Hilbert Space Methods for Reduced-Rank Gaussian Process Regression

Arno Solin ARNO.SOLIN@AALTO.FI
Simo Särkkä SIMO.SARKKA@AALTO.FI
Department of Biomedical Engineering and Computational Science (BECS)
Aalto University, School of Science
P.O. Box 12200, FI-00076 Aalto, Finland

Abstract

This paper proposes a novel scheme for reduced-rank Gaussian process regression. The method is based on an approximate series expansion of the covariance function in terms of an eigenfunction expansion of the Laplace operator in a compact subset of $\mathbb{R}^d$. On this approximate eigenbasis the eigenvalues of the covariance function can be expressed as simple functions of the spectral density of the Gaussian process, which allows the GP inference to be solved under a computational cost scaling as $\mathcal{O}(nm^2)$ (initial) and $\mathcal{O}(m^3)$ (hyperparameter learning) with m basis functions and n data points. The approach also allows for rigorous error analysis with Hilbert space theory, and we show that the approximation becomes exact when the size of the compact subset and the number of eigenfunctions go to infinity. The expansion generalizes to Hilbert spaces with an inner product which is defined as an integral over a specified input density. The method is compared to previously proposed methods theoretically and through empirical tests with simulated and real data.

Keywords: Gaussian process regression, Laplace operator, eigenfunction expansion, pseudo-differential operator, reduced-rank approximation

1. Introduction

Gaussian processes (GPs, Rasmussen and Williams, 2006) are powerful tools for non-parametric Bayesian inference and learning. In GP regression the model functions $f(\mathbf{x})$ are assumed to be realizations from a Gaussian random process prior with a given covariance function $k(\mathbf{x}, \mathbf{x}')$, and learning amounts to solving the posterior process given a set of noisy measurements $y_1, y_2, \ldots, y_n$ at some given test inputs. This model is often written in the form

$$\begin{aligned} f &\sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')), \\ y_i &= f(\mathbf{x}_i) + \varepsilon_i, \end{aligned} \tag{1}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_\text{n}^2)$, for $i = 1, 2, \ldots, n$. One of the main limitations of GPs in machine learning is the computational and memory requirements that scale as $\mathcal{O}(n^3)$ and $\mathcal{O}(n^2)$ in a direct implementation. This limits the applicability of GPs when the number of training samples n grows large. The computational requirements arise because in solving the GP regression problem we need to invert the $n \times n$ Gram matrix $\mathbf{K} + \sigma_\text{n}^2\mathbf{I}$, where $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, which is an $\mathcal{O}(n^3)$ operation in general.

To overcome this problem, over the years, several schemes have been proposed. They typically reduce the storage requirements to $\mathcal{O}(nm)$ and complexity to $\mathcal{O}(nm^2)$, where $m < n$. Some early methods have been reviewed in Rasmussen and Williams (2006), and Quiñonero-Candela and Rasmussen (2005a) provide a unifying view on several methods. From a spectral point of view, several of these methods (*e.g.*, SOR, DTC, VAR, FIC) can be interpreted as modifications to the so-called *Nyström method* (see Baker, 1977; Williams and Seeger, 2001), a scheme for approximating the eigenspectrum.

For stationary covariance functions the spectral density of the covariance function can be employed: in this context the spectral approach has mainly been considered in regular grids, as this allows for the use of FFT-based methods for fast solutions (see Paciorek, 2007; Fritz et al., 2009), and more recently in terms of converting GPs to state space models (Särkkä and Hartikainen, 2012; Särkkä et al., 2013). Recently, Lázaro-Gredilla et al. (2010) proposed a sparse spectrum method where a randomly chosen set of spectral points span a trigonometric basis for the problem.

The methods proposed in this article fall into the class of methods called reduced-rank approximations (see, *e.g.*, Rasmussen and Williams, 2006) which are based on approximating the Gram matrix $\mathbf{K}$ with a matrix $\tilde{\mathbf{K}}$ with a smaller rank $m < n$. This allows for the use of matrix inversion lemma (Woodbury formula) to speed up the computations. It is well-known that the optimal reduced-rank approximation of the Gram (covariance) matrix $\mathbf{K}$ with respect to the Frobenius norm is $\tilde{\mathbf{K}} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^\mathsf{T}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix of the leading m eigenvalues of $\mathbf{K}$ and $\boldsymbol{\Phi}$ is the matrix of the corresponding orthonormal eigenvectors (Golub and Van Loan, 1996; Rasmussen and Williams, 2006). Yet, as computing the eigendecomposition is an $\mathcal{O}(n^3)$ operation, this provides no remedy as such.

In this work we propose a novel method for obtaining approximate eigendecompositions of covariance functions in terms of an eigenfunction expansion of the Laplace operator in a compact subset of $\mathbb{R}^d$. The method is based on interpreting the covariance function as the kernel of a pseudo-differential operator (Shubin, 1987) and approximating it using Hilbert space methods (Courant and Hilbert, 2008; Showalter, 2010). This results in a reduced-rank approximation for the covariance function. This path has not been explored in GP regression context before, although the approach is closely related to the stochastic partial differential equation based methods recently introduced to spatial statistics and GP regression (Lindgren et al., 2011; Särkkä and Hartikainen, 2012; Särkkä et al., 2013). We also show how the solution formally converges to the exact solution in well-defined conditions, and provide theoretical and experimental comparisons to existing state-of-the-art methods.

This paper is structured as follows: In Section 2 we derive the approximative series expansion of the covariance functions. Section 3 is dedicated to applying the approximation scheme to GP regression and providing details of the computational benefits. We provide a detailed analysis of the convergence of the method in Section 4. Section 5 and 6 provide comparisons to existing methods, the former from a more theoretical point of view, whereas the latter contains examples and comparative evaluation on several datasets. Finally the properties of the method are summarized and discussed in Section 7.

2. Approximating the Covariance Function

In this section, we start by stating the assumptions and properties of the class of covariance functions that we are considering, and show how a homogenous covariance function can be considered as a pseudo-differential operator constructed as a series of Laplace operators. Then we show how the pseudo-differential operators can be approximated with Hilbert space methods on compact subsets of $\mathbb{R}^d$ or via inner products with integrable weight functions, and discuss connections to Sturm–Liouville theory.

2.1 Spectral Densities of Homogeneous and Isotropic Gaussian Processes

In this work it is assumed that the covariance function is homogeneous (stationary), which means that the covariance function $k(\mathbf{x}, \mathbf{x}')$ is actually a function of $\mathbf{r} = \mathbf{x} - \mathbf{x}'$ only. This means that the covariance structure of the model function $f(\mathbf{x})$ is the same regardless of the absolute position in the input space (*cf.* Rasmussen and Williams, 2006). In this case the covariance function can be equivalently represented in terms of the spectral density. This results from the *Bochner's theorem* (see, *e.g.*, Akhiezer and Glazman, 1993; Da Prato and Zabczyk, 1992) which states that an arbitrary positive definite function $k(\mathbf{r})$ can be represented as

$$k(\mathbf{r}) = \frac{1}{(2\pi)^d} \int \exp\left(\mathrm{i}\, \boldsymbol{\omega}^\mathsf{T} \mathbf{r}\right)\, \mu(\mathrm{d}\boldsymbol{\omega}), \tag{2}$$

where μ is a positive measure.

If the measure $\mu(\boldsymbol{\omega})$ has a density, it is called the *spectral density* $S(\boldsymbol{\omega})$ corresponding to the covariance function $k(\mathbf{r})$. This gives rise to the Fourier duality of covariance and spectral density, which is known as the *Wiener–Khintchin theorem* (Rasmussen and Williams, 2006), giving the identities

$$k(\mathbf{r}) = \frac{1}{(2\pi)^d} \int S(\boldsymbol{\omega})\, e^{\mathrm{i}\, \boldsymbol{\omega}^\mathsf{T} \mathbf{r}}\, \mathrm{d}\boldsymbol{\omega} \quad \text{and} \quad S(\boldsymbol{\omega}) = \int k(\mathbf{r})\, e^{-\mathrm{i}\, \boldsymbol{\omega}^\mathsf{T} \mathbf{r}}\, \mathrm{d}\mathbf{r}. \tag{3}$$

From these identities it is easy to see that if the covariance function is *isotropic*, that is, it only depends on the Euclidean norm $\|\mathbf{r}\|$ such that $k(\mathbf{r}) \triangleq k(\|\mathbf{r}\|)$, then the spectral density will also only depend on the norm of $\boldsymbol{\omega}$ such that we can write $S(\boldsymbol{\omega}) \triangleq S(\|\boldsymbol{\omega}\|)$. In the following we assume that the considered covariance functions are indeed isotropic, but the approach can be generalized to more general homogenous covariance functions.

2.2 The Covariance Operator As a Pseudo-Differential Operator

Associated to each covariance function $k(\mathbf{x}, \mathbf{x}')$ we can also define a covariance operator $\mathcal{K}$ as follows:

$$\mathcal{K}\, \phi = \int k(\cdot, \mathbf{x}')\, \phi(\mathbf{x}')\, \mathrm{d}\mathbf{x}'. \tag{4}$$

As we show in the next section, this interpretation allows us to approximate the covariance operator using Hilbert space methods which are typically used for approximating differential and pseudo-differential operators in the context of partial differential equations (Showalter, 2010). When the covariance function is homogenous, the corresponding operator will be translation invariant thus allowing for Fourier-representation as a transfer function. This transfer function is just the spectral density of the Gaussian process.

Consider an isotropic covariance function $k(\mathbf{x}, \mathbf{x}') \triangleq k(\|\mathbf{r}\|)$ (recall that $\|\cdot\|$ denotes the Euclidean norm). The spectral density of the Gaussian process and thus the transfer function corresponding to the covariance operator will now have the form $S(\|\boldsymbol{\omega}\|)$. We can formally write it as a function of $\|\boldsymbol{\omega}\|^2$ such that

$$S(\|\boldsymbol{\omega}\|) = \psi(\|\boldsymbol{\omega}\|^2). \tag{5}$$

Assume that the spectral density $S(\cdot)$ and hence $\psi(\cdot)$ are regular enough so that the spectral density has the following polynomial expansion:

$$\psi(\|\boldsymbol{\omega}\|^2) = a_0 + a_1\|\boldsymbol{\omega}\|^2 + a_2(\|\boldsymbol{\omega}\|^2)^2 + a_3(\|\boldsymbol{\omega}\|^2)^3 + \cdots. \tag{6}$$

Thus we also have

$$S(\|\boldsymbol{\omega}\|) = a_0 + a_1\|\boldsymbol{\omega}\|^2 + a_2(\|\boldsymbol{\omega}\|^2)^2 + a_3(\|\boldsymbol{\omega}\|^2)^3 + \cdots. \tag{7}$$

Recall that the transfer function corresponding to the Laplacian operator ∇^2 is $-\|\boldsymbol{\omega}\|^2$ in the sense that

$$\mathscr{F}[\nabla^2 f](\boldsymbol{\omega}) = -\|\boldsymbol{\omega}\|^2 \mathscr{F}[f](\boldsymbol{\omega}), \tag{8}$$

where $\mathscr{F}[\cdot]$ denotes the Fourier transform of its argument. If we take the Fourier transform of (7), we get the following representation for the covariance operator $\mathcal{K}$, which defines a pseudo-differential operator (Shubin, 1987) as a formal series of Laplacian operators:

$$\mathcal{K} = a_0 + a_1(-\nabla^2) + a_2(-\nabla^2)^2 + a_3(-\nabla^2)^3 + \cdots. \tag{9}$$

In the next section we will use this representation to form a series expansion approximation for the covariance function.

2.3 Hilbert-Space Approximation of the Covariance Operator

We will now form a Hilbert-space approximation for the pseudo-differential operator defined by (9). Let $\Omega \subset \mathbb{R}^d$ be a compact set, and consider the eigenvalue problem for the Laplacian operators with Dirichlet boundary conditions (we could use other boundary conditions as well):

$$\begin{cases} -\nabla^2 \phi_j(\mathbf{x}) = \lambda_j \, \phi_j(\mathbf{x}), & \mathbf{x} \in \Omega \\ \phi_j(\mathbf{x}) = 0, & \mathbf{x} \in \partial\Omega. \end{cases} \tag{10}$$

Because $-\nabla^2$ is a positive definite Hermitian operator, the set of eigenfunctions $\phi_j(\cdot)$ is orthonormal with respect to the inner product

$$\langle f, g \rangle = \int_\Omega f(\mathbf{x}) \, g(\mathbf{x}) \, \mathrm{d}\mathbf{x} \tag{11}$$

that is,

$$\int_\Omega \phi_i(\mathbf{x}) \, \phi_j(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \delta_{ij}, \tag{12}$$

and all the eigenvalues λ_j are real and positive. The negative Laplace operator can then be assigned the formal kernel

$$l(\mathbf{x}, \mathbf{x}') = \sum_j \lambda_j \, \phi_j(\mathbf{x}) \, \phi_j(\mathbf{x}') \tag{13}$$

in the sense that

$$-\nabla^2 f(\mathbf{x}) = \int l(\mathbf{x}, \mathbf{x}')\, f(\mathbf{x}')\, \mathrm{d}\mathbf{x}', \tag{14}$$

for sufficiently (weakly) differentiable functions f in the domain Ω assuming Dirichlet boundary conditions.

If we consider the formal powers of this representation, due to orthonormality of the basis, we can write the arbitrary operator power $s = 1, 2, \ldots$ of the kernel as

$$l^s(\mathbf{x}, \mathbf{x}') = \sum_j \lambda_j^s\, \phi_j(\mathbf{x})\, \phi_j(\mathbf{x}'). \tag{15}$$

This is again to be interpreted to mean that

$$-(\nabla^2)^s f(\mathbf{x}) = \int l^s(\mathbf{x}, \mathbf{x}')\, f(\mathbf{x}')\, \mathrm{d}\mathbf{x}', \tag{16}$$

for regular enough functions f and in the current domain with the assumed boundary conditions.

This implies that on the domain Ω, assuming the boundary conditions, we also have

$$\begin{aligned}\big[a_0 + a_1(-\nabla^2) + a_2(-\nabla^2)^2 + a_3(-\nabla^2)^3 + \cdots\big]\, f(\mathbf{x}) \\ = \int \big[a_0 + a_1\, l^1(\mathbf{x}, \mathbf{x}') + a_2\, l^2(\mathbf{x}, \mathbf{x}') + a_3\, l^3(\mathbf{x}, \mathbf{x}') + \cdots\big]\, f(\mathbf{x}')\, \mathrm{d}\mathbf{x}'.\end{aligned} \tag{17}$$

The left hand side is just $\mathcal{K} f$ via (9), on the domain with the boundary conditions, and thus by comparing to (4) and using (15) we can conclude that

$$\begin{aligned}k(\mathbf{x}, \mathbf{x}') &\approx a_0 + a_1\, l^1(\mathbf{x}, \mathbf{x}') + a_2\, l^2(\mathbf{x}, \mathbf{x}') + a_3\, l^3(\mathbf{x}, \mathbf{x}') + \cdots \\ &= \sum_j \big[a_0 + a_1\, \lambda_j^1 + a_2\, \lambda_j^2 + a_3\, \lambda_j^3 + \cdots\big]\, \phi_j(\mathbf{x})\, \phi_j(\mathbf{x}'),\end{aligned} \tag{18}$$

which is only an approximation to the covariance function due to restriction of the domain to Ω and the boundary conditions. By letting $\|\boldsymbol{\omega}\|^2 = \lambda_j$ in (7) we now obtain

$$S(\sqrt{\lambda_j}) = a_0 + a_1\lambda_j^1 + a_2\lambda_j^2 + a_3\lambda_j^3 + \cdots\,. \tag{19}$$

and substituting this into (18) then leads to the approximation

$$\boxed{k(\mathbf{x}, \mathbf{x}') \approx \sum_j S(\sqrt{\lambda_j})\, \phi_j(\mathbf{x})\, \phi_j(\mathbf{x}'),} \tag{20}$$

where $S(\cdot)$ is the spectral density of the covariance function, λ_j is the jth eigenvalue and $\phi_j(\cdot)$ the eigenfunction of the Laplace operator in a given domain. These expressions tend to be simple closed-form expressions.

The right hand side of (20) is very easy to evaluate, because it corresponds to evaluating the spectral density at the square roots of the eigenvalues and multiplying them with the eigenfunctions of the Laplace operator. Because the eigenvalues of the Laplacian operator

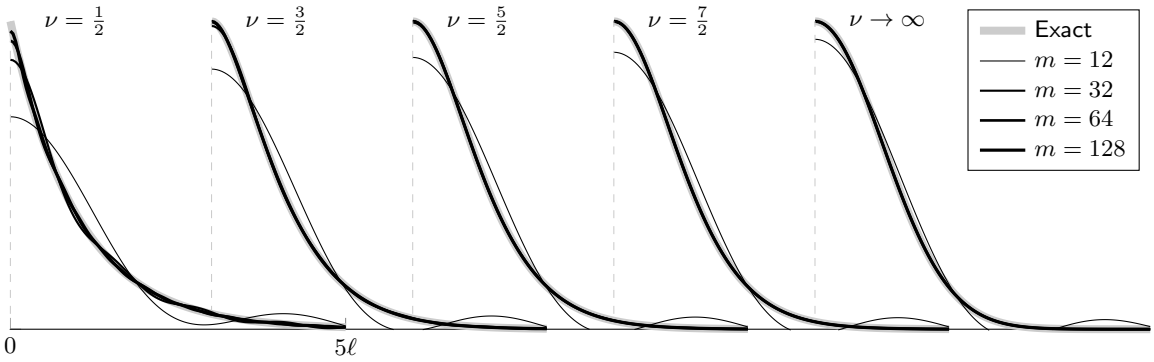


Figure 1: Approximations to covariance functions of the Matérn class of various degrees of smoothness; $\nu = 1/2$ corresponds to the exponential Ornstein–Uhlenbeck covariance function, and $\nu \to \infty$ to the squared exponential (exponentiated quadratic) covariance function. Approximations are shown for 12, 32, 64, and 128 eigenfunctions.

are monotonically increasing with j and for bounded covariance functions the spectral density goes to zero fast with higher frequencies, we can expect to obtain a good approximation of the right hand side by retaining only a finite number of terms in the series. However, even with an infinite number of terms this is only an approximation, because we assumed a compact domain with boundary conditions. The approximation can be, though, expected to be good at the input values which are not near the boundary of Ω, where the Laplacian was taken to be zero.

As an example, Figure 1 shows Matérn covariance functions of various degrees of smoothness ν (see, *e.g.*, Rasmussen and Williams, 2006) and approximations for different numbers of basis functions in the approximation. The basis consists of the functions $\phi_j(x) = L^{-1/2} \sin(\pi j(x + L)/(2L))$ and the eigenvalues were $\lambda_j = (\pi\, j/(2L))^2$ with $L = 1$ and $\ell = 0.1$. For the squared exponential the approximation is indistinguishable from the exact curve already at $m = 12$, whereas the less smooth functions require more terms.

2.4 Inner Product Point of View

Instead of considering a compact bounded set Ω, we can consider the same approximation in terms of an inner product defined by an input density (Williams and Seeger, 2000). Let the inner product be defined as

$$\langle f, g \rangle = \int f(\mathbf{x})\, g(\mathbf{x})\, w(\mathbf{x})\, \mathrm{d}\mathbf{x}, \tag{21}$$

where $w(\mathbf{x})$ is some positive weight function such that $\int w(\mathbf{x})\, \mathrm{d}\mathbf{x} < \infty$. In terms of this inner product, we define the operator

$$\mathcal{K} f = \int k(\cdot, \mathbf{x})\, f(\mathbf{x})\, w(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{22}$$

This operator is self-adjoint with respect to the inner product, $\langle \mathcal{K} f, g \rangle = \langle f, \mathcal{K} g \rangle$, and according to the spectral theorem there exists an orthonormal set of basis functions and

positive constants, $\{\varphi_j(\mathbf{x}), \gamma_j \mid j = 1, 2, \ldots\}$, that satisfies the eigenvalue equation

$$(\mathcal{K}\varphi_j)(\mathbf{x}) = \gamma_j \, \varphi_j(\mathbf{x}). \tag{23}$$

Thus $k(\mathbf{x}, \mathbf{x}')$ has the series expansion

$$k(\mathbf{x}, \mathbf{x}') = \sum_j \gamma_j \, \varphi_j(\mathbf{x}) \, \varphi_j(\mathbf{x}'). \tag{24}$$

Similarly, we also have the *Karhunen–Loeve expansion* for a sample function $f(\mathbf{x})$ with zero mean and the above covariance function:

$$f(\mathbf{x}) = \sum_j f_j \, \varphi_j(\mathbf{x}), \tag{25}$$

where f_js are independent zero mean Gaussian random variables with variances γ_j (see, *e.g.*, Lenk, 1991).

For the negative Laplacian the corresponding definition is

$$\mathcal{D}f = -\nabla^2[f \, w], \tag{26}$$

which implies

$$\langle \mathcal{D}f, g \rangle = -\int f(\mathbf{x}) \, w(\mathbf{x}) \nabla^2[g(\mathbf{x}) \, w(\mathbf{x})] \, \mathrm{d}\mathbf{x}, \tag{27}$$

and the operator defined by (26) can be seen to be self-adjoint. Again, there exists an orthonormal basis $\{\phi_j(\mathbf{x}) \mid j = 1, 2, \ldots\}$ and positive eigenvalues λ_j which satisfy the eigenvalue equation

$$(\mathcal{D} \, \phi_j)(\mathbf{x}) = \lambda_j \, \phi_j(\mathbf{x}). \tag{28}$$

Thus the kernel of $\mathcal{D}$ has a series expansion similar to Equation (13) and thus an approximation can be given in the same form as in Equation (20). In this case the approximation error comes from approximating the Laplacian operator with the more smooth operator,

$$\nabla^2 f \approx \nabla^2[f \, w], \tag{29}$$

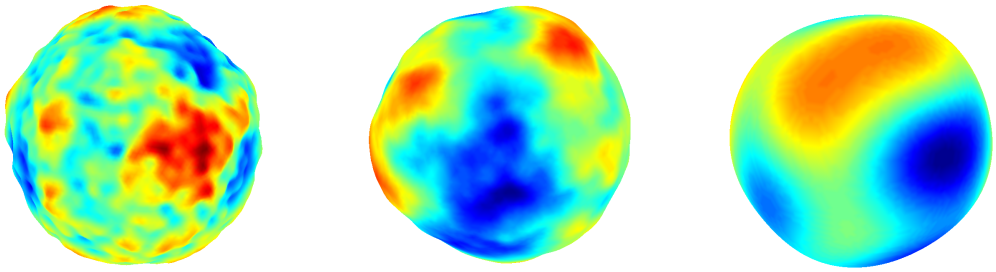
which is closely related to assumption of an input density $w(\mathbf{x})$ for the Gaussian process. However, when the weight function $w(\cdot)$ is close to constant in the area where the inputs points are located, the approximation is accurate.

2.5 Connection to Sturm–Liouville Theory

The presented methodology is also related to the Sturm–Liouville theory arising in the theory of partial differential equations (Courant and Hilbert, 2008). When the input x is scalar, the eigenvalue problem in Equation (23) can be written in Sturm–Liouville form as follows:

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left[w^2(x) \, \frac{\mathrm{d}\phi_j(x)}{\mathrm{d}x}\right] - w(x) \, \frac{\mathrm{d}^2 w(x)}{\mathrm{d}x^2} \, \phi_j(x) = \lambda_j \, w(x) \, \phi_j(x). \tag{30}$$

The above equation can be solved for $\phi_j(x)$ and λ_j using numerical methods for Sturm–Liouville equations. Also note that if we select $w(x) = 1$ in a finite set, we obtain the



(a) $\nu = \frac{1}{2}$ and $\ell = 0.5$ (b) $\nu = \frac{3}{2}$ and $\ell = 0.5$ (c) $\nu \to \infty$ and $\ell = 0.5$

Figure 2: Approximate random draws of Gaussian processes with the Matérn covariance function on the hull of a unit sphere. The color scale and radius follow the process.

equation $-\mathrm{d}^2/\,\mathrm{d}x^2\,\phi_j(x) = \lambda_j\,\phi_j(x)$ which is compatible with the results in the previous section.

We consider the case where $\mathbf{x} \in \mathbb{R}^d$ and $w(\mathbf{x})$ is symmetric around the origin and thus is only a function of the norm $r = \|\mathbf{x}\|$ (*i.e.* has the form $w(r)$). The Laplacian in spherical coordinates is

$$\nabla^2 f = \frac{1}{r^{d-1}} \frac{\partial}{\partial r} \left(r^{d-1} \frac{\partial f}{\partial r} \right) + \frac{1}{r^2} \Delta_{S^{d-1}} f, \tag{31}$$

where $\Delta_{S^{d-1}}$ is the Laplace–Beltrami operator on S^{d-1}. Let us assume that $\phi_j(r, \boldsymbol{\xi}) = h_j(r)\, g(\boldsymbol{\xi})$, where $\boldsymbol{\xi}$ denotes the angular variables. After some algebra, writing the equations into Sturm–Liouville form yields for the radial part

$$-\frac{\mathrm{d}}{\mathrm{d}r} \left(w^2(r)\, r\, \frac{\mathrm{d}h_j(r)}{\mathrm{d}r} \right) - \left(\frac{\mathrm{d}w(r)}{\mathrm{d}r}\, w(r) + \frac{\mathrm{d}^2 w(r)}{\mathrm{d}r^2}\, w(r)\, r \right) h_j(r) = \lambda_j\, w(r)\, r\, h_j(r), \tag{32}$$

and $\Delta_{S^{d-1}} g(\boldsymbol{\xi}) = 0$ for the angular part. The solutions to the angular part are the Laplace's spherical harmonics. Note that if we assume that we have $w(r) = 1$ on some area of finite radius, the first equation becomes (when $d > 1$):

$$r^2\, \frac{\mathrm{d}^2 h_j(r)}{\mathrm{d}r^2} + r\, \frac{\mathrm{d}h_j(r)}{\mathrm{d}r} + r^2\, \lambda_j\, h_j(r) = 0. \tag{33}$$

Figure 2 shows example Gaussian random field draws on a unit sphere, where the basis functions are the Laplace spherical harmonics and the covariance functions of the Matérn class with different degrees of smoothness ν. Our approximation is straight-forward to apply in any domain, where the eigendecomposition of the Laplacian can be formed.

3. Application of the Method to GP Regression

In this section we show how the approximation (20) can be used in Gaussian process regression. We also write down the expressions needed for hyperparameter learning and discuss the computational requirements of the methods.

3.1 Gaussian Process Regression

GP regression is usually formulated as predicting an unknown scalar output $f(\mathbf{x}_*)$ associated with a known input $\mathbf{x}_* \in \mathbb{R}^d$, given a training data set $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid i = 1, 2, \ldots, n\}$. The model functions f are assumed to be realizations of a Gaussian random process prior and the observations corrupted by Gaussian noise:

$$\begin{aligned} f &\sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')) \\ y_i &= f(\mathbf{x}_i) + \varepsilon_i, \end{aligned} \tag{34}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_\mathrm{n}^2)$. For notational simplicity the functions in the above model are *a priori* zero mean and the measurement errors are independent Gaussian, but the results of this paper can be easily generalized to arbitrary mean functions and dependent Gaussian errors. The direct solution to the GP regression problem (34) gives the predictions $p(f(\mathbf{x}_*) \mid \mathcal{D}) = \mathcal{N}(f(\mathbf{x}_*) \mid \mathbb{E}[f(\mathbf{x}_*)], \mathbb{V}[f(\mathbf{x}_*)])$. The conditional mean and variance can be computed in closed-form as (Rasmussen and Williams, 2006)

$$\begin{aligned} \mathbb{E}[f(\mathbf{x}_*)] &= \mathbf{k}_*^\mathsf{T} (\mathbf{K} + \sigma_\mathrm{n}^2 \mathbf{I})^{-1} \mathbf{y}, \\ \mathbb{V}[f(\mathbf{x}_*)] &= k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\mathsf{T} (\mathbf{K} + \sigma_\mathrm{n}^2 \mathbf{I})^{-1} \mathbf{k}_*, \end{aligned} \tag{35}$$

where $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{k}_*$ is an n-dimensional vector with the ith entry being $k(\mathbf{x}_*, \mathbf{x}_i)$, and $\mathbf{y}$ is a vector of the n observations.

In order to avoid the $n \times n$ matrix inversion in (35), we use the approximation scheme presented in the previous section and project the process to a truncated set of m basis functions of the Laplacian as given in Equation (20) such that

$$f(\mathbf{x}) \approx \sum_{j=1}^{m} f_j \, \phi_j(\mathbf{x}), \tag{36}$$

where $f_j \sim \mathcal{N}(0, S(\sqrt{\lambda_j}))$. We can then form an approximate eigendecomposition of the matrix $\mathbf{K} \approx \mathbf{\Phi} \mathbf{\Lambda} \mathbf{\Phi}^\mathsf{T}$, where $\mathbf{\Lambda}$ is a diagonal matrix of the leading m approximate eigenvalues such that $\mathbf{\Lambda}_{jj} = S(\sqrt{\lambda_j}), j = 1, 2, \ldots, m$. Here $S(\cdot)$ is the spectral density of the Gaussian process and λ_j the jth eigenvalue of the Laplace operator. The corresponding eigenvectors in the decomposition are given by the eigenvectors $\phi_j(\mathbf{x})$ of the Laplacian such that $\mathbf{\Phi}_{ij} = \phi_j(\mathbf{x}_i)$.

Using the matrix inversion lemma we rewrite (35) as follows:

$$\begin{aligned} \mathbb{E}[f_*] &\approx \boldsymbol{\phi}_*^\mathsf{T} (\mathbf{\Phi}^\mathsf{T} \mathbf{\Phi} + \sigma_\mathrm{n}^2 \mathbf{\Lambda}^{-1})^{-1} \mathbf{\Phi}^\mathsf{T} \mathbf{y}, \\ \mathbb{V}[f_*] &\approx \sigma_\mathrm{n}^2 \boldsymbol{\phi}_*^\mathsf{T} (\mathbf{\Phi}^\mathsf{T} \mathbf{\Phi} + \sigma_\mathrm{n}^2 \mathbf{\Lambda}^{-1})^{-1} \boldsymbol{\phi}_*, \end{aligned} \tag{37}$$

where $\boldsymbol{\phi}_*$ is an m-dimensional vector with the jth entry being $\phi_j(\mathbf{x}_*)$. Thus, when the size of the training set is higher than the number of required basis functions $n > m$, the use of this approximation is advantageous.

3.2 Learning the Hyperparameters

A common way to learn the hyperparameters $\boldsymbol{\theta}$ of the covariance function (suppressed earlier in the notation for brevity) and the noise variance σ_n^2 is by maximizing the marginal

likelihood function (Rasmussen and Williams, 2006; Quiñonero-Candela and Rasmussen, 2005b). Let $\mathbf{Q} = \mathbf{K} + \sigma_{\mathrm{n}}^2\mathbf{I}$ for the full model, then the negative log marginal likelihood and its derivatives are

$$\mathcal{L} = \frac{1}{2}\log|\mathbf{Q}| + \frac{1}{2}\mathbf{y}^{\mathsf{T}}\mathbf{Q}^{-1}\mathbf{y} + \frac{n}{2}\log(2\pi), \tag{38}$$

$$\frac{\partial\mathcal{L}}{\partial\theta_k} = \frac{1}{2}\mathrm{Tr}\left(\mathbf{Q}^{-1}\frac{\partial\mathbf{Q}}{\partial\theta_k}\right) - \frac{1}{2}\mathbf{y}^{\mathsf{T}}\mathbf{Q}^{-1}\frac{\partial\mathbf{Q}}{\partial\theta_k}\mathbf{Q}^{-1}\mathbf{y}, \tag{39}$$

$$\frac{\partial\mathcal{L}}{\partial\sigma_{\mathrm{n}}^2} = \frac{1}{2}\mathrm{Tr}\big(\mathbf{Q}^{-1}\big) - \frac{1}{2}\mathbf{y}^{\mathsf{T}}\mathbf{Q}^{-1}\mathbf{Q}^{-1}\mathbf{y}, \tag{40}$$

and they can be combined with a conjugate gradient optimizer. The problem in this case is the inversion of $\mathbf{Q}$, which is an $n \times n$ matrix. And thus each step of running the optimizer is $\mathcal{O}(n^3)$. For our approximation scheme, let $\tilde{\mathbf{Q}} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^{\mathsf{T}} + \sigma_{\mathrm{n}}^2\mathbf{I}$. Now replacing $\mathbf{Q}$ with $\tilde{\mathbf{Q}}$ in the above expressions gives us the following:

$$\tilde{\mathcal{L}} = \frac{1}{2}\log|\tilde{\mathbf{Q}}| + \frac{1}{2}\mathbf{y}^{\mathsf{T}}\tilde{\mathbf{Q}}^{-1}\mathbf{y} + \frac{n}{2}\log(2\pi), \tag{41}$$

$$\frac{\partial\tilde{\mathcal{L}}}{\partial\theta_k} = \frac{1}{2}\frac{\partial\log|\tilde{\mathbf{Q}}|}{\partial\theta_k} + \frac{1}{2}\frac{\partial\mathbf{y}^{\mathsf{T}}\tilde{\mathbf{Q}}^{-1}\mathbf{y}}{\partial\theta_k}, \tag{42}$$

$$\frac{\partial\tilde{\mathcal{L}}}{\partial\sigma_{\mathrm{n}}^2} = \frac{1}{2}\frac{\partial\log|\tilde{\mathbf{Q}}|}{\partial\sigma_{\mathrm{n}}^2} + \frac{1}{2}\frac{\partial\mathbf{y}^{\mathsf{T}}\tilde{\mathbf{Q}}^{-1}\mathbf{y}}{\partial\sigma_{\mathrm{n}}^2}, \tag{43}$$

where for the terms involving $\log|\tilde{\mathbf{Q}}|$:

$$\log|\tilde{\mathbf{Q}}| = (n-m)\log\sigma_{\mathrm{n}}^2 + \log|\mathbf{Z}| + \sum_{j=1}^{m}\log S(\sqrt{\lambda_j}), \tag{44}$$

$$\frac{\partial\log|\tilde{\mathbf{Q}}|}{\partial\theta_k} = \sum_{j=1}^{m} S(\sqrt{\lambda_j})^{-1}\frac{\partial S(\sqrt{\lambda_j})}{\partial\theta_k} - \sigma_{\mathrm{n}}^2\mathrm{Tr}\left(\mathbf{Z}^{-1}\mathbf{\Lambda}^{-2}\frac{\partial\mathbf{\Lambda}}{\partial\theta_k}\right), \tag{45}$$

$$\frac{\partial\log|\tilde{\mathbf{Q}}|}{\partial\sigma_{\mathrm{n}}^2} = \frac{n-m}{\sigma_{\mathrm{n}}^2} + \mathrm{Tr}\big(\mathbf{Z}^{-1}\mathbf{\Lambda}^{-1}\big), \tag{46}$$

and for the terms involving $\tilde{\mathbf{Q}}^{-1}$:

$$\mathbf{y}^{\mathsf{T}}\tilde{\mathbf{Q}}^{-1}\mathbf{y} = \frac{1}{\sigma_{\mathrm{n}}^2}\left(\mathbf{y}^{\mathsf{T}}\mathbf{y} - \mathbf{y}\mathbf{\Phi}\mathbf{Z}^{-1}\mathbf{\Phi}^{\mathsf{T}}\mathbf{y}\right), \tag{47}$$

$$\frac{\partial\mathbf{y}^{\mathsf{T}}\tilde{\mathbf{Q}}^{-1}\mathbf{y}}{\partial\theta_k} = -\mathbf{y}^{\mathsf{T}}\mathbf{Z}^{-1}\left(\mathbf{\Lambda}^{-2}\frac{\partial\mathbf{\Lambda}}{\partial\theta_k}\right)\mathbf{Z}^{-1}\mathbf{y}, \tag{48}$$

$$\frac{\partial\mathbf{y}^{\mathsf{T}}\tilde{\mathbf{Q}}^{-1}\mathbf{y}}{\partial\sigma_{\mathrm{n}}^2} = \frac{1}{\sigma_{\mathrm{n}}^2}\mathbf{y}^{\mathsf{T}}\mathbf{\Phi}\mathbf{Z}^{-1}\mathbf{\Lambda}^{-1}\mathbf{Z}^{-1}\mathbf{\Phi}^{\mathsf{T}}\mathbf{y} - \frac{1}{\sigma_{\mathrm{n}}^2}\mathbf{y}^{\mathsf{T}}\tilde{\mathbf{Q}}\mathbf{y}, \tag{49}$$

where $\mathbf{Z} = \sigma_{\mathrm{n}}^2\mathbf{\Lambda}^{-1} + \mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi}$. For efficient implementation, matrix-to-matrix multiplications can be avoided in many cases, and the inversion of $\mathbf{Z}$ can be carried out through Cholesky factorization for numerical stability. This factorization ($\mathbf{L}\mathbf{L}^{\mathsf{T}} = \mathbf{Z}$) can also be used for

the term $\log|\mathbf{Z}| = 2\sum_j \log \mathbf{L}_{jj}$, and $\mathrm{Tr}\big(\mathbf{Z}^{-1}\boldsymbol{\Lambda}^{-1}\big) = \sum_j 1/(\mathbf{Z}_{jj}\boldsymbol{\Lambda}_{jj})$ can be evaluated by element-wise multiplication.

Once the marginal likelihood and its derivatives are available, it is also possible to use other methods for parameter inference such as Markov chain Monte Carlo methods (Liu, 2001; Brooks et al., 2011) including Hamiltonian Monte Carlo (HMC, Duane et al., 1987; Neal, 2011) as well as numerous others.

3.3 Discussion on the Computational Complexity

As can be noted from Equation (20), the basis functions in the reduced-rank approximation do not depend on the hyperparameters of the covariance function. Thus it is enough to calculate the product $\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}$ only once, which means that the method has a overall asymptotic computational complexity of $\mathcal{O}(nm^2)$. After this initial cost, evaluating the marginal likelihood and the marginal likelihood gradient is an $\mathcal{O}(m^3)$ operation—which in practice comes from the Cholesky factorization of $\mathbf{Z}$ on each step.

If the number of observations n is so large that storing the $n \times m$ matrix $\boldsymbol{\Phi}$ is not feasible, the computations of $\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}$ can be carried out in blocks. Storing the evaluated eigenfunctions in $\boldsymbol{\Phi}$ is not necessary, because the $\phi_j(\mathbf{x})$ are closed-form expressions that can be evaluated when necessary. In practice, it might be preferable to cache the result of $\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}$ (causing a memory requirement scaling as $\mathcal{O}(m^2)$), but this is not required.

The computational complexity of conventional sparse GP approximations typically scale as $\mathcal{O}(nm^2)$ in time for each step of evaluating the marginal likelihood. The scaling in demand of storage is $\mathcal{O}(nm)$. This comes from the inevitable cost of re-evaluating all results involving the basis functions on each step and storing the matrices required for doing this. This applies to all the methods that will be discussed in Section 5, with the exception of SSGP, where the storage demand can be relaxed by re-evaluating the basis functions on demand.

We can also consider the rather restricting, but in certain applications often encountered case, where the measurements are constrained to a regular grid. This causes the product of the orthonormal eigenfunction matrices $\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}$ to be diagonal, avoiding the calculation of the matrix inverse altogether. This relates to the FFT-based methods for GP regression (Paciorek, 2007; Fritz et al., 2009), and the projections to the basis functions can be evaluated by fast Fourier transform in $\mathcal{O}(n \log n)$ time complexity.

4. Convergence Analysis

In this section we analyze the convergence of the proposed approximation when the size of the domain Ω and the number of terms in the series grows to infinity. We start by analyzing a univariate problem in the domain $\Omega = [-L, L]$ and with Dirichlet boundary conditions and then generalize the result to d-dimensional cubes $\Omega = [-L_1, L_1] \times \cdots \times [-L_d, L_d]$. We also discuss how the analysis could be extended to other types of basis functions.

4.1 Univariate Dirichlet Case

In the univariate case, the m-term approximation has the form

$$\widetilde{k}_m(x, x') = \sum_{j=1}^{m} S(\sqrt{\lambda_j})\, \phi_j(x)\, \phi_j(x'), \tag{50}$$

where the eigenfunctions and eigenvalues are:

$$\phi_j(x) = \frac{1}{\sqrt{L}} \sin\left(\frac{\pi\, j\, (x + L)}{2L}\right) \quad \text{and} \quad \lambda_j = \left(\frac{\pi\, j}{2L}\right)^2, \quad \text{for } j = 1, 2, \ldots. \tag{51}$$

The true covariance function $k(x, x')$ is assumed to be stationary and have a spectral density which is uniformly bounded $S(\omega) < \infty$, has at least two bounded derivatives $|S'(\omega)| < \infty$, $|S''(\omega)| < \infty$, and has a bounded integral over the real axis $\int_{-\infty}^{\infty} S(\omega)\, \mathrm{d}\omega < \infty$. We also assume that our training and test sets are constrained in the area $[-\widetilde{L}, \widetilde{L}]$, where $\widetilde{L} < L$, and thus we are only interested in the case $x, x' \in [-\widetilde{L}, \widetilde{L}]$. For the purposes of analysis we also assume that L is bounded below by a constant.

The univariate convergence result can be summarized as the following theorem which is proved in Appendix A.1.

Theorem 4.1. *There exists a constant C such that*

$$\left| k(x, x') - \widetilde{k}_m(x, x') \right| \leq \frac{C}{L} + \frac{2}{\pi} \int_{\frac{\pi\, m}{2L}}^{\infty} S(\omega)\, \mathrm{d}\omega, \tag{52}$$

which in turn implies that uniformly

$$\lim_{L \to \infty} \left[\lim_{m \to \infty} \widetilde{k}_m(x, x') \right] = k(x, x'). \tag{53}$$

Remark 4.2. *Note that we cannot simply exchange the order of the limits in the above theorem. However, the theorem does ensure the convergence of the approximation in the joint limit $m, L \to \infty$ provided that we add terms to the series fast enough such that $m/L \to \infty$. That is, in this limit, the approximation $\widetilde{k}_m(x, x')$ converges uniformly to $k(x, x')$.*

As such, the results above only ensure the convergence of the prior covariance functions. However, it turns out that this also ensures the convergence of the posterior as is summarized in the following corollary.

Corollary 4.3. *Because the Gaussian process regression equations only involve pointwise evaluations of the kernels, it also follows that the posterior mean and covariance functions converge uniformly to the exact solutions in the limit $m, L \to \infty$.*

4.2 Multivariate Cartesian Dirichlet Case

In order to generalize the results from the previous section, we turn our attention to a d-dimensional inputs space with rectangular domain $\Omega = [-L_1, L_1] \times \cdots \times [-L_d, L_d]$ with

Dirichlet boundary conditions. In this case we consider a truncated $m = \hat{m}^d$ term approximation of the form

$$\widetilde{k}_m(\mathbf{x}, \mathbf{x}') = \sum_{j_1,\ldots,j_d=1}^{\hat{m}} S(\sqrt{\lambda_{j_1,\ldots,j_d}})\, \phi_{j_1,\ldots,j_d}(\mathbf{x})\, \phi_{j_1,\ldots,j_d}(\mathbf{x}') \tag{54}$$

with the eigenfunctions and eigenvalues

$$\phi_{j_1,\ldots,j_d}(x) = \prod_{k=1}^{d} \frac{1}{\sqrt{L_k}} \sin\left(\frac{\pi\, j_k\, (x_k + L_k)}{2L_k}\right) \quad \text{and} \quad \lambda_{j_1,\ldots,j_d} = \sum_{k=1}^{d} \left(\frac{\pi\, j_k}{2L_k}\right)^2. \tag{55}$$

The true covariance function $k(\mathbf{x}, \mathbf{x}')$ is assumed to be stationary and have a spectral density $S(\boldsymbol{\omega})$ which is two times differentiable and the derivatives are assumed to be bounded. We also assume that the single-variable integrals are finite $\int_{-\infty}^{\infty} S(\boldsymbol{\omega})\, \mathrm{d}\omega_k < \infty$, which in this case is equivalent to requiring that the integral over the whole space is finite $\int_{\mathbb{R}^d} S(\boldsymbol{\omega})\, \mathrm{d}\boldsymbol{\omega} < \infty$. Furthermore, we assume that the training and test sets are contained in the d-dimensional cube $[-\widetilde{L}, \widetilde{L}]^d$ and that L_ks are bounded from below.

The following result for this d-dimensional case is proved in Appendix A.2.

Theorem 4.4. *There exists a constant C_d such that*

$$\left| k(\mathbf{x}, \mathbf{x}') - \widetilde{k}_m(\mathbf{x}, \mathbf{x}') \right| \leq \frac{C_d}{L} + \frac{1}{\pi^d} \int_{\|\boldsymbol{\omega}\| \geq \frac{\pi\, \hat{m}}{2L}} S(\boldsymbol{\omega})\, \mathrm{d}\boldsymbol{\omega}, \tag{56}$$

where $L = \min_k L_k$, which in turn implies that uniformly

$$\lim_{L_1,\ldots,L_d \to \infty} \left[\lim_{m \to \infty} \widetilde{k}_m(\mathbf{x}, \mathbf{x}') \right] = k(\mathbf{x}, \mathbf{x}'). \tag{57}$$

Remark 4.5. *Analogously as in the one-dimensional case we cannot simply exchange the order of the limits above. Furthermore, we need to add terms fast enough so that $\hat{m}/L_k \to \infty$ when $m, L_1, \ldots, L_d \to \infty$.*

Corollary 4.6. *As in the one-dimensional case, the uniform convergence of the prior covariance function also implies uniform convergence of the posterior mean and covariance in the limit $m, L_1, \ldots, L_d \to \infty$.*

4.3 Other Domains

It would also be possible carry out similar convergence analysis, for example, in a spherical domain. In that case the technical details become slightly more complicated, because instead sinusoidals we will have Bessel functions and the eigenvalues no longer form a uniform grid. This means that instead of Riemann integrals we need to consider weighted integrals where the distribution of the zeros of Bessel functions is explicitly accounted for. It might also be possible to use some more general theoretical results from mathematical analysis to obtain the convergence results. However, due to these technical challenges more general convergence proof will be developed elsewhere.

There is also a similar technical challenge in the analysis when the basis functions are formed by assuming an input density (see Section 2.4) instead of a bounded domain. Because explicit expressions for eigenfunctions and eigenvalues cannot be obtained in general, the elementary proof methods which we used here cannot be applied. Therefore the convergence analysis of this case is also left as a topic for future research.

5. Relationship to Other Methods

In this section we compare our method to existing sparse GP methods from a theoretical point of view. We consider two different classes of approaches: a class of inducing input methods based on the Nyström approximation (following the interpretation of Quiñonero-Candela and Rasmussen, 2005a), and direct spectral approximations.

5.1 Methods from the Nyström Family

A crude but rather effective scheme for approximating the eigendecomposition of the Gram matrix is the Nyström method (see, *e.g.*, Baker, 1977, for the integral approximation scheme). This method is based on choosing a set of m inducing inputs $\mathbf{x}_\mathrm{u}$ and scaling the corresponding eigendecomposition of their corresponding covariance matrix $\mathbf{K}_\mathrm{u,u}$ to match that of the actual covariance. The Nyström approximations to the jth eigenvalue and eigenfunction are

$$\widetilde{\lambda}_j = \frac{1}{m}\,\lambda_{\mathrm{u},j}, \tag{58}$$

$$\widetilde{\phi}_j(\mathbf{x}) = \frac{\sqrt{m}}{\lambda_{\mathrm{u},j}}\,k(\mathbf{x},\mathbf{x}_\mathrm{u})\,\boldsymbol{\phi}_{\mathrm{u},j}, \tag{59}$$

where $\lambda_{\mathrm{u},j}$ and $\boldsymbol{\phi}_{\mathrm{u},j}$ correspond to the jth eigenvalue and eigenvector of $\mathbf{K}_\mathrm{u,u}$. This scheme was originally introduced to the GP context by Williams and Seeger (2001). They presented a sparse scheme, where the resulting approximate prior covariance over the latent variables is $\mathbf{K}_\mathrm{f,u}\mathbf{K}_\mathrm{u,u}^{-1}\mathbf{K}_\mathrm{u,f}$, which can be derived directly from Equations (58) and (59).

As discussed by Quiñonero-Candela and Rasmussen (2005a), the Nyström method by Williams and Seeger (2001) does not correspond to a well-formed probabilistic model. However, several methods modifying the inducing point approach are widely used. The *Subset of Regressors* (SOR, Smola and Bartlett, 2001) method uses the Nyström approximation scheme for approximating the whole covariance function,

$$k_\mathrm{SOR}(\mathbf{x},\mathbf{x}') = \sum_{j=1}^{m} \widetilde{\lambda}_j\,\widetilde{\phi}_j(\mathbf{x})\,\widetilde{\phi}_j(\mathbf{x}'), \tag{60}$$

whereas the sparse Nyström method (Williams and Seeger, 2001) only replaces the training data covariance matrix. The SOR method is in this sense a complete Nyström approximation to the full GP problem. A method in-between is the *Deterministic Training Conditional* (DTC, Csató and Opper, 2002; Seeger et al., 2003) method which retains the true covariance for the training data, but uses the approximate cross-covariances between training and test data. For DTC, tampering with the covariance matrix causes the result not to actually be a Gaussian process. The *Variational Approximation* (VAR, Titsias, 2009) method

modifies the DTC method by an additional trace term in the likelihood that comes from the variational bound.

The *Fully Independent (Training) Conditional* (FIC, Quiñonero-Candela and Rasmussen, 2005a) method (originally introduced as *Sparse Pseudo-Input GP* by Snelson and Ghahramani, 2006) is also based on the Nyström approximation but contains an additional diagonal term replacing the diagonal of the approximate covariance matrix with the values from the true covariance. The corresponding prior covariance function for FIC, is thus

$$k_{\mathrm{FIC}}(\mathbf{x}_i, \mathbf{x}_j) = k_{\mathrm{SOR}}(\mathbf{x}_i, \mathbf{x}_j) + \delta_{i,j}(k(\mathbf{x}_i, \mathbf{x}_j) - k_{\mathrm{SOR}}(\mathbf{x}_i, \mathbf{x}_j)), \tag{61}$$

where $\delta_{i,j}$ is the Kronecker delta.

Figure 3 illustrates the effect of the approximations compared to the exact correlation structure in the GP. The dashed contours show the exact correlation contours computed for three locations with the squared exponential covariance function. Figure 3a shows the results for the FIC approximation with 16 inducing points (locations shown in the figure). It is clear that the number of inducing points or their locations are not sufficient to capture the correlation structure. For similar figures and discussion on the effects of the inducing points, see Vanhatalo et al. (2010). This behavior is not unique to SOR or FIC, but applies to all the methods from the Nyström family.

5.2 Direct Spectral Methods

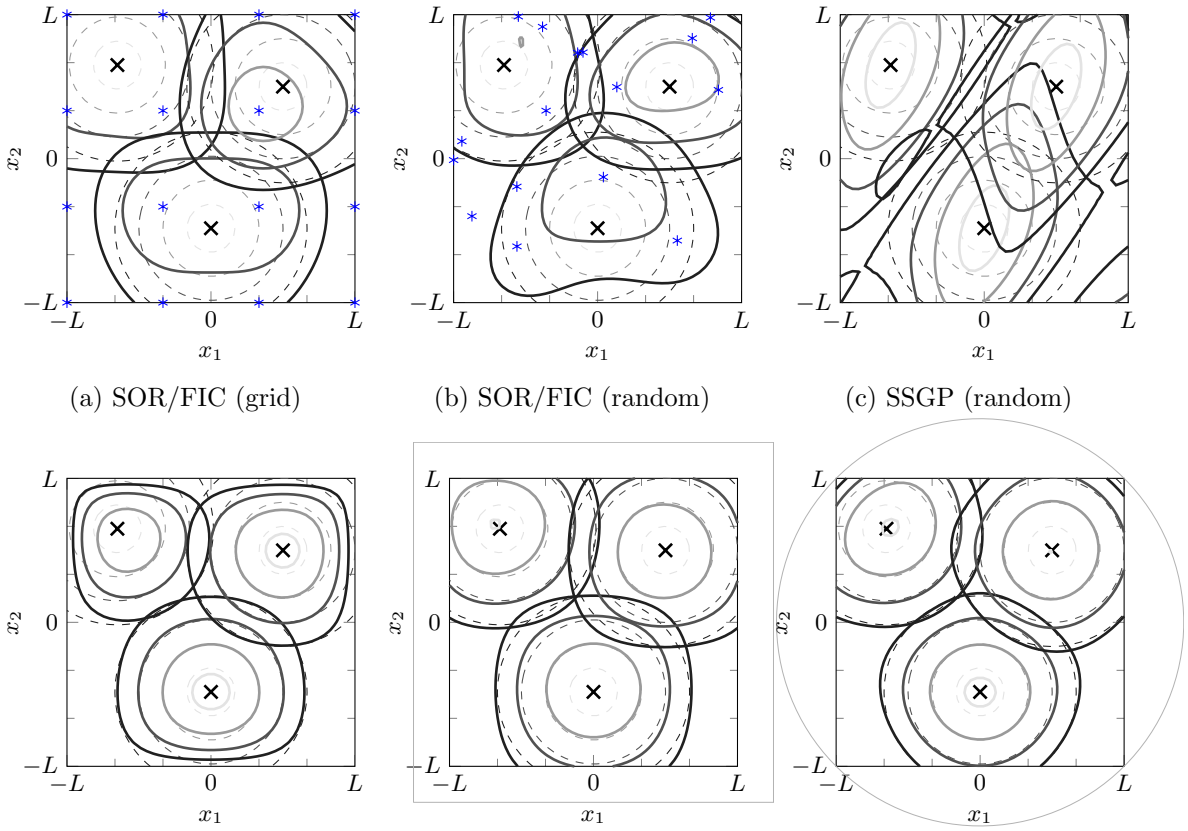
The sparse spectrum GP (SSGP) method introduced by Lázaro-Gredilla et al. (2010) uses the spectral representation of the covariance function for drawing random samples from the spectrum. These samples are used for representing the GP on a trigonometric basis

$$\boldsymbol{\phi}(\mathbf{x}) = \begin{pmatrix} \cos(2\pi\, \mathbf{s}_1^{\mathsf{T}}\mathbf{x}) & \sin(2\pi\, \mathbf{s}_1^{\mathsf{T}}\mathbf{x}) & \dots & \cos(2\pi\, \mathbf{s}_h^{\mathsf{T}}\mathbf{x}) & \sin(2\pi\, \mathbf{s}_h^{\mathsf{T}}\mathbf{x}) \end{pmatrix}, \tag{62}$$

where the spectral points $\mathbf{s}_r, r = 1, 2, \dots, h$ ($2h = m$), are sampled from the spectral density of the original stationary covariance function (following the normalization convention used in the original paper). The covariance function corresponding to the SSGP scheme is now of the form

$$k_{\mathrm{SSGP}}(\mathbf{x}, \mathbf{x}') = \frac{2\sigma^2}{m}\, \boldsymbol{\phi}(\mathbf{x})\, \boldsymbol{\phi}^{\mathsf{T}}(\mathbf{x}') = \frac{\sigma^2}{h} \sum_{r=1}^{h} \cos\left(2\pi\, \mathbf{s}_r^{\mathsf{T}}(\mathbf{x} - \mathbf{x}')\right), \tag{63}$$

where σ^2 is the magnitude scale hyperparameter. This representation of the sparse spectrum method converges to the full GP in the limit of the number of spectral points going to infinity, and it is the preferred formulation of the method in one or two dimensions (see Lázaro-Gredilla, 2010, for discussion). We can interpret the SSGP method in (63) as a Monte Carlo approximation of the Wiener–Khintchin integral. In order to have a representative sample of the spectrum, the method typically requires the number of spectral points to be large. For high-dimensional inputs the number of required spectral points becomes overwhelming, and optimizing the spectral locations along with the hyperparameters attractive. However, as argued by Lázaro-Gredilla et al. (2010), this option does not converge to the full GP and suffers from overfitting to the training data.



(a) SOR/FIC (grid)

(b) SOR/FIC (random)

(c) SSGP (random)

(d) The new method (Cartesian)

(e) The new method (extended Cartesian)

(f) The new method (polar)

Figure 3: Correlation contours computed for three locations (×) corresponding to the squared exponential covariance function (exact contours dashed). The rank of each approximation is $m = 16$, and the locations of the inducing inputs are marked with blue stars (∗). The hyperparameters are the same in each figure. The domain boundary is shown in thin grey (——) if extended outside the box.

Contours for the sparse spectrum SSGP method are visualized in Figure 3c. Here the spectral points were chosen at random following Lázaro-Gredilla (2010). Because the basis functions are spanned using both sines and cosines, the number of spectral points was $h = 8$ in order to match the rank $m = 16$. These results agree well with those presented in the Lázaro-Gredilla et al. (2010) for a one-dimensional example. For this particular set of spectral points some directions of the contours happen to match the true values very well, while other directions are completely off. Increasing the rank from 16 to 100 would give comparable results to the other methods.

While SSGP is based on a sparse spectrum, the reduced-rank method proposed in this paper aims to make the spectrum as 'full' as possible at a given rank. While SSGP can be interpreted as a Monte Carlo integral approximation, the corresponding interpretation to the proposed method would a numerical quadrature-based integral approximation (*cf.* the

convergence proof in Appendix A.1). Figure 3d shows the same contours obtained by the proposed reduced-rank method. Here the eigendecomposition of the Laplace operator has been obtained for the square $\Omega = [-L, L] \times [-L, L]$ with Dirichlet boundary conditions. The contours match well with the full solution towards the middle of the domain. The boundary effects drive the process to zero, which is seen as distortion near the edges.

Figure 3e shows how extending the boundaries just by 25% and keeping the number of basis functions fixed at 16, gives good results. The last Figure 3f corresponds to using a disk shaped domain instead of the rectangular. The eigendecomposition of the Laplace operator is done in polar coordinates, and the Dirichlet boundary is visualized by a circle in the figure.

6. Experiments

In this section we aim to provide examples of the practical use of the proposed method, and to compare it against other methods that are typically used in a similar setting. We start with a small simulated one-dimensional dataset, and then provide more extensive comparisons by using real-world data. We also consider an example of data, where the input domain is the surface of a sphere, and conclude our comparison by using a very large dataset to demonstrate what possibilities the computational benefits open.

6.1 Experimental Setup

For assessing the performance of different methods we use 10-fold cross-validation and evaluate the following measures based on the validation set: the *standardized mean squared error* (SMSE) and the *mean standardized log loss* (MSLL), respectively defined as:

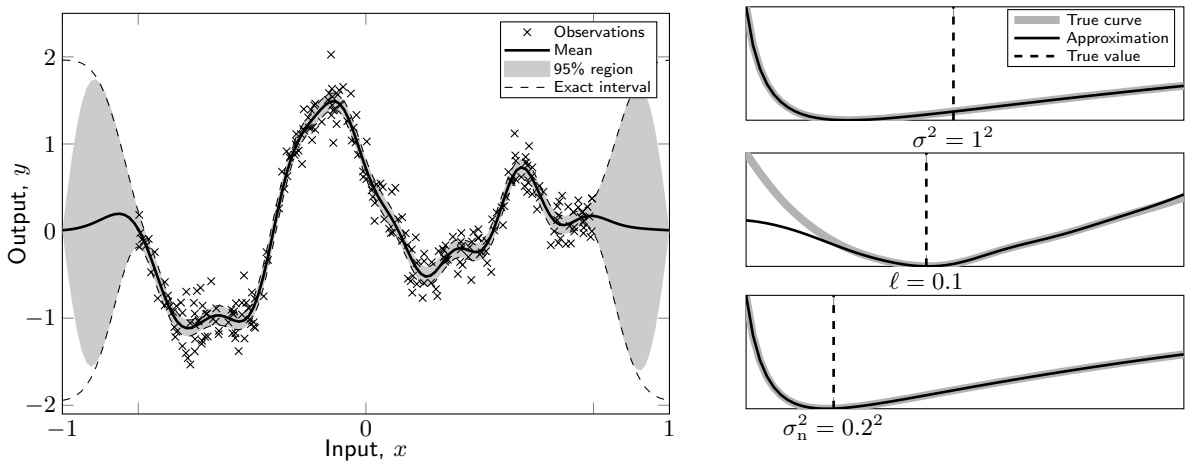
$$\text{SMSE} = \sum_{i=1}^{n_*} \frac{(y_{*i} - \mu_{*i})^2}{\text{Var}[y]}, \quad \text{and} \quad \text{MSLL} = \frac{1}{2n_*} \sum_{i=1}^{n_*} \left(\frac{(y_{*i} - \mu_{*i})^2}{\sigma_{*i}^2} + \log 2\pi\sigma_{*i}^2 \right),$$

where $\mu_{*i} = \mathbb{E}[f(\mathbf{x}_{*i})]$ and $\sigma_{*i}^2 = \mathbb{V}[f(\mathbf{x}_{*i})] + \sigma_{\text{n}}^2$ are the predictive mean and variance for test sample $i = 1, 2, \ldots, n_*$, and y_{*i} is the actual test value. The training data variance is denoted by $\text{Var}[y]$. For all experiments, the values reported are averages over ten repetitions.

We compare our solution to SOR, DTC, VAR, and FIC using the implementations provided in the GPstuff software package (version 4.3.1, see Vanhatalo et al., 2013) for Mathworks Matlab. The sparse spectrum SSGP method by Lázaro-Gredilla et al. (2010) was implemented into the GPstuff toolbox for the comparisons.[1] The reference implementation was modified such that also non-ARD covariances could be accounted for.

The m inducing inputs for SOR, DTC, VAR, and FIC were chosen at random as a subset from the training data and kept fixed between the methods. For low-dimensional inputs, this tends to lead to good results and avoid over-fitting to the training data, while optimizing the input locations alongside hyperparameters becomes the preferred approach in high input dimensions (Quiñonero-Candela and Rasmussen, 2005a). The results are averaged over ten repetitions in order to present the average performance of the methods. In Sections 6.2, 6.3, and 6.5, we used a Cartesian domain with Dirichlet boundary conditions

1. The implementation is based on the code available from Miguel Lázaro-Gredilla: `http://www.tsc.uc3m.es/~miguel/downloads.php`.



(a) Gaussian process regression solution

(b) Log marginal likelihood curves

Figure 4: (a) 256 data points generated from a GP with hyperparameters $(\sigma^2, \ell, \sigma_n^2) = (1^2, 0.1, 0.2^2)$, the full GP solution, and an approximate solution with $m = 32$. (b) Negative marginal likelihood curves for the signal variance σ^2, length-scale ℓ, and noise variance σ_n^2.

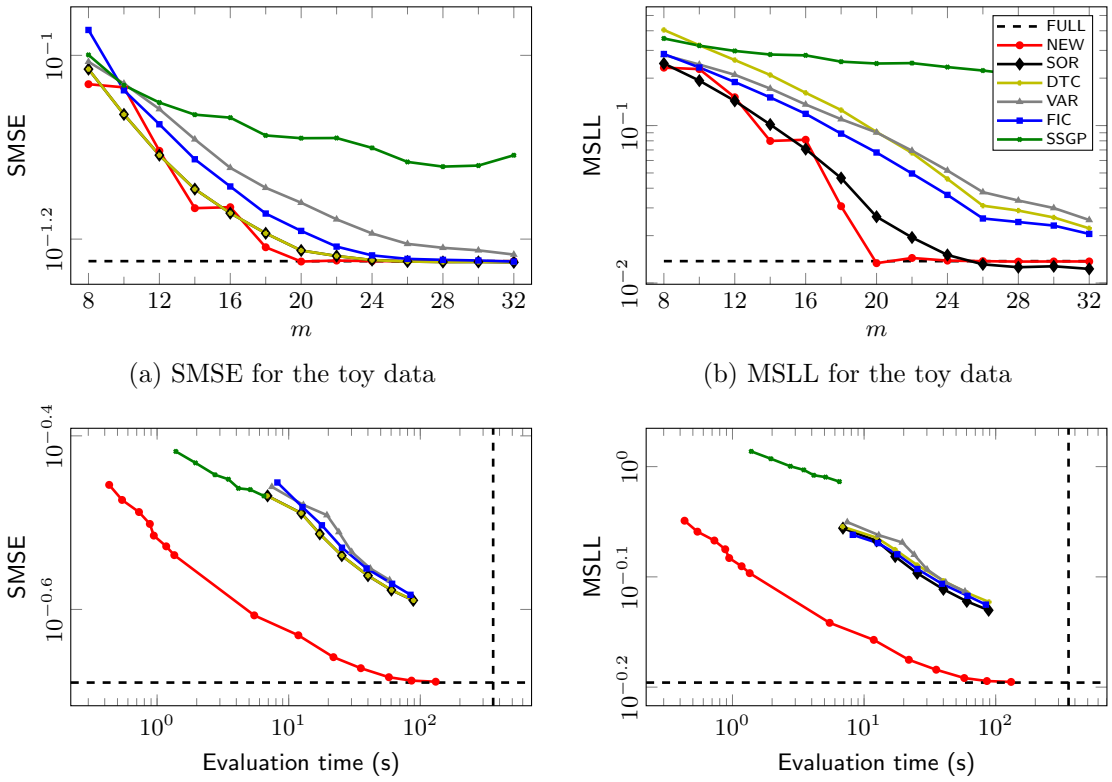
for the new reduced-rank method. To avoid boundary effects, the domain was extended by 10% outside the inputs in each direction.

In the comparisons we followed the guidelines given by Chalupka et al. (2013) for making comparisons between the actual performance of different methods. For hyperparameter optimization we used the `fminunc` routine in Matlab with a Quasi-Newton optimizer. We also tested several other algorithms, but the results were not sensitive to the choice of optimizer. The optimizer was run with a termination tolerance of 10^{-5} on the target function value and on the optimizer inputs. The number of required target function evaluations stayed fairly constant for all the comparisons, making the comparisons for the hyperparameter learning bespoke.

6.2 Toy Example

Figure 4 shows a simulated example, where 256 data points are drawn from a Gaussian process prior with a squared exponential covariance function. We use the same parametrization as Rasmussen and Williams (2006) and denote the signal variance σ^2, length-scale ℓ, and noise variance σ_n^2. Figure 4b shows the negative marginal log likelihood curves both for the full GP and the approximation with $m = 32$ basis functions. The likelihood curve approximations are almost exact and only differs from the full GP likelihood for small length-scales (roughly for values smaller than $2L/m$). Figure 4a shows the approximate GP solution. The mean estimate follows the exact GP mean, and the shaded region showing the 95% confidence area differs from the exact solution (dashed) only near the boundaries.

Figures 5a and 5b show the SMSE and MSLL values for $m = 8, 10, \ldots, 32$ inducing inputs and basis functions for the toy dataset from Figure 4. The convergence of the proposed reduced rank method is fast and a soon as the number of eigenfunctions is large



(a) SMSE for the toy data

(b) MSLL for the toy data

(c) SMSE for the precipitation data

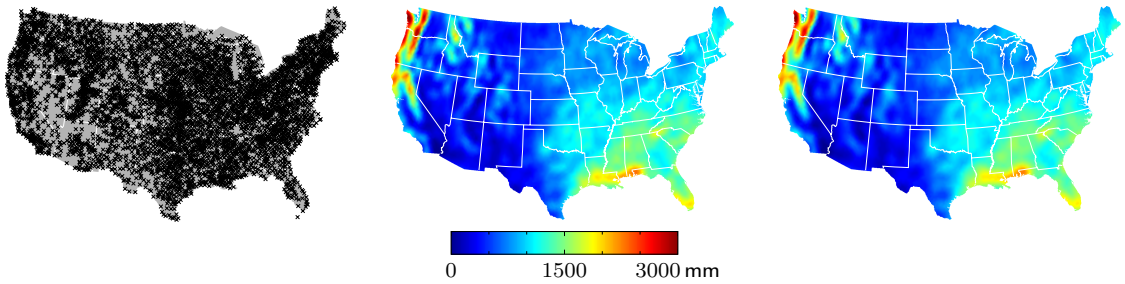
(d) MSLL for the precipitation data

Figure 5: Standardized mean squared error (SMSE) and mean standardized log loss (MSLL) results for the toy data ($d = 1$, $n = 256$) from Figure 4 and the precipitation data ($d = 2$, $n = 5776$) evaluated by 10-fold cross-validation and averaged over ten repetitions. The evaluation time includes hyperparameter learning.

enough ($m = 20$) to account for the short length-scales, the approximation converges to the exact full GP solution (shown by the dashed line).

In this case the SOR method that uses the Nyström approximation to directly approximate the spectrum of the full GP (see Section 5) seems to give good results. However, as the resulting approximation in SOR corresponds to a singular Gaussian distribution, the predictive variance is underestimated. This can be seen in Figure 5b, where SOR seems to give better results than the full GP. These results are however due to the smaller predictive variance on the test set. DTC tries to fix this shortcoming of SOR—they are identical in other respects except predictive variance evaluation—and while SOR and DTC give identical results in terms of SMSE, they differ in MSLL. We also note that additional trace term in the marginal likelihood in VAR makes the likelihood surface flat, which explains the differences in the results in comparison to DTC.

The sparse spectrum SSGP method did not perform well on average. Still, it can be seen that it converges towards the performance of the full GP. The dependence on the number of spectral points differs from the rest of the methods, and a rank of $m = 32$ is not enough to meet the other methods. However, in terms of best case performance



(a) Observation locations (b) The full GP (c) The reduced-rank method

Figure 6: Interpolation of the yearly precipitation levels using reduced-rank GP regression. Subfigure 6a shows the $n = 5776$ weather station locations. Subfigures 6b and 6c show the results for the full GP model and the new reduced-rank GP method.

over the ten repetitions with different inducing inputs and spectral points, both FIC and SSGP outperformed SOR, DTC, and VAR. Because of its 'dense spectrum' approach, the proposed reduced-rank method is not sensitive to the choice of spectral points, and thus the performance remained the same between repetitions. In terms of variance over the 10-fold cross-validation folds, the methods in order of growing variance in the figure legend (the variance approximately doubling between FULL and SSGP).

6.3 Precipitation Data

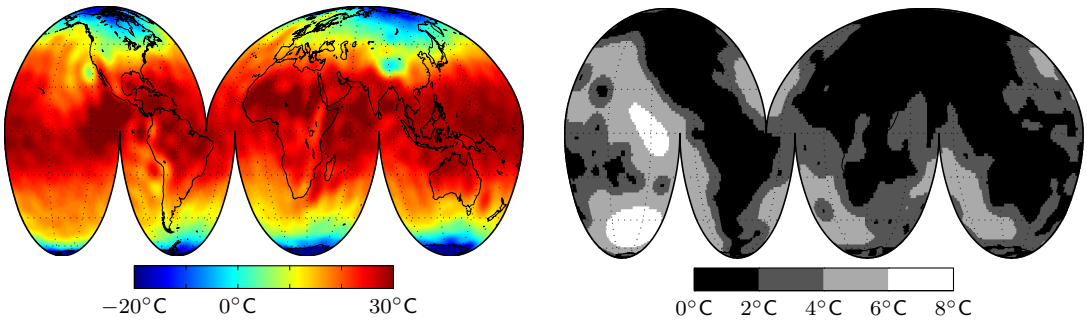
As a real-data example, we consider a precipitation data set that contain US annual precipitation summaries for year 1995 ($d = 2$ and $n = 5776$, available online, see Vanhatalo et al., 2013). The observation locations are shown on a map in Figure 6a.

We limit the number of inducing inputs and spectral points to $m = 128, 192, \ldots, 512$. For the new method we additionally consider ranks $m = 1024, 1536, \ldots, 4096$, and show that this causes a computational burden of the same order as the conventional sparse GP methods with smaller ms.

In order to demonstrate the computational benefits of the proposed model, we also present the running time of the GP inference (including hyperparameter optimization). All methods were implemented under a similar framework in the GPstuff package, and they all employ similar reformulations for numerical stability. The key difference in the evaluation times comes from hyperparameter optimization, where SOR, DTC, VAR, FIC, and SSGP scale as $\mathcal{O}(nm^2)$ for each evaluation of the marginal likelihood. The proposed reduced-rank method scales as $\mathcal{O}(m^3)$ for each evaluation (after an initial cost of $\mathcal{O}(nm^2)$).

Figures 5c and 5d show the SMSE and MSLL results for this data against evaluation time. On this scale we note that the evaluation time and accuracy, both in terms of SMSE and MSLL, are alike for SOR, DTC, VAR, and FIC. SSGP is faster to evaluate in comparison with the Nyström family of methods, which comes from the simpler structure of the approximation. Still, the number of required spectral points to meet a certain average error level is larger for SSGP.

The results for the proposed reduced-rank method (NEW) show that with two input dimensions, the required number of basis functions is larger. For the first seven points,



(a) The mean temperature (b) Standard deviation contours

Figure 7: Modeling of the yearly mean temperature on the spherical surface of the Earth ($n = 11\,028$). Figure 7b shows the standard deviation contours which match well with the continents.

we notice that even though the evaluation is two orders of magnitude faster, the method performs only slightly worse in comparison to conventional sparse methods. By considering higher ranks (the next seven points), the new method converges to the performance of the full GP (both in SMSE and MSLL), while retaining a computational time comparable to the conventional methods. This type of spatial medium-size GP regression problems can thus be solved in seconds.

Figures 6b and 6c show interpolation of the precipitation levels using a full GP model and the reduced-rank method ($m = 1728$), respectively. The results are practically identical, as is easy to confirm from the color surfaces. Obtaining the reduced-rank result (including initialization and hyperparameter learning) took slightly less than 30 seconds on a laptop computer (MacBook Air, Late 2010 model, 2.13 GHz, 4 GB RAM), while the full GP inference took approximately 18 minutes.

6.4 Temperature Data on the Surface of the Globe

We also demonstrate the use of the method in non-Cartesian coordinates. We consider modeling of the spatial mean temperature over a number of $n = 11\,028$ locations around the globe.[2]

As earlier demonstrated in Figure 2, we use the Laplace operator in spherical coordinates as defined in (31). The eigenfunctions for the angular part are the Laplace's spherical harmonics. The evaluation of the approximation does not depend on the coordinate system, and thus all the equations presented in the earlier sections remain valid. We use the squared exponential covariance function and $m = 1\,089$ basis functions.

Figure 7 visualizes the modeling outcome. The results are visualized using an interrupted projection (an adaption of the Goode homolosine projection) in order to preserve the length-scale structure across the map. The uncertainty is visualized in Figure 7b, which corresponds to the $n = 11\,028$ observation locations that are mostly spread over the continents and western countries (the white areas in Figure 7b contain no observations).

2. The data are available for download from US National Climatic Data Center: `http://www7.ncdc.noaa.gov/CDO/cdoselect.cmd` (accessed January 3, 2014).

Method	SMSE	MSLL
The reduced-rank method	0.388 (0.007)	0.608 (0.009)
Random subset ($n = 500$)	0.419 (0.035)	0.648 (0.014)
Random subset ($n = 1000$)	0.392 (0.022)	0.614 (0.010)

Table 1: Results for the apartment data set ($d = 2$, $n = 102\,890$) for predicting the log-apartment prices across England and Wales. The results for the Standardized mean squared error (SMSE) and mean standardized log loss (MSLL) were obtained by 10-fold cross-validation, where the shown values are the mean (standard deviation parenthesised).

Obtaining the reduced-rank result (including initialization and hyperparameter learning) took approximately 50 seconds on a laptop computer (MacBook Air, Late 2010 model, 2.13 GHz, 4 GB RAM), which scales with n in comparison to the evaluation time in the previous section.

6.5 Apartment Price Data

In order to fully use the computational benefits of the method, we consider a large dataset. We use records of sold apartments[3] in the UK for the period of February to October 2013. The data consist of $n = 102\,890$ records for apartments, which were cross-referenced against a postcode database to get the geographical coordinates on which the normalized logarithmic prices were regressed. The dataset is similar to that used in Hensman et al. (2013), where the records were for year 2012.

To account for both the national and regional variations in apartment prices, we used two squared exponential covariance functions with different length-scales and magnitudes. Additionally, a Gaussian noise term captures the variation that is not related to location alone. Applying the reduced-rank methodology to a sum of covariances is straight-forward, as the the kernel approximations share basis functions and only the spectra have to be summed.

To validate the results, because the full GP solution is infeasible, we used the subset of data approach as was done in Hensman et al. (2013). We solved the full GP problem by considering subsets of $n = 500$ and $n = 1000$ data points randomly chosen from the training set. For each fold in the cross-validation the results were averaged over ten choices of subsets. The rank of the reduced-rank approximation was fixed at $m = 1000$ in order to match with the larger of the two subsets.

Table 1 shows the SMSE and MSLL values for the apartment data. The results show that the reduced rank method. The results show that the proposed method gives good results in terms of both SMSE and MSLL, and the standard deviation between the folds is also small. In this case the reduced-rank result (including initialization and hyperparameter learning) took approximately 130 seconds on a laptop computer (MacBook Air, Late 2010 model, 2.13 GHz, 4 GB RAM).

3. The data are available from `http://data.gov.uk/dataset/land-registry-monthly-price-paid-data/` (accessed January 6, 2014).

7. Conclusion and Discussion

In this paper we have proposed a novel approximation scheme for forming approximate eigendecompositions of covariance functions in terms of the Laplace operator eigenbasis and the spectral density of the covariance function. The eigenfunction decomposition of the Laplacian can easily be formed in various domains, and the eigenfunctions are independent of the choice of hyperparameters of the covariance.

An advantage of the method is that it has the ability to approximate the eigendecomposition using only the eigendecomposition of the Laplacian and the spectral density of the covariance function, both of which are closed-from expressions. This together with having the eigenvectors in $\mathbf{\Phi}$ mutually orthogonal and independent of the hyperparameters, is the key to efficiency. This allows an implementation with a computational cost of $\mathcal{O}(nm^2)$ (initial) and $\mathcal{O}(m^3)$ (marginal likelihood evaluation), with negligible memory requirements.

Of the infinite number of possible basis functions only an extremely small subset are of any relevance to the GP being approximated. In GP regression the model functions are conditioned on a covariance function (kernel), which imposes desired properties on the solutions. We choose the basis functions such that they are as close as possible (w.r.t. the Frobenius norm) to those of the particular covariance function. Our method gives the exact eigendecomposition of a GP that has been constrained to be zero at the boundary of the domain.

The method allows for theoretical analysis of the error induced by the truncation of the series and the boundary effects. This is something new in this context and extremely important, for example, in medical imaging applications. The approximative eigendecomposition also opens a range of interesting possibilities for further analysis. In *learning curve* estimation, the eigenvalues of the Gaussian process can now be directly approximated. For example, we can approximate the Opper–Vivarelli bound (Opper and Vivarelli, 1999) as

$$\epsilon_{\mathrm{OV}}(n) \approx \sigma_{\mathrm{n}}^2 \sum_j \frac{S(\sqrt{\lambda_j})}{\sigma_{\mathrm{n}}^2 + n\, S(\sqrt{\lambda_j})}. \tag{64}$$

Sollich's eigenvalue based bounds (Sollich and Halees, 2002) can be approximated and analyzed in an analogous way.

However, some of these abilities come with a cost. As demonstrated throughout the paper, restraining the domain to boundary conditions introduces edge effects. These are, however, known and can be accounted for. Extrapolating with a stationary covariance function outside the training inputs only causes the predictions to revert to the prior mean and variance. Therefore we consider the boundary effects a minor problem for practical use.

A more severe limitation for applicability is the 'full' nature of the spectrum. For high-dimensional inputs the required number of basis functions grows large. There is, however, a substantial call for GPs in low-dimensional problems, especially in medical imaging applications (typical number of training data points in millions) and spatial problems. Furthermore, the mathematical formulation of the method provides a foundation for future sparse methods to build upon. A step in this direction has been taken by Lázaro-Gredilla et al. (2010), which has shown good results in high-dimensional input spaces.

Appendix A. Proofs of Convergence Theorems

A.1 Proof of Theorem 4.1

The Wiener–Khinchin identity and the symmetry of the spectral density allows us to write

$$\begin{aligned} k(x,x') &= \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)\,\exp(-\mathrm{i}\,\omega\,(x-x'))\,\mathrm{d}\omega \\ &= \frac{1}{\pi}\int_{0}^{\infty} S(\omega)\,\cos(\omega\,(x-x'))\,\mathrm{d}\omega. \end{aligned} \tag{65}$$

In a one-dimensional domain $\Omega = [-L, L]$ with Dirichet boundary conditions we have an m-term approximation of the form

$$\widetilde{k}_m(x,x') = \sum_{j=1}^{m} S\left(\frac{\pi\,j}{2L}\right)\,\frac{1}{L}\,\sin\left(\frac{\pi\,j\,(x+L)}{2L}\right)\,\sin\left(\frac{\pi\,j\,(x'+L)}{2L}\right). \tag{66}$$

We start by showing the convergence by growing the domain and therefore first consider an approximation with an infinite number of terms $m = \infty$:

$$\widetilde{k}_\infty(x,x') = \sum_{j=1}^{\infty} S(\sqrt{\lambda_j})\,\phi_j(x)\,\phi_j(x'). \tag{67}$$

Lemma A.1. *There exists a constant D_1 such that for all $x, x' \in [-\widetilde{L}, \widetilde{L}]$ we have*

$$\begin{aligned} \left|\sum_{j=1}^{\infty} S\left(\frac{\pi\,j}{2L}\right)\,\frac{1}{L}\,\sin\left(\frac{\pi\,j\,(x+L)}{2L}\right)\,\sin\left(\frac{\pi\,j\,(x'+L)}{2L}\right)\right. \\ \left. -\,\frac{1}{\pi}\int_{0}^{\infty} S(\omega)\,\cos(\omega\,(x-x'))\,\mathrm{d}\omega\right| \leq \frac{D_1}{L}. \end{aligned} \tag{68}$$

That is,

$$\left|\widetilde{k}_\infty(x,x') - k(x,x')\right| \leq \frac{D_1}{L}, \quad \text{for } x, x' \in [-\widetilde{L}, \widetilde{L}]. \tag{69}$$

Proof. We can rewrite the summation in (68) as

$$\sum_{j=1}^{\infty} S\left(\frac{\pi\,j}{2L}\right)\,\frac{1}{L}\,\sin\left(\frac{\pi\,j\,(x+L)}{2L}\right)\,\sin\left(\frac{\pi\,j\,(x'+L)}{2L}\right) \tag{70}$$

$$\begin{aligned} &= \sum_{j=1}^{\infty} S\left(\frac{\pi\,j}{2L}\right)\,\cos\left(\frac{\pi\,j\,(x-x')}{2L}\right)\,\frac{1}{2L} \\ &\quad - \frac{1}{2L}\sum_{j=1}^{\infty}\left[S\left(\frac{\pi\,2j}{2L}\right) - S\left(\frac{\pi\,(2j-1)}{2L}\right)\right]\,\cos\left(\frac{\pi\,2j\,(x+x')}{2L}\right) \\ &\quad - \frac{1}{2L}\sum_{j=1}^{\infty} S\left(\frac{\pi\,(2j-1)}{2L}\right)\,\left[\cos\left(\frac{\pi\,2j\,(x+x')}{2L}\right) - \cos\left(\frac{\pi\,(2j-1)\,(x+x')}{2L}\right)\right] \end{aligned} \tag{71}$$

First consider the first term above in Equation (71). Let $\Delta = \frac{\pi}{2L}$, and thus it can be seen to have the form

$$\frac{1}{\pi} \sum_{j=1}^{\infty} S\left(\Delta\, j\right) \cos\left(\Delta\, j\, (x - x')\right)\, \Delta, \tag{72}$$

which can be recognized as a Riemannian sum approximation to the integral $\frac{1}{\pi} \int_0^{\infty} S(\omega)\, \cos(\omega\, (x - x'))\, \mathrm{d}\omega$. Because we assume that $x, x' \in [-\widetilde{L}, \widetilde{L}]$, the integrand and its derivatives are bounded and because the integral $\int_{-\infty}^{\infty} S(\omega)\, \mathrm{d}\omega < \infty$, the Riemannian integral converges, and hence we conclude that

$$\left| \sum_{j=1}^{\infty} S\left(\frac{\pi\, j}{2L}\right) \cos\left(\frac{\pi\, j\, (x - x')}{2L}\right)\, \frac{1}{2L} - \frac{1}{\pi} \int_0^{\infty} S(\omega)\, \cos(\omega\, (x - x'))\, \mathrm{d}\omega \right| \leq \frac{D_2}{L} \tag{73}$$

for some constant D_2.

The second summation term in Equation (71) can also be interpreted as a Riemann sum if we set $\Delta = \frac{\pi}{L}$:

$$\begin{aligned}
&\frac{1}{2L} \sum_{j=1}^{\infty} \left[S\left(\Delta\, j\right)\, - S\left(\Delta\, j - \Delta/2\right)\right] \cos\left(\Delta\, (x + x')\right) \\
&\quad = \frac{1}{2L} \sum_{j=1}^{\infty} \frac{1}{\Delta}\, \left[S\left(\Delta\, j\right)\, - S\left(\Delta\, j - \Delta/2\right)\right] \cos\left(\Delta\, (x + x')\right)\, \Delta \\
&\quad \approx \frac{1}{2L} \int_0^{\infty} 2\, S'(\omega)\, \cos(\omega\, (x + x'))\, \mathrm{d}\omega.
\end{aligned} \tag{74}$$

Because we assumed that also the second derivative of $S(\cdot)$ is bounded, the derivative and the Riemann sum converge (alternatively, we could analyze the sums as a Stieltjes integral with respect to a differentiable function), and hence the exists a constant D_3' such that

$$\begin{aligned}
\Bigg| &\frac{1}{2L} \sum_{j=1}^{\infty} \left[S\left(\Delta\, j\right)\, - S\left(\Delta\, j - \Delta/2\right)\right] \cos\left(\Delta\, (x + x')\right) \\
&\qquad - \frac{1}{2L} \int_0^{\infty} 2\, S'(\omega)\, \cos(\omega\, (x + x'))\, \mathrm{d}\omega \Bigg| \leq \frac{D_3'}{L}
\end{aligned} \tag{75}$$

But now because $\int_0^{\infty} 2\, S'(\omega)\, \cos(\omega\, (x + x'))\, \mathrm{d}\omega < \infty$, this actually implies that

$$\left| \frac{1}{2L} \sum_{j=1}^{\infty} \left[S\left(\frac{\pi\, 2j}{2L}\right)\, - S\left(\frac{\pi\, (2j - 1)}{2L}\right)\right] \cos\left(\frac{\pi\, 2j\, (x + x')}{2L}\right) \right| \leq \frac{D_3}{L} \tag{76}$$

for some constant D_3. For the last summation term in Equation (71) we get the interpretation

$$\begin{aligned}
&\frac{1}{2L}\sum_{j=1}^{\infty} S\left(\frac{\pi\,(2j-1)}{2L}\right)\,\left[\cos\left(\frac{\pi\,2j\,(x+x')}{2L}\right)-\cos\left(\frac{\pi\,(2j-1)\,(x+x')}{2L}\right)\right]\\
&\quad\approx \frac{1}{2L}\int_0^{\infty} S(\omega)\,2\,\left[\frac{\mathrm{d}}{\mathrm{d}\omega}\,\cos(\omega\,(x+x'))\right]\,\mathrm{d}\omega\\
&\quad= -\frac{1}{2L}\int_0^{\infty} S(\omega)\,2\,(x+x')\,\sin(\omega\,(x+x'))\,\mathrm{d}\omega,
\end{aligned} \tag{77}$$

which by boundedness of x and x' implies

$$\left|\frac{1}{2L}\sum_{j=1}^{\infty} S\left(\frac{\pi\,(2j-1)}{2L}\right)\,\left[\cos\left(\frac{\pi\,2j\,(x+x')}{2L}\right)-\cos\left(\frac{\pi\,(2j-1)\,(x+x')}{2L}\right)\right]\right| \leq \frac{D_4}{L} \tag{78}$$

for some constant D_4. The result now follows by combining (73), (76), and (78) via the triangle inequality. □

Let us now return to the original question, and consider what happens when we replace the infinite sum approximation with a finite m number of terms. We are now interested in

$$\widetilde{k}_{\infty}(x,x') - \widetilde{k}_m(x,x') = \sum_{j=m+1}^{\infty} S\left(\frac{\pi\,j}{2L}\right)\,\frac{1}{L}\,\sin\left(\frac{\pi\,j\,(x+L)}{2L}\right)\,\sin\left(\frac{\pi\,j\,(x'+L)}{2L}\right). \tag{79}$$

Lemma A.2. *There exists a constant D_5 such that for all $x, x' \in [-\widetilde{L}, \widetilde{L}]$ we have*

$$\left|\widetilde{k}_{\infty}(x,x') - \widetilde{k}_m(x,x')\right| \leq \frac{D_5}{L} + \frac{2}{\pi}\int_{\frac{\pi\,m}{2L}}^{\infty} S(\omega)\,\mathrm{d}\omega. \tag{80}$$

Proof. Because the sinusoidals are bounded by unity, we get

$$\left|\sum_{j=m+1}^{\infty} S\left(\frac{\pi\,j}{2L}\right)\,\frac{1}{L}\,\sin\left(\frac{\pi\,j\,(x+L)}{2L}\right)\,\sin\left(\frac{\pi\,j\,(x'+L)}{2L}\right)\right| \leq \left|\sum_{j=m+1}^{\infty} S\left(\frac{\pi\,j}{2L}\right)\,\frac{1}{L}\right|. \tag{81}$$

The right-hand term can now be seen as Riemann sum approximation to the integral

$$\sum_{j=m+1}^{\infty} S\left(\frac{\pi\,j}{2L}\right)\,\frac{1}{L} \approx \frac{2}{\pi}\int_{\frac{\pi\,m}{2L}}^{\infty} S(\omega)\,\mathrm{d}\omega. \tag{82}$$

Our assumptions ensure that this integral converges and hence there exists a constant D_5 such that

$$\left|\sum_{j=m+1}^{\infty} S\left(\frac{\pi\,j}{2L}\right)\,\frac{1}{L} - \frac{2}{\pi}\int_{\frac{\pi\,m}{2L}}^{\infty} S(\omega)\,\mathrm{d}\omega\right| \leq \frac{D_5}{L}. \tag{83}$$

Hence by the triangle inequality we get

$$\begin{aligned}\left|\sum_{j=m+1}^{\infty} S\left(\frac{\pi\, j}{2L}\right)\, \frac{1}{L}\right| &= \left|\sum_{j=m+1}^{\infty} S\left(\frac{\pi\, j}{2L}\right)\, \frac{1}{L} - \frac{2}{\pi}\int_{\frac{\pi\, m}{2L}}^{\infty} S(\omega)\,\mathrm{d}\omega + \frac{2}{\pi}\int_{\frac{\pi\, m}{2L}}^{\infty} S(\omega)\,\mathrm{d}\omega\right| \\ &\leq \left|\sum_{j=m+1}^{\infty} S\left(\frac{\pi\, j}{2L}\right)\, \frac{1}{L} - \frac{2}{\pi}\int_{\frac{\pi\, m}{2L}}^{\infty} S(\omega)\,\mathrm{d}\omega\right| + \frac{2}{\pi}\int_{\frac{\pi\, m}{2L}}^{\infty} S(\omega)\,\mathrm{d}\omega \\ &\leq \frac{D_5}{L} + \frac{2}{\pi}\int_{\frac{\pi\, m}{2L}}^{\infty} S(\omega)\,\mathrm{d}\omega \end{aligned} \tag{84}$$

and thus the result follows. □

The above result can now easily be combined to a proof of the one-dimensional convergence theorem as follows:

Proof of Theorem 4.1. The first result follows by combining Lemmas A.1 and A.2 via the triangle inequality. Because our assumptions imply that

$$\lim_{x\to\infty}\int_{x}^{\infty} S(\omega)\,\mathrm{d}\omega = 0, \tag{85}$$

for any fixed L we have

$$\lim_{m\to\infty}\left[\frac{E}{L} + \frac{2}{\pi}\int_{\frac{\pi\, m}{2L}}^{\infty} S(\omega)\,\mathrm{d}\omega\right] \to \frac{E}{L}. \tag{86}$$

If we now take the limit $L \to \infty$, the second result in the theorem follows. □

A.2 Proof of Theorem 4.4

When $\mathbf{x} \in \mathbb{R}^d$, the Wiener–Khinchin identity and symmetry of the spectral density imply that

$$\begin{aligned} k(\mathbf{x},\mathbf{x}') &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} S(\boldsymbol{\omega})\, \exp(-\mathrm{i}\,\boldsymbol{\omega}^{\mathsf{T}}(\mathbf{x}-\mathbf{x}'))\,\mathrm{d}\boldsymbol{\omega} \\ &= \frac{1}{\pi^d}\int_0^{\infty}\cdots\int_0^{\infty} S(\boldsymbol{\omega})\, \prod_{k=1}^{d}\cos(\omega_k\,(x_k - x_k'))\,\mathrm{d}\omega_1\cdots\mathrm{d}\omega_d. \end{aligned} \tag{87}$$

The $m = \hat{m}^d$ term approximation now has the form

$$\widetilde{k}_m(\mathbf{x},\mathbf{x}') = \sum_{j_1,\ldots,j_d=1}^{\hat{m}} S\left(\frac{\pi\, j_1}{2L_1},\ldots,\frac{\pi\, j_d}{2L_d}\right)\prod_{k=1}^{d}\frac{1}{L_k}\,\sin\left(\frac{\pi\, j_k\,(x_k + L_k)}{2L_k}\right)\sin\left(\frac{\pi\, j_k\,(x_k' + L_k)}{2L_k}\right). \tag{88}$$

As in the one-dimensional problem we start by considering the case where $\hat{m} = \infty$.

Lemma A.3. *There exists a constant D_1 such that for all $\mathbf{x}, \mathbf{x}' \in [-\widetilde{L}, \widetilde{L}]^d$ we have*

$$\left| \sum_{j_1,\ldots,j_d=1}^{\infty} S\left(\frac{\pi\, j_1}{2L_1}, \ldots, \frac{\pi\, j_d}{2L_d}\right) \prod_{k=1}^{d} \frac{1}{L_k} \sin\left(\frac{\pi\, j_k\, (x_k + L_k)}{2L_k}\right) \sin\left(\frac{\pi\, j_k\, (x'_k + L_k)}{2L_k}\right) - \frac{1}{\pi^d} \int_0^\infty \cdots \int_0^\infty S(\boldsymbol{\omega}) \prod_{k=1}^{d} \cos(\omega_k\, (x - x'))\, \mathrm{d}\omega_1 \cdots \mathrm{d}\omega_d \right| \leq D_1 \sum_{k=1}^{d} \frac{1}{L_k}. \quad (89)$$

That is,

$$\left| \widetilde{k}_\infty(\mathbf{x}, \mathbf{x}') - k(\mathbf{x}, \mathbf{x}') \right| \leq D_1 \sum_{k=1}^{d} \frac{1}{L_k} \quad \text{for } \mathbf{x}, \mathbf{x}' \in [-\widetilde{L}, \widetilde{L}]^d. \quad (90)$$

Proof. We can separate the summation over j_1 in the summation term above as follows:

$$\sum_{j_2,\ldots,j_d=1}^{\infty} \left[\sum_{j_1=1}^{\infty} S\left(\frac{\pi\, j_1}{2L_1}, \ldots, \frac{\pi\, j_d}{2L_d}\right) \sin\left(\frac{\pi\, j_1\, (x_1 + L_1)}{2L_1}\right) \sin\left(\frac{\pi\, j_1\, (x'_1 + L_1)}{2L_1}\right) \right] \times \prod_{k=2}^{d} \frac{1}{L_k} \sin\left(\frac{\pi\, j_k\, (x_k + L_k)}{2L_k}\right) \sin\left(\frac{\pi\, j_k\, (x'_k + L_k)}{2L_k}\right). \quad (91)$$

By Lemma A.1 there now exists a constant $D_{1,1}$ such that

$$\left| \sum_{j_1=1}^{\infty} S\left(\frac{\pi\, j_1}{2L_1}, \ldots, \frac{\pi\, j_d}{2L_d}\right) \sin\left(\frac{\pi\, j_1\, (x_1 + L_1)}{2L_1}\right) \sin\left(\frac{\pi\, j_1\, (x'_1 + L_1)}{2L_1}\right) - \frac{1}{\pi} \int_0^\infty S\left(\omega_1, \frac{\pi\, j_2}{2L_2}, \ldots, \frac{\pi\, j_d}{2L_d}\right) \cos(\omega_1\, (x_1 - x'_1))\, \mathrm{d}\omega_1 \right| \leq \frac{D_{1,1}}{L_1}. \quad (92)$$

The triangle inequality then gives

$$\begin{aligned} &\left| \sum_{j_1,\ldots,j_d=1}^{\infty} S\left(\frac{\pi\, j_1}{2L_1}, \ldots, \frac{\pi\, j_d}{2L_d}\right) \prod_{k=1}^{d} \frac{1}{L_k} \sin\left(\frac{\pi\, j_k\, (x_k + L_k)}{2L_k}\right) \sin\left(\frac{\pi\, j_k\, (x'_k + L_k)}{2L_k}\right) - \frac{1}{\pi^d} \int_0^\infty \cdots \int_0^\infty S(\boldsymbol{\omega}) \prod_{k=1}^{d} \cos(\omega_j\, (x_k - x'_k))\, \mathrm{d}\omega_1 \cdots \mathrm{d}\omega_d \right| \quad (93) \\ &\leq \frac{D_{1,1}}{L_1} + \left| \frac{1}{\pi} \sum_{j_2,\ldots,j_d=1}^{\infty} \int_0^\infty S\left(\omega_1, \frac{\pi\, j_2}{2L_2}, \ldots, \frac{\pi\, j_d}{2L_d}\right) \cos(\omega_1\, (x_1 - x'_1))\, \mathrm{d}\omega_1 \right. \\ &\qquad \times \prod_{k=2}^{d} \frac{1}{L_k} \sin\left(\frac{\pi\, j_k\, (x_k + L_k)}{2L_k}\right) \sin\left(\frac{\pi\, j_k\, (x'_k + L_k)}{2L_k}\right) \\ &\qquad \left. - \frac{1}{\pi^d} \int_0^\infty \cdots \int_0^\infty S(\boldsymbol{\omega}) \prod_{k=1}^{d} \cos(\omega_k\, (x_k - x'_k))\, \mathrm{d}\omega_1 \cdots \mathrm{d}\omega_d \right|. \quad (94) \end{aligned}$$

We can now similarly bound with respect to the summations over $j_2, \ldots, j_d$ which leads to a bound of the form $\frac{D_{1,1}}{L_1} + \cdots + \frac{D_{1,d}}{L_d}$. Taking $D_1 = \max_k D_{1,k}$ leads to the desired result. □

Now we can consider what happens in the finite truncation of the series. That is, we analyze the following residual sum

$$\begin{aligned}&\widetilde{k}_\infty(\mathbf{x},\mathbf{x}') - \widetilde{k}_m(\mathbf{x},\mathbf{x}') \\ &\quad= \sum_{j_1,\ldots,j_d=\hat{m}+1}^{\infty} S\left(\frac{\pi\, j_1}{2L_1},\ldots,\frac{\pi\, j_d}{2L_d}\right) \prod_{k=1}^{d} \frac{1}{L_k} \sin\left(\frac{\pi\, j_k\,(x_k+L_k)}{2L_k}\right) \sin\left(\frac{\pi\, j_k\,(x'_k+L_k)}{2L_k}\right).\end{aligned} \tag{95}$$

Lemma A.4. *The exists a constant D_2 such that for all $\mathbf{x},\mathbf{x}' \in [-\widetilde{L},\widetilde{L}]^d$ we have*

$$\left|\widetilde{k}_\infty(\mathbf{x},\mathbf{x}') - \widetilde{k}_m(\mathbf{x},\mathbf{x}')\right| \leq \frac{D_2}{L} + \frac{1}{\pi^d} \int_{\|\boldsymbol{\omega}\| \geq \frac{\pi\,\hat{m}}{2L}} S(\boldsymbol{\omega})\,\mathrm{d}\boldsymbol{\omega}, \tag{96}$$

where $L = \min_k L_k$.

Proof. We can write the following bound

$$\begin{aligned}&\left| \sum_{j_1,\ldots,j_d=\hat{m}+1}^{\infty} S\left(\frac{\pi\, j_1}{2L_1},\ldots,\frac{\pi\, j_d}{2L_d}\right) \prod_{k=1}^{d} \frac{1}{L_k} \sin\left(\frac{\pi\, j_k\,(x_k+L_k)}{2L_k}\right) \sin\left(\frac{\pi\, j_k\,(x'_k+L_k)}{2L_k}\right) \right| \\ &\qquad\leq \left| \sum_{j_1,\ldots,j_d=\hat{m}+1}^{\infty} S\left(\frac{\pi\, j_1}{2L_1},\ldots,\frac{\pi\, j_d}{2L_d}\right) \prod_{k=1}^{d} \frac{1}{L_k} \right|.\end{aligned} \tag{97}$$

The summation over the index j_1 can now be interpreted as a Riemann integral approximation with $\Delta = \frac{\pi}{2L_1}$ giving

$$\begin{aligned}&\left| \sum_{j_1,\ldots,j_d=\hat{m}+1}^{\infty} S\left(\frac{\pi\, j_1}{2L_1},\ldots,\frac{\pi\, j_d}{2L_d}\right) \prod_{k=1}^{d} \frac{1}{L_k} \right. \\ &\qquad\left. - \frac{2}{\pi} \sum_{j_2,\ldots,j_d=\hat{m}+1}^{\infty} \int_{\frac{\pi\,\hat{m}}{2L_1}}^{\infty} S\left(\omega_1, \frac{\pi\, j_2}{2L_2},\ldots,\frac{\pi\, j_d}{2L_d}\right) \mathrm{d}\omega_1 \prod_{k=2}^{d} \frac{1}{L_k} \right| \leq \frac{D_{2,1}}{L_1}.\end{aligned} \tag{98}$$

Using a similar argument again, we get

$$\begin{aligned}&\left| \frac{2}{\pi} \sum_{j_2,\ldots,j_d=\hat{m}+1}^{\infty} \int_{\frac{\pi\,\hat{m}}{2L_1}}^{\infty} S\left(\omega_1, \frac{\pi\, j_2}{2L_2},\ldots,\frac{\pi\, j_d}{2L_d}\right) \mathrm{d}\omega_1 \prod_{k=2}^{d} \frac{1}{L_k} \right. \\ &\qquad\left. - \frac{2^2}{\pi^2} \sum_{j_3,\ldots,j_d=\hat{m}+1}^{\infty} \int_{\frac{\pi\,\hat{m}}{2L_1}}^{\infty} \int_{\frac{\pi\,\hat{m}}{2L_2}}^{\infty} S\left(\omega_1, \omega_2, \frac{\pi\, j_3}{2L_3},\ldots,\frac{\pi\, j_d}{2L_d}\right) \mathrm{d}\omega_1\,\mathrm{d}\omega_2 \prod_{k=3}^{d} \frac{1}{L_k} \right| \leq \frac{D_{2,2}}{L_2}.\end{aligned} \tag{99}$$

After repeating this for all the indexes, by forming a telescoping sum of the terms and applying the triangle inequality then gives

$$\begin{aligned}&\left| \sum_{j_1,\ldots,j_d=\hat{m}+1}^{\infty} S\left(\frac{\pi\, j_1}{2L_1},\ldots,\frac{\pi\, j_d}{2L_d}\right) \prod_{k=1}^{d} \frac{1}{L_k} \right. \\ &\qquad\left. - \left(\frac{2}{\pi}\right)^d \int_{\frac{\pi\,\hat{m}}{2L_1}}^{\infty} \cdots \int_{\frac{\pi\,\hat{m}}{2L_d}}^{\infty} S(\omega_1,\ldots,\omega_d)\,\mathrm{d}\omega_1 \cdots \mathrm{d}\omega_d \right| \leq \sum_{k=1}^{d} \frac{D_{2,k}}{L_k}.\end{aligned} \tag{100}$$

Applying the triangle inequality again gives

$$\left| \sum_{j_1,\ldots,j_d=\hat{m}+1}^{\infty} S\left(\frac{\pi\, j_1}{2L_1},\ldots,\frac{\pi\, j_d}{2L_d}\right) \prod_{k=1}^{d} \frac{1}{L_k} \right| \leq \sum_{k=1}^{d} \frac{D_{2,k}}{L_k} + \left(\frac{2}{\pi}\right)^d \int_{\frac{\pi\,\hat{m}}{2L_1}}^{\infty} \cdots \int_{\frac{\pi\,\hat{m}}{2L_d}}^{\infty} S(\omega_1,\ldots,\omega_d)\,\mathrm{d}\omega_1 \cdots \mathrm{d}\omega_d. \quad (101)$$

By interpreting the latter integral as being over the positive exterior of a rectangular hypercuboid and bounding it by a integral over exterior of a hypersphere which fits inside the cuboid, we can bound the expression by

$$\sum_{k=1}^{d} \frac{D_{2,k}}{L_k} + \frac{1}{\pi^d} \int_{\|\boldsymbol{\omega}\| \geq \frac{\pi\,\hat{m}}{2L}} S(\boldsymbol{\omega})\,\mathrm{d}\boldsymbol{\omega}. \quad (102)$$

The first term can be further bounded by replacing L_ks with their minimum L and by defining a new constant D_2 which is d times the maximum of $D_{2,k}$. This leads to the final form of the result. □

Proof of Theorem 4.4. Analogous to the one-dimensional case. That is, we combine the results of the above lemmas using the triangle inequality. □

References

Naum I. Akhiezer and Izrail' M. Glazman. *Theory of Linear Operators in Hilbert Space*. Dover, New York, 1993.

Christopher T. H. Baker. *The Numerical Treatment of Integral Equations*. Clarendon press, Oxford, 1977.

Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC, 2011.

Krzysztof Chalupka, Christopher K. I. Williams, and Iain Murray. A framework for evaluating approximation methods for Gaussian process regression. *Journal of Machine Learning Research*, 14:333–350, 2013.

Richard Courant and David Hilbert. *Methods of Mathematical Physics*, volume 1. Wiley-VCH, 2008.

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

Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic Equations in Infinite Dimensions*, volume 45 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 1992.

Simon Duane, Anthony D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

Jochen Fritz, Insa Neuweiler, and Wolfgang Nowak. Application of FFT-based algorithms for large-scale universal kriging problems. *Mathematical Geosciences*, 41(5):509–533, 2009.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, third edition, 1996.

James Hensman, Nicolò Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI 2013)*, pages 282–290, 2013.

Miguel Lázaro-Gredilla. *Sparse Gaussian Processes for Large-Scale Machine Learning*. PhD thesis, Universidad Carlos III de Madrid, 2010.

Miguel Lázaro-Gredilla, Joaquin Quiñonero-Candela, Carl Edward Rasmussen, and Aníbal R. Figueiras-Vidal. Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, 11:1865–1881, 2010.

Peter J. Lenk. Towards a practicable Bayesian nonparametric density estimator. *Biometrika*, 78(3):531–543, 1991.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73 (4):423–498, 2011.

Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, New York, 2001.

Radford M. Neal. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, chapter 5. Chapman & Hall/CRC, 2011.

Manfred Opper and Francesco Vivarelli. General bounds on Bayes errors for regression with Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 11, pages 302–308, 1999.

Christopher J. Paciorek. Bayesian smoothing with Gaussian processes using Fourier basis functions in the spectralGP package. *Journal of Statistical Software*, 19(2):1–38, 2007.

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

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. Analysis of some methods for reduced rank Gaussian process regression. In *Switching and Learning in Feedback Systems*, volume 3355 of *Lecture Notes in Computer Science*, pages 98–127. Springer, 2005b.

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

Simo Särkkä and Jouni Hartikainen. Infinite-dimensional Kalman filtering approach to spatio-temporal Gaussian process regression. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*, volume 22 of *JMLR Workshop and Conference Proceedings*, pages 993–1001, 2012.

Simo Särkkä, Arno Solin, and Jouni Hartikainen. Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing. *IEEE Signal Processing Magazine*, 30(4): 51–61, 2013.

Matthias Seeger, Christopher K. I. Williams, and Neil D. Lawrence. Fast forward selection to speed up sparse Gaussian process regression. In *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics (AISTATS 2003)*, 2003.

Ralph E. Showalter. *Hilbert Space Methods in Partial Differential Equations*. Dover Publications, 2010.

Mikhail A. Shubin. *Pseudodifferential Operators and Spectral Theory*. Springer Series in Soviet Mathematics. Springer-Verlag, 1987.

Alex J. Smola and Peter Bartlett. Sparse greedy Gaussian process regression. In *Advances in Neural Information Processing Systems*, volume 13, 2001.

Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, volume 18, pages 1259–1266, 2006.

Peter Sollich and Anason Halees. Learning curves for Gaussian process regression: Approximations and bounds. *Neural Computation*, 14(6):1393–1428, 2002.

Michalis K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS 2009)*, volume 5 of *JMLR Workshop and Conference Proceedings*, pages 567–574, 2009.

Jarno Vanhatalo, Ville Pietiläinen, and Aki Vehtari. Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in Medicine*, 29(15):1580–1607, 2010.

Jarno Vanhatalo, Jaakko Riihimäki, Jouni Hartikainen, Pasi Jylänki, Ville Tolvanen, and Aki Vehtari. GPstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research*, 14:1175–1179, 2013.

Christopher K. I. Williams and Matthias Seeger. The effect of the input density distribution on kernel-based classifiers. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.

Christopher K. I. Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, volume 13, 2001.