{ff}} - \mathrm{Q_{ff}}\|_{\mathrm{op}}\mathrm{I} + \sigma^2\mathrm{I})^{-1}y - \frac{N}{2}\log 2\pi \\
&= \mathcal{U}_1.
\end{aligned} \tag{33}$$

The first inequality uses $\mathrm{K_{ff}} + \sigma^2\mathrm{I} \succ \mathrm{Q_{ff}} + +\sigma^2\mathrm{I}$, which implies $\log\det(\mathrm{K_{ff}} + \sigma^2\mathrm{I}) \geq \log\det(\mathrm{Q_{ff}} + \sigma^2\mathrm{I})$. The second inequality uses that $\mathrm{K_{ff}} \prec \mathrm{Q_{ff}} + \|\mathrm{K_{ff}} - \mathrm{Q_{ff}}\|_{\mathrm{op}}\mathrm{I}$ and the second part of Proposition 35.

In problems where sparse GP regression is applied, computing the largest eigenvalue of $\mathrm{K_{ff}} - \mathrm{Q_{ff}}$ is computationally prohibitive. However, we can use the upper bound $\|\mathrm{K_{ff}} - \mathrm{Q_{ff}}\|_{\mathrm{op}} \leq \mathrm{tr}(\mathrm{K_{ff}} - \mathrm{Q_{ff}})$, yielding

$$\mathcal{U}_1 \leq -\frac{1}{2}\log\det(\mathrm{Q_{ff}} + \sigma^2\mathrm{I}) - \frac{1}{2}y^\intercal(\mathrm{Q_{ff}} + \mathrm{tr}(\mathrm{K_{ff}} - \mathrm{Q_{ff}})\mathrm{I} + \sigma^2\mathrm{I})^{-1}y - \frac{N}{2}\log 2\pi = \mathcal{U}_2.$$

The bound $\mathcal{U}_2$ can be computed in time $\mathcal{O}(NM^2)$ with memory $\mathcal{O}(NM)$ in much the same way as the ELBO is computed, as it only depends on the low-rank matrix $\mathrm{Q_{ff}}$ and the diagonal entries of $\mathrm{K_{ff}}$. ■

Appendix C. Proofs for Results Leading to Upper bounds on the KL-divergence

In this appendix, we restate and provide proofs for the results in Section 4.

Lemma 3 *For any $y \in \mathbb{R}^N, X \in \mathcal{X}^N$, and any $Z \in \mathcal{X}^M$*

$$\mathrm{KL}[Q||P] \leq \mathcal{U}_1 - \mathcal{L} \leq \frac{1}{2\sigma^2}\left(t + \frac{\zeta\|y\|_2^2}{\zeta + \sigma^2}\right) \leq \frac{1}{2\sigma^2}\left(t + \frac{t\|y\|_2^2}{t + \sigma^2}\right),$$

with $t = \mathrm{tr}(\mathrm{K_{ff}} - \mathrm{Q_{ff}})$ and $\zeta = \|\mathrm{K_{ff}} - \mathrm{Q_{ff}}\|_{\mathrm{op}}$.

Proof We apply the matrix identity $(A+B)^{-1} = A^{-1} - A^{-1}B(A+B)^{-1}$ to the expression

$$\mathcal{U}_1 - \mathcal{L} = \frac{t}{2\sigma^2} + \frac{1}{2}y^\intercal\big((\mathrm{Q_{ff}} + \sigma^2\mathrm{I})^{-1} - (\mathrm{Q_{ff}} + \zeta\mathrm{I} + \sigma^2\mathrm{I})^{-1}\big)y,$$

with $A = \mathrm{Q_{ff}} + \sigma^2\mathrm{I}$ and $B = \zeta\mathrm{I}$. This gives

$$\begin{aligned}\mathcal{U}_1 - \mathcal{L} &= \frac{t}{2\sigma^2} + \frac{\zeta}{2}y^\intercal\big((\mathrm{Q_{ff}} + \sigma^2\mathrm{I})^{-1}(\mathrm{Q_{ff}} + (\zeta + \sigma^2)\mathrm{I})^{-1}\big)y \\ &= \frac{t}{2\sigma^2} + \frac{\zeta}{2}y^\intercal(\mathrm{Q_{ff}^2} + (\zeta + \sigma^2)\mathrm{Q_{ff}} + \sigma^2(\zeta + \sigma^2)\mathrm{I}))^{-1}y.\end{aligned}$$

The matrix $\mathrm{Q_{ff}^2} + (\zeta + \sigma^2)\mathrm{Q_{ff}}$ is SPSD, as it is the product of SPSD matrices that commute. This implies that the eigenvalues of $\mathrm{Q_{ff}^2} + (\zeta + \sigma^2)\mathrm{Q_{ff}} + \sigma^2(\zeta + \sigma^2)\mathrm{I}$ are bounded below by $\sigma^2(\zeta + \sigma^2)$. As the eigenvalues of the inverse of a SPSD matrix are the inverse of the eigenvalues of the original matrix, the largest eigenvalue of $(\mathrm{Q_{ff}^2} + (\zeta + \sigma^2)\mathrm{Q_{ff}} + \sigma^2(\zeta + \sigma^2)\mathrm{I}))^{-1}$ is bounded above by $(\sigma^2(\zeta + \sigma^2))^{-1}$. Therefore,

$$\mathrm{KL}[Q||P] \leq \mathcal{U}_1 - \mathcal{L} \leq \frac{t}{2\sigma^2} + \frac{\zeta\|y\|_2^2}{2\sigma^2(\zeta + \sigma^2)}. \tag{34}$$

This proves the second inequality in Lemma 3. The same argument using $\mathcal{U}_2$ in place of $\mathcal{U}_1$ yields,

$$\mathrm{KL}[Q||P] \leq \frac{t}{2\sigma^2} + \frac{t\|y\|_2^2}{2\sigma^2(t+\sigma^2)}.$$

■

Lemma 4 *Suppose* $\mathbf{y}|\mathbf{X},\mathbf{Z} \sim \mathcal{N}(0, \mathrm{K_{ff}} + \sigma^2\mathrm{I})$. *For any* $X \in \mathcal{X}^N$ *and* $Z \in \mathcal{X}^M$,

$$t/(2\sigma^2) \leq \mathbb{E}[\mathrm{KL}[Q||P] \,|\, \mathbf{Z} = Z, \mathbf{X} = X] \leq t/\sigma^2$$

where $t = \mathrm{tr}(\mathrm{K_{ff}} - \mathrm{Q_{ff}})$ *and* $\mathrm{K_{ff}}$ *and* $\mathrm{Q_{ff}}$ *are defined with respect to this* X, Z *as in Section 2.*

We have already proven the lower bound in the main body. In order to prove the upper bound in Lemma 4, we use a Hölder-type inequality, Tao (2012, Exercise 1.3.26).

Proposition 36 *For any matrix, let* $\|A\|_p := (\sum_i |\sigma_i(A)|^p)^{1/p}$ *if* p *is finite and* $\|A\|_\infty = \max_i |\sigma_i(A)|$, *where* $\sigma_i(A)$ *are singular values of* A. *Then for* $A, B \in \mathbb{R}^{n\times n}$ *and any* $1 \leq p, q \leq \infty$ *such that* $\frac{1}{p} + \frac{1}{q} = 1$,

$$\mathrm{tr}(AB) \leq \|A\|_p \|B\|_q.$$

In particular, if A and B are SPSD (so that the singular values agree with the eigenvalues), taking $p = 1$, $q = \infty$,

$$\mathrm{tr}(AB^\intercal) \leq \mathrm{tr}(A)\|B\|_{op}.$$

where $\|B\|_{\mathrm{op}}$ is the largest eigenvalue of B.

