Variational Gaussian Processes: A Functional Analysis View

Veit Wild*
University of Oxford

George Wynne*
Imperial College London

Abstract

Variational Gaussian process (GP) approximations have become a standard tool in fast GP inference. This technique requires a user to select variational features to increase efficiency. So far the common choices in the literature are disparate and lacking generality. We propose to view the GP as lying in a Banach space which then facilitates a unified perspective. This is used to understand the relationship between existing features and to draw a connection between kernel ridge regression and variational GP approximations.

1 INTRODUCTION

Gaussian processes (GPs) are a ubiquitous modelling paradigm within machine learning (Rasmussen and Williams, 2005). They are random functions with the useful property that their pointwise evaluations form a multivariate Gaussian random vector. Within the Bayesian framework one may use a Gaussian process as a prior for an unknown function, condition on observed information and then a posterior over the unknown function is obtained. Examples of applications include regression, classification, reinforcement learning and optimisation. See Rasmussen and Williams (2005) for an introduction. The GP framework enjoys such popularity since it is flexible, interpretable, has closed form expressions in some scenarios and offers a degree of uncertainty quantification.

An issue of the GP framework is the naive computational cost $O(N^3)$ to perform predictions, where N is the number of observed data points. This is due to a matrix inversion term. Many methods have been proposed to reduce this computational cost to something more palatable, both through theoretical (Csató and Opper, 2002; Seeger et al., 2003; Snelson and Ghahramani, 2006, 2007; Titsias, 2009a) and computational innovations (Gardner et al., 2018; Wang et al., 2019).

The focus of this paper is the variational inference paradigm where the posterior GP is approximated by an element of a candidate family through an optimisation routine where distance from the posterior GP is measured with the Kullback-Leibler (KL) divergence. The common candidate family is a family of GPs formed by conditioning the prior on M surrogate *features*, not necessarily equal to the observed information, resulting in $O(NM^2)$ complexity rather than the aforementioned $O(N^3)$. For example, features could be point evaluations or values of inner products against some user chosen set of functions. Foundational papers regarding variational inference for GPs include Titsias (2009a); Matthews et al. (2016) and for a survey consult Leibfried et al. (2020).

Choosing features to condition on is often conducted with the aim of closed form, or at least easy to compute, expressions. Therefore the features chosen can be very dependent on the given GP of interest through the covariance kernel, mean function or space the GP takes values in. This has led to a heuristic and somewhat ad-hoc approach in the literature to deriving features. Indeed, there is little in the way of a unified view of the choice of features and how different choices relate to each other.

Contributions: We present a unified perspective of existing features used in variational Gaussian process approximations by embracing the fact that GPs can be viewed as Gaussian random elements in Banach spaces. This perspective reveals generalisations of, and equivalences between, commonly used variational features. In particular we generalise the derivation for the popular variational Fourier features (Hensman et al., 2018). This injects rigour and clarity into the features used. Finally, a connection to kernel ridge regression is made which clarifies the role that the variational features play in the posterior approximation.

Existing work: The commonly employed framework of variational GP approximation was derived by Tit-

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. *** Equal contribution**

sias (2009a). Two other types of features we focus on are inter-domain (Lázaro-Gredilla and Figueiras-Vidal, 2009) and Fourier (Hensman et al., 2018). A formalism of the use of Kullback-Liebler divergence over infinite dimensions, a key part of the variational GP methodology, was clarified by Matthews et al. (2016). The Fourier features have recently been combined with spherical harmonics (Dutordoir et al., 2020) and rough path theory (Lemercier et al., 2021). An appeal to Gaussian measures is made in Cheng and Boots (2016, 2017); Salimbeni et al. (2018). The Gaussian process is associated with a Gaussian measure in the RKHS. This *dual* formulation is then used to derive more efficient variational GP approximations. However, although Gaussian measures are referenced in their work this avenue is not consistently pursued which separates their line of work from ours.

2 GAUSSIAN PROCESSES AND GAUSSIAN RANDOM ELEMENTS

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a the underlying probability space on which all random quantities are defined. Let $\mathcal{X}$ be a set. A family of random variables $G\colon \mathcal{X} \times \Omega \to \mathbb{R}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ is called a random process. Let $G(x)$ denote the random variable $G(x, \cdot)$. A random process is called a *Gaussian process* (GP) if for every $N \in \mathbb{N}$ and $\{x_n\}_{n=1}^N \subset \mathcal{X}$ the random vector $(G(x_1), \ldots, G(x_N))$ is Gaussian. A Gaussian process is entirely determined by its mean function $m\colon \mathcal{X} \to \mathbb{R}$ defined $m(x) \coloneqq \mathbb{E}[G(x)]$ and covariance function, also know as covariance kernel, $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined $k(x, x') \coloneqq \mathbb{E}[(G(x) - m(x))(G(x') - m(x'))]$. We denote the GP with mean m and covariance kernel k as $G \sim GP(m, k)$. For background on Gaussian processes consult the works by Rasmussen and Williams (2005); Lifshits (2012); Adler (1990).

Let E be a separable Banach space, $\mathcal{B}(E)$ the Borel σ-algebra and $\mathcal{P}(E)$ the set of Borel probability measures on E. We will always assume that E is separable without explicitly stating it every time. The dual of E is defined as $E^* \coloneqq \{x^* : E \to \mathbb{R} \,|\, x^* \text{ is linear and continuous}\}$ and for $x^* \in E^*, y \in E$ we write $(y, x^*)_E \coloneqq x^*(y)$ for the so called *dual pairing*. A mapping $F\colon \Omega \to E$ is called a *Gaussian random element* (GRE) if for every $x^* \in E^*$ the real valued random variable $x^*(F) = (F, x^*)_E$ is Gaussian. Each GRE has an associated mean $m \in E$ which is uniquely characterised by satisfying $(m, x^*)_E = \mathbb{E}[(F, x^*)_E]$ for all $x^* \in E^*$ and covariance operator $C\colon E^* \to E$ uniquely characterised by satisfying $(Cx^*, y^*)_E = \mathrm{Cov}[(F, x^*)_E, (F, y^*)_E]$ for all $x^*, y^* \in E^*$. We denote $F \sim \mathcal{N}(m, C)$ for a GRE with mean m and covariance operator C. Note that for $E = \mathbb{R}^N$ this coincides with the standard definition for the normal distribution in $\mathbb{R}^N$.

A measure $P \in \mathcal{P}(E)$ is called a *Gaussian measure* (GM) if for every $x^* \in E^*$ the pushforward measure $P^{x^*} \in \mathcal{P}(\mathbb{R})$ defined by $P^{x^*}(\cdot) \coloneqq (P \circ (x^*)^{-1})(\cdot)$ is a Gaussian measure on $\mathcal{B}(\mathbb{R})$. As with random variables in $\mathbb{R}$ and probability measures on $\mathcal{B}(\mathbb{R})$ there is a one-to-one correspondence between GREs and GMs on E. GREs and GMs can be studied in far more generality, or indeed with more specificity, than E being a Banach space (Bogachev, 1998; Da Prato, 2006).For a gentle introduction into Gaussian measures on Banach spaces see chapter 3 in Hairer (2009).

When a GP G satisfies $\mathbb{P}(\{\omega\colon G(\cdot, \omega) \in E\}) = 1$ we say that its sample paths lie almost surely in E. This has been studied for numerous common choices of E (Rajput and Cambanis, 1972; Rajput, 1972; Lukić and Beder, 2001) and one can then identify the GP with a GRE, or equivalently with a GM, over E. Throughout the rest of this paper we shall be dealing purely in terms of GREs and later specific examples equating these to GPs shall be given. The use of GREs facilitates a general view of variational inference and will be the vehicle of our results.

2.1 The Problem With The Path View

This subsection will motivate the use of GREs in the analysis of variational GP approximation methods by arguing why the path view is insufficient.

Let $G \sim GP(m, k)$ be a Gaussian process on $\mathcal{X}$ and $\mathbb{R}^{\mathcal{X}} \coloneqq \{f : f : \mathcal{X} \to \mathbb{R}\}$ the vector-space of functions from $\mathcal{X}$ to $\mathbb{R}$. Define $\pi_x : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$, $f \mapsto f(x)$ the pointwise evaluation map for some $x \in \mathcal{X}$. Let $\mathcal{S}$ be the cylindrical σ-algebra, which is defined as the smallest σ-algebra such that all projections π_x, $x \in \mathcal{X}$, are measurable. Every random process G with paths over $\mathcal{X}$ can be canonically identified with a random element in $\mathbb{R}^{\mathcal{X}}$ $F : \big(\Omega, \mathcal{A}, \mathbb{P}\big) \to \big(\mathbb{R}^{\mathcal{X}}, \mathcal{S}\big)$ via $F(\omega) = G(\cdot, \omega)$ (Kallenberg and Kallenberg, 1997, Chapter 3). In short, a stochastic process without any additional assumptions about the path is nothing but a random element in the *large* space $\mathbb{R}^{\mathcal{X}}$ with a *small* σ-algebra $\mathcal{S}$.

This presents two main issues. First, $\mathcal{S}$ is so small that while pointwise evaluations are measurable, operations we want to perform in the variational GP framework may not be, such as integration of GPs. Secondly, the theoretical framework of variational GP approximations outlined by Matthews et al. (2016) requires the paths to lie in a Polish space so that Bayes theorem may be used. The space $\mathbb{R}^{\mathcal{X}}$ is not a Polish space so is not an appropriate space to view our GP paths in.

The remedy will be to view the GP paths in a Banach space, which is Polish and has enough structure so that operations we want to perform on the GP are valid. To do this we will equate the GP to a GRE.

3 GAUSSIAN RANDOM ELEMENT REGRESSION

In this section we outline Gaussian random element regression which is the random element view of standard Gaussian process regression. This view is commonly employed in areas such as Bayesian inverse problems (Stuart, 2010). At first glance the framework may appear (indulgently) abstract. However, we believe this is the most natural framework to investigate variational GP approximation. There is an important example at the end of the section showing all that follows does in fact coincide with standard GP regression.

Let F be a GRE in E with mean m and covariance operator C and denote by $P \in \mathcal{P}(\mathcal{X})$ its corresponding GM. Suppose we have observations $Y = \{Y_n\}_{n=1}^N$ which are the image of F under some $\{D_n\}_{n=1}^N \subset E^*$, corrupted by independent scalar Gaussian noise

$$Y_n = (F, D_n)_E + \epsilon_n,$$

where $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$ independently for $n = 1, \dots, N$. This can be equivalently expressed in perhaps more familiar notation as the probability density function (pdf) of Y given $F = f$ is

$$p(y|F = f) := \mathcal{N}\big(y|(f, D)_E, \sigma^2 I_N\big),$$

for $y \in \mathbb{R}^N, f \in E$, $(f, D)_E := \big((f, D_n)_E\big)_{n=1,\dots,N}$ and $\mathcal{N}\big(\,\cdot\,|\mu, \Sigma\big)$ denotes the pdf of a Gaussian distribution on $\mathbb{R}^N$ with mean vector $\mu \in \mathbb{R}^N$ and covariance matrix $\Sigma \in \mathbb{R}^{N\times N}$.

In the Bayesian paradigm, one updates their beliefs about F after observing $Y := (Y_1, \dots, Y_N)$ by combining the prior $F \sim \mathcal{N}(0, C)$ with the likelihood $p(y|F = f)$ to form a posterior. This can be a delicate task since E could be infinite dimensional. However, since in our scenario E is a Banach space and the measures corresponding to $Y|F = f$ are all dominated by the Lebesgue measure on $\mathbb{R}^N$ with a jointly measurable map $(y, f) \in \mathbb{R}^N \times E \to p(y|f) \in \mathbb{R}$, an infinite dimensional version of Bayes theorem applies (Ghosal and van der Vaart, 2017, Chapter 1.3).

It states that a regular version (Klenke, 2013, chapter 8.3) of the posterior measure exists, denoted $P^{F|Y} : \mathbb{R}^N \times \mathcal{B}(E) \to [0, \infty)$, $(y, A) \mapsto P^{F|Y=y}(A)$ and the measure $P^{F|Y=y}$ on $\mathcal{B}(E)$, which is the posterior measure of F given $Y = y$, is dominated by the prior measure P for any $y \in \mathbb{R}^N$ with Radon-Nikodym density $\frac{p(y|f)}{p(y)}$.

What all these technicalities really mean is that for $A \in \mathcal{B}(E)$

$$P^{F|Y=y}(A) = \int_A \frac{p(y|f)}{p(y)} dP(f), \tag{1}$$

where

$$\begin{aligned} p(y) &= \int_E p(y|F = f) dP(f) \\ &= \mathcal{N}\big(y|(m, D)_E, C_{DD} + \sigma^2 I_N\big) \end{aligned}$$

with

$$\big((m, D)_E\big)_n := (m, D_n) \tag{2}$$

$$(C_{DD})_{n,n'} := (CD_n, D_{n'})_E, \tag{3}$$

for all $n, n' = 1, \dots, N$.

This posterior measure $P^{F|Y=y}$ is a GM since it is formed from the Gaussian likelihood $p(y|F = f)$ and Gaussian prior $F \sim \mathcal{N}(m, C)$ (details in supplementary material section A). Denote the mean and covariance operator of $P^{F|Y=y}$ by $\widetilde{m}, \widetilde{C}$ respectively. As is usually the case with Bayesian techniques, the user is often not interested in the posterior measure itself but its pushforward through some prediction operation.

We focus on the case of two linear maps $T, T' \in E^*$ since the general case of $S \in \mathbb{N}$ elements can be handled analogously. The posterior mean $\widetilde{m} \in E$ satisfies

$$\big(\widetilde{m}, T\big)_E = (m, T)_E + C_{TD}(C_{DD} + \sigma^2 I_N)^{-1} y, \tag{4}$$

for any $T \in E^*$ and the posterior covariance operator $\widetilde{C} : E^* \to E$ satisfies

$$\begin{aligned} &(\widetilde{C}T, T')_E \\ &\quad = (CT, T')_E - C_{TD}(C_{DD} + \sigma^2 I_N)^{-1} C_{DT'}, \end{aligned} \tag{5}$$

for all $T, T' \in E^*$, where C_{DD} as in (3) and $(C_{TD})_{1,n} := (CT, D_n)_E$ for $n = 1, \dots, N$ and $C_{DT'} = C_{T'D}^\top \in \mathbb{R}^{N\times 1}$. The proof for these statements is given in section A of the supplementary materials.

In summary, given a prior GRE and some observed values Y via some maps $\{D_n\}_{n=1}^N \subset E^*$ we can get a Gaussian posterior measure $P^{F|Y=y}$ for any $y \in \mathbb{R}^N$ on E. Two, or equivalently finitely many, linear functionals $T, T' \in E^*$ of F under the posterior will follow a multivariate Gaussian distribution and one can use (4) and (5) to calculate the mean vector and the covariance matrix.

3.1 Example: GPs With Continuous Paths

Before we give the promised example that links GP regression to GRE regression, we need to introduce some key results from functional analysis.

Let $\mathcal{X} \subset \mathbb{R}^D$ be compact with Borel σ-algebra $\mathcal{B}(\mathcal{X})$ and $C(\mathcal{X}, \mathbb{R})$ the space of continuous functions from $\mathcal{X}$ to $\mathbb{R}$ equipped with the standard supremum norm. Note that $C(\mathcal{X}, \mathbb{R})$ is a Banach space as long as $\mathcal{X}$ is compact. Denote by $R(\mathcal{X})$ the space of finite regular signed measures over $\mathcal{X}$ equipped with total variation norm (Rao and Rao, 1983, Section 2.4). The Riesz-Markov theorem (Rao and Rao, 1983; Royden and Fitzpatrick, 2010, Corollary 4.7.6) states that $C(\mathcal{X}, \mathbb{R})^* = R(\mathcal{X})$ in the sense that each $\mu \in R(\mathcal{X})$ gives an element of $C(\mathcal{X}, \mathbb{R})^*$ via $f \mapsto \int_{\mathcal{X}} f(x) d\mu(x)$ and for every $T \in C(\mathcal{X}, \mathbb{R})^*$ there exists a unique $\mu \in R(\mathcal{X})$ such that $Tf = \int_{\mathcal{X}} f(x) d\mu(x)$. For example the pointwise evaluation map $\pi_x(f) = f(x)$ corresponds to the Dirac measure δ_x based at x for which we write $\pi_x = \delta_x$. For a covariance operator $C \colon C(\mathcal{X}, \mathbb{R})^* \to C(\mathcal{X}, \mathbb{R})$ we set $C\mu$ to be CT where T is the unique element of $C(\mathcal{X}, \mathbb{R})^*$ such that $Tf = \int_{\mathcal{X}} f(x) d\mu(x)$.

We now establish the connection between GREs and GPs when $E = C(\mathcal{X}, \mathbb{R})$. Let $G \sim GP(0, k)$ be a Gaussian process over a compact subset $\mathcal{X} \subset \mathbb{R}^d$ with kernel k and zero mean. Zero mean is for simplicity, non-zero can be handled straightforwardly.

Assume the GP has paths in $C(\mathcal{X}, \mathbb{R})$ with probability one. A standard result which provide a sufficient condition is the Kolmogorov continuity theorem (Øksendal, 2003, Theorem 2.2.3), if k is translation invariant then a condition regarding the decay of k is provided by Adler and Taylor (2007, Corollary 1.5.5) and a condition regarding the spectral measure (Rasmussen and Williams, 2005, Chapter 4.2.1) of k by Adler and Taylor (2007, Page 22).

Following Lifshits (2012, Example 2.4), see also Rajput and Cambanis (1972), any GP with almost surely continuous paths can be identified with a GRE F taking values in $E = C(\mathcal{X}, \mathbb{R})$, denote the corresponding GM by P. The covariance operator C of F is given as

$$C\nu(\cdot) = \int_{\mathcal{X}} k(\cdot, x') d\nu(x') \tag{6}$$

$$(C\nu, \mu)_E = \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, x') d\nu(x') d\mu(x), \tag{7}$$

for $\mu, \nu \in R(\mathcal{X})$. Using the identification of pointwise evaluation and Dirac measures mentioned above

$$\begin{aligned} \mathrm{Cov}_P[F(x), F(x')] &= \mathrm{Cov}[(F, \delta_x)_E, (F, \delta_{x'})_E] \\ &= (C\delta_x, \delta_{x'})_E \\ &= \int \int k(t, t') d\delta_x(t) d\delta_{x'}(t') \\ &= k(x, x'), \end{aligned}$$

for any $x, x' \in \mathcal{X}$ as expected.

In standard GP regression one observes corrupted pointwise information about the unknown function at a collection of points $X = \{x_n\}_{n=1}^N \subset \mathcal{X}$. So in the notation of the previous subsection $D_n = \delta_{x_n}$ is the map through which our observations are viewed.

Suppose we want to make a prediction at two new points $x, x' \in \mathcal{X}$. This corresponds to the measures $T = \delta_x$ and $T' = \delta_{x'}$ and we know that $\big(F(x), F(x')\big)|Y = y$ is multivariate Gaussian. From (4) we calculate the mean as

$$\widetilde{m}(x) = (\widetilde{m}, \delta_x)_E = k_{xX}(k_{XX} + \sigma^2 I_N)^{-1} y,$$

and similarly for $m(x')$. Furthermore the covariance between $F(x)$ and $F(x')$ under the posterior is given by formula (5) as

$$(\widetilde{C}\delta_x, \delta_{x'})_E = k(x, x') - k_{xX}(k_{XX} + \sigma^2 I_N)^{-1} k_{Xx'},$$

where k_{XX} is the matrix with n, n'-th entry $k(x_n, x_{n'})$ and $k_{xX} = (k(x, x_1), \ldots, k(x, x_N)) = k_{Xx}^\top$. This is the standard formula for the posterior mean and covariance of a GP given noisy pointwise observations (Rasmussen and Williams, 2005).

In summary, GRE regression on $E = C(\mathcal{X}, \mathbb{R})$ with observation functionals $D_n = \delta_{x_n}$, $n = 1, \ldots, N$ recovers standard GPR.

4 VARIATIONAL INFERENCE FOR GAUSSIAN RANDOM ELEMENTS

The posterior expressions (4) and (5) can have high computational cost since the matrix inverse term has naive cost $O(N^3)$. To avoid this cost the variational approximation paradigm is often used where $P^{F|Y=y}$ is approximated by selecting a measure in a candidate family $\mathcal{Q} \subset \mathcal{P}(E)$ that is optimal according to some divergence. The earliest works on this method are Titsias (2009a); Hensman et al. (2013) and, as discussed in the introduction, this area has received a lot of attention and innovation in the GP community recently (Hensman et al., 2018; Dutordoir et al., 2020; Lemercier et al., 2021).

We will now describe variational inference for Gaussian random elements in an abstract Banach space E. Much is owed to the work of Matthews et al. (2016), which formulated the important equations in the context of Gaussian processes. The following presentation applies in generality of Banach spaces which is the most general derivation the authors are aware of.

The following are desired properties of the family $\mathcal{Q}$ of candidates to approximate the posterior

1. Predictions involving any $Q \in \mathcal{Q}$ must be computationally tractable and less expensive than the true posterior.
2. $\mathcal{Q}$ contains measures that give a good approximation for the true posterior $P^{F|Y=y}$.
3. A measure of *closeness* between each $Q \in \mathcal{Q}$ and $P^{F|Y=y}$ must be tractable and cheap to evaluate.

Variational family: The idea in the construction of $\mathcal{Q}$ is to parameterise certain features of the target posterior with a multivariate Gaussian, the hope being that these features will represent the target posterior well even though there may be less features than the number of points observed.

First choose $M \in \mathbb{N}$ elements from the dual $\{L_m\}_{m=1}^M \subset E^*$, the *features*, and set $L = (L_1, \dots, L_M)$, $L\colon E \to \mathbb{R}^M$. Denote $U_m := (F, L_m)_E$, $m = 1, \dots, M$ and $U := (U_1, \dots, U_M)$. Define $Q^L := \mathcal{N}(\mu, \Sigma) \in \mathcal{P}(\mathbb{R}^M)$ for some mean vector $\mu \in \mathbb{R}^M$ and covariance matrix $\Sigma \in \mathbb{R}^{M\times M}$. Starting with a Q^L of this form we obtain a member of the approximating family Q as

$$Q(A) = \int_A \left(\frac{dQ^L}{dP^L} \circ L\right)(f) dP(f)$$

where $A \in \mathcal{B}(E)$ and dQ^L/dP^L is the Radon-Nikodym derivative of Q^L with respect to P^L and P^L is the law of U under the prior equal to $\mathcal{N}((m, L)_E, C_{LL})$. The idea is that Q^L dictates the behaviour of Q on the features L.

While this formulation of a candidate $Q \in \mathcal{Q}$ may seem obtuse, in Theorem 2 in the Supplement it is shown

$$Q(A) = \int_{\mathbb{R}^M} \mathbb{P}\big(F \in A \big| U = u\big) dQ^L(u),$$

which is used to deduce that each $Q \in \mathcal{Q}$ is a GM with mean m_Q

$$(m_Q, T)_E = (m, T)_E + C_{LL}^{-1}\big(\mu - (m, L)_E\big)C_{LT} \quad (8)$$

for all $T \in E^*$ and covariance operator C_Q

$$\begin{aligned}&\big(C_Q T, T'\big)_E \\ &\quad = (CT, T')_E + C_{TL}C_{LL}^{-1}(\Sigma - C_{LL})C_{LL}^{-1}C_{LT'} \quad (9)\end{aligned}$$

for all $T, T' \in E^*$.

The variational parameters of this family are μ, Σ and potentially parameters that appear in the specification of the inducing features L. For ease of notation denote all of these parameters by η and the Q corresponding to this choice by Q_η.

Measure of closeness: The Kullback-Leibler (KL) divergence is the measure of closeness employed. For $P, Q \in \mathcal{P}(E)$ with Q absolutely continuous with respect to P, denoted $Q \ll P$

$$KL(Q, P) = \int_E \log\left(\frac{dQ}{dP}(f)\right) dQ(f),$$

and $KL(Q, P)$ is infinite if Q is not absolutely continuous with respect to P. The variational parameter $\eta \in \Gamma$ is then selected by minimising the KL

$$\eta^* \in \arg\min_\eta KL\big(Q_\eta, P^{F|Y=y}\big).$$

After η^* has been determined, the posterior is approximated with Q_{η^*} which we denote by Q^* for ease of notation.

The choice of $\mathcal{Q}$ means one may rewrite the KL as

$$\begin{aligned}&KL(Q, P^{F|Y=y}) \\ &= KL(Q, P) - \mathbb{E}_Q[\log p(y|F)] + \log p(y)\end{aligned}$$

where $\mathbb{E}_Q[\log p(y|F)] := \int \log p(y|F = f) dQ(f)$, see Theorem 4 in the Supplement.

Optimisation: Optimisation with respect to KL is performed by optimising the evidence based lower bound (ELBO) defined $\mathcal{L} := -KL(Q, P) + \mathbb{E}_Q[\log p(y|F)]$.

The user can now optimise the parameters η, which we recall are μ, Σ, analytically to obtain the optimal $Q^* \in \mathcal{Q}$ within the candidate family. The parameters μ, Σ have a closed form expression and cost $\mathcal{O}(NM^2 + M^3)$ (Titsias, 2009a). These optimal choices for μ, Σ, given fixed L, are given in Theorem 4 in the Supplement. The resulting optimal mean and covariance operators, denoted m_{Q^*} and C_{Q^*} satisfy

$$\begin{aligned}&(m_{Q^*}, T)_E \\ &\qquad = C_{TL}\big(\sigma^2 C_{LL} + C_{LD}C_{DL}\big)^{-1}C_{LD}y \quad (10)\\ &(C_{Q^*}T, T')_E \\ &\qquad = (CT, T')_E - C_{TL}C_{LL}^{-1}C_{LT'} \\ &\qquad + C_{TL}\big(C_{LL} + \frac{1}{\sigma^2}C_{LD}C_{DL}\big)^{-1}\big)C_{LT'},\end{aligned}$$

for all $T, T' \in E^*$. See Theorem 4 in the Supplement for a proof.

Alternatively, a user could numerically optimise η using a factorisation of $\mathcal{L}$ over N to make use of batch size optimisation (Hensman et al., 2013). This leads to complexity $\mathcal{O}(N_B M^2 + M^3)$, where $N_B \in \mathbb{N}$ is the batch size.

The factorised ELBO is normally used for really large data sets and in this case the bottleneck is the inversion of C_{LL} which causes the $\mathcal{O}(M^3)$ complexity term. Therefore it is vital for practitioners to choose L which

result in C_{LL} being easy to invert, for example making C_{LL} be diagonal.

The next section investigates common choices of L in the literature and derives a unifying perspective using GREs, crucial for understanding the different choices.

5 FUNCTIONAL ANALYSIS VIEW ON INDUCING FEATURES

In this section several variational approaches are recovered within the GRE framework. The goal is obtaining a unified perspective and greater generality of the derivations.

As described in Section 3.1, the starting point for our analysis is a GRE $F \sim \mathcal{N}(m, C)$ in $E = C(\mathcal{X}, \mathbb{R})$ whose corresponding measure on E is denoted P. This GRE is then conditioned upon corrupted pointwise observations $\{x_n\}_{n=1}^N \subset \mathcal{X}$ such that $Y_n = (F, \delta_{x_n})_E + \epsilon_n$ with $\epsilon_1, \dots, \epsilon_N \sim \mathcal{N}(0, \sigma^2)$.

Different choices of features $\{L_m\}_{m=1}^M \subset C(\mathcal{X}, \mathbb{R})^*$ shall lead to the original inducing point approach (Titsias, 2009a), inter-domain features (Lázaro-Gredilla and Figueiras-Vidal, 2009) and variational Fourier features (Hensman et al., 2018).

While the first two examples appear pedestrian the third crucially relies upon the GRE to reveal how Fourier features actually behave and under what conditions they are valid, greatly expanding their scope beyond the example in Hensman et al. (2018).

Only the covariances C_{LL}, C_{LT} of the features are derived since this is all that is needed to compute the variational mean m_Q and covariance operator C_Q of the approximating Q, see (8) and (9).

The prediction map T will always be a single point evaluation at an arbitrary point $x \in \mathcal{X}$. The case of other choices of T, in particular point evaluation at multiple locations, is straightforward. The term C_{LL} is the bottleneck term in the computation, as discussed in the previous section.

5.1 Background in Functional Analysis

In this section we introduce some basic terminology and results from functional analysis. The reader unfamiliar with these tools is referred to Chapter 13 of Royden and Fitzpatrick (2010) for additional information.

Let E be a Banach space and $\mathcal{X} \subset \mathbb{R}^D$ be compact[1]. Typical examples of Banach spaces E are the space of continuous functions $C(\mathcal{X}, \mathbb{R})$ endowed with the supremums norm and the space $L^2(\mathcal{X}, \mathbb{R})$ of equivalence classes of square integral functions, which is even a Hilbert space.

Linear operators A linear map $L : E \to W$ (typically called operator) between two Banach spaces $\big(E, \|\cdot\|_E\big)$ and $\big(W, \|\cdot\|_W\big)$ is called bounded iff

$$\|L\| := \sup_{v \in E} \frac{\|Lv\|_W}{\|v\|_E}$$

is finite. For every operator $L : E \to W$ we define the adjoint operator $L^* : W^* \to E^*$ via $L^*(\phi)(v) := \phi(Lv)$ for all $\phi \in W^*$, $v \in E$. This can be equivalently expressed with the help of the dual pairing as $\big(v, L^*\phi\big)_E = \big(Lv, \phi\big)_W$. If E and W are Hilbert spaces we can interpret the dual pairing as the respective Hilbert space inner-products via the Riesz-representation theorem.

Kernel integral operator Let k be a (continuous) kernel on $\mathcal{X}$ and define $T_k : L^2(\mathcal{X}, \mathbb{R}) \to L^2(\mathcal{X}, \mathbb{R})$ as $T_k f(x) = \int_{\mathcal{X}} k(x, x') f(x') dx'$. The operator T_k is called kernel (integral) operator and is well-studied in machine learning (cp. Steinwart and Scovel, 2012). It can be easily shown that T_k is bounded and self-adjoint, i.e. $T_k^* = T_k$.

Linear transformation of Gaussian Measures Let $L : E \to W$ be a bounded linear operator and $F \sim \mathcal{N}(m, C)$ be a GRE in E. Then $L(F)$ is a GRE in W with mean Lm and covariance operator LCL^* (Hairer, 2009, chapter 3.3).

5.2 Inducing Points

The inducing-points framework of Titsias (2009a) chooses the inducing features as pointwise evaluations $L_m = \pi_{z_m}$, which corresponds to the measure δ_{z_m} for some set of points $\{z_m\}_{m=1}^M \subset \mathcal{X}$.

Substituting this choice of L into (7)

$$\begin{aligned}(C_{LL})_{mm'} &= \mathrm{Cov}_P(L_m F, L_{m'} F) \\ &= \mathrm{Cov}_P[(F, \delta_{z_m})_E, (F, \delta_{z_m})_E] \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, x') d\delta_{z_m}(x) \delta_{z_{m'}}(x') \\ &= k(z_m, z_{m'}),\end{aligned}$$

and

$$\begin{aligned}(C_{TL})_m &= \mathrm{Cov}_P(TF, L_m F) \\ &= \mathrm{Cov}_P[(F, \delta_x)_E, (F, \delta_{z_m})_E] \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} k(t, t') d\delta_x(t) \delta_{z_m}(t') = k(x, z_m).\end{aligned}$$

[1] The compactness of $\mathcal{X}$ is required to make $C(\mathcal{X}, \mathbb{R})$ a Banach space, but is not necessary for the Hilbert space structure of $\mathcal{L}^2(\mathcal{X}, \mathbb{R})$.

Combining these calculations with (8), (9) we recover the same formula as seen in the original inducing point derivation (Titsias, 2009a).

5.3 Inter-domain Features

The natural space to realise inter-domain features (Lázaro-Gredilla and Figueiras-Vidal, 2009) is $L^2(\mathcal{X}, \mathbb{R})$ since they involve an inner product on $L^2(\mathcal{X}, \mathbb{R})$. To this end map the GRE F that takes values in $C(\mathcal{X}, \mathbb{R})$ into $L^2(\mathcal{X}, \mathbb{R})$ via the canonical embedding $\iota : C(\mathcal{X}, \mathbb{R}) \to L^2(\mathcal{X}, \mathbb{R})$, $f \mapsto f$. The adjoint operator of ι is given as $\iota^* : L^2(\mathcal{X}, \mathbb{R}) \to R(\mathcal{X})$, $f \mapsto \int_{(\cdot)} f(x)dx$, which can be easily verified. As explained in Section 3.1 the space $C(\mathcal{X}, \mathbb{R})^*$ can be identified with the space of signed Borel measures on $\mathcal{B}(\mathcal{X})$ which explains why $\iota^*(f)$ is an element of $R(\mathcal{X})$ and therefore accepts as input a set to integrate over.

It follows from $\|f\|_{L^2(\mathcal{X},\mathbb{R})} \leq \sqrt{\lambda(\mathcal{X})}\|f\|_\infty$, where $\|\cdot\|_\infty$ is the supremumsnorm and λ the Lebesgue measure, that ι is a bounded, linear operator from $C(\mathcal{X}, \mathbb{R})$ to $L^2(\mathcal{X}, \mathbb{R})$. We can use the result mentioned in Section 5.1 to conclude ιF is a GRE in $L^2(\mathcal{X}, \mathbb{R})$ with covariance operator $\iota C \iota^*$ given as

$$(\iota C \iota^*)(g) = C\big(\iota^*(g)\big) \tag{11}$$
$$= \int_{\mathcal{X}} k(\cdot, x')g(x')dx', \tag{12}$$

where (12) is simply from the definition of the covariance operator of the GRE on $C(\mathcal{X}, \mathbb{R})$. By (12) $\iota C \iota^*$ coincides with the well-studied integral operator described in Section 5.1 and denoted as T_k.

Inter-domain features can be written in our notation as

$$L_m F = \langle \iota F, g_m \rangle_{L^2(\mathcal{X},\mathbb{R})} = \int_{\mathcal{X}} F(x) g_m(x) dx$$

for some collection $\{g_m\}_{m=1}^M \subset L^2(\mathcal{X}, \mathbb{R})$. Using the definition of an adjoint operator this is equal to $L_m F = (F, \iota^* g_m,)_E$ so $L_m = \iota^* g_m$.

Substituting into (7)

$$\begin{aligned}(C_{LL})_{mm'} &= \mathrm{Cov}_P(L_m F, L_{m'} F) \\ &= \mathrm{Cov}_P[(F, \iota^* g_m)_E, (F, \iota^* g_{m'})_E] \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, x') g_m(x) g_{m'}(x') dx dx', \end{aligned} \tag{13}$$

and

$$\begin{aligned}(C_{TL})_m &= \mathrm{Cov}_P(TF, L_m F) \\ &= \mathrm{Cov}_P[(F, \delta_x)_E, (F, \iota^* g_m)_E] \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} k(t, x') g_m(x') d\delta_x(t) dx' = T_k(x), \end{aligned} \tag{14}$$

which agrees with the original derivation of inter-domain features (Lázaro-Gredilla and Figueiras-Vidal, 2009).

5.4 Fourier Features

Fourier features are defined as a reproducing kernel Hilbert space (RKHS) inner product between F and trigonometric functions (Hensman et al., 2018).

An RKHS is a Hilbert space of functions that, as the name suggests, is associated with a kernel. Namely, given a kernel k the RKHS is the unique Hilbert space of functions mapping from $\mathcal{X}$ to $\mathbb{R}$, which we denote H_k, such that $k(\cdot, x) \in H_k$ for all $x \in \mathcal{X}$ and $\langle f, k(\cdot, x)\rangle_k = f(x)$ for all $f \in H_k$ and $x \in \mathcal{X}$, where $\langle \cdot, \cdot \rangle_k$ denotes the inner product on H_k. This latter property is called the reproducing property. For more on the theory of RKHS consult Berlinet and Thomas-Agnan (2004).

The idea of Fourier features is to observe an inner product not in $L^2(\mathcal{X}, \mathbb{R})$, as was done in inter-domain features, but instead to observe an inner product in H_k. Namely, Hensman et al. (2018) use features that would be written in our framework as $L_m F = \langle F, g_m \rangle_k$ where $\{g_m\}_{m=1}^M$ are the first M elements of the Fourier basis. However, it is well known that $F \notin H_k$ almost surely (Lukić and Beder, 2001) therefore the aforementioned choice of L *cannot* be used without extra justification.

Hensman et al. (2018) provided justification in the particular case when k is a Matérn kernel with certain parameters over $\mathcal{X} = \mathbb{R}$. An explicit form of the Matérn RKHS inner product is used and the core of the argument is the g_m are "very regular" in an appropriate sense to compensate for the way that almost surely F is not regular enough to be contained in H_k (Hensman et al., 2018, Section 3.3.1).

A rigorous justification is now given for these type of features. The derivation is more general than just the one-dimensional Matérn case and a condition on which functions g_m can be used instead of the Fourier basis is provided.

For the rest of this section we will assume that k is continuous. Recall the kernel integral operator T_k defined above, the square root $T_k^{1/2}$ is an isomet-

ric isomorphism between $L^2(\mathcal{X},\mathbb{R})$ and H_k (Steinwart and Christmann, 2008, Theorem 4.51) meaning that $T_k^{1/2}(L^2(\mathcal{X},\mathbb{R})) = H_k$ and

$$\langle T_k^{1/2} f, T_k^{1/2} g\rangle_k = \langle f, g\rangle_{L^2(\mathcal{X},\mathbb{R})}, \tag{15}$$

for all $f, g \in L^2(\mathcal{X},\mathbb{R})$. The moral of this result is that $T_k^{1/2}$ bestows upon element of $L^2(\mathcal{X},\mathbb{R})$ just enough regularity to be in H_k. This notion of adding regularity will tie into the notion of "very regular" employed by Hensman et al. (2018).

Define features $L_m = \iota^* f_m$ for some $\{f_m\}_{m=1}^M \subset L^2(\mathcal{X},\mathbb{R})$ so

$$L_m F = (F, \iota^* f_m)_E = \langle \iota F, f_m\rangle_{L^2(\mathcal{X},\mathbb{R})}$$

where $\iota\colon C(\mathcal{X},\mathbb{R}) \to L^2(\mathcal{X},\mathbb{R})$ is the inclusion operator used in the previous subsection. Then

$$\begin{aligned} \text{“}\langle F, T_k f_m\rangle_k = \text{”} &\langle T_k^{1/2}\iota F, T_k^{1/2} f_m\rangle_k && (16)\\ &= \langle \iota F, f_m\rangle_{L^2(\mathcal{X},\mathbb{R})} && (17)\\ &= (F, \iota^* f_m)_E = L_m F. && (18)\end{aligned}$$

The first expression is in quotes since $F \notin H_k$ almost surely (Lukić and Beder, 2001) so the expression has no real meaning. We include it though since if one were to suspend reality then the equality in quotes would be valid since $T_k^{1/2}$ is self-adjoint on H_k so it can be borrowed from $T_k f_m$. The second term in (16) is well defined since $T_k^{1/2}$ maps from $L^2(\mathcal{X},\mathbb{R})$ to H_k. The move to (17) is facilitated by the isometry (15).

The main idea of what is happening in (16) is F is borrowing a $T_k^{1/2}$ from $T_k f_m$ to be able to live in the RKHS. This is happening explicitly in the calculations done by Hensman et al. (2018) in the Matérn case. Indeed, the way that $T_k f_m$ is the image of f_m under *two* applications of $T_k^{1/2}$, rather than just the standard one needed to be in the RKHS, is the mathematical explanation of the notion of "very regular" that was alluded to by Hensman et al. (2018) since it has had two portions of the regularity provided by $T_k^{1/2}$.

It is interesting to see that RKHS inner product feature (16) can be reduced to (18) which is simply using $T_k f_m$ as an inter-domain feature. So using the inter-domain formula for C_{LL} (13) gives

$$\begin{aligned} (C_{LL})_{mm'} &= \text{“}\mathrm{Cov}_P(\langle F, T_k f_m\rangle_k, \langle F, T_k f_{m'}\rangle_k)\text{”}\\ &= \mathrm{Cov}_P[(F, \iota^* f_m)_E, (F, \iota^* f_{m'})_E]\\ &= \langle T_k f_m, f_{m'}\rangle_{L^2(\mathcal{X},\mathbb{R})}\\ &= \langle T_k^{1/2} f_m, T_k^{1/2} f_{m'}\rangle_{L^2(\mathcal{X},\mathbb{R})} && (19)\\ &= \langle T_k f_m, T_k f_{m'}\rangle_k, && (20)\\ & && (21)\end{aligned}$$

where (19) from the fact that $T_k^{1/2}$ is self-adjoint considered as operator on $L^2(\mathcal{X},\mathbb{R})$ and (20) is using the isometry between $L^2(\mathcal{X},\mathbb{R})$ and H_k and. Similarly,

$$\begin{aligned} (C_{TL})_m &= \text{“}\mathrm{Cov}_P(\langle F, k(x,\cdot)\rangle_k, \langle F, T_k f_m\rangle_k)\text{”}\\ &= \mathrm{Cov}_P[(F, \delta_x)_E, (F, \iota^* f_m)_E]\\ &= T_k f_m(x) && (22)\end{aligned}$$

where (22) is using (14). The connection to Hensman et al. (2018) is revealed once one sets $g_m = T_k f_m$ where g_m are the features employed by Hensman et al. (2018). In particular (20) and (22) are equal to Equation 61 and Equation 60, respectively, in Hensman et al. (2018).

As was done by Hensman et al. (2018) it makes sense from a practical point of view to define $T_k f_m$ without explicitly choosing f_m. Our derivation shows this may be performed without dependence on kernel parameters as long as g_m is in the range of T_k for all the kernel parameters that could be considered.

Our derivation is general enough to justify the use of Fourier features in Dutordoir et al. (2020), where zonal kernels in more than one input dimension are used. Their choice of g_m corresponds to eigenfunctions in the Mercer expansion of the kernel which is always in the image of T_k.

6 GAUSSIAN RANDOM ELEMENTS AND THE NYSTRÖM METHOD

In this section we demonstrate how the GRE framework reveals further connections between Gaussian process regression and the Nyström approximation for kernel Ridge regression (KRR). These connections have been known for a while and received some attention recently (Parzen, 1961; Wahba, 1990; Kanagawa et al., 2018; Wild et al., 2021)

Let k be a kernel, H_k the corresponding RKHS and $\{x_n, y_n\}_{n=1}^N \subset \mathcal{X}\times\mathbb{R}$ be paired observations. In Nyström KRR (Williams and Seeger, 2001) we seek to minimise the empirical risk over a finite dimensional subspace $\mathcal{M} \subset H_k$

$$\widehat{f} \coloneqq \operatorname*{argmin}_{f\in\mathcal{M}} \frac{1}{N}\sum_{n=1}^N \big(f(x_n) - y_n\big)^2 + \lambda\|f\|_k^2, \tag{23}$$

where $\lambda > 0$ is a regularisation parameter and $\widehat{f}$ is called the Nyström approximation over $\mathcal{M}$.

The subspace $\mathcal{M}$ that is typically selected is $\mathrm{span}\big(\{k(\cdot, z_m)\}_{m=1}^M\big)$, where $\{z_m\}_{m=1}^M \subset \mathcal{X}$ are user chosen points, referred to as landmark points. Most research has focused on different sampling approaches

for the landmark points to guarantee high quality approximations (Rudi et al., 2015; Musco and Musco, 2016; Li et al., 2016).

The GRE perspective facilitates a generalised view in which it becomes clear how the choice of features L in variational GRE regression, see Section 4, corresponds to the choice of subspace $\mathcal{M}$ in Nyström KRR.

Theorem 1. *Let $F \sim \mathcal{N}(0, C)$ be a GRE in $E = C(\mathcal{X}, \mathbb{R})$ with covariance operator C as defined in* (6) *and assumed pointwise noisy data is observed as described in Section 3.1. For $\{\mu_m\}_{m=1}^M \subset R(\mathcal{X})$ let $L_m = \mu_m$ be the features used in the variational approximation. Set $\mathcal{M} = \{C\mu_m\}_{m=1}^M$ where $C\mu_m = \int k(\cdot, x')d\mu_m(x')$ as the approximating family in the Nyström approximation. Then for $\sigma^2 = N\lambda$ the Nyström KRR estimator $\widehat{f}$ in Section 6 is equal to the mean m_{Q^*}, given by* (10)*, of the optimal Q^* from the variational family $\mathcal{Q}$.*

We give the proof and an explicit form of $\widehat{f}$ and m_{Q^*} in Section C of the Supplement.

The GRE view reveals that the connecting element between the Nyström approximations and variational GPs is via the signed measures $\{\mu_m\}_{m=1}^M$. Knowledge of the signed measures either from inspecting the elements of the set $\mathcal{M}$ in the Nyström approximation or the projections L_m in variational GPs will allow to translate results from one field to another. This connection was previously only known for the special case of inducing inducing points $\mu_m := \delta_{z_m}$ outlined in Wild et al. (2021). This opens the door to applying the theory of Nyström KRR error bounds to variational GPs to gain a better understanding of the latter's approximation properties. Vice versa, recent advances in variational GPR approaches, for example, variational Fourier features, could be leveraged in the context of KRR Nyström, as they simply correspond to a particular choice of $\mathcal{M}$.

7 CONCLUSION

We have outlined the GRE framework as a technical tool to provide a unified and generalising perspective to variational GP approximation. Along the way we have seen how existing choices of features, previously thought distinct, are in fact highly related and how they can be derived in wider settings than those currently employed. Finally, we related the posterior mean of variational GP approximations with a Nyström KRR approximation which offers a new lens to view the impact of different feature choices in variational GP approximations.

Acknowledgements

The research of George Wynne was supported by an EPSRC Industrial CASE award [EP/S513635/1] in partnership with Shell UK Ltd.

References

R. J. Adler. An introduction to continuity, extrema, and related topics for general gaussian processes. *Lecture Notes-Monograph Series*, 12:i–155, 1990.

R. J. Adler and R. Taylor. *Random Fields and Geometry*. Springer New York, 2007.

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

V. Bogachev. *Gaussian Measures*. American Mathematical Society, 1998.

C.-A. Cheng and B. Boots. Incremental variational sparse gaussian process regression. *Advances in Neural Information Processing Systems*, 29:4410–4418, 2016.

C.-A. Cheng and B. Boots. Variational inference for gaussian process models with linear complexity. *arXiv preprint arXiv:1711.10127*, 2017.

L. Csató and M. Opper. Sparse on-line gaussian processes. *Neural computation*, 14(3):641–668, 2002.

G. Da Prato. *An Introduction to Infinite-Dimensional Analysis*. Springer Berlin Heidelberg, 2006.

V. Dutordoir, N. Durrande, and J. Hensman. Sparse Gaussian processes with spherical harmonic features. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2793–2802. PMLR, 2020.

J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *arXiv preprint arXiv:1809.11165*, 2018.

S. Ghosal and A. van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

M. Hairer. An introduction to stochastic pdes. *arXiv preprint arXiv:0907.4178*, 2009.

P. R. Halmos. *Measure theory*, volume 18. Springer, 2013.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

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

O. Kallenberg and O. Kallenberg. *Foundations of modern probability*, volume 2. Springer, 1997.

M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv:1807.02582*, 2018.

A. Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.

M. Lázaro-Gredilla and A. R. Figueiras-Vidal. Inter-domain gaussian processes for sparse inference using inducing features. In *NIPS*, volume 22, pages 1087–1095. Citeseer, 2009.

F. Leibfried, V. Dutordoir, S. John, and N. Durrande. A tutorial on sparse gaussian processes and variational inference. *arXiv preprint arXiv:2012.13962*, 2020.

M. Lemercier, C. Salvi, T. Cass, E. V. Bonilla, T. Damoulas, and T. Lyons. Siggpde: Scaling sparse gaussian processes on sequential data. In *International Conference on Machine Learning*. PMLR, 2021.

C. Li, S. Jegelka, and S. Sra. Fast dpp sampling for nystrom with application to kernel methods. In *International Conference on Machine Learning*, pages 2061–2070. PMLR, 2016.

M. Lifshits. *Lectures on Gaussian Processes*. Springer Berlin Heidelberg, 2012.

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

A. G. d. G. Matthews, J. Hensman, R. Turner, and Z. Ghahramani. On sparse variational methods and the kullback-leibler divergence between stochastic processes. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 231–239. PMLR, 2016.

C. Musco and C. Musco. Recursive sampling for the nystr\" om method. *arXiv preprint arXiv:1605.07583*, 2016.

B. Øksendal. *Stochastic Differential Equations*. Springer Berlin Heidelberg, 2003.

E. Parzen. An approach to time series analysis. *The Annals of Mathematical Statistics*, pages 951–989, 1961.

B. S. Rajput. Gaussian measures on lp spaces, $1 \leq p < \infty$. *Journal of Multivariate Analysis*, 2(4):382–403, 1972.

B. S. Rajput and S. Cambanis. Gaussian processes and gaussian measures. *The Annals of Mathematical Statistics*, 43(6):1944–1952, 1972.

K. B. Rao and M. B. Rao. *Theory of charges: a study of finitely additive measures*. Academic Press, 1983.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

H. L. Royden and M. P. Fitzpatrick. *Real Analysis - Fourth Edition*. Pearson, 2010.

A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. In *NIPS*, pages 1657–1665, 2015.

H. Salimbeni, C.-A. Cheng, B. Boots, and M. Deisenroth. Orthogonally decoupled variational gaussian processes. *arXiv preprint arXiv:1809.08820*, 2018.

M. W. Seeger, C. K. Williams, and N. D. Lawrence. Fast forward selection to speed up sparse gaussian process regression. In *International Workshop on Artificial Intelligence and Statistics*, pages 254–261. PMLR, 2003.

E. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. *Advances in neural information processing systems*, 18:1257, 2006.

E. Snelson and Z. Ghahramani. Local and global sparse gaussian process approximations. In *Artificial Intelligence and Statistics*, pages 524–531. PMLR, 2007.

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

I. Steinwart and C. Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

A. M. Stuart. Inverse problems: A bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

M. Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*, pages 567–574. PMLR, 2009a.

M. K. Titsias. Variational model selection for sparse gaussian process regression. *Report, University of Manchester, UK*, 2009b.

G. Wahba. *Spline models for observational data*. SIAM, 1990.

K. Wang, G. Pleiss, J. Gardner, S. Tyree, K. Q. Weinberger, and A. G. Wilson. Exact gaussian processes on a million data points. *Advances in Neural Information Processing Systems*, 32:14648–14659, 2019.

V. Wild, M. Kanagawa, and D. Sejdinovic. Connections and equivalences between the nystr\" om method and sparse variational gaussian processes. *arXiv preprint arXiv:2106.01121*, 2021.

C. Williams and M. Seeger. Using the nyström method to speed up kernel machines. In *Proceedings of the 14th annual conference on neural information processing systems*, pages 682–688, 2001.

Supplementary Material: Variational Gaussian Processes: A Functional Analysis View

A Proofs of Section 3: Gaussian Random Element Regression

Recall that the posterior measure in Section 3 satisfies $P^{F|Y=y}(A) = \int_A \frac{p(y|f)}{p(y)} dP(f)$ for all $A \in \mathcal{B}(E)$.

Theorem 2. *1. For any measurable $h : E \to \mathbb{R}^S$*

$$P^{F|Y=y}\big(h^{-1}(B)\big) = \mathbb{P}(h(F) \in B | Y = y). \tag{24}$$

2. The posterior measure $P^{F|Y=y}$ is Gaussian with mean $\widetilde{m}$ satisfying

$$\big(\widetilde{m}, T\big)_E = (m, T)_E + C_{TD}(C_{DD} + \sigma^2 I_N)^{-1} y, \tag{25}$$

for any $T \in E^*$ and posterior covariance operator $\widetilde{C} : E^* \to E$ satisfying

$$(\widetilde{C}T, T')_E = (CT, T')_E - C_{TD}(C_{DD} + \sigma^2 I_N)^{-1} C_{DT'}, \tag{26}$$

for all $T, T' \in E^*$.

Proof. As $P^{F|Y=Y}$ is a regular version of the conditional probability of $F|Y = y$ it satisfies

$$P^{F|Y=y}(A) = \mathbb{P}(F \in A | Y = y) \tag{27}$$

for any fixed $y \in \mathbb{R}^N$, $A \in \mathcal{B}(E)$ (Klenke, 2013, Definition 8.28). Hence for $A = h^{-1}(B)$ with $h : E \to \mathbb{R}^S$ measurable and $B \in \mathcal{B}(\mathbb{R}^N)$

$$\begin{aligned} P^{F|Y=y}(h^{-1}(B)) &= \mathbb{P}(F \in h^{-1}(B) | Y = y) && (28)\\ &= \mathbb{P}(h(F) \in B | Y = y), && (29) \end{aligned}$$

where the last line is using $\{\omega \in \Omega \colon F(\omega) \in h^{-1}(B)\} = \{\omega \in \Omega \colon h(F(\omega)) \in B\}$, this proves the first statement.

For the second statement, we can use equation (29) with the choice $h = (T, T') \colon E \to \mathbb{R}^2$ for two arbitrary $T, T' \in E^*$. Recall that the random vector $X \colon \Omega \to \mathbb{R}^N$ is Gaussian if and only if $\alpha^\top X \colon \Omega \to \mathbb{R}$ is Gaussian for all $\alpha \in \mathbb{R}^N$. Using this we will show random vector $V \coloneqq \big((F, T)_E, (F, T')_E, Y\big)$ in $\mathbb{R}^{N+2}$ is Gaussian. Take any $\alpha \in \mathbb{R}^{N+2}$ and set $\varphi = \sum_{n=1}^N \alpha_n D_n + \alpha_{N+1} T + \alpha_{N+2} T' \in E^*$, then

$$\alpha^\top V = \sum_{n=1}^{N+2} \alpha_n V_n = (F, \sum_{n=1}^N \alpha_n D_n + \alpha_{N+1} T + \alpha_{N+2} T')_E + \sum_{n=1}^N \alpha_n \epsilon_n = (F, \varphi)_E + \sum_{n=1}^N \alpha_n \epsilon_n, \tag{30}$$

is Gaussian for all $\alpha \in \mathbb{R}^{N+2}$ as it is sum of two independent Gaussian distributions, therefore V is Gaussian.

The mean μ_V is given by

$$\mathbb{E}[\alpha^\top V] = (m, \varphi)_E + 0 = \sum_{n=1}^N \alpha_n (m, D_n)_E + \alpha_{N+1}(m, T)_E + \alpha_{N+2}(m, T')_E =: \alpha^\top \mu_V, \tag{31}$$

and by the characterising property of the covariance operator C we get the covariance matrix Σ_V

$$\mathrm{Cov}[\alpha^\top V, \alpha^\top V] = (C\varphi, \varphi)_E + \sigma^2 \sum_{n=1}^N \alpha_n^2 = \alpha^\top \Sigma_V \alpha, \tag{32}$$

where the covariance matrix $\Sigma_V \in \mathbb{R}^{(N+2)\times(N+2)}$ is defined as

$$\Sigma_V \coloneqq \begin{bmatrix} (CT,T)_E & (CT,T')_E & C_{TD} \\ (CT',T)_E & (CT',T')_E & C_{T'D} \\ C_{DT} & C_{DT'} & C_{DD}+\sigma^2 I_N. \end{bmatrix} \tag{33}$$

We have showed $V \sim \mathcal{N}(\mu_V, \Sigma_V)$ and so using standard conditioning rules for multivariate Gaussians to condition on the last entry of V it is clear that $h(F)|Y = y$ is Gaussian with the desired mean and covariance. □

B Proofs of Section 4: Variational Inference for Gaussian Random Elements

Heavy use is made of the transformation rule for measures (Halmos, 2013, Theorem C). Let $(E, \mathcal{B}_E, \mu)$ be a measure space and $(W, \mathcal{B}_W)$ a measurable space. If $\mathcal{T} : E \to W$ and $\mathcal{G} : E \to [-\infty, \infty]$ are measurable, then

$$\int (\mathcal{G} \circ \mathcal{T})(x)\, d\mu(x) = \int \mathcal{G}(t)\, d\mu^{\mathcal{T}}(t),$$

where the left hand-side exists if and only if the right hand-side exists. Here again $\mu^{\mathcal{T}}(\cdot) \coloneqq \mu\big(\mathcal{T}^{-1}(\cdot)\big)$ denotes the image measure of μ induced by $\mathcal{T}$.

