Journal of Machine Learning Research 24 (2023) 1-72 Submitted 11/21; Revised 11/22; Published 1/23

Learning Partial Differential Equations in Reproducing Kernel Hilbert Spaces

George Stepaniants GSTEPAN@MIT.EDU
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge, MA 02139

Editor: Jean-Philippe Vert

Abstract

We propose a new data-driven approach for learning the fundamental solutions (Green's functions) of various linear partial differential equations (PDEs) given sample pairs of input-output functions. Building off the theory of functional linear regression (FLR), we estimate the best-fit Green's function and bias term of the fundamental solution in a reproducing kernel Hilbert space (RKHS) which allows us to regularize their smoothness and impose various structural constraints. We derive a general representer theorem for operator RKHSs to approximate the original infinite-dimensional regression problem by a finite-dimensional one, reducing the search space to a parametric class of Green's functions. In order to study the prediction error of our Green's function estimator, we extend prior results on FLR with scalar outputs to the case with functional outputs. Finally, we demonstrate our method on several linear PDEs including the Poisson, Helmholtz, Schrödinger, Fokker–Planck, and heat equation. We highlight its robustness to noise as well as its ability to generalize to new data with varying degrees of smoothness and mesh discretization without any additional training.

Keywords: Green's functions, partial differential equations, reproducing kernel Hilbert spaces, functional linear regression, simultaneous diagonalization

1. Introduction

The rapid development of data-driven scientific discovery holds the promise of new and faster methods to analyze, understand, and predict various complex phenomena whose physical laws are still beyond our grasp. Of central interest to this development is the ability to solve efficiently a broad range of differential equations, and more precisely partial differential equations (PDEs) which still largely require advanced numerical techniques tailored for specific problems.

In this paper, we study what is certainly one of the most inspiring outcomes of this program: solving PDEs from input-output data. Let $u(x,t)$ denote the state at time t and location x of a system evolving according to a PDE such as the dynamics of a swarm of particles or the propagation of a wave through a complex medium. The goal of this paper is to predict $u(x,t)$ under initial conditions $u(x,0)$, under boundary conditions $u(x,t) = b(x,t)$ for x on the boundary of a domain, or under an external forcing $f(x,t)$ that represents the ambient conditions of an evolving system. In other words, we propose to learn an operator that maps these inputs $\{u(x,0), b(x,t), f(x,t)\}$ to output solutions $u(x,t)$ from input-output data. While such operators can be, and often are, nonlinear, we focus on linear operators in this paper. Not only is this a natural first step for this program but linear approximations to nonlinear phenomena are often useful and are, in general, more robust to model misspecification.

The study of learning input-output maps where the inputs or the outputs (or both) are functions traditionally falls under the umbrella of *functional data analysis* introduced in the seminal paper of Ramsay and Dalzell (1991) with much of the theory and practical applications to physical, biological, and economic data reviewed in the monographs by Ramsay (2004); Ramsay and Silverman (2007).

As a concrete driving example, consider a system whose state u is the solution of a PDE on a compact domain $D \subset \mathbb{R}^d$,

$$\begin{aligned} \mathcal{P}u &= f \text{ on } D \\ \mathcal{B}u &= 0 \text{ on } \partial D\,. \end{aligned} \tag{1}$$

Here $\mathcal{P}$ is a differential operator and $\mathcal{B}$ encodes the boundary conditions. Given input f, and knowledge of the operators $\mathcal{P}, \mathcal{B}$, solving the PDE (1) requires sophisticated numerical methods such as finite differences, spectral decompositions, or finite elements (Zienkiewicz et al., 1977; Trefethen, 2000; LeVeque, 2007). These methods can, and will, be used to create input-output pairs $\{(f_i, u_i), i = 1, \ldots, n\}$ on the domain D usually at some fixed level of discretization.

Building on this observation, we propose a natural goal: to learn a *surrogate model* $\mathcal{T} : f \mapsto u$ which takes input $f \in L^2(D_\mathcal{X})$ to output $u \in L^2(D_\mathcal{Y})$. Solving this supervised learning problem is of paramount importance to unlock the potential of data-driven methods and understand real-world systems under new and unseen conditions. Surrogate modeling combines elements of numerical analysis, statistics, and machine learning to efficiently learn solution maps $\mathcal{T}$ that are both physically relevant and fast to evaluate.

Two approaches have been predominantly used in the study of surrogate models. The first approach discretizes the functions $f \in \mathbb{R}^{m_y}, u \in \mathbb{R}^{m_x}$ and regresses a map (i.e. neural network) $F : \mathbb{R}^{m_y} \to \mathbb{R}^{m_x}$ on the data. This methodology has been successfully applied to surrogate modeling of flow fields, computed tomography, and porous media (Guo et al., 2016; Adler and Öktem, 2017; Zhu and Zabaras, 2018; Bhatnagar et al., 2019). However, this approach is not robust to mesh-refinement which is a serious issue as it becomes data-hungry with increasing mesh sizes m_x, m_y (Bhattacharya et al., 2021). Furthermore, evaluating such a model on more finely sampled input-output data requires an entire retraining of the architecture. The second predominant approach to surrogate modeling attempts to learn the solution u of the PDE by parameterizing u itself as a map $F_\theta : D_\mathcal{Y} \to \mathbb{R}$ where θ is a set of model parameters. For example, Chen et al. (2021b) optimize the solution map $F_\theta : D_\mathcal{Y} \to \mathbb{R}$ in a reproducing kernel Hilbert space (RKHS) which can be viewed as the maximum a posteriori estimator of a Gaussian process. Using RKHSs to estimate solutions of PDEs is a large area of research stemming from the foundational work of Fasshauer on mesh-free approximation methods (Fasshauer, 2007, Chapter 38) with many recent extensions to fractional PDEs and integro-differential equations (Arqub, 2018, 2019; Al-Smadi and Arqub, 2019; Arqub and Al-Smadi, 2020). A review of kernel-based numerical methods for PDEs can be found in Fornberg and Flyer (2015) and the classical text of Saitoh and Sawano (2016, Chapters 5 & 6) contains numerous examples of kernel methods specialized for solving forward and inverse problems for ODEs and PDEs. As an alternative to kernel methods, recent approaches have proposed to learn the solutions map $F_\theta : D_\mathcal{Y} \to \mathbb{R}$ of a PDE as a neural network where θ are the network weights. This idea has found applications in many applied problems such as the study of electrical impedance tomography, reaction-diffusion systems, and wave propagation (E and Yu, 2018; Raissi et al., 2019;

Bar and Sochen, 2019) along with grid-independent generative modeling of images (Dupont et al., 2022). In general, this second approach of learning PDE solution maps $F_\theta : D_\mathcal{Y} \to \mathbb{R}$ is indeed independent of mesh discretization. However, its dependence on the initial conditions, boundary conditions, and forcings of the PDE are all fixed thus requiring complete retraining for a new set of parameters. Furthermore, this approach requires knowledge of the underlying PDE which is not always available.

The first results to propose surrogate models between function spaces which are independent of mesh discretization and do not rely on stringent modeling assumptions have appeared in works on neural operators (Lu et al., 2019; Bhattacharya et al., 2021; Li et al., 2020, 2021) and operator-valued kernels (Nelsen and Stuart, 2021; Bao et al., 2022) for estimating PDE solution maps. By making little to no assumptions on the domain geometry and mesh discretization of the data, these works produced efficient surrogate models which could be applied to general nonlinear PDEs. Below we follow the same guiding principle to learn surrogate maps for a large class of linear PDEs by means of learning their Green's function. Restricting ourselves to linear systems enables us to prove rates on the prediction error of our model as in de Hoop et al. (2021); Boullé and Townsend (2022) and places us in a setting where the learned surrogate models become interpretable.

A broad class of linear PDEs including the Poisson equation, wave equation, and heat equation are solved by an integral operator or *fundamental solution* of the form

$$u(y) = \mathcal{T}(f)(y) = \beta(y) + \int_{D_\mathcal{X}} G(x,y)f(x)\mathrm{d}x \tag{2}$$

where $G \in L^2(D_\mathcal{X} \times D_\mathcal{Y})$ is called the *Green's function* and $\beta \in L^2(D_\mathcal{Y})$ is a bias term that satisfies the boundary conditions of the PDE and solves the homogeneous equation $\mathcal{P}u = 0$. In general, the domain $D_\mathcal{X}$ of the input function f can be different from the domain $D_\mathcal{Y}$ of the solution u, if for example f is an initial or boundary condition of a PDE.

In recent years, a large body of work has developed fast and accurate approximations to Green's functions of linear PDEs. Simulating the solution of a linear PDE with a new input function is equivalent to integrating this input against the Green's function of the PDE as in (2). If the Green's function can be efficiently constructed from input-output pairs (f_i, u_i) where $u_i \approx Gf_i$, then it can be used to solve the PDE under new forcings, initial, and boundary conditions. *Matrix probing* algorithms discretize the Green's function at a set of collocation points or in a basis and reconstruct the matrix G from a small number of matrix-vector products Gf_i. By assuming that G lies in the span of prespecified basis matrices $\{B_1, \ldots, B_m\}$, Chiu and Demanet (2012) solve a least squares problem to learn a Green's function independent of the mesh discretization. This methodology is applied to solve the Helmholtz equation with absorbing boundary conditions (Bélanger-Rioux and Demanet, 2015) as well as linearized seismic inversion problems (Demanet et al., 2012). For a large class of elliptic PDEs, the Green's function under mild regularity assumptions exhibits hierarchical low-rank structure (Bebendorf and Hackbusch, 2003). Several matrix probing algorithms (Lin et al., 2011; Boullé and Townsend, 2022; Boullé et al., 2022b) leverage this structure by evaluating the PDE with a few random forcings and use randomized SVD to efficiently learn the low-rank sub-blocks of G. For elliptic PDEs with symmetric Green's functions, sparse Cholesky factorization and operator-adapted wavelets (Schäfer et al., 2021b; Owhadi et al., 2019; Schäfer et al., 2021a) can be applied to compress G and G^{-1} in order to expedite simulations of boundary layer problems and sparse ill-conditioned PDEs arising from computer graphics (Chen et al., 2021a). We refer the readers to Owhadi (2015, 2017); Owhadi and Scovel (2019) for a comprehensive review of Bayesian numerical homogenization, operator-adapted wavelets and their application to fast multigrid and multiresolution methods for linear PDEs. Finally, recent approaches have modeled Green's functions of PDEs using autoencoders (Gin et al., 2021) and rational neural networks (Boullé et al., 2022a).

Similar to the approaches above, we are interested in learning the Green's function G of a linear PDE in order to learn a surrogate model $\mathcal{T} : f \mapsto u$ from the input of a PDE (e.g. forcing, initial condition, boundary condition) to its solution. We restrict ourselves to only observing input-output samples (f_i, u_i) of the PDE and develop a Green's function estimator that is robust to high levels of noise in the data, a property that is crucial for learning on real data sets but is less addressed in prior work on Green's function estimation. Compared to the learning methods outlined above, we do not make assumptions on the form of the underlying PDE or its Green's function (e.g. elliptic, hyperbolic, hierarchical low-rank).

For different physical problems, the Green's function of a PDE satisfies certain sparsity, continuity, smoothness conditions, or combinations thereof. This motivates us to search for linear forward maps $\mathcal{T}$ of the form given in (2) where the Green's function G and bias β belong to a pair of reproducing kernel Hilbert spaces (RKHSs) $\mathcal{G}, \mathcal{B}$ respectively. This setting belongs to a subclass of operator RKHSs studied in Kadri et al. (2016) and restricting ourselves to the space of integral operators leads to more flexibility and insight about the choice of the RKHSs $\mathcal{G}, \mathcal{B}$ for many physical problems.

Optimizing $\mathcal{T}$ over the space of integral operators with the Green's function G and bias β in an RKHS offers four concrete advantages:

1. The estimators $\widehat{G}, \widehat{\beta}$ of the Green's function and bias term are robust to significant levels of noise in the input-output samples $\{(f_i, u_i)\}_{i=1}^n$ due to the penalization of their RKHS norm.

2. It gives an explicit closed-form (representer theorem) for the best-fit functions $\widehat{G}, \widehat{\beta}$ over data samples $\{(f_i, u_i)\}_{i=1}^n$ which is independent of the mesh discretization of the samples and, crucially, can extrapolate to new meshes.

3. It allows us to interpret our learned model by inspecting the estimated Green's function $\widehat{G}$ which specifies the impulse response of the system.

4. The RKHSs can be designed to enforce specific structure and symmetries in $\widehat{G}, \widehat{\beta}$ based on prior knowledge about the system.

The rest of this paper is organized as follows. We discuss our data generating model in Section 2 along with the full representer theorem for the best-fit Green's function and bias term. The implementation of our Green's function and bias term RKHS estimators are detailed in Section 3. In Section 4, we extend the analysis of Yuan et al. (2010) for real outputs to the case of functional outputs and derive the corresponding error bounds for the RKHS Green's function estimator. A concrete list of examples to estimate Green's functions and bias terms of linear PDEs in different RKHSs are given in Section 5 with all proofs deferred to the appendix.

1.1 Motivating Example

All approaches for learning Green's function, including our proposed method, begin by taking a set of input-output functions $\{(f_i(x), u_i(y))\}_{i=1}^n$ discretized on a finite set of grid points $\{x_j\}_{j=1}^{m_x}$ and $\{y_k\}_{k=1}^{m_y}$ respectively. The goal is to learn a function $G(x, y)$ such that numerical integrations of G

against the discretized inputs f_i are as close as possible to the discretized outputs u_i. This can be written as

$$u(y_k) \approx \sum_{j=1}^{m_x} G(x_j, y_k) f_i(x_j) \Delta_j^x \tag{3}$$

where Δ_j^x are numerical quadrature weights.

The classical grid-based methods outlined above for Green's function estimation including matrix probing and sparse factorization have focused almost exclusively on the noiseless setting when data collected from the underlying PDE can be perfectly measured. These methods aim to learn G solely on the grid points $G(x_j, y_k)$ and do not enforce smoothness by constraining the values of G at neighboring grid points to be close. They are then able to fully exploit the linearity of (3) to obtain remarkably efficient and accurate algorithms on noiseless data. In practice however, measured data from real-world systems are often corrupted with high levels of noise which make these prior approaches inapplicable. On a moderate number of input-output samples, these classical grid based approaches tend to overfit noisy data regardless of the number of grid points $\{x_j\}_{j=1}^{m_x}$, $\{y_k\}_{k=1}^{m_y}$ used to construct the estimator.

Here we show how estimating the Green's function of a PDE in an RKHS allows us to fit (3) while also penalizing the RKHS norm of our estimator. The additional RKHS norm leads to a convex objective for G and naturally penalizes the smoothness of the learned Green's function allowing for significant robustness to noise (see Sections 2 & 3).

In Figure 1, we compare our approach to a classical grid-based method for learning the Green's function of the Poisson equation. Taking the Poisson equation $\Delta u = f$ on $[0, 1]$ with zero Dirichlet boundary conditions, we estimate it's Green's function from 500 functional samples (f_i, u_i) discretized on a 100 point uniform grid $\{x_j = y_j = \frac{j-1}{99}\}_{j=1}^{100}$. The input forcings f_i are simulated using a Karhunen–Loeve expansion (KLE) with a squared exponential kernel of lengthscale $\ell = 0.01$ and the solutions u_i are generated with a standard finite difference solver and corrupted with 10% Gaussian noise (see Appendix A for details). The true analytic Green's function of the Poisson equation is given by

$$G_{\text{Poisson}}(x, y) = \frac{1}{2}(x + y - |x - y|) - xy \tag{4}$$

depicted in the rightmost plot. With noise corrupted data, naively learning the Green's function G as a discretized matrix $G_{jk} = G(x_j, x_k)$ by solving the least squares problem $\sum_{i=1}^{500} \|\mathbf{G}^T f_i - u_i\|_2^2$ leads to a nonsmooth estimator that is corrupted by the noise in the train samples (left plot). Instead, by learning $G(x, y)$ as a *function* in a squared exponential RKHS (e.g. sum of 2D Gaussian kernels of width $\sigma = 5 \times 10^{-2}$) we can penalize the smoothness of G to learn a much more faithful estimate of the true Green's function (center plot). Our learned estimator is smooth and has an analytic closed form which can be reevaluated on finer mesh sizes. As opposed to the matrix estimator (left plot), the RKHS estimator (center plot) is much more interpretable as it allows us to conclude that perturbations f concentrated around a point $x_0 \in [0, 1]$ produce a smoothed response in u around that same point and that such perturbations get weaker as x_0 approaches the boundary of the domain. This example demonstrates the importance of learning Green's functions of PDEs in function spaces that enforce structure such as continuity and smoothness.

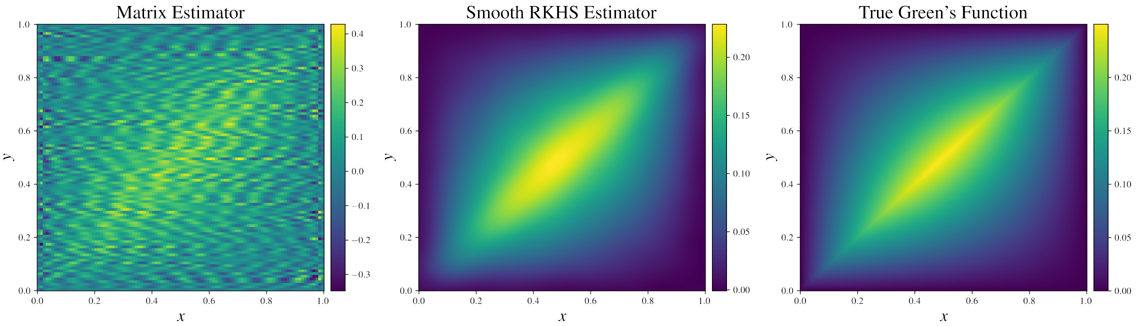


Figure 1: Estimating Green's function of the Poisson equation with zero Dirichlet boundary conditions on 500 noisy functional samples (f_i, u_i) discretized on a 100 point uniform grid. Learning a Green's function as a 100×100 matrix without enforcing smoothness leads to a noise corrupted estimator that overfits the training samples. Alternatively, learning the Green's function as a sum of 2D Gaussian kernels (e.g. in a squared exponential RKHS) and penalizing its smoothness with regularization $\lambda = 10^{-4}$ faithfully estimates the true Green's function. This estimator has a known functional form which allows us to resample it on a finer mesh size of 500×500 (center plot). The true Green's function of the Poisson equation is shown in the rightmost plot.

1.2 Definitions and Notation

Before we describe our data model and estimator, we introduce mathematical definitions and notation that will be used throughout the paper.

1.2.1 Function Spaces

Definition 1 (Real Hilbert Space) *A real Hilbert space $\mathcal{H}$ is a real inner product space with inner product $\langle x, y\rangle_{\mathcal{H}}$ for all $x, y \in \mathcal{H}$. The inner product induces a norm given by $\|x\|_{\mathcal{H}} = \langle x, x\rangle_{\mathcal{H}}$ such that the Hilbert space $\mathcal{H}$ is a complete metric space with respect to the metric $d(x, y) = \|x - y\|_{\mathcal{H}}$.*

The space $L^2(D)$ denotes the Hilbert space of square-integrable functions on the domain D with the standard inner product $\langle f, g\rangle_{L^2(D)} = \int_D f(x)g(x)\mathrm{d}x$ and norm $\|f\|_{L^2(D)}^2 = \int_D f(x)^2\mathrm{d}x$ for all $f, g \in L^2(D)$. Likewise, the Euclidean space $\mathbb{R}^d$ equipped with the inner product $\langle u, v\rangle_2 = \sum_{i=1}^d u_i v_i$ is a simple example of a Hilbert space.

Definition 2 (Reproducing Kernel Hilbert Space) *A reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ is a Hilbert space of functions on a domain D with an inner product $\langle f, g\rangle_{\mathcal{H}}$ for all $f, g \in \mathcal{H}$. For each $x \in D$, the Hilbert space $\mathcal{H}$ has a unique element $K_x \in \mathcal{H}$ such that*

$$\langle f, K_x\rangle_{\mathcal{H}} = f(x) \quad \forall f \in \mathcal{H} \tag{5}$$

which is called the reproducing property. The function $K : D \times D \to \mathbb{R}$ defined as $K(x, y) = \langle K_x, K_y\rangle$ for all $x, y \in D$ is called the reproducing kernel. Often we will denote the Hilbert space norm by $\|\cdot\|_{\mathcal{H}} = \|\cdot\|_K$ and the inner product by $\langle\cdot,\cdot\rangle_{\mathcal{H}} = \langle\cdot,\cdot\rangle_K$.

1.2.2 Kernel Functions

Given a set or domain D, a real-valued kernel is a function $K : D \times D \to \mathbb{R}$ that is symmetric which means $K(x, y) = K(y, x)$ for all $x, y \in D$. Furthermore, a kernel is called *positive semidefinite* if

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0 \tag{6}$$

for all $x_1, \ldots, x_n \in D$ for any integer $n \geq 1$ and $c_1, \ldots, c_n \in \mathbb{R}$. A kernel is called *positive definite* if the inequality in (6) is strict.

For a domain $D \subseteq \mathbb{R}^d$, we say that K is a *Mercer kernel* if it is a continuous function on $D \times D$ that is symmetric and positive semidefinite.

1.2.3 Probability and Expectation

- For a distribution $\mathbb{P}$ over functions in $L^2(D)$, the notation $F \sim \mathbb{P}$ denotes that F is a functional sample from this distribution.

- The expectation with respect to a distribution $\mathbb{P}$ is denoted by $\mathbb{E}_{\mathbb{P}}[\cdot]$ and the subscript is dropped for convenience when it is clear which distribution is being used.

1.2.4 Functions and Operators

- Given two functions $f(x)$ and $g(y)$ we define their tensor product as $(f \otimes g)(x, y) = f(x)g(y)$.

- Given two domains $D_{\mathcal{X}}$ and $D_{\mathcal{Y}}$ we can define a function $M \in L^2(D_{\mathcal{X}} \times D_{\mathcal{Y}})$. For any $f \in L^2(D_{\mathcal{Y}})$ we use the shorthand $M(f) = \int_{D_{\mathcal{Y}}} M(x, y) f(y) \mathrm{d}y$ and similarly for any $f \in L^2(D_{\mathcal{X}})$ we write $M^T(f) = \int_{D_{\mathcal{X}}} M(x, y) f(x) \mathrm{d}x$.

- An operator $\Gamma : L^2(D) \to L^2(D)$ for some domain D is called positive semidefinite if $\langle \Gamma(f), f \rangle_{L^2(D))} \geq 0$ for all $f \in L^2(D)$. If this inequality is always strictly larger than zero, then it is called positive definite. For two operators $\Sigma : L^2(D) \to L^2(D)$ and $\Gamma : L^2(D) \to L^2(D)$, we use the notation $\Sigma \preceq \Gamma$ when the difference $\Gamma - \Sigma$ is a positive semidefinite operator and the notation $\Sigma \prec \Gamma$ when their difference is strictly positive definite.

1.2.5 Sets and Sequences

- For a positive integer m, we use the notation $[m]$ to denote the set of integers from 1 to m.

- We use $\mathbb{R}$ to denote the set of real numbers and $\mathbb{R}_+$ to denote the set of nonnegative real numbers.

- For two positive real sequences $a_n, b_n \in \mathbb{R}_+$ for $n \geq 1$ we use the symbol $a_n \asymp b_n$ to denote that the ratio a_n/b_n is bounded away from zero and infinity as $n \to \infty$.

- We write $a_n \lesssim b_n$ to signify that $a_n \leq C b_n$ for all $n \geq 1$ for some real constant $C > 0$.

2. Model and Estimator

Given compact domains $D_{\mathcal{X}} \subset \mathbb{R}^{d_{\mathcal{X}}}$ and $D_{\mathcal{Y}} \subset \mathbb{R}^{d_{\mathcal{Y}}}$, our goal is to learn a map from input functions $f : D_{\mathcal{X}} \to \mathbb{R}$ to output solutions $u : D_{\mathcal{Y}} \to \mathbb{R}$. We note that $D_{\mathcal{X}}$ and $D_{\mathcal{Y}}$ can be different domains such as when f is the boundary condition of a PDE and u is the solution on the interior of the domain.

We consider the model for a random input-output pair $(F, U) \in L^2(D_{\mathcal{X}}) \times L^2(D_{\mathcal{Y}})$:

$$U = \mathcal{T}^*(F) + \varepsilon \tag{7}$$

where $\mathcal{T}^* : L^2(D_{\mathcal{X}}) \to L^2(D_{\mathcal{Y}})$ is a possibly nonlinear operator (parameter of interest) and ε is a centered random variable in $L^2(D_{\mathcal{Y}})$. We assume further that $F \sim \mathsf{subG}(\Gamma_F)$ and $\varepsilon \sim \mathsf{subG}(\Gamma_\varepsilon)$ are subgaussian where $\Gamma_F : L^2(D_{\mathcal{X}}) \to L^2(D_{\mathcal{X}})$ and $\Gamma_\varepsilon : L^2(D_{\mathcal{Y}}) \to L^2(D_{\mathcal{Y}})$ are positive semidefinite trace-class linear operators known as *covariance proxies*; see Appendix E. The *covariance operator* of F is the function $\Sigma_F \in L^2(D_{\mathcal{X}} \times D_{\mathcal{X}})$ given by

$$\Sigma_F(x, x') = \mathbb{E}[(F(x) - \mathbb{E}[F(x)]) \cdot (F(x') - \mathbb{E}[F(x')])] \tag{8}$$

where the expectation is taken over the randomness of the process F. The covariance can also be interpreted as a linear operator $\Sigma_F : L^2(D_{\mathcal{X}}) \to L^2(D_{\mathcal{X}})$ and as shorthand we will write $\Sigma_F = \mathbb{E}[(F - \mathbb{E}[F]) \otimes (F - \mathbb{E}[F])]$ where $\otimes$ is the tensor product. For our theoretical analysis, we make the assumption that F is strictly subgaussian as defined in Appendix E.

Assumption 1 *The subgaussian process $F \sim \mathsf{subG}(\Gamma_F)$ is strictly subgaussian meaning that $\Gamma_F \preceq C\Sigma_F$ for some constant $C > 0$. Additionally, we assume that Γ_F is strictly positive definite. In other words, the covariance function and covariance proxy of F are both strictly positive definite and on the same order*

$$0 \prec c\Sigma_F \preceq \Gamma_F \preceq C\Sigma_F \tag{9}$$

for some $0 < c < C$.

Modeling the input functions F by a subgaussian distribution includes as a subset all inputs which can be constructed from random subgaussian-weighted combinations of basis functions with sufficient decay in their weights (see Karhunen–Loeve expansion in Appendix A.1). Such random functions are used extensively as initial conditions, boundary conditions, and forcing functions for learning PDEs (Bhattacharya et al., 2021; Boullé and Townsend, 2022).

To learn an operator for the solution map of a linear PDE, the affine representation (2) suggests to consider operators of the form

$$\mathcal{T}_{\beta,G}(f) = \beta + \int_{D_{\mathcal{X}}} G(x, \cdot) f(x) \mathrm{d}x\,, \qquad \beta \in L^2(D_{\mathcal{Y}}),\ G \in L^2(D_{\mathcal{X}} \times D_{\mathcal{Y}}) \tag{10}$$

This simple representation is still too flexible to be learned from a finite amount of data. To overcome this limitation, we impose additional regularity on G and β, namely that $G \in \mathcal{G}$ and $\beta \in \mathcal{B}$, where $\mathcal{G}$ and $\mathcal{B}$ are two RKHSs with continuous, square integrable, and strictly positive definite reproducing kernels

$$K : (D_{\mathcal{X}} \times D_{\mathcal{Y}}) \times (D_{\mathcal{X}} \times D_{\mathcal{Y}}) \to \mathbb{R} \text{ and } Q : D_{\mathcal{Y}} \times D_{\mathcal{Y}} \to \mathbb{R}. \tag{11}$$

In fact we establish below an oracle inequality that holds for a much more general class of estimators for $\mathcal{T}^*$. In particular, it shows that our estimator performs well not only when $\mathcal{T}^*$ is affine as in (10) (well-specified model) but also if it is well approximated by such estimators (mis-specified case); see Theorem 7 below. In general, our results hold under the following sublinear growth condition.