Proof of Lemma 4

For the upper bound, it remains to bound

$$\mathrm{kl}(\mathrm{K_{ff}}, \mathrm{Q_{ff}}) \coloneqq \mathrm{KL}\big[\mathcal{N}(0, \mathrm{K_{ff}} + \sigma^2\mathrm{I})||\mathcal{N}(0, \mathrm{Q_{ff}} + \sigma^2\mathrm{I})\big].$$

$$\begin{aligned}
\mathrm{kl}(\mathrm{K_{ff}}, \mathrm{Q_{ff}}) &= \frac{1}{2}\Big(\log\det(\mathrm{Q_{ff}} + \sigma^2\mathrm{I}) - \log\det(\mathrm{K_{ff}} + \sigma^2\mathrm{I}) - N + \mathrm{tr}((\mathrm{Q_{ff}} + \sigma^2\mathrm{I})^{-1}(\mathrm{K_{ff}} + \sigma^2\mathrm{I}))\Big) \\
&\leq \frac{1}{2}\Big(- N + \mathrm{tr}((\mathrm{Q_{ff}} + \sigma^2\mathrm{I})^{-1}(\mathbf{K_{ff}} + \sigma^2\mathrm{I})\Big) \\
&= \frac{1}{2}\Big(- N + \mathrm{tr}((\mathrm{Q_{ff}} + \sigma^2\mathrm{I})^{-1}((\mathrm{Q_{ff}} + \sigma^2 I) + (\mathrm{K_{ff}} - \mathrm{Q_{ff}}))\Big) \\
&= \frac{1}{2}\mathrm{tr}((\mathrm{Q_{ff}} + \sigma^2\mathrm{I})^{-1}(\mathrm{K_{ff}} - \mathrm{Q_{ff}})). \qquad (35)
\end{aligned}$$

The inequality uses that $\mathrm{Q_{ff}} + \sigma^2\mathrm{I} \prec \mathrm{K_{ff}} + \sigma^2\mathrm{I}$, so $\det(\mathrm{Q_{ff}} + \sigma^2\mathrm{I}) \leq \det(\mathrm{K_{ff}} + \sigma^2\mathrm{I})$ by Proposition 35. We can now apply Proposition 36 with $p = 1, q = \infty$ to Eq. (35) giving,

$$\frac{1}{2}\mathrm{tr}((\mathrm{Q_{ff}} + \sigma^2\mathrm{I})^{-1}(\mathrm{K_{ff}} - \mathrm{Q_{ff}})) \leq \frac{t}{2}\|(\mathrm{Q_{ff}} + \sigma^2\mathrm{I})^{-1}\|_{\mathrm{op}} \leq \frac{t}{2\sigma^2}.$$

Using this bound in Eq. (35) and combining with Eq. (15) completes the proof of the upper bound. ■

Appendix D. Derivations of bounds for specific kernels and covariate distributions

In this appendix, we restate and provide proofs for the results in Section 5.

D.1. Bounds for Univariate Gaussian distributions and Squared Exponential Kernel

Corollary 19 *Let k be a squared exponential kernel. Suppose that N real-valued (one-dimensional) covariates are observed, with identical Gaussian marginal distributions. Suppose the conditions of Theorem 13 are satisfied for some $R > 0$. Fix any $\gamma \in (0, 1]$. Then there exists an $M = \mathcal{O}(\log(N^3/\gamma))$ and an $\epsilon = \Theta(\gamma/N^2)$ such if inducing points are distributed according to an ϵ-approximate M-DPP with kernel matrix $\mathrm{K_{ff}}$,*

$$\mathbb{E}[\mathrm{KL}[Q||P]] \leq \gamma.$$

Similarly, for any $\delta \in (0, 1/32)$ using the ridge leverage algorithm of Musco and Musco (2017) and choosing S appropriately, with probability $1-5\delta$, $\mathbf{M} = \mathcal{O}\Big(\log \frac{N^2}{\delta^2\gamma} \log \frac{\log(N^2/\delta^2\gamma)}{\delta}\Big)$ and

$$\mathrm{KL}[Q||P] \leq \gamma.$$

The implicit constants depend on the kernel hyperparameters, the likelihood variance, the variance of the covariate distribution and R.

Proof of Corollary 19 Using Eq. (24) and applying the geometric series formula,

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

We can use this equation in Theorem 13 (a similar result could be obtained using Theorem 14) yielding,

$$\mathbb{E}[\mathrm{KL}[Q||P]] \leq \left(\sqrt{\frac{2a}{A}} \frac{(M+1)NB^M}{2\sigma^2(1-B)} + \frac{Nv\epsilon}{\sigma^2}\right)\left(1 + \frac{RN}{\sigma^2}\right).$$

Choose $\epsilon = \frac{\gamma\sigma^2}{2Nv(1+RN/\sigma^2)} = \Theta(\gamma/N^2)$. By Lemma 9, an M-DPP can be sampled to this level of accuracy using not more than $\mathcal{O}(NM(\log \frac{N^2}{\gamma\delta}))$ iterations of MCMC, making the computational cost of selecting inducing inputs $\mathcal{O}(NM^3(\log \frac{N^2}{\gamma\delta}))$. We may assume that $M < N$, otherwise by choosing $Z = X$ the KL-divergence is zero and nothing more needs to be shown. Then,

$$\mathbb{E}[\mathrm{KL}[Q||P]] \leq \sqrt{\frac{2a}{A}} \frac{N^2B^M}{2\sigma^2(1-B)}\left(1 + \frac{RN}{\sigma^2}\right) + \frac{\gamma}{2}$$

Take $M = \log_B \sqrt{\frac{A}{2a}} \frac{\gamma\delta\sigma^2(1-B)}{N^2(1+RN/\sigma^2)} = \mathcal{O}(\log(N^3/\gamma\delta))$, then

$$\mathbb{E}[\mathrm{KL}[Q||P]] \leq \gamma.$$

In the case of ridge leverage score initializations, from Theorem 17 we have with probability $1-5\delta$,

$$\mathrm{KL}[Q||P] \leq \sqrt{\frac{2a}{A}}\frac{N^2 B^S}{S(1-B)\delta^2\sigma^2}(1+R/\sigma^2)$$

and $\mathbf{M} \leq S\log\frac{S}{\delta}$. Choose $S = \log_B \sqrt{\frac{A}{2a}}\frac{\gamma\sigma^2(1-B)\delta^2}{N^2(1+R/\sigma^2)}$. Then on the event where these bounds hold, $\mathrm{KL}[Q||P] \leq \frac{\gamma}{S} \leq \gamma$ and $\mathbf{M} \leq S\log\frac{S}{\delta} = \mathcal{O}\Big(\log\frac{N^2}{\delta^2\gamma}\log\frac{\log(N^2/\delta^2\gamma)}{\delta}\Big)$. If we allow the implicit constant to depend on δ and γ as well this becomes $\mathcal{O}(\log N\log\log N)$. ■

D.2. Bounds for Multivariate Gaussian distributions and Squared Exponential Kernel

Proposition 21 *For a SE-kernel and Gaussian distributed covariates in $\mathbb{R}^D$, for $M \geq \frac{1}{\alpha}D^D + D - 1$, $\sum_{m=M+1}^{\infty}\lambda_m = \mathcal{O}(M\exp(-\alpha M^{1/D}))$, where $\alpha = -\log B > 0$ and the implicit constant depends on the dimension of the covariates, the kernel parameters and the covariance matrix of the covariate distribution.*

Proof of Proposition 21 The proof of this proposition is nearly identical to an argument in Seeger et al. (2008). Consider the upper bound,

$$\lambda_{M+D-1} \leq \left(\frac{2a}{A}\right)^{\frac{D}{2}} B^{M^{1/D}}.$$

Define $\tilde{M} = M - D + 1$, then for $M > D - 1$,

$$\begin{aligned}\sum_{m=M+1}^{\infty}\lambda_m &\leq \left(\frac{2a}{A}\right)^{\frac{D}{2}}\sum_{m=\widetilde{M}+1}^{\infty}B^{m^{1/D}} \leq \left(\frac{2a}{A}\right)^{\frac{D}{2}}\int_{s=\widetilde{M}}^{\infty}B^{s^{1/D}}\,\mathrm{d}s \\ &= \left(\frac{2a}{A}\right)^{\frac{D}{2}} D\alpha^{-D}\int_{t=\alpha\widetilde{M}^{1/D}}^{\infty}\exp(-t)t^{D-1}\,\mathrm{d}t \\ &= \left(\frac{2a}{A}\right)^{\frac{D}{2}} D\alpha^{-D}\Gamma(D,\alpha(M-D+1)^{1/D})\end{aligned}$$

where in the second to last line we make the substitution $t = \alpha s^{1/D}$ and in the final line we recognized the integral as an incomplete Γ-function.

From Gradshteyn and Ryzhik (2014, 8.352) for integer D and $r > 0$,

$$\Gamma(D,r) = (D-1)!e^{-r}\sum_{k=0}^{D-1}\frac{r^k}{k!}.$$

For fixed D and r large (which is satisfied by the condition $M \geq \frac{1}{\alpha}D^D + D - 1$), we have that the final term in the above sum is the largest, so that

$$\Gamma(D,r) \leq D!e^{-r}\frac{r^{D-1}}{(D-1)!} = De^{-r}y^{D-1}$$

Using this bound, we arrive at

$$\sum_{m=M+1}^{\infty} \lambda_m \leq \left(\frac{2a}{A}\right)^{\frac{D}{2}} D^2 \alpha^{-D} \exp(-\alpha(M-D+1)^{1/D})(\alpha(M-D+1)^{1/D})^{D-1}$$
$$\leq \left(\frac{2a}{A}\right)^{\frac{D}{2}} \frac{D^2(M-D+1)}{\alpha} \exp(-\alpha(M-D)^{1/D}) = \mathcal{O}(M\exp(-\alpha M^{1/D})).$$

■

Corollary 22 *Let k be a SE-ARD kernel in D-dimensions. Suppose that N D-dimensional covariates are observed, so that each covariate has an identical multivariate Gaussian distribution, and that the distribution of training outputs satisfies $\mathbb{E}\left[\|\mathbf{y}\|^2 \,\middle|\, \mathbf{X}\right] \leq RN$. Fix any $\gamma \in (0,1]$. Then there exists an $M = \mathcal{O}((\log N/\gamma)^D)$ and an $\epsilon = \mathcal{O}(N^2/\gamma)$ such if inducing inputs are distributed according to an ϵ-approximate M-DPP with kernel matrix $\mathrm{K_{ff}}$,*

$$\mathbb{E}[\mathrm{KL}[Q||P]] \leq \gamma.$$

The implicit constant depends on the kernel hyperparameters, the variance matrix of the covariate distribution, D and R. With the same assumptions but applying the RLS algorithm of Musco and Musco (2017) to selecting inducing inputs, for any $\delta \in (0, 1/32)$ there exists a choice of S such that with probability $1-5\delta$, $\mathbf{M} = \mathcal{O}\left(\left(\log \frac{N^2}{\delta\gamma}\right)^D (\log\log \frac{N^2}{\delta\gamma} + \log(1/\delta)\right)$ and

$$\mathrm{KL}[Q||P] \leq \gamma.$$

Proof Corollary 22 is a consequence of Theorem 13 and Proposition 21. In the case of the M-DPP, we take $\epsilon = \frac{\gamma\sigma^2}{2Nv(1+RN/\sigma^2)} = \Theta(\gamma/N^2)$ as in the proof of Corollary 19. It then remains to choose M so that

$$\left(\frac{N^2}{2\sigma^2(1-B)}\left(1+\frac{RN}{\sigma^2}\right)\sum_{m=M+1}^{\infty}\lambda_m\right) \leq \gamma/2.$$

From Proposition 21, there exists an $M = \mathcal{O}((\log \frac{N^3}{\gamma})^D)$ that satisfies this criteria. In the case of ridge leverage scores, it is sufficient to choose $S = \mathcal{O}\left(\left(\log \frac{N^2}{\delta\gamma}\right)^D\right)$, which means that with probability at least $1-5\delta$, $\mathbf{M} = \mathcal{O}\left(\left(\log \frac{N^2}{\delta\gamma}\right)^D (\log\log \frac{N^2}{\delta\gamma} + \log(1/\delta))\right)$. ■

D.3. Conditions for Widom's Theorem

Widom's Theorem (Widom, 1963), states that for stationary kernels on compact subsets of Euclidean space, the eigenvalues of the operator $\mathcal{K}$ are closely linked to the decay of the spectral density of the kernel function. The theorem applies to any compactly supported covariate distribution with Lebesgue density and stationary kernel with spectral density satisfying the following three conditions:

1. For all $i \in \{1, \cdots, D\}$, fixing all $\omega^{(j)}, j \neq i$, there exists an $\omega_0^{(i)} \in \mathbb{R}$ such that $s(\omega)$ is monotonically increasing as a function of $\omega^{(i)}$ for all $\omega^{(i)} < \omega_0^{(i)}$ and is monotonically decreasing as a function of $\omega^{(i)}$ for $\omega^{(i)} \geq \omega_0^{(i)}$.

2. Let $\{\xi_i\}_{i=1}^\infty, \{\eta_i\}_{i=1}^\infty$, be sequences in $\mathbb{R}^D$ such that $\lim_{i\to\infty} \frac{\|\eta_i - \xi_i\|}{\|\eta_i\|} = 0$ and $\lim_{i\to\infty} \|\xi_i\| = \infty$, then $\lim_{i\to\infty} \frac{|s(\xi_i)|}{|s(\eta_i)|} = 1$.

3. Let $\{\xi_i\}_{i=1}^\infty, \{\eta_i\}_{i=1}^\infty$, be sequences in $\mathbb{R}^D$ such that $\lim_{i\to\infty} \|\xi_i\|, \|\eta_i\| = \infty$ and $\lim_{i\to\infty} \frac{\|\xi_i\|}{\|\eta_i\|} = 0$, then $\lim_{i\to\infty} \frac{|s(\xi_i)|}{|s(\eta_i)|} = 0$.

