Connections and Equivalences between the Nyström Method and Sparse Variational Gaussian Processes

Veit Wild veit.wild@keble.ox.ac.uk
Department of Statistics, University of Oxford, UK

Motonobu Kanagawa motonobu.kanagawa@eurecom.fr
Data Science Department, EURECOM, France

Dino Sejdinovic dino.sejdinovic@stats.ox.ac.uk
Department of Statistics, University of Oxford, UK

Abstract

We investigate the connections between sparse approximation methods for making kernel methods and Gaussian processes (GPs) scalable to massive data, focusing on the Nyström method and the Sparse Variational Gaussian Processes (SVGP). While sparse approximation methods for GPs and kernel methods share some algebraic similarities, the literature lacks a deep understanding of how and why they are related. This is a possible obstacle for the communications between the GP and kernel communities, making it difficult to transfer results from one side to the other. Our motivation is to remove this possible obstacle, by clarifying the connections between the sparse approximations for GPs and kernel methods. In this work, we study the two popular approaches, the Nyström and SVGP approximations, in the context of a regression problem, and establish various connections and equivalences between them. In particular, we provide an RKHS interpretation of the SVGP approximation, and show that the Evidence Lower Bound of the SVGP contains the objective function of the Nyström approximation, revealing the origin of the algebraic equivalence between the two approaches. We also study recently established convergence results for the SVGP and how they are related to the approximation quality of the Nyström method.

Keywords: Gaussian Processes, Kernel Methods, Sparse Approximation, Nyström Method, Sparse Variational Gaussian Processes

1. Introduction

Gaussian processes (GPs) and kernel methods are the two principled learning approaches that make use of *positive definite kernels*, and have been studied extensively in statistics and machine learning. On one hand, GP-based approaches (Rasmussen and Williams, 2006) employ a kernel to induce the corresponding GP, in oder to define a prior distribution of the ground-truth latent function of interest. Given data, Bayes' rule is then applied to obtain the posterior distribution of the latent function. On the other hand, kernel methods (Schölkopf and Smola, 2002) make use of a kernel to induce the corresponding Reproducing Kernel Hilbert Space (RKHS) as a "hypothesis space." Given data, empirical risk minimization is then performed in the RKHS to estimate the ground-truth function of interest. Although the GP and kernel approaches have different modeling philosophies, there are indeed deep connections and equivalences between them, which extend beyond a superficial

similarity (Parzen, 1961; Kimeldorf and Wahba, 1970; Berlinet and Thomas-Agnan, 2004; Kanagawa et al., 2018).

The elegance of the GP and kernel approaches are that the infinite dimensional learning problems can be reduced to the corresponding finite dimensional problems. However, this comes with a cost: the computational complexity of either approach is usually cubic or at least quadratic with respect to the data size. This unfavorable scaling property has motivated the developments of several approximation methods to make the GP and kernel approaches scalable. *Sparse approximation* methods, which approximate the solution of interest using a set of input points smaller than training data, are among the most popular and successful approximation approaches. These approaches have been studied since the earliest developments of the GP and kernel approaches (e.g., Williams and Seeger, 2001; Csató and Opper, 2002; Smola and Schölkopf, 2000; Seeger et al., 2003).

As the GP and kernel communities grow, sparse approximation methods for either approach tend to be developed independently to those for the other approach. For instance, consider the Sparse Variational Gaussian Process (SVGP) approach of Titsias (2009a,b), which is one of the most successful and widely used sparse approximation methods for GPs. The SVGP is derived in the framework of variational Bayesian inference, so that the sparse approximation is to be chosen to minimize the KL divergence to the exact GP posterior. As such, the developments in SVGP (e.g., Hensman et al., 2013, 2015a; Matthews et al., 2016; Burt et al., 2019; Rossi et al., 2021) have proceeded almost independently of the corresponding literature on sparse approximations for kernel methods. Similarly, the recent advances in using and understanding the Nyström method (Williams and Seeger, 2001), which is one of the most popular sparse approximations in kernel methods, have been made independently to those of sparse GP approximations. The majority of these advances focus on an efficient approximation of the kernel matrix (e.g., Drineas and Mahoney, 2005; Belabbas and Wolfe, 2009; Gittens and Mahoney, 2016; Derezinski et al., 2020) or empirical risk minimization in the RKHS with a reduced basis (e.g, Bach, 2013; El Alaoui and Mahoney, 2015; Rudi et al., 2015, 2017; Meanti et al., 2020). This separation of two lines of research are arguably due to the difference in the notations and modeling philosophies of GPs and kernel methods. The separation makes it difficult to transfer useful and interesting results from one side to the other, and the communities might have missed an important advance that may be obtained otherwise. The motivation of the current work is to overcome this potential difficulty by bridging the two lines of research.

In this work, we investigate the connections between the sparse approximation methods for GPs and kernel methods. Specifically, we focus on the regression setting, and study the relationships between the SVGP and Nyström approximations. We summarize below our contributions and main findings:

- In Section 3, we establish an equivalence between the SVGP posterior mean function and kernel ridge regression (KRR) using the Nyström method. To understand this equivalence, we analyze the Evidence Lower Bound (ELBO) that is optimized by the SVGP approximation. We show that the ELBO contains the objective function that the Nyström KRR essentially minimizes. In this sense, the equivalence is not a "coincidence."

- We make an RKHS interpretation for the ELBO of the SVGP approximation. We reformulate the ELBO in terms of orthogonal projections onto the subspace spanned by the inducing inputs (or landmark points). This formulation enables one to understand the ELBO geometrically. Specifically, it shows that the SVGP posterior covariance function is given via the solution to a certain optimization problem in the RKHS. Moreover, it enables a geometric understanding of the ELBO as an objective function for choosing inducing inputs. These results may be useful for inspiring more advanced SVGP approaches, in a similar manner as Shi et al. (2020) who used a certain geometric argument for justifying their approach.

- We present a worst case error interpretation of the SVGP posterior variance function. We show that it is the sum of two worst case errors in RKHSs: one is that of kernel-based interpolation at inducing inputs, and the other is that of the Nyström KRR based on noisy observations. This interpretation may be useful for understanding the impacts of the choice of a kernel and inducing points on the SVGP uncertainty estimates, as the RKHS formulation enables a discussion of the "capacity" of the model. Moreover, these RKHS interpretations enable kernel researchers to understand the SVGP approximation in their own terminology and provide a tool for uncertainty quantification.

- In Section 4, we discuss convergence results for the SVGP recently established by Burt et al. (2019, 2020), and investigate how they are related to the approximation quality of the Nyström KRR. To this end, we first show that the "data fit term" in the marginal likelihood of Gaussian process regression (GPR) is essentially identical to the objective function value of the corresponding KRR estimator. This result enables us to rewrite relevant terms in the KL divergence for the SVGP approximation as the "excess risk" of the Nyström KRR over the exact KRR estimator. Since the theoretical arguments of Burt et al. (2020) are essentially based on bounding this "excess risk," many of their results can be directly translated to the corresponding results on the Nyström KRR. Moreover, the existence of the "excess risk" of the Nyström KRR suggests that a more refined analysis may be done for the KL divergence for the SVGP approximation by employing sharper theoretical results on the Nyström KRR (e.g., Bach, 2013; El Alaoui and Mahoney, 2015; Rudi et al., 2015).

- We also establish a novel approximation error bound for the Nyström KRR in terms of its RKHS distance to the exact KRR solution. This bound may be interesting in its own right, since it shares a certain structural similarity to a fundamental bound in Burt et al. (2020) on the KL divergence for the SVGP approximation. Moreover, because of the equivalence between the SVGP posterior mean function and the Nyström KRR estimator, and that between the exact GP posterior mean function and the exact KRR estimator, the novel bound also holds as a bound on the RKHS distance between the SVGP and exact posterior mean functions. This result is useful in that it leads to an approximation bound for the *derivatives* of the SVGP posterior mean function, when the kernel is continuously differentiable. This demonstrates the usefulness of the RKHS interpretations.

- Finally, we study the *lower bound* of the averaged KL divergence for the SVGP approximation, where the average is taken with respect to the distribution of training outputs under the prior model (Burt et al., 2020, Lemma 4). By the identity of the "data fit terms" and the "excess risk" of the Nyström KRR, this lower bound can be directly transformed to a lower bound for the excess risk of the Nyström KRR. This new lower bound is given in terms of the ratio of the "complexities" of the exact and approximation kernel models, and provides a novel theoretical understanding of the Nyström KRR. This is another manifestation of benefits of studying the connections between the GP and kernel approaches.

This paper is organized as follows. Section 2 provides relevant background on Gaussian processes and kernel methods. Section 3 investigates the connections and equivalences between the SVGP and the Nyström approximations. Section 4 studies the approximation properties of the SVGP and Nyström approximations and their connections. Section 5 concludes.

1.1 Notation

We use the following notation in this paper. Let $\mathbb{N}$ be the set of natural numbers, $\mathbb{R}$ be the real line, and $\mathbb{R}^d$ for $d \in \mathbb{N}$ be the d-dimensional Euclidean space. For any $v \in \mathbb{R}^d$, $\|v\|$ denotes the Euclidean norm.

Let $\mathcal{X}$ be a nonempty set. For a function $f : \mathcal{X} \to \mathbb{R}$ and $X := (x_1, \ldots, x_n) \in \mathcal{X}^n$ with $n \in \mathbb{N}$, denote by f_X the n-vector consisting of function values evaluated at points in X: $f_X := (f(x_1), \ldots, f(x_n))^\top \in \mathbb{R}^n$. Similarly, for a function with two arguments $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and $X := (x_1, ..., x_n) \in \mathcal{X}^n$ and $Z := (z_1, ..., z_m) \in \mathcal{X}^m$ with $n, m \in \mathbb{N}$, define $k_{XZ} \in \mathbb{R}^{n \times m}$ by $(k_{XZ})_{i,j} = k(x_i, z_j)$ for $i = 1, \ldots, n$, $j = 1, \ldots, m$. For $X := (x_1, ..., x_n) \in \mathcal{X}^n$, let $k_X(x) := (k(x_1, x), \ldots, k(x_n, x))^\top \in \mathbb{R}^n$ for any $x \in \mathcal{X}$ and denote by $k_X(\cdot)$ the vector-valued function $x \in \mathcal{X} \mapsto k_X(x) \in \mathbb{R}^n$.

For a symmetric matrix Σ, denote by $\Sigma \succ 0$ and $\Sigma \succeq 0$ that Σ is positive definite and positive semi-definite, respectively. For $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$ with $\Sigma \succeq 0$, denote by $\mathcal{N}(\mu, \Sigma)$ the Gaussian distribution on $\mathbb{R}^n$ with mean vector μ and covariance matrix Σ. Let $\mathcal{N}(\cdot \mid \mu, \Sigma)$ be its probability density function. For a matrix $A \in \mathbb{R}^{n \times n}$, $\mathrm{tr}(A)$ and $\det(A)$ denote its trace and determinant, respectively.

Let $\mathcal{Y}$ be a measurable space, $Y \in \mathcal{Y}$ be a random variable, and $\mathbb{P}$ be a probability measure on $\mathcal{Y}$. We write $Y \sim \mathbb{P}$ to mean that Y follows $\mathbb{P}$. For a measurable function $g : \mathcal{Y} \to \mathbb{R}$, denote by $\int g(y) d\mathbb{P}(y)$ its integral with respect to $\mathbb{P}$ and by $\mathbb{E}[g(Y)]$ the expectation of $g(Y)$. When $\mathbb{P}$ has a density function $p : \mathcal{Y} \to \mathbb{R}$ with respect to a reference measure λ on $\mathcal{Y}$ (e.g., the Lebesgue measure when $\mathcal{Y} = \mathbb{R}^n$), the integral is denoted by $\int g(y) p(y) d\lambda(y)$.

2. Background

This section briefly reviews reproducing kernel Hilbert spaces (RKHS) and Gaussian processes (GP). In particular, we focus on the respective approaches to *regression*, namely kernel ridge regression (KRR) and Gaussian process regression (GPR).

We first describe the regression problem. Let $\mathcal{X}$ be a non-empty set. Suppose we are given $n \in \mathbb{N}$ paired observations

$$(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}.$$

We assume that there exists a function $f_0 : \mathcal{X} \to \mathbb{R}$ such that

$$y_i = f_0(x_i) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

where $\varepsilon_1, \ldots, \varepsilon_n \in \mathbb{R}$ are independent, zero-mean, noise variables. f_0 is called *regression function*. The task of regression is to estimate (or learn) f_0 from the training data $(x_i, y_i)_{i=1}^n$. We will often write $X := (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $y := (y_1, \ldots, y_n) \in \mathbb{R}^n$.

2.1 Kernel Ridge Regression (KRR)

2.1.1 Kernels and RKHSs

We review here basics of kernels and RKHSs. For details, we refer to Schölkopf and Smola (2002); Hofmann et al. (2008); Steinwart and Christmann (2008).

Let $\mathcal{X}$ be an arbitrary non-empty set. A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *positive definite kernel*, if for every $n \in \mathbb{N}$ and every $X = (x_1, ..., x_n) \in \mathcal{X}^n$, the induced kernel matrix $k_{XX} = (k(x_i, x_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}$ is positive semi-definite. We may simply call such k *kernel*. By the Moore-Aronszajn theorem (Aronszajn, 1950), for any such kernel k there exists a uniquely associated Hilbert space $\left(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}}\right)$ of real-valued functions $f : \mathcal{X} \to \mathbb{R}$ called *reproducing kernel Hilbert space (RKHS)* such that

1. $k(\cdot, x) \in \mathcal{H}$ for every $x \in \mathcal{X}$ and
2. $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$ for every $f \in \mathcal{H}$ and $x \in \mathcal{X}$,

where $k(\cdot, x)$ denotes the function of the first argument with x being fixed: $x' \in \mathcal{X} \to k(x', x)$. The kernel k is called *reproducing kernel* of $\mathcal{H}$.

Examples of kernels on $\mathcal{X} \subset \mathbb{R}^d$ include the following. For $\gamma > 0$ the Gaussian kernel or square-exponential kernel is defined as $k_\gamma(x, x') := \exp(-\frac{\|x - x'\|^2}{\gamma^2})$ for $x, x' \in \mathcal{X}$. For constants $\alpha > 0$ and $h > 0$ the Matérn kernel is defined for $x, x' \in \mathcal{X}$ as $k_{\alpha,h}(x, x') := \frac{1}{2^{\alpha-1}\Gamma(\alpha)}(\frac{2\alpha\|x-x'\|}{h})^\alpha K_\alpha(\frac{2\alpha\|x-x'\|}{h})$, where Γ is the gamma function and K_α is the modified Bessel function of the second kind of order α.

The RKHS $\mathcal{H}_k$ may be "explicitly" constructed from k as follows (Hofmann et al., 2008, Section 2.2.1). Define a function space

$$\mathcal{H}_0 := \Big\{ f = \sum_{i=1}^n \alpha_i k(\cdot, x_i) \mid \ n \in \mathbb{N},\ \alpha_1, ..., \alpha_n \in \mathbb{R},\ x_1, ..., x_n \in \mathcal{X} \Big\}.$$

For $f = \sum_{i=1}^n \alpha_i k(\cdot, x_i) \in \mathcal{H}_0$ and $g = \sum_{i=1}^m \beta_i k(\cdot, y_i) \in \mathcal{H}_0$, define an inner product

$$\langle f, g \rangle_{\mathcal{H}_0} := \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j k(x_i, y_j).$$

Then we have

$$\mathcal{H}_k = \overline{\mathcal{H}_0},$$

i.e., the RKHS $\mathcal{H}_k$ is the closure of $\mathcal{H}_0$ with respect to the norm induced by the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$.