Assumption 2 *The true solution map $\mathcal{T}^* : L^2(D_{\mathcal{X}}) \to L^2(D_{\mathcal{Y}})$ of the PDE has at most linear growth, that is, for all $f \in L^2(D_{\mathcal{X}})$,*

$$\|\mathcal{T}^*(f)\|_{L^2(D_{\mathcal{Y}})} \leq c + M\|f\|_{L^2(D_{\mathcal{X}})} \tag{12}$$

where $c, M \geq 0$ are constants.

If we set $c = 0$ then Assumption 2 is equivalent to requiring that $\mathcal{T}^*$ be a bounded operator. As an example, a general class of elliptic PDEs on compact domains have bounded solution maps $\mathcal{T}^*$ by the bounded inverse theorem (Evans, 1998, Section 6.2, Theorem 6). The condition above is also clearly satisfied by all linear PDEs that have a square integrable Green's function.

We are now in a position to describe our estimator. Assume that we observe independent samples $(F_1, U_1), \ldots, (F_n, U_n)$ of (F, U) from (7) and define the empirical risk

$$\widehat{R}(\beta, G) := \frac{1}{n}\sum_{i=1}^{n}\left\|U_i - \beta - \int_{D_{\mathcal{X}}} G(x, \cdot)F_i(x)\mathrm{d}x\right\|^2_{L^2(D_{\mathcal{Y}})}. \tag{13}$$

Likewise, define the penalized empirical risk

$$\widehat{R}_{\rho,\lambda}(\beta, G) := \widehat{R}(\beta, G) + \rho P(\beta) + \lambda J(G) \tag{14}$$

where P and J are the penalty functionals for the Green's function and bias term respectively. Then our estimators are defined as the RKHS minimizers

$$\widehat{\beta}_{n,\rho,\lambda}, \widehat{G}_{n,\rho,\lambda} := \underset{\beta\in\mathcal{B}, G\in\mathcal{G}}{\arg\min}\, \widehat{R}_{\rho,\lambda}(\beta, G) \tag{15}$$

Here, the subscripts n, ρ, λ on our estimators indicate the number of samples n and the regularization values ρ, λ for which our estimators were optimally chosen. We remind the reader that $\mathcal{G}$ is the RKHS of Green's functions with reproducing kernel $K : (D_{\mathcal{X}} \times D_{\mathcal{Y}})^2 \to \mathbb{R}$ and $\mathcal{B}$ is the RKHS of bias terms with reproducing kernel $Q : D_{\mathcal{Y}} \times D_{\mathcal{Y}} \to \mathbb{R}$ where both reproducing kernels are continuous, symmetric, and strictly positive definite. The penalty functionals with which we will regularize the Green's function and bias term are the respective RKHS norms

$$J(G) = \|G\|^2_{\mathcal{G}} = \|G\|^2_K, \quad P(\beta) = \|\beta\|^2_{\mathcal{B}} = \|\beta\|^2_Q \tag{16}$$

which make the penalized empirical risk $\widehat{R}_{\rho,\lambda}(\beta, G)$ a strictly convex objective such that our estimators $\widehat{\beta}_{n,\rho,\lambda}, \widehat{G}_{n,\rho,\lambda}$ are unique.

2.1 Common Examples of RKHS Kernels

By the classical Moore-Aronzajn theorem (Aronszajn, 1950), any positive-semidefinite kernel defines a unique RKHS. Hence, for nonzero measure sets $D_{\mathcal{X}} \subset \mathbb{R}^{dx}, D_{\mathcal{Y}} \subset \mathbb{R}^{dy}$ the *squared-exponential (SE) kernels*

$$K(x, y, \xi, \eta) = \exp\Big(-\frac{\|x - \xi\|^2}{2\sigma_x^2}\Big)\exp\Big(-\frac{\|y - \eta\|^2}{2\sigma_y^2}\Big), \quad Q(y, \eta) = \exp\Big(-\frac{\|y - \eta\|^2}{2\sigma_y^2}\Big) \tag{17}$$

generate unique RKHSs $\mathcal{G}$ and $\mathcal{B}$ for the Green's function and bias term respectively. The RKHS norms defined by such SE kernels strongly penalize derivatives of a function and hence bias the choice of G and β towards very smooth functions. In the example above, the kernel $K(x, y, \xi, \eta)$ is constructed as a products of simpler SE kernels in (x, ξ) and (y, η). Constructing kernels through tensor products is a standard procedure outlined in Saitoh and Sawano (2016, Theorem 2.20) and is an important tool for building RKHSs of higher-dimensional functions.

Another popular example are RKHSs generated by *exponential kernels*

$$K(x, y, \xi, \eta) = \exp\Big(-\sqrt{\frac{\|x-\xi\|^2}{\sigma_x^2} + \frac{\|y-\eta\|^2}{\sigma_y^2}}\Big), \quad Q(y, \eta) = \exp\Big(-\frac{\|y-\eta\|}{\sigma_y}\Big) \tag{18}$$

whose RKHS norms do not penalize any derivatives and allow us to represent functions G and β which are nondifferentiable.

An important family of kernel functions known as *Matérn kernels* (Genton, 2001) are given by

$$\begin{gathered} C_\nu(d) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\Big(\frac{\sqrt{2\nu}}{l}d\Big)^\nu K_\nu\Big(\frac{\sqrt{2\nu}}{l}d\Big) \\ K(x, y, \xi, \eta) = C_\nu\Big(\sqrt{\frac{\|x-\xi\|^2}{\sigma_x^2} + \frac{\|y-\eta\|^2}{\sigma_y^2}}\Big), \quad Q(y, \eta) = C_\nu\Big(\frac{\|y-\eta\|}{\sigma_y}\Big) \end{gathered} \tag{19}$$

and interpolate between the exponential kernel at $\nu = 0$ and the Gaussian kernel as $\nu \to \infty$. The parameter ν controls how strongly the magnitudes of higher-order derivatives of G and β are penalized.

The positive definite kernel functions described above are all examples of *radial basis functions* (RBFs) or anisotropic variants of RBFs; functions that only depend on the distances $\|x - \xi\|$ and $\|y - \eta\|$. In general, reproducing kernels are not restricted to be of this form. For example, given any finite or infinite set of orthonormal functions $\{\psi_k\}_{k=1}^m$ on $L^2(D)$ we have that

$$K(x, y) = \sum_{k=1}^{m} \lambda_k \psi_k(x)\psi_k(y) \tag{20}$$

with $\lambda_k > 0$ and $\sum_{k=1}^m \lambda_k^2 < \infty$ defines an RKHS of functions on the domain D with inner product

$$\langle f, g\rangle = \sum_{k=1}^{m} \frac{\langle f, \psi_k\rangle_{L^2(D)}\langle g, \psi_k\rangle_{L^2(D)}}{\lambda_k}. \tag{21}$$

for all f, g in this RKHS. In particular, a basis of $L^2(D)$ such as a Fourier or polynomial basis truncated to a finite number of terms is an example of an RKHS. As a concrete example, on the box domain $D = [0, 1]^d$ the kernel

$$K(x_1, \ldots, x_d, \xi_1, \ldots, \xi_d) = 2^d \sum_{k_1,\ldots,k_d=1}^{\infty} \frac{\sin(\pi k_1 x_1)\ldots\sin(\pi k_d x_d)\sin(\pi k_1 \xi_1)\ldots\sin(\pi k_d \xi_d)}{\pi^2(k_1^2 + \ldots + k_d^2)}. \tag{22}$$

is a reproducing kernel for the Sobolev-Hilbert space

$$W_1^2(D) = \Big\{f : f \text{ absolutely continuous}, \ f \equiv 0 \text{ on } \partial D, \ \frac{\partial f}{\partial x_i} \in L^2(D), \ \forall 1 \le i \le d\Big\} \tag{23}$$

with inner product

$$\langle f, g \rangle_{W_1^2(D)} = \int_D \nabla f(x) \cdot \nabla g(x) \mathrm{d}x. \tag{24}$$

This can be seen by noting that K is the Green's function of the Poisson equation

$$-\Delta u(x) = s(x) \quad \forall x \in D, \qquad u(x) = 0 \quad \forall x \in \partial D \tag{25}$$

with homogeneous Dirichlet boundary conditions (Polyanin and Nazaikinskii, 2015, Section 8.2.2-16).

In this paper, all experiments described in Section 5 only use the exponential, squared exponential, and Matérn kernels as they are used ubiquitously in the kernel methods literature, are simple to implement, and can be computed efficiently through fast kernel matrix-vector products (see Section 3). These kernels are strictly positive definite although our analysis can also be extended to degenerate kernels which are positive semidefinite.

Remark 3 *We refer the reader to the classical texts of (Wahba, 1990, Chapters 1, 2, 10), Berlinet and Thomas-Agnan (2011, Chapters 1, 7) and Saitoh and Sawano (2016, Chapter 1) for detailed examples of reproducing kernels and their associated Hilbert spaces. In particular, these texts outline the deep connection between Green's functions of differential equations and reproducing kernels. As shown on the example of the Poisson equation in* (25)*, Green's functions of classical PDEs can be seen as natural reproducing kernels over the space of their solutions. Similar reproducing kernels can be derived from Green's functions of the Helmholtz and heat equations (Saitoh and Sawano, 2016, Sections 1.7.2-1.7.3). In this paper, we take the opposite perspective and use a reproducing kernel to learn a Green's function of an unknown PDE from data. Our choice of reproducing kernel for the RKHS gives rise to a best-fit estimator for the Green's function of a PDE.*

2.2 Representer Theorem

Given random input-output function samples $\{(F_i, U_i)\}_{i=1}^n$ from (7), we would like to minimize the regularized cost function $\widehat{R}_{\rho,\lambda}$ defined in (14). Our cost function is composed of a convex mean-squared error $\widehat{R}(\beta, G)$ given in (13) as well as two strictly convex RKHS regularizers $J(G) = \|G\|_{\mathcal{G}}^2$ and $P(\beta) = \|\beta\|_{\mathcal{B}}^2$. Hence, it is strictly convex implying that it has a unique minimizer

$$\widehat{\beta}_{n,\rho,\lambda}, \widehat{G}_{n,\rho,\lambda} := \underset{\beta \in \mathcal{B}, G \in \mathcal{G}}{\arg\min}\, \widehat{R}_{\rho,\lambda}(\beta, G). \tag{26}$$

which are the estimators for the Green's function and bias term of our PDE.

In practice, the functional inputs $F_i(x)$ and ouputs $U_i(y)$ are given to us at discretized mesh points $\{x_j\}_{j=1}^{m_x}$ and $\{y_k\}_{k=1}^{m_y}$. Given discretized data, our original objective function

$$\widehat{R}_{\rho,\lambda}(\beta, G) = \frac{1}{n} \sum_{i=1}^{n} \left\| U_i - \beta - \int_{D_{\mathcal{X}}} G(x, \cdot) F_i(x) \mathrm{d}x \right\|_{L^2(D_{\mathcal{Y}})}^2 + \rho \|\beta\|_Q^2 + \lambda \|G\|_K^2 \tag{27}$$

is numerically approximated by a Riemann sum

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{m_y} \Big(U_i(y_k) - \beta(y_k) - \sum_{j=1}^{m_x} G(x_j, y_k) F_i(x_j) \Delta_j^x \Big)^2 \Delta_k^y + \rho \|\beta\|_Q^2 + \lambda \|G\|_K^2 \tag{28}$$

where Δ_j^x and Δ_k^y are the quadrature weights for the Riemann sums in x and y respectively. Importantly, here we discretize the square loss term using a Riemann sum but the Hilbert norms $\|\beta\|_Q^2$, $\|G\|_K^2$ are kept continuous. The semi-discrete objective function above is now in the right form for us to apply the traditional representer theorem to β and G. First we see that any minimizer β must have the form

$$\widehat{\beta}_{n,\rho,\lambda}(y) = \sum_{k=1}^{m_y} Q(y, y_k) w_k \Delta_k^y \tag{29}$$

where $\mathbf{w} = (w_1, \ldots, w_{m_y})^T \in \mathbb{R}^{m_y}$ is any weight vector and Q is once again the reproducing kernel for the RKHS $\mathcal{B}$ of the bias term. Now fixing β, our discretized loss function (28) in G has the form

$$L\Big(\Big\{\Big\langle \mathbf{G}, \mathbf{A}_{ik}\Big\rangle_2\Big\}_{i\in[n],k\in[m_y]}\Big) + \lambda\|G\|_K^2, \qquad \mathbf{G} = \{G(x_j, y_k)\},\ \ \mathbf{A}_{ik} = (\mathbf{F}_i \odot \boldsymbol{\Delta}^x)\mathbf{e}_k^T \in \mathbb{R}^{m_x\times m_y} \tag{30}$$

for some loss function $L : \mathbb{R}^n \to \mathbb{R}$ where $\mathbf{F}_i = (F_i(x_1), \ldots, F_i(x_{m_x}))^T \in \mathbb{R}^{m_x}$ and $\boldsymbol{\Delta}^x = (\Delta_1^x, \ldots, \Delta_{m_x}^x)^T \in \mathbb{R}^{m_x}$ and $\mathbf{e}_k \in \mathbb{R}^{m_y}$ denotes the unit vector with a one in the kth position. Here $\odot$ denotes the element-wise product and $\langle \mathbf{G}, \mathbf{A}_{ik}\rangle_2$ is the matrix trace inner product. The loss function as written above is a function of the $m_x \times m_y$ function evaluations $\{G(x_j, y_k)\}$ plus a regularization term so it is directly amenable to the classical representer theorem in G. Hence, any minimizer G must have the form

$$\widehat{G}_{n,\rho,\lambda}(x, y) = \sum_{j=1}^{m_x}\sum_{k=1}^{m_y} K(x, y, x_j, y_k) W_{jk} \tag{31}$$

where $\mathbf{W} = \{W_{jk}\} \in \mathbb{R}^{m_x\times m_y}$ is any weight matrix and K is the reproducing kernel for the RKHS $\mathcal{G}$ of the Green's function. Surprisingly, the particular form of our loss function $L(\{\langle \mathbf{G}, \mathbf{A}_{ik}\rangle\}_{i\in[n],k\in[m_y]})$ allows us to give a more constrained description of G. Since the function evaluations $\mathbf{G} = \{G(x_j, y_k)\}$ only enter our loss function through inner products with $\{\mathbf{A}_{ik}\}$, we can in fact show that

$$\widehat{G}_{n,\rho,\lambda}(x, y) = \sum_{j=1}^{m_x}\sum_{k=1}^{m_y} K(x, y, x_j, y_k) W_{jk}, \quad \mathbf{W} \in \operatorname{span}\{\mathbf{A}_{ik} : i \in [n], k \in [m_y]\}. \tag{32}$$

A concise proof of this statement is detailed at the start of Appendix B. If we expand

$$\mathbf{W} = \sum_{i=1}^{n}\sum_{k=1}^{m_y} \mathbf{A}_{ik} c_{ik} \Delta_k^y \tag{33}$$

for any constants $c_{ik} \in \mathbb{R}$ then we can finally write

$$\widehat{G}_{n,\rho,\lambda}(x, y) = \sum_{j=1}^{m_x}\sum_{k=1}^{m_y} K(x, y, x_j, y_k) F_i(x_j) c_{ik} \Delta_j^x \Delta_k^y. \tag{34}$$

Hence we have derived a representer theorem for our Green's function G that minimizes the discretized loss in (28). Building on the derivations in Wahba (1990, Section 1.3, Theorem 1.3.1), we also present a continuous version of this result when the mesh discretizations $\{x_j\}_{j=1}^{m_x}$ and $\{y_k\}_{k=1}^{m_y}$ are taken to the continuum limit, i.e. minimizing $\widehat{R}_{\rho,\lambda}(\beta, G)$ from (27) directly without a Riemann sum approximation.

Theorem 4 (Green's Function Representer Theorem) *For any minimizer $\widehat{\beta}_{n,\rho,\lambda}, \widehat{G}_{n,\rho,\lambda}$ of the empirical risk $\widehat{R}_{\rho,\lambda}(\beta, G)$ from* (27) *on functional data $\{(F_i, U_i)\}_{i=1}^n \subset L^2(D_\mathcal{X}) \times L^2(D_\mathcal{Y})$, the function $\widehat{G}_{n,\rho,\lambda}$ must have the form*

$$\widehat{G}_{n,\rho,\lambda}(x, y) = \sum_{i=1}^{n} \int_{D_\mathcal{X}} \int_{D_\mathcal{Y}} K(x, y, \xi, \eta) F_i(\xi) c_i(\eta) \mathrm{d}\xi \mathrm{d}\eta \tag{35}$$

where $c_i \in L^2(D_\mathcal{Y})$ for $i \in [n]$ are coefficient functions which are free to be determined.

The theorem above can be seen as a consequence of the general representer theorem derived in the seminal work of Micchelli and Pontil (2005, Section 4, Theorem 4.1) which laid the groundwork for vector-valued RKHSs, sometimes called operator RKHSs. We give an alternative proof of this result in Appendix B at the same level of generality for any loss function optimized over the space of an operator RKHS (i.e. not just integral operators). A similar representer theorem is also proven in Kadri et al. (2016, Appendix B, Theorem 9) albeit with restrictive assumptions that the loss function and regularization term are quadratic.

In the following sections, we describe the numerical implementation of our Green's function and bias term estimators and derive error bounds for approximating the Green's function G of a PDE in an RKHS $\mathcal{G}$ given functional data samples (F_i, U_i).

3. Implementation

In practice, functional input-output data $\{(F_i, U_i)\}_{i=1}^n$ are almost always discretized on a set of mesh points $\{x_j\}_{j=1}^{m_x}, \{y_k\}_{k=1}^{m_y}$ so our implementation of the Green's function and bias term estimators are based on the discrete representer theorems from equations (29) and (31) respectively

$$\beta_\mathbf{w}(y) = \sum_{k=1}^{m_y} Q(y, y_k) w_k \Delta_k^y, \quad G_\mathbf{W}(x, y) = \sum_{j=1}^{m_x} \sum_{k=1}^{m_y} K(x, y, x_j, y_k) W_{jk} \Delta_j^x \Delta_k^y. \tag{36}$$

Note that we do not use the more restricted form of the discrete representer theorem (32) for the Green's function G as it is only efficient in the small data limit when $n \ll \min(m_x, m_y)$ but in practice we take m_x, m_y on the order of 10^2 and the number of training samples n range from 100 to 500.

The full forward map $\mathcal{T} : f \to u$ of the PDE from (10) is estimated by

$$[\mathcal{T}_{\mathbf{w},\mathbf{W}}(f)](y) = [\mathcal{T}_{\beta_\mathbf{w}, G_\mathbf{W}}(f)](y) = \beta_\mathbf{w}(y) + \sum_{j=1}^{m_x} G_\mathbf{W}(x_j, y) f(x_j) \Delta_j^x. \tag{37}$$

where $\mathbf{w}, \mathbf{W}$ are the weights of our estimator. Given that (28) is a convex objective, a natural choice is to learn $\mathbf{w}, \mathbf{W}$ through convex optimization. However, computation of the above estimators require fast evaluation of kernel matrix-vector products which are not supported by traditional Python convex optimization libraries. Instead we efficiently evaluate these summations on GPUs with the KeOps Python libraries (Charlier et al., 2021) and obtain derivatives with respect to $\mathbf{w}, \mathbf{W}$ which seamlessly integrate with the PyTorch automatic differentiation library (Paszke et al., 2019). Optimization of these weights is performed by Adam with amsgrad, a popular gradient descent method, which uses gradients from previous iterations to stabilize its convergence (Kingma and Ba, 2015; Reddi et al., 2018). As an additional benefit, Pytorch libraries offer a *parametrizations* class which allows us to easily constrain our RHKS estimators $\beta_{\mathbf{w}}, G_{\mathbf{W}}$ to satisfy various properties such as coordinate symmetries and time causality (see Section 5).

We train the estimators $\beta_{\mathbf{w}}, G_{\mathbf{W}}$ stochastically on batches of size 100 with $n = 100$-500 training pairs $(F_i, U_i) \in \mathbb{R}^{m_y}, \mathbb{R}^{m_x}$ using between 100-1000 epochs such that the solution converges. The loss function minimized by gradient descent on the training data is

$$\text{Loss}(\mathcal{T}_{\beta_{\mathbf{w}}, G_{\mathbf{W}}}) = \text{MSE}(\mathcal{T}_{\beta_{\mathbf{w}}, G_{\mathbf{W}}}) + \lambda P(\beta_{\mathbf{w}}) + \rho J(G_{\mathbf{W}}) \tag{38}$$

which is the mean squared L^2 error regularized by the RKHS norms of the Green's function and bias term $P(\beta) = \|\beta\|_Q$ and $J(G) = \|G\|_K$. The mean squared error above is approximated by the Riemann sum

$$\text{MSE}(\mathcal{T}) = \frac{1}{n}\sum_{i=1}^{n}\int_{D_{\mathcal{Y}}}\Big(U_i(y) - [\mathcal{T}(F_i)](y)\Big)^2 \mathrm{d}y \approx \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m_y}\Big(U_i(y_k) - [\mathcal{T}(F_i)](y_k)\Big)^2 \Delta_k^y \tag{39}$$

and the regularization terms can be exactly evaluated as

$$\begin{aligned} P(\beta_{\mathbf{w}}) &= \sum_{j=1}^{m_y}\sum_{l=1}^{m_y} w_j Q(y_j, y_l) w_l \Delta_j^y \Delta_l^y \\ J(G_{\mathbf{W}}) &= \sum_{i=1}^{m_x}\sum_{j=1}^{m_y}\sum_{k=1}^{m_x}\sum_{l=1}^{m_y} W_{ij} K(x_i, y_j, x_k, y_l) W_{kl} \Delta_i^x \Delta_j^y \Delta_k^x \Delta_l^y. \end{aligned} \tag{40}$$

Since the samples (F_i, U_i) generated can vary significantly in magnitude, to evaluate the performance of our estimator $\mathcal{T}_{\mathbf{w},\mathbf{W}}$ we also investigate its L^2 mean relative error which is computed by

$$\text{RE}(\mathcal{T}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{\Big\|U_i - \mathcal{T}(F_i)\Big\|_{L^2(D_{\mathcal{Y}})}}{\Big\|U_i\Big\|_{L^2(D_{\mathcal{Y}})}}} \approx \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{k=1}^{m_y}\Big(U_i(y_k) - [\mathcal{T}(F_i)](y_k)\Big)^2 \Delta_k^y}{\sum_{k=1}^{m_y} U_i(y_k)^2 \Delta_k^y}}. \tag{41}$$

Finally, given new meshes $\{\overline{x}_j\}_{j=1}^{\overline{m}_x}, \{\overline{y}_k\}_{k=1}^{\overline{m}_y}$ with quadrature weights $\overline{\Delta}_j^x, \overline{\Delta}_k^y$ we can extrapolate the predictions of $\mathcal{T}_{\mathbf{w},\mathbf{W}} : f \to u$ on these new meshes by writing

$$[\mathcal{T}_{\mathbf{w},\mathbf{W}}(f)](y) = [\mathcal{T}_{\beta_{\mathbf{w}}, G_{\mathbf{W}}}(f)](y) = \beta_{\mathbf{w}}(y) + \sum_{j=1}^{\overline{m}_x} G_{\mathbf{W}}(\overline{x}_j, y) f(\overline{x}_j) \overline{\Delta}_j^x. \tag{42}$$

The mean and relative squared errors on these new meshes can be computed in the same way as shown above.

In Section 5 we implement the Green's function and bias term estimators described above and learn the solution maps of various space and time-varying PDEs. In all examples, the domains $D_{\mathcal{X}}, D_{\mathcal{Y}}$ of the input and output data are rectangular domains of the form $\Pi[a_i, b_i]$ and the discretization meshes $\{x_j\}_{j=1}^{m_x}, \{y_k\}_{k=1}^{m_y}$ are equispaced with quadrature weights Δ_j^x, Δ_k^y defined by the trapezoid rule. Our approach however can easily be extended to nonuniform grids by setting suitable quadrature weights or using Monte Carlo integration methods such as importance sampling.

In examples where we know the true Green's function and bias term, we can also compute the relative error of our estimated Green's functions and bias terms G, β to the true functions $G_{\text{true}}, \beta_{\text{true}}$. In practice, we compute these relative errors as Riemann sums

$$\begin{aligned} \text{RE}(G) &= \sqrt{\frac{\left\|G - G_{\text{true}}\right\|_{L^2(D_{\mathcal{X}}\times D_{\mathcal{Y}})}}{\left\|G_{\text{true}}\right\|_{L^2(D_{\mathcal{X}}\times D_{\mathcal{Y}})}}} \approx \sqrt{\frac{\sum_{j=1}^{\overline{m}_x}\sum_{k=1}^{\overline{m}_y}\left(G(\overline{x}_j,\overline{y}_k) - G_{\text{true}}(\overline{x}_j,\overline{y}_k)\right)^2\overline{\Delta}_j^x\overline{\Delta}_k^y}{\sum_{j=1}^{\overline{m}_x}\sum_{k=1}^{\overline{m}_y} G_{\text{true}}(\overline{x}_j,\overline{y}_k)^2\overline{\Delta}_j^x\overline{\Delta}_k^y}} \\ \text{RE}(\beta) &= \sqrt{\frac{\left\|\beta - \beta_{\text{true}}\right\|_{L^2(D_{\mathcal{Y}})}}{\left\|\beta_{\text{true}}\right\|_{L^2(D_{\mathcal{Y}})}}} \approx \sqrt{\frac{\sum_{k=1}^{\overline{m}_y}\left(\beta(\overline{y}_k) - \beta_{\text{true}}(\overline{y}_k)\right)^2\overline{\Delta}_k^y}{\sum_{k=1}^{\overline{m}_y}\beta_{\text{true}}(\overline{y}_k)^2\overline{\Delta}_k^y}} \end{aligned} \tag{43}$$

where the uniform meshes $\{\overline{x}_j\}_{j=1}^{\overline{m}_x}, \{\overline{y}_k\}_{k=1}^{\overline{m}_y}$ are finely discretized with mesh sizes 10 times larger (in each dimension) than the meshes $\{x_j\}_{j=1}^{m_x}, \{y_k\}_{k=1}^{m_y}$ on which our estimators G and β were trained.

4. Error Analysis

In this section, we establish error bounds for the Green's function of a PDE when it is estimated in an RKHS from a finite number of samples. Throughout this section, we assume that the input-output data $\{(F_i, U_i)\}_{i=1}^n \subset L^2(D_{\mathcal{X}}) \times L^2(D_{\mathcal{Y}})$ are truly functional data and are not discretized on a finite mesh. Furthermore, we limit our theoretical analysis to the simpler case of only estimating the Green's function of a PDE. We assume that the bias term $\beta \in \mathcal{B}$ is known and hence, can be subtracted from the observations U, which significantly simplifies the notation in our analysis. For the interested reader, we briefly discuss in Section 4.1 below how our theoretical framework can be extended to the case where the bias term is also estimated.

Focusing on the Green's function estimation problem, we can define the empirical risk of our Green's function estimator as

$$\widehat{R}(G) := \frac{1}{n}\sum_{i=1}^{n}\left\|U_i - \int_{D_{\mathcal{X}}} G(x,\cdot)F_i(x)\mathrm{d}x\right\|_{L^2(D_{\mathcal{Y}})}^2. \tag{44}$$

the penalized empirical risk as

$$\widehat{R}_\lambda(G) := \widehat{R}(G) + \lambda J(G) \tag{45}$$

and, most importantly, the population risk as

$$R(G) := \mathbb{E}\Big[\Big\|U - \int_{D_{\mathcal{X}}} G(x,\cdot)F(x)\mathrm{d}x\Big\|_{L^2(D_{\mathcal{Y}})}^2\Big]. \tag{46}$$

where the expectation above is taken with respect to the randomness of (F, U). This last objective, the population risk, is the quantity of interest that we study in order to bound the prediction error of our Green's function estimator. In Sections 4.2-4.4 we show how the empirical RKHS estimator

$$\widehat{G}_{n,\lambda} := \underset{G\in\mathcal{G}}{\arg\min}\,\widehat{R}_\lambda(G) \tag{47}$$

compares to any oracle

$$G_{\mathcal{G}} \in \arg\min_{G \in \mathcal{G}} R(G) \tag{48}$$

We prove an oracle inequality in Section 4.4 that controls the difference $R(\widehat{G}_{n,\lambda}) - R(G_{\mathcal{G}})$. For appropriately chosen regularizer λ, this prediction error tends to zero as $n \to \infty$. We assume throughout that $F \in L^2(D_{\mathcal{X}}), U \in L^2(D_{\mathcal{Y}})$ are mean zero random variables.

As shorthand, we denote $G^T \in L^2(D_{\mathcal{Y}} \times D_{\mathcal{X}})$ as the function $G^T(x, y) = G(y, x)$ for all $x \in D_{\mathcal{X}}, y \in D_{\mathcal{Y}}$. We frequently use the notation $G^T(F) = \int_{D_{\mathcal{X}}} G(x, \cdot)F(x)\mathrm{d}x$ as the Green's function $G(x, y)$ integrated against $F(x)$ in its first coordinate.

4.1 Accommodating the Bias Term

The error analysis for $\widehat{G}_{n,\lambda}$ derived in the following sections can be generalized to include the bias term $\beta \in \mathcal{B}$. In this case, we need to bound the full population risk $R(\widehat{\beta}_{n,\rho,\lambda}, \widehat{G}_{n,\rho,\lambda})$ of the RKHS estimators

$$\widehat{\beta}_{n,\rho,\lambda}, \widehat{G}_{n,\rho,\lambda} := \arg\min_{\beta \in \mathcal{B}, G \in \mathcal{G}} \widehat{R}_{\rho,\lambda}(\beta, G). \tag{49}$$

Here the full population risk is defined as

$$R(\beta, G) := \mathbb{E}\left[\left\| U - \beta - \int_{D_{\mathcal{X}}} G(x, \cdot)F(x)\mathrm{d}x \right\|^2_{L^2(D_{\mathcal{Y}})}\right]. \tag{50}$$

Including the bias term, our full affine operator is

$$\mathcal{T}_{\beta,G}(f) = \beta + \int_{D_{\mathcal{X}}} G(x, \cdot)f(x)\mathrm{d}x. \tag{51}$$

Denote $|D_{\mathcal{X}}|$ by the Lebesgue measure of $D_{\mathcal{X}}$. By taking our input function $f : \mathbb{R} \to \mathbb{R}$ and appending a constant to obtain $\tilde{f} : \mathbb{R} \to \mathbb{R}^2$ where $\tilde{f}(x) = [f(x), \frac{1}{|D_{\mathcal{X}}|}]^T$ we can rewrite our affine operator as

$$\mathcal{T}_{\tilde{G}}(\tilde{f}) = \int_{D_{\mathcal{X}}} \langle \tilde{G}(x, \cdot), \tilde{f}(x) \rangle_2 \mathrm{d}x, \quad \tilde{G}(x, y) = [G(x, y), \beta(y)]^T \tag{52}$$

where now $\tilde{G} : D_{\mathcal{X}} \times D_{\mathcal{Y}} \to \mathbb{R}^2$ and $\langle \cdot, \cdot \rangle_2$ denotes the standard Euclidean vector inner product. Here, $\tilde{G}$ is an element of a new Cartesian product RKHS of vector-valued functions

$$\tilde{\mathcal{G}} := \left\{ [G(x, y), \beta(y)]^T : G \in \mathcal{G}, \beta \in \mathcal{B} \right\} \tag{53}$$

with inner product defined as

$$\langle \tilde{G}, \tilde{H} \rangle_{\tilde{\mathcal{G}}} = \langle \beta, \gamma \rangle_{\mathcal{B}} + \langle G, H \rangle_{\mathcal{G}} \tag{54}$$

for all $\tilde{G}(x, y) = [G(x, y), \beta(y)]^T$ and $\tilde{H}(x, y) = [H(x, y), \gamma(y)]^T$ in $\tilde{\mathcal{G}}$ where $G, H \in \mathcal{G}$ and $\beta, \gamma \in \mathcal{B}$. As defined in Micchelli and Pontil (2005, Section 2, Theorem 2.1), the reproducing kernel of this RKHS of vector-valued functions $\tilde{\mathcal{G}}$ is $\tilde{K} : (D_{\mathcal{X}} \times D_{\mathcal{Y}})^2 \to \mathbb{R}^{2\times 2}$ given by

$$\tilde{K}(x, y, \xi, \eta) = \begin{bmatrix} K(x, y, \xi, \eta) & 0 \\ 0 & Q(y, \eta) \end{bmatrix} \tag{55}$$

where, as before, $K : (D_{\mathcal{X}} \times D_{\mathcal{Y}})^2 \to \mathbb{R}$ and $Q : D_{\mathcal{Y}}^2 \to \mathbb{R}$ are the reproducing kernels of $\mathcal{G}$ and $\mathcal{B}$ respectively. Using these definitions, the error analysis for Green's functions with no bias term extends to the case when the Green's function and bias term are jointly optimized. In this more general setting, proving error bounds for $\widehat{\beta}_{n,\rho,\lambda}, \widehat{G}_{n\rho\lambda}$ similarly involves studying the eigenvalues and eigenvectors of $\tilde{K}, \tilde{\Sigma}_F$ to bound the population risk $R(\widehat{\beta}_{n,\rho,\lambda}, \widehat{G}_{n\rho\lambda})$. The covariance operator of the input data $\tilde{F} = [F, \frac{1}{|D_{\mathcal{X}}|}]^T$ now becomes a matrix-valued function $\tilde{\Sigma}_F : D_{\mathcal{X}}^2 \to \mathbb{R}^{2\times 2}$. Furthermore, the spectra and vector-valued eigenfunctions of $\tilde{K}$ become concatenations of the eigenvalues and eigenfunctions of the original reproducing kernels K and Q. In this paper, we choose to avoid these additional notational complexities by studying error bounds for the Green's function only.

4.2 Eigenbases of Mercer Kernels Σ_F and K

To derive an oracle inequality that bounds $R(\widehat{G}_{n,\lambda}) - R(G_{\mathcal{G}})$, we make the following assumptions on the covariance operator $\Sigma_F = \mathbb{E}[(F - \mathbb{E}[F]) \otimes (F - \mathbb{E}[F])]$ and the reproducing kernel K of $\mathcal{G}$.

Assumption 3 (Mercer Kernels) *We assume that the covariance operator $\Sigma_F \in L^2(D_{\mathcal{X}} \times D_{\mathcal{X}})$ and the reproducing kernel $K \in L^2((D_{\mathcal{X}} \times D_{\mathcal{Y}})^2)$ are continuous, square integrable, and positive definite. Kernels that satisfy these three conditions are called Mercer kernels. Since Σ_F is a Mercer kernel, we know by Mercer's theorem (Cucker and Smale, 2002, Section 2, Theorem 1) that it has nonnegative eigenvalues $\mu_1 \geq \mu_2 \geq \dots$ and L^2 orthonormal eigenfunctions $\{\phi_k\}_{k=1}^{\infty} \subseteq L^2(D_{\mathcal{X}})$ with the spectral decomposition*

$$\Sigma_F(x, \xi) = \sum_{k=1}^{\infty} \mu_k \phi_k(x) \phi_k(\xi). \tag{56}$$

Similarly, K is a Mercer kernel so it has nonnegative eigenvalues $\rho_1 \geq \rho_2 \geq \dots$ and L^2 orthonormal eigenfunctions $\{\Psi_k\}_{k=1}^{\infty} \subset L^2(D_{\mathcal{X}} \times D_{\mathcal{Y}})$ with the spectral decomposition

$$K(x, y, \xi, \eta) = \sum_{k=1}^{\infty} \rho_k \Psi_k(x, y) \Psi_k(\xi, \eta) \tag{57}$$

where $\langle \Psi_i, \Psi_j \rangle_{L^2(D_{\mathcal{X}} \times D_{\mathcal{Y}})} = \delta_{ij}$ and $\rho_i \langle \Psi_i, \Psi_j \rangle_{\mathcal{G}} = \delta_{ij}$.

The decay of the eigenvalues of the reproducing kernel K play a key role in the estimation of $\widehat{G}_{n,\lambda}$. The rate at which these eigenvalues decay to zero determines the rate at which $\widehat{G}_{n,\lambda}$ converges to $G_{\mathcal{G}}$ in terms of their prediction error.

Assumption 4 *The eigenvalues of the reproducing kernel K satisfy $\rho_k \lesssim k^{-r}$ for some $r > \frac{1}{2}$.*

In Appendix D, we provide several examples of reproducing kernels which have this rate of decay in their spectrum. For the RKHSs considered in the following sections, we focus our analysis on kernels with polynomial decay in their eigenvalues. The same proof technique in Sections 4.3 and 4.4 for deriving the error bounds can be applied to RKHSs whose kernels have a stricter, exponential decay in their eigenvalues such as smooth radial kernels.

4.3 Simultaneous Diagonalization

The variance of our RKHS estimator $\widehat{G}_{n,\lambda}$ is related to the quadratic form $\langle(\Sigma_F \otimes I)G, G\rangle_{L^2(D_\mathcal{X}\times D_\mathcal{Y})}$ where $\Sigma_F : D_\mathcal{X} \to D_\mathcal{X}$ is the covariance operator of the input data and $I : D_\mathcal{Y} \to D_\mathcal{Y}$ is the identity operator of the output data. Likewise, the bias of our estimator is determined by the regularization term $J(\widehat{G}_{n,\lambda}) = \|\widehat{G}_{n,\lambda}\|_K^2$ which implies that we need to study the spectrum of the reproducing kernel K of $\mathcal{G}$. These statements, formalized in Section 4.4, suggest an approach to studying the bias-variance tradeoff of our estimator. Namely, we approach this problem by simultaneously diagonalizing the operators $\Sigma_F \otimes I$ and K. This allows us to write any $G \in \mathcal{G}$ as a sum of basis functions where $\langle(\Sigma_F \otimes I)G, G\rangle_{L^2(D_\mathcal{X}\times D_\mathcal{Y})}$ and $J(G) = \|G\|_K^2$ are expanded into series whose terms depend on the basis coefficients of G.

First, for any $G, H \in L^2(D_\mathcal{X} \times D_\mathcal{Y})$ define the semi-inner product

$$\langle G, H\rangle_{\Sigma_F} = \langle(\Sigma_F \otimes I)G, H\rangle_{L^2(D_\mathcal{X}\times D_\mathcal{Y})} = \int_{D_\mathcal{X}}\int_{D_\mathcal{X}}\int_{D_\mathcal{Y}} G(x,y)\Sigma_F(x,z)H(z,y)\mathrm{d}x\mathrm{d}y\mathrm{d}z \quad (58)$$

Because K is a Mercer kernel, its can be expanded as in (57), and then by Cucker and Smale (2002, Chapter 3, Section 3) we can write

$$\langle G, F\rangle_\mathcal{G} = \langle G, F\rangle_K = \sum_{k=1}^{\infty} \frac{\langle G, \Psi_k\rangle_{L^2(D_\mathcal{X}\times D_\mathcal{Y})}\langle F, \Psi_k\rangle_{L^2(D_\mathcal{X}\times D_\mathcal{Y})}}{\rho_k}. \quad (59)$$

Define a sum of the two inner products

$$\langle G, H\rangle_{\overline{K}} = \langle G, H\rangle_{\Sigma_F} + \langle G, H\rangle_K \quad (60)$$

with corresponding norm

$$\|G\|_{\overline{K}}^2 = \|G\|_{\Sigma_F}^2 + \|G\|_K^2 = \|G\|_{\Sigma_F}^2 + J(G). \quad (61)$$

Note that $\langle\cdot,\cdot\rangle_{\overline{K}}$ is indeed an inner product over $\mathcal{G}$ because it is clearly linear and conjugate symmetric, and it is positive definite since $\|G\|_K^2 \le \|G\|_{\overline{K}}^2$.

Now we use this new norm $\|\cdot\|_{\overline{K}}$ to define a basis for $\mathcal{G}$ that simultaneously diagonalizes the quadratic forms $\|G\|_{\Sigma_F}^2$ and $J(G)$. First in Appendix F we prove the following proposition.

Proposition 1 *The norms $\|\cdot\|_\mathcal{G}$ and $\|\cdot\|_{\overline{K}}$ are equivalent. Hence, the Hilbert space $\mathcal{G}$ with inner product $\langle\cdot,\cdot\rangle_{\overline{K}}$ is also an RKHS.*

Let us denote $\overline{K} : (D_\mathcal{X} \times D_\mathcal{Y})^2 \to \mathbb{R}$ as the reproducing kernel associated with $\|\cdot\|_{\overline{K}}$. The kernel $\overline{K}$ can be viewed as a positive definite operator over the space $\mathcal{G}$. Denoting the eigenvalues and eigenfunctions of $\overline{K}$ by $\{(\rho_k', \Psi_k')\}_{k=1}^{\infty}$ we can interpret $\overline{K} : L^2(D_\mathcal{X} \times D_\mathcal{Y}) \to L^2(D_\mathcal{X} \times D_\mathcal{Y})$ as a positive operator defined as $\overline{K}(\Psi_k') := \int_{D_\mathcal{X}}\int_{D_\mathcal{Y}} \overline{K}(\cdot,\cdot,\xi,\eta)\Psi_k'(\xi,\eta)\mathrm{d}\xi\mathrm{d}\eta = \rho_k'\Psi_k'$. Now we can define the square root of this positive operator as $\overline{K}^{\frac{1}{2}} : L^2(D_\mathcal{X} \times D_\mathcal{Y}) \to L^2(D_\mathcal{X} \times D_\mathcal{Y})$ which satisfies $\overline{K}^{\frac{1}{2}}(\Psi_k') = (\rho_k')^{\frac{1}{2}}\Psi_k'$.

To summarize, we have defined above two norms $\|\cdot\|_{\Sigma_F}^2$ and $\|\cdot\|_{\overline{K}}^2$ where the former is the quadratic form of $\Sigma_F \otimes I$ and the latter is roughly the quadratic form of the inverse kernel $\overline{K}^{-1}$. For notational convenience, we have chosen to drop the tensor product with the identity and the kernel inverse from the norm subscripts.

Following the ideas of Horn and Johnson (2012, Section 7.6, Theorem 7.6.4) we simultaneously diagonalize the quadratic forms $\|\cdot\|^2_{\Sigma_F}$ and $\|\cdot\|^2_{\overline{K}}$. Defining the linear operator $\overline{K}^{\frac{1}{2}}(\Sigma_F \otimes I)\overline{K}^{\frac{1}{2}}$, denote its eigenvalues and L^2 orthonormal eigenfunctions by $\nu_1 \geq \nu_2 \geq \dots$ and $\{\Gamma_k\}_{k=1}^{\infty} \subset \mathcal{G}$ respectively. Note that $\overline{K}$ is positive definite over $\mathcal{G}$ and so is $\Sigma_F \otimes I$ by Assumption 1 which implies that $\{\Gamma_k\}_{k=1}^{\infty}$ spans $\mathcal{G}$ and that $\nu_k > 0$ for all $k \geq 1$. In fact, we have that

$$\nu_k = \|\overline{K}^{\frac{1}{2}}\Gamma_k\|^2_{\Sigma_F} \leq \|\overline{K}^{\frac{1}{2}}\Gamma_k\|^2_{\overline{K}} = \|\Gamma_k\|^2_{L^2(D_{\mathcal{X}}\times D_{\mathcal{Y}})} = 1. \tag{62}$$

Now we define the basis functions

$$\Omega_k = \nu_k^{-\frac{1}{2}}\overline{K}^{\frac{1}{2}}\Gamma_k \in \mathcal{G}, \quad k \geq 1. \tag{63}$$

Then using the induced inner product $\langle\cdot,\cdot\rangle_{\overline{K}}$ we can write out

$$\langle\Omega_i, \Omega_j\rangle_{\overline{K}} = \langle\nu_i^{-\frac{1}{2}}\overline{K}^{\frac{1}{2}}\Gamma_i, \nu_j^{-\frac{1}{2}}\overline{K}^{\frac{1}{2}}\Gamma_j\rangle_{\overline{K}} = \langle\nu_i^{-\frac{1}{2}}\Gamma_i, \nu_j^{-\frac{1}{2}}\Gamma_j\rangle_{L^2(D_{\mathcal{X}}\times D_{\mathcal{Y}})} = \nu_i^{-1}\delta_{ij} \tag{64}$$

and similarly

$$\begin{aligned}\langle(\Sigma_F \otimes I)\Omega_i, \Omega_j\rangle_{L^2(D_{\mathcal{X}}\times D_{\mathcal{Y}})} &= \langle\nu_i^{-\frac{1}{2}}(\Sigma_F \otimes I)\overline{K}^{\frac{1}{2}}\Gamma_i, \nu_j^{-\frac{1}{2}}\overline{K}^{\frac{1}{2}}\Gamma_j\rangle_{L^2(D_{\mathcal{X}}\times D_{\mathcal{Y}})} \\ &= \nu_i^{-\frac{1}{2}}\nu_j^{-\frac{1}{2}}\langle\overline{K}^{\frac{1}{2}}(\Sigma_F \otimes I)\overline{K}^{\frac{1}{2}}\Gamma_i, \Gamma_j\rangle_{L^2(D_{\mathcal{X}}\times D_{\mathcal{Y}})} = \delta_{ij}.\end{aligned} \tag{65}$$

We emphasize that the basis $\{\Omega_k\}_{k=1}^{\infty}$ defined above is not an orthogonal basis in $L^2(D_{\mathcal{X}} \times D_{\mathcal{Y}})$ but it does form an orthogonal basis for $\mathcal{G}$ as we shall show next. Furthermore, we can simultaneously diagonalize the quadratic forms $\|\cdot\|^2_{\Sigma_F}$ and $\|\cdot\|^2_{\overline{K}}$ on the basis $\{\Omega_k\}_{k=1}^{\infty}$.

Theorem 5 *For any $G \in \mathcal{G}$,*

$$G = \sum_{k=1}^{\infty} g_k\Omega_k \tag{66}$$

which converges absolutely and where $g_k = \nu_k\langle G, \Omega_k\rangle_{\overline{K}}$. Furthermore, setting $\gamma_k = (\nu_k^{-1} - 1)^{-1}$ then we can write

$$\|G\|^2_{\overline{K}} = \sum_{k=1}^{\infty}\nu_k^{-1}g_k^2 = \sum_{k=1}^{\infty}(1 + \gamma_k^{-1})g_k^2, \qquad \|G\|^2_{\Sigma_F} = \sum_{k=1}^{\infty} g_k^2 \tag{67}$$

and similarly

$$J(G) = \|G\|^2_{\overline{K}} - \|G\|^2_{\Sigma_F} = \sum_{k=1}^{\infty}\gamma_k^{-1}g_k^2. \tag{68}$$

We do not necessarily assume that the γ_k coefficients are ordered in decreasing order since the series in Theorem 5 converge absolutely. To acquire some intuition for this simultaneous diagonalization we discuss it in a useful setting where the operators K and $\Sigma_F \otimes I$ commute.

Proposition 2 *Recall that the eigenvalues and eigenbasis of Σ_F are $\mu_i, \phi_i \in L^2(D_\mathcal{X})$ for $i \geq 1$. Assume that the eigenbasis for K is $\{\Psi_{ij} := \phi_i \otimes \varphi_j\}_{k=1}^\infty \subseteq \mathcal{G}$ with eigenvalues $\{\rho_{ij}\}_{i,j=1}^\infty$ where $\{\varphi_j\}_{j=1}^\infty$ is any orthonormal basis of $L^2(D_\mathcal{Y})$. Then we know that $\{\gamma_{ij} := \mu_i \rho_{ij}\}_{i,j=1}^\infty$ and $\{\Omega_{ij} := \mu_i^{-\frac{1}{2}} \Psi_{ij}\}_{i,j=1}^\infty$ are the coefficients and basis functions given in Theorem 5.*

Sort the eigenvalues ρ_{ij} of K in decreasing order as $\rho_1 \geq \rho_2 \geq \ldots$ and assume that $\rho_k \lesssim k^{-r}$ for some $r > \frac{1}{2}$ as in Assumption 4. If we enumerate the coefficients γ_k and basis functions Ω_k in the same order as the decreasing ρ_k, then it holds that $\gamma_k \lesssim \rho_k \lesssim k^{-r}$.

We prove the proposition above in Appendix F. As an example, the assumptions in this proposition hold when $D_\mathcal{X} = D_\mathcal{Y} = [0, 1]$ and F is a Brownian bridge with variance σ^2. In this case, $\Sigma_F(x, \xi) = \sigma^2[\min(x, \xi) - x\xi]$ for any $x, \xi \in [0, 1]$ with eigenvalues $\mu_i = \frac{\sigma^2}{\pi i}$ and eigenfunctions $\phi_i(x) = \sqrt{2}\sin(\pi i x)$. If we choose $\mathcal{G} = W_1^2([0, 1]^2)$ to be the space of Sobolev-1 Green's functions on $D_\mathcal{X} \times D_\mathcal{Y} = [0, 1]^2$ with Dirichlet boundary conditions at $\partial(D_\mathcal{X} \times D_\mathcal{Y})$, then as described in Section 2.1 it has reproducing kernel

$$K(x, y, \xi, \eta) = 2 \sum_{i=1}^\infty \sum_{j=1}^\infty \frac{\sin(\pi i x)\sin(\pi j y)\sin(\pi i \xi)\sin(\pi j \eta)}{\pi^2(i^2 + j^2)} \tag{69}$$

whose eigenvalues are $\rho_{ij} = \frac{1}{2\pi^2(i^2+j^2)}$ with eigenfunctions $\Psi_{ij}(x, y) = \sqrt{2}\sin(\pi i x) \cdot \sqrt{2}\sin(\pi j y)$. Since we know that $\Psi_{ij} = \phi_i \otimes \phi_j$ then the assumptions of Proposition 2 are satisfied. Because K is a bounded Mercer kernel on $[0, 1]^2$ then from Example 1 of Appendix D we can check that $\rho_k \lesssim k^{-1}$. Finally, by applying Proposition 2 this tells us that $\gamma_k \lesssim k^{-1}$.

In order to prove an oracle inequality for $\widehat{G}_{n,\lambda}$, we must require the coefficients of the simultaneous diagonalization γ_k to decrease at a certain rate. Here we take inspiration from the setting discussed in Proposition 2 and make the following assumption on the coefficients γ_k.

Assumption 5 *When simultaneously diagonalizing $\|\cdot\|_{\overline{K}}$ and $\|\cdot\|_{\Sigma_F}$ in Theorem 5, we assume that $\gamma_k \lesssim k^{-r}$ for some $r > \frac{1}{2}$.*

4.4 Oracle Inequality for $\widehat{G}_{n,\lambda}$

Now we are in the setting to prove the oracle inequality which bounds the difference of the expected risks $R(\widehat{G}_{n,\lambda}) - R(G_\mathcal{G})$ between our Green's function estimator and the oracle. Here we again use the notation

$$G^T(F) = \int_{D_\mathcal{X}} G(x, y)F(x)\mathrm{d}x \tag{70}$$

to denote a Green's function $G \in L^2(D_\mathcal{X} \times D_\mathcal{Y})$ integrated against an input $F \in L^2(D_\mathcal{X})$.

By definition of the oracle $G_\mathcal{G} \in \arg\min_{G \in \mathcal{G}} R(G)$, we know that $\mathbb{E}\|U - G(F)\|^2_{L^2(D_\mathcal{Y})}$ for all $G \in \mathcal{G}$ is minimized at $G = G_\mathcal{G}$. Hence, for any $G \in \mathcal{G}$ by the optimality of $G_\mathcal{G}$ and the Pythagorean theorem we see that

$$\begin{aligned} R(G) &= \mathbb{E}\|U - G^T(F)\|^2_{L^2(D_\mathcal{Y})} \\ &= \mathbb{E}\|U - G^T_\mathcal{G}(F)\|^2_{L^2(D_\mathcal{Y})} + \mathbb{E}\|G^T(F) - G^T_\mathcal{G}(F)\|^2_{L^2(D_\mathcal{Y})} \\ &= R(G_\mathcal{G}) + \|G - G_\mathcal{G}\|^2_{\Sigma_F} \end{aligned} \tag{71}$$

In particular, setting $G = \widehat{G}_{n,\lambda}$ in the equation above, we need to bound the norm $\|\widehat{G}_{n,\lambda} - G_{\mathcal{G}}\|_{\Sigma_F}^2$. First we define the intermediate oracle

$$\overline{G}_{\infty,\lambda} = \operatorname*{arg\,min}_{G \in \mathcal{G}} \Big\{ R(G) + \lambda J(G) \Big\} \tag{72}$$

which is unique since $\|G - G_{\mathcal{G}}\|_{\Sigma_F}$ is strictly convex because $\Sigma_F \otimes I$ is positive definite over $\mathcal{G}$. Then by the Cauchy–Schwarz inequality we can decompose

$$\|\widehat{G}_{n,\lambda} - G_{\mathcal{G}}\|_{\Sigma_F}^2 \lesssim \underbrace{\|\widehat{G}_{n,\lambda} - \overline{G}_{\infty,\lambda}\|_{\Sigma_F}^2}_{\text{stochastic error}} + \underbrace{\|\overline{G}_{\infty,\lambda} - G_{\mathcal{G}}\|_{\Sigma_F}^2}_{\text{deterministic error}}. \tag{73}$$

We bound the deterministic error simply by writing

$$\|\overline{G}_{\infty,\lambda} - G_{\mathcal{G}}\|_{\Sigma_F}^2 = R(\overline{G}_{\infty,\lambda}) - R(G_{\mathcal{G}}) \tag{74}$$

Since we know that $R(\overline{G}_{\infty,\lambda}) + \lambda J(\overline{G}_{\infty,\lambda}) \leq R(G_{\mathcal{G}}) + \lambda J(G_{\mathcal{G}})$ then this proves that

$$\|\overline{G}_{\infty,\lambda} - G_{\mathcal{G}}\|_{\Sigma_F}^2 \leq \lambda\Big(J(G_{\mathcal{G}}) - J(\overline{G}_{\infty,\lambda})\Big) \leq \lambda J(G_{\mathcal{G}}). \tag{75}$$

For the stochastic error term, it takes a bit more work to show the following bound.

Lemma 6 *If* $\frac{\log(1/\delta)}{n}\lambda^{-\frac{1}{r}} \lesssim 1$*, then for the estimator* $\widehat{G}_{n,\lambda}$ *we have that*

$$\|\widehat{G}_{n,\lambda} - \overline{G}_{\infty,\lambda}\|_{\Sigma_F}^2 \lesssim \max\Big(1, \|G_{\mathcal{G}}\|_{op}, \lambda J(G_{\mathcal{G}})\Big)\frac{\log(1/\delta)}{n}\lambda^{-\frac{1}{r}} \tag{76}$$

with probability at least $1 - \delta$.

Combining (73), (75) and Lemma 6 we get

$$\begin{aligned} \|\widehat{G}_{n,\lambda} - G_{\mathcal{G}}\|_{\Sigma_F}^2 &\lesssim \max\Big(1, \|G_{\mathcal{G}}\|_{\mathrm{op}}, \lambda J(G_{\mathcal{G}})\Big)\frac{\log(1/\delta)}{n}\lambda^{-\frac{1}{r}} + \lambda J(G_{\mathcal{G}}) \\ &\lesssim \max\Big(1, \|G_{\mathcal{G}}\|_{\mathrm{op}}\Big)\frac{\log(1/\delta)}{n}\lambda^{-\frac{1}{r}} + \lambda J(G_{\mathcal{G}}) \end{aligned} \tag{77}$$

since we assumed that $\frac{\log(1/\delta)}{n}\lambda^{-\frac{1}{r}} \lesssim 1$. Taking $\lambda \asymp (\frac{n}{\log(1/\delta)})^{-\frac{r}{r+1}}$ yields

$$\begin{aligned} \|\widehat{G}_{n,\lambda} - G_{\mathcal{G}}\|_{\Sigma_F}^2 &\lesssim \max\Big(1, \|G_{\mathcal{G}}\|_{\mathrm{op}}, J(G_{\mathcal{G}})\Big)\Big(\frac{n}{\log(1/\delta)}\Big)^{-\frac{r}{r+1}} \\ &\lesssim \max\Big(1, \|G_{\mathcal{G}}\|_{\mathrm{op}}, J(G_{\mathcal{G}})\Big)n^{-\frac{r}{r+1}}\log(1/\delta) \end{aligned} \tag{78}$$

with probability at least $1 - \delta$. Therefore, we have shown the following oracle inequality.

Theorem 7 (Oracle Inequality) *The regularized RKHS estimator* $\widehat{G}_{n,\lambda}$ *from Theorem 4 with* $\lambda \asymp \Big(\frac{n}{\log(1/\delta)}\Big)^{-\frac{r}{r+1}}$ *satisfies the oracle inequality*

$$R(\widehat{G}_{n,\lambda}) \leq R(G_{\mathcal{G}}) + C\max\Big(1, \|G_{\mathcal{G}}\|_{op}, J(G_{\mathcal{G}})\Big)n^{-\frac{r}{r+1}}\log(1/\delta) \tag{79}$$

with probability at least $1 - \delta$ *for some numerical constant* $C > 0$.

As a final note, if the reproducing kernel K of $\mathcal{G}$ is squared exponential or generally a smooth radial kernel, its eigenvalues decay exponentially (Belkin, 2018, Section 3, Theorem 2). This implies that $\gamma_k \lesssim \rho_k \lesssim k^{-r}$ for all $r > 0$. Hence, the optimal choice of regularizer is $\lambda \asymp \frac{\log(1/\delta)}{n}$ and the prediction error in Theorem 7 above has rate $R(\widehat{G}_{n,\lambda}) - R(G_{\mathcal{G}}) \lesssim \frac{\log(1/\delta)}{n}$.

5. Examples

In this section, we show how the Green's functions several linear PDEs can be learned in an RKHS using the estimators defined in Section 3. On several examples, we show how RKHSs can be designed to enforce important physical constraints such as coordinate symmetries, time causality, and time invariance in these Green's function estimators.

5.1 Poisson Equation

We begin with the one-dimensional Poisson equation

$$-\Delta u(x) = f(x) \text{ on } D = [0,1], \qquad u(0) = -0.1,\ u(1) = 0.1 \tag{80}$$

with Dirichlet boundary conditions. The input and output domains of $f(x)$ and $u(y)$ respectively are $D_{\mathcal{X}} = D_{\mathcal{Y}} = [0,1]$. The random input forcings $f(x)$ are generated from a squared exponential KLE with lengthscale $\ell = 0.01$ (see Appendix A for details) and the corresponding solutions $u(y)$ are simulated with a finite difference solver. All input and output functions are discretized on a uniform $m_x = m_y = 100$ point grid on the unit interval. From this procedure we build up $n = 500$ input-output pairs $\{(f_i, u_i)\}_{i=1}^n$ on which we learn the Green's function $G(x, y)$ and bias term $\beta(y)$ of the Poisson equation. Here we do not add any noise to our data as we are interested in perfectly recovering the true Green's function and bias term of our PDE. In this example, we know the true form of the bias term is $\beta_{\text{Poisson}}(y) = 0.2x - 0.1$ and the Green's function is $G_{\text{Poisson}}(x, y)$ as given in (4).

In Figure 2 we compute the relative error of our learned Green's function and bias term estimators from (43), as well as their combined relative error on the train data from (41), and we study how these errors behave as a function of training epochs in our optimization (top row). Each line color in the top row represents a choice of kernel $K(x, y, \xi, \eta)$ and $Q(y, \eta)$ which defines the RKHS $\mathcal{G}$ for our Green's function and the RKHS $\mathcal{B}$ for our bias term respectively. We take K and Q to be of the same type for each line plot (both exponential, both Matérn, etc.) and refer the reader to Section 2.1 for the definitions of these kernels. The kernel lengthscales of K and Q in the x and y directions are set to $\sigma_x = \sigma_y = 0.2$ (see Appendix A.3 for details). In the second and third rows, we plot the absolute difference of the learned estimators (after 500 training epochs) compared to the true Green's function and bias term of the Poisson equation.

From these results, we find that all estimators of the Green's function incur the most error around the diagonal $x = y$ where the function is not smooth. The exponential RKHS estimator gives the best approximation near the diagonal but suffers large approximation errors away from the diagonal where the Green's function is very smooth. The smoother squared exponential and Matérn 5/2 kernels give an improved fit away from the diagonal. Finally, the Matérn 3/2 gives the best estimator of the Green's function as it balances the degree of smoothness correctly and is able to nicely approximate the function both near and away from the diagonal $x = y$. All RKHS kernel estimators of the bias term provide a reasonable fit to the true linear bias term with expected ringing phenomena near the boundaries of the domain.

5.2 Helmholtz Equation and Coordinate Symmetries

Now we study the one-dimensional Helmholtz equation

$$-\Delta u(x) - \omega^2 u(x) = f(x) \text{ on } D = [0,1], \qquad u(0) = -0.1,\ u(1) = 0.1 \tag{81}$$

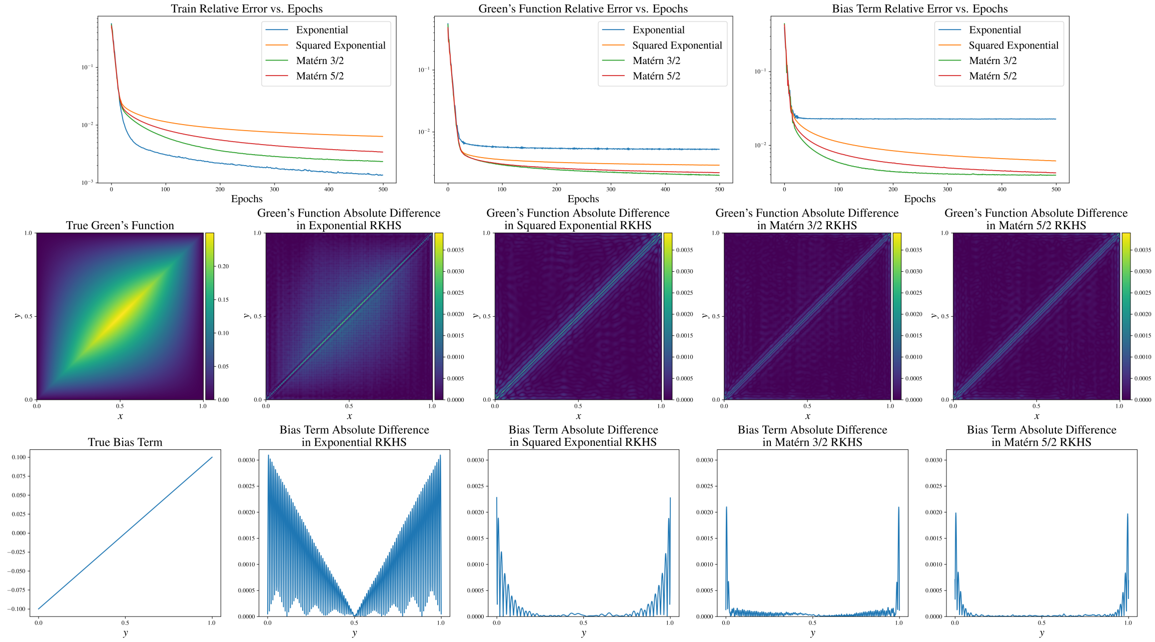


Figure 2: Learning the Green's function and bias term of the Poisson equation in exponential, squared exponential, and Matérn RKHSs. Top row shows the relative error of the Green's function and bias term estimators as well as their combined relative error on the train data as a function of training epochs. Middle and bottom row show absolute differences of learned Green's functions and bias terms in each RKHS.

with Dirichlet boundary conditions and high wavenumber $\omega = 20$. Here the input and output domains for $f(x)$ and $u(y)$ are simply the unit interval $D_{\mathcal{X}} = D_{\mathcal{Y}} = [0, 1]$. The input forcings $f(x)$ to the Helmholtz equation are generated from a squared exponential KLE with lengthscale $\ell = 0.01$ and the corresponding solutions $u(y)$ are simulated with a finite difference solver and corrupted with a fixed amount of i.i.d additive Gaussian noise as in the previous example. We are interested in learning the Green's function and bias term of this PDE, where we know their closed form expressions to be

$$\begin{aligned} G_{\text{Helmholtz}}(x, y) &= 2 \sum_{k=1}^{\infty} \frac{\sin(\pi k x)\sin(\pi k y)}{\pi^2 k^2 - \omega^2} \\ \beta_{\text{Helmholtz}}(y) &= 0.1 \frac{(1 + \cos(\omega))}{\sin(\omega)} \sin(\omega y) - 0.1 \cos(\omega y). \end{aligned} \tag{82}$$

Note that the true Green's function of the Helmholtz equation is coordinate symmetric $G(x, y) = G(y, x)$ due to the self-adjointness of the differential operator $-(\Delta + \omega^2)$.

In Figure 3 we study the robustness of our learning method to the number of samples, mesh discretization, and noise corruption in the Helmholtz equation. In these experiments, we choose as a reference $n = 100$ input-output samples, set a uniform mesh discretization at $m = m_x = m_y = 100$ points, and begin with no noise corruption. Then fixing two of the three parameters to their reference values, we vary one-at-a-time the number of input samples from $n = 1$ to 100, the mesh discretization from $m = 1$ to 100, and the noise corruption from $p = 0\%$ to 50%. We set the regularization parameter to $\lambda = 10^{-5}$ except for the noise experiments where we set $\lambda = 10^{-3}$ to achieve better noise robustness. For all experiments, our estimators for the Green's function $G(x, y)$ and bias term $\beta(y)$ are learned in a Matérn 5/2 RKHS with kernels $K(x, y, \xi, \eta)$ and $Q(y, \eta)$ defined as in Section 2.1 with lengthscales $\sigma_x = \sigma_y = 0.02$.

In Figure 3 we compute the relative errors of our Green's function and bias term estimators compared to their true functional form as defined in (43). We see that the relative errors (shown in blue) decrease exponentially as a function of the number of input-ouput samples (n), and their mesh discretization/measurements (m). Most importantly, in the third column we show that our estimators (blue lines) are noise robust and scale linearly with the amount of noise present in the train data.

We perform an identical set of experiments where we explicitly enforce the coordinate symmetry $G(x, y) = G(y, x)$ in our Green's function estimator due to the self-adjointness of the differential operator in the Helmholtz equation. As described in Appendix G.1, this is done by transforming the Matérn 5/2 reproducing kernel $K(x, y, \xi, \eta)$ of our Green's function estimator into

$$K_{\text{symm}}(x, y, \xi, \eta) = \frac{1}{4}\Big[K(x, y, \xi, \eta) + K(x, y, \eta, \xi) + K(y, x, \xi, \eta) + K(y, x, \eta, \xi)\Big]. \tag{83}$$

The Green's function G of the RKHS defined by this symmetrized kernel will necessarily be coordinate symmetric. As shown in the first column of Figure 3, enforcing this coordinate symmetry into our Green's function leads to faster convergence of both the Green's function and bias term estimators when the number of samples are increased (orange lines). Furthermore, symmetrizing the Green's function significantly improves its robustness to noise as shown in the top rightmost plot.

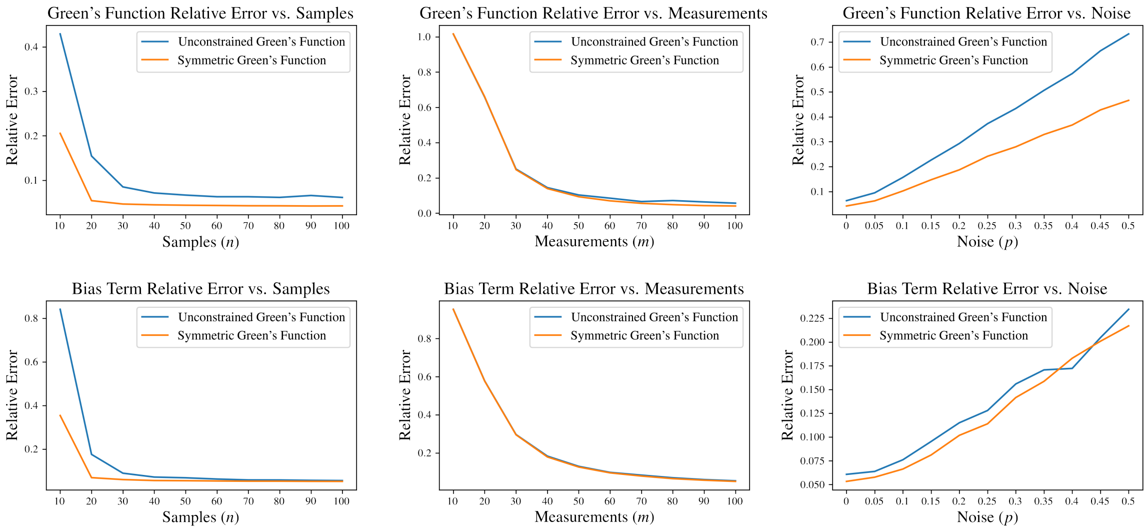


Figure 3: Plots of the relative errors of the Green's function and bias term estimators on the Helmholtz equation as the number of sample (n), measurements (m), and noise corruption (p) are increased. Two sets of identical experiments are performed where in the first experiment the Green's function is learned in a Matérn 5/2 RKHS with no constraints, and in a second experiment in a symmetrized Matérn 5/2 RKHS such that the coordinate symmetries $G(x, y) = G(y, x)$ are enforced.

5.3 Schrödinger Equation

We now study the time-independent Schrödinger equation in two spatial dimensions

$$\Delta u(x_1, x_2) - V(x_1, x_2)u(x_1, x_2) = 0 \text{ on } D = [0,1]^2, \qquad u(x_1, x_2) = b(x_1, x_2) \text{ on } \partial D \tag{84}$$

where we are interested in learning the map from the boundary condition $b(x_1, x_2)$ on domain $D_{\mathcal{X}} = \partial D$ to the solution $u(y_1, y_2)$ on domain $D_{\mathcal{Y}} = D$. Instead of defining the boundary conditions $b(x_1, x_2)$ as a function on the unit square D, we parametrize them as a function of arc length $b(x)$ for $x \in [0, 4]$ clockwise along the boundary of the unit square ∂D. Random input boundary conditions are generated by a KLE with squared exponential *periodic* kernel of lengthscale $\ell = 0.01$ and the corresponding solutions $u(y)$ are simulated with a finite difference solver and corrupted with 20% i.i.d additive Gaussian noise (see Appendix A). The output solutions u are discretized at $m_y = m = 50$ uniform grid points in both the y_1 and y_2 dimensions. Since the boundary conditions $b(x)$ wrap around the unit square, we discretize them correspondingly at $m_x = 4m - 3 = 197$ uniform grid points.

The potential function $V(x_1, x_2)$ for this example depicted in Figure 4 (top left) is chosen as a step function which is positive in a hexagonal-shaped well and zero outside. In Figure 4 (top center and right) we learn the Green's function $\widehat{G}(x, y_1, y_2) : \partial D \to D$ mapping the boundary condition $b(x)$ of the PDE to the solution $u(y_1, y_2)$ in a Matérn 5/2 RKHS with kernel

$$K(x, y_1, y_2, \xi, \eta_1, \eta_2) = C_\nu\Big(\sqrt{\frac{(x-\xi)^2}{\sigma_x^2} + \frac{(y_1-\eta_1)^2}{\sigma_y^2} + \frac{(y_2-\eta_2)^2}{\sigma_y^2}}\Big). \tag{85}$$

for lengthscales $\sigma_x = \sigma_y = 0.02$. Here $\nu = 5/2$ where C_ν is the Matérn covariance function defined in (19) of Section 2.1. We train on 500 noisy samples (b_i, u_i) and regularize our estimator with $\lambda = 10^{-4}$ penalty which allows us to learn a smooth Green's function even with 20% noise in our output samples.

In the top center plot of Figure 4 we see that the estimator has learned an impulse response from perturbing the boundary condition at a given point. By integrating the learned Green's function along the boundary ∂D of the unit square (top right), we clearly see the hexagonal shape of the potential $V(x_1, x_2)$ implying that solutions of the PDE have smaller magnitude in this hexagonal region as expected from the form of the PDE. In the bottom of Figure 4 we study how our learned Green's function estimator performs on 500 new test samples when we vary the lengthscale ℓ of the boundary condition from 0.01 to 10.0 and the mesh discretization from $m = 50$ to 150. We remind the reader that the mesh discretization in the y_1 and y_2 directions both scale as $m_y = m$ and the mesh discretization of the boundary condition scales as $m_x = 4m - 3$. From the botom plot of Figure 4 we see that our learned Green's function is independent of mesh discretization as the relative test error plateaus quickly as we increase m. Furthermore, the relative error of our estimator only decreases as we raise the boundary condition lengthscale from $\ell = 0.1$ on which it was trained to $\ell = 10.0$ (e.g. very smooth boundary conditions). When we evaluate our Green's function estimator on test boundary conditions with lengthscale below $\ell = 0.1$ (on which it was trained), then the predictive ability of our estimator gradually worsens. This behavior is expected as the Green's function estimator cannot make perfect predictions on inputs which exceed the lengthscale of the boundary conditions it was trained. We include example predictions of our Green's function estimator for input boundary conditions of several lengthscales in Appendix A.4

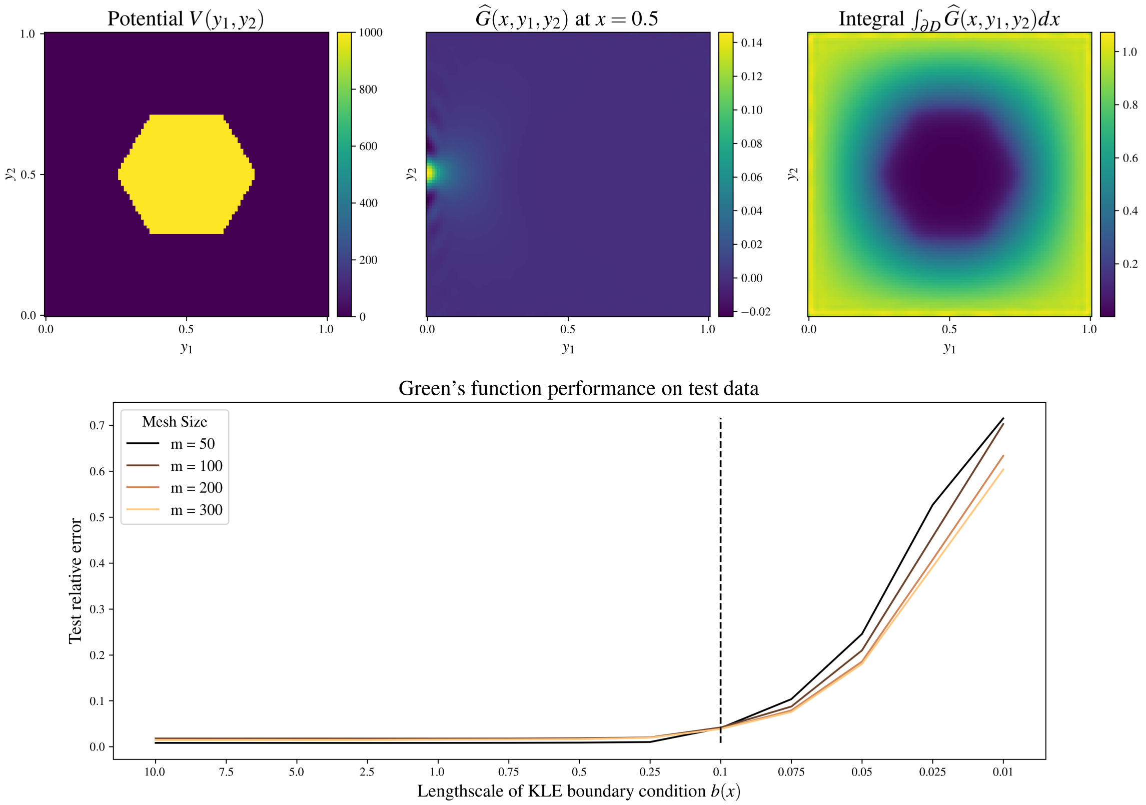


Figure 4: Learning the Schrödinger equation Green's function $\widehat{G} : b \mapsto u$ from Dirichlet boundary conditions on the boundary of the unit square to the solution in the interior. The Schrödinger equation has a hexagonal step-function potential shown in the top left plot. Here we estimate $\widehat{G}$ in a Matérn 5/2 RKHS and several views of the learned Green's function are shown in the top center and right plots. In the bottom row we plot the test error of our estimator and show that with no additional training, it can simulate the PDE on finer grids and generalizes to input boundary conditions with varying lengthscales. Black dashed line indicates the lengthscale of the boundary conditions b which $\widehat{G}$ was trained on.

5.4 Fokker–Planck Equation

Now we show how to estimate the Green's function or fundamental solution of the *Fokker–Planck equation* (FPE) which is a generalization of the diffusion (heat) equation and describes the evolution of a distribution of Brownian particles in a potential $V(x)$. Each particle's position X_t is governed by a stochastic differential equation (SDE) of the form

$$dX_t = -\frac{dV}{dx}(X_t)dt + \sqrt{2\alpha}dW_t \tag{86}$$

where the *diffusivity* $\alpha > 0$ is a constant and W_t is a Wiener process. Assuming that the particles at time $t = 0$ are distributed by $X_0 \sim u_0$ then the Fokker–Planck equation for the probability density $u(x, t)$ is

$$\frac{\partial}{\partial t}u(x,t) = \frac{\partial}{\partial x}\left[\frac{dV}{dx}(x)u(x,t)\right] + \alpha\frac{\partial^2}{\partial x^2}u(x,t), \quad u(x,0) = u_0(x). \tag{87}$$

Defining the *probability flux*

$$j(x,t) = \frac{dV}{dx}(x)u(x,t) + \alpha \frac{\partial}{\partial x} u(x,t) \tag{88}$$

we can rewrite the FPE above as

$$\frac{\partial}{\partial t} u(x,t) = \frac{\partial}{\partial x} j(x,t), \quad u(x,0) = u_0(x). \tag{89}$$

Now take the space-time domain $(x,t) \in D = [a,b] \times \mathbb{R}_+$ where the spatial domain is a finite interval. In order for the solution of (87) to be well-specified we impose *reflecting* boundary conditions $j(a,t) = j(b,t) = 0$. This enforces that particles which reach the boundary are reflected back into the domain such that no mass leaves the domain (i.e. probability flux is zero).

In the following example, we simulate (87) on the domain $D = [-2,2] \times [0,1]$ with the potential function

$$V(x) = x^4 - 3x^2. \tag{90}$$

We learn a *fundamental solution* $\widehat{G}(x,y,t)$ that maps the initial distribution $u_0(x)$ on the domain $D_{\mathcal{X}} = [-2,2]$ to the distribution at all future times $u(y,t)$ on the domain $D_{\mathcal{Y}} = D$. Here we learn our Green's function in a Matérn 3/2 RKHS with the kernel

$$K(x,y,t,\xi,\eta,\tau) = C_\nu\Big(\sqrt{\frac{(x-\xi)^2}{\sigma_x^2} + \frac{(y-\eta)^2}{\sigma_y^2} + \frac{(t-\tau)^2}{\sigma_t^2}}\Big) \tag{91}$$

for $\nu = 3/2$ where $\sigma_x = \sigma_y = 0.16$ and $\sigma_t = 0.04$. Here C_ν is the Matérn covariance function defined in (19) of Section 2.1.

Our Green's function estimator is trained on 500 samples where the inputs $(u_0)_i$ are generated from a Gaussian process KLE with a squared exponential kernel of lengthscale $\ell = 0.1$. For each initial condition $(u_0)_i$, we simulate the output solutions u_i by a matrix numerical method (Holubec et al., 2019) and corrupt our outputs with 20% additive Gaussian noise. In the training data, the input initial conditions are discretized on $m_x = m = 50$ uniform grid points and the output solutions are discretized on $m_y \times m_t$ grid points where $m_y = m_t = m = 50$. The estimator is trained to convergence for 100 epochs with a $\lambda = 10^{-5}$ penalty on its RKHS norm.

Figure 5 shows cross-sections of our learned Green's function $\widehat{G}(x,y,t)$ at several timepoints t. From the learned Green's function we extract important features of the Fokker–Planck dynamics. At $t = 0$, the Green's function learns a map from $u_0(x) \mapsto u(y,0)$ which is as close as possible to a delta function, limited only by the fixed lengthscale σ_x, σ_y of our RKHS kernel. As time t increases, the Green's function maps all the mass in $u_0(x)$ for $x > 0$ and $x < 0$ near the points ± 1.225 respectively which correspond to the basins of the potential $V(x)$. As expected, our Green's function has learned that movement of mass for the FPE tends to the basins of the potential function.

The second row of Figure 5 shows the relative test error of our estimator when it is evaluated on 500 new test samples generated from a different distribution than the train data. For each test data set, we generate the initial conditions from a KLE with squared exponential kernel and vary the lengthscale of this process from $\ell = 0.01$ to 10.0. We also vary the mesh discretization of the input and output samples jointly from $m_x = m_y = m_t = m = 50$ to 150. We observe that the relative test error drops as we raise the lengthscale of the initial condition u_0 from $\ell = 0.1$ to 10.0 but, as expected, increases if the lengthscale becomes smaller than $\ell = 0.1$ on which our estimator was trained. Additionally, we observe that our Green's function is insensitive to the mesh discretization and quickly plateaus as the mesh size m is increased. Example predictions of our Green's function estimator on initial conditions with varied lengthscales are shown in Appendix A.4.

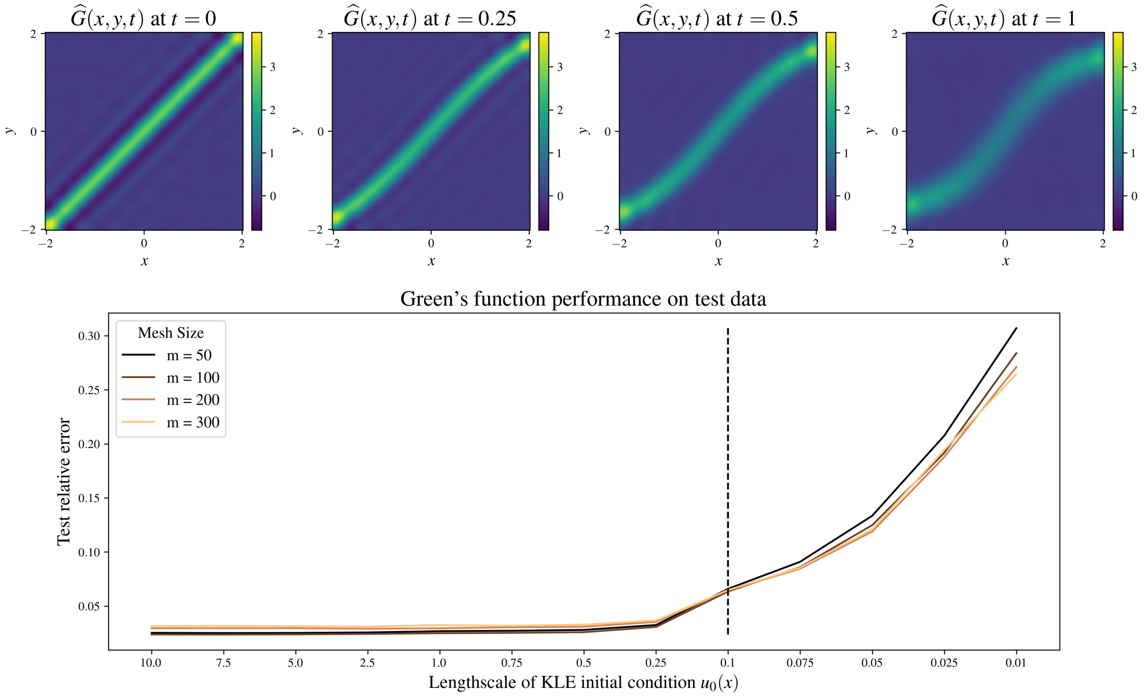


Figure 5: Learning the Green's function of the Fokker–Planck equation in a Matérn 3/2 RKHS. Top row shows the Green's function estimator $\widehat{G}(x, y, t)$ at different time slices. Bottom row shows that our estimator generalizes to test data with inputs discretized on finer grids and with varying lengthscales. Black dashed line indicates the lengthscale of the initial conditions u_0 which $\widehat{G}$ was trained on.

5.5 Heat Equation with Time Invariance & Causality Constraints

Many physical and biological systems are time-dependent where the inputs to the system $f(x, t)$ and output solutions $u(y, t)$ can be functions of time (here denoted by t). For example, the heat equation on the space-time domain $D = [0, 1] \times [0, \infty)$ with Dirichlet boundary conditions is

$$\begin{cases} \frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} = f(x,t), \quad (x,t) \in [0,1] \times [0,\infty) \\ u(0,t) = u(1,t) = 0 \\ u(x,0) = 0 \end{cases} \tag{92}$$

for some constant $\alpha > 0$. The function $u(x,t)$ is the distribution of heat in space at time t and $f(x,t)$ is a heat source. Its fundamental solution has the form

$$\begin{aligned} G_{\text{Heat}}(x,s,y,t) &= \mathbf{1}_{\{t\geq s\}} \sum_{k=1}^{\infty} 2\sin(\pi kx)\sin(\pi ky)e^{-\alpha k^2\pi^2(t-s)} \\ u(y,t) &= \int_0^{\infty}\int_0^{\infty} G_{\text{Heat}}(x,s,y,t)f(x,s)\mathrm{d}x\mathrm{d}s. \end{aligned} \tag{93}$$

Another example is the damped harmonic oscillator for $t \in \mathbb{R}$,

$$\frac{\partial^2 u}{\partial t^2} + c\frac{\partial u}{\partial t} + \omega^2 u = f(t) \tag{94}$$

where $u(t)$ is the position of the oscillator at time t, $f(t)$ is the forcing, c is the damping coefficient, and ω is the frequency of the oscillator. Letting r_1, r_2 be the roots of the quadratic equation $x^2 + cx + \omega^2$ we can write the solution as

$$u(t) = \int_{-\infty}^{\infty} G_{\text{Harmonic}}(s,t)f(s)ds, \quad G_{\text{Harmonic}}(s,t) = \mathbf{1}_{\{t\geq s\}}\frac{1}{r_1 - r_2}[e^{r_1(t-s)} - e^{r_2(t-s)}]. \tag{95}$$

Note that in both these examples, the Green's functions $G(\cdot,t,\cdot,s)$ satisfied the conditions

$$G(\cdot,s,\cdot,t) = G(\cdot,\cdot,t-s) \qquad\qquad G(\cdot,s,\cdot,t) = 0 \text{ for } t < s. \tag{96}$$

The first condition is called *time invariance* and it enforces that the Green's function performs a convolution in the time variable. The Green's function is time invariant because it only depends on the difference $t - s$ between the time a perturbation was applied and the time of the response. This is a common feature of memoryless systems where a perturbation at time s influences a response at time t in precisely the same way that a perturbation at time $s + \Delta t$ influences a response at time $t + \Delta t$ for any $\Delta t > 0$. The second condition above forces the Green's function to be zero for all $t < s$ which ensures that the Green's function is *causal* in time. This means that an impulse from f applied at time s cannot affect the solution u at an earlier time $t < s$.

If such prior knowledge about the system is available, as is the case for many diffusive processes, it is advantageous to optimize over Green's functions that satisfy these constraints. Here we use these constraints to learn the Green's function of the heat equation (92) in one spatial dimension with diffusivity constant $\alpha = 0.01$ and zero initial and boundary conditions. We aim to learn the map from the function of heat sources $f(x,s)$ to the distribution of heat $u(y,t)$ which both live on domains $D_{\mathcal{X}} = D_{\mathcal{Y}} = [0,1] \times [0,1]$. As discussed above, we study two Green's functions estimators

$$\widehat{G}_1(x,s,y,t) = \widehat{g}(x,y,t-s) \tag{97a}$$

$$\widehat{G}_2(x,s,y,t) = \mathbf{1}_{\{t\geq s\}}\widehat{g}(x,y,t-s) \tag{97b}$$

where the first estimator is time invariant while the second estimator is time invariant as well as time causal. As detailed in Appendices G.2- G.4, we can learn Green's function in these two forms by designing RKHSs with reproducing kernels

$$K(x,s,y,t,\xi,\sigma,\eta,\tau) = k(x,y,t-s,\xi,\eta,\tau-\sigma) \tag{98a}$$

$$K(x,s,y,t,\xi,\sigma,\eta,\tau) = \mathbf{1}_{t\geq s}\mathbf{1}_{\tau\geq\sigma} \cdot k_{\text{symm}}(x,y,t-s,\xi,\eta,\tau-\sigma) \tag{98b}$$

respectively where $k : ([0,1]^2 \times \mathbb{R})^2 \to \mathbb{R}$ is some Mercer kernel function and the kernel $k_{\text{symm}} : ([0,1]^2 \times \mathbb{R})^2 \to \mathbb{R}$ is its flip-symmetrized version given by

$$\begin{aligned} k_{\text{symm}}(x,y,t,\xi,\eta,\tau) = \frac{1}{4}\Big(& k(x,y,t,\xi,\eta,\tau) + k(x,y,t,\xi,\eta,-\tau) \\ &+ k(x,y,-t,\xi,\eta,\tau) + k(x,y,-t,\xi,\eta,-\tau)\Big). \end{aligned} \tag{99}$$

In the experiments for this section, we take the base kernel

$$k(x,y,t,\xi,\eta,\tau) = \exp\left(\sqrt{\frac{(x-\xi)^2}{\sigma_x^2} + \frac{(y-\eta)^2}{\sigma_y^2} + \frac{(t-\tau)^2}{\sigma_t^2}}\right) \tag{100}$$

to be exponential with lengthscales $\sigma_x = \sigma_y = \sigma_t = 0.04$.

To train our Green's function estimators, we use a small train set of 100 samples with inputs $f(x,s)$ generated by an exponential KLE with lengthscale $\ell = 0.1$ on the $[0,1]^2$ unit square. Outputs $u(y,t)$ are simulated using a backward Euler scheme and corrupted with 20% additive i.i.d. Gaussian noise. Input functions f are discretized on a grid of $m_x \times m_t$ points and output functions u on a grid of $m_y \times m_t$ points where $m_x = m_y = m_t = m = 50$. During training, we regularize our estimators with $\lambda = 2 \times 10^{-7}$ penalty which is chosen to produce smooth Green's functions $\widehat{G}_1, \widehat{G}_2$ while still small enough to approximate the exponential growth of the true Green's function around $s = t$.

In the top of Figure 6 we display cross-sections of the true and estimated Green's functions, G_{Heat} and $\widehat{G}_1, \widehat{G}_2$ respectively, for different time intervals $t = 0, 0.25$ and 0.5. As expected, for $t = 0$ both of our Green's function estimators attempt to approximate the continuous delta function $\delta(x - y)$ and are solely limited by the small but fixed bandwidth of their RKHS kernel K. For $t > 0$ we observe a good agreement with the true Green's function and correctly identify the key features of the heat equation. Namely, both estimators $\widehat{G}_1, \widehat{G}_2$ smooth the input forcings $f(x,s)$ for positive times $t > 0$, fit the zero Dirichlet boundary conditions at $y = 0$ and $y = 1$ by tending to zero at the edges, and correctly learn that the Green's function is symmetric in its spatial (x, y) coordinates due to the the positive definite property of the Laplacian operator in (92).

In the bottom of Figure 6 we show how our estimators perform on 500 new test samples when we vary the lengthscale ℓ of the boundary condition from 0.01 to 10.0 and the mesh discretization $m = m_x = m_y = m_t$ from 50 to 150. Our estimators are indeed independent of mesh discretization as their relative test errors plateau quickly as we increase m. As with previous examples, we see that the relative errors decrease as we raise the forcing input lengthscales from $\ell = 0.1$ to 10.0 (e.g. very smooth boundary conditions) but gradually increase if the input lengthscales becomes smaller than $\ell = 0.1$ on which our estimators were trained. Once again, this behavior is expected since the true Green's function of the heat equation grows exponentially near the line $t = s$, and hence cannot be estimated perfectly at resolutions which exceed the mesh discretization and the lengthscale of the forcing inputs in the train data. Importantly, we see in the bottom plot of Figure 6 that the Green's function $\widehat{G}_2$ from (97b) which is causal as well as time invariant is able to generalize better on inputs of finer lengthscales compared to the Green's function $\widehat{G}_1$ from (97a) which is simply time invariant but not causal. This again highlights the benefits of learning Green's functions of PDEs in RKHSs that encode physical constraints. Example test predictions of our time-invariant and time-causal Green's function estimator $\widehat{G}_2$ are shown in Appendix A.4 for input forcings of varied lengthscales.

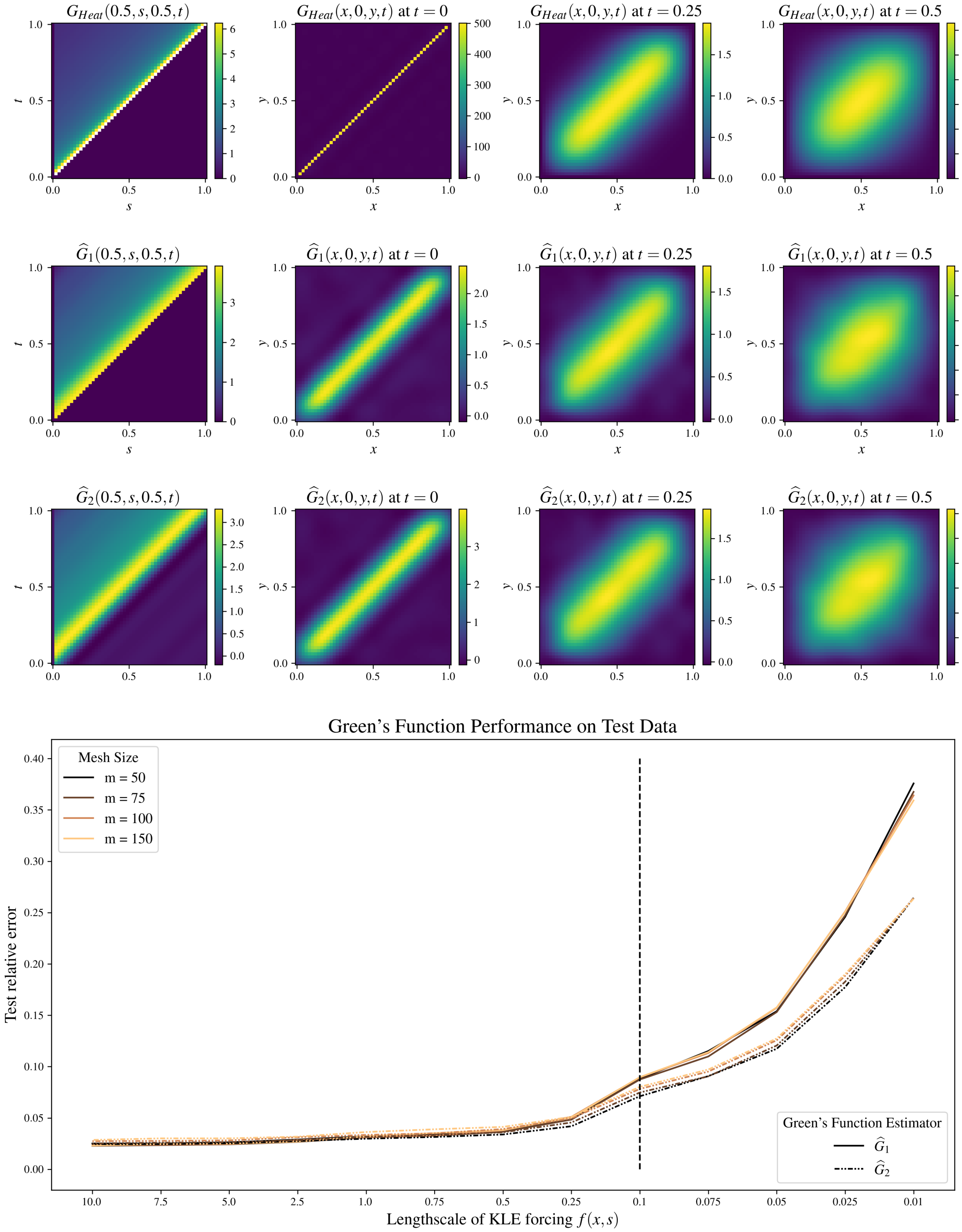


Figure 6: Estimating the Green's function of the heat equation which maps a function of heat sources $f(x, s)$ to the distribution of heat $u(y, t)$. Top row shows multiple views of the true Green's function. Next two rows show the learned Green's function estimators $\widehat{G}_1$ from (97a) and $\widehat{G}_2$ from (97b) which satisfy time invariant and causal constraints respectively. Bottom plot shows how both Green's function estimators (shown in solid and dash-dotted lines) generalize to test forcing inputs discretized on finer grids with varying levels of smoothness. The Green's function $\widehat{G}_2$ with causal constraints generalizes better to input forcings with smaller lengthscales. Black dashed line indicates the lengthscale of the training inputs f.

6. Conclusion

In this paper, we developed a data-driven method for estimating the fundamental solution operator of a linear PDE which maps an input of the PDE (initial condition, boundary condition, or forcing) to its solution. Our method estimates the Green's function and bias term of the fundamental solution operator in an RKHS solely from input-output samples and with no detailed knowledge of the PDE. Through RKHS theory, we showed that our estimator is provably optimal on the training data and can be learned by minimizing a convex objective over a finite set of weights. We extended prior results on functional linear regression to bound the prediction error of our Green's function estimator trained on a finite number of samples. On the Poisson, Helmholtz, Schrödinger, Fokker–Planck, and heat equation, we showed numerically that our RKHS estimators are significantly robust to noise in the training data and generalize to out-of-distribution test inputs with varying degrees of smoothness and mesh discretization.

An important direction of research is to extend our RKHS framework to learn nonlinear operators. This is necessary to model solution operators of nonlinear PDEs but also arises in linear PDEs when learning a map from a coefficient function to a solution. Recent advances in neural operators (Lu et al., 2019; Li et al., 2021) and operator RKHS theory (Kadri et al., 2016; Nelsen and Stuart, 2021) have been instrumental in modeling such nonlinear maps. In light of these works, a natural extension of our framework would be the following *Reproducing Kernel Network* (RKN)

$$u_l(y) = \sigma_l\Big(\int_{D_{l-1}} G_l(x,y)u_{l-1}(x)\mathrm{d}x + \beta_l(y)\Big) \text{ for all } l = 1,\dots,L \tag{101}$$

where the input to the PDE is $f(x) = u_0(x)$ at layer $l = 0$ and the solution is modeled by $u(y) = u_L(y)$ at layer $l = L$. Here the lth layer takes a function from $L_2(D_{l-1})$ to $L_2(D_l)$ where it is composed of an integral operator G_l in an RKHS $\mathcal{G}_l \subset L^2(D_{l-1} \times D_l)$ and a bias term β_l in an RKHS $\mathcal{B}_l \subset L^2(D_l)$ as well as a pointwise activation function $\sigma_l : \mathbb{R} \to \mathbb{R}$. This general architecture is truly a continuous operator map between function spaces that can be implemented on any PDE domain geometry and irregular mesh. Furthermore, it promises various benefits including robustness to noise and mesh discretization as seen in our linear Green's function setting. Developing optimization methods and statistical guarantees for such nonlinear operators pose an interesting direction for future research.

Acknowledgments

We thank Philippe Rigollet for encouraging us to study the statistical aspects of surrogate modeling and for helpful advice and comments on our manuscript. We also thank Alasdair Hastewell, Adam Block, and Jörn Dunkel for helpful discussions and feedback on our final draft as well as Nicholas Nelsen and Houman Owhadi for useful comments and for pointing out relevant related work. Lastly, we greatly appreciate the many insightful suggestions of the anonymous referees. George Stepaniants was supported by the National Science Foundation under Grant No. 1745302. We acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center (Reuther et al., 2018) for providing HPC resources that have contributed to the numerical experiments reported within this paper. All code and experiments can be found at
`https://github.com/sgstepaniants/OperatorLearning`.

Appendix A. Experimental Details and Supplementary Results

Here we give further details on our experimental setup for all experiments studied in Section 5. For all numerical examples of the Poisson, Helmholtz, and heat equations we studied forced linear PDEs of the form

$$\begin{aligned} \mathcal{L}u &= f \\ \mathcal{B}u &= 0 \end{aligned} \tag{102}$$

where $\mathcal{L}$ is a linear differential operator and $\mathcal{B}$ is a linear operator which enforces the zero Dirichlet or Neumann boundary conditions of the PDE. On the examples of the Schrödinger and Fokker–Planck equations we studied linear boundary value problems of the form

$$\begin{aligned} \mathcal{L}u &= 0 \\ \mathcal{B}u &= f \end{aligned} \tag{103}$$

where $\mathcal{B}$ enforces either the initial conditions or the boundary conditions of the PDE. In both settings, we assume that f and u are square integrable functions in possibly different domains $D_{\mathcal{X}}$ and $D_{\mathcal{Y}}$ respectively and are interested in learning the linear operator $\mathcal{T}$ which maps f to u.

A.1 Data Generation

For every experiment, the input functions $f(x)$ are discretized on some mesh $\{x_j\}_{j=1}^{m_x}$ of m_x points. These inputs are generated using a Karhunen–Loeve expansion up to order m_x given by

$$f(x_j) = \sum_{k=1}^{m_x} Z_k \phi_k(x_j), \quad Z_k \sim \mathcal{N}(0, \lambda_k) \tag{104}$$

for i.i.d. normal random variables Z_k where λ_k, ϕ_k are the eigenvalues and eigenvectors of a continuous, symmetric, positive definite kernel function $K(x, x')$. Note that the kernel K here is used to define the distribution of our random inputs f, it is not to be confused with the reproducing kernel of the Green's function and bias term RKHSs.

By Mercer's theorem (Cucker and Smale, 2002, Section 2, Theorem 1) we know that K has the spectral decomposition

$$K(x, x') = \sum_{k=1}^{\infty} \lambda_k \phi_k(x) \phi_k(y). \tag{105}$$

To compute the expansion in (104), the eigenvalues and eigenfunctions of K are computed numerically from the discrete matrix $\mathbf{K} \in \mathbb{R}^{m_x \times m_x}$ where $K_{ij} = K(x_i, x_j)$.

In Sections 5.1 and 5.2, for the Poisson and Helmholtz equations we use the squared exponential kernel $K(x, x') = \exp(\|x - x'\|^2/2\ell^2)$ with a predefined lengthscale ℓ to generate all input forcings $f(x)$. Likewise, for the Fokker–Planck equation in Section 5.4 we use the squared exponential kernel to generate all initial conditions $u_0(x)$. On the example of Schrödinger's equation in Section 5.3, the random boundary conditions $b(x)$ are generated using the squared exponential *periodic* kernel $K(x, x') = \exp(-2\sin(\pi\|x - x'\|/4)^2/\ell^2)$ to ensure that the boundary conditions are indeed periodic when wrapped around the boundary of the unit square. Lastly, for the heat equation in Section 5.5 we use the exponential kernel $K(x, x') = \exp(\|x - x'\|/\ell)$ to generate the input forcings $f(x)$.

A.2 Noise Model

Throughout all experiments, we corrupt our output samples $u(y)$ with a fraction p of i.i.d. Gaussian noise. Specifically, given n output samples $\mathbf{u}_1, \ldots, \mathbf{u}_n \in \mathbb{R}^{m_y}$ discretized on a mesh $\{y_k\}_{k=1}^{m_y}$ we compute the average standard deviation of all the samples

$$\sigma_u = \sqrt{\frac{1}{nm_y}\sum_{i=1}^{n}\sum_{k=1}^{m_y}(u_{ik} - \overline{u}_k)^2}, \quad \overline{u}_k = \frac{1}{n}\sum_{i=1}^{n} u_{ik} \tag{106}$$

and then corrupt our outputs with a fraction p of i.i.d. Gaussian noise with standard deviation σ_u by

$$\tilde{u}_{ik} = u_{ik} + p\varepsilon_{ik}, \quad \varepsilon_{ik} \sim \mathcal{N}(0, \sigma_u^2) \tag{107}$$

for $i = 1, \ldots, n$ and $k = 1, \ldots, m_y$. Here ε_{ik} are i.i.d. Gaussian random variables. This noise setup assumes that all locations in the domain $D_\mathcal{Y}$ of the functional outputs u receive the same level of noise corruption which is a realistic model of sensor recordings where the sensor has a fixed level of noise in its measurements.

A.3 Setting the RKHS Kernel Lengthscale Parameter

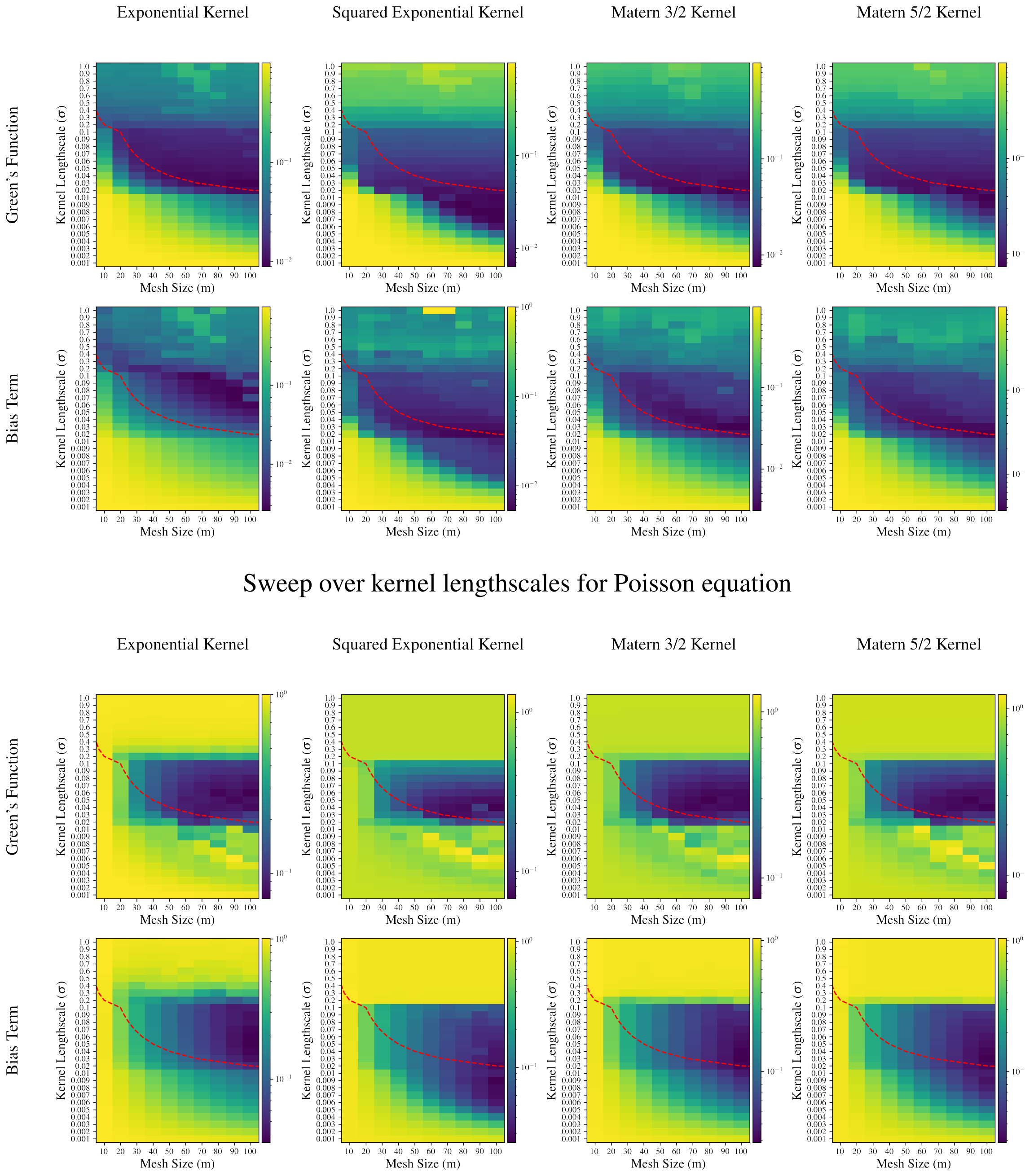
Recall from Section 3 that given input functions $f_i(x)$ discretized on the mesh $\{x_j\}_{j=1}^{m_x}$ and output functions $u_i(y)$ discretized on the mesh $\{y_k\}_{k=1}^{m_y}$, our Green's function and bias term estimators take the form

$$G_\mathbf{W}(x, y) = \sum_{j=1}^{m_x}\sum_{k=1}^{m_y} K(x, y, x_j, y_k)W_{jk}\Delta_j^x\Delta_k^y, \quad \beta_\mathbf{w}(y) = \sum_{k=1}^{m_y} Q(y, y_k)w_k\Delta_k^y \tag{108}$$

where Δ_j^x, Δ_k^y are fixed numerical quadrature weights of the input-output function grids respectively. During training, we optimize the weights $\mathbf{w}, \mathbf{W}$ of our estimators such that they minimize (38) which consists of a mean squared error term on the training data plus RKHS norm penalties on the estimators. In all experiments, the input-output grids $\{x_j\}_{j=1}^{m_x}, \{y_k\}_{k=1}^{m_y}$ are chosen to be uniform with equal spacing in all dimensions.

Here we study how the relative errors of the learned Green's function and bias term depend on the number of grid points m and the lengthscale σ of the kernels K and Q that define these estimators. We perform this experiment on the 1D Poisson equation from (80) and on the 1D Helmholtz equation with $\omega = 20$ from (81). Our estimators are trained on 500 input-output samples with regularization $\lambda = 10^{-6}$ because we do not add noise to our data. In this experiment, $m_x = m_y = m$ and $\{x_j\}_{j=1}^{m_x}, \{y_k\}_{k=1}^{m_y} \subset [0, 1]$ since we are solving these PDEs on the unit interval. From Section 2.1, K and Q are both taken to be either exponential, squared exponential, Matérn 3/2 or Matérn 5/2 kernels and their lengthscales in the x and y directions are set equal to each other $\sigma_x = \sigma_y = \sigma$.

In Figure 7, we show the relative error of the learned Green's function and bias term on the Poisson and Helmholtz equations across several different kernels, mesh sizes m, and kernel lengthscales σ. We see that the relation $\sigma = \frac{2}{m}$ (red dashed line) is a consistently good choice for the kernel lengthscale across all kernels types as it is quite small while still achieving a low relative error. In all our numerical examples in Section 5 we have found this to be a reliable choice of the kernel lengthscale.



Sweep over kernel lengthscales for Poisson equation

Sweep over kernel lengthscales for Helmholtz equation

Figure 7: Each heatmap in this figure shows for a particular RKHS kernel, the relative error of the Green's function and bias term of the Poisson or Helmholtz equation learned in that RKHS. In each heatmap plot, we show how the relative error of the learned Green's function and bias term depend on the number of grid discretization points $m_x = m_y = m$ and the lengthscale $\sigma_x = \sigma_y = \sigma$ of the reproducing kernel used to construct the estimator on that grid. We find that the scaling relation $\sigma = \frac{2}{m}$ (red dashed line) is a consistently good choice for the lengthscale across all RKHS kernels as it is as small as possible while still achieving a low relative error.

In general, for multidimensional problems we may be given input functions $f(x)$ with $x \in \mathbb{R}^{d_\mathcal{X}}$ discretized on a uniform grid of size $m_x^1 \times \dots m_x^{d_\mathcal{X}}$ defined on a product domain $D_\mathcal{X} = \Pi_{i=1}^{d_\mathcal{X}} [a_i^x, b_i^x]$. Likewise, the output functions $u(y)$ with $y \in \mathbb{R}^{d_\mathcal{Y}}$ can be discretized on a uniform grid of size $m_y^1 \times \dots m_y^{d_\mathcal{Y}}$ defined on a product domain $D_\mathcal{Y} = \Pi_{i=1}^{d_\mathcal{Y}} [a_i^y, b_i^y]$. For multidimensional problems, the kernels K and Q have vector-valued lengthscales which specify the width in each dimension as $[\boldsymbol{\sigma}_x, \boldsymbol{\sigma}_y] \in \mathbb{R}^{d_\mathcal{X} + d_\mathcal{Y}}$ and $\boldsymbol{\sigma}_y \in \mathbb{R}^{d_\mathcal{Y}}$ respectively. Generalizing our rule above to multiple dimensions, we set the kernel lengthscales to $(\boldsymbol{\sigma}_x)_i \approx 2\frac{b_i^x - a_i^x}{m_x^i}$ for $i = 1, \dots, d_\mathcal{X}$ as well as $(\boldsymbol{\sigma}_y)_i \approx 2\frac{b_i^y - a_i^y}{m_y^i}$ for $i = 1, \dots, d_\mathcal{Y}$.

A.4 Test Predictions of Learned Green's Functions

In this section, we show how Green's functions learned on the Schrödinger, Fokker–Planck, and heat equations predict the solutions of these PDEs given finely discretized test inputs of varying lengthscales. For each lengthscale ℓ, we generate 500 test inputs $f(x)$ at that lengthscale and simulate the corresponding solutions $u(y)$. We then show the best (lowest relative error) and worst (highest relative error) predictions of PDE solutions that are observed across those test samples at that lengthscale.

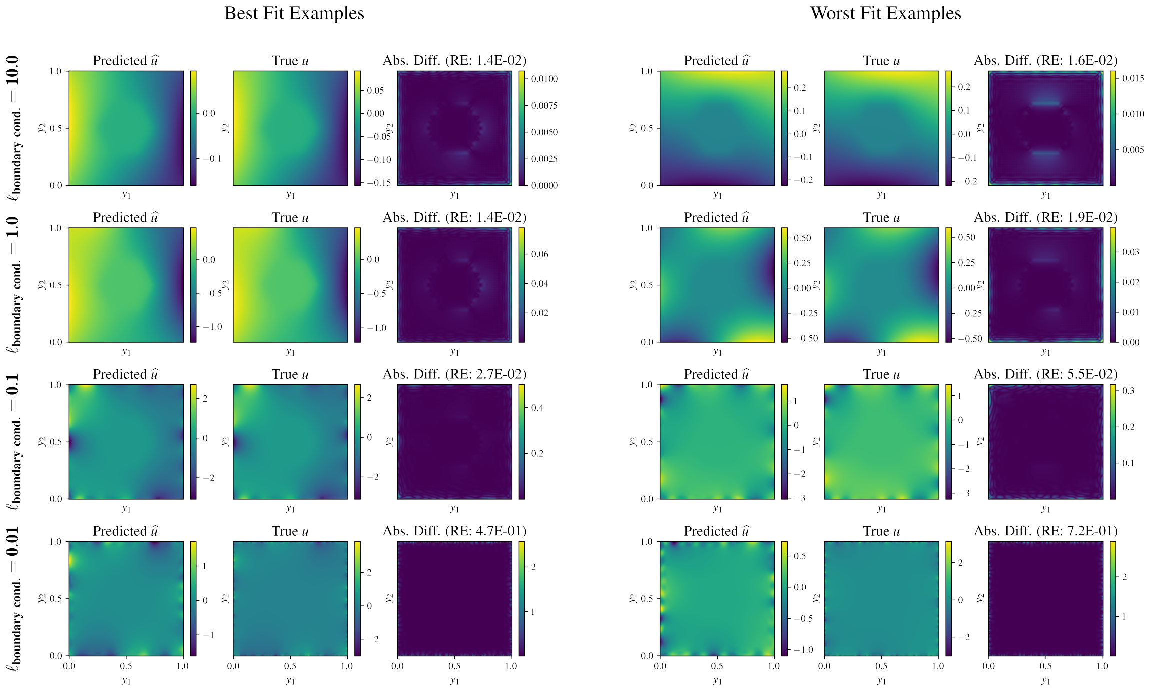


Figure 8: Example test predictions from Section 5.3 for the Schrödinger equation $\Delta u - Vu = 0$ on $D = [0, 1]^2$ with $u = b$ on ∂D for the linear map $G : b \mapsto u$ from boundary condition to solution. All input boundary conditions b and solutions u in the plots above are discretized on a grid of size 300 × 300.

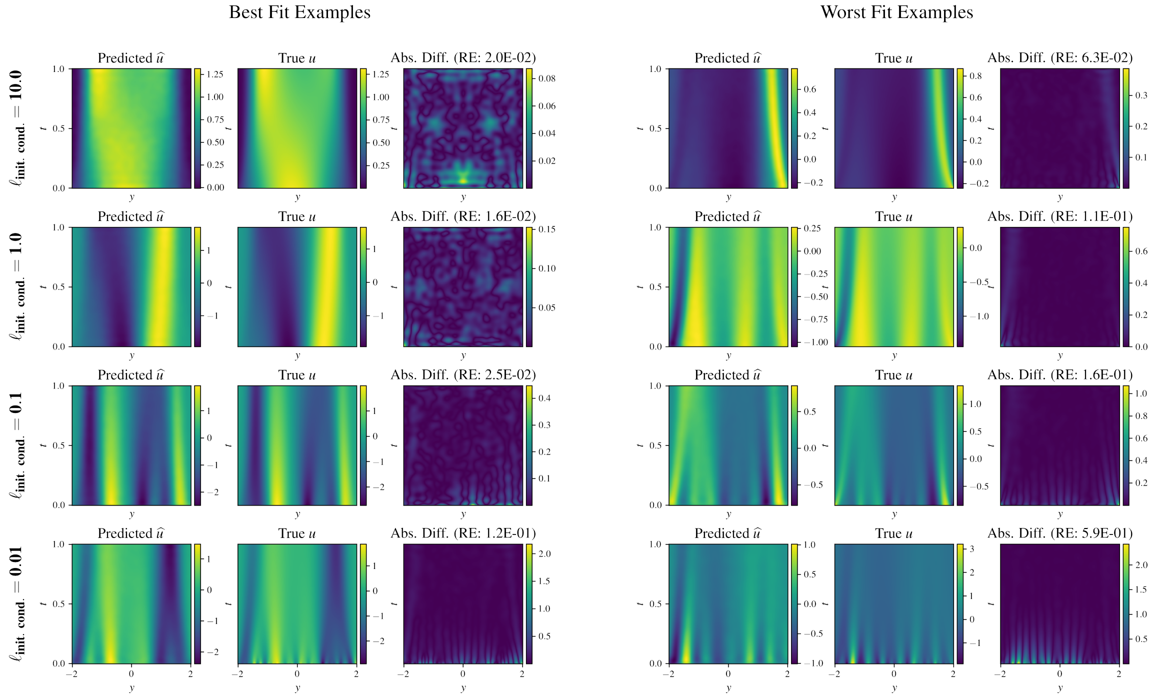


Figure 9: Example test predictions from Section 5.4 of $\widehat{G} : u_0 \mapsto u$ for the Fokker–Planck equation $\partial_t u = \nabla \cdot (u\nabla V) + \alpha\Delta u$ in one space and time dimension on the domain $(x, t) \in D = [-2, 2] \times [0, 1]$. The input initial conditions $u_0(x)$ are discretized on 150 points in [-2, 2] while the solutions plotted above on the domain D are discretized on a space-time grid of 300×300 points.

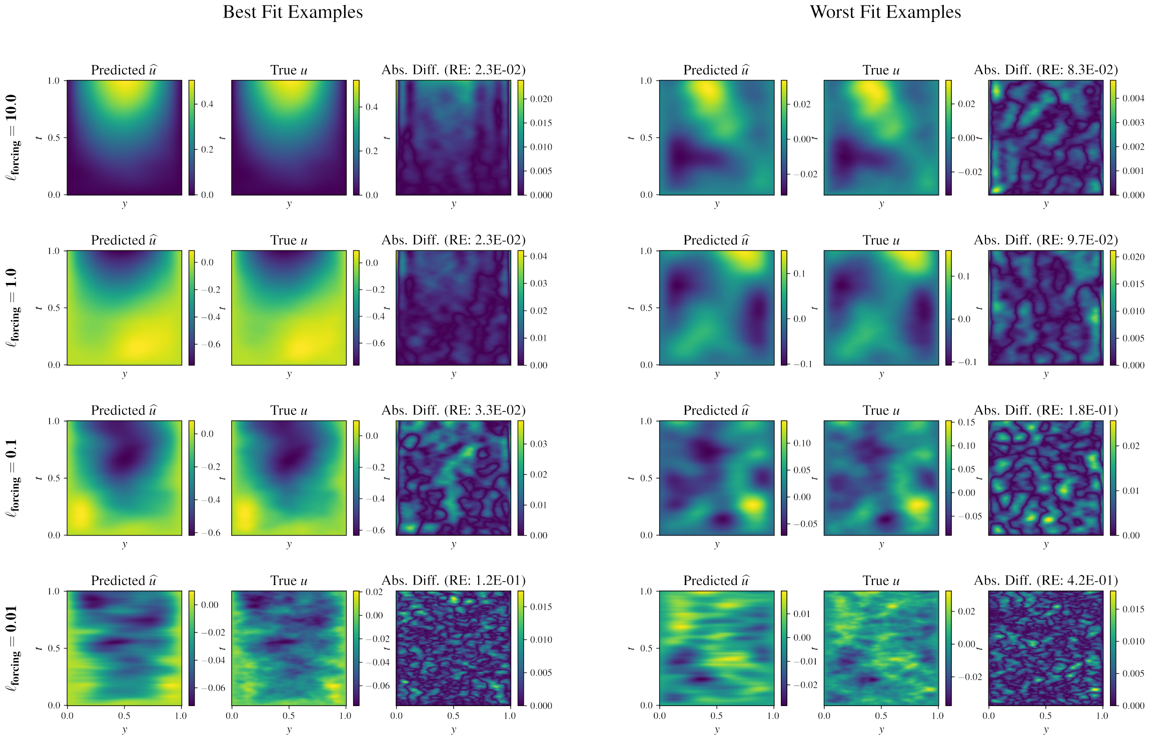


Figure 10: Example test predictions from Section 5.5 of $\widehat{G}_2 : f \mapsto u$ for the heat equation $\partial_t u = \alpha \Delta u$ with $\alpha = 0.01$ in one space and time dimension on the domain $(x, t) \in D = [0, 1] \times [0, 1]$. The input forcings $f(x, s)$ and output solutions $u(y, t)$ on the domain D plotted above are discretized on a space-time grid of 150×150 points.

Appendix B. Representer Theorems

At the start of Section 2.2, we began by proving a simple representer theorem for our Green's function estimator when the functional data given to us was discretized on a set of grid points. Deriving the closed-form of the Green's function estimator relied on a key result that we restate and prove here.

Theorem 8 *Suppose we are given an RKHS $\mathcal{H} \subset L^2(D)$ on a domain D with continuous, symmetric, and strictly positive definite kernel $K : D^2 \to \mathbb{R}$. We denote the RKHS Hilbert space norm of any function $f \in \mathcal{H}$ by $\|f\|_{\mathcal{H}}$. Take any finite set of m points $\{x_k\}_{k=1}^m \subset D$ in the domain and n weight vectors $\{\mathbf{a}_i\}_{i=1}^n \subset \mathbb{R}^m$. We denote $\mathbf{f} = (f(x_1), \dots, f(x_m))^T \in \mathbb{R}^m$ as the function f evaluated on the set of grid points. Finally, let $\psi : \mathbb{R}_+ \to \mathbb{R}$ be a strictly increasing real-valued function and let $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}$ be an arbitrary error function. Then any minimizer*

$$\widehat{f} \in \arg\min_{f \in \mathcal{H}} L(\{\langle \mathbf{f}, \mathbf{a}_i \rangle_2\}_{i=1}^n) + \psi(\|f\|_{\mathcal{H}}) \tag{109}$$

must be of the form

$$\widehat{f}(x) = \sum_{k=1}^m K(x, x_k) w_k, \quad \mathbf{w} \in \operatorname{span}(\{\mathbf{a}_i\}_{i=1}^n) \subset \mathbb{R}^m. \tag{110}$$

We assume here for simplicity that K is a strictly positive definite kernel although the proof of this theorem can be easily extended to degenerate RKHSs.

Proof We begin by defining two vector spaces

$$V = \operatorname{span}(\{\mathbf{a}_i\}_{i=1}^n), \quad V^\perp = \{\mathbf{v} \in \mathbb{R}^m : \mathbf{v}^T \mathbf{K} \mathbf{w} = 0, \, \forall \mathbf{w} \in V\} \tag{111}$$

where $\mathbf{K} = \{K(x_i, x_j)\}_{i,j=1}^m \in \mathbb{R}^{m \times m}$. Because $\mathbf{K}$ is strictly positive definite, we know that $V \oplus V^\perp = \mathbb{R}^m$.

Since our objective function depends only on $f \in \mathcal{H}$ evaluated at the points $x_1, \dots, x_m$ then we know by the standard representer theorem that $\widehat{f}$ must take the form

$$\widehat{f}(x) = \sum_{k=1}^m K(x, x_k) w_k \tag{112}$$

where $\mathbf{w} = (w_1, \dots, w_m)^T \in \mathbb{R}^m$. Now we can decompose our weight vector as $\mathbf{w} = \mathbf{w}^{\|} + \mathbf{w}^\perp$ for $\mathbf{w}^{\|} \in V$ and $\mathbf{w}^\perp \in V^\perp$ to write

$$\widehat{f}(x) = \sum_{k=1}^m K(x, x_k) w_k^{\|} + \sum_{k=1}^m K(x, x_k) w_k^\perp. \tag{113}$$

Using the definition of the RKHS norm we can write out

$$\begin{aligned} \|\widehat{f}\|_{\mathcal{H}}^2 &= (\mathbf{w}^{\|})^T \mathbf{K} \mathbf{w}^{\|} + (\mathbf{w}^\perp)^T \mathbf{K} \mathbf{w}^\perp + 2(\mathbf{w}^\perp)^T \mathbf{K} \mathbf{w}^{\|} \\ &= (\mathbf{w}^{\|})^T \mathbf{K} \mathbf{w}^{\|} + (\mathbf{w}^\perp)^T \mathbf{K} \mathbf{w}^\perp \end{aligned} \tag{114}$$

where the final term $(\mathbf{w}^\perp)^T \mathbf{K} \mathbf{w}^{\|}$ is zero since $\mathbf{w}^\perp \in V^\perp$.

Additionally, writing as shorthand $\widehat{\mathbf{f}} = (\widehat{f}(x_1), \ldots, \widehat{f}(x_m))^T \in \mathbb{R}^m$ we see for each $i = 1, \ldots, n$ that

$$\langle \widehat{\mathbf{f}}, \mathbf{a}_i \rangle_2 = \mathbf{a}_i^T \widehat{\mathbf{f}} = \mathbf{a}_i^T \mathbf{K} \mathbf{w}^{\parallel} + \mathbf{a}_i^T \mathbf{K} \mathbf{w}^{\perp} = \mathbf{a}_i^T \mathbf{K} \mathbf{w}^{\parallel} \tag{115}$$

where again the last equality follows since $\mathbf{a}_i \in V$ and $\mathbf{w}^{\perp} \in V^{\perp}$. Finally, we can write our objective function as

$$L(\{\langle \mathbf{f}, \mathbf{a}_i \rangle_2\}_{i=1}^n) + \psi(\|f\|_{\mathcal{H}}) = L(\{\mathbf{a}_i^T \mathbf{K} \mathbf{w}^{\parallel}\}_{i=1}^n) + \psi(\sqrt{(\mathbf{w}^{\parallel})^T \mathbf{K} \mathbf{w}^{\parallel} + (\mathbf{w}^{\perp})^T \mathbf{K} \mathbf{w}^{\perp}}). \tag{116}$$

Since ψ is strictly increasing this proves that the optimal $\mathbf{w} \in \mathbb{R}^m$ must have $\mathbf{w}^{\perp} = 0$ implying that

$$\mathbf{w} = \mathbf{w}^{\parallel} \in V = \operatorname{span}(\{\mathbf{a}_i\}_{i=1}^n). \tag{117}$$

■

As an important note, from the theorem above we can recover the traditional representer theorem by setting $n = m$ and $\mathbf{a}_i = \mathbf{e}_i$ for all $i = 1, \ldots, n$.

Now we lay the groundwork and derive the Green's function representer theorem stated in Theorem 4 for functional (non-discrete) input-output samples.

Proof of Operator Representer Theorem

We state a more general representer theorem than that which is given in Theorem 4. It holds for any operator (function-valued) RKHS with arbitrary loss function and regularization term. It is a generalization of the representer theorem proven in Kadri et al. (2016, Appendix B, Theorem 9). The original theory of operator RKHSs and the corresponding representer theorem was first presented in the seminal work of Micchelli and Pontil (2005, Section 4, Theorem 4.1). We follow an alternative derivation of their result using a similar analysis to that of Wahba (1990, Section 1.3, Theorem 1.3.1).

To establish notation, we let $\mathcal{X}$ be a separable Hilbert space of functions from $D_{\mathcal{X}} \to \mathbb{R}$ and we let $\mathcal{Y}$ be a separable Hilbert space of functions from $D_{\mathcal{Y}} \to \mathbb{R}$. We denote $\mathcal{L}(\mathcal{Y})$ as the space of bounded linear operators from $\mathcal{Y}$ to $\mathcal{Y}$. First we define an RKHS over operators as posed by Kadri et al. (2016, Section 4, Definitions 3, 5).

Definition 9 (Operator-valued kernel) *An operator-valued kernel is a function $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ satisfying*

(i) K is Hermitian if $\forall f, g \in \mathcal{X}$ we have that $K(f, g) = K(g, f)^*$ where $*$ denotes the adjoint operator.

(ii) K is nonnegative (positive semidefinite) on $\mathcal{X}$ if it is Hermitian and for all $r \in \mathbb{N}$ and any $\{(f_i, u_i)\}_{i=1}^r \subset \mathcal{X} \times \mathcal{Y}$ we have that the matrix $M \in \mathbb{R}^{r \times r}$ with $M_{ij} = \langle K(f_i, f_j)u_i, u_j \rangle_{\mathcal{Y}}$ is positive semidefinite.

Definition 10 (Operator RKHS) *Let $\mathcal{O}$ be a Hilbert space of operators $O : \mathcal{X} \to \mathcal{Y}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{O}}$. We call $\mathcal{O}$ an operator RKHS if there exists an operator-valued kernel $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ such that*

(i) The function $g \mapsto K(f, g)u$ for $g \in \mathcal{X}$ belongs to $\mathcal{O}$ for all $f \in \mathcal{X}, u \in \mathcal{Y}$.

(ii) K satisfies the reproducing property

$$\langle O, K(f,\cdot)u\rangle_{\mathcal{O}} = \langle O(f), u\rangle_{\mathcal{Y}} \tag{118}$$

for all $O \in \mathcal{O}$ and $f \in \mathcal{X}, u \in \mathcal{Y}$.

We assume now that the operator RKHS $\mathcal{O}$ can be decomposed orthogonally into $\mathcal{O} = \mathcal{O}_0 \oplus \mathcal{O}_1$ where $\mathcal{O}_0$ is a finite-dimensional Hilbert space spanned by the operators $\{E_k\}_{k=1}^r$ and $\mathcal{O}_1$ is its orthogonal complement under the inner product $\langle\cdot,\cdot\rangle_{\mathcal{O}}$. We denote the inner product $\langle\cdot,\cdot\rangle_{\mathcal{O}}$ restricted to $\mathcal{O}_0, \mathcal{O}_1$ as $\langle\cdot,\cdot\rangle_{\mathcal{O}_0}, \langle\cdot,\cdot\rangle_{\mathcal{O}_1}$ respectively.

Theorem 11 *Let $\psi : \mathbb{R}_+ \to \mathbb{R}$ be a strictly increasing real-valued function and let $\mathcal{L} : (\mathcal{X} \times \mathcal{Y} \times \mathcal{Y}) \to \mathbb{R}$ be an arbitrary error function. Then any minimizer*

$$\widehat{O}_{n\psi} \in \operatorname*{arg\,min}_{O \in \mathcal{O}} \Big[\mathcal{L}\Big(\{(f_i, u_i, O(f_i))\}_{i=1}^n\Big) + \psi(\|\operatorname*{proj}_{\mathcal{O}_1} O\|_{\mathcal{O}_1})\Big] \tag{119}$$

must be of the form

$$\widehat{O}_{n\psi}(x,y) = \sum_{k=1}^r d_k E_k + \sum_{i=1}^n K(f_i,\cdot)c_i \tag{120}$$

for some $\mathbf{d} \in \mathbb{R}^r$ and $c_i \in \mathcal{Y}, i \in [n]$.

Proof [Theorem 11]
Define the linear functional $L_i^v : \mathcal{O} \to \mathbb{R}$ for any $v \in \mathcal{Y}, i \in [n]$ where $L_i^v O = \langle O(f_i), v\rangle_{\mathcal{Y}}$. Note that for each $v \in \mathcal{Y}$ we have that L_i^v is continuous (bounded) because

$$\begin{aligned} |L_i^v O| &= |\langle O(f_i), v\rangle_{\mathcal{Y}}| = |\langle O, K(f_i,\cdot)v\rangle_{\mathcal{O}}| \le \|O\|_{\mathcal{O}}\|K(f_i,\cdot)v\|_{\mathcal{O}} \\ &= \|O\|_{\mathcal{O}}\sqrt{\langle K(f_i,\cdot)v, K(f_i,\cdot)v\rangle_{\mathcal{O}}} = \|O\|_{\mathcal{O}}\sqrt{\langle K(f_i,f_i)v, v\rangle_{\mathcal{Y}}} \\ &\le \|v\|_{\mathcal{Y}}\|K(f_i,f_i)\|_{\text{op}}\|O\|_{\mathcal{O}} \end{aligned} \tag{121}$$

Note that the operator norm $\|K(f_i,f_i)\|_{\text{op}} < \infty$ because by definition $K(f_i,f_i) \in \mathcal{L}(\mathcal{Y})$ is bounded and $\|v\|_{\mathcal{Y}} < \infty$ since $v \in \mathcal{Y}$.

Therefore, by the Riesz representation theorem we know that for all $i \in [n], v \in \mathcal{Y}$ there exists a representer $N_i^v \in \mathcal{G}$ for L_i^v such that

$$\langle N_i^v, O\rangle_{\mathcal{O}} = L_i^v O. \tag{122}$$

for every $O \in \mathcal{O}$.

Define $\Xi_i^v = \operatorname{proj}_{\mathcal{O}_1} N_i^v \in \mathcal{O}_1$ for all $i \in [n]$ and $v \in \mathcal{Y}$. Then for every $f \in \mathcal{X}, u \in \mathcal{Y}$ we can write

$$\begin{aligned} \langle \Xi_i^v(f), u\rangle_{\mathcal{Y}} &= \langle \Xi_i^v, K(f,\cdot)u\rangle_{\mathcal{O}_1} = \langle \Xi_i^v, K(f,\cdot)u\rangle_{\mathcal{O}} = \langle \operatorname*{proj}_{\mathcal{O}_1} N_i^v, K(f,\cdot)u\rangle_{\mathcal{O}} \\ &= \langle N_i^v, \operatorname*{proj}_{\mathcal{O}_1} K(f,\cdot)u\rangle_{\mathcal{O}} = \langle N_i^v, K(f,\cdot)u\rangle_{\mathcal{O}} = L_i^v\Big(K(f,\cdot)u\Big) \\ &= \langle K(f,f_i)u, v\rangle_{\mathcal{Y}} = \langle K(f_i,f)v, u\rangle_{\mathcal{Y}}. \end{aligned} \tag{123}$$

We have used the fact above that the projection operator $\mathsf{proj}_{\mathcal{O}_1}$ is self-adjoint and that $K(f, f_i)^* = K(f_i, f)$ by definition of an operator reproducing kernel. Since the equality above holds for all $f \in \mathcal{X}, u \in \mathcal{Y}$ this proves that $\Xi_i^v = K(f_i, \cdot)v$ for all $v \in \mathcal{Y}$.

Now let $Q : \mathcal{X} \to \mathcal{Y} \in \mathcal{O}$ be any operator perpendicular in the norm of $\mathcal{O}$ to the subspace of $\mathcal{O}$ spanned by $\{E_k\}_{k=1}^r$ and $\{\Xi_i^v : i \in [n], v \in \mathcal{Y}\}$. Note that $\operatorname{span}\{\Xi_i^v : i \in [n], v \in \mathcal{Y}\} = \{K(f_i, \cdot)v : v \in \mathcal{Y}\}$. Hence, for any $\mathbf{d} \in \mathbb{R}^r$ and $c_i \in \mathcal{Y}, i \in [n]$ we can decompose any $O \in \mathcal{O}$ as

$$O = \sum_{k=1}^{r} d_k E_k + \sum_{i=1}^{n} K(f_i, \cdot)c_i + Q. \tag{124}$$

Since Q is orthogonal to the span of $\{E_k\}_{k=1}^r$ this immediately implies that $Q \in \mathcal{O}_1$. Hence, for any $i \in [n], v \in \mathcal{Y}$,

$$\langle Q(f_i), v\rangle_{\mathcal{Y}} = L_i^v Q = \langle N_i^v, Q\rangle_{\mathcal{O}} = \langle \Xi_i^v, Q\rangle_{\mathcal{O}_1} = 0. \tag{125}$$

This proves that $Q(f_i) = 0$ so

$$O(f_i) = \sum_{k=1}^{r} d_k E_k(f_i) + \sum_{i=1}^{n} K(f_i, f_i)c_i \tag{126}$$

does not depend on Q. Lastly, we can see that $\mathsf{proj}_{\mathcal{O}_1} O = \sum_{i=1}^n K(f_i, \cdot)c_i + Q$ where $\sum_{i=1}^n K(f_i, \cdot)c_i$ and Q are orthogonal under the $\mathcal{O}_1$ inner product.

Then, the objective of (119) with the representation of O in (124) becomes

$$\begin{aligned} &\mathcal{L}\Big(\{(f_i, u_i, O(f_i))\}_{i=1}^n\Big) + \psi\Big(\|\operatorname*{\mathsf{proj}}_{\mathcal{O}_1} O\|_{\mathcal{O}_1}\Big) \\ &= \mathcal{L}\Big(\{(f_i, u_i, O(f_i))\}_{i=1}^n\Big) + \psi\Big(\sqrt{\Big\|\sum_{i=1}^{n} K(f_i, \cdot)c_i\Big\|_{\mathcal{O}_1}^2 + \|Q\|_{\mathcal{O}_1}^2}\Big). \end{aligned} \tag{127}$$

Clearly, the optimal choice of Q is zero since Ψ is strictly increasing so this proves that the minimizer of the objective in (119) must be of the form

$$\widehat{O}_{n\psi} = \sum_{k=1}^{r} d_k E_k + \sum_{i=1}^{n} K(f_i, \cdot)c_i. \tag{128}$$

This completes the proof of the representer theorem for operator RKHSs $\mathcal{O} = \mathcal{O}_0 \oplus \mathcal{O}_1$. ■

Remark 12 *Note that the space of Green's function integral operators*

$$\mathcal{O} = \Big\{O_G(f) := \int_{D_{\mathcal{X}}} G(x, \cdot)f(x)\mathrm{d}x \Big| G \in \mathcal{G}\Big\} \tag{129}$$

for some RKHS $\mathcal{G} \subset L^2(D_{\mathcal{X}} \times D_{\mathcal{Y}})$ actually defines an operator RKHS. This is because we can make $\mathcal{O}$ a Hilbert space by defining on it the induced inner product

$$\langle O_F, O_G\rangle_{\mathcal{O}} = \langle F, G\rangle_{\mathcal{G}}. \tag{130}$$

Furthermore, given that $\mathcal{G}$ has the continuous reproducing kernel $K : (D_\mathcal{X} \times D_\mathcal{Y})^2 \to \mathbb{R}$, it is easy to check that the operator-valued reproducing kernel for $\mathcal{O}$ is $\mathcal{K} : L^2(D_\mathcal{X}) \times L^2(D_\mathcal{X}) \to \mathcal{L}(L^2(D_\mathcal{Y}))$ given by

$$[\mathcal{K}(f,g)u](y) = \int_{D_\mathcal{X}} \int_{D_\mathcal{X}} \int_{D_\mathcal{Y}} K(x,y,\xi,\eta) g(x) f(\xi) u(\eta) \tag{131}$$

for all $f, g \in L^2(D_\mathcal{X})$ and $u \in L^2(D_\mathcal{Y})$. Note that $\mathcal{K}(f,g)$ is always a bounded linear operator since K is continuous on the bounded set $(D_\mathcal{X} \times D_\mathcal{Y})^2$. Using the definition of $\mathcal{K}$ we can write for any $O_F \in \mathcal{O}$,

$$\begin{aligned}\langle O_F, \mathcal{K}(f,\cdot)u\rangle &= \Big\langle F, \int_{D_\mathcal{X}} \int_{D_\mathcal{Y}} K(\cdot,\cdot,\xi,\eta) f(\xi) u(\eta) \mathrm{d}\xi \mathrm{d}\eta \Big\rangle_\mathcal{G} \\ &= \int_{D_\mathcal{X}} \int_{D_\mathcal{Y}} \langle F, K_{(\xi,\eta)}\rangle_\mathcal{G} f(\xi) u(\eta) \mathrm{d}\xi \mathrm{d}\eta \\ &= \int_{D_\mathcal{X}} \int_{D_\mathcal{Y}} F(\xi,\eta) f(\xi) u(\eta) \mathrm{d}\xi \mathrm{d}\eta \\ &= \langle O_F(f), u\rangle_{L^2(D_\mathcal{Y})}.\end{aligned} \tag{132}$$

Hence, the operator representer theorem derived above is immediately applicable to the setting of Theorem 4.

In the following corollary, we derive an expression for the closed form of these weights d_k, c_i in the case of ridge regression. Throughout, we denote $\mathcal{Y}^n$ as the Hilbert space of vector-valued functions where every $\mathbf{u} \in \mathcal{Y}^n$ is a vector of functions $(u_1, \dots, u_n)^T$ with each $u_i \in \mathcal{Y}^n$. For all $h, g \in \mathcal{Y}^n$ we define the inner product $\langle \mathbf{h}, \mathbf{g}\rangle_{\mathcal{Y}^n} = \sum_{i=1}^n \langle h_i, g_i\rangle_\mathcal{Y}$. Similar to before, $\mathcal{L}(\mathcal{Y}^n)$ denotes the space of bounded linear operators from $\mathcal{Y}^n$ to $\mathcal{Y}^n$.

Corollary 1 *Define the positive semidefinite self-adjoint linear operators $\mathcal{M}, \mathcal{M}_\lambda \in \mathcal{L}(\mathcal{Y}^n)$ for all $\mathbf{h} \in \mathcal{Y}^n$ by*

$$\begin{aligned}[\mathcal{M}_\lambda(\mathbf{h})]_i &= [\mathcal{M}(\mathbf{h})]_i + n\lambda h_i \\ [\mathcal{M}(\mathbf{h})]_i &= \sum_{j=1}^n K(f_j, f_i) h_j.\end{aligned} \tag{133}$$

Any minimizer of

$$\widehat{O}_{n\lambda} \in \operatorname*{arg\,min}_{O \in \mathcal{O}} \Big[\frac{1}{n}\sum_{i=1}^n \Big\| u_i - O(f_i)) \Big\|_\mathcal{Y}^2 + \lambda \| \operatorname*{proj}_{\mathcal{O}_1} O\|_{\mathcal{O}_1}^2\Big] \tag{134}$$

must be of the form

$$\widehat{O}_{n\lambda}(x,y) = \sum_{k=1}^r d_k E_k + \sum_{i=1}^n K(f_i, \cdot) c_i \tag{135}$$

Assuming that the matrix $\mathbf{A} \in \mathbb{R}^{r \times r}$ defined by

$$\mathbf{A} = \begin{bmatrix} \langle \mathbf{T}^1, \mathcal{M}_\lambda^{-1}(\mathbf{T}^1)\rangle_{\mathcal{Y}^n} & \dots & \langle \mathbf{T}^1, \mathcal{M}_\lambda^{-1}(\mathbf{T}^r)\rangle_{\mathcal{Y}^n} \\ \vdots & & \vdots \\ \langle \mathbf{T}^r, \mathcal{M}_\lambda^{-1}(\mathbf{T}^1)\rangle_{\mathcal{Y}^n} & \dots & \langle \mathbf{T}^r, \mathcal{M}_\lambda^{-1}(\mathbf{T}^r)\rangle_{\mathcal{Y}^n} \end{bmatrix}. \tag{136}$$

is invertible, an optimal choice of the weights $\mathbf{d} \in \mathbb{R}^n, \mathbf{c} \in \mathcal{Y}^n$ *is*

$$\mathbf{c} = \mathcal{M}_\lambda^{-1}(\mathbf{u}) - \sum_{k=1}^{r} d_k \mathcal{M}_\lambda^{-1}(\mathbf{T}^k), \quad \mathbf{d} = \mathbf{A}^{-1} \begin{bmatrix} \langle \mathbf{T}^1, \mathcal{M}_\lambda^{-1}(\mathbf{u}) \rangle_{\mathcal{Y}^n} \\ \vdots \\ \langle \mathbf{T}^r, \mathcal{M}_\lambda^{-1}(\mathbf{u}) \rangle_{\mathcal{Y}^n} \end{bmatrix} \tag{137}$$

Proof [Corollary 1]
As shorthand, let us define the functions $T_i^k \in \mathcal{Y}$ for $i \in [n], k \in [r]$ where

$$T_i^k = E_k(f_i) \tag{138}$$

and the bounded linear operators $\Sigma_{ij} \in \mathcal{L}(\mathcal{Y})$ given by

$$\Sigma_{ij} = K(f_j, f_i). \tag{139}$$

Note that in fact,

$$\langle K(f_i, \cdot)c_i, K(f_j, \cdot)c_j \rangle_{\mathcal{O}_1} = \langle c_i, K(f_j, f_i)c_j \rangle_{\mathcal{Y}} = \langle c_i, \Sigma_{ij} c_j \rangle_{\mathcal{Y}} \tag{140}$$

by the reproducing property of K and $\Sigma_{ij}^* = \Sigma_{ji}$. Also, all of the Σ_{ij} are positive semidefinite linear operators by property (ii) of Definition 9. Hence, from the derivations in Theorem 11 we know that the optimal $\widehat{O}_{n\lambda}$ in (128) satisfies

$$\widehat{O}_{n\lambda}(f_i) = \sum_{k=1}^{r} d_k T_i^k + \sum_{j=1}^{n} \Sigma_{ij} c_j, \qquad \| \underset{\mathcal{O}_1}{\mathsf{proj}}\, \widehat{O}_{n\lambda} \|_{\mathcal{O}_1}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \langle c_i, \Sigma_{ij} c_j \rangle_{\mathcal{Y}}. \tag{141}$$

Now fixing $\mathbf{d} \in \mathbb{R}^r$ we need to find the functions $c_i \in \mathcal{Y}$ that minimize

$$\begin{aligned} \sum_{i=1}^{n} \Big\| u_i - \sum_{k=1}^{r} d_k T_i^k - \sum_{j=1}^{n} \Sigma_{ij} c_j \Big\|_{\mathcal{Y}}^2 + n\lambda \sum_{i=1}^{n} \sum_{j=1}^{n} \langle c_i, \Sigma_{ij} c_j \rangle_{\mathcal{Y}} \\ = -2 \sum_{i=1}^{n} \sum_{j=1}^{n} \Big\langle u_i - \sum_{k=1}^{r} d_k T_i^k, \Sigma_{ij} c_j \Big\rangle_{\mathcal{Y}} \\ + \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \Big\langle \Sigma_{ij} c_j, \Sigma_{ik} c_k \Big\rangle_{\mathcal{Y}} \\ + n\lambda \sum_{i=1}^{n} \sum_{j=1}^{n} \langle c_i, \Sigma_{ij} c_j \rangle_{\mathcal{Y}} + \text{const.} \end{aligned} \tag{142}$$

Taking the Hilbert space variational derivative of the expression above in $c_i \in \mathcal{Y}$ and setting it to zero we get that

$$-\sum_{j=1}^{n} \Sigma_{ij} \Big(u_j - \sum_{k=1}^{r} d_k T_j^k \Big) + \sum_{j=1}^{n} \sum_{k=1}^{n} \Sigma_{ij} \Sigma_{jk} c_k + n\lambda \sum_{j=1}^{n} \Sigma_{ij} c_j = 0 \tag{143}$$

where we have repeatedly used the fact that $\Sigma_{ij}^* = \Sigma_{ji}$. We can rewrite this as

$$\sum_{j=1}^{n} \sum_{k=1}^{n} \Sigma_{ij} \Big(\Sigma_{jk} + n\lambda \delta_{jk} \Big) c_k = \sum_{j=1}^{n} \Sigma_{ij} \Big(u_j - \sum_{k=1}^{r} d_k T_j^k \Big) \tag{144}$$

where δ_{jk} is the Kronecker delta function.

Now define the bounded linear operators $\mathcal{M}, \mathcal{M}_\lambda \in \mathcal{L}(\mathcal{Y}^n)$ for all $\mathbf{h} \in \mathcal{Y}^n$ by

$$\begin{aligned} [\mathcal{M}_\lambda(\mathbf{h})]_i &= [\mathcal{M}(\mathbf{h})]_i + n\lambda h_i \\ [\mathcal{M}(\mathbf{h})]_i &= \sum_{j=1}^{n} \Sigma_{ij} h_j. \end{aligned} \tag{145}$$

We can use the definition of $\mathcal{M}_\lambda$ to write (144) as

$$\sum_{j=1}^{n} \Sigma_{ij} [\mathcal{M}_\lambda(\mathbf{c})]_j = \sum_{j=1}^{n} \Sigma_{ij} \Big(u_j - \sum_{k=1}^{r} d_k T_j^k \Big). \tag{146}$$

and using the definition of $\mathcal{M}$ this can further be written as

$$\mathcal{M}\Big(\mathcal{M}_\lambda(\mathbf{c})\Big) = \mathcal{M}\Big(\mathbf{u} - \sum_{k=1}^{r} d_k \mathbf{T}^k\Big). \tag{147}$$

It is clear from the expression above that if the linear operator $\mathcal{M}_\lambda \in \mathcal{L}(\mathcal{Y})$ is invertible then

$$\mathbf{c} = \mathcal{M}_\lambda^{-1}(\mathbf{u}) - \sum_{k=1}^{r} d_k \mathcal{M}_\lambda^{-1}(\mathbf{T}^k). \tag{148}$$

is a solution. Note that $\mathcal{M}_\lambda$ is a sum of positive semidefinite operators Σ_{ij} plus λ times the identity operator. Hence $\mathcal{M}_\lambda$ is indeed strictly positive definite and invertible for $\lambda > 0$. It is also self-adjoint under the inner product $\langle \mathbf{h}, \mathbf{g} \rangle_{\mathcal{Y}^n} = \sum_{i=1}^{n} \langle h_i, g_i \rangle_{\mathcal{Y}}$ since $\Sigma_{ij}^* = \Sigma_{ji}$. This also immediately implies that $\mathcal{M}_\lambda^{-1}$ is strictly positive definite and self-adjoint.

Taking this value of $\mathbf{c}$ in (148) we need to find the coefficients $\mathbf{d} \in \mathbb{R}^r$ that minimize

$$\begin{aligned} \sum_{i=1}^{n} \Big\| u_i - \sum_{k=1}^{r} d_k T_i^k - \sum_{j=1}^{n} \Sigma_{ij} c_j \Big\|_{\mathcal{Y}}^2 &+ n\lambda \sum_{i=1}^{n} \sum_{j=1}^{n} \langle c_i, \Sigma_{ij} c_j \rangle_{\mathcal{Y}} \\ = \sum_{i=1}^{n} \Big\| \Big(u_i - \sum_{j=1}^{n} \Sigma_{ij} [\mathcal{M}_\lambda^{-1}(\mathbf{u})]_j \Big) &- \sum_{k=1}^{r} d_k \Big(T_i^k - \sum_{j=1}^{n} \Sigma_{ij} [\mathcal{M}_\lambda^{-1}(\mathbf{T}^k)]_j \Big) \Big\|_{\mathcal{Y}}^2 \\ &- 2n\lambda \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{r} d_k \Big\langle [\mathcal{M}_\lambda^{-1}(\mathbf{u})]_i, \Sigma_{ij} [\mathcal{M}_\lambda^{-1}(\mathbf{T}^k)]_j \Big\rangle_{\mathcal{Y}} \\ &+ n\lambda \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{r} \sum_{l=1}^{r} d_k d_l \Big\langle [\mathcal{M}_\lambda^{-1}(\mathbf{T}^k)]_i, \Sigma_{ij} [\mathcal{M}_\lambda^{-1}(\mathbf{T}^l)]_j \Big\rangle_{\mathcal{Y}} + \text{const.} \end{aligned} \tag{149}$$

Noting that $u_i = \sum_{j=1}^n \Sigma_{ij}[\mathcal{M}_\lambda^{-1}(\mathbf{u})]_j + n\lambda[\mathcal{M}_\lambda^{-1}(\mathbf{u})]_i$ and $T_i^k = \sum_{j=1}^n \Sigma_{ij}[\mathcal{M}_\lambda^{-1}(\mathbf{T}^k)]_j + n\lambda[\mathcal{M}_\lambda^{-1}(\mathbf{T}^k)]_i$ we can rewrite the above expression as

$$\begin{aligned} n^2\lambda^2 \sum_{i=1}^n \Big\| \mathcal{M}_\lambda^{-1}(\mathbf{u})]_i - \sum_{k=1}^r d_k \mathcal{M}_\lambda^{-1}(\mathbf{T}^k)]_i \Big\|^2_{L^2(D_\mathcal{Y})} \\ - 2n\lambda \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^r d_k \Big\langle [\mathcal{M}_\lambda^{-1}(\mathbf{u})]_i, \Sigma_{ij}[\mathcal{M}_\lambda^{-1}(\mathbf{T}^k)]_j \Big\rangle_\mathcal{Y} \\ + n\lambda \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^r \sum_{l=1}^r d_k d_l \Big\langle [\mathcal{M}_\lambda^{-1}(\mathbf{T}^k)]_i, \Sigma_{ij}[\mathcal{M}_\lambda^{-1}(\mathbf{T}^l)]_j \Big\rangle_\mathcal{Y} + \text{const.} \end{aligned} \tag{150}$$

Dividing through by $n\lambda$ and setting the derivative in d_k to zero we get

$$\begin{aligned} - n\lambda \sum_{i=1}^n \Big\langle \mathcal{M}_\lambda^{-1}(\mathbf{u})]_i, \mathcal{M}_\lambda^{-1}(\mathbf{T}^k)]_i \Big\rangle_\mathcal{Y} + n\lambda \sum_{i=1}^n \sum_{l=1}^r d_l \Big\langle \mathcal{M}_\lambda^{-1}(\mathbf{T}^k)]_i, \mathcal{M}_\lambda^{-1}(\mathbf{T}^l)]_i \Big\rangle_\mathcal{Y} \\ - \sum_{i=1}^n \sum_{j=1}^n \Big\langle [\mathcal{M}_\lambda^{-1}(\mathbf{u})]_i, \Sigma_{ij}[\mathcal{M}_\lambda^{-1}(\mathbf{T}^k)]_j \Big\rangle_\mathcal{Y} \\ + \sum_{i=1}^n \sum_{j=1}^n \sum_{l=1}^r d_l \Big\langle [\mathcal{M}_\lambda^{-1}(\mathbf{T}^k)]_i, \Sigma_{ij}[\mathcal{M}_\lambda^{-1}(\mathbf{T}^l)]_j \Big\rangle_\mathcal{Y} = 0 \end{aligned} \tag{151}$$

which we can further rewrite as

$$\begin{aligned} \sum_{i=1}^n \sum_{l=1}^r d_l \Big\langle [\mathcal{M}_\lambda^{-1}(\mathbf{T}^k)]_i, \Big(\sum_{j=1}^n \Sigma_{ij}[\mathcal{M}_\lambda^{-1}(\mathbf{T}^l)]_j + n\lambda[\mathcal{M}_\lambda^{-1}(\mathbf{T}^l)]_i \Big) \Big\rangle_\mathcal{Y} \\ = \sum_{i=1}^n \Big\langle [\mathcal{M}_\lambda^{-1}(\mathbf{u})]_i, \Big(\sum_{j=1}^n \Sigma_{ij}[\mathcal{M}_\lambda^{-1}(\mathbf{T}^k)]_j + n\lambda[\mathcal{M}_\lambda^{-1}(\mathbf{T}^k)]_i \Big) \Big\rangle_\mathcal{Y}. \end{aligned} \tag{152}$$

Finally, noting that the terms in the parentheses are T_i^l, T_i^k respectively we can write

$$\sum_{l=1}^r d_l \sum_{i=1}^n \langle T_i^l, [\mathcal{M}_\lambda^{-1}(\mathbf{T}^k)]_i \rangle_\mathcal{Y} = \sum_{i=1}^n \langle T_i^k, [\mathcal{M}_\lambda^{-1}(\mathbf{u})]_i \rangle_\mathcal{Y}. \tag{153}$$

Using the shorthand notation $\langle \mathbf{h}, \mathbf{g} \rangle_{\mathcal{Y}^n} = \sum_{i=1}^n \langle h_i, g_i \rangle_\mathcal{Y}$ and the fact that $\mathcal{M}_\lambda^{-1}$ is self-adjoint this becomes

$$\sum_{l=1}^r \langle \mathbf{T}^k, \mathcal{M}_\lambda^{-1}(\mathbf{T}^l) \rangle_{\mathcal{Y}^n} d_l = \langle \mathbf{T}^k, \mathcal{M}_\lambda^{-1}(\mathbf{u}) \rangle_{\mathcal{Y}^n} \tag{154}$$

for $k = 1, \ldots, n$. Hence, assuming that

$$\mathbf{A} = \begin{bmatrix} \langle \mathbf{T}^1, \mathcal{M}_\lambda^{-1}(\mathbf{T}^1) \rangle_{\mathcal{Y}^n} & \ldots & \langle \mathbf{T}^1, \mathcal{M}_\lambda^{-1}(\mathbf{T}^r) \rangle_{\mathcal{Y}^n} \\ \vdots & & \vdots \\ \langle \mathbf{T}^r, \mathcal{M}_\lambda^{-1}(\mathbf{T}^1) \rangle_{\mathcal{Y}^n} & \ldots & \langle \mathbf{T}^r, \mathcal{M}_\lambda^{-1}(\mathbf{T}^r) \rangle_{\mathcal{Y}^n} \end{bmatrix}. \tag{155}$$

is invertible we find that

$$\mathbf{d} = \mathbf{A}^{-1} \begin{bmatrix} \langle \mathbf{T}^1, \mathcal{M}_\lambda^{-1}(\mathbf{u}) \rangle_{\mathcal{Y}^n} \\ \vdots \\ \langle \mathbf{T}^r, \mathcal{M}_\lambda^{-1}(\mathbf{u}) \rangle_{\mathcal{Y}^n} \end{bmatrix} \tag{156}$$

which is what we desired to prove. ■

Appendix C. Rearrangements of Eigenvalues

In Proposition 2 we encountered the problem where it was necessary to bound the rate of decay of eigenvalues ρ_{ij} of a multidimensional kernel. For a simple example, if we have a kernel $K = k_1 \otimes k_2$ where the eigenvalues of k_1 decay with rate i^{-r_1} and the eigenvalues of k_2 decay with rate j^{-r_2} then the eigenvalues of K decay as $\rho_{ij} \asymp i^{-r_1} j^{-r_2}$. Sorting the eigenvalues ρ_{ij} and determining their rate of decay exactly is a difficult problem. However, as in Example 2 of Section 4.2 it is possible to enumerate these eigenvalues using a bijection $\pi : \mathbb{N}^2 \to \mathbb{N}$ such as the Cantor pairing function. This allows us to easily obtain bounds on the rate of decay of the eigenvalues $\rho_k := \rho_{\pi^{-1}(k)}$ under this ordering. However, these error bounds cannot be immediately applied to bound the rate of decrease of the sorted eigenvalues $\rho_1 \geq \rho_2 \geq \dots$

Here we show that any monotonic upper bound for a sequence is still an upper bound for the sequence sorted in decreasing order. First we prove the following lemma to show that any sequence of eigenvalues tending to zero can be reordered into sorted decreasing order by some bijection.

Lemma 13 *A sequence $\{a_k\}_{k=1}^{\infty}$ satisfies $\lim_{k\to\infty} a_k = \inf_{k\geq 1} a_k$ if and only if there exists a bijection $\pi : \mathbb{N} \to \mathbb{N}$ such that $a_{\pi(1)} \geq a_\pi(2) \geq \dots$*

Proof First let's assume that a_k satisfies $\lim_{k\to\infty} a_k = \inf_{k\geq 1} a_k$. Then we can construct the following bijection π. Define $U = \sup_{k\geq 1} a_k$ and $L = \inf_{k\geq 1} a_k$. Take all the $a_k \in [U, \frac{U}{2} + \frac{L}{2})$ and sort them. Continue this process by appending all of the $a_k \in [\frac{U}{2^n} + \frac{(2^n-1)L}{2^n}, \frac{U}{2^{n+1}} + \frac{(2^{n+1}-1)L}{2^{n+1}})$ in sorted order for all $n \geq 1$. This defines the bijection π from the original sequence of a_k's to the sorted decreasing sequence $a_1 \geq a_2 \geq \dots$

Now in the opposite direction, let's assume there exists a bijection π such that $a_{\pi(1)} \geq a_\pi(2) \geq \dots$ Then by the monotone convergence theorem we have that $\lim_{k\to\infty} a_{\pi(k)} = \inf_{k\geq 1} a_k$. Since a sequence's limit is invariant under rearrangements π, this proves that $\lim_{k\to\infty} a_k = \inf_{k\geq 1} a_k$. ■

Lemma 14 *Assume that a sequence $\{a_k\}_{k=1}^{\infty}$ is nonincreasing $a_1 \geq a_2 \geq \dots$ and bounded from below. Take any bijection $\pi : \mathbb{N} \to \mathbb{N}$ and assume that under this reordering we know that $a_{\pi(k)} \leq b_k$ for some nonincreasing sequence $b_1 \geq b_2 \geq \dots$ (i.e. an upper bound). Then this implies that $a_k \leq b_k$.*

Proof Note again that the a_k have a limit by the monotone convergence theorem and are bounded from above. Without loss of generality, the tightest nonincreasing upper bound for the sequence $a_{\pi(k)}$ is $b_k = \sup_{m\geq k} a_{\pi(m)} < \infty$ so it suffices to prove the lemma for this upper bound. We proceed to prove this result by contradiction. Assume there exists a $k \geq 1$ such that $a_k > b_k = \sup_{m\geq k} a_{\pi(m)}$. Then by the monotonicity of the a_k sequence this implies that $k < \pi(m)$ for all $m \geq k$. This can be rewritten as $\{\pi(m) : m \geq k\} \subseteq \{k+1, k+2, \dots\}$ which implies that $\{\pi^{-1}(m) : 1 \leq m \leq k\} \subseteq \{1, \dots, k-1\}$. But since π^{-1} is also a bijection then we have reached a contradication since the cardinality of $\{1, \dots, k\}$ is larger than $\{1, \dots, k-1\}$ so π^{-1} is not injective. ■

Appendix D. Decay of Kernel Eigenvalues

In this section, we describe two important cases in which eigenvalues of kernel functions decay at polynomial or exponential rates.

Example 1: Assume we have an RKHS with a reproducing kernel $K : D \times D \to \mathbb{R}$ on a compact domain D that is a Mercer kernel (e.g. continuous, square integrable, and nonnegative definite). Then we know that K is a bounded function on D so

$$\mathsf{Tr}(K) = \int_D K(x,x)\mathrm{d}x < \infty \tag{157}$$

Furthermore, by Mercer's theorem we know that K has the eigenexpansion

$$K(x,y) = \sum_{n=1}^{\infty} \lambda_n \phi_n(x) \phi_n(y) \tag{158}$$

for nonnegative eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$ and $L^2(D)$ orthonormal eigenfunctions ϕ_k. This implies that

$$\mathsf{Tr}(K) = \int_D K(x,x)\mathrm{d}x = \sum_{n=1}^{\infty} \lambda_n. \tag{159}$$

Since the series $\sum_{n=1}^{\infty} \lambda_n$ converges and the sequence $\{\lambda_n\}_{n=1}^{\infty}$ is nonnegative and decreasing, this implies that $n\lambda_n \to 0$ as $n \to \infty$. Hence, this proves that

$$\lambda_n \lesssim \frac{1}{n}. \tag{160}$$

where again the notation $a \lesssim b$ means that $a \leq Cb$ for some $C \geq 0$. In general, the spectrum of a reproducing kernel (assumed to be positive semidefinite) decreases at rate at least $\frac{1}{n}$ if it is a trace class linear operator. Also, this statement is sharp over the space of reproducing kernels as we can construct for any $\epsilon > 0$ a symmetric, positive semidefinite, square integrable reproducing kernel of the form

$$K(x,y) = \sum_{n=1}^{\infty} n^{-(1+\epsilon)} \phi_n(x) \phi_n(y). \tag{161}$$

for any $\epsilon > 0$. For an arbitrary domain D, as long as $L^2(D)$ is a separable Hilbert space we can always construct an infinitely large countable basis. Note that $\mathsf{Tr}(K) = \sum_{n=1}^{\infty} n^{-(1+\epsilon)} < \infty$ by the integral test.

Example 2: A large class of reproducing kernels can be built by taking tensor products of simpler ones. For example, taking the product of two squared exponential kernels $\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\xi)^2}{2\sigma_1^2}}$ and $\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{(y-\eta)^2}{2\sigma_2^2}}$ gives us another squared exponential kernel. In general, take m RKHSs with reproducing kernels $K_i : D_i \times D_i \to \mathbb{R}$ where $D_i \subset \mathbb{R}^{d_i}$ for $i = 1, \ldots, m$. Defining the product kernel $K : (D_1 \times \ldots \times D_m)^2 \to \mathbb{R}$ where

$$K(x_1, \ldots, x_m, y_1, \ldots, y_m) = \prod_{i=1}^{m} K_i(x_i, y_i) \tag{162}$$

then by Paulsen and Raghupathi (2016, Section 5, Theorem 5.24) we know that K is also a reproducing kernel of some RKHS. If we have information about the rate of decay of the eigenvalues of each kernel K_i, then we can bound the rate of decay of the tensor product kernel K.

Let us assume that the sorted eigenvalues of each K_i are $\{\lambda_k^{(i)}\}_{k=1}^{\infty}$ and that they decay polynmially at the rate $\lambda_k^{(i)} \lesssim k^{-r_i}$ for some $r_i > 0$. Note that the eigenvalues of the tensor product kernel K are $\rho_{k_1\ldots k_m} = \prod_{i=1}^{m} \lambda_{k_i}^{(i)}$ for all $k_1 \ldots k_m \geq 1$. To quantify the rate of decrease of the eigenvalues of K, we first order them as $\{\rho_n\}_{n=1}^{\infty}$ using the m-tupling function

$$\pi^{(m)}(k_1, \ldots, k_m) = \pi(\pi^{(m-1)}(k_1, \ldots, k_{m-1}), k_m), \quad m > 2 \tag{163}$$

defined recursively where

$$\pi^{(2)}(k_1, k_2) = \pi(k_1, k_2) = \frac{1}{2}(k_1 + k_2 - 2)(k_1 + k_2 - 1) + k_2 \tag{164}$$

is the Cantor pairing function. For any $\rho_n = \rho_{k_1\ldots k_m}$ in our ordered sequence where $n = \pi^{(m)}(k_1, \ldots, k_m)$, define the sum $s(n) = \sum_{i=1}^{m} k_i$. It is not hard to check that $\prod_{i=1}^{m} x_i^{-r_i}$ for $(x_1, \ldots, x_m) \in \mathbb{R}^m$ is a convex function on the convex constraint set $x_i \geq 1$ with $\sum_{i=1}^{m} x_i = s(n)$. Therefore, the global maximum of this function is located at one of the extremal points $x_i = s(n) - m + 1$ with $x_j = 1$ for all $j \neq i$. As shorthand, write $[m] = \{1, \ldots, m\}$. Defining $i_{\min} = \arg\min_{i\in[m]} r_i$ and $r_{\min} = r_{i_{\min}}$ we know that this function over the convex constraint set is maximized at the extremal point $x_{i_{\min}} = s(n) - m + 1$ with $x_j = 1$ for all $j \neq i_{\min}$. The maximal value it takes there is $(s(n) - m + 1)^{-r_{\min}}$. This implies that

$$\rho_n = \rho_{k_1,\ldots,k_m} = \prod_{i=1}^{m} \lambda_{k_i}^{(i)} \lesssim \prod_{i=1}^{m} k_i^{-r_i} \lesssim s(n)^{-r_{\min}} \tag{165}$$

where again the notation $a \lesssim b$ means that $a \leq Cb$ for some $C \geq 0$. If we iterate over m-tuples $(k_1, \ldots, k_m)$ in the order prescribed by the pairing function $\pi^{(m)}$, then by the time we have reached a tuple where $n = \pi^{(m)}(k_1, \ldots, k_m)$ we must have iterated over *at most* all of the positive m-tuples with sums $m, \ldots, s(n)$. The number of such m-tuples is exactly $\sum_{k=m}^{s(n)} \binom{k-1}{m-1}$ because $\binom{k-1}{m-1}$ is the number of tuples whose sum is exactly k through a stars and bars argument. Now by the Hockey-stick identity we see that $\sum_{k=m}^{s(n)} \binom{k-1}{m-1} = \binom{s(n)}{m} \lesssim s(n)^m$. Finally, this implies that $n \lesssim s(n)^m$ so $s(n) \gtrsim n^{\frac{1}{m}}$ and hence

$$\rho_n \lesssim n^{-\frac{r_{\min}}{m}}. \tag{166}$$

Since the ρ_n enumerated by the m-tupling function $\pi^{(m)}$ tend to zero, then by Lemma 14 the sorted eigenvalues $\rho_1 \geq \rho_2 \geq \dots$ of K also decrease at the same rate as above.

Following a similar analysis, if all of the sorted eigenvalues $\{\lambda_k^{(i)}\}_{k=1}^{\infty}$ of K_i decay exponentially at the rate $\lambda_k^{(i)} \lesssim \exp(-a_i k^{r_i})$ for all $i \in [d]$ then under the constraint $s(n) = \sum_{i=1}^{m} k_i$ we have that

$$\rho_n = \rho_{k_1,\dots,k_m} = \prod_{i=1}^{m} \lambda_{k_i}^{(i)} = e^{-\sum_{i=1}^{m} r_i k_i} \lesssim e^{-a_{i_{\min}} s(n)^{r_{i_{\min}}}}. \tag{167}$$

where $i_{\min} = \arg\min_{i \in [m]} r_i$. Since we know that $s(n) \gtrsim n^{\frac{1}{m}}$ then finally

$$\rho_n \lesssim \exp\Big(- a_{i_{\min}} \cdot n^{\frac{r_{i_{\min}}}{m}} \Big). \tag{168}$$

Again by Lemma 14 the sorted eigenvalues $\rho_1 \geq \rho_2 \geq \dots$ of K also decrease at the same rate as above.

Appendix E. SubGaussian Functions

In this appendix we define what it means for a random function to be subgaussian with respect to a variance proxy operator, similar to the definition of subgaussianity for real random variables.

Definition 15 *Let $\Gamma : L^2(D) \to L^2(D)$ be a positive semidefinite trace class linear operator. A random variable $F \in L^2(D)$ is subgaussian with respect to variance proxy Γ (written as $F \sim \mathsf{subG}(\Gamma)$) if there exists an $\alpha \geq 0$ such that for all $f \in L^2(D)$,*

$$\mathbb{E}\Big[e^{\langle f, F-\mathbb{E}[F]\rangle_{L^2(D)}}\Big] \leq e^{\alpha^2 \langle \Gamma f, f\rangle_{L^2(D)}/2} \tag{169}$$

Furthermore, if $F \sim \mathsf{subG}(\Gamma)$ then the ψ_2-norm of F with respect to Γ is defined as

$$\|F\|_{\psi_2,\Gamma} = \inf\Big\{\alpha \geq 0 : \mathbb{E}\Big[e^{\langle f, F-\mathbb{E}[F]\rangle_{L^2(D)}}\Big] \leq e^{\alpha^2 \langle \Gamma f, f\rangle_{L^2(D)}/2},\ \forall f \in L^2(D)\Big\}. \tag{170}$$

We denote the covariance operator of F as the positive semidefinite linear operator $\Sigma_F : L^2(D_\mathcal{Y}) \to L^2(D_\mathcal{Y})$ which is identified with the function

$$\Sigma_F = \mathbb{E}[(F - \mathbb{E}[F]) \otimes (F - \mathbb{E}[F])]. \tag{171}$$

By Chen and Yang (2021, Section 2.1, Lemma 2.4) we know that

$$\Sigma_F \preceq 4\|F\|^2_{\psi_2,\Gamma_F}\Gamma_F \tag{172}$$

and we also make the following key assumption.

Definition 16 (Strict subgaussianity) *The random variable $F \sim \mathsf{subG}(\Gamma_F)$ is called strictly subgaussian if its covariance operator and covariance proxy satisfy*

$$\Gamma_F \preceq C\Sigma_F \tag{173}$$

for some fixed constant $C > 0$.

The definition above is a natural extension of *strict subgaussianity* to random variables in Hilbert spaces. It is trivially satisfied for Gaussian random variables $Z \in L^2(D)$ for which it is easy to check that $Z \sim \mathsf{subG}(\Sigma)$ where $\Sigma = \mathbb{E}[(Z - \mathbb{E}[Z]) \otimes (Z - \mathbb{E}[Z])]$ and $\|Z\|_{\psi_2,\Sigma} = 1$.

Lemma 17 *Take any subgaussian random variable $F \sim \mathsf{subG}(\Gamma)$ in $L^2(D_\mathcal{X})$ where $\Gamma \in L^2(D_\mathcal{X} \times D_\mathcal{X})$ is a trace class linear operator. Then for any operator $H \in L^2(D_\mathcal{X} \times D_\mathcal{Y})$ we have that $H(F) \sim \mathsf{subG}(H\Gamma H^*)$ and $\|H(F)\|_{\psi_2,H\Gamma H^*} \leq \|F\|_{\psi_2,\Gamma}$.*

Proof First since $F \sim \mathsf{subG}(\Gamma)$ then there exists an $\alpha \geq 0$ such that for all $f \in L^2(D_\mathcal{X})$,

$$\mathbb{E}\Big[e^{\langle f, F-\mathbb{E}[F]\rangle_{L^2(D_\mathcal{X})}}\Big] \leq e^{\alpha^2 \langle \Gamma f, f\rangle_{L^2(D_\mathcal{X})}/2}. \tag{174}$$

So for the same α we can write for all $g \in L^2(D_\mathcal{Y})$,

$$\mathbb{E}\Big[e^{\langle g, H(F)-\mathbb{E}[H(F)]\rangle_{L^2(D_\mathcal{Y})}}\Big] = \mathbb{E}\Big[e^{\langle H^*(g), F-\mathbb{E}[F]\rangle_{L^2(D_\mathcal{X})}}\Big] \leq e^{\alpha^2 \langle H\Gamma H^*(g), g\rangle_{L^2(D_\mathcal{Y})}/2} \tag{175}$$

which proves that $H(F) \sim \mathsf{subG}(H\Gamma H^*)$. Furthermore, we know that the ψ_2-norm of F with respect to Γ is defined as

$$\|F\|_{\psi_2,\Gamma} = \inf\Big\{\alpha \geq 0 : \mathbb{E}\Big[e^{\langle g, F-\mathbb{E}[F]\rangle_{L^2(D_\mathcal{X})}}\Big] \leq e^{\alpha^2\langle \Gamma g, g\rangle_{L^2(D_\mathcal{X})}/2},\ \forall g \in L^2(D_\mathcal{X})\Big\}. \tag{176}$$

which implies by the derivations above that

$$\begin{aligned}\|F\|_{\psi_2,\Gamma} &\geq \inf\Big\{\alpha \geq 0 : \mathbb{E}\Big[e^{\langle g, H(F)-\mathbb{E}[H(F)]\rangle_{L^2(D_\mathcal{Y})}}\Big] \leq e^{\alpha^2\langle H\Gamma H^*(g), g\rangle_{L^2(D_\mathcal{Y})}/2},\ \forall g \in L^2(D_\mathcal{Y})\Big\}\\ &= \|H(F)\|_{\psi_2,H\Gamma H^*}.\end{aligned} \tag{177}$$

■

Remark 18 *Define the usual ψ_1-norm for a (possibly noncentered) subexponential real-valued random variable R,*

$$\|R\|_{\psi_1} = \inf\Big\{t > 0 : \mathbb{E}\Big[\exp\Big(\frac{|R|}{t}\Big)\Big] \leq 2\Big\}. \tag{178}$$

Then for any centered subgaussian vectors in $L^2(D)$ denoted by $X \sim \mathsf{subG}(\Sigma)$ and $Y \sim \mathsf{subG}(\Gamma)$ we have that

$$\begin{aligned}\|\langle X,Y\rangle_{L^2(D)}\|_{\psi_1} &:= \inf\Big\{t > 0 : \mathbb{E}\Big[\exp\Big(\frac{|\langle X, Y\rangle_{L^2(D)}|}{t}\Big)\Big] \leq 2\Big\}\\ &\leq \inf\Big\{t > 0 : \mathbb{E}\Big[\exp\Big(\frac{\|X\|_{L^2(D)}\|Y\|_{L^2(D)}}{t}\Big)\Big] \leq 2\Big\}\\ &\leq \Big\|\|X\|_{L^2(D)}\Big\|_{\psi_2}\Big\|\|Y\|_{L^2(D)}\Big\|_{\psi_2}\\ &= \Big\|\|X\|^2_{L^2(D)}\Big\|^{\frac{1}{2}}_{\psi_1}\Big\|\|Y\|^2_{L^2(D)}\Big\|^{\frac{1}{2}}_{\psi_1}\end{aligned} \tag{179}$$

By Chen and Yang (2021, Appendix A.2, Lemma A.4) we know there exists a universal constant $c > 0$ such that $\|\|X\|^2_{L^2(D)}\Big\|_{\psi_1} \leq c\|X\|^2_{\psi_2,\Sigma}\,\mathsf{Tr}(\Sigma)$ which implies that

$$\|\langle X, Y\rangle_{L^2(D)}\|_{\psi_1} \leq c\|X\|_{\psi_2,\Sigma}\|Y\|_{\psi_2,\Gamma}\,\mathsf{Tr}(\Sigma)^{\frac{1}{2}}\,\mathsf{Tr}(\Gamma)^{\frac{1}{2}}. \tag{180}$$

Appendix F. Results for Error Analysis

In this appendix we derive all of the results for the error analysis of our Green's function estimator. The following results closely follow the derivations in Yuan et al. (2010) for bounding the error of scalar-output functional linear regression and extend them to the case of functional-output functional linear regression.

Proof [Proposition 1] First note that

$$\|G\|^2_{\Sigma_F} \leq \mu_1 \int_{D_\mathcal{Y}} \|G(x,\cdot)\|^2_{L^2(D_\mathcal{X})} \mathrm{d}x = \mu_1 \|G\|^2_{L^2(D_\mathcal{X}\times D_\mathcal{Y})} \leq c_1 \|G\|^2_\mathcal{G} \tag{181}$$

where the last inequality follows by Cauchy–Schwarz since

$$\begin{aligned}\|G\|^2_{L^2(D_\mathcal{X}\times D_\mathcal{Y})} &= \int_{D_\mathcal{Y}}\int_{D_\mathcal{X}} \langle K_{(x,y)}, G\rangle^2_\mathcal{G} \mathrm{d}x\mathrm{d}y \leq \int_{D_\mathcal{Y}}\int_{D_\mathcal{X}} \|K_{(x,y)}\|^2_\mathcal{G}\|G\|^2_\mathcal{G}\mathrm{d}x\mathrm{d}y \\ &= \Big(\int_{D_\mathcal{Y}}\int_{D_\mathcal{X}} K(x,y,x,y)^2 \mathrm{d}x\mathrm{d}y\Big)\|G\|^2_\mathcal{G}\end{aligned} \tag{182}$$

where the integral on the right hand side is finite since D is closed and bounded and K is continuous. From this, since $J(G) = \|G\|^2_\mathcal{G}$ we immediately have that

$$\|G\|^2_{\overline{K}} = \|G\|^2_{\Sigma_F} + J(G) \leq (c_1+1)\|G\|^2_\mathcal{G}. \tag{183}$$

The reverse inequality follows immediately since by definition

$$\|G\|^2_\mathcal{G} = J(G) \leq \|G\|^2_{\overline{K}}. \tag{184}$$

for any $G \in \mathcal{G}$. This proves that $\|\cdot\|_\mathcal{G}$ and $\|\cdot\|_{\overline{K}}$ are equivalent norms on $\mathcal{G}$.

From this we immediately see that $\|\cdot\|_{\overline{K}}$ is a valid norm on $\mathcal{G}$ since it is zero only at $0 \in \mathcal{G}$ and finite for all $G \in \mathcal{G}$. Furthermore, $\mathcal{G}$ equipped with $\|\cdot\|_\mathcal{G}$ is an RKHS iff all the linear evaluation functionals $L_{(x,y)} : G \to G(x,y)$ are bounded

$$|L_{(x,y)}(G)| := |G(x,y)| \leq M\|G\|_\mathcal{G},\ \forall G \in \mathcal{G}. \tag{185}$$

This proves that $\mathcal{G}$ equipped with $\|\cdot\|_{\overline{K}}$ is also an RKHS because all of the linear functionals are once again bounded

$$|L_{(x,y)}(G)| := |G(x,y)| \leq M\|G\|_\mathcal{G} \leq M\|G\|_{\overline{K}},\ \forall G \in \mathcal{G}. \tag{186}$$

■

Proof [Theorem 5]
First we write out for any $G \in \mathcal{G}$,

$$\begin{aligned}\overline{K}^{-\frac{1}{2}}G &= \sum_{k=1}^\infty \langle \overline{K}^{-\frac{1}{2}}G, \Gamma_k\rangle_{L^2(D_\mathcal{X}\times D_\mathcal{Y})}\Gamma_k = \sum_{k=1}^\infty \langle \overline{K}^{-\frac{1}{2}}G, \nu_k^{\frac{1}{2}}\overline{K}^{-\frac{1}{2}}\Omega_k\rangle_{L^2(D_\mathcal{X}\times D_\mathcal{Y})}\nu_k^{\frac{1}{2}}\overline{K}^{-\frac{1}{2}}\Omega_k \\ &= \sum_{k=1}^\infty \nu_k \langle \overline{K}^{-1}G, \Omega_k\rangle_{L^2(D_\mathcal{X}\times D_\mathcal{Y})}\overline{K}^{-\frac{1}{2}}\Omega_k = \overline{K}^{-\frac{1}{2}}\Big(\sum_{k=1}^\infty \nu_k\langle G, \Omega_k\rangle_{\overline{K}}\Omega_k\Big).\end{aligned} \tag{187}$$

Applying $\overline{K}^{\frac{1}{2}}$ to both sides we see that

$$G = \sum_{k=1}^{\infty} g_k \Omega_k, \quad g_k = \nu_k \langle G, \Omega_k \rangle_{\overline{K}} \tag{188}$$

which converges absolutely. Now let $\gamma_k = (\nu_k^{-1} - 1)^{-1}$. Then we can use the fact that $\langle \Omega_k, \Omega_j \rangle_{\overline{K}} = \nu_k^{-1}\delta_{kj}$ to write

$$\|G\|_{\overline{K}}^2 = \Big\langle \sum_{k=1}^{\infty} g_k \Omega_k, \sum_{j=1}^{\infty} g_j \Omega_j \Big\rangle_{\overline{K}} = \sum_{k=1}^{\infty} \nu_k^{-1} g_k^2 = \sum_{k=1}^{\infty} (1 + \gamma_k^{-1}) g_k^2. \tag{189}$$

Similarly, since $\langle (\Sigma_F \otimes I)\Omega_k, \Omega_j \rangle_{L^2(D_\mathcal{X} \times D_\mathcal{Y})} = \delta_{kj}$ we have

$$\langle (\Sigma_F \otimes I)G, G \rangle_{L^2(D_\mathcal{X} \times D_\mathcal{Y})} = \Big\langle \sum_{k=1}^{\infty} g_k (\Sigma_F \otimes I)\Omega_k, \sum_{j=1}^{\infty} g_j \Omega_j \Big\rangle_{L^2(D_\mathcal{X} \times D_\mathcal{Y})} = \sum_{k=1}^{\infty} g_k^2. \tag{190}$$

■

Proof [Proposition 2]
For any $\{\varphi_j\}_{j=1}^{\infty}$ which is an orthonormal basis of $L^2(D_\mathcal{X})$, it is not hard to see that $\{\phi_i \otimes \varphi_j\}_{i,j=1}^{\infty}$ is an eigenbasis of $\Sigma_F \otimes I$ where

$$(\Sigma_F \otimes I)(\phi_i \otimes \varphi_j) = \mu_i \phi_i \otimes \varphi_j. \tag{191}$$

By definition of the reproducing kernel K for $\mathcal{G}$, since $\mathrm{span}\{\phi_i \otimes \varphi_j\}_{i,j=1}^{\infty}$ are the eigenfunctions of K then we must have that $\mathcal{G} = \mathrm{span}\{\phi_i \otimes \varphi_j\}_{i,j=1}^{\infty}$. By the definition of the induced inner product, we know that for all $F, G \in \mathcal{G}$,

$$\langle F, G \rangle_{\overline{K}} = \langle (\Sigma_F \otimes I)F, G \rangle_{L^2(D_\mathcal{X} \times D_\mathcal{Y})} + \langle F, G \rangle_K. \tag{192}$$

Therefore, we can write out for all $1 \leq i, i', j, j' < \infty$,

$$\langle \phi_i \otimes \varphi_j, \phi_{i'} \otimes \varphi_{j'} \rangle_{\overline{K}} = \mu_i \delta_{ii'}\delta_{jj'} + \rho_{ij}^{-1}\delta_{ii'}\delta_{jj'} = (\mu_i + \rho_{ij}^{-1})\delta_{ii'}\delta_{jj'}. \tag{193}$$

Since $\overline{K}$ is invertible over $\mathcal{G}$ and $\langle \phi_i \otimes \varphi_j, \phi_{i'} \otimes \varphi_{j'} \rangle_{\overline{K}} = \langle \overline{K}^{-1}(\phi_i \otimes \varphi_j), \phi_{i'} \otimes \varphi_{j'} \rangle_{L^2(D_\mathcal{X} \times D_\mathcal{Y})}$ this implies that $\overline{K}$ has the eigendecomposition

$$\overline{K}(x, y, \xi, \eta) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} (\mu_j + \rho_{ij}^{-1})^{-1} \phi_i(x)\phi_i(\xi)\varphi_j(y)\varphi_j(\eta) \tag{194}$$

and since the $\{\phi_i\}_{i=1}^{\infty}$ and $\{\varphi_j\}_{j=1}^{\infty}$ are orthonormal in $L^2(D_\mathcal{X})$ we have that

$$\overline{K}(\phi_i \otimes \varphi_j) = (\mu_i + \rho_{ij}^{-1})^{-1} \phi_i \otimes \varphi_j \tag{195}$$

which implies that

$$\begin{aligned} \overline{K}^{\frac{1}{2}}(\Sigma_F \otimes I)\overline{K}^{\frac{1}{2}}(\phi_i \otimes \varphi_j) &= (\mu_i + \rho_{ij}^{-1})^{-\frac{1}{2}} \overline{K}^{\frac{1}{2}}(\Sigma_F \otimes I)(\phi_i \otimes \varphi_j) \\ &= \mu_i(\mu_i + \rho_{ij}^{-1})^{-\frac{1}{2}} \overline{K}^{\frac{1}{2}}(\phi_i \otimes \varphi_j) \\ &= (1 + \mu_i^{-1}\rho_{ij}^{-1})^{-1} \phi_i \otimes \varphi_j. \end{aligned} \tag{196}$$

This proves that the eigenfunctions of $\overline{K}^{\frac{1}{2}}(\Sigma_F \otimes I)\overline{K}^{\frac{1}{2}}$ are $\Gamma_{ij} = \Psi_{ij} = \phi_i \otimes \varphi_j$ with eigenvalues $\nu_{ij} = (1 + \mu_i^{-1}\rho_{ij}^{-1})^{-1}$. Therefore, we have that $\gamma_{ij} = (\nu_{ij}^{-1} - 1)^{-1} = \mu_i\rho_{ij}$ and $\Omega_{ij} = \nu_{ij}^{-\frac{1}{2}}\overline{K}^{\frac{1}{2}}\Gamma_{ij} = \mu_i^{-\frac{1}{2}}\phi_i \otimes \varphi_j = \mu_i^{-\frac{1}{2}}\Psi_{ij}$.

Now reorder the ρ_{ij} in decreasing order and enumerate them as ρ_k. Assume that these sorted decreasing eigenvalues satisfy $\rho_k \lesssim k^{-r}$ for some $r > \frac{1}{2}$ as in Assumption 4.

Let us reorder the coefficients γ_{ij} and enumerate them as γ_k in the same way as the ρ_k eigenvalues. Since the covariance Σ_F is a bounded operator then all of the eigenvalues μ_i are bounded. Finally, this proves that $\gamma_k \lesssim \rho_k \lesssim k^{-r}$ for some $r > \frac{1}{2}$. ■

Proof [Lemma 6] We would like to bound the term $\|\widehat{G}_{n,\lambda} - \overline{G}_{\infty,\lambda}\|_{\Sigma_F}$. Define the variational derivatives for any $G, P, H \in \mathcal{G}$,

$$\begin{aligned}
\delta\widehat{R}(G)[H] &= -\frac{2}{n}\sum_{i=1}^{n}\langle U_i - G(F_i), H(F_i)\rangle_{L^2(D_\mathcal{Y})} \\
\delta R(G)[H] &= -2\mathbb{E}\Big[\langle U - G(F), H(F)\rangle_{L^2(D_\mathcal{Y})}\Big] \\
\delta^2\widehat{R}[P,H] &= \frac{2}{n}\sum_{i=1}^{n}\langle P(F_i), H(F_i)\rangle_{L^2(D_\mathcal{Y})} = 2\langle(\widehat{\Sigma}_F \otimes I)P, H\rangle_{L^2(D_\mathcal{X}\times D_\mathcal{Y})} \\
\delta^2 R[P,H] &= 2\mathbb{E}\Big[\langle P(F), H(F)\rangle_{L^2(D_\mathcal{Y})}\Big] = 2\langle(\Sigma_F \otimes I)P, H\rangle_{L^2(D_\mathcal{X}\times D_\mathcal{Y})}.
\end{aligned} \tag{197}$$

where $\widehat{\Sigma}_F = \frac{1}{n}\sum_{i=1}^{n} F_i \otimes F_i$ and $\Sigma_F = \mathbb{E}[F \otimes F]$. Remember that since $\mathcal{G}$ equipped with $\|\cdot\|_{\overline{K}}$ is an RKHS, then $\|H\|_{L^2(D_\mathcal{X}\times D_\mathcal{Y})} \le c\|H\|_{\overline{K}}$ for some universal constant $c > 0$. Viewing the variational derivatives above as functionals of $H \in \mathcal{G}$, it is not hard to see that they are bounded in the $\|\cdot\|_{\overline{K}}$ norm. Hence, by the Riesz representation theorem there exist $\nabla\widehat{R}(G), \nabla R(G), \nabla^2\widehat{R}(P)$, and $\nabla^2 R(P) \in \mathcal{G}$ such that

$$\begin{aligned}
\delta\widehat{R}(G)[H] = \langle\nabla\widehat{R}(G), H\rangle_{\overline{K}}, \quad \delta R(G)[H] = \langle\nabla R(G), H\rangle_{\overline{K}} \\
\delta^2\widehat{R}[P,H] = \langle\nabla^2\widehat{R}(P), H\rangle_{\overline{K}}, \quad \delta^2 R[P,H] = \langle\nabla^2 R(P), H\rangle_{\overline{K}}.
\end{aligned} \tag{198}$$

Denoting $\widehat{R}_\lambda(G) = \widehat{R}(G) + \lambda J(G)$ and $R_\lambda(G) = R(G) + \lambda J(G)$ we can similarly compute the first and second variational derivatives $\delta\widehat{R}_\lambda(G)[H], \delta R_\lambda(G)[H], \delta^2\widehat{R}_\lambda[P,H], \delta^2 R_\lambda[P,H]$ and show that these functionals of $H \in \mathcal{G}$ are bounded in norm $\|\cdot\|_{\overline{K}}$. Hence, there exist $\nabla\widehat{R}_\lambda(G), \delta R_\lambda(G), \nabla^2\widehat{R}_\lambda(P)$, and $\nabla^2 R_\lambda(P) \in \mathcal{G}$ such that

$$\begin{aligned}
\delta\widehat{R}_\lambda(G)[H] = \langle\nabla\widehat{R}_\lambda(G), H\rangle_{\overline{K}}, \quad \delta R_\lambda(G)[H] = \langle\nabla R_\lambda(G), H\rangle_{\overline{K}} \\
\delta^2\widehat{R}_\lambda[P,H] = \langle\nabla^2\widehat{R}_\lambda(P), H\rangle_{\overline{K}}, \quad \delta^2 R_\lambda[P,H] = \langle\nabla^2 R_\lambda(P), H\rangle_{\overline{K}}.
\end{aligned} \tag{199}$$

We interpret the Hessians $\nabla^2\widehat{R}, \nabla^2 R, \nabla^2\widehat{R}_\lambda$, and $\nabla^2 R_\lambda$ as maps from $\mathcal{G}$ to $\mathcal{G}$. Since $\langle\Omega_k, \Omega_j\rangle_{\overline{K}} = \nu_k^{-1}\delta_{kj}$, this immediately implies for all $k \ge 1$ and $G \in \mathcal{G}$ that

$$\langle\nabla R(G), \Omega_k\rangle_{\overline{K}} = \delta R(G)[\Omega_k] \implies \nabla R(G) = \sum_{k=1}^{\infty}\nu_k\delta R(G)[\Omega_k]\Omega_k \tag{200}$$

and likewise for $\nabla\widehat{R}, \nabla\widehat{R}_\lambda, \nabla R_\lambda, \nabla^2\widehat{R}, \nabla^2 R, \nabla^2\widehat{R}_\lambda, \nabla^2 R_\lambda$. For all $G = \sum_{k=1}^{\infty} g_k\Omega_k \in \mathcal{G}$ we can write out a series expansion for $\nabla^2 R_\lambda$ as

$$\nabla^2 R_\lambda(G) = 2\sum_{k=1}^{\infty}\nu_k(1 + \lambda\gamma_k^{-1})g_k\Omega_k. \tag{201}$$

It is not hard to check using the series expansions from Theorem 5 that indeed

$$\langle\nabla^2 R_\lambda(G), H\rangle_{\overline{K}} = 2\langle G, H\rangle_{\Sigma_F}^2 + 2\lambda\langle G, H\rangle_K = \delta^2 R_\lambda[G, H] \tag{202}$$

for all $G, H \in \mathcal{G}$. Finally, we can define the "linearization" of $\widehat{G}_{n,\lambda}$ as $\tilde{G} \in \mathcal{G}$ where

$$\tilde{G} = \overline{G}_{\infty,\lambda} - \nabla^2 R_\lambda^{-1}\Big(\nabla \widehat{R}_\lambda(\overline{G}_{\infty,\lambda})\Big), \quad \nabla^2 R_\lambda^{-1}(G) = \frac{1}{2}\sum_{k=1}^{\infty} \nu_k^{-1}(1+\lambda\gamma_k^{-1})^{-1} g_k \Omega_k \tag{203}$$

following the analysis of the rate of convergence of penalized likelihood estimators in Cox and O'Sullivan (1990, Section 1, Equation 1.13). Now decomposing

$$\widehat{G}_{n,\lambda} - \overline{G}_{\infty,\lambda} = (\widehat{G}_{n,\lambda} - \tilde{G}) + (\tilde{G} - \overline{G}_{\infty,\lambda}) \tag{204}$$

we bound both terms on the right-hand side.

1. Bounding $\|\tilde{G} