Poisson Intensity Estimation with Reproducing Kernels

Seth Flaxman, Yee Whye Teh, Dino Sejdinovic
Department of Statistics
University of Oxford
{flaxman,y.w.teh,dino.sejdinovic}@stats.ox.ac.uk

Abstract

Despite the fundamental nature of the Poisson process in the theory and application of stochastic processes, and its attractive generalizations (e.g. Cox process), few tractable nonparametric modeling approaches exist, especially in multiple dimensions. In this paper we develop a new Reproducing Kernel Hilbert Space (RKHS) formulation for the inhomogeneous Poisson process. We model the square root of the intensity as an RKHS function. The modeling challenge is that the usual representer theorem arguments no longer apply due to the form of the inhomogeneous Poisson process likelihood. However, we prove that the representer theorem does hold in an appropriately transformed RKHS, guaranteeing that the optimization of the penalized likelihood can be cast as a finite-dimensional problem. The resulting approach is simple to implement, scales to multiple dimensions and can readily be extended to handle large-scale datasets.

1 Introduction

Poisson processes are ubiquitous in statistical science, with a long history spanning both theory (e.g. [16]) and applications (e.g. [10]), especially in the spatial statistics and time series literature. Despite their ubiquity, fundamental questions in their application to real datasets remain open. Namely, scalable nonparametric models for intensity functions of inhomogeneous Poisson processes are not well understood, especially in multiple dimensions since the standard approaches are akin to density estimation. In this contribution, we propose a step towards such scalable nonparametric modeling and introduce a new Reproducing Kernel Hilbert Space (RKHS) formulation for inhomogeneous Poisson process modeling, which is based on the Empirical Risk Minimization (ERM) framework. We model the square root of the intensity as an RKHS function and consider a risk functional given by a penalized version of the inhomogeneous Poisson process likelihood. While standard representer theorem arguments do not apply directly due to the form of the likelihood—as a counting process, a Poisson process is fundamentally different from standard probability distributions because the observation that *no points* occur in some region is just as important as the locations of the points that do occur, and thus the likelihood depends not only on the evaluations of the intensity at the observed points, but also on its integral across the domain of interest. Nevertheless, we prove a version of the representer theorem in an appropriately adjusted RKHS. The adjusted RKHS coincides with the original RKHS as a space of functions but has a different inner product structure. This allows us to cast the estimation problem as an optimization over a finite-dimensional subspace of the adjusted RKHS. The derived method is demonstrated to give better performance than a naïve unadjusted RKHS method which resorts to an optimization over a subspace without representer theorem guarantees. We describe cases where adjusted RKHS can be described with explicit Mercer expansions as well as numerical approximations where Mercer expansions are not available. We observe strong performance of the proposed method on a variety of synthetic, environmental and bioinformatics data.

2 Background and Related Work

2.1 Poisson process

We briefly state relevant definitions for point processes over domains $S \subset \mathbb{R}^D$, following [7]. For Lebesgue measurable subsets $T \subset S$, $N(T)$ denotes the number of events in $T \subset S$. $N(\cdot)$ is a stochastic process characterizing the point process. Our focus is on providing a nonparametric estimator for the first-order intensity of a point process, which is defined as:

$$\lambda(s) = \lim_{|ds| \to 0} \mathbb{E}[N(ds))]/|ds| \tag{1}$$

The inhomogeneous Poisson process is driven solely by the intensity function $\lambda(\cdot)$:

$$N(T) \sim \text{Poisson}(\int_T \lambda(x)dx) \tag{2}$$

In the homogeneous Poisson process, $\lambda(x) = \lambda$ is constant, so the number of points in any region T simply depends on the volume of T, which we denote $|T|$:

$$N(T) \sim \text{Poisson}(\lambda|T|) \tag{3}$$

Assuming that $\lambda(x)$ is known (either deterministically parameterized with known parameters or if we are in a hierarchical model like the Cox process, then conditional on a realization), we have the following likelihood function for a set of $N = N(S)$ points $x_1, \ldots, x_N$ observed over a fixed window S:

$$\mathcal{L}(x_1, \ldots, x_N | \lambda(x)) = \prod_{i=1}^{N} \lambda(x_i) \text{exp}\left(-\int_S \lambda(x)dx\right) \tag{4}$$

2.2 Reproducing Kernel Hilbert Spaces

Given a non-empty domain S and a positive definite kernel function $k : S \times S \to \mathbb{R}$, there exists a unique reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$. RKHS is a space of functions $f : S \to \mathbb{R}$ where evaluation is a continuous functional, and can thus be represented by an inner product $f(x) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}_k}$ for all $f \in \mathcal{H}_k, x \in S$ (reproducing property). While $\mathcal{H}_k$ is in most interesting cases an infinite-dimensional space of functions, due to the classical representer theorem [15], [24, Section 4.2], optimization over $\mathcal{H}_k$ is typically a tractable finite-dimensional problem. In particular, if we have a set of N observations $x_1, \ldots, x_N$, $x_i \in S$ and consider the problem

$$\min_{f \in \mathcal{H}_k} \left\{R\left(f(x_1), \ldots, f(x_N)\right) + \Omega\left(\|f\|_{\mathcal{H}_k}\right)\right\}. \tag{5}$$

where $R\left(f(x_1), \ldots, f(x_N)\right)$ depends on f through its evaluations on the set of observations only, and Ω is a non-decreasing function of the RKHS norm of f, there exists a solution to Eq. (5) of the form $f^*(\cdot) = \sum_{i=1}^{N} \alpha_i k(x_i, \cdot)$, and the optimization can thus be cast in terms of $\alpha \in \mathbb{R}^N$. This formulation is widely used in the framework of regularized Empirical Risk Minimization (ERM) for supervised learning, where $R\left(f(x_1), \ldots, f(x_N)\right) = \frac{1}{N}\sum_{i=1}^{N} L(f(x_i), y_i)$ is the empirical risk corresponding to a loss function L.

If domain S is compact and kernel k is continuous, one can assign to k its integral kernel operator $T_k : \mathcal{L}_2(S) \to \mathcal{L}_2(S)$, given by $T_k g = \int_S k(x, \cdot)g(x)dx$, which is positive, self-adjoint and compact. There thus exists an orthonormal set of eigenfunctions $\{e_j\}_{j=1}^{\infty}$ of T_k, and the corresponding eigenvalues $\{\eta_j\}_{j=1}^{\infty}$. This spectral decomposition of T_k leads to Mercer's representation of kernel function k [24, Section 2.2]:

$$k(x, x') = \sum_{j=1}^{\infty} \eta_j e_j(x) e_j(x'), \qquad x, x' \in S \tag{6}$$

with uniform convergence on $S \times S$. Any function $f \in \mathcal{H}_k$ can then be written as $f = \sum_j b_j e_j$ where $\|f\|_{\mathcal{H}_k}^2 = \sum_j b_j^2/\eta_j < \infty$.

2.3 Related work

The classic approach to nonparametric intensity estimation is based on smoothing kernels [22, 9] and has a form closely related to the kernel density estimator:

$$\hat{\lambda}(x) = \sum_{i=1}^{N} k(x_i - x) \tag{7}$$

where k is a smoothing kernel, that is, any bounded function integrating to 1. Early work in this area focused on edge-corrections and methods for choosing the bandwidth [9, 5, 6]. Connections with RKHS have been considered by, for example, [4] who use a maximum penalized likelihood approach based on Hilbert spaces to estimate the intensity of a Poisson process. There is long literature on maximum penalized likelihood approaches to density estimation, which also contain interesting connections with RKHS, e.g. [25].

Much recent work on estimating intensities for point processes has focused on Bayesian approaches to modeling Cox processes. The log Gaussian Cox Process [20] and related parameterizations of Cox (doubly stochastic) Poisson processes in terms of Gaussian processes have been proposed, along with Monte Carlo [1, 10, 26], Laplace approximate [14, 8, 12] and variational [18, 17] inference schemes.

3 Proposed Method and Kernel Transformation

Let S be a compact domain of observations, e.g. the interval $[0, T]$ for a time series dataset observed between times 0 and T. Let $k : S \times S \to \mathbb{R}$ be a continuous positive definite kernel, and $\mathcal{H}_k$ its corresponding RKHS of functions $f : S \to \mathbb{R}$. We wish to parameterize the intensity of an inhomogeneous Poisson process using a function $f \in \mathcal{H}_k$. We define our intensity as:

$$\lambda(x) := af^2(x), \quad x \in S. \tag{8}$$

where $a > 0$ is a scale parameter and we have squared f to ensure that the intensity is non-negative on S. The exact rationale for including a will become clear later—the intuition is that it allows us to decouple the overall scale of the intensity (which depends on the units of the problem, e.g. number of points per hour versus number of points per year) from the penalty on the complexity of f which arises from the classical regularized Empirical Risk Minimization framework (and which should depend only on how complex, i.e. "wiggly" f is).

We use the inhomogeneous Poisson process likelihood from Eq. (4) to write the log-likelihood of a Poisson process corresponding to the observations $\{x_1, \ldots, x_N\}$, for $x_i \in S$, and intensity $\lambda(\cdot)$:

$$\ell(x_1, \ldots, x_N | \lambda) = \sum_{i=1}^{N} \log(\lambda(x_i)) - \int_S \lambda(x) dx \tag{9}$$

We will consider the problem of minimization of the penalized negative log likelihood, where the regularization term corresponds to the squared Hilbert space norm of f in parametrization Eq. (8):

$$\min_{f \in \mathcal{H}_k} \left\{ -\sum_{i=1}^{N} \log(af^2(x_i)) + a \int_S f^2(x) dx + \gamma \|f\|^2_{\mathcal{H}_k} \right\}. \tag{10}$$

This objective is akin to a classical regularized empirical risk minimization framework over RKHS: there is a term that depends on evaluations of f at the observed points $x_1, \ldots, x_N$ as well as a term corresponding to the RKHS norm. However, the representer theorem does not apply directly to Eq. (10), since there is also a term given by the L_2-norm of f, and so there is no guarantee that there is a solution of Eq. (10) that lies in span$\{k(x_i, \cdot)\}$. We will show that Eq. (10) fortunately still reduces to a finite-dimensional optimization problem corresponding to a different kernel function $\tilde{k}$ which we define below.

Using the Mercer expansion of k in Eq. (6), we can write the objective Eq. (10) as follows:

$$J[f] = -\sum_{i=1}^{N} \log(af^2(x_i)) + a\|f\|^2_{\mathcal{L}_2(S)} + \gamma \|f\|^2_{\mathcal{H}_k} \tag{11}$$

$$= -\sum_{i=1}^{N} \log(af^2(x_i)) + a \sum_{j=1}^{\infty} b_j^2 + \gamma \sum_{j=1}^{\infty} \frac{b_j^2}{\eta_j}. \tag{12}$$

The last two terms can now be merged together, giving

$$a\sum_{j=1}^{\infty} b_j^2 + \gamma \sum_{j=1}^{\infty} \frac{b_j^2}{e_j} = \sum_{j=1}^{\infty} b_j^2 \frac{a\eta_j + \gamma}{\eta_j} = \sum_{j=1}^{\infty} \frac{b_j^2}{\eta_j (a\eta_j + \gamma)^{-1}}. \tag{13}$$

Now, if we define kernel $\tilde{k}$ to be the kernel corresponding to the integral operator $T_{\tilde{k}} := T_k(aT_k + \gamma I)^{-1}$, i.e., $\tilde{k}$ is given by:

$$\tilde{k}(x, x') = \sum_{j=1}^{\infty} \frac{\eta_j}{a\eta_j + \gamma} e_j(x) e_j(x'), \qquad x, x' \in S, \tag{14}$$

we see that:

$$J[f] = -\sum_{i=1}^{N} \log(af^2(x_i)) + \|f\|^2_{\mathcal{H}_{\tilde{k}}}. \tag{15}$$

We are now ready to state the representer theorem in terms of kernel $\tilde{k}$.

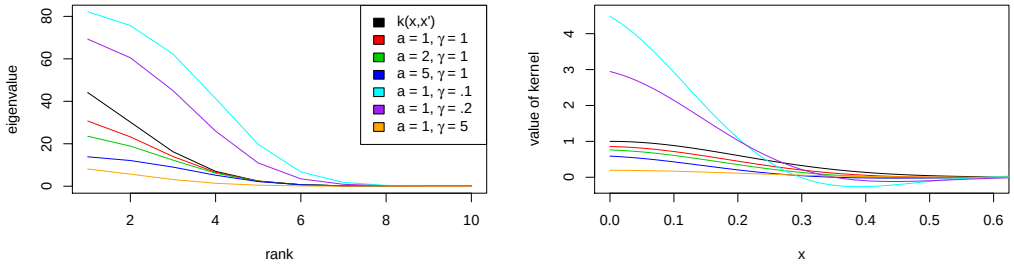
Theorem 1. *There exists a solution of* $\min_{f \in \mathcal{H}_k} \left\{ -\sum_{i=1}^{N} \log(af^2(x_i)) + a \int_S f^2(x)dx + \gamma \|f\|^2_{\mathcal{H}_k} \right\}$ *for observations* $x_1, \ldots, x_N$*, which takes the form* $f^*(\cdot) = \sum_{i=1}^{N} \alpha_i \tilde{k}(x_i, \cdot)$.