2.1.2 Regression Approach

Kernel ridge regression (KRR) is an approach to regression using a kernel k and its RKHS $\mathcal{H}_k$. The KRR estimator $\hat{f}$ of the regression function f_0 in (1) is defined as the solution of the following *regularized empirical risk minimization (ERM)* problem

$$\hat{f} = \underset{f \in \mathcal{H}_k}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda ||f||^2_{\mathcal{H}_k}, \tag{2}$$

where $\lambda > 0$ is a regularization constant. To gain an intuitive understanding for the necessity of the regularization, suppose that the function space $\mathcal{H}_k$ is potentially very large and therefore the unregularised ERM tends to interpolate the data points. The regularization term in (2) imposes a certain degree of smoothness onto the solution $\hat{f}$ and hence prevents it from adapting too strongly to the training data $(x_1, y_1), \ldots, (x_n, y_n)$.

Let $X := (x_1, ..., x_n) \in \mathcal{X}^n$ and $y := (y_1, ..., y_n)^\top \in \mathbb{R}^n$. By the representer theorem (Schölkopf et al., 2001), the solution $\hat{f}$ is given as a linear combination of $k(\cdot, x_1)$, $\ldots, k(\cdot, x_n)$. Hence, the optimization problem (2) reduces to that of the coefficients of the linear combination. As a result, the estimator is given by

$$\hat{f} = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i), \tag{3}$$

where $\alpha := (\alpha_1, ..., \alpha_n)^\top \in \mathbb{R}^n$ is given by

$$\alpha = (k_{XX} + n\lambda I_n)^{-1} y,$$

where $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix. The prediction of KRR at any $x \in \mathcal{X}$ is compactly written as

$$\hat{f}(x) = k_X(x)^\top \alpha = k_X(x)^\top (k_{XX} + n\lambda I_n)^{-1} y, \tag{4}$$

where $k_X(x) = (k(x_1, x), \ldots, k(x_n, x))^\top \in \mathbb{R}^n$.

The elegance of KRR is that we arrive at a simple, closed form expression for an infinite-dimensional optimisation problem. As we shall see soon, the same expression will arise in the context of non-parametric Bayesian learning.

2.2 Gaussian Process Regression (GPR)

2.2.1 Gaussian Processes

Gaussian processes (GPs) are one of the main workhorses of Bayesian nonparametric statistics and machine learning, as they can be used to place a prior distribution over functions. See Rasmussen and Williams (2006) for more details.

Let $\mathcal{X}$ be a non-empty set, $m : \mathcal{X} \to \mathbb{R}$ a function and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a positive definite kernel. A random function $F : \mathcal{X} \to \mathbb{R}$ is called *Gaussian process (GP)* with mean function m and covariance kernel k, if for all $n \in \mathbb{N}$ and all $X = (x_1, ..., x_n) \in \mathcal{X}^n$, the random vector $F_X := (F(x_1), \ldots, F(x_n))^\top \in \mathbb{R}^n$ satisfies

$$F_X \sim \mathcal{N}(m_X, k_{XX}),$$

i.e., F_X follows the Gaussian distribution with mean vector $m_X = (m(x_1), \ldots, m(x_n))^\top \in \mathbb{R}^n$ and covariance matrix $k_{XX} = (k(x_i, x_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}$. In this case, we write

$$F \sim GP(m, k).$$

By definition, we have

$$\begin{aligned} m(x) &= \mathbb{E}\left[F(x)\right], \quad x \in \mathcal{X}, \\ k(x, x') &= \mathbb{E}\left[(F(x) - m(x))(F(x') - m(x'))\right], \quad x, x' \in \mathcal{X} \end{aligned}$$

For any function $m : \mathcal{X} \to \mathbb{R}$ and kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists[1] a GP whose mean function is m and covariance function is k. Therefore, by choosing m and k, one can implicitly define the corresponding GP, $F \sim GP(m, k)$. This is how a GP is used to define a prior distribution in Bayesian nonparametrics.

2.2.2 Regression Approach

Gaussian process regression (GPR) is a Bayesian nonparametric approach to the regression problem. In GPR, the regression function f_0 in Eq. (1) is the quantity of interest and modeled as a random function F. The prior distribution is given by a GP

$$F \sim GP(m, k). \tag{5}$$

where the modeler choses the mean function m and covariance function k to encode his/her prior knowledge/assumption about the regression function f_0.

The likelihood model of F for the observations $y = (y_1, \ldots, y_n)^\top$ is given by

$$y_i = F(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{6}$$

where[2] $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is an independent Gaussian noise with variance $\sigma^2 > 0$.

By Bayes' rule, the posterior distribution of F given y, under the prior (5), is given by again a GP[3]

$$F \mid y \sim GP(\bar{m}, \bar{k}),$$

where $\bar{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\bar{m} : \mathcal{X} \to \mathbb{R}$ are defined as

$$\bar{m}(x) := m(x) + k_X(x)^\top (k_{XX} + \sigma^2 I_n)^{-1}(y - m_X), \tag{7}$$

$$\bar{k}(x, x') := k(x, x') - k_X(x)^\top (k_{XX} + \sigma^2 I_n)^{-1} k_X(x'), \tag{8}$$

1. This is a consequence of the Kolmogorov consistency theorem (see e.g. Tao, 2011, Chapter 2.4).
2. In GPR, the noise assumption can be weaker, e.g., $\varepsilon_1, \ldots, \varepsilon_n$ can be dependent and/or their variances can be different. In this paper, we consider this simplest noise model to investigate connections to the KRR.
3. See e.g. Rasmussen and Williams (2006) for derivation.

where $k_X(x) = (k(x_1, x), \ldots, k(x_n, x))^\top \in \mathbb{R}^n$. We call $GP(\bar{m}, \bar{k})$ the *posterior GP*, $\bar{m}$ the *posterior mean function* and $\bar{k}$ the *posterior covariance function*. We use the following notation for the probability measure of the posterior GP:

$$\mathbb{P}^{F|y} := GP(\bar{m}, \bar{k}). \tag{9}$$

The posterior mean function $\bar{m}$ serves as an estimator of the regression function f_0. On the other hand, the posterior covariance function $\bar{k}$ provides a way of uncertainty quantification: $\bar{k}(x, x)$ is the *posterior variance* of $F(x)$ given y. Its square root $\sqrt{k(x, x)}$ is the *posterior standard deviation* and may be used to construct a Bayesian credible interval for $f_0(x)$.

2.3 Connections between KRR and GPR

There is a well-known equivalence between the KRR estimator and the posterior mean function of GPR (Kimeldorf and Wahba, 1970; Kanagawa et al., 2018). This is summarized in the following theorem.

Theorem 1 *Let k be a kernel, and suppose that data $(x_i, y_i)_{i=1}^n \subset \mathcal{X} \times \mathbb{R}$ are given.*

- *Let $\bar{m}$ be the posterior mean function (7) of GPR with the prior being the zero-mean Gaussian process $F \sim GP(0, k)$ (i.e., $m(x) = 0$, $\forall x \in \mathcal{X}$, in (5)).*
- *Let $\hat{f}$ be the KRR estimator (4) performed on the RKHS $\mathcal{H}_k$.*

Then if $\sigma^2 = n\lambda$, we have $\hat{f} = \bar{m}$.

This result provides a Bayesian interpretation for the KRR estimator, and a least-squares interpretation for the GPR posterior mean function. In particular, the condition $\sigma^2 = n\lambda$ shows that specifying the noise variance σ^2 in GPR is equivalent to specifying a regularization constant λ in the KRR. For GPR, this equivalence implies that assuming observation noises works as regularization or smoothing. For KRR, it implies that the regularization constant may be learned from the data via interpreting it as a (scaled) noise variance. The equivalence between KRR and GPR enables an alternative interpretation of either approach and opens up a possibility of devising novel learning algorithms.

RKHS Interpretation of Posterior Variances. There is also an RKHS interpretation of the posterior variance $\bar{k}(x, x)$ of GPR (Kanagawa et al., 2018, Section 3.4). To describe this, define $w^\sigma : \mathcal{X} \to \mathbb{R}^n$ by

$$w^\sigma(x) := (k_{XX} + \sigma^2 I_n)^{-1} k_X(x), \quad x \in \mathcal{X},$$

Then, provided $\sigma^2 = n\lambda$, the KRR estimator in (4) can be written as

$$\hat{f}(x) = \sum_{i=1}^n w_i^\sigma(x) y_i = y^\top w^\sigma(x), \quad x \in \mathcal{X}. \tag{10}$$

which is linear in the training outputs $y = (y_1, \ldots, y_n)^\top$. Define also an augmented kernel

$$k^\sigma(x, x') = k(x, x') + \sigma^2 \mathbb{1}\{x = x'\}, \quad x, x' \in \mathcal{X},$$

where $\mathbb{1}\{x = x'\} = 1$ if $x = x'$ and $\mathbb{1}\{x = x'\} = 0$ otherwise. Intuitively, its RKHS $\mathcal{H}_{k^\sigma}$ may be understood as an "corrupted" version of $\mathcal{H}_k$, in the sense that each function in $\mathcal{H}_{k^\sigma}$ is the sum of a function from $\mathcal{H}_k$ and an independent noise function; see Kanagawa et al. (2018, Section 3.4) for a discussion.

The following theorem relates that the posterior variance $\bar{k}(x, x)$ of GPR to a worst case error of predictions by KRR in $\mathcal{H}_{k^\sigma}$.

Theorem 2 *(Kanagawa et al., 2018, Proposition 3.8) Let $\bar{k}$ be the posterior covariance function* (8) *of GPR based on a prior $F \sim GP(0, k)$ and observation noise variance $\sigma^2 \geq 0$ (Suppose the kernel matrix k_{XX} is invertible if $\sigma^2 = 0$). If $\sigma^2 = n\lambda$ and $x \neq x_i$ for all $i = 1, ..., n$, then*

$$\sqrt{\bar{k}(x, x) + \sigma^2} = \sup_{g \in \mathcal{H}_{k^\sigma} : \|g\|_{\mathcal{H}_{k^\sigma}} \leq 1} \Big(g(x) - \sum_{i=1}^{n} w_i^\sigma(x) g(x_i) \Big).$$

In the right hand side, $\sum_{i=1}^n w_i^\sigma(x) g(x_i)$ can be interpreted as the prediction of $g(x)$ by KRR trained with data $(x_i, y_i)_{i=1}^n$ where $y_i = g(x_i)$ (see Eq. (10)). Thus, the right hand side is the worst case error of KRR predictions at input x, for functions g from the unit ball in $\mathcal{H}_{k^\delta}$. This result suggests that uncertainty quantification can be also done in the RKHS framework.

In this paper, we investigate whether these parallels between KRR and GPR extend to their *sparse approximations*, which have been developed (largely independently within the two research communities) to deal with the unfavourable computational properties of KRR and GPR. The following sections are devoted to this question.

3. Sparse Approximations

The elegance of KRR and GPR is that closed form expressions are respectively available for the solution of optimization or Bayesian inference in potentially infinite dimensional function spaces. Unfortunately, this comes with high computational costs, since both methods involve the inversion of the regularized kernel matrix, which leads to the computational complexity of $\mathcal{O}(n^3)$, where n is the size of training data $(x_i, y_i)_{i=1}^n$. Both kernel and GP communities have been developing a variety of approximation methods for making their respective approaches scalable. One of the most successful approaches is *sparse approximation*, which approximates the solution of interest using a smaller set of input points $z_1, \ldots, z_m \in \mathcal{X}$, where $m \in \mathbb{N}$ may be much smaller than the original data size n.

We investigate here connections between the sparse approximation methods for KRR and GPR. Specifically, we focus on the connections between the *Nyström approximation* and the *Sparse Variational Gaussian Process (SVGP) approximation*, which are respectively popular sparse approximation methods for the kernel and GP-based approaches. Sections 3.1 and 3.2 review the Nyström and the SVGP methods, respectively. Section 3.3 describes an equivalence between the SVGP posterior mean function and the Nyström KRR, and investigates the original of the equivalence by studying the ELBO for the SVGP. Section 3.4 summarizes an equivalence between the Nyström and the *Deterministic Training Conditional (DTC)* approximation, a classic sparse GP approximation approach. Section 3.5 provides a geometric interpretation of the SVGP posterior covariance function, based

on which Section 3.6 investigates further the ELBO. Section 3.7 provides an RKHS interpretation of the SVGP posterior *variance* function as consisting of worst case errors of kernel-based interpolation and the Nyström KRR.

3.1 Nyström Approximation

The Nyström method was first proposed by Williams and Seeger (2001) for scaling up kernel-based learning algorithms. It has been successfully used in a variety of applications including manifold leaning (Talwalkar et al., 2008, 2013), computer vision (Fowlkes et al., 2004; Belabbas and Wolfe, 2009), and approximate sampling (Affandi et al., 2013), to name a few. Recent studies make use of the Nyström method to enable KRR to handle millions to billions of data points (e.g. Rudi et al., 2017; Meanti et al., 2020).

We describe here the use of the Nyström approximation in KRR. In particular, we consider a popular version classically known as the *subset of regressors* (Wahba, 1990, Chapter 7), which has been widely used both in practice and theory (e.g., Smola and Schölkopf, 2000; Rudi et al., 2015, 2017; Meanti et al., 2020). As before, let $(x_i, y_i)_{i=1}^n \subset \mathcal{X} \times \mathbb{R}$ be training data, and let $X := (x_1, ..., x_n) \in \mathcal{X}^n$ and $y := (y_1, ..., y_n)^\top \in \mathbb{R}^n$.

For $m \in \mathbb{N}$, let $z_1, \ldots, z_m \in \mathcal{X}$ be a set of input points based on which we approximate the KRR solution. These points $z_1, \ldots, z_m$ are usually a subset of training input points $x_1, \ldots, x_n$ in the kernel literature, but we allow for $z_1, \ldots, z_m$ to be generic points in $\mathcal{X}$ for a later comparison with the GP counterpart. Write $Z = (z_1, \ldots, z_m) \in \mathcal{X}^m$. Suppose that the kernel matrix $k_{ZZ} = (k(z_i, z_j))_{i,j=1}^m \in \mathbb{R}^{m \times m}$ is invertible.