If the kernel and spectral density satisfy these conditions, the number of eigenvalues of $\mathcal{K}$ greater than ϵ is asymptotic (as $\epsilon \to 0$) to the volume of the collection of points in $\mathbb{R}^d \times \mathbb{R}^d$ such that $p(x)s(\omega) > \epsilon$. A precise statement of the result can be found in Widom (1963), and more discussion of the result is given in Seeger et al. (2008). Because of the second condition, Widom's theorem cannot be applied to kernels with rapidly decaying spectral densities, such as the SE-kernel (though more stationary kernels are analyzed in Widom (1964) for uniformly distributed covariates).

Appendix E. Lower bounds on the number of features

In this appendix, we restate and prove the results stated in Section 6.

E.1. General Lower Bound on KL-divergence

Lemma 27 *Given a kernel k, likelihood model with variance σ^2 and random covariates $\mathbf{X}$. Then,*

$$\min_{Z\in\mathcal{X}^M} \min_{y\in\mathbb{R}^N} \mathrm{KL}[Q||P] \geq \frac{1}{2} \sum_{m=M+1}^{N} \frac{\tilde{\boldsymbol{\lambda}}_m}{\sigma^2} - \log\left(1 + \frac{\tilde{\boldsymbol{\lambda}}_m}{\sigma^2}\right)$$

where $\tilde{\boldsymbol{\lambda}}_m$ denotes the m^{th} largest eigenvalue of the matrix $\mathbf{K_{ff}}$ determined by the covariates and kernel.

Proof of Lemma 27 Define $\mathcal{L}(y, Z)$ to be the evidence lower bound assuming y are the observations and inducing points are placed at locations Z. Then for any $y \in \mathbb{R}^N$ and $Z \in \mathcal{X}^M$,

$$\begin{aligned}\log p(y) - \mathcal{L}(y, Z) &\geq \log p(0) - \mathcal{L}(0, Z) \\ &= \frac{1}{2}\left(\frac{1}{\sigma^2}\mathrm{tr}(\mathbf{K_{ff}} - \mathbf{Q_{ff}}) - \log\frac{\det(\mathbf{K_{ff}} + \sigma^2\mathrm{I})}{\det(\mathbf{Q_{ff}} + \sigma^2\mathrm{I})}\right). \qquad (36)\end{aligned}$$

The inequality uses that the only term in $\log p(y) - \mathcal{L}(y, Z)$ that depends on y is the quadratic $\frac{1}{2}y^\intercal\big((\mathbf{Q_{ff}} + \sigma^2\mathrm{I})^{-1} - (\mathbf{K_{ff}} + \sigma^2\mathrm{I})^{-1}\big)y \geq 0$ since $(\mathbf{Q_{ff}} + \sigma^2\mathrm{I})^{-1} \succ (\mathbf{K_{ff}} + \sigma^2\mathrm{I})^{-1}$.

We can rewrite Eq. (36) as a sum over the eigenvalues of $\mathbf{K_{ff}}$ and $\mathbf{Q_{ff}}$, which we denote by $\tilde{\boldsymbol{\lambda}}_m$ and $\boldsymbol{\psi}_m$ respectively. Also, since $\mathbf{K_{ff}} \succ \mathbf{Q_{ff}}$, $\tilde{\boldsymbol{\lambda}}_m \geq \boldsymbol{\psi}_m$ for all $1 \leq m \leq N$

(Proposition 35). This yields,

$$\begin{aligned}\log p(0) - \mathcal{L}(0, Z) &= \frac{1}{2}\sum_{m=1}^{N} \frac{\tilde{\boldsymbol{\lambda}}_m - \boldsymbol{\psi}_m}{\sigma^2} - \log\left(1 + \frac{\tilde{\boldsymbol{\lambda}}_m - \boldsymbol{\psi}_m}{\boldsymbol{\psi}_m + \sigma^2}\right) \\ &\geq \frac{1}{2}\sum_{m=1}^{N} \frac{\tilde{\boldsymbol{\lambda}}_m - \boldsymbol{\psi}_m}{\sigma^2} - \log\left(1 + \frac{\tilde{\boldsymbol{\lambda}}_m - \boldsymbol{\psi}_m}{\sigma^2}\right) \qquad (*) \\ &= \frac{1}{2}\left(\sum_{m=1}^{M} \frac{\tilde{\boldsymbol{\lambda}}_m - \boldsymbol{\psi}_m}{\sigma^2} - \log\left(1 + \frac{\tilde{\boldsymbol{\lambda}}_m - \boldsymbol{\psi}_m}{\sigma^2}\right) + \sum_{m=M+1}^{N} \frac{\tilde{\boldsymbol{\lambda}}_m}{\sigma^2} - \log\left(1 + \frac{\tilde{\boldsymbol{\lambda}}_m}{\sigma^2}\right)\right).\end{aligned}$$

In the final line, we use that $\mathbf{Q_{ff}}$ is at most rank M, so that $\boldsymbol{\psi}_m = 0$ for all $m > M$. It follows from the inequality $\log(1 + a) \leq a$ for $a \geq 0$ that each term in the first sum is non-negative. Hence,

$$\log p(y) - \mathcal{L}(y, Z) \geq \frac{1}{2}\sum_{m=M+1}^{N} \frac{\tilde{\boldsymbol{\lambda}}_m}{\sigma^2} - \log\left(1 + \frac{\tilde{\boldsymbol{\lambda}}_m}{\sigma^2}\right). \tag{37}$$

■

E.2. Lower Bound on Eigenvalues of Multivariate Gaussian Inputs and Squared Exponential Kernel

Proposition 30 *Suppose k is an isotropic SE-kernel in D dimensions with lengthscale ℓ and variance v. Suppose the training covariates are independently identically distributed according to an isotropic Gaussian measure, μ, on $\mathbb{R}^D$ with covariance matrix $\beta^2\mathrm{I}$. For any $r \in \mathbb{N}$, we have*

$$\lambda_r \geq \left(\frac{2a}{A}\right)^{D/2} B^{Dr^{1/D}}.$$

where λ_r denotes the r^{th} largest eigenvalue of the operator $\mathcal{K} : L^2(\mathbb{R}^D, \mu) \to L^2(\mathbb{R}^D, \mu)$ defined by $(\mathcal{K}g)(x') = \int g(x)k(x, x')p(x)dx$ with $p(x)$ the density of the multivariate Gaussian at x.

Proof Recall from Section 5.1.1 that the eigenvalues of this operator are of the form,

$$\lambda_r = \left(\frac{2a}{A}\right)^{D/2} B^s$$

where the number of times each eigenvalue is repeated is equal to the number of ways to write s as a sum of D non-negative integers, where the order of the summands matters. This is equal to $\binom{s+D-1}{D-1}$. The number of eigenvalues greater than $(2a/A)^{D/2}B^s$ is therefore,

$$\sum_{t=1}^{s} \binom{t + D - 1}{D - 1} = \binom{s + D}{D}.$$

The equality follows from observing that the right hand side is equal to the number of way to write s as a sum of $D+1$ non-negative integers. For each of these representations, the first D integers sum to some $t \leq s$, and once these are fixed there is a unique choice for the final integer. This is equivalent to the left hand side. We therefore conclude $\lambda_{\binom{s+D}{D}} = \left(\frac{2a}{A}\right)^{D/2} B^s$. Define

$$\tilde{r} = \min_{s \in \{0\} \cup \mathbb{N}} \left\{ \binom{s+D}{D} : \binom{s+D}{D} > r \right\},$$

and let $\tilde{s}$ denote the corresponding s. Then,

$$\lambda_r = \lambda_{\tilde{r}} = \left(\frac{2a}{A}\right)^{D/2} B^{\tilde{s}} \quad \text{and} \quad \binom{\tilde{s}}{D} \leq \binom{\tilde{s}-1+D}{D} \leq r.$$

Using the lower bound, $\left(\frac{\tilde{s}}{D}\right)^D \leq \binom{\tilde{s}}{D}$, we obtain $\tilde{s} \leq D r^{1/D}$, completing the proof of the lower bound. ■

Proposition 31 *Let k be an isotropic SE-kernel. Suppose N covariates are sampled independent and identically from an isotropic Gaussian density with variance β^2 along each dimension. Define $M(N)$ to be any function of N such that $\lim_{N\to\infty} M(N)/(\log N)^D = 0$; i.e. $M(N) = o((\log N)^D)$. Suppose inference is performed using any set of inducing inputs, Z such that $|Z| = M(N)$. Then for any $\mathbf{y} \in \mathbb{R}^N$, for any $\epsilon > 0$ and for any $\delta \in (0,1)$, with probability at least $1-\delta$, $\mathrm{KL}[Q||P] = \Omega(N^{1-\epsilon})$.*

Proof By Lemma 28 and Proposition 21, for $\delta \in (0,1)$, with probability $1-\delta$,

$$\frac{|\lambda_m - \frac{1}{N}\tilde{\boldsymbol{\lambda}}_m|}{\lambda_m} = \mathcal{O}\left(r^2 \lambda_m \lambda_r^{-1/2} N^{-1/2} \delta^{-1/2} + r \exp(-\alpha r^{1/D}) + \sqrt{\frac{r \exp(-\alpha r^{1/D})}{N\delta}} \right). \quad (38)$$

with $\alpha = -\log B$ for any $1 \leq r \leq N$. Using Proposition 30 we have,

$$\frac{|\lambda_m - \frac{1}{N}\tilde{\boldsymbol{\lambda}}_m|}{\lambda_m} = \mathcal{O}\left(r^2 N^{-\frac{1}{2}} \exp(\alpha D r^{\frac{1}{D}}/2) \delta^{-\frac{1}{2}} + \frac{r}{\lambda_m} \exp(-\alpha r^{\frac{1}{D}}) + \frac{1}{\lambda_m} \sqrt{\frac{r \exp(-\alpha r^{1/D})}{N\delta}} \right).$$

For $\gamma \in (0, 1/2)$, choose $r = \lceil (\frac{1}{\alpha D} \log N^\gamma)^D \rceil$, then noting that for this choice of r, the third term in the sum is smaller than the second term,

$$\frac{|\lambda_m - \frac{1}{N}\tilde{\boldsymbol{\lambda}}_m|}{\lambda_m} = \mathcal{O}\Big(\delta^{-1/2} \Big(r^2 N^{(\gamma-1)/2} + \lambda_m^{-1} r N^{\frac{-\gamma}{D}} \Big) \Big).$$

Applying Proposition 30, with $M+1 = \lfloor (\frac{1}{D} \log_B N^{-\zeta/D})^D \rfloor$ with $\zeta \in (0, \gamma)$. We have

$$\lambda_{M+1}^{-1} r N^{\frac{-\gamma}{D}} \leq (M+1) \left(\frac{A}{2a}\right)^{D/2} B^{-D(M+1)^{1/D}} N^{-\gamma/D} \leq (M+1) \left(\frac{A}{2a}\right)^{D/2} N^{(\zeta-\gamma)/D}.$$

Thus, for such a choice of M, we have with probability at least $1-\delta$, $N\lambda_{M+1} = \tilde{\boldsymbol{\lambda}}_{M+1}(1 + o(1))$. It follows that with probability $1-\delta$, $\mathrm{KL}[Q||P] \geq N\lambda_{M+1}(1+o(1))$ and $N\lambda_{M+1} = \Omega(N^{1-\zeta/D})$. Choosing $\gamma = 1/4$ and $\zeta = \min\{\gamma/2, D\epsilon\}$, completes the proof. ■

Appendix F. Effect of Jitter on Bounds

In this section, we restate and prove Proposition 34. Recall that for $\epsilon > 0$, we define $\mathrm{Q}_{\mathrm{ff}}(\epsilon) \coloneqq \mathrm{K}_{\mathrm{uf}}^{\intercal}(\mathrm{K}_{\mathrm{uu}} + \epsilon \mathrm{I})^{-1}\mathrm{K}_{\mathrm{uf}}$.

Proposition 34 *Let $\mathcal{L}_\epsilon$ denote the evidence lower bound computed with jitter $\epsilon \geq 0$ added to K_{uu}, that is*

$$\mathcal{L}_\epsilon = -\frac{1}{2}\log\det(\mathrm{Q}_{\mathrm{ff}}(\epsilon) + \sigma^2\mathrm{I}) - \frac{1}{2}y^{\intercal}(\mathrm{Q}_{\mathrm{ff}}(\epsilon) + \sigma^2\mathrm{I})^{-1}y - \frac{N}{2}\log 2\pi - \frac{1}{2\sigma^2}\mathrm{tr}(\mathrm{K}_{\mathrm{ff}} - \mathrm{Q}_{\mathrm{ff}}(\epsilon)).$$

Then $\mathcal{L}_\epsilon$ is monotonically decreasing in ϵ. Similarly if $\mathcal{U}_\epsilon$ denotes the upper bound Eq. (12) computed with added jitter to K_{uu}, that is

$$\mathcal{U}_\epsilon \coloneqq -\frac{1}{2}\log\det(\mathrm{Q}_{\mathrm{ff}}(\epsilon) + \sigma^2\mathrm{I}) - \frac{1}{2}y^{\intercal}(\mathrm{Q}_{\mathrm{ff}}(\epsilon) + \mathrm{tr}(\mathrm{K}_{\mathrm{ff}} - \mathrm{Q}_{\mathrm{ff}}(\epsilon))\mathrm{I} + \sigma^2\mathrm{I})^{-1}y - \frac{N}{2}\log 2\pi.$$

Then $\mathcal{U}_\epsilon$ is monotonically increasing in ϵ. In particular, adding jitter can only make the upper bound on the log marginal likelihood larger and the ELBO smaller.

Proof Let $0 \leq \epsilon < \epsilon'$. For an arbitrary $v \in \mathbb{R}^n$,

$$v^{\intercal}(\mathrm{Q}_{\mathrm{ff}}(\epsilon) - \mathrm{Q}_{\mathrm{ff}}(\epsilon'))v = (\mathrm{K}_{\mathrm{uf}}v)^{\intercal}\big((\mathrm{K}_{\mathrm{uu}} + \epsilon\mathrm{I})^{-1} - (\mathrm{K}_{\mathrm{uu}} + \epsilon'\mathrm{I})^{-1}\big)(\mathrm{K}_{\mathrm{uf}}v) \geq 0,$$

The final inequality follows from $\mathrm{K}_{\mathrm{uu}} + \epsilon\mathrm{I} \prec \mathrm{K}_{\mathrm{uu}} + \epsilon'\mathrm{I}$ and Proposition 35. Therefore, $\mathrm{Q}_{\mathrm{ff}}(\epsilon') \prec \mathrm{Q}_{\mathrm{ff}}(\epsilon)$. From Proposition 35, we have

$$-\frac{1}{2}y^{\intercal}(\mathrm{Q}_{\mathrm{ff}}(\epsilon) + \sigma^2\mathrm{I})^{-1}y \geq -\frac{1}{2}y^{\intercal}(\mathrm{Q}_{\mathrm{ff}}(\epsilon') + \sigma^2\mathrm{I})^{-1}y. \tag{39}$$

Let A, B arbitrary $N \times N$ SPSD matrices with $A \succ B \succ \sigma^2\mathrm{I}$. Denote the eigenvalues of A and B respectively as $\lambda_1(A) \geq \ldots \lambda_N(A)$ and $\lambda_1(B) \geq \ldots \lambda_N(B)$. Then,

$$\begin{aligned}
\log\det A &= \sum_{i=1}^{N}\log\lambda_i(A) \\
&= \sum_{i=1}^{N}\log\lambda_i(B) + \sum_{i=1}^{n}\log\frac{\lambda_i(A)}{\lambda_i(B)} \\
&= \log\det B + \sum_{i=1}^{N}\log\left(1 + \frac{\lambda_i(A) - \lambda_i(B)}{\lambda_i(B)}\right) \\
&\leq \log\det B + \sum_{i=1}^{N}\frac{\lambda_i(A) - \lambda_i(B)}{\lambda_i(B)} \\
&\leq \log\det B + \frac{1}{\sigma^2}\mathrm{tr}(A - B).
\end{aligned} \tag{40}$$

The first inequality follows applying $\log(1+a) \leq a$ to each term in the sum. The second inequality used that $\lambda_i(B) \geq \sigma^2$ since $B \succ \sigma^2 \mathrm{I}$. Then,

$$\begin{aligned}
&-\frac{1}{2}\log\det(\mathrm{Q_{ff}}(\epsilon)+\sigma^2\mathrm{I}) - \frac{1}{2\sigma^2}\mathrm{tr}(\mathrm{K_{ff}}-\mathrm{Q_{ff}}(\epsilon)) \\
&\quad = -\frac{1}{2}\log\det(\mathrm{Q_{ff}}(\epsilon)+\sigma^2\mathrm{I}) - \frac{1}{2\sigma^2}\mathrm{tr}(\mathrm{Q_{ff}}(\epsilon')-\mathrm{Q_{ff}}(\epsilon)) - \frac{1}{2\sigma^2}\mathrm{tr}(\mathrm{K_{ff}}-\mathrm{Q_{ff}}(\epsilon')) \\
&\quad \geq -\frac{1}{2}\log\det(\mathrm{Q_{ff}}(\epsilon')+\sigma^2\mathrm{I}) - \frac{1}{2\sigma^2}\mathrm{tr}(\mathrm{K_{ff}}-\mathrm{Q_{ff}}(\epsilon')). \qquad (41)
\end{aligned}$$

where the final inequality follows from Eq. (40) with $A = \mathrm{Q_{ff}}(\epsilon)+\sigma^2\mathrm{I}$ and $B = \mathrm{Q_{ff}}(\epsilon')+\sigma^2\mathrm{I}$. Combining Eq. (39) with Eq. (41) proves the monotonicity of the lower bound in ϵ. The upper bound follows from Proposition 35 noting that in the quadratic form

$$\begin{aligned}
\mathrm{Q_{ff}}(\epsilon) + (\mathrm{tr}(\mathrm{K_{ff}}-\mathrm{Q_{ff}}(\epsilon))+\sigma^2)\mathrm{I} - \mathrm{Q_{ff}}(\epsilon') + (\mathrm{tr}(\mathrm{K_{ff}}-\mathrm{Q_{ff}}(\epsilon'))+\sigma^2)\mathrm{I} \\
= \mathrm{Q_{ff}}(\epsilon) - \mathrm{Q_{ff}}(\epsilon') + \mathrm{tr}(\mathrm{Q_{ff}}(\epsilon)-\mathrm{Q_{ff}}(\epsilon'))\mathrm{I} \\
\succ 0.
\end{aligned}$$

■

Appendix G. An Alternative Ridge Leverage Sampling Initialization

Many implementations of leverage score sampling allow for adaptively selecting the number of inducing points to achieve a desired level of accuracy. We briefly discuss the application of Algorithm 2 in Musco and Musco (2017) to the problem of sparse variational inference in Gaussian processes.

G.1. Effective Dimension

The number of points sampled by ridge leverage score methods to achieve a desired level of accuracy is closely related to the *effective dimension* of the kernel matrix, which can be thought of as measure of the complexity of the non-parameteric regression model. The effective dimension is defined as the sum of the ridge leverage scores,

$$d_{\mathrm{eff}}^{\omega} \coloneqq \sum_{n=1}^{N} \ell^{\omega}(x_n) = \sum_{m=1}^{N} \frac{\tilde{\lambda}_m}{\tilde{\lambda}_m + \omega}, \qquad (42)$$

and depends on the choice of kernel, the distribution of the covariates and the regularization parameter.

In order to compare such an adaptive method with the bounds discussed in Section 4, we need to consider the typical size of the effective dimension, assuming a fixed kernel and a random set of covariates with identical marginal distributions (or marginal distributions satisfying the conditions in Lemma 11).

For any fixed set of covariates, we can split the sum in Eq. (42) into two parts, yielding

$$d_{\mathrm{eff}}^{\omega} \leq S + \frac{1}{\omega}\sum_{m=S+1}^{N} \tilde{\lambda}_m, \qquad (43)$$

where S is an arbitrary positive integer. Upper bounds on the effective dimension can be obtained by choosing S so that the two terms on the right hand side of Eq. (43) are of the same order of magnitude.

G.2. Adaptively Selecting the Number of Inducing Points with Leverage Scores

We consider the application of Musco and Musco (2017, Algorithm 2) to the problem of selecting inducing inputs for sparse variational inference in GP models. This algorithm comes with the following bounds on the quality of the resulting Nyström approximation.

Lemma 37 (Musco and Musco (2017), Theorem 7) *Fix $\delta \in (0, \frac{1}{32})$. There exists an algorithm with run time $\mathcal{O}(NM^2)$ and memory complexity $\mathcal{O}(NM)$ that with probability $1 - 3\delta$ returns $M < 384 d_{\text{eff}}^{\omega} \log(d_{\text{eff}}^{\omega}/\delta)$ columns of $\mathrm{K_{ff}}$ such that the resulting Nyström approximation, $\mathrm{Q_{ff}}$, satisfies*

$$\|\mathrm{K_{ff}} - \mathrm{Q_{ff}}\|_{\text{op}} \leq \omega$$

where d_{eff}^{ω} denotes the effective dimension of the Gaussian process regressor with $\sigma^2 = \omega$.[14]

We can now consider the implications of this bound on sparse variational GP regression using Lemmas 3 and 4.

G.3. Ridge Leverage Scores and Sparse Variational Inference

We begin by considering the resulting error from employing Lemma 37 in Lemma 4. Noting that $\mathrm{tr}(\mathrm{K_{ff}} - \mathrm{Q_{ff}}) \leq N \|\mathrm{K_{ff}} - \mathrm{Q_{ff}}\|_{\text{op}}$, Lemma 4 gives us the bound

$$\mathbb{E}[\mathrm{KL}[Q||P] \mid \mathbf{Z}, \mathbf{X}] \leq N \frac{\|\mathbf{K_{ff}} - \mathbf{Q_{ff}}\|_{op}}{\sigma^2}. \tag{44}$$

Similarly, Lemma 3 becomes

$$\mathrm{KL}[Q||P] \leq \frac{\|\mathbf{K_{ff}} - \mathbf{Q_{ff}}\|_{\text{op}}}{2\sigma^2} \left(N + \frac{\|\mathbf{y}\|_2^2}{\sigma^2} \right).$$

Both of these bounds are small if $\|\mathbf{K_{ff}} - \mathbf{Q_{ff}}\|_{\text{op}} \ll 1/N$.

For simplicity, we consider the case when $\mathbf{y}$ is assumed to have a conditional distribution that agrees with the GP prior. Fix $\delta \in (0, 1/32)$ and $\gamma > 0$. Applying Markov's inequality to Eq. (44), with probability at least $1 - \delta$,

$$\mathrm{KL}[Q||P] \leq N \frac{\|\mathbf{K_{ff}} - \mathbf{Q_{ff}}\|_{\text{op}}}{\delta\sigma^2}. \tag{45}$$

We can apply the algorithm referred to in Lemma 37 with $\omega = \sigma^2 \delta\gamma/N$, so that with probability at least $1 - 3\delta$ a set of inducing inputs is chosen such that,

$$\|\mathbf{K_{ff}} - \mathbf{Q_{ff}}\|_{\text{op}} \leq \delta\sigma^2\gamma/N.$$

14. Note that $\mathrm{K_{ff}}$ and $\mathrm{Q_{ff}}$ are both independent of the noise parameter, so there is no requirement that the 'noise parameter' used for initializing inducing points matches the noise parameter used in performing regression.

We can then apply a union bound to conclude with probability at least $1 - 4\delta$,

$$\mathrm{KL}[Q||P] \leq \gamma. \tag{46}$$

By Corollary 12 and Markov's inequality, with probability at least $1 - \delta$,

$$\frac{1}{\omega} \sum_{m=S+1}^{N} \tilde{\boldsymbol{\lambda}}_m \leq \frac{N}{\delta\omega} \sum_{m=S+1}^{\infty} \lambda_m$$

for any $1 \leq S \leq N$. On the event where this holds and recalling we chose the parameter $\omega = \sigma^2\delta\gamma/N$, Eq. (43) implies that,

$$d_{\text{eff}}^{\omega} \leq S + \frac{1}{\delta\omega} \sum_{m=S+1}^{N} \tilde{\boldsymbol{\lambda}}_m \leq S + \frac{N^2\gamma}{\sigma^2\delta^2} \sum_{m=S+1}^{\infty} \lambda_m. \tag{47}$$

We can again apply the union bound to lower bound the probability that both the effective dimension is less that the bound in Eq. (47) and that Eq. (46) holds. This yields the following probabilistic bounds on the quality of sparse VI in GP regression with inducing points placed according to approximate ridge leverage scores.

Theorem 38 *Fix $\delta \in (0, \frac{1}{32})$, $\gamma > 0$. Under the same assumptions on the covariate distribution and the distribution of $\mathbf{y}$ as in Theorem 14 if inducing points are placed according to Musco and Musco (2017, Algorithm 2) with $\omega = \sigma^2\delta\gamma/N$, then with probability $1 - 5\delta$, $\mathbf{M} < 384d\log(d/\delta)$ and*

$$\mathrm{KL}[Q||P] \leq \gamma$$

where $d = \min_{S\in\mathbb{N}, S\leq N}\Big(S + \frac{N^2}{\sigma^2\delta^2\gamma}\sum_{m=S+1}^{\infty}\lambda_m\Big)$.

A similar argument in the case when we do not assume $\mathbf{y}$ is distributed according to the prior model leads to the following result:

Theorem 39 *Fix $\delta \in (0, \frac{1}{32})$, $\gamma > 0$. Under the same assumptions on the covariate distribution and the distribution of $\mathbf{y}$ as in Theorem 13 if inducing points are placed according to Musco and Musco (2017, Algorithm 2) with $\omega = \frac{2\sigma^2\delta\gamma}{N(1+R/\sigma^2)}$ then with probability $1 - 5\delta$, $\mathbf{M} < 384d'\log(d'/\delta)$ and*

$$\mathrm{KL}[Q||P] \leq \gamma$$

where $d' = \min_{S\in\mathbb{N}, S\leq N}\Big(S + \frac{N^2(1+R^2/\sigma^2)}{2\sigma^2\delta\gamma}\sum_{m=S+1}^{\infty}\lambda_m\Big)$.

Note that while the resulting bounds on $\mathbf{M}$ depend on the kernel and covariate distribution, the quality of the resulting approximation in both Theorems 38 and 39 does not.

The bounds implied by these results for various kernels are given in Table 3. Note that the asymptotic rates implied by both Theorems 38 and 39 are the same. This is because, unlike in the case of the M-DPP initialization in which the trace is bounded and this is used as an upper bound on the operator norm, the operator norm is bounded directly.

Kernel	Input Distribution	M
SE-Kernel	Compact support	$\mathcal{O}((\log N)^D \log\log(N))$
SE-Kernel	Gaussian	$\mathcal{O}((\log N)^D \log\log(N))$
Matérn ν	Compact support	$\mathcal{O}(N^{\frac{2D}{2\nu+D}} \log N)$

Table 3: Bounds on the number of inducing points used in Theorems 38 and 39.

References

Ahmed El Alaoui and Michael W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems (NIPS)*, pages 775–783, 2015.

Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Niles-Weed. Massively scalable Sinkhorn distances via the Nyström method. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4429–4439, 2019.

Nima Anari, Shayan Oveis Gharan, and Alireza Rezaei. Monte Carlo Markov chain algorithms for sampling strongly Rayleigh distributions and determinantal point processes. In *Conference on Learning Theory (COLT)*, pages 103–115, 2016.

Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory (COLT)*, pages 185–209, 2013.

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

Matthew J. Beal and Zoubin Ghahramani. The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statistics*, 7:453–464, 2003.

Mohamed-Ali Belabbas and Patrick J. Wolfe. Spectral methods in machine learning and new strategies for very large datasets. *Proceedings of the National Academy of Sciences (PNAS)*, 106(2):369–374, 2009.

Mikio L. Braun. Accurate error bounds for the eigenvalues of the kernel matrix. *Journal of Machine Learning Research (JMLR)*, 7(Nov):2303–2328, 2006.

David R. Burt, Carl E. Rasmussen, and Mark van der Wilk. Rates of convergence for sparse variational Gaussian process regression. In *International Conference on Machine Learning (ICML)*, pages 862–871, 2019.

Daniele Calandriello, Alessandro Lazaric, and Michal Valko. Distributed adaptive sampling for kernel matrix approximation. In *Artificial Intelligence and Statistics (AISTATS)*, pages 1421–1429, 2017.

Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, and Lorenzo Rosasco. Gaussian process optimization with adaptive sketching: Scalable and no regret. In *Conference on Learning Theory (COLT)*, pages 1–25, 2019.

Laming Chen, Guoxin Zhang, and Eric Zhou. Fast greedy MAP inference for determinantal point process to improve recommendation diversity. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5622–5633, 2018.

Ali Çivril and Malik Magdon-Ismail. On selecting a maximum volume sub-matrix of a matrix and related problems. *Theoretical Computer Science*, 410(47-49):4801–4811, 2009.

Luis Pedro Coelho. Jug: Software for parallel reproducible computation in Python. *Journal of Open Research Software*, 5(1), 2017.

Alexander Davies. *Effective Implementation of Gaussian Process Regression for Machine Learning*. PhD Thesis, University of Cambridge, 2015.

Marc P. Deisenroth, Yicheng Luo, and Mark van der Wilk. A practical guide to Gaussian processes. `https://drafts.distill.pub/gp/`, 2019.

Michał Dereziński, Daniele Calandriello, and Michal Valko. Exact sampling of determinantal point processes with sublinear time preprocessing. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11542–11554, 2019.

Giancarlo Ferrari-Trecate, Christopher K.I. Williams, and Manfred Opper. Finite-dimensional approximation of Gaussian processes. In *Advances in Neural Information Processing Systems (NIPS)*, pages 218–224, 1999.

Shai Fine and Katya Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2(Dec):243–264, 2001.

Leslie Foster, Alex Waagen, Nabeela Aijaz, Michael Hurley, Apolonio Luis, Joel Rinsky, Chandrika Satyavolu, Michael J. Way, Paul Gazis, and Ashok Srivastava. Stable and efficient Gaussian process calculations. *Journal of Machine Learning Research*, 10(4), 2009.

Jean Gallier. The Schur complement and symmetric positive semidefinite (and definite) matrices. `https://www.cis.upenn.edu/~jean/schur-comp.pdf`, 2010.

Jacob Gardner, Geoff Pleiss, Kilian Weinberger, David Bindel, and Andrew G. Wilson. Gpytorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7576–7586, 2018.

Guillaume Gautier, Guillermo Polito, Rémi Bardenet, and Michal Valko. DPPy: DPP sampling with Python. *Journal of Machine Learning Research*, 20(180):1–7, 2019.

Mark Gibbs and David Mackay. Efficient implementation of Gaussian processes. Technical report, Cavendish Laboratory, University of Cambridge, 1997.

Alex Gittens and Michael W. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. *Journal of Machine Learning Research (JMLR)*, 17(117):1–65, 2016.

GPy. GPy: A Gaussian process framework in Python. `http://github.com/SheffieldML/GPy`, since 2012.

Izrail S. Gradshteyn and Iosif M. Ryzhik. *Table of Integrals, Series, and Products*. Academic press, 2014.

James Hensman, Nicolo Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *Uncertainty in Artificial Intelligence (UAI)*, pages 282–290, 2013.

James Hensman, Alexander G. de G. Matthews, and Zoubin Ghahramani. Scalable variational Gaussian process classification. In *Artificial Intelligence and Statistics (AISTATS)*, pages 351–360, 2015.

James Hensman, Nicolas Durrande, and Arno Solin. Variational fourier features for gaussian processes. *The Journal of Machine Learning Research*, 18(1):5537–5588, 2018.

Jonathan Hermon and Justin Salez. Modified log-Sobolev inequalities for strong-Rayleigh measures, 2019.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

Jonathan H. Huggins, Trevor Campbell, Mikołaj Kasprzak, and Tamara Broderick. Scalable Gaussian process inference with finite-data mean and variance guarantees. In *Artificial Intelligence and Statistics (AISTATS)*, pages 796–805, 2019.

Jonathan H. Huggins, Mikołaj Kasprzak, Trevor Campbell, and Tamara Broderick. Practical posterior error bounds from variational objectives. In *Artificial Intelligence and Statistics (AISTATS)*, 2020.

Hyunjik Kim and Yee Whye Teh. Scaling up the automatic statistician: Scalable structure discovery using Gaussian processes. In *Artificial Intelligence and Statistics (AISTATS)*, pages 575–584, 2018.

Vladimir Koltchinskii and Evarist Giné. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167, 2000.

Hermann König. *Eigenvalue Distribution of Compact Operators*. Birkhäuser, 1986.

Alex Kulesza and Ben Taskar. k-DPPs: Fixed-size determinantal point processes. In *International Conference on Machine Learning (ICML)*, pages 1193–1200, 2011.

Miguel Lázaro-Gredilla and Anbal Figueiras-Vidal. Inter-domain Gaussian processes for sparse inference using inducing features. In *Advances in Neural Information Processing Systems (NIPS) 22*, pages 1087–1095, 2009.

Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Fast DPP sampling for Nyström with application to kernel methods. In *International Conference on Machine Learning (ICML)*, pages 2061–2070, 2016.

Alexander G. de G. Matthews. *Scalable Gaussian Process Inference using Variational Methods*. PhD Thesis, University of Cambridge, 2016.

Alexander G. de G. Matthews, James Hensman, Richard E. Turner, and Zoubin Ghahramani. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In *Artificial Intelligence and Statistics (AISTATS)*, pages 231–239, 2016.

Alexander G. de G. Matthews, Mark van der Wilk, Tom Nickson, Keisuke Fujii, Alexis Boukouvalas, Pablo León-Villagrá, Zoubin Ghahramani, and James Hensman. GPflow: A Gaussian process library using Tensorflow. *The Journal of Machine Learning Research*, 18(1):1299–1304, 2017.

James Mercer. Functions of positive and negative type, and their connection the theory of integral equations. *Philosophical Transactions of the Royal Society of London.*, 209 (441-458):415–446, 1909.

Charles A. Micchelli and Grace Wahba. Design problems for optimal surface interpolation. Technical report, Department Of Statistics, University of Wisconsin Madison, 1979.

Cameron Musco and Christopher Musco. Recursive sampling for the Nyström method. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3833–3845, 2017.

Leszek Plaskota. *Noisy Information and Computational Complexity*. Cambridge University Press, 1996.

Joaquin Quiñonero-Candela and Carl E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research (JMLR)*, 6((12)): 1939–1959, 2005.

Carl E. Rasmussen and Christopher K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1657–1665, 2015.

Matthias W. Seeger, Christopher K.I. Williams, and Neil D. Lawrence. Fast forward selection to speed up sparse Gaussian process regression. In *Artificial Intelligence and Statistics (AISTATS)*, pages 205–212, 2003.

Matthias W. Seeger, Sham M. Kakade, and Dean P. Foster. Information consistency of nonparametric Gaussian process methods. *IEEE Transactions on Information Theory*, 54(5):2376–2382, 2008.

John Shawe-Taylor, Christopher K.I. Williams, Nello Cristianini, and Jaz Kandola. On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Transactions on Information Theory*, 51(7):2510–2522, 2005.

Alexander J. Smola and Bernard Schölkopf. Sparse greedy matrix approximation for machine learning. In *International Conference on Machine Learning (ICML)*, pages 911–918, 2000.

Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1257–1264, 2006.

Arno Solin and Simo Srkk. Hilbert space methods for reduced-rank Gaussian process regression. *Statistics and Computing*, 30(2):419–446, 2020.

Terence Tao. *Topics in Random Matrix Theory.* American Mathematical Society, 2012.

Michalis K. Titsias. Variational model selection for sparse Gaussian process regression. Technical report, University of Manchester, UK, 2009a.

Michalis K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics (AISTATS)*, pages 567–574, 2009b.

Michalis K. Titsias. Variational inference for Gaussian and determinantal point processes. In *Workshop on Advances in Variational Inference (NIPS)*, 2014.

Richard E. Turner and Maneesh Sahani. Two problems with variational expectation maximisation for time-series models. In *Bayesian Time Series Models*, pages 104–124. Cambridge University Press, 2011.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

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

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

Huaiyu Zhu, Christopher K. I. Williams, Richard Rohwer, and Michal Morciniec. Gaussian regression and optimal finite dimensional linear models. In *Neural Networks and Machine Learning*, pages 167–184. Springer-Verlag, 1997.