Proof. Since $\sum_j \frac{b_j^2}{\eta_j} < \infty$ if and only if $\sum_j \frac{b_j^2}{\eta_j (a\eta_j + \gamma)^{-1}} < \infty$, i.e. $f \in \mathcal{H}_k \iff f \in \mathcal{H}_{\tilde{k}}$, we have that the two spaces correspond to exactly the same set of functions. Optimization over $\mathcal{H}_k$ is therefore equivalent to optimization over $\mathcal{H}_{\tilde{k}}$. The proof now follows by applying the classical representer theorem in $\tilde{k}$ to the representation of the objective function in Eq. (15). For completeness, this is given in Appendix C. □

Remark 1. The notions of the inner product in $\mathcal{H}_k$ and $\mathcal{H}_{\tilde{k}}$ are different and thus in general $\text{span}\{k(x_i, \cdot)\} \neq \text{span}\{\tilde{k}(x_i, \cdot)\}$.

Remark 2. Notice that unlike in a standard ERM setting, $\gamma = 0$ does not recover the unpenalized risk, because γ appears in $\tilde{k}$. Notice further that the overall scale parameter a also appears in $\tilde{k}$. This is important in practice, because it allows us to decouple the scale of the intensity (which is controlled by a) from its complexity (which is controlled by γ).

Illustration. The eigenspectrum of $\tilde{k}$ where k is a squared exponential kernel is shown below for various settings of a and γ. Reminiscent of spectral filtering, we see that depending on the settings of a and γ, eigenvalues are shrunk or inflated as compared to $k(x, x')$ which is shown in black. On the right, the values of $k(0, x)$ are shown for the same set of kernels.



4 Computation of $\tilde{k}$

In this section, we consider first the case in which an explicit Mercer expansion is known, and then we consider the more commonly encountered situation in which we only have access to the parametric form of the kernel $k(x, x')$, so we must approximate $\tilde{k}$. We show experimentally that our approximation is very accurate by considering the Sobolev kernel, which can be expressed in both ways.

4.1 Explicit Mercer Expansion

We start by assuming that we have a kernel k with an explicit Mercer expansion, so we have eigenvectors $\{e_j(x)\}_{j\in J}$ and eigenvalues $\{\eta_j\}_{j\in J}$:

$$k(x,x') = \sum_{j\in J} \eta_j e_j(x) e_j(x'), \tag{16}$$

with an at most countable index set J. Given a and γ we can calculate:

$$\tilde{k}(x,x') = \sum_{j\in J} \frac{\eta_j}{a\eta_j + \gamma} e_j(x) e_j(x') \tag{17}$$

up to a desired precision as informed by the spectral decay in $\{\eta_j\}_{j\in J}$. We consider a kernel on the Sobolev space on $[0,1]$ with a periodic boundary condition, proposed by Wahba in [27, chapter 2] and recently used in [2]:

$$k(x,x') = 1 + \sum_{j=1}^{\infty} \frac{2\cos\left(2\pi j\left(x-x'\right)\right)}{(2\pi j)^{2s}} \tag{18}$$

where $s = 1, 2, \ldots$ denotes the order of the Sobolev space (larger s means existence of a larger number of square-integrable derivatives). We will return to this kernel in the experiments and use it to model point process data on periodic domains, including dihedral angles in protein structures. The Mercer expansion is given by:

$$k(x,x') = \sum_{j\in\mathbb{Z}} \eta_j e_j(x) e_j(x') \tag{19}$$

where the eigenfunctions are $e_0(x) = 1$ and $e_j(x) = \sqrt{2}\cos(2\pi jx)$, $e_{-j}(x) = \sqrt{2}\sin(2\pi jx)$ for $j = \{1, 2, \ldots\}$ with the corresponding eigenvalues $\eta_0 = 1$, $\eta_j = \eta_{-j} = (2\pi j)^{-2s}$. Further details are in the Appendix in Section B.1. We derive:

$$\tilde{k}(x,x') = \frac{1}{1+c} + \sum_{j=1}^{\infty} \frac{2\cos\left(2\pi j\left(x-x'\right)\right)}{a + \gamma(2\pi j)^{2s}}. \tag{20}$$

We discuss a Mercer expansion of the squared exponential kernel in the Appendix in Section B.2 and extensions of the Mercer expansion to multiple dimensions using a tensor product formulation in the Appendix in Section B.4. Although not practical for large datasets, we can use the Mercer expansion with summing terms up to $j > 50$ (for which the error is less than 10^{-5}) to evaluate the further approximations where Mercer expansion is not available, which we develop next.

4.2 Numerical Approximation

We propose an approximation to $\tilde{k}$ given access only to a kernel k for which we do not have an explicit Mercer expansion with respect to Lebesgue measure. We only assume that we can form Gram matrices corresponding to k and calculate their eigenvectors and eigenvalues. As a side benefit, this representation will also enable scalable computations through Toeplitz / Kronecker algebra or primal reduced rank approximations.

Let us first consider the one-dimensional case and construct a uniform grid $\mathbf{u} = (u_1, \ldots, u_m)$. Then the integral kernel operator T_k can be approximated with the (scaled) kernel matrix $\frac{1}{m}K_{\mathbf{uu}} : \mathbb{R}^m \to \mathbb{R}^m$, where $[K_{\mathbf{uu}}]_{ij} = k(u_i, u_j)$, and thus $\tilde{K}_{\mathbf{uu}}$ is approximately $K_{\mathbf{uu}}\left(\frac{a}{m}K_{\mathbf{uu}} + \gamma I\right)^{-1}$. However, we are not primarily interested in evaluations of $\tilde{k}$ on this grid, but on the observations $x_1, \ldots, x_N$. Simply adding the observations into the kernel matrix is not an option however, as it changes the base measure with respect to which the integral kernel operator is to be computed (Lebesgue measure on $[0,T]$). Thus, we consider the relationship between the eigendecomposition of $K_{\mathbf{uu}}$ and the eigenvalues and eigenfunctions of the integral kernel operator T_k.

Let $\lambda_i^u, \mathbf{e}_i^u$ be the eigenvalue/eigenvector pairs of the matrix $K_{\mathbf{uu}}$, i.e., its eigendecomposition is given by $K_{\mathbf{uu}} = Q\Lambda Q^\top = \sum_{i=1}^m \lambda_i^u \mathbf{e}_i^u (\mathbf{e}_i^u)^\top$. Then the estimates of the eigenvalues/eigenfunctions

of the integral operator T_k are given by the Nyström method (see [23, Section 4.3] and references therein, especially [3]):

$$\hat{\eta}_i = \frac{1}{m}\lambda_i^u, \qquad \hat{e}_i(x) = \frac{\sqrt{m}}{\lambda_i^u} K_{x\mathbf{u}} \mathbf{e}_i^u, \tag{21}$$

with $K_{x\mathbf{u}} = [k(x, u_1), \ldots, k(x, u_m)]$, leading to:

$$\widehat{\tilde{k}}(x, x') = \sum_{i=1}^{m} \frac{\hat{\eta}_i}{a\hat{\eta}_i + \gamma} \hat{e}_i(x)\hat{e}_i(x') = \sum_{i=1}^{m} \frac{\frac{1}{m}\lambda_i^u}{\frac{a}{m}\lambda_i^u + \gamma} \cdot \frac{m}{(\lambda_i^u)^2} K_{x\mathbf{u}} \mathbf{e}_i^u (\mathbf{e}_i^u)^\top K_{\mathbf{u}x'} \tag{22}$$

$$= K_{x\mathbf{u}} \left\{ \sum_{i=1}^{m} \frac{1}{\left(\frac{a}{m}\lambda_i^u + \gamma\right)\lambda_i^u} \mathbf{e}_i^u (\mathbf{e}_i^u)^\top \right\} K_{\mathbf{u}x'}. \tag{23}$$

For an estimate of the whole matrix $\tilde{K}_{\mathbf{xx}}$ we thus have

$$\widehat{\tilde{K}}_{\mathbf{xx}} = K_{\mathbf{xu}} \left\{ \sum_{i=1}^{m} \frac{1}{\left(\frac{a}{m}\lambda_i^u + \gamma\right)\lambda_i^u} \mathbf{e}_i^u (\mathbf{e}_i^u)^\top \right\} K_{\mathbf{ux}} = K_{\mathbf{xu}} Q \left(\frac{a}{m}\Lambda^2 + \gamma\Lambda\right)^{-1} Q^\top K_{\mathbf{ux}}. \tag{24}$$

The above is reminiscent of the Nyström method [28] proposed for speeding up Gaussian process regression. A reduced rank representation for Eq. (24) is straightforward by considering only the top p eigenvalues/eigenvectors of $K_{\mathbf{uu}}$. Computational cost is thus $O(m^3 + N^2 m)$. Furthermore, a primal representation with the features corresponding to kernel $\tilde{k}$ is readily available and is given by

$$\tilde{\phi}(x) = \left(\frac{a}{m}\Lambda^2 + \gamma\Lambda\right)^{-1/2} Q^\top K_{\mathbf{u}x}, \tag{25}$$

which allows linear computational cost in the number N of observations.

For $D > 1$ dimensions, one can exploit Kronecker and Toeplitz algebra approaches. Assuming that the $K_{\mathbf{uu}}$ matrix corresponds to a Cartesian product structure of the one-dimensional grids of size m, one can write $K_{\mathbf{uu}} = K_1 \otimes K_2 \cdots \otimes K_D$. Thus, the eigenspectrum can be efficiently calculated by eigendecomposing each of the smaller $m \times m$ matrices $K_1, \ldots, K_D$ and then applying standard Kronecker algebra, thereby avoiding ever having to form the prohibitively large $m^D \times m^D$ matrix $K_{\mathbf{uu}}$. For regular grids and stationary kernels, each small matrix will be Toeplitz structured, yielding further efficiency gains [29]. The resulting approach therefore scales linearly in dimension D.

We compared the exact calculation of $\tilde{K}_{\mathbf{uu}}$ with $s = 1$, $a = 10$, and $\gamma = .5$ to our approximate calculation. For illustration we tried a coarse grid of size 10 on the unit interval (left) to a finer grid of size 100. The RMSE was 2E-3 for the coarse grid and 1.6E-5 for the fine grid, as shown in the Appendix in Figure A4. In the same figure we compared the exact calculation of $\tilde{K}_{\mathbf{xx}}$ with $s = 1$, $a = 10$, and $\gamma = .5$ to our Nyström-based approximation, where $x_1, \ldots, x_{400} \sim \text{Beta}(.5, .5)$ distribution. The RMSE was 0.98E-3. A low-rank approximation using only the top 5 eigenvalues gives the RMSE of 1.6E-2.

5 Inference

The penalized risk can be readily minimized with gradient descent. Let $\alpha = [\alpha_1, \ldots, \alpha_N]^\top$ and $\tilde{K}$ be the Gram matrix corresponding to $\tilde{k}$ such that $\tilde{K}_{ij} = \tilde{k}(x_i, x_j)$. Then $[f(x_1), \ldots, f(x_N)]^\top = \tilde{K}\alpha$ and the gradient of the objective function J from (15) is calculated as follows, where $\log(\cdot)$ is understood to be applied element-wise to a vector.

$$\begin{aligned}
\nabla_\alpha J &= -\nabla_\alpha \sum_i \log(af^2(x_i)) + \gamma \nabla_\alpha \|f\|^2_{\mathcal{H}_{\tilde{k}}} = -\nabla_\alpha \sum_i \log(a(\sum_j \tilde{k}_{ij}\alpha_j)^2) + \gamma \nabla_\alpha \alpha^\top \tilde{K}\alpha \\
&= -\sum_i \frac{2a(\sum_j \tilde{k}_{ij}\alpha_j)\nabla_\alpha \sum_j \tilde{k}_{ij}\alpha_j}{a(\sum_j \tilde{k}_{ij}\alpha_j)^2} + 2\gamma\tilde{K}\alpha = -\sum_i \frac{2\tilde{K}_{\cdot i}}{\sum_j \tilde{k}_{ij}\alpha_j} + 2\gamma\tilde{K}\alpha \\
&= -2\sum_i (\tilde{K}_{\cdot i}./(\tilde{K}\alpha)) + 2\gamma\tilde{K}\alpha
\end{aligned}$$

where $./$ denotes element-wise division. We use L-BFGS-B to maximize R. Computing $\tilde{K}$ requires $\mathcal{O}(N^2)$ time and memory, and each gradient and likelihood computation requires matrix-vector

multiplications which are also $\mathcal{O}(N^2)$. Overall, the running time is $\mathcal{O}(qN^2)$ for q iterations of the gradient descent method, where q is usually very small in practice.