Let $M \subset \mathcal{H}_k$ be the finite dimensional subspace spanned by $k(\cdot, z_1), \ldots, k(\cdot, z_m)$:

$$M := \operatorname{span}(k(\cdot, z_1), \ldots, k(\cdot, z_m)) := \left\{ \sum_{j=1}^m \alpha_j k(\cdot, z_j) \mid \alpha_1, ..., \alpha_m \in \mathbb{R} \right\}. \tag{11}$$

We replace the hypothesis space $\mathcal{H}_k$ in the KRR objective function (2) by this subspace M, and define its solution $\bar{f}$ as the Nyström approximation of the KRR solution $\hat{f}$:

$$\bar{f} := \operatorname*{arg\,min}_{f \in M} \frac{1}{n} \sum_{i=1}^n \big(y_i - f(x_i)\big)^2 + \lambda \|f\|_{\mathcal{H}_k}^2, \tag{12}$$

In other words, we approximately solve the minimization problem of KRR by searching for the solution of the form

$$f = \sum_{i=1}^m \beta_i k(\cdot, z_i) = k_Z(\cdot)^\top \beta$$

for some coefficients $\beta := (\beta_1, \ldots, \beta_m)^\top \in \mathbb{R}^m$, where $k_Z(\cdot) := (k(\cdot, z_1), \ldots, k(\cdot, z_m))^\top$. Inserting this expression in (12), the optimization problem now becomes

$$\min_{\beta \in \mathbb{R}^m} \frac{1}{n} \|y - k_{XZ}\beta\|^2 + \lambda \beta^\top k_{ZZ} \beta,$$

where $k_{XZ} \in \mathbb{R}^{n \times m}$ with $(k_{XZ})_{i,j} = k(x_i, z_j)$ and $k_{ZZ} \in \mathbb{R}^{m \times m}$ with $(k_{ZZ})_{i,j} = k(z_i, z_j)$. Taking the first order derivative with respect to β leads to the condition

$$-\frac{2}{n} k_{ZX} y + \frac{2}{n} k_{ZX} k_{XZ} \beta + 2\lambda k_{ZZ} \beta = 0,$$

which is satisfied for

$$\beta = \big(k_{ZX}k_{XZ} + n\lambda k_{ZZ}\big)^{-1}k_{ZX}y.$$

This leads to the following expression of the Nyström approximation:

$$\bar{f}(x) = k_Z(x)^\top (n\lambda k_{ZZ} + k_{ZX}k_{XZ})^{-1}k_{ZX}y. \tag{13}$$

This approximation can be computed with the complexity of $\mathcal{O}(nm^2+m^3)$ instead of $\mathcal{O}(n^3)$, since the inversion of a $n \times n$ matrix is replaced by that of a $m \times m$ matrix. This grants significant computational gains, if m is much smaller than n and hence allows KRR to be applied to large data sets. Of course, how to choose m and the input points $z_1, \ldots, z_m$ depends not only on the computational budget but also on how accurately $\bar{f}$ approximates the KRR solution $\hat{f}$. We discuss this issue in Section 4.

3.1.1 Characterization with an Approximate Kernel

We study here another characterization of the Nyström approximation based on a certain approximate kernel. This characterization provides a natural connection of the Nyström method to one of sparse approximation methods for GPR, as we will see later.

For any $f \in \mathcal{H}_k$, denote by $P_M\big(f\big) \in M$ the *orthogonal projection* of f onto the subspace M:

$$P_M\big(f\big) := \operatorname*{argmin}_{g \in M} \|f - g\|_{\mathcal{H}_k},$$

which is the best approximation of f by an element in M. The projection is given as $P_m(f) = \sum_{j=1}^m \alpha_j^* k(\cdot, z_j)$, where the $\alpha^* = (\alpha_1^*, \ldots, \alpha_m^*)^\top \in \mathbb{R}^m$ is the solution of

$$\min_{\alpha \in \mathbb{R}^m} \|f - \sum_{j=1}^m \alpha_j k(\cdot, z_j)\|_{\mathcal{H}_k}.$$

Given that the kernel matrix k_{ZZ} is invertible, the solution can be shown to be $\alpha^* = k_{ZZ}^{-1} f_Z$ with $f_Z = \big(f(z_1), ..., f(z_m)\big)^\top \in \mathbb{R}^m$. Thus the projection of f is given by

$$P_M\big(f\big) = k_Z(\cdot)^\top k_{ZZ}^{-1} f_Z. \tag{14}$$

We can use the orthogonal projection P_M to define an approximate kernel. Note that by definition,

$$k(x, x') = \big\langle k(\cdot, x), k(\cdot, x')\big\rangle_{\mathcal{H}_k}, \quad x, x' \in \mathcal{X}.$$

We then define a new kernel $q : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as the inner product between the projections of $k(\cdot, x)$ and $k(\cdot, x')$ onto the subspace M:

$$\begin{aligned} q(x, x') &:= \big\langle P_M\big(k(\cdot, x)\big), P_M\big(k(\cdot, x')\big)\big\rangle_{\mathcal{H}_k} \\ &= \big\langle k_Z(x)^\top k_{ZZ}^{-1} k_Z(\cdot), k_Z(x')^\top k_{ZZ}^{-1} k_Z(\cdot)\big\rangle_{\mathcal{H}_k} \\ &= k_Z(x)^\top k_{ZZ}^{-1} k_Z(x'), \quad x, x' \in \mathcal{X}. \end{aligned} \tag{15}$$

Since q is a positive definite kernel, it induces its own RKHS $\mathcal{H}_q$. As the following lemma shows, this RKHS $\mathcal{H}_q$ is nothing but the subspace M, with the inner product of $\mathcal{H}_q$ being identical to that of the original RKHS $\mathcal{H}_k$. The proof can be found in Appendix A.

Lemma 3 *Let $Z = (z_1, \ldots, z_m) \in \mathcal{X}^m$ be such that the kernel matrix k_{ZZ} is invertible. Then we have $M = \mathcal{H}_q$ as a set, and*

$$\langle f, g \rangle_{\mathcal{H}_q} = \langle f, g \rangle_{\mathcal{H}_k}, \quad \forall f, g \in M = \mathcal{H}_q.$$

In particular, Lemma 3 implies that

$$\|f\|_{\mathcal{H}_q} = \|f\|_{\mathcal{H}_k}, \quad \forall f \in \mathcal{H}_q = M$$

By using this identity and $\mathcal{H}_q = M$ in the Nyström KRR objective function (12), we immediately have the following characterization of the Nyström approximation in (13).

Theorem 4 *Let $X := (x_1, ..., x_n) \in \mathcal{X}^n$ and $y := (y_1, ..., y_n)^\top \in \mathbb{R}^n$ be given. Let $Z = (z_1, \ldots, z_m) \in \mathcal{X}^m$ be such that the kernel matrix k_{ZZ} is invertible, and $\bar{f}$ be the Nyström approximation $\bar{f}$ of KRR in* (13)*. Then we have*

$$\bar{f} = \operatorname*{arg\,min}_{f \in \mathcal{H}_q} \frac{1}{n} \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 + \lambda \|f\|^2_{\mathcal{H}_q}$$

where q is the approximate kernel defined in (15)*.*

Theorem 4 shows that the Nyström approximation (13) is the solution of the KRR with the approximate kernel q in (15). Note that, by Lemma 3, any $f \in \mathcal{H}_q$ can be written as

$$\begin{aligned} f(x) &= \langle f, q(\cdot, x) \rangle_{\mathcal{H}_q} = \langle f, q(\cdot, x) \rangle_{\mathcal{H}_k} \\ &= \Big\langle f, k_Z(\cdot)^\top k_{ZZ}^{-1} k_Z(x) \Big\rangle_{\mathcal{H}_k} = f_Z^\top k_{ZZ}^{-1} k_Z(x). \end{aligned}$$

The last expression coincides with the expression (14) of the orthogonal projection onto M, and also with the *kernel-based interporator*[4] obtained from noise-free observations $(z_j, f(z_j))_{j=1}^m$. Therefore, each function $\mathcal{H}_q$ is the best approximation of functions in $\mathcal{H}_k$ that pass $(z_j, f(z_j))_{j=1}^m$. In this sense, $\mathcal{H}_q$ consists of functions that approximate the functions in $\mathcal{H}_k$ on the landmark points $z_1, \ldots, z_m$. Hence, Theorem 4 shows that the Nyström approximation is the solution of the KRR where the hypothesis space $\mathcal{H}_q$ consists of such approximate functions.

3.2 Sparse Approximations for GPR

We review here the Sparse Variational Gaussian Process (SVGP) approach by Titsias (2009a) based on a measure-theoretic formulation suggested by Matthews et al. (2016). There have been many works on sparse approximations for scaling up GP-based methods. In a nutshell, there are two common approaches: either the generative model is approximated and inference is performed exactly (Seeger et al., 2003; Snelson and Ghahramani, 2006, 2007) or the generative model is left unaltered and inference is done approximately

4. This corresponds to the KRR estimator (4) with $\lambda := 0$, $X := Z$ and $y := f_Z$. Setting $\lambda = 0$ leads to the *minimum-norm interpolation* in the RKHS; see e.g. Kanagawa et al. (2018, Section 3.2) and references therein.

(Csató and Opper, 2002; Titsias, 2009a). In this work, we mainly focus on the SVGP approximation by Titsias (2009a), which is the latter approach, but we also discuss the *Deterministic Training Conditional (DTC)* approximation (Seeger et al., 2003), one of the former approaches, as this will provide us a more insight about the SVGP approximation. We refer to Bauer et al. (2016) for a systematic comparison of the two approaches.

Since we focus on the basic framework of Titsias (2009a) and its comparison to the kernel counterpart in a regression setting, we do not discuss sparse variational GP approaches to the classification problem (Hensman et al., 2015a) and other (more recent) developments (e.g., Hensman et al., 2015b, 2018; Dutordoir et al., 2020; Adam et al., 2020; Shi et al., 2020; Rossi et al., 2021; Tran et al., 2021). See e.g., Leibfried et al. (2020) for an overview over variational GP approaches.

We first recall the setting of GPR using a measure-theoretic notation. As before, let $(x_i, y_i)_{i=1}^n \subset \mathcal{X} \times \mathbb{R}$ be training data and let $X = (x_1, \dots, x_n) \in \mathcal{X}^n$ and $y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$. For simplicity, we assume the zero prior mean function, $m(x) = 0$. We denote by $\mathbb{P}$ the probability measure of a Gaussian process $F \sim GP(0, k)$. For any finite set of points $D := (d_1, \dots, d_\ell) \in \mathcal{X}^\ell$ with $\ell \in \mathbb{N}$, let $\mathbb{P}_D$ be the corresponding distribution of $F_D := (F(d_1), \dots, F(d_\ell))^\top$ on $\mathbb{R}^\ell$, which is $\mathbb{P}_D = \mathcal{N}(0, k_{DD})$ by definition.

3.2.1 Variational Family

We first introduce a variational family of probability measures of functions on $\mathcal{X}$, from which we search for a computationally tractable approximation of the GP posterior $\mathbb{P}^{F|y} = GP(\bar{m}, \bar{k})$ in (9). Let $m \in \mathbb{N}$ be fixed, and Γ be a set of *variational parameters* defined by

$$\Gamma := \{\nu := (Z, \mu, \Sigma) \mid Z := (z_1, \dots, z_m) \in \mathcal{X}^m, k_{ZZ} \text{ is invertible}, \\ \mu \in \mathbb{R}^m, \Sigma \in \mathbb{R}^{m \times m}_{\succ 0}\}$$

where $\mathbb{R}^{m \times m}_{\succ 0}$ stands for symmetric and positive definite matrices in $\mathbb{R}^{m \times m}$. The points $Z = (z_1, \dots, z_m)$ are the so-called *inducing inputs*, based on which we approximate the posterior GP. On the other hand, μ and Σ are parameters for the distribution of function values at $z_1, \dots, z_m$.

We then define a variational family

$$\mathcal{Q}_\Gamma := \{\mathbb{Q}^\nu \mid \nu \in \Gamma\}$$

as a set of Gaussian processes parametrized by the tuple $\nu = (Z, \mu, \Sigma)$ defined as follows:

$$\mathbb{Q}^\nu := GP(m^\nu, k^\nu), \tag{16}$$

$$m^\nu(x) := k_Z(x)^\top k_{ZZ}^{-1} \mu, \tag{17}$$

$$k^\nu(x, x') := k(x, x') - k_Z(x)^\top k_{ZZ}^{-1} k_Z(x') \\ + k_Z(x)^\top k_{ZZ}^{-1} \Sigma k_{ZZ}^{-1} k_Z(x'). \tag{18}$$

Each variational distribution (16) is defined so as to have the following properties, where $F^\nu \sim GP(m^\nu, k^\nu)$ denotes the corresponding GP sample function:

1. The function values[5] $F_Z^\nu := (F^\nu(z_1), \ldots, F^\nu(z_m))^\top \in \mathbb{R}^m$ at the inducing inputs $z_1, \ldots, z_m$ follow the Gaussian distribution with mean vector $\mu \in \mathbb{R}^m$ and covariance matrix $\Sigma \in \mathbb{R}^{m\times m}_{\succ 0}$, i.e., $F_Z^\nu \sim \mathcal{N}(\mu, \Sigma)$. We denote by $\mathbb{Q}_Z^\nu$ by the distribution of F_Z^ν i.e., $\mathbb{Q}_Z^\nu = \mathcal{N}(\mu, \Sigma)$.

2. The conditional distribution of the process F^ν given $(z_i, F^\nu(z_i))_{i=1}^m$ is identical to the conditional distribution of $F \sim GP(m, k)$ given $(z_i, F(z_i))_{i=1}^m$:

$$F^\nu \mid (z_i, F^\nu(z_i))_{i=1}^m \overset{d}{=} F \mid (z_i, F(z_i))_{i=1}^m. \tag{19}$$

In fact, starting from the expressions (16) (17) (18), one can check that $F^\nu \sim GP(m^\nu, k^\nu)$ satisfies the above two requirements as follows:

1. $m_Z^\nu = \mu$ and $k_{ZZ}^\nu = \Sigma$, and thus $F_Z^\nu \sim \mathcal{N}(\mu, \Sigma)$;

2. By using (7) and (8) with $m := m^\nu$, $k := k^\nu$, $X := Z$, $Y := F_Z^\nu$ and $\sigma^2 := 0$,[6] the conditional distribution of F^ν given $(z_i, F^\nu(z_i))_{i=1}^m$ is given by the Gaussian process

$$F^\nu \mid (z_i, F^\nu(z_i))_{i=1}^m \sim GP(\tilde{m}^\nu, \tilde{k}^\nu),$$

with mean function $\tilde{m}^\nu : \mathcal{X} \to \mathbb{R}$ and covariance function $\tilde{k}^\nu : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ given by

$$\begin{aligned} \tilde{m}^\nu(x) &:= m^\nu(x) + k_Z^\nu(x)^\top (k_{ZZ}^\nu)^{-1}(F_Z - m_Z^\nu) \\ &= k_Z(x)^\top k_{ZZ}^{-1} F_Z, \end{aligned} \tag{20}$$

$$\begin{aligned} \tilde{k}^\nu(x, x') &:= k^\nu(x, x') - k_Z^\nu(x)^\top (k_{ZZ}^\nu)^{-1} k_Z^\nu(x') \\ &= k(x, x') - k_Z(x)^\top k_{ZZ}^{-1} k_Z(x') \end{aligned} \tag{21}$$

Since the expressions (20) and (21) are respectively the mean function and covariance function of the conditional distribution of F given $(z_i, F(z_i))_{i=1}^m$ i.e., $F \mid (z_i, F(z_i))_{i=1}^m \sim GP(\tilde{m}, \tilde{k})$, we have the distributional identity (19).

3.2.2 Evidence Lower Bound and Optimal Vartional Parameters

The aim of variational inference is to obtain a distribution $\mathbb{Q}^{\nu^*}$ from the variational family $\mathcal{Q}_\Gamma$ that best approximates the posterior measure $\mathbb{P}^{F|y}$ in terms of the Kullback-Leibler (KL) divergence, without explicitly computing the posterior. That is, we want to compute $\nu^* \in \Gamma$ such that

$$\nu^* \in \arg\min_{\nu\in\Gamma} \; KL(\mathbb{Q}^\nu \parallel \mathbb{P}^{F|y}). \tag{22}$$

where $KL(\mathbb{Q}^\nu \parallel \mathbb{P}^{F|y})$ is the KL divergence between $\mathbb{Q}^\nu$ and $\mathbb{P}^{F|y}$ defined by

$$KL(\mathbb{Q}^\nu || \mathbb{P}^{F|y}) := \int \log\left(\frac{d\mathbb{Q}^\nu}{d\mathbb{P}^{F|y}}(f)\right) d\mathbb{Q}^\nu(f).$$

5. Note that F_Z^ν is usually called *inducing variables* and is denoted with symbol $\mathbf{u}$ in the literature.
6. The case $\sigma^2 = 0$ is well defined as long as the kernel matrix k_Z is invertible. In this case, the regression problem becomes that of *interpolation*, i.e., function approximation from noise-free observations. See e.g. Kanagawa et al. (2018) and references therein.

with $\frac{d\mathbb{Q}^\nu}{d\mathbb{P}^{F|y}}$ being the Radon-Nikodym derivative of $\mathbb{Q}_\nu$ with respect to $\mathbb{P}^{F|y}$, which exists by the construction of $\mathbb{Q}^\nu$ (Matthews et al., 2016, Section 3.3).

Matthews et al. (2016, Eq. 15) show that this KL divergence can be written as

$$KL\big(\mathbb{Q}^\nu \parallel \mathbb{P}^{F|y}\big) = \log p(y) - \mathcal{L}(\nu), \tag{23}$$

where $p(y)$ is the marginal likelihood, or the *Evidence*, of observing $y = (y_1, \ldots, y_n)^\top$ under the prior $F \sim GP(m, k)$ and the likelihood model $y_i \sim \mathcal{N}(F(x_i), \sigma^2)$, while $\mathcal{L}(\nu)$ is the *Evidence Lower Bound (ELBO)* defined as

$$\mathcal{L}(\nu) := -KL\big(\mathbb{Q}^\nu_Z \parallel \mathbb{P}_Z\big) + \mathbb{E}_{F^\nu \sim \mathbb{Q}^\nu}\big[\log p(y|F^\nu_X)\big], \tag{24}$$

where

- $KL\big(\mathbb{Q}^\nu_Z \parallel \mathbb{P}_Z\big)$ is the standard KL divergence between $\mathbb{Q}^\nu_Z = \mathcal{N}(\mu, \Sigma)$, which is the marginal distribution of $F^\nu_Z \in \mathbb{R}^m$ of the parameterized Gaussian process $F^\nu \sim \mathbb{Q}^\nu = GP(m^\nu, k^\nu)$ in (16), and $\mathbb{P}_Z = \mathcal{N}(0, k_{ZZ})$, which is the marginal distribution of $F_Z \in \mathbb{R}^m$ of the prior Gaussian process $F \sim \mathbb{P} = GP(0, k)$:

$$\begin{aligned} KL\big(\mathbb{Q}^\nu_Z \parallel \mathbb{P}_Z\big) &= \int_{\mathbb{R}^m} \log\left(\frac{d\mathbb{Q}^\nu_Z}{d\mathbb{P}_Z}(f_Z)\right) d\mathbb{Q}^\nu_Z(f_Z) \\ &= \frac{1}{2}\left(\mathrm{tr}(k_{ZZ}^{-1}\Sigma) + \mu^\top k_{ZZ}^{-1}\mu - m + \log\left(\frac{\det k_{ZZ}}{\det \Sigma}\right)\right), \end{aligned} \tag{25}$$

  where the last identity is the well-known expression of the KL divergence between multivariate Gaussian densities (see, e.g., Appendix A.5 of Rasmussen and Williams 2006).

- $\mathbb{E}_{F^\nu \sim \mathbb{Q}^\nu}[\log p(y|F^\nu_X)]$ is the *marginal log likelihood* of observing $y = (y_1, \ldots, y_n)^\top$ under the likelihood model $y_i = F^\nu(x_i) + \varepsilon_i$ with independent $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and the parametrized process $F^\nu \sim \mathbb{Q}^\nu$:

$$\mathbb{E}_{F^\nu \sim \mathbb{Q}^\nu}[\log p(y|F^\nu_X)] = -n\log(\sqrt{2\pi\sigma^2}) - \mathbb{E}_{F^\nu \sim \mathbb{Q}^\nu} \sum_{i=1}^n \frac{(y_i - F^\nu(x_i))^2}{2\sigma^2} \tag{26}$$

  where $p(y|F^\nu_X) := \mathcal{N}(y; F^\nu_X, \sigma^2 I_n)$ is the Gausian density of the likelihood function (6).

Since the marginal likelihood $p(y)$ under the original GP prior does not depend on the variational parameters ν, the minimization of the KL divergence (23) is equivalent to the maximization of the ELBO $\mathcal{L}(\nu)$ in (24). Titsias (2009a, Eq.(10)) show that,[7] for fixed inducing points Z, the optimal parameters μ^* and Σ^* that maximize the ELBO are given analytically as

$$\mu^* := k_{ZZ}(\sigma^2 k_{ZZ} + k_{ZX}k_{XZ})^{-1}k_{ZX}y \tag{27}$$

$$\Sigma^* := k_{ZZ}(k_{ZZ} + \sigma^{-2}k_{ZX}k_{XZ})^{-1}k_{ZZ} \tag{28}$$

7. See Appendix A of Titsias (2009b) or Hensman et al. (2013, Section 2) for the derivation

and the resulting ELBO, denoted by $\mathcal{L}^*$, is

$$\mathcal{L}^* = -\frac{1}{2}\log\det(q_{XX} + \sigma^2 I_n) - \frac{1}{2}y^\top(q_{XX} + \sigma^2 I)^{-1}y \\ - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\mathrm{tr}(k_{XX} - q_{XX}), \tag{29}$$

where q is the approximate kernel in (15). Inserting these expressions in the definition of the variational distribution (16), the optimal variational approximation (for fixed inducing points Z) is given by $GP(m^*, k^*)$ with:

$$m^*(x) := k_Z(x)^\top(\sigma^2 k_{ZZ} + k_{ZX}k_{XZ})^{-1}k_{ZX}y \tag{30}$$

$$k^*(x,x') := k(x,x') - k_Z(x)^\top k_{ZZ}^{-1}k_Z(x') \\ + k_Z(x)^\top(k_{ZZ} + \sigma^{-2}k_{ZX}k_{XZ})^{-1}k_Z(x') \tag{31}$$

The computational complexity of obtaining the mean function m^* and the covariance function k^* is $\mathcal{O}(nm^2 + m^3)$, which can be much smaller than the complexity $\mathcal{O}(n^3)$ of the exact posterior as long as the number of inducing points m is much smaller than the training data size n.

The ELBO (29) with optimal μ^* and Σ^* is a key quantity, as it can be used i) as a criterion for optimizing the inducing inputs $z_1, \ldots, z_m$ and ii) for theoretically analyzing the quality of variational approximation.

3.3 Equivalence between the Nyström and SVGP Approximations

We now focus on the relations between the Nyström method for KRR and the variational approximation for GPR. Like Theorem 1, our result below summarizes the equivalence between the predictors in the two approaches. It directly follows from the corresponding expressions (13) and (30).

Theorem 5 *Let k be a kernel, and suppose that data $(x_i, y_i)_{i=1}^n \subset \mathcal{X} \times \mathbb{R}$ are given. Let $Z = (z_1, \ldots, z_m) \in \mathcal{X}^m$ be fixed inducing inputs such that the kernel matrix $k_{ZZ} = (k(z_i, z_j))_{i,j=1}^m \in \mathbb{R}^{m\times m}$ is invertible.*

- *Let m^* be the mean function* (30) *of the variational posterior for GPR with the prior being the zero-mean Gaussian process $F \sim GP(0, k)$.*
- *Let $\bar{f}$ be the Nyström approximation* (13) *of KRR performed on the RKHS $\mathcal{H}_k$.*

Then if $\sigma^2 = n\lambda$, we have $m^* = \bar{f}$.

Theorem 5 shows that the two approximate regressors are identical, although their derivations are (seemingly) quite different. The condition $n\lambda = \sigma^2$ is the same as that required for the equivalence between KRR and GPR in Theorem 1.

The question now is why there is this equivalence in the two ways of approximation. To investigation this, we first inspect closely the ELBO (24) of the variational approximation, to reveal the source of the equivalence. In the expression (24), the first term $-KL(\mathbb{Q}_Z^\nu \| \mathbb{P}_Z)$ can be understood as a regularizer, constraining the variational approximation $\mathbb{Q}^\nu$ not to

deviate too much from the GP prior $\mathbb{P} = GP(0, k)$ on the inducing points $Z = (z_1, \ldots, z_n)$. The second term $\mathbb{E}_{F^\nu \sim \mathbb{Q}^\nu}[\log p(y|F_X^\nu)]$ represents the quality of predictions evaluated at the training data $(x_i, y_i)_{i=1}^n$. Therefore, the ELBO may be understood as a regularized empirical risk for the (distributional) regression problem.

The following result reveals the source of the equivalence between the two approaches. The proof can be found in Appendix A.2.

Theorem 6 *Let $\nu = (Z, \mu, \Sigma) \in \mathcal{X}^m \times \mathbb{R}^m \times \mathbb{R}^{m\times m}_{>0}$ be such that the kernel matrix $k_{ZZ} \in \mathbb{R}^{m\times m}$ is invertible, and let $\mathcal{L}(\nu)$ be the ELBO in (24). Then we have*

$$-2\sigma^2\mathcal{L}(\nu) = \sum_{i=1}^n \left(y_i - k_Z(x_i)^\top k_{ZZ}^{-1}\mu\right)^2 + \sigma^2 \mu k_{ZZ}^{-1}\mu \tag{32}$$

$$+ \sum_{i=1}^n k_Z(x_i)^\top k_{ZZ}^{-1}\Sigma k_{ZZ}^{-1} k_Z(x_i) \tag{33}$$

$$+ \sigma^2\left(\mathrm{tr}(k_{ZZ}^{-1}\Sigma) + \log(\det k_{ZZ}/\det\Sigma) - m\right) \tag{34}$$

$$+ \sum_{i=1}^n \left(k(x_i, x_i) - k_Z(x_i)^\top k_{ZZ}^{-1} k_Z(x_i)\right). \tag{35}$$

An important consequence of Theorem 6 is that, given Z being fixed, the optimization of μ and Σ can be decoupled, since there is no term that depends on both of μ and Σ. Consider the term (32), which does not depend on Σ but on μ and Z. Since k_{ZZ} is invertible, k_{ZZ} as an operator defines a one-to-one mapping from $\mathbb{R}^m$ to $\mathbb{R}^m$. Thus we can consider the reparametrization $\alpha := k_{ZZ}^{-1}\mu$. With this, the μ-dependent term (32) can be written as

$$\begin{aligned}&\sum_{i=1}^n (y_i - k_Z(x_i)^\top\alpha)^2 + \sigma^2\alpha^\top k_{ZZ}\alpha \\ &= \sum_{i=1}^n (y_i - f_m(x_i))^2 + \sigma^2\|f_m\|_{\mathcal{H}_k}^2,\end{aligned} \tag{36}$$

where we defined

$$f_m := \sum_{j=1}^m \alpha_j k(\cdot, z_j) \in M \subset \mathcal{H}_k,$$

with $M = \mathrm{span}(k(\cdot, z_1), \ldots, k(\cdot, z_m))$ being the subspace in (11). Notice that (36) is the objective function of KRR if $\sigma^2 = n\lambda$, evaluated for the function f_m from the subspace M. In other words, the maximization of the ELBO with respect to μ is equivalent to the Nyström KRR (12) with $\sigma^2 = n\lambda$. This is summarized in the following corollarly.

Corollary 7 *Let $\nu = (Z, \mu, \Sigma) \in \mathcal{X}^m \times \mathbb{R}^m \times \mathbb{R}^{m\times m}_{\succ 0}$ be such that the kernel matrix $k_{ZZ} \in \mathbb{R}^{m\times m}$ is invertible, and let $\mathcal{L}(\nu)$ be the ELBO in (24). For any fixed Z and Σ, let*

$$\mu^* = \arg\max_{\mu\in\mathbb{R}^m} L(\nu)$$

and define $\alpha^* = k_{ZZ}^{-1}\mu^*$. *Then* $f^* = \sum_{j=1}^m \alpha_j^* k(\cdot, z_j)$ *is the solution of the Nyström KRR*

$$f^* = \operatorname*{arg\,min}_{f \in M} \sum_{i=1}^n (y_i - f(x_i))^2 + \sigma^2 \|f\|_{\mathcal{H}_k}^2,$$

where $M = \operatorname{span}(k(\cdot, z_1), \ldots, k(\cdot, z_m)) \subset \mathcal{H}_k$.

Thus, Theorem 6 shows the origin of the equivalence between the SVGP approach of Titsias (2009a) for GPR the Nyström approximation for KRR.

Note that $\mu = (\mu_1, \ldots, \mu_m)^\top$ can be interpreted as noise-free observations at the inducing points $z_1, \ldots, z_m$. In fact, $f_m = \sum_{j=1}^m \alpha_j k(\cdot, z_j)$ with $\alpha = (\alpha_1, \ldots, \alpha_m)^\top = k_{ZZ}^{-1}\mu$, which is the mean function (17) of the variational distribution, is the kernel interpolator[8] obtained from training data $(z_j, \mu_j)_{j=1}^m$. Thus, the specification of $\mu_1, \ldots, \mu_m$ can be understood as the specification of pseudo observations at $z_1, \ldots, z_m$. This interpretation is consistent with the fact that μ is the mean vector of "inducing variables" $F_Z^\nu = (F^\nu(z_1), \ldots, F^\nu(z_m))^\top$.

3.4 Relation to the Deterministic Training Conditional

Before proceeding further, we mention here the equivalence between the Nyström approximation and the *Deterministic Training Conditional (DTC)*, a classic sparse approximation approach to GPR by Seeger et al. (2003); see also Quiñonero-Candela and Rasmussen (2005, Section 5). This discussion will be useful in our investigation of the connections between the Nyström and SVGP.

With DTC, one performs GPR using the approximate kernel q in (15) for the prior, $F \sim GP(0, q)$, instead of the original kernel k. Given observations y, the resulting GP posterior is $F|y \sim GP(\bar{m}, \bar{k})$ with the posterior mean function $\bar{m}$ and posterior covariance function $\bar{k}$ given by

$$\begin{aligned}
\bar{m}(x) &= q_X(x)^\top (q_{XX} + \sigma^2 I_n)^{-1} y \\
&= k_Z(x)^\top (\sigma^2 k_{ZZ} + k_{ZX} k_{XZ})^{-1} k_{ZX} y, \\
\bar{q}(x, x') &= q(x, x') - q_X(x)(q_{XX} + \sigma^2 I_n)^{-1} q_X(x') \\
&= k_Z(x)^\top (k_{ZZ} + \sigma^{-2} k_{ZX} k_{XZ})^{-1} k_Z(x'). \qquad (37)
\end{aligned}$$

Notice that the posterior mean function $\bar{m}$ here and the Nyström approximation (13) are the same if $\sigma^2 = n\lambda$. In fact, this identity immediately follows from Theorem 1 on the equivalence between the KRR and GPR and Theorem 4 on the formulation of the Nyström as the KRR with the approximate kernel q.

3.5 Geometric Interpretation of the Variational Covariance Function

To further investigate the connections between the sparse approximation methods, we make a geometric interpretation for the covariance function (18) of a variational distribution

8. This corresponds to the KRR estimator (4) with $\lambda = 0$, $X = Z$, and $y = \mu$, which is well-defined as long as the kernel matrix k_{ZZ} is invertible. In this case, the solution (4) is that of *minimum-norm interpolation*: $\min_{f \in \mathcal{H}_k} \|f\|_{\mathcal{H}_k}$ subject to $f(z_j) = \mu_j$, $j = 1, \ldots, m$. See e.g. Kanagawa et al. (2018) and references therein.