Measures in the variational family are Gaussian measures

Recall the definition of a measure in the variational family

$$Q(A) \coloneqq \int_A \left(\frac{dQ^L}{dP^L} \circ L\right)(f)\, dP(f)$$

for all $A \in \mathcal{B}(E)$ with $Q^L = \mathcal{N}(\mu, \Sigma)$ and $P^L = \mathcal{N}((m, L)_E, C_{LL})$.

Theorem 3. *1. For any $A \in \mathcal{B}(E)$*

$$Q(A) = \int_{\mathbb{R}^M} \mathbb{P}\big(F \in A \big| U = u\big) dQ^L(u).$$

2. The measure Q is a Gaussian measure with mean m_Q satisfying

$$(m_Q, T)_E = (m, T)_E + C_{LL}^{-1}\big(\mu - (m, L)_E\big), \tag{34}$$

for all $T \in E^*$ and covariance operator C_Q satisfying

$$\big(C_Q T, T'\big)_E = (CT, T')_E + C_{TL} C_{LL}^{-1}(\Sigma - C_{LL}) C_{LL}^{-1} C_{LT'}, \tag{35}$$

for all $T, T' \in E^*$.

Proof. Firstly, it suffices to prove the statement for sets of the form $A = T^{-1}(B)$, $B \in \mathcal{B}(\mathbb{R})$, $T \in E^*$, as two measures on a Banach space coincide if and only if they coincide for all sets of this form.

We want to apply the transformation rule for measures. To this end set $\mathcal{T} = (T, L) : E \to \mathbb{R}^{M+1}, \mathcal{T}(F) = (T(F), L(F))$ and $V = T(F)$ and $\pi_L\big((x_1, \ldots, x_{M+1})\big) = (x_2, \ldots, x_{M+1})$ for $x \in \mathbb{R}^{M+1}$. Clearly, $\{\omega \in \Omega \colon T(F(\omega)) \in B\} = \{\omega \in \Omega \colon \mathcal{T}(F(\omega)) \in B \times \mathbb{R}^M\}$ and $\pi_L \circ \mathcal{T} = L$.

From this we use the transformation rule

$$\begin{aligned} Q(\{T \in B\}) &= Q(\{\mathcal{T} \in B \times \mathbb{R}^M\}) \\ &= \int_{\mathcal{T} \in B \times \mathbb{R}^M} \left(\frac{dQ^L}{dP^L} \circ L\right)(f) dP(f) \\ &= \int_{\mathcal{T} \in B \times \mathbb{R}^M} \left(\frac{dQ^L}{dP^L} \circ \pi_L \circ \mathcal{T}\right)(f) dP(f) \\ &= \int_{B \times \mathbb{R}^M} \frac{dQ^L}{dP^L}(u) dP^{\mathcal{T}}(u, v) \\ &= \int_{B \times \mathbb{R}^M} \frac{q(u)}{p(u)} p(u, v) d(u, v), \end{aligned}$$

where we denote by $p(u)$ the probability density function (pdf) corresponding to P^L, by $q(u)$ the pdf corresponding to Q^L and $p(u,v)$ the joint pdf corresponding to $P^{\mathcal{T}}$. By $p(u,v) = p(v|U=u)p(u)$ and an application of Fubini

$$\begin{aligned} Q(\{T \in B\}) &= \int_{\mathbb{R}^M} \Big(\int_B p(v|U=u)dv \Big) q(u)du \\ &= \int_{\mathbb{R}^M} \mathbb{P}(V \in B|U=u)dQ^L(u), \end{aligned} \tag{36}$$

which proves the claim.

For the second statement we maintain the notation $V = (F,T)_E$. We need to show that (36) is Gaussian for any choice of $T \in E^*$. It is well known that the conditional distribution $V|U=u$ can be written as

$$V|(U=u) \overset{\mathcal{D}}{=} (m,T)_E + C_{TL}C_{LL}^{-1}\big(u - (m,L)_E\big) + W =: h(u,W), \tag{37}$$

where $\overset{\mathcal{D}}{=}$ means equality in distribution and $W \sim \mathcal{N}\big(0, (CT,T)_E - C_{TL}C_{LL}^{-1}C_{LT}\big)$ independently of U.

Since h is linear in U, we know that $h(U,W)$ is Gaussian for $U \sim \mathcal{N}(\mu,\Sigma)$ and the mean and variance can easily be calculated as

$$\mathbb{E}_Q[h(U,W)] = (m,T)_E + C_{TL}C_{LL}^{-1}\big(\mu - (m,L)_E\big) \tag{38}$$

$$\mathrm{Cov}_Q[h(U,W)] = \big(C_Q T, T\big)_E = (CT,T)_E + C_{TL}C_{LL}^{-1}(\Sigma - C_{LL})C_{LL}^{-1}C_{LT}. \tag{39}$$

In other words Q^T is Gaussian for any $T \in E^*$ and we conclude that Q is a Gaussian measure.

To deduce $(C_Q T, T')_E$ for two arbitrary elements $T, T' \in E$ we reduce everything to the one-dimensional case. For $\alpha, \beta \in \mathbb{R}$ let $\varphi = \alpha T + \beta T'$ then since $\varphi \in E^*$ and F is Gaussian

$$(F,\varphi)_E = \alpha(F,T)_E + \beta(F,T')_E,$$

is Gaussian. This proves, by definition, that $(F,T)_E$ and $(F,T')_E$ are jointly Gaussian under Q. The mean and variance of $(F,\varphi)_E$, can be calculated from (38) and (39). Using standard linear algebra

$$\mathrm{Cov}_Q[(F,T)_E, (F,T')_E] = (CT,T')_E + C_{TL}C_{LL}^{-1}(\Sigma - C_{LL})C_{LL}^{-1}C_{LT'}$$

which shows C_Q is as described in (35). □

The Kullback-Leibler divergence is tractable

In this section we show that the Kullback-Leibler divergence between the variational measure Q and $P^{F|Y=y}$ can be re-written in a convenient form. This is well-known for finite dimensional Gaussians and has been done for the process view in Matthews et al. (2016) but we have not seen such a derivation for Gaussian measures in Banach spaces so we include it here for completeness.

First, recall the chain rule for Radon-Nikodym derivatives (Halmos, 2013, Chapter 32). Let μ, ν and η be σ-finite measures on the same measure space. If $\mu \ll \nu$ and $\nu \ll \eta$, then $\mu \ll \eta$ with Radon-Nikodym derrivative given as

$$\frac{d\mu}{d\eta}(f) = \frac{d\mu}{d\nu}(f)\frac{d\nu}{d\eta}(f)$$

for η-almost every f.

Theorem 4. *1. The Kullback-Leibler divergence satisfies*

$$\begin{aligned} KL(Q, P^{F|Y=y}) &= KL(Q,P) - \mathbb{E}_Q[\log p(y|F)] + \log p(y) \\ &= -\mathcal{L} + \log p(y), \end{aligned}$$

for any $y \in \mathbb{R}^N$.

2. *The ELBO is tractable and given as*

$$\begin{aligned}\mathcal{L} = \sum_{n=1}^{N} \Bigg(& \log \mathcal{N}\big(y_n | (m, D_n)_E + C_{LL}^{-1}(\mu - (m, L)_E) C_{LD_n}, \sigma^2\big) \\ & - \frac{1}{2\sigma^2}\Big((CD_n, D_n)_E + C_{D_n L} C_{LL}^{-1}(\Sigma - C_{LL}) C_{LL}^{-1} C_{LD_n}\Big)\Bigg) \\ & - KL\big(\mathcal{N}(\mu, \Sigma), \mathcal{N}((F, m)_E, C_{LL})\big)\end{aligned}$$

3. *If the prior mean is zero, then the optimal values used in the variational family $\mathcal{Q}$ for μ and Σ are*

$$\begin{aligned}\mu^* &= C_{LL}\big(\sigma^2 C_{LL} + C_{LD} C_{DL}\big)^{-1} C_{LD} y \\ \Sigma^* &= C_{LL}\big(C_{LL} + \frac{1}{\sigma^2} C_{LD} C_{DL}\big)^{-1} C_{LL}\end{aligned}$$

which then leads to the optimal mean and covariance satisfying

$$(m_{Q^*}, T)_E = C_{TL}\big(\sigma^2 C_{LL} + C_{LD} C_{DL}\big)^{-1} C_{LD} y \tag{40}$$

$$(C_{Q^*} T, T')_E = (CT, T')_E - C_{TL} C_{LL}^{-1} C_{LT'} + C_{TL}\big(C_{LL} + \frac{1}{\sigma^2} C_{LD} C_{DL}\big)^{-1}) C_{LT'}, \tag{41}$$

for all $T, T' \in E^*$.

Proof. For the proof of the first statement, by Bayes theorem we know that $P^{F|Y=y}$ is dominated by the prior P with $\frac{dP^{F|Y=y}}{dP}(f) = \frac{p(y|F=f)}{p(y)}$ for any $y \in \mathbb{R}^N$. The reverse statement is also true, that P is dominated by $P^{F|=y}$ for fixed $y \in \mathbb{R}^N$. This is a consequence of $f \mapsto \frac{p(y|F=f)}{p(y)} > 0$ since then for any $A \in \mathcal{B}(E)$ with $P(A) > 0$, $P^{F|Y=y}(A) = \int_A \frac{p(y|F=f)}{p(y)} dP(f) > 0$, which the contrapositive of P being dominated by $P^{F|Y=y}$. The Radon-Nikodym in this situation is given as $\frac{p(y)}{p(y|F=f)}$. Finally, by definition of Q it is dominated by P and $\frac{dQ}{dP}(f) = (\frac{dQ^L}{dP^L} \circ L)(f)$ for any $f \in E$.

The chain-rule for Radon-Nikodym derivatives therefore tells us that Q is dominated by $P^{F|Y=y}$ with

$$\frac{dQ}{dP^{F|Y=y}}(f) = \frac{dQ}{dP}(f) \frac{dP}{dP^{F|Y=y}}(f) = \big(\frac{dQ^L}{dP^L} \circ L\big)(f) \frac{p(y)}{p(y|F=f)}.$$

This lets us rewrite the KL divergence as

$$\begin{aligned}KL(Q, P^{F|Y=y}) &= \int_E \log\left(\frac{dQ}{dP^{F|Y=y}}\right)(f) dQ(f) \\ &= \int_E \log\left(\big(\frac{dQ^L}{dP^L} \circ L\big)(f) \frac{p(y)}{p(y|F=f)}\right) dQ(f) \\ &= \int_E \log\left(\big(\frac{dQ^L}{dP^L} \circ L\big)(f)\right) dQ(f) + \int_E \log\left(\frac{p(y)}{p(y|F=f)}\right) dQ(f) \\ &= \int_E \log\left(\frac{dQ^L}{dP^L}(u)\right) dQ^L(u) - \int_E \log p(y|F=f) Q(f) + \log p(y) \\ &= KL(Q^L, P^L) - \mathbb{E}_Q[\log p(y|F)] + \log p(y),\end{aligned}$$

which proves the first statement.

For the second statement, recall the ELBO is $\mathcal{L} = -KL(Q^L, P^L) + \mathbb{E}_Q[\log(y|F)]$. The KL term is clear, since Q^L and P^L are both Gaussian measures with the required mean and covariance. We therefore investigate the log-likelihood term now. Note that $Y_1, \ldots, Y_N$ are conditionally independent given $F = f$. The expected log-likelihood term therefore factorises as

$$\mathbb{E}_Q[\log p(y|F=f)] = \sum_{n=1}^{N} \mathbb{E}_Q[\log p(y_n|F)].$$

Setting $V = D_n(F) = (F, D_n)_E$ and noting that $p(y_n|F = f)$ depends on F only through D_n, meaning $p(y_n|F = f) = p(y_n|V = D_n(f))$, we see

$$\mathbb{E}_Q[\log p(y_n|F)] = \int_E \log p\big(y_n|V = D_n(f)\big)dQ(f) = \int_E \log p\big(y_n|V = v\big)dQ^{D_n}(v).$$

Note that Q^{D_n} is Gaussian with mean $\mu_V := (m, D_n)_E + C_{LL}^{-1}\big(\mu - (m, L)_E\big)C_{LD_n}$ and variance $\sigma_V^2 := (CD_n, D_n)_E + C_{D_nL}C_{LL}^{-1}(\Sigma - C_{LL})C_{LL}^{-1}C_{LD_n}$ as Q is a Gaussian measure. So using the parametric form of the pdf of a Gaussian

$$\begin{aligned}
\mathbb{E}_Q[\log p(y_n|F)] &= \mathbb{E}_Q\big[-\frac{1}{2\sigma^2}(y_n - V)^2 - \log(\sigma) - \frac{1}{2}\log(2\pi)\big] \\
&= \mathbb{E}_Q\big[-\frac{1}{2\sigma^2}(y_n - \mu_V)^2 - \log(\sigma) - \frac{1}{2}\log(2\pi) - \frac{1}{2\sigma^2}(\mu_V - V_n)^2\big] \\
&= \log \mathcal{N}(y_n|\mu_V, \sigma^2) - \frac{1}{2\sigma^2}\mathbb{E}_Q\big[(\mu_v - V_n)^2\big] \\
&= \log \mathcal{N}(y_n|\mu_V, \sigma^2) - \frac{1}{2\sigma^2}\,\sigma_V^2,
\end{aligned}$$

which proves the claim.

Finally, for the third statement, in Appendix A of Titsias (2009b) the optimal form of μ and Σ are given. Note that we have the same objective function as Titsias (2009b) with the only difference that the kernel matrices k_{nm} and k_{mm} need to be replaced with the covariance matrices C_{LD} and C_{DD}. Plugging in the optimal form for μ^* and Σ^* into (34) and (35) gives rise to m_{Q^*} and C_{Q^*}. □

C Proof of Section 6: Connections between GRE Regression and KRR Nyström

Theorem 5. *Let $F \sim \mathcal{N}(0, C)$ be a GRE in $E = C(\mathcal{X}, \mathbb{R})$ with covariance operator C as defined in (6) and assumed pointwise noisy data is observed as described in Section 3.1. Let $L_m = \mu_m$, where $\{\mu_m\}_{m=1}^M \subset R(\mathcal{X})$ be the features used in the variational approximation. Set $\mathcal{M} = \{C\mu_m\}_{m=1}^M$ where $C\mu_m = \int k(\cdot, x')d\mu_m(x')$ as the approximating family in the Nyström approximation. Then for $\sigma^2 = N\lambda$ the Nyström KRR estimator $\widehat{f}$ in Section 6 is equal to the mean m_{Q^*}, given by (40), of the optimal Q^* from the variational family $\mathcal{Q}$.*

Proof. First of all note that $\mathcal{M} \subset H_k$, the RKHS of k. For a proof see (Ghosal and van der Vaart, 2017, Lemma 11.4). This means that the structure of H_k can be leveraged to deduce $\widehat{f}$. Specifically, as every $f \in \mathcal{M}$ can be expressed as $f = \sum_{m=1}^M \alpha_m C\mu_m$ for some $\alpha \in \mathbb{R}^M$ we can solve the finite dimensional optimisation problem

$$J(\alpha) := \frac{1}{N}\sum_{n=1}^N \big(y_n - \sum_{m=1}^M \alpha_m C\mu_m(x_n)\big)^2 + \lambda \| \sum_{m=1}^M \alpha_m C\mu_m\|_k^2,$$

in $\alpha \in \mathbb{R}^M$ to find the KRR Nyström estimator. Expanding $J(\alpha)$

$$\begin{aligned}
J(\alpha) &= \frac{1}{N}\sum_{n=1}^N y_n^2 - 2\frac{1}{N}\sum_{n=1}^N\sum_{m=1}^M y_n\alpha_m g_m(x_n) + \frac{1}{N}\sum_{n=1}^N\sum_{m,m'=1}^M \alpha_m\alpha_{m'}C\mu_m(x_n)C\mu_{m'}(x_n) \\
&\quad + \lambda\sum_{m,m'=1}^M \alpha_m\alpha_{m'}\langle C\mu_m, C\mu_{m'}\rangle_k \\
&= \frac{1}{N}y^\top y - 2\frac{1}{N}y^\top K_{X\mathcal{M}}\alpha + \frac{1}{N}\alpha^\top K_{\mathcal{M}X}K_{X\mathcal{M}}\alpha + \lambda\alpha^\top K_{\mathcal{M}\mathcal{M}}\alpha,
\end{aligned}$$

where $(K_{\mathcal{M}X})_{mn} = C\mu_m(x)$ and $(K_{\mathcal{M}\mathcal{M}})_{mm'} = \langle C\mu_m, C\mu_{m'}\rangle_{H_k}$ for $n = 1, \dots, N$ and $m, m' = 1, \dots, M$. Standard rules for differentiation give

$$\begin{aligned}
J'(\alpha) &= -\frac{2}{N}K_{X\mathcal{M}}y + \frac{2}{N}K_{X\mathcal{M}}K_{\mathcal{M}X}\alpha + 2\lambda K_{\mathcal{M}\mathcal{M}}\alpha \\
J''(\alpha) &= \frac{2}{N}K_{X\mathcal{M}}K_{\mathcal{M}X} + 2\lambda K_{\mathcal{M}\mathcal{M}}.
\end{aligned}$$

It is easy to see that $J'(\alpha) = 0$ for $\alpha = (K_{\mathcal{M}X}K_{X\mathcal{M}} + N\lambda K_{\mathcal{M}\mathcal{M}})^{-1}K_{\mathcal{M}X}y$ and that $J''(\alpha)$ is positive definite. Hence the KRR estimator is given as

$$\widehat{f}(x) = \sum_{m=1}^{M} \alpha_m C\mu_m(x),$$

with $\alpha = (\alpha_1, \ldots, \alpha_M)$ given as $\alpha = (K_{\mathcal{M}X}K_{X\mathcal{M}} + N\lambda K_{\mathcal{M}\mathcal{M}})^{-1}K_{\mathcal{M}X}y$.

On the other hand, from (40) with $T = \delta_x$ and $D_n = \delta_{x_n}$

$$m_{Q^*}(x) = \sum_{m=1}^{M} \beta_m C\mu_m(x),$$

with $\beta = (\sigma^2 C_{LL} + C_{LD}C_{DL})^{-1}C_{LD}y$.

The only thing left to show is $K_{\mathcal{M}X} = C_{LD}$ and $K_{\mathcal{M}\mathcal{M}} = C_{LL}$. This is a consequence of the relationship between the so-called Cameron-Martin space of a GRE and the RKHS. In particular this exact equivalence is outlined by Ghosal and van der Vaart (2017, Page 316) which completes the proof. □