6 Naïve RKHS model

In this section, we compare the proposed approach, which uses the representer theorem in the transformed kernel $\tilde{k}$, to the naïve one, where a solution to Eq. (10) of the form $f(\cdot) = \sum_{j=1}^{N} \alpha_j k(x_j, \cdot)$ is sought even though the representer theorem in k need not hold. Despite being theoretically suboptimal, this is a natural model to consider, and it might perform well in practice. The corresponding optimization problem is:

$$\min_{f \in \text{span}\{k(x_i,\cdot)\}} \left\{ -\sum_{i=1}^{N} \log(af^2(x_i)) + a\int_S f^2(x)dx + \gamma \|f\|^2_{\mathcal{H}_k} \right\}. \tag{26}$$

While the first and the last term are straightforward to calculate for any $f(\cdot) = \sum_j \alpha_j k(x_j, \cdot)$, $\int_S f^2(x)dx$ needs to be estimated. As before, we construct a uniform grid of fineness h, $\mathbf{u} = (u_1, \ldots, u_n)$ covering the domain. Then

$$\int_S f^2(u)du = \int_S \left(\alpha^\top K_{\mathbf{x}u}\right)^2 du = \alpha^\top \left\{ \int_S K_{\mathbf{x}u} K_{u\mathbf{x}} du \right\} \alpha \approx h\alpha^\top K_{\mathbf{xu}} K_{\mathbf{ux}} \alpha, \tag{27}$$

and the optimization problem reads:

$$\min_{\alpha \in \mathbb{R}^N} \left\{ -\sum_{i=1}^{N} \log(a(\alpha^\top K_{\mathbf{x}x_i})^2) + \alpha^\top \left(ahK_{\mathbf{xu}} K_{\mathbf{ux}} + \gamma K_{\mathbf{xx}}\right) \alpha \right\}. \tag{28}$$

We carried out a small simulation study using simulated intensities drawn from $\mathcal{H}_k$ where k was the squared exponential (SE) kernel in order to compare the performance of the correct model using $\tilde{k}$ as previously developed to the naïve model discussed above. In both cases we used crossvalidation to tune the hyperparameters on the same synthetic dataset. For testing, we calculated the RMSE of each method in reconstructing the true underlying intensity. In 97 out of 100 repetitions, the correct model had a smaller average test log-likelihood than the incorrect model (paired t-test p-value < 0.001). We report further comparisons in the next section.

7 Experiments

Synthetic Example. We generated a synthetic intensity using the Mercer expansion of a SE kernel with lengthscale 0.5, producing a random linear combination of 64 basis functions, weighted with iid draws $\alpha \sim \mathcal{N}(0, 1)$. In Figure 1 we compare ground truth to estimates made with: our RKHS method with SE kernel, the naïve RKHS approach with SE kernel, a log-Gaussian Cox process method [13], and classical kernel smoothing with bandwidth selected by crossvalidation (`bw.ucv` in R). The RMSE of our method was 34 which compared to 75 for kernel intensity estimation, 75 for the Cox process, and 86 for the naïve RKHS approach.

Environmental datasets. Next we demonstrate our method on a collection of two-dimensional environmental datasets giving the locations of trees. The datasets were obtained from the R package `spatstat`. For illustration, we visualize the locations and intensity estimates for black oak trees in Lansing, Michigan in Figure 2, where we calculated the intensity using various approaches: our proposed method with squared exponential kernel, our naïve RKHS method with squared exponential kernel, and classical kernel intensity estimation, with a crossvalidated bandwidth and edge correction. We used cross-validation to tune our methods and we provide cross-validated likelihoods in Table 1.

Dihedral angles as point process on a torus. Finally, we consider a novel application of Poisson process estimation, suited to our periodic Sobolev kernel. The tensor product construction in two dimensions is appropriate for data observed on a torus. An example from protein bioinformatics is shown in Figure A3 using data included with the R package `MDplot`, visualizing the dihedral torsion angles $[\psi, \phi]$ of amino acids in proteins [21, 19]. Classically, datasets of observed angle pairs have been binned using two-dimensional histograms. We propose to treat a set of observed angles as an inhomogeneous Poisson process, enabling intensity estimation as shown.

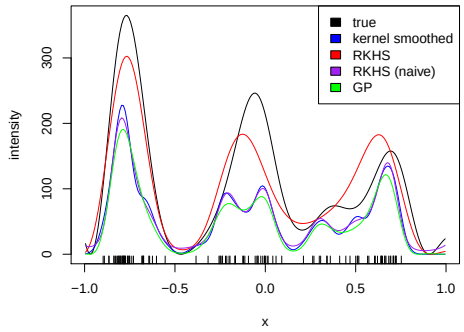
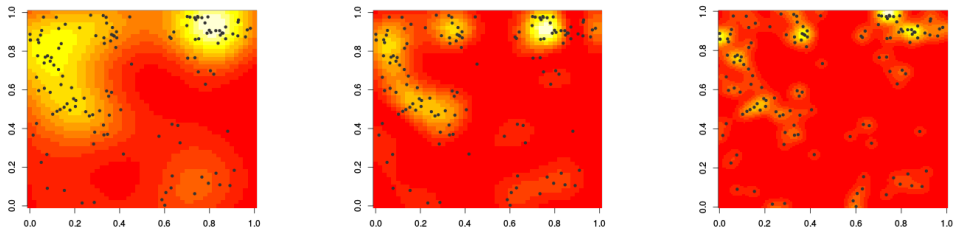


Figure 1: A synthetic dataset, comparing our RKHS method, the naïve model, a Cox process model, and kernel smoothing. The RMSE of the model using $\tilde{k}$ was significantly better than the competing approaches, as described in text.

Dataset	Kernel intensity estimation	Naïve approach	Our approach with $\tilde{k}$
Lansing: Black oak (n = 135)	-236	-233	-232
Hickory (n = 703)	**-1759**	-1748	-1752
Maple (n = 514)	**-1243**	**-1237**	-1227
Misc (n = 105)	-172	-173	-171
Red oak (n = 346)	**-722**	-727	-727
White oak (n = 448)	-994	-990	**-996**
Spruces in Saxonia (n = 134)	-213	-213	-213
Waka national park (n = 504)	-1136	**-1139**	**-1139**
New Zealand (n = 86)	-117	**-118**	**-118**
Swedish pines (n = 71)	-91	-91	-91

Table 1: Various datasets from the R package `spatstat` give the locations of trees. We compared kernel intensity estimation with bandwidth selected by crossvalidation to the naïve RKHS approach and our proposed approach. Bolded results were significant at $p < 0.01$ with paired two-sample tests.



(a) SE kernel: RKHS (b) SE kernel: naïve (c) Kernel smoothing

Figure 2: Location of black oak trees in Lansing michigan smoothed with various approaches. As shown in Table 1, none of the approaches were significantly better than the others on this dataset.

8 Conclusion

We presented a novel approach to inhomogeneous Poisson process intensity estimation using a Representor Theorem formulation in an appropriately transformed RKHS, providing a computationally tractable approach giving comparable performance to existing methods on low dimensionality datasets. In future work, we will consider marked Poisson processes and other more complex point process models, as well as Bayesian extensions akin to Cox process modeling.

References

[1] Adams, R. P., Murray, I., and MacKay, D. J. Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 9–16. ACM, 2009.

[2] Bach, F. On the equivalence between quadrature rules and random features. *arXiv:1502.06800*, 2015.

[3] Baker, C. *The Numerical Treatment of Integral Equations*. Monographs on Numerical Analysis Series. Oxford : Clarendon Press, 1977. ISBN 9780198534068.

[4] Bartoszynski, R., Brown, B. W., McBride, C. M., and Thompson, J. R. Some nonparametric techniques for estimating the intensity function of a cancer related nonstationary poisson process. *The Annals of Statistics*, pages 1050–1060, 1981.

[5] Berman, M. and Diggle, P. Estimating weighted integrals of the second-order intensity of a spatial point process. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 81–92, 1989.

[6] Brooks, M. M. and Marron, J. S. Asymptotic optimality of the least-squares cross-validation bandwidth for kernel estimates of intensity functions. *Stochastic Processes and their Applications*, 38(1):157–165, 1991.

[7] Cressie, N. and Wikle, C. *Statistics for spatio-temporal data*, volume 465. Wiley, 2011.

[8] Cunningham, J. P., Shenoy, K. V., and Sahani, M. Fast gaussian process methods for point process intensity estimation. In *ICML*, pages 192–199. ACM, 2008.

[9] Diggle, P. A kernel method for smoothing point process data. *Applied Statistics*, pages 138–147, 1985.

[10] Diggle, P. J., Moraga, P., Rowlingson, B., Taylor, B. M., et al. Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563, 2013.

[11] Fasshauer, G. E. and McCourt, M. J. Stable evaluation of gaussian radial basis function interpolants. *SIAM Journal on Scientific Computing*, 34(2):A737–A762, 2012.

[12] Flaxman, S. R., Wilson, A. G., Neill, D. B., Nickisch, H., and Smola, A. J. Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods. *International Conference on Machine Learning*, 2015.

[13] Flaxman, S. R., Neill, D. B., and Smola, A. J. Gaussian processes for independence tests with non-iid data in causal inference. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2016.

[14] Illian, J. B., Sørbye, S. H., Rue, H., et al. A toolbox for fitting complex spatial point process models using integrated nested laplace approximation (inla). *The Annals of Applied Statistics*, 6(4):1499–1530, 2012.

[15] Kimeldorf, G. and Wahba, G. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82 – 95, 1971. ISSN 0022-247X.

[16] Kingman, J. F. C. *Poisson processes*, volume 3 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, 1993. ISBN 0-19-853693-3. Oxford Science Publications.

[17] Kom Samo, Y.-L. and Roberts, S. Scalable nonparametric bayesian inference on point processes with gaussian processes. In *ICML*, pages 2227–2236, 2015.

[18] Lloyd, C., Gunter, T., Osborne, M., and Roberts, S. Variational inference for gaussian process modulated poisson processes. In *ICML*, pages 1814–1822, 2015.

[19] Mardia, K. V. Statistical approaches to three key challenges in protein structural bioinformatics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(3):487–514, 2013.

[20] Møller, J., Syversveen, A., and Waagepetersen, R. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998.

[21] Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *Journal of molecular biology*, 7(1):95–99, 1963.

[22] Ramlau-Hansen, H. Smoothing counting process intensities by means of kernel functions. *Ann. Statist.*, 11 (2):453–466, 06 1983. doi: 10.1214/aos/1176346152.

[23] Rasmussen, C. E. and Williams, C. K. Gaussian processes for machine learning, 2006.

[24] Schölkopf, B. and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization and beyond*. the MIT Press, 2002.

[25] Silverman, B. W. On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.*, 10(3):795–810, 09 1982. doi: 10.1214/aos/1176345872.

[26] Teh, Y. W. and Rao, V. Gaussian process modulated renewal processes. In *Advances in Neural Information Processing Systems*, pages 2474–2482, 2011.

[27] Wahba, G. *Spline models for observational data*, volume 59. Siam, 1990.

[28] Williams, C. and Seeger, M. Using the nyström method to speed up kernel machines. In *Proceedings of the 14th Annual Conference on Neural Information Processing Systems*, pages 682–688, 2001.

[29] Wilson, A. G., Dann, C., and Nickisch, H. Thoughts on massively scalable gaussian processes. *arXiv:1511.01870*, 2015.

[30] Zhu, H., Williams, C. K., Rohwer, R., and Morciniec, M. Gaussian regression and optimal finite dimensional linear models. 1997.

A Supplementary results

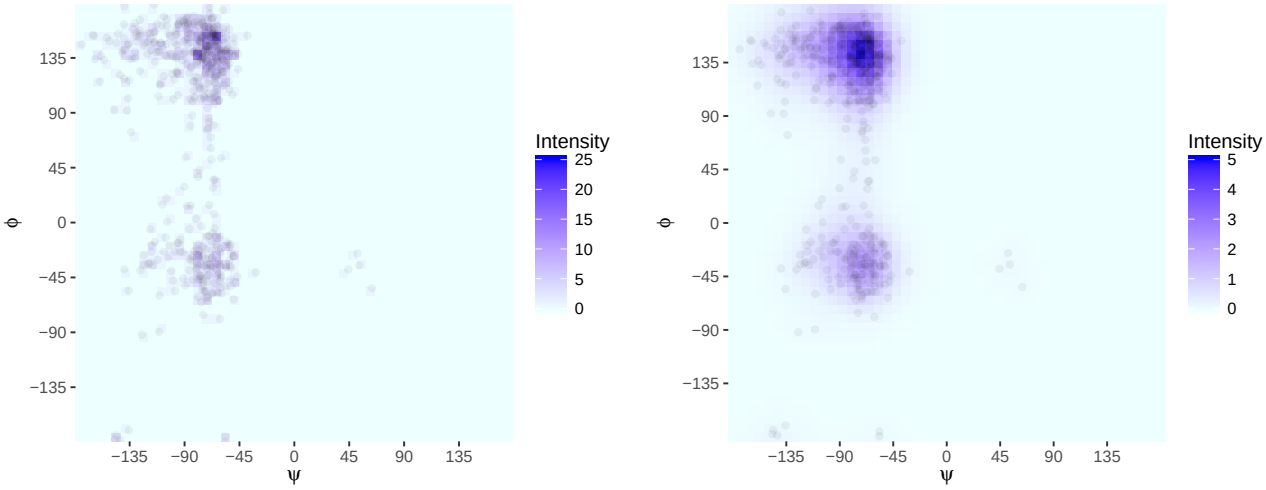


Figure A3: Standard Ramachandran plot (left) based on a two-dimensional histogram versus our proposed Ramachandran plot based on an intensity estimate with a two-dimensional Sobolev kernel

B Kernels with Explicit Mercer Expansions

B.1 Sobolev space on $[0,1]$ with a periodic boundary condition

We consider domain $S = [0,1]$. The kernel is given by:

$$\begin{aligned} k(x,y) &= 1 + \sum_{m=1}^{\infty} \frac{2\cos\left(2\pi m\left(x-y\right)\right)}{(2\pi m)^{2s}} \\ &= 1 + \sum_{m=1}^{\infty} \frac{2}{(2\pi m)^{2s}} \left[\cos\left(2\pi mx\right)\cos\left(2\pi my\right) + \sin\left(2\pi mx\right)\sin\left(2\pi my\right)\right], \\ &= 1 + \frac{(-1)^{s-1}}{(2s)!} B_{2s}(\{x-y\}), \end{aligned}$$

where $s = 1, 2, \ldots$ denotes the order of the Sobolev space and $B_{2s}(\{x-y\})$ is the Bernoulli polynomial of degree $2s$ applied to the fractional part of $x-y$. The corresponding RKHS is the space of functions on $[0,1]$ with absolutely continuous $f, f', \ldots, f^{(s-1)}$ and square integrable $f^{(s)}$ satisfying a periodic boundary condition $f^{(l)}(0) = f^{(l)}(1)$, $l = 0, \ldots, s-1$. For more details, see [27, Chapter 2] Bernoulli polynomials admit a simple form for low degrees. In particular,

$$\begin{aligned} B_2(t) &= t^2 - t + \frac{1}{6}, \\ B_4(t) &= t^4 - 2t^3 + t^2 - \frac{1}{30}, \\ B_6(t) &= t^6 - 3t^5 + \frac{5}{2}t^4 - \frac{1}{2}t^2 + \frac{1}{42}. \end{aligned}$$

If we consider the Mercer expansion where the underlying measure ρ is uniform on $[0,1]$: $d\rho(x) = dx$, we have

$$\begin{aligned}
\int_0^1 2\cos(2\pi mx)\sin(2\pi m'x)\,dx &= 0 \\
\int_0^1 2\cos(2\pi mx)\cos(2\pi m'x)\,dx &= \delta(m-m') \\
\int_0^1 2\sin(2\pi mx)\sin(2\pi m'x)\,dx &= \delta(m-m').
\end{aligned}$$

Thus, the desired Mercer expansion $k(x,y) = \sum_{m\in\mathbb{Z}} \eta_m e_m(x) e_m(y)$ has eigenfunctions $e_0(x) = 1$ and for $m = \{1,2,\ldots\}$, $e_m(x) = \sqrt{2}\cos(2\pi mx)$, $e_{-m}(x) = \sqrt{2}\sin(2\pi mx)$ and corresponding eigenvalues $\eta_0 = 1$, $\eta_m = \eta_{-m} = (2\pi m)^{-2s}$.

- $\tilde{k}(x,y)$ is the kernel of $T_k(T_k + cI)^{-1}$ and has form

$$\begin{aligned}
\tilde{k}(x,y) &= \sum_{m\in\mathbb{Z}} \frac{\eta_m}{\eta_m + c} e_m(x) e_m(y) \\
&= \frac{1}{1+c} + \sum_{m=1}^{\infty} \frac{2\cos(2\pi m(x-y))}{1 + c(2\pi m)^{2s}}
\end{aligned}$$

- To compute $\int_0^1 f^2(x)dx$ for $f = \sum_i \alpha_i \tilde{k}(\cdot, x_i)$ we have

$$\begin{aligned}
\int_0^1 f^2(x)dx &= \sum_{i,j} \alpha_i\alpha_j \int_0^1 \tilde{k}(x_i,u)\tilde{k}(u,x_j)du \\
&= \sum_{i,j} \alpha_i\alpha_j \sum_{m,m'} \frac{\eta_m\eta'_m}{(\eta_m + c)(\eta_{m'} + c)} e_m(x_i) e_{m'}(x_j) \int_0^1 e_m(u) e_{m'}(u)du \\
&= \sum_{i,j} \alpha_i\alpha_j \sum_{m} \frac{\eta_m^2}{(\eta_m + c)^2} e_m(x_i) e_m(x_j) \\
&= \alpha^\top \tilde{R}\alpha,
\end{aligned}$$

  where kernel matrix $\tilde{R}$ is computed using kernel $\tilde{r}$ of $T_k^2(T_k + cI)^{-2}$, i.e.

$$\begin{aligned}
\tilde{r}(x,y) &= \sum_{m\in\mathbb{Z}} \frac{\eta_m^2}{(\eta_m + c)^2} e_m(x) e_m(y) \\
&= \frac{1}{(1+c)^2} + \sum_{m=1}^{\infty} \frac{2\cos(2\pi m(x-y))}{(1 + c(2\pi m)^{2s})^2}.
\end{aligned}$$

- To generate a function $f \in \mathcal{H}_k$ of unit norm $\|f\|_{\mathcal{H}_k} = 1$, one takes

$$f(x) = a_0 + \sqrt{2}\sum_{m=1}^{M} \left(a_m \cos(2\pi mx) + a_{-m}\sin(2\pi mx)\right), \tag{29}$$

  for which the norm is given by

$$\|f\|_{\mathcal{H}_k}^2 = a_0^2 + \sum_{m=1}^{M} (a_m^2 + a_{-m}^2)(2\pi m)^{2s}. \tag{30}$$

  Thus we can simply generate $\mathbf{z} = (z_{-M}, \ldots, z_0, \ldots, z_M) \sim \mathcal{N}(0, I_{2M+1})$, set $\tilde{\mathbf{z}} = \mathbf{z}/\|\mathbf{z}\|$ and then $a_0 = \tilde{z}_0$, $a_m = \tilde{z}_m(2\pi|m|)^{-s}$, for $m \neq 0$.

B.2 Squared exponential kernel

A Mercer expansion for the squared exponential kernel was proposed in [30] and refined in [11]. However, this expansion is with respect to a Gaussian measure on $\mathbb{R}$, i.e., it consists of eigenfunctions which form an orthonormal set in $\mathcal{L}^2(\mathbb{R}, \nu)$ where $\nu = \mathcal{N}(0, \ell^2 I)$. The formalism can therefore be used to estimate Poisson intensity functions with respect to such Gaussian measure. In the classical

framework, where the intensity is with respect to a Lebesgue measure, numerical approximations of Mercer expansion, as described in Section 4.2 are needed. Following the exposition in [23, section 4.3.1] and the relevant errata[1] we parameterize the kernel as:

$$k(x, x') = \exp(-\frac{\|x - x'\|^2}{2\sigma^2}) \tag{31}$$

The Mercer expansion with respect to $\nu = \mathcal{N}(0, \ell^2 I)$ then has the following eigenvalues:

$$\eta_i = \sqrt{\frac{2a}{A}} B^i \tag{32}$$

And eigenfunctions:

$$e_i(x) = \frac{1}{\sqrt{\sqrt{a/c}\, 2^i i!}} \exp(-(c-a)x^2) H_i(\sqrt{2c}x) \tag{33}$$

where H_i is the i-th order (physicist's) Hermite polynomial, $a = \frac{1}{4\sigma^2}$, $b = \frac{1}{2\ell^2}$, $c = \sqrt{a^2 + 2ab}$, $A = a + b + c$, and $B = b/A$. Thus we have the following eigenvalues for $\tilde{k}$:

$$\widetilde{\eta_i} = \frac{\eta_i}{a\eta_i + \gamma} = \frac{1}{a + \gamma\sqrt{\frac{A}{2a}} B^{-i}} \tag{34}$$

while the eigenfunctions remain the same.

B.3 Brownian Bridge kernel

This is the kernel

$$k(x, y) = \min(x, y) - xy = \sum_{m=1}^{\infty} \frac{2\sin(\pi m x)\sin(\pi m y)}{\pi^2 m^2},$$

with

$$\eta_m = \frac{1}{\pi^2 m^2},\ e_m(x) = \sqrt{2}\sin(\pi m x), \quad m = 1, 2, \ldots. \tag{35}$$

Thus

$$\begin{aligned} \tilde{k}(x, y) &= \sum_{m=1}^{\infty} \frac{\eta_m}{\eta_m + c} e_m(x) e_m(y) \\ &= \sum_{m=1}^{\infty} \frac{2\sin(\pi m x)\sin(\pi m y)}{1 + c\pi^2 m^2} \end{aligned}$$

B.4 Extending the Mercer expansion to multiple dimensions

The extension of any kernel to higher dimensions can be constructed by considering tensor product spaces: $\mathcal{H}_{k_1 \otimes k_2}$ (where k_1 and k_2 could potentially be different kernels with different hyperparameters). If k_1 has eigenvalues η_i and eigenfunctions e_i and k_2 has eigenvalues δ_j and eigenfunctions f_j, then the eigenvalues of the product space are then given by the Cartesian product $\eta_i \delta_j, \forall i, j$, and similarly the eigenfunctions are given by $e_i(x) f_j(y)$. Our regularized kernel has the following Mercer expansion:

$$\widetilde{k_1 \otimes k_2}((x, y), (x', y')) = \sum_{ij} \frac{\eta_i \delta_j}{a\eta_i \delta_j + \gamma} e_i(x) e_i(x') f_j(y) f_j(y') \tag{36}$$

Notice that $\widetilde{k_1 \otimes k_2}$ is the kernel corresponding to the integral operator $(T_{k_1} \otimes T_{k_2})(aT_{k_1} \otimes T_{k_2} + \gamma I)^{-1}$ which is different than $\tilde{k}_1 \otimes \tilde{k}_2$.

[1] http://www.gaussianprocess.org/gpml/errata.html

C Proof of the Representer Theorem

We decompose $f \in \mathcal{H}_{\tilde{k}}$ as the sum of two functions:

$$f(\cdot) = \sum_{j=1}^{N} \alpha_j \tilde{k}(x_j, \cdot) + v \tag{37}$$

where v is orthogonal to the span of $\{\tilde{k}(x_j, \cdot)\}_j$. We prove that the first term in the objective $J[f]$ given in Eq. (15), $-\sum_{i=1}^{N} \log(af^2(x_i))$, is independent of v. It depends on f only through the evaluations $f(x_i)$ for all i. Using the reproducing property we have:

$$f(x_i) = \langle f, \tilde{k}(x_i, \cdot)\rangle = \sum_j \alpha_j \tilde{k}(x_j, x_i) + \langle v, \tilde{k}(x_i, \cdot)\rangle = \sum_j \alpha_j \tilde{k}(x_j, x_i) \tag{38}$$

where the last step is by orthogonality. Next we substitute into the regularization term:

$$\gamma \| \sum_j \alpha_j \tilde{k}(x_j, \cdot) + v\|^2_{\mathcal{H}_{\tilde{k}}} = \gamma \| \sum_j \alpha_j \tilde{k}(x_j, \cdot)\|^2_{\mathcal{H}_{\tilde{k}}} + \|v\|^2_{\mathcal{H}_{\tilde{k}}} \geq \gamma \| \sum_j \alpha_j \tilde{k}(x_j, \cdot)\|^2_{\mathcal{H}_{\tilde{k}}}. \tag{39}$$

Thus, the choice of v has no effect on the first term in $J[f]$ and a non-zero v can only increase the second term $\|f\|^2_{\mathcal{H}_{\tilde{k}}}$, so we conclude that $v = 0$ and that $f^* = \sum_{j=1}^{N} \alpha_j \tilde{k}(x_j, \cdot)$ is the minimizer.

D Numerical evaluation of kernel approximations

Here we present an evaluation of the numerical approximation to $\tilde{k}$ described in 4.2 on the case of the Sobolev kernel where Mercer expansion is also available so that truncated Mercer expansion representation of $\tilde{k}$ can be treated as a ground truth. As Figure A4, demonstrates, good approximation is possible with a fairly coarse grid $\mathbf{u} = (u_1, \ldots, u_m)$ as well as with a low-rank approximation.

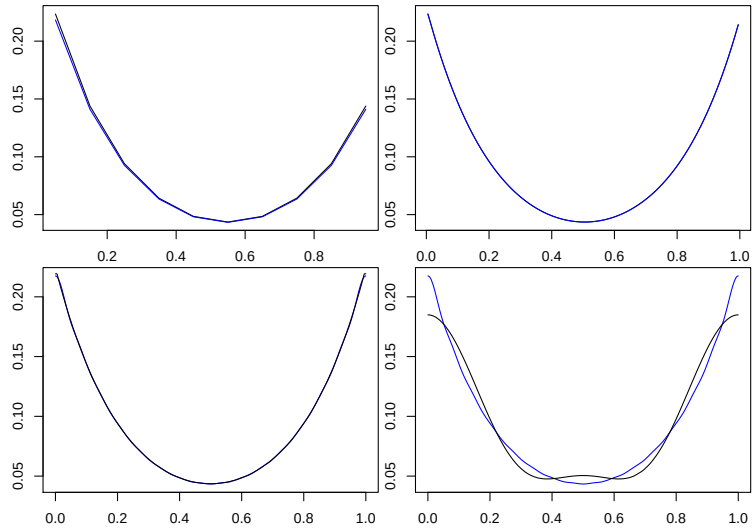


Figure A4: We compared the exact calculation of $\tilde{K}_{\mathbf{uu}}$ with $s = 1$, $a = 10$, and $\gamma = .5$ to our approximate calculation. For illustration we tried a coarse grid of size 10 on the unit interval (top left) to a finer grid of size 100 (top right). The RMSE was 2E-3 for the coarse grid and 1.6E-5 for the fine grid. We compare the exact calculation of $\tilde{K}_{\mathbf{xx}}$ with $s = 1$, $a = 10$, and $\gamma = .5$ to our Nyström-based approximation, where $x_1, \ldots, x_{400} \sim$ Beta$(.5, .5)$ distribution (bottom left). The RMSE was 0.98E-3. A low-rank approximation using only the top 5 eigenvalues gives the RMSE of 1.6E-2 (bottom right).