$\mathbb{Q}^\nu = GP(m^\nu, k^\nu)$:

$$k^\nu(x,x') = k(x,x') - k_Z(x)^\top k_{ZZ}^{-1} k_Z(x') \tag{38}$$
$$+ k_Z(x)^\top k_{ZZ}^{-1} \Sigma k_{ZZ}^{-1} k_Z(x'). \tag{39}$$

Part (38) is the posterior covariance function of $F \sim GP(0,k)$ given noise-free observations $(z_j, F(z_j))_{j=1}^m$ and can be written as

$$\begin{aligned} &k(x,x') - k_Z(x)^\top k_{ZZ}^{-1} k_Z(x') \\ &= k(x,x') - q(x,x') \\ &= k(x,x') - \left\langle P_M(k(\cdot,x)), P_M(k(\cdot,x')) \right\rangle_{\mathcal{H}_k} \\ &= \left\langle k(\cdot,x) - P_M(k(\cdot,x)),\ k(\cdot,x') - P_M(k(\cdot,x')) \right\rangle_{\mathcal{H}_k}, \end{aligned} \tag{40}$$

where q is the approximate kernel (15) and $P_M : \mathcal{H}_k \to M$ is the orthogonal projection (14) onto the subspace $M = \operatorname{span}(k(\cdot,z_1), \ldots, k(\cdot,z_m))$.

Note that $k(\cdot,x) - P_M(k(\cdot,x))$ is the *residual* of the orthogonal projection $P_M(k(\cdot,x))$. Therefore geometrically, part (38) is the inner product between the two residuals. Intuitively, (38) represents the part of the kernel k that is not captured by the approximate kernel q or, equivalently, by the subspace M.

We next consider part (39). To this end, let $\mathcal{H}_{k^\nu}$ be the RKHS of k^ν, M^ν be the subspace spanned by $k^\nu(\cdot,z_1), \ldots, k^\nu(\cdot,z_m)$:

$$M^\nu := \operatorname{span}(k^\nu(\cdot,z_1), \ldots, k^\nu(\cdot,z_m)) \subset \mathcal{H}_{k^\nu}, \tag{41}$$

and $P_{M^\nu} : \mathcal{H}_{k^\nu} \to M^\nu$ be the orthogonal projection onto M^ν:

$$P_{M^\nu}(f) := \operatorname*{arg\,min}_{g \in M^\nu} \|f - g\|_{\mathcal{H}_{k^\nu}} = k_Z(\cdot)^\top (k_{ZZ}^\nu)^{-1} f_Z,$$

for any $f \in \mathcal{H}_{k^\nu}$. (See (14)) It can be easily verified that

$$k_{ZZ}^\nu = \Sigma, \quad k_Z^\nu(x) = \Sigma k_{ZZ}^{-1} k_Z(x), \quad \forall x \in \mathcal{X}.$$

We can now rewrite part (39) as

$$\begin{aligned} k_Z(x)^\top k_{ZZ}^{-1} \Sigma k_{ZZ}^{-1} k_Z(x') &= k_Z^\nu(x)^\top (k_{ZZ}^\nu)^{-1} k_Z^\nu(x') \\ &= \left\langle P_{M^\nu}(k^\nu(\cdot,x)), P_{M^\nu}(k^\nu(\cdot,x')) \right\rangle_{\mathcal{H}_{k^\nu}} =: q^\nu(x,x') \end{aligned} \tag{42}$$

Thus, part (39) is the inner product in $\mathcal{H}_{k^\nu}$ between the projections $P_{M^\nu}(k^\nu(\cdot,x))$ and $P_{M^\nu}(k^\nu(\cdot,x'))$ on M^ν. Notice that this structure is the same as the definition of the approximate kernel q in (15), but is given with the kernel k^ν parametrized by Σ and Z. Therefore, part (39) can be understood as the approximation, denoted by q^ν, of k^ν using the inducing inputs $z_1, \ldots, z_m$.

3.6 Further Investigation of the ELBO

We now investigate further the expression of ELBO in Theorem 6, focusing on the terms (33) (34) that depend on the covariance matrix Σ, and the term (35) that depends only on Z.

Using (42), the term (33) can be written as

$$\sum_{i=1}^{n} k_Z(x_i)^\top k_{ZZ}^{-1} \Sigma k_{ZZ}^{-1} k_Z(x_i) = \sum_{i=1}^{n} \|P_{M^\nu}(k^\nu(\cdot, x_i))\|^2_{\mathcal{H}_{k^\nu}} = \sum_{i=1}^{n} q^\nu(x_i, x_i).$$

Thus, this term represents the variances (or uncertainties) of a GP with the approximate kernel q^ν at training inputs $x_1, \ldots, x_n$. On the other hand, the term (34) can be written as

$$\sigma^2 \left(\mathrm{tr}(k_{ZZ}^{-1}\Sigma) + \log(\det k_{ZZ}/\det\Sigma) - m\right) = \sigma^2 KL(\mathcal{N}(0,\Sigma) \parallel \mathcal{N}(0, k_{ZZ})).$$

This is the KL divergence between the two zero-mean multivariate Gaussians with covariance matrix Σ and k_{ZZ}. Thus, the sum of (33) and (34) in ELBO is

$$\sum_{i=1}^{n} q^\nu(x_i, x_i) + \sigma^2 KL(\mathcal{N}(0,\Sigma) \parallel \mathcal{N}(0, k_{ZZ})) \tag{43}$$

The first term represent uncertainties at training input $x_1, \ldots, x_n$, and the second term can be interpreted as a regularizer that encourages Σ not to deviate from prior covariance matrix k_{ZZ} too much. Thus intuitively, the optimal Σ^*, which minimizes the sum, is such that the uncertainties at $x_1, \ldots, x_n$ are small while Σ is not very different from k_{ZZ}. This interpretation is consistent with the fact that Σ^* is the one that produces a best approximation to the GP posterior given observations $(x_i, y_i)_{i=1}^n$.

The last term (35) depends only on Z, and therefore it only works when optimizing Z. This term represents the posterior variances of $F \sim GP(0, k)$ at $x_1, \ldots, x_n$ given noise-free observations $(z_j, F(z_j))_{j=1}^m$. By (40) it can be written as

$$\begin{aligned}
&\sum_{i=1}^{n} \left(k(x_i, x_i) - k_Z(x_i)^\top k_{ZZ}^{-1} k_Z(x_i)\right) \\
&= \sum_{i=1}^{n} \left(k(x_i, x_i) - q(x_i, x_i)\right) = \sum_{i=1}^{n} \|k(\cdot, x_i) - P_M(k(\cdot, x_i))\|^2_{\mathcal{H}_k}
\end{aligned}$$

Geometrically, this is the sum of the squared length of residuals $k(\cdot, x_i) - P_M(k(\cdot, x_i))$, and becomes small when the subspace $M = \mathrm{span}(k(\cdot, z_1), \ldots, k(\cdot, z_m))$ approximates well the feature representations $k(\cdot, x_1), \ldots, k(\cdot, x_n)$ of training inputs $x_1, \ldots, x_n$.

3.7 RKHS Interpretation of the Variational Posterior Covariance Function

We present here an RKHS interpretation of the posterior covariance function (31) of the variational distribution with optimal Σ^* in (28):

$$k^*(x, x') = k(x, x') - k_Z(x)^\top k_{ZZ}^{-1} k_Z(x') \tag{44}$$

$$+ k_Z(x)^\top (k_{ZZ} + \sigma^{-2} k_{ZX} k_{XZ})^{-1} k_Z(x'). \tag{45}$$

As mentioned, the first term (44) is the posterior covariance function of $F \sim GP(0, k)$ given noise-free observations $(z_j, F(z_j))_{j=1}^m$. It is the same as the corresponding term (38) of the covariance function of a variational GP with generic Σ, as this term does not depend on Σ. The second term (45) is given by the approximate kernel in (42) with the optimal Σ^* in (28), which minimizes the KL-regularized objective function (43).

Interestingly, the term (45) coincides with the posterior covariance function (37) of the DTC approximation. We summarize these observations in the following proposition:

Proposition 8 *Let Σ^* be the minimizer of the KL-regularized objective function* (43)*, and let q^ν be the resulting approximate kernel in* (42)*. Then we have $q^\nu = \bar{q}$, where $\bar{q}$ is the posterior covariance function* (37) *of the DTC approximation.*

Thus, the covariance function (31) of the variational posterior is the sum of two posterior covariance functions:[9]: (i) one is that of $F \sim GP(0, k)$ given noise-free observations $(z_j, F(z_j))_{j=1}^m$, and (ii) the other is that of $F \sim GP(0, q)$ given noisy training data $(x_i, y_i)_{i=1}^n$, where q is the approximate kernel (15) based on inducing inputs $z_1, \ldots, z_m$.

$$k^*(x, x') = \underbrace{k(x, x') - k_Z(x)^\top k_{ZZ}^{-1} k_Z(x')}_{(i)} + \underbrace{\bar{q}(x, x')}_{(ii)}.$$

By using Theorem 2, we can now provide the following RKHS interpretation of the variational posterior *variance* function (i.e., (31) with $x = x'$). For part (i) with $x = x'$, we have by Theorem 2, for $x \notin \{z_1, \ldots, z_m\}$,

$$k(x, x) - k_Z(x)^\top k_{ZZ}^{-1} k_Z(x) = (\sup_{\|f\|_{\mathcal{H}_k} \leq 1} \{f(x) - k_Z(x)^\top k_{ZZ}^{-1} f_Z\})^2.$$

The is the worst case error of kernel interpolation from noise-free observations $(z_j, f(z_j))_{j=1}^m$, where f is from the unit ball in $\mathcal{H}_k$.

For part (ii) with $x = x'$, define an augmented kernel

$$q^\sigma(x, x') = q(x, x') + \sigma^2 \mathbb{1}\{x = x'\} \tag{46}$$

and let $\mathcal{H}_{q^\sigma}$ be its RKHS. Then, again by Theorem 2, for $x \notin \{x_1, \ldots, x_n\}$ we have

$$\bar{q}(x, x) + \sigma^2 = (\sup_{\|h\|_{\mathcal{H}_{q^\sigma}} \leq 1} \{h(x) + q_X(x)^\top (q_{XX} + \sigma^2 I_n)^{-1} h_X\})^2.$$

Note that the expression $q_X(x)^\top (q_{XX} + \sigma^2 I_n)^{-1} h_X$ is the solution to the KRR for $\sigma^2 = n\lambda$ with kernel q and training data $(x_i, h(x_i))_{i=1}^n$. By Theorem 4, this is equal to the Nyström approximation (13) and thus

$$q_X(x)^\top (q_{XX} + \sigma^2 I_n)^{-1} h_X = k_Z(x)^\top (\sigma^2 k_{ZZ} + k_{ZX} k_{XZ})^{-1} k_{ZX} h_X$$

9. Therefore, $k^*(x, x') - \bar{q}(x, x') = k(x, x') - k_Z(x)^\top k_{ZZ}^{-1} k_Z(x')$. From this expression, it follows that $k^* - \bar{q}$ is a positive definite kernel. In particular, this implies that $k^*(x, x) \geq \bar{q}(x, x)$ for all $x \in \mathcal{X}$. In this sense, the uncertainty estimate of the SVGP approach is more conservative than the DTC, as the posterior variance $k^*(x, x)$ is always larger or equal to that of the DTC $\bar{q}(x, x)$.

Hence, part (ii) with $x = x'$ is the worst case error of the Nyström KRR predicting $h(x)$ using data $(x_i, h(x_i))_{i=1}^n$, where h is from the unit ball in the RKHS $\mathcal{H}_{q^\sigma}$ and may be interpreted as a noisy version of a certain function from the RKHS $\mathcal{H}_q$ of the approximate kernel q.

To summarize, we have the following RKHS interpretation the variational posterior variance:

Theorem 9 *Let k be a kernel with RKHS $\mathcal{H}_k$, and q^σ be the augmented approximate kernel defined in (46) with RKHS $\mathcal{H}_{q^\sigma}$. Suppose that data $(x_i, y_i)_{i=1}^n \subset \mathcal{X} \times \mathbb{R}$ are given, and that $Z = (z_1, \ldots, z_m) \in \mathcal{X}^m$ are fixed inducing inputs such that the kernel matrix $k_{ZZ} = (k(z_i, z_j))_{i,j=1}^m \in \mathbb{R}^{m \times m}$ is invertible. Let k^* be the covariance function (31) of the variational posterior. Then, for $x \notin \{x_1, \ldots, x_n, z_1, \ldots, z_m\}$, we have*

$$k^*(x,x) + \sigma^2 = \Big(\sup_{\|f\|_{\mathcal{H}_k} \leq 1} \{ f(x) - \underbrace{k_Z(x)^\top k_{ZZ}^{-1} f_Z}_{\textit{Kernel Interpolation}} \} \Big)^2$$
$$+ \Big(\sup_{\|h\|_{\mathcal{H}_{q^\sigma}} \leq 1} \{ h(x) - \underbrace{k_Z(x)^\top (\sigma^2 k_{ZZ} + k_{ZX} k_{XZ})^{-1} k_{ZX} h_X}_{\textit{Nyström KRR}} \} \Big)^2$$

The posterior variance function of the SVGP provides a means for uncertainty quantification, and thus it is important to understand its behaviors. Theorem 9 shows that this posterior variance (plus the noise variance) is equal to the sum of two worst case errors: (a) the worst case error of kernel interpolation in the original RKHS $\mathcal{H}_k$ given noise-free observations at inducing inputs $z_1, \ldots, z_m$, and (b) the worst case error of the Nyström KRR in the RKHS $\mathcal{H}_{q^\sigma}$ of the augmented approximate kernel q^σ given noisy observations at training inputs $x_1, \ldots, x_n$. The first part (a) becomes large when the test input point x is far from inducing inputs $z_1, \ldots, z_m$, and the second part (b) becomes large when the test input x is far from training inputs $x_1, \ldots, x_n$. Note that the part (b) depends on the capacity of the RKHS $\mathcal{H}_{q^\sigma}$, which depends on the inducing inputs $z_1, \ldots z_m$ since q^σ is defined from the approximate kernel q. Therefore, in general, as the number m of inducing inputs increases, the capacity of $\mathcal{H}_{q^\sigma}$ increases and approaches that of the RKHS $\mathcal{H}_{k^\sigma}$ in Theorem 2, and thus the part (b) would also increase.

We have discussed the equivalences and connections between the Nyström and SVGP approximations. In the next section, we focus on their theoretical properties, focusing on the quality of approximation.

4. Connections in Theoretical Properties of Sparse Approximations

We investigate here connections between the theoretical properties of the Nyström and SVGP approximations. The Nyström method provides an approximation to the exact KRR solution, and the SVGP approximates the exact GP posterior. The quality of approximation of either approach depends on the choice of inducing inputs $Z = (z_1, \ldots, z_m)$. We focus here on theoretical error bounds for the approximation quality of either approach, and investigate how they are related.

For the Nyström approximation, researchers have studied various approaches for subsampling inducing inputs $z_1, \ldots, z_m$ from training inputs $x_1, \ldots, x_n$ and their theoretical

properties. These range from uniform subsampling to subsampling methods based on leverage scores (Rudi et al., 2015; Musco and Musco, 2017; Chen and Yang, 2021), determinantal point processes (DPPs) (Li et al., 2016), and to ensemble methods (Kumar et al., 2009, 2012). Theoretical works either quantify a (relative) deviation of the approximate kernel matrix from the exact one and its impact on downstream tasks (e.g., Cortes et al., 2010; Musco and Musco, 2017), or more directly bound the expected loss of the resulting approximate KRR estimator (e.g. Bach, 2013; El Alaoui and Mahoney, 2015; Rudi et al., 2015). On the other hand, for the SVGP approach, Burt et al. (2019, 2020) recently provided a theoretical analysis of its quality of approximation for the first time.

In Section 4.1, we first discuss a fundamental result of Burt et al. (2020, Lemma 3) on bounding the KL divergence between the approximate and true GP posteriors. In Section 4.2, we then study how their theoretical results are related to the approximation properties of the Nyström KRR. To this end, we first show that the "data fit" term in the marginal likelihood of GPR is identical to the regularized empirical risk of the corresponding KRR estimator. Using this, we show that the analysis of Burt et al. (2020) is essentially based on bounding the difference between the regularized empirical risks of the approximate and exact KRR solutions. Thus, many of the theoretical arguments of Burt et al. (2020), such as the analysis of impacts of using certain inducing inputs (e.g., DPPs and leverage scores), can be directly translated to the analysis of the Nyström KRR. On the other hand, our finding suggests that more sophisticated theoretical arguments for the Nyström KRR (e.g. Bach, 2013; El Alaoui and Mahoney, 2015; Rudi et al., 2015; Chen and Yang, 2021) may be used for obtaining sharper bounds for the SVGP approach. This investigation is left for future research.

Moreover, in Section 4.3, we establish a novel error bound for the Nyström KRR in terms of the RKHS distance to the exact KRR. Note that the RKHS distance is stronger than standard error metrics such as the L_2 or L_1 errors. This new error bound is parallel to a fundamental bound of Burt et al. (2020) for the SVGP, and thus the theoretical arguments of Burt et al. (2020) can also be applied to bounding the RKHS error for the Nyström KRR.

Lastly, in Section 4.4, we study a lower bound of the KL divergence for the SVGP approximation by Burt et al. (2020, Lemma 4) by considering an average case performance. We show that, under the same setting, this lower bound leads to a lower bound of the approximation error for the Nyström KRR, by the established equivalence result. This new lower bound may be useful in a further analysis of the performance limit of the Nyström approximation.

4.1 A Fundamental Result of Burt et al. (2020)

We first consider the approximation quality of the SVGP approach. In particular, we study a fundamental result of Burt (Lemma 3), from which many other results in Burt et al. (2020) are derived. As before, let μ^* and Σ^* be the optimal variational parameters in (27) and (28), respectively, and let $Z = (z_1, \ldots, z_m) \in \mathcal{X}^m$ be m inducing inputs such that the kernel matrix k_{ZZ} is invertible. Let $\nu^* := (Z, \mu^*, \Sigma^*)$ and $\mathbb{Q}^{\nu^*} = GP(m^*, k^*)$ be the resulting variational GP posterior with mean function m^* and covariance function k^* in (17) and (18), respectively.

A natural metric of quantifying the approximation quality of $\mathbb{Q}^{\nu^*}$ is the KL divergence to the exact GP posterior $\mathbb{P}^{F|y}$, which is given by (23) with $\nu^* = (Z, \mu^*, \Sigma^*)$

$$KL\big(\mathbb{Q}^{\nu^*} \parallel \mathbb{P}^{F|y}\big) = \log p(y) - \mathcal{L}^*,$$

where $p(y)$ is the marginal likelihood and $\mathcal{L}^*$ is the ELBO in (29). We know that (e.g., Rasmussen and Williams, 2006, Eq. (5.8))

$$\log p(y) = -\frac{1}{2}\log\det(k_{XX} + \sigma^2 I_n) - \frac{1}{2} y^\top (k_{XX} + \sigma^2 I_n)^{-1} y - \frac{n}{2}\log 2\pi$$

Therefore,

$$\begin{aligned}
&2KL\big(\mathbb{Q}^{\nu^*} \parallel \mathbb{P}^{F|y}\big) = 2\log p(y) - 2\mathcal{L}^* \\
&= -\log\det(k_{XX} + \sigma^2 I_n) + \log\det(q_{XX} + \sigma^2 I_n) \qquad (47)\\
&\quad - y^\top (k_{XX} + \sigma^2 I_n)^{-1} y + y^\top (q_{XX} + \sigma^2 I_n)^{-1} y + \frac{1}{\sigma^2}\mathrm{tr}(k_{XX} - q_{XX}) \\
&\leq -y^\top (k_{XX} + \sigma^2 I_n)^{-1} y + y^\top (q_{XX} + \sigma^2 I_n)^{-1} y + \frac{1}{\sigma^2}\mathrm{tr}(k_{XX} - q_{XX}), \qquad (48)
\end{aligned}$$

where the inequality follows from $k_{XX} - q_{XX}$ being positive semi-definite.

Burt et al. (2020, Proof of Lemma 3) proceed to bound the first two terms in (48) as

$$\begin{aligned}
&- y^\top (k_{XX} + \sigma^2 I_n)^{-1} y + y^\top (q_{XX} + \sigma^2 I_n)^{-1} y \\
&\leq \frac{\|y\|^2 \|k_{XX} - q_{XX}\|_{\mathrm{op}}}{\sigma^2(\|k_{XX} - q_{XX}\|_{\mathrm{op}} + \sigma^2)} \leq \frac{\|y\|^2 \mathrm{tr}(k_{XX} - q_{XX})}{\sigma^2(\mathrm{tr}(k_{XX} - q_{XX}) + \sigma^2)}, \qquad (49)
\end{aligned}$$

where $\|\cdot\|_{\mathrm{op}}$ is the operator norm. Thus, we arrive at the following bound (Burt et al., 2020, Lemma 3):

$$\begin{aligned}
2KL\big(\mathbb{Q}^{\nu^*} \parallel \mathbb{P}^{F|y}\big) &\leq \frac{\mathrm{tr}(k_{XX} - q_{XX})}{\sigma^2}\left(\frac{\|y\|^2}{\mathrm{tr}(k_{XX} - q_{XX}) + \sigma^2} + 1\right). \\
&\leq \frac{\mathrm{tr}(k_{XX} - q_{XX})}{\sigma^2}\left(\frac{\|y\|^2}{\sigma^2} + 1\right). \qquad (50)
\end{aligned}$$

This result shows that the KL divergence becomes small if $\mathrm{tr}(k_{XX} - q_{XX})$ is small. The latter quantity becomes small if, intuitively, the approximate kernel matrix $q_{XX} = k_{XZ} k_{ZZ}^{-1} k_{ZX}$ is close to the exact one k_{XX}. See also Section 3.6 for probabilistic and geometric interpretations of $\mathrm{tr}(k_{XX} - q_{XX}) = \sum_{i=1}^n k(x_i, x_i) - k_Z(x_i)^\top k_{ZZ}^{-1} k_Z(x_i)$.

Burt et al. (2020) then establish various results on the KL divergence for the SVGP approximation by i) relating $\mathrm{tr}(k_{XX} - q_{XX})$ to the eigenvalues of k_{XX} by considering a specific sampling scheme for Z, such as DPPs and leverage score sampling, ii) relating the eigenvalues of k_{XX} to those of the corresponding kernel integral operator, and iii) bounding the eigenvalue decays of the integral operator by considering specific choices of the kernel k and the probability distribution of training input points $x_1, \ldots, x_n$.

4.2 Connection to the Nyström KRR

We now investigate how the bounds on the KL divergence for the SVGP approach are related to the Nyström KRR. The key is the following lemma, which provides an RKHS interpretation of $y^\top (k_{XX} + \sigma^2 I)^{-1} y$ (and $y^\top (q_{XX} + \sigma^2 I)^{-1} y$) appearing in the bound (48). The proof can be found in Appendix B.1.

Lemma 10 *Let k be a kernel with RKHS $\mathcal{H}_k$. Let $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ be given. Then for any $\sigma^2 > 0$, we have*

$$y^\top (k_{XX} + \sigma^2 I_n)^{-1} y = \min_{f \in \mathcal{H}_k} \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - f(x_i))^2 + \|f\|_{\mathcal{H}_k}^2.$$

Lemma 10 shows that $y^\top (k_{XX} + \sigma^2 I_n)^{-1} y$ is a scaled version of the KRR objective function in (2) with $\sigma^2 = n\lambda$. Since the solution to the KRR is given by $\hat{f} = k_X(\cdot)^\top (k_{XX} + \sigma^2 I_n)^{-1} y$, Lemma 10 implies that

$$y^\top (k_{XX} + \sigma^2 I_n)^{-1} y = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + \|\hat{f}\|_{\mathcal{H}_k}^2,$$

Similarly, Lemma 10 implies that $y^\top (q_{XX} + \sigma^2 I_n) y$ can be written as

$$\begin{aligned} y^\top (q_{XX} + \sigma^2 I_n)^{-1} y &= \min_{f \in \mathcal{H}_q} \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \bar{f}(x_i))^2 + \|f\|_{\mathcal{H}_q}^2 \\ &= \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \bar{f}(x_i))^2 + \|\bar{f}\|_{\mathcal{H}_k}^2, \end{aligned}$$

where $\bar{f} = k_Z(\cdot)^\top (\sigma^2 k_{ZZ} + k_{ZX} k_{XZ})^{-1} k_{ZX} y$ is the Nyström approximation by Theorem 4, and we used $\|\bar{f}\|_{\mathcal{H}_k} = \|\bar{f}\|_{\mathcal{H}_q}$ from Lemma 3. Therefore, $y^\top (q_{XX} + \sigma^2 I)^{-1} y$ is a scaled version of the regularized empirical risk of the Nyström KRR.

Therefore, the first two terms in the upper bound (48) of the KL divergence can be written as

$$\begin{aligned} &y^\top (q_{XX} + \sigma^2 I_n)^{-1} y - y^\top (k_{XX} + \sigma^2 I_n)^{-1} y \\ &= \left(\frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \bar{f}(x_i))^2 + \|\bar{f}\|_{\mathcal{H}_k}^2 \right) - \left(\frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + \|\hat{f}\|_{\mathcal{H}_k}^2 \right) \end{aligned} \tag{51}$$

which is essentially the difference between the KRR objectives for the Nyström and exact estimators. Note that the difference (51) is non-negative since the Nyström approximation is obtained from the subspace M of the RKHS $\mathcal{H}_k$;

$$(51) = \min_{f \in M} A(f) - \min_{f \in \mathcal{H}_k} A(f) \geq 0$$

where $A(f) := \left(\frac{1}{\sigma^2} \sum_{i=1}^n (y_i - f(x_i))^2 + \|f\|_{\mathcal{H}_k}^2 \right)$. We can interpret the difference (51) as quantifying the accuracy of the Nyström approximation to the exact KRR solution. That is,

if (51) is small, the Nyström approximation is accurate in the sense that the approximation makes the objective function as small as the exact solution.

As we saw, Burt et al. (2020, Proof of Lemma 3) provides upper-bounds on this difference (51) as in (49). Thus we immediately obtain the following corollary by multiplying (49) by σ^2/n and setting $\sigma^2 = n\lambda$.

Corollary 11 *Let k be a kernel with RKHS $\mathcal{H}_k$. Let $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ be given, and let $Z = (z_1, \ldots, z_m) \in \mathcal{X}^m$ be such that the kernel matrix $k_{ZZ} \in \mathbb{R}^{m \times m}$ is invertible. Let $\bar{f}$ and $\hat{f}$ be the Nyström and exact KRR estimators in* (13) *and* (4)*, respectively, with a regularization constant $\lambda > 0$. Then we have*

$$R_n(\bar{f}; y) - R_n(\hat{f}; y) \leq \frac{\|y\|^2 \|k_{XX} - q_{XX}\|_{\mathrm{op}}}{n(\|k_{XX} - q_{XX}\|_{\mathrm{op}} + n\lambda)} \leq \frac{\|y\|^2 \mathrm{tr}(k_{XX} - q_{XX})}{n(\mathrm{tr}(k_{XX} - q_{XX}) + n\lambda)},$$

where $R_n(f; y) := \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|^2_{\mathcal{H}_k}$.

Therefore, all the results in Burt et al. (2020) on the KL divergence for the SVGP based on Burt et al. (2020, Lemma 3) can be directly translated to the corresponding results for the Nyström KRR using Corollary 11. This is one useful consequence of the RKHS interpretation.

On the other hand, since we now know that there exist KRR objective functions in the expression (47) of the KL divergence, it may be possible to use more sophisticated theoretical arguments for the Nyström KRR (e.g. Bach, 2013; El Alaoui and Mahoney, 2015; Rudi et al., 2015; Chen and Yang, 2021) to obtain sharper bounds on the KL divergence for the SVGP approximation. This investigation is reserved for future research.

4.3 An RKHS Error Bound and its Implications

We present here an upper-bound on the RKHS distance between the Nyström and exact KRR estimators, which is novel to the best of our knowledge. This bound shares a structural similarity to the bound of Burt et al. (2020, Lemma 3) on the KL divergence between the SVGP and exact GP posteriors (50). The proof can be found in Appendix B.2.

Theorem 12 *Let k be a kernel with RKHS $\mathcal{H}_k$. Let $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ be given, and let $Z = (z_1, \ldots, z_m) \in \mathcal{X}^m$ be such that the kernel matrix $k_{ZZ} \in \mathbb{R}^{m \times m}$ is invertible. Let $\bar{f}$ and $\hat{f}$ be the Nyström and exact KRR estimators in* (13) *and* (4)*, respectively, with a regularization constant $\lambda > 0$. Then we have*

$$\|\hat{f} - \bar{f}\|^2_{\mathcal{H}_k} \leq \frac{2\ \mathrm{tr}(k_{XX} - q_{XX})\|y\|^2}{(n\lambda)^2}.$$

The upper-bound takes a similar form as the bound (50) on KL divergence for the SVGP approximation in terms of the dependence on $\mathrm{tr}(k_{XX} - q_{XX})$, $\|y\|^2$ and $\sigma^2 = n\lambda$. Similar to our discussion on Corollary 11, one can thus translate the results from Burt et al. (2020) to the correspond bounds on the RKHS distance between the Nyström and exact KRR estimators.

By the equivalence results in Theorems 1 and 5, we immediately obtain the following corollary for the SVGP and exact GP posterior mean functions:

Corollary 13 *Let k be a kernel with RKHS $\mathcal{H}_k$. Let $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ be given, and let $Z = (z_1, \ldots, z_m) \in \mathcal{X}^m$ be such that the kernel matrix $k_{ZZ} \in \mathbb{R}^{m \times m}$ is invertible. Let m^* and $\bar{m}$ be the SVGP and exact GP posterior mean functions in* (30) *and* (7)*, respectively, with prior $F \sim GP(0, k)$ and likelihood model* (6) *with noise variance $\sigma^2 > 0$. Then we have*

$$\|\bar{m} - m^*\|_{\mathcal{H}_k}^2 \leq \frac{2 \operatorname{tr}(k_{XX} - q_{XX})\|y\|^2}{\sigma^4}.$$

Note that the RKHS distance is stronger than the supremum norm between two functions. In fact, by the reproducing property, it can be shown that

$$(\bar{f}(x) - \hat{f}(x))^2 \leq \|\bar{f} - \hat{f}\|_{\mathcal{H}_k}^2 k(x, x), \quad \forall x \in \mathcal{X}.$$

Moreover, if the kernel k is smooth, then the RKHS distance upper-bounds the derivatives of the RKHS functions. To describe this, let $\mathcal{X} \subset \mathbb{R}^d$ be an open set. Suppose that the kernel k is continuously differntiable[10] on $\mathcal{X}$ in the sense that, for any $j = 1, \ldots, d$, the partial derivative $\partial_j \partial_j' k(x, x')$ exists and is continuous on $\mathcal{X}$, where ∂_j and ∂_j' denote the partial derivatives with respect to the j-th coordinate of the first and second arguments of $k(x, x')$, respectively. Then Steinwart and Christmann (2008, Corollary 4.36) implies that, for all $j = 1, \ldots, d$ and all $x \in \mathcal{X}$,

$$(\partial_j \bar{f}(x) - \partial_j \hat{f}(x))^2 \leq \|\bar{f} - \hat{f}\|_{\mathcal{H}_k}^2 \partial_j \partial_j' k(x, x),$$

Thus, the bound in Theorem 12 implies that, if $\operatorname{tr}(k_{XX} - q_{XX})$ is small, then the partial derivatives (and thus the gradients) of the Nyström KRR approximate well those of the exact KRR. By the same argument and Corollary 13, we immediately obtain the following corollary on the equivalent result for the SVGP approximation.

Corollary 14 *Suppose the same notation and assumptions in Corollary 13. Let $\mathcal{X} \subset \mathbb{R}^d$ be an open set and assume that k is continuously differentiable on $\mathcal{X}$. Then we have for all $j = 1, \ldots, d$ and all $x \in \mathcal{X}$,*

$$(\partial_j m^*(x) - \partial_j \bar{m}(x))^2 \leq \frac{2 \operatorname{tr}(k_{XX} - q_{XX})\|y\|^2 \partial_j \partial_j' k(x, x)}{\sigma^4}.$$

This shows that the SVGP can approximate not only the exact posterior mean function but also its derivatives, if $\operatorname{tr}(k_{XX} - q_{XX})$ is small. In applications where the derivative estimates are used (e.g., see Wu et al. 2017), this result provides a support for using the SVGP approximation in place of the exact GP posterior means of derivatives.

4.4 Lower Bounds for Approximation Errors

We discuss here lower bounds for the average case errors of sparse approximations, by assuming a probabilistic model for training outputs $y = (y_1, \ldots, y_n)^\top$. As before, we fix training inputs $X = (x_1, \ldots, x_n)$ and inducing inputs $Z = (z_1, \ldots, z_m)$. Following Burt et al. (2020), we consider the following model for y:

$$y|X \sim \mathcal{N}(0, k_{XX} + \sigma^2 I_n) \tag{52}$$

10. Many commonly used kernels, such as the Gaussian kernel, satisfy this requirement.

which is given by the likelihood model (6) and by marginalizing the latent prior GP, $F \sim GP(0, k)$. Burt et al. (2020, Lemma 4) shows the following lower and upper bounds for the averaged KL divergence between the SGVP and exact GP posteriors:

$$\frac{\mathrm{tr}(k_{XX} - q_{XX})}{2\sigma^2} \leq \mathbb{E}_y \left[KL\big(\mathbb{Q}^{\nu^*} \parallel \mathbb{P}^{F|y}\big) \right] \leq \frac{\mathrm{tr}(k_{XX} - q_{XX})}{\sigma^2}, \tag{53}$$

where $\mathbb{E}_y$ denotes the expectation with respect to y generated according to (52).

These lower and upper bounds are *a priori* bounds in the sense that they hold for the average with respect to the model and thus are informative before observing the actual training outputs $y_1, \ldots, y_n$. While this performance measure (the averaged KL divergence) is less informative for the approximation accuracy after one has observed actual training outputs $y_1, \ldots, y_n$ (the *a posteori* setting), the lower bound still provides a useful insight. Specifically, the lower bound (53) is proportional to $\mathrm{tr}(k_{XX} - q_{XX})$. Thus if $\mathrm{tr}(k_{XX} - q_{XX})$ is large, then the SGVP posterior $\mathbb{Q}^{\nu^*}$ *cannot* accurately approximate the exact posterior $\mathbb{P}^{F|y}$ on average. This is intuitively the case where the inducing inputs $Z = (z_1, \ldots, z_m)$ do not effectively "cover" the training inputs $X = (x_1, \ldots, x_n)$. Since $\mathrm{tr}(k_{XX} - q_{XX})$ appears both in the upper and lower bounds, the above result shows that $\mathrm{tr}(k_{XX} - q_{XX})$ can serve as an average performance metric for the SVGP approximation.

By combining (47), (51) and (53), we can obtain the corresponding lower bound on the difference of the KRR objectives for the Nyström and exact KRR estimators:

Corollary 15 *Let k be a kernel with RKHS $\mathcal{H}_k$. Let $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and let $Z = (z_1, \ldots, z_m) \in \mathcal{X}^m$ be such that the kernel matrix $k_{ZZ} \in \mathbb{R}^{m \times m}$ is invertible. Suppose $y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ are generated as* (52)*. Let $\bar{f}$ and $\hat{f}$ be the Nyström and exact KRR estimators in* (13) *and* (4)*, respectively, with a regularization constant $\lambda = \sigma^2/n$. Then we have*

$$\lambda \log \frac{\det(k_{XX} + \sigma^2 I_n)}{\det(q_{XX} + \sigma^2 I_n)} \leq \mathbb{E}_y \left[R_n(\bar{f}; y) - R_n(\hat{f}; y) \right]$$

where $R_n(f; y) := \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}_k}^2$.

In the left hand side of Corollary 15, $\log\det(k_{XX} + \sigma^2 I)$ and $\log\det(q_{XX} + \sigma^2 I)$ can intuitively be interpreted as the complexities of the models associated with the kernels k and q, respectively. Thus, Corollary shows that, if the complexity for q is much smaller than that for k, then the difference of the KRR objectives cannot be small on average. This suggests that the left hand side of Corollary 15 may be useful as a quality metric for the Nyström approximation in the *a priori* setting.

5. Conclusions

We have established various connections and equivalences between sparse approximation methods for GPR and KRR, namely the SVGP and Nyström approximations. We believe that these connections and equivalences will be useful for researchers working on either approach to understand the other. As we demonstrated, framing the approach of interest in the language of the other deepens understanding of that approach and provides new insights. For instance, the RKHS formulation is useful in providing geometric interpretation

for various concepts in the GP approach and its sparse approximations. We hope that our investigation will inspire further advances in sparse approximation methods as well as new theoretical studies on the topic.

Appendix A. Proofs for Section 3

A.1 Proof of Lemma 3

Proof We first show $\mathcal{H}_q = M$ as a set of functions. First note that

$$q(\cdot, x) = k_Z(\cdot)^\top k_{ZZ}^{-1} k_Z(x) = P_M\big(k(\cdot, x)\big), \quad \forall x \in \mathcal{X}.$$

Define $\mathcal{H}_{0,q}$ as the vector space

$$\mathcal{H}_{0,q} := \Big\{ f = \sum_{i=1}^{n} \alpha_i q(\cdot, d_i) \mid n \in \mathbb{N},\ \alpha = (\alpha_1, ..., \alpha_n)^\top \in \mathbb{R}^n,\ D = (d_1, ..., d_n) \in \mathcal{X}^n \Big\}.$$

Let $f := \sum_{i=1}^n \alpha_i q(\cdot, d_i) \in \mathcal{H}_{0,q}$ be arbitrary. Since $q(\cdot, d_i) = P_M\big(k(\cdot, d_i)\big) \in M$, we have $f = \sum_{i=1}^n \alpha_i q(\cdot, d_i) \in M$ by the linearity of M. Therefore $\mathcal{H}_{0,q} \subset M$. On the other hand, for any $f = \sum_{j=1}^m \beta_j k(\cdot, z_j) \in M$ with some $\beta_1, \ldots, \beta_m \in \mathbb{R}$, we have $f = \sum_{j=1}^m \beta_j k(\cdot, z_j) = \sum_{j=1}^m \beta_j P_M(k(\cdot, z_j)) = \sum_{j=1}^m \beta_j q(\cdot, z_j) \in \mathcal{H}_{q,0}$. Therefore $M \subset \mathcal{H}_{q,0}$. Thus we have shown $\mathcal{H}_{0,q} = M$ as a set.

Note that the RKHS $\mathcal{H}_q$ is the closure of $\mathcal{H}_{0,q}$ with respect to the norm

$$\begin{aligned}
&\left\| \sum_{i=1}^n \alpha_i q(\cdot, d_i) \right\|_{\mathcal{H}_q}^2 = \alpha^\top q_{DD} \alpha = \alpha^\top k_{DZ} k_{ZZ}^{-1} k_{ZD} \alpha = \alpha^\top k_{DZ} k_{ZZ}^{-1} k_{ZZ} k_{ZZ}^{-1} k_{ZD} \alpha \\
&= \left\| k_Z(\cdot)^\top k_{ZZ}^{-1} k_{ZD} \alpha \right\|_{\mathcal{H}_k}^2 = \left\| q_D(\cdot)^\top \alpha \right\|_{\mathcal{H}_k}^2 = \left\| \sum_{i=1}^n \alpha_i q(\cdot, d_i) \right\|_{\mathcal{H}_k}^2,
\end{aligned}$$

which coincides with the norm of $\mathcal{H}_k$. As $\mathcal{H}_q = M$ is a finite-dimensional subspace of $\mathcal{H}_k$, it is closed and therefore

$$\mathcal{H}_q = \overline{\mathcal{H}_{0,q}} = \overline{M} = M.$$

where the closure is with respect to the norm $\|\cdot\|_{\mathcal{H}_k} = \|\cdot\|_{\mathcal{H}_q}$.

Next we show that the scalar products on M and $\mathcal{H}_q$ also coincide. Take arbitrary f and g from $\mathcal{H}_q$. As $\mathcal{H}_q = \mathcal{H}_{0,q}$, we find a representation of the form

$$\begin{aligned}
f &= q_D(\cdot)^\top \alpha = k_Z(\cdot)^\top k_{ZZ}^{-1} k_{ZD} \alpha = k_Z(\cdot)^\top \tilde{\alpha} \\
g &= q_E(\cdot)^\top \beta = k_Z(\cdot)^\top k_{ZZ}^{-1} k_{ZE} \beta = k_Z(\cdot)^\top \tilde{\beta},
\end{aligned}$$

where $D = (d_1, ..., d_n) \in \mathcal{X}^n$, $E = (e_1, ..., e_\ell) \in \mathcal{X}^\ell$, $\alpha \in \mathbb{R}^n$, $\beta \in \mathbb{R}^\ell$, $\tilde{\alpha} := k_{ZZ}^{-1} k_{ZD} \alpha$ and $\tilde{\beta} := k_{ZZ}^{-1} k_{ZE} \beta$. This leads to

$$\langle f, g \rangle_{\mathcal{H}_q} = \alpha^\top q_{DE} \beta = \alpha^\top k_{DZ} k_{ZZ}^{-1} k_{ZE} \beta$$

and

$$\langle f, g \rangle_{\mathcal{H}_k} = \tilde{\alpha}^\top k_{ZZ} \tilde{\beta} = \alpha^\top k_{DZ} k_{ZZ}^{-1} k_{ZE} \beta,$$

which shows that the scalar products are the same.

■

A.2 Proof of Theorem 6

Proof We first analyze the term (26) in the ELBO. Let $f_Z \in \mathbb{R}^m$ be an arbitrary vector in the support of $\mathbb{Q}_Z^\nu$. Define a notation for the conditional expectation

$$F_m(x) := \mathbb{E}[F(x) \mid F_Z = f_Z] = \mathbb{E}[F^\nu(x) \mid F_Z^\nu = f_Z].$$

where the identity follows from the definition of $F^\nu \sim \mathbb{Q}^\nu$. Then, from the standard bias-variance decomposition argument, we have

$$\begin{aligned}
&\sum_{i=1}^n \mathbb{E}\left[(y_i - F^\nu(x_i))^2 \mid F_Z^\nu = f_Z\right] = \sum_{i=1}^n \mathbb{E}\left[(y_i - F(x_i))^2 \mid F_Z = f_Z\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[(y_i - F_m(x_i))^2 \mid F_Z = f_Z\right] + \sum_{i=1}^n \mathbb{E}\left[(F_m(x_i) - F(x_i))^2 \mid F_Z = f_Z\right] \\
&= \sum_{i=1}^n (y_i - F_m(x_i))^2 + \sum_{i=1}^n \mathbb{E}\left[(F_m(x_i) - F(x_i))^2 \mid F_Z = f_Z\right]. \qquad (54)
\end{aligned}$$

Note that, because $F_m(x)$ is the conditional expectation of $F(x)$ given $F_Z = f_Z$, it is equivalent to the kernel interpolator with training data $(z_i, F(z_i))_{i=1}^n$ and can be written as

$$F_m(x) = k_Z(x)^\top k_{ZZ}^{-1} f_Z.$$

Therefore

$$\begin{aligned}
&\int (y_i - F_m(x_i))^2 dQ_Z^\nu(f_Z) = \int (y_i - k_Z(x_i)^\top k_{ZZ}^{-1} f_Z)^2 dQ_Z^\nu(f_Z) \\
&= \int (y_i - k_Z(x_i)^\top k_{ZZ}^{-1}\mu)^2 dQ_Z^\nu(f_Z) + \int (k_Z(x_i)^\top k_{ZZ}^{-1}\mu - k_Z(x_i)^\top k_{ZZ}^{-1} f_Z)^2 dQ_Z^\nu(f_Z) \\
&= (y_i - k_Z(x_i)^\top k_{ZZ}^{-1}\mu)^2 + k_Z(x_i)^\top k_{ZZ}^{-1}\Sigma k_{ZZ}^{-1} k_Z(x_i),
\end{aligned}$$

where the last identity follows from $Q_Z^\nu = \mathcal{N}(\mu, \Sigma)$ by definition.

On the other hand, the second term in (54) is the conditional variance of $F(x_i)$ given $F_Z = f_Z$, and thus given by

$$\mathbb{E}\left[(F_m(x_i) - F(x_i))^2 \mid F_Z = f_Z\right] = k(x_i, x_i) - k_Z(x_i)^\top k_{ZZ}^{-1} k_Z(x_i),$$

which is independent to the "observations" f_Z. Therefore

$$\int \mathbb{E}\left[(F_m(x_i) - F(x_i))^2 \mid F_Z = f_Z\right] dQ_Z^\nu(f_Z) = k(x_i, x_i) - k_Z(x_i)^\top k_{ZZ}^{-1} k_Z(x_i).$$

Using these identities, we have

$$\begin{aligned}
&\int \left(\sum_{i=1}^{n} \mathbb{E}\left[(y_i - F^\nu(x_i))^2 \mid F_Z^\nu = f_Z \right] \right) dQ_Z^\nu(f_Z) \\
&= \sum_{i=1}^{n} (y_i - k_Z(x_i)^\top k_{ZZ}^{-1} \mu)^2 + \sum_{i=1}^{n} k_Z(x_i)^\top k_{ZZ}^{-1} \Sigma k_{ZZ}^{-1} k_Z(x_i) \\
&+ \sum_{i=1}^{n} \Big(k(x_i, x_i) - k_Z(x_i)^\top k_{ZZ}^{-1} k_Z(x_i) \Big)
\end{aligned}$$

The proof completes by inserting this last expression of (26) and the expression of the KL divergence (25) in the ELBO (24) and rearranging. ■

Appendix B. Proofs for Section 4

B.1 Proof of Lemma 10

Proof Recall that $\hat{f} := k_X(\cdot)^\top (k_{XX} + \sigma^2 I_n)^{-1} y$ is the solution of KRR. We have

$$\hat{f}_X = k_{XX}(k_{XX} + \sigma^2 I_n)^{-1} y = (I_n - \sigma^2 (k_{XX} + \sigma^2 I_n)^{-1}) y,$$

where we used the formula $A(A + \sigma^2 I)^{-1} = I_n - \sigma^2 (A + \sigma^2 I_n)^{-1}$ that holds for any positive semidefinte matrix A. Now we have

$$\min_{f \in \mathcal{H}_k} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \sigma^2 \|f\|_{\mathcal{H}_k}^2 = \|y - \hat{f}_X\|^2 + \sigma^2 \|\hat{f}\|_{\mathcal{H}_k}^2$$

The first term can be expanded as

$$\begin{aligned}
&\|y - \hat{f}_X\|^2 = \|y\|^2 - 2y^\top \hat{f}_X + \|\hat{f}_X\|^2 \\
&= \|y\|^2 - 2y^\top (I_n - \sigma^2 (k_{XX} + \sigma^2 I_n)^{-1}) y + y^\top (I_n - \sigma^2 (k_{XX} + \sigma^2 I_n)^{-1})^2 y \\
&= \|y\|^2 - 2y^\top (I_n - \sigma^2 (k_{XX} + \sigma^2 I_n)^{-1}) y \\
&+ y^\top (I_n - 2\sigma^2 (k_{XX} + \sigma^2 I_n)^{-1} + \sigma^4 (k_{XX} + \sigma^2 I_n)^{-2}) y \\
&= \sigma^4 y^\top (k_{XX} + \sigma^2 I_n)^{-2} y.
\end{aligned}$$

The second term is

$$\begin{aligned}
\sigma^2 \|\hat{f}\|_{\mathcal{H}_k}^2 &= \sigma^2 y^\top (k_{XX} + \sigma^2 I_n)^{-1} k_{XX} (k_{XX} + \sigma^2 I_n)^{-1} y \\
&= \sigma^2 y^\top (k_{XX} + \sigma^2 I_n)^{-1} (I_n - \sigma^2 (k_{XX} + \sigma^2 I_n)^{-1}) y \\
&= \sigma^2 y^\top (k_{XX} + \sigma^2 I_n)^{-1} y - \sigma^4 y^\top (k_{XX} + \sigma^2 I_n)^{-2} y.
\end{aligned}$$

Therefore,

$$\|y - \hat{f}_X\|^2 + \sigma^2 \|\hat{f}\|_{\mathcal{H}_k}^2 = \sigma^2 y^\top (k_{XX} + \sigma^2 I_n)^{-1} y.$$

■

B.2 Proof of Theorem 12

Proof We first make preliminaries for proving the theorem. For a symmetric matrix $B \in \mathbb{R}^{n\times n}$ with $n \in \mathbb{N}$, denote by $\lambda_1(B) \geq \cdots \geq \lambda_n(B)$ its eigenvalues with multiplicities in the decreasing order. For any symmetric and positive semi-definite (SPSD) matrix $A \in \mathbb{R}^{n\times n}$ and any $B \in \mathbb{R}^{n\times n}$, we have (see Saniuk and Rhodes, 1987)

$$\mathrm{tr}(AB) \leq \mathrm{tr}(A)\|B\|_{\mathrm{op}}, \tag{55}$$

where $\|B\|_{\mathrm{op}} := \sup_{v\in\mathbb{R}:\ \|v\|\leq 1} \|Bv\|$ denotes the operator norm (or spectral norm). If B is symmetric, we have $\|B\|_{\mathrm{op}} = \max(|\lambda_1(B)|, |\lambda_n(B)|)$.

For any SPSD matrix $A \in \mathbb{R}^{n\times n}$ and any symmetric and negative semi-definite (SNSD) matrix $B \in \mathbb{R}^{n\times n}$, we have

$$\mathrm{tr}(AB) = \mathrm{tr}(BA) = \mathrm{tr}\big(A^{1/2}BA^{1/2}\big) = \sum_{i=1}^{n} \lambda_i(A^{1/2}BA^{1/2}) \leq 0, \tag{56}$$

where the inequality follows from the fact that $A^{1/2}BA^{1/2}$ is SNSD and hence all its eigenvalues are non-positive.

We also use the following short hand notation

$$\begin{aligned} &K := k_{XX}, \quad Q := q_{XX}, \quad \tilde{K} := K + n\lambda I_n, \quad \tilde{Q} := Q + n\lambda I_n, \\ &\alpha = \tilde{K}^{-1}y, \quad \beta = \tilde{Q}^{-1}y, \quad \tilde{\beta} = \big(n\lambda k_{ZZ} + k_{ZX}k_{XZ}\big)^{-1}k_{ZX}y. \end{aligned}$$

Note that the matrices $K, Q, \tilde{K}, \tilde{Q}$ are SPSD. It holds that

$$\|\tilde{K}^{-1}\|_{\mathrm{op}} \leq \frac{1}{n\lambda}, \quad \|\tilde{Q}^{-1}\|_{\mathrm{op}} \leq \frac{1}{n\lambda}, \quad \|K\tilde{K}^{-1}\|_{\mathrm{op}} \leq 1, \quad \|Q\tilde{Q}^{-1}\|_{\mathrm{op}} \leq 1.$$

Using the above notation, the KRR estimator $\hat{f}$ and the Nyström approximation $\bar{f}$ can be written for any $x \in \mathcal{X}$ as

$$\begin{aligned} \hat{f}(x) &= k_X(x)^\top \tilde{K}^{-1}y, \\ \bar{f}(x) &= q_X(x)^\top \tilde{Q}^{-1}y = q_X(x)^\top \beta \\ &= k_Z(x)^\top \big(n\lambda k_{ZZ} + k_{ZX}k_{XZ}\big)^{-1}k_{ZX}y = k_Z(x)^\top \tilde{\beta}. \end{aligned}$$

We will use the following identity:

$$k_{XZ}\tilde{\beta} = \bar{f}_X = q_{XX}\tilde{Q}^{-1}y = Q\tilde{Q}^{-1}y.$$

With these preparations, we now prove the assertion. First we have

$$\begin{aligned} \|\hat{f} - \bar{f}\|_{\mathcal{H}_k}^2 &= \|\hat{f}\|_{\mathcal{H}_k}^2 - 2\langle \hat{f}, \bar{f}\rangle_{\mathcal{H}_k} + \|\bar{f}\|_{\mathcal{H}_k}^2 \\ &= \|\hat{f}\|_{\mathcal{H}_k}^2 - 2\langle \hat{f}, \bar{f}\rangle_{\mathcal{H}_k} + \|\bar{f}\|_{\mathcal{H}_q}^2, \end{aligned}$$

where we used $\|\bar{f}\|^2_{\mathcal{H}_q} = \|\bar{f}\|^2_{\mathcal{H}_k}$, which holds from $\bar{f} \in M$ and Lemma 3. The expression is equal to

$$
\begin{aligned}
&= \alpha^\top K \alpha - 2\alpha^\top k_{XZ}\tilde{\beta} + \beta^\top Q \beta \\
&= \mathrm{tr}(K\alpha\alpha^\top) - 2\mathrm{tr}(k_{XZ}\tilde{\beta}\alpha^\top) + \mathrm{tr}(Q\beta\beta^\top) \\
&= \mathrm{tr}(K\tilde{K}^{-1}yy^\top\tilde{K}^{-1}) - \mathrm{tr}(Q\tilde{Q}^{-1}yy^\top\tilde{K}^{-1}) + \mathrm{tr}(Q\tilde{Q}^{-1}yy^\top\tilde{Q}^{-1}) - \mathrm{tr}(Q\tilde{Q}^{-1}yy^\top\tilde{K}^{-1}) \\
&= \mathrm{tr}\big((K\tilde{K}^{-1} - Q\tilde{Q}^{-1})yy^\top\tilde{K}^{-1}\big) + \mathrm{tr}\big(Q\tilde{Q}^{-1}yy^\top(\tilde{Q}^{-1} - \tilde{K}^{-1})\big) \\
&\leq \mathrm{tr}(K\tilde{K}^{-1} - Q\tilde{Q}^{-1})\|yy^\top\|_{\mathrm{op}}\|\tilde{K}^{-1}\|_{\mathrm{op}} + \mathrm{tr}(\tilde{Q}^{-1} - \tilde{K}^{-1})\|Q\tilde{Q}^{-1}\|_{\mathrm{op}}\|yy^\top\|_{\mathrm{op}} \\
&\leq \frac{1}{n\lambda}\mathrm{tr}(K\tilde{K}^{-1} - Q\tilde{Q}^{-1})\|y\|^2 + \mathrm{tr}(\tilde{Q}^{-1} - \tilde{K}^{-1})\|y\|^2 \\
&= \frac{1}{n\lambda}\Big(\mathrm{tr}\big((K-Q)\tilde{K}^{-1}\big) + \mathrm{tr}\big(Q(\tilde{K}^{-1} - \tilde{Q}^{-1})\big)\Big)\|y\|^2 + \mathrm{tr}(\tilde{Q}^{-1}(K-Q)\tilde{K}^{-1})\|y\|^2
\end{aligned}
$$

Since Q is SPSD and $\tilde{K}^{-1} - \tilde{Q}^{-1}$ is SNSD, we have

$$\mathrm{tr}\big(Q(\tilde{K}^{-1} - \tilde{Q}^{-1})\big) \leq 0$$

due to (56). Using this and (55), we obtain

$$
\begin{aligned}
&\leq \frac{1}{n\lambda}\mathrm{tr}(K-Q)\|\tilde{K}^{-1}\|_{\mathrm{op}}\|y\|^2 + \mathrm{tr}(K-Q)\|\tilde{K}^{-1}\|_{\mathrm{op}}\|\tilde{Q}^{-1}\|_{\mathrm{op}}\|y\|^2 \\
&\leq \frac{2}{(n\lambda)^2}\mathrm{tr}(K-Q)\|y\|^2,
\end{aligned}
$$

which concludes the proof. ■

References

Vincent Adam, Stefanos Eleftheriadis, Artem Artemev, Nicolas Durrande, and James Hensman. Doubly sparse variational gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 2874–2884. PMLR, 2020.

Raja Hafiz Affandi, Alex Kulesza, Emily Fox, and Ben Taskar. Nyström approximation for large-scale determinantal processes. In *Artificial Intelligence and Statistics*, pages 85–98. PMLR, 2013.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

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

Matthias Bauer, Mark van der Wilk, and Carl Edward Rasmussen. Understanding probabilistic sparse Gaussian process approximations. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

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*, 106(2):369–374, 2009.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

David Burt, Carl Edward Rasmussen, and Mark Van Der Wilk. Rates of convergence for sparse variational Gaussian process regression. In *Proceedings of the 36th International Conference on Machine Learning*, pages 862–871, 2019.

David R. Burt, Carl Edward Rasmussen, and Mark van der Wilk. Convergence of sparse variational inference in Gaussian processes regression. *Journal of Machine Learning Research*, 21(131):1–63, 2020.

Yifan Chen and Yun Yang. Fast statistical leverage score approximation in kernel ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pages 2935–2943. PMLR, 2021.

Corinna Cortes, Mehryar Mohri, and Ameet Talwalkar. On the impact of kernel approximation on learning accuracy. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 113–120. JMLR Workshop and Conference Proceedings, 2010.

Lehel Csató and Manfred Opper. Sparse on-line Gaussian processes. *Neural Computation*, 14(3):641–668, 2002.

Michal Derezinski, Rajiv Khanna, and Michael W Mahoney. Improved guarantees and a multiple-descent curve for Column Subset Selection and the Nyström method. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4953–4964. Curran Associates, Inc., 2020.

Petros Drineas and Michael W. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6(72):2153–2175, 2005.

Vincent Dutordoir, Nicolas Durrande, and James Hensman. Sparse Gaussian processes with spherical harmonic features. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2793–2802. PMLR, 2020.

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

Charless Fowlkes, Serge Belongie, Fan Chung, and Jitendra Malik. Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, 2004.

Alex Gittens and Michael W Mahoney. Revisiting the nyström method for improved large-scale machine learning. *The Journal of Machine Learning Research*, 17(1):3977–4041, 2016.

James Hensman, Nicolò Fusi, and Neil D Lawrence. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 282–290, 2013.

James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable variational Gaussian process classification. In *Artificial Intelligence and Statistics*, pages 351–360. PMLR, 2015a.

James Hensman, Alexander G Matthews, Maurizio Filippone, and Zoubin Ghahramani. MCMC for variationally sparse Gaussian processes. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015b.

James Hensman, Nicolas Durrande, and Arno Solin. Variational Fourier features for Gaussian processes. *Journal of Machine Learning Research*, 18(151):1–52, 2018.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

G. S. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2): 495–502, 1970.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Ensemble Nyström method. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling methods for the nyström method. *Journal of Machine Learning Research*, 13(34):981–1006, 2012.

Felix Leibfried, Vincent Dutordoir, ST John, and Nicolas Durrande. A tutorial on sparse Gaussian processes and variational inference. *arXiv preprint arXiv:2012.13962*, 2020.

Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Fast DPP sampling for Nyström with application to kernel methods. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2061–2070, New York, New York, USA, 20–22 Jun 2016. PMLR.

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

Giacomo Meanti, Luigi Carratino, Lorenzo Rosasco, and Alessandro Rudi. Kernel methods through the roof: Handling billions of points efficiently. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14410–14422. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/a59afb1b7d82ec353921a55c579ee26d-Paper.pdf

Cameron Musco and Christopher Musco. Recursive sampling for the Nyström method. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

E. Parzen. An approach to time series analysis. *The Annals of Mathematical Statistics*, 32 (4):951–989, 1961.

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

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

Simone Rossi, Markus Heinonen, Edwin Bonilla, Zheyang Shen, and Maurizio Filippone. Sparse Gaussian processes revisited: Bayesian approaches to inducing-variable approximations. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1837–1845. PMLR, 13–15 Apr 2021.

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

Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. FALKON: An optimal large scale kernel method. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

J Saniuk and I Rhodes. A matrix inequality associated with bounds on solutions of algebraic Riccati and Lyapunov equations. *IEEE Transactions on Automatic Control*, 32(8):739–740, 1987.

Bernhard Schölkopf and Alexander J Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2002.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pages 416–426. Springer, 2001.

Matthias Seeger, Christopher Williams, and Neil Lawrence. Fast forward selection to speed up sparse Gaussian process regression. In *Artificial Intelligence and Statistics*, 2003.

Jiaxin Shi, Michalis Titsias, and Andriy Mnih. Sparse orthogonal variational inference for Gaussian processes. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1932–1942. PMLR, 26–28 Aug 2020.

Alex J Smola and Bernhard Schölkopf. Sparse greedy matrix approximation for machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 911–918, 2000.

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

Edward Snelson and Zoubin Ghahramani. Local and global sparse Gaussian process approximations. In *Artificial Intelligence and Statistics*, pages 524–531, 2007.

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

Ameet Talwalkar, Sanjiv Kumar, and Henry Rowley. Large-scale manifold learning. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008.

Ameet Talwalkar, Sanjiv Kumar, Mehryar Mohri, and Henry Rowley. Large-scale SVD and manifold learning. *Journal of Machine Learning Research*, 14(60):3129–3152, 2013.

Terence Tao. *An Introduction to Measure Theory*, volume 126. American Mathematical Society Providence, 2011.

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

Michalis K Titsias. Variational model selection for sparse Gaussian process regression. *Technical Report, University of Manchester, UK*, 2009b.

Gia-Lac Tran, Dimitrios Milios, Pietro Michiardi, and Maurizio Filippone. Sparse within sparse Gaussian processes using neighbor information. In *Proceedings of the Thirty-eighth International Conference on Machine Learning*, 2021.

Grace Wahba. *Spline Models for Observational Data*. SIAM, 1990.

Christopher KI Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, pages 682–688, 2001.

Jian Wu, Matthias Poloczek, Andrew G Wilson, and Peter Frazier. Bayesian optimization with gradients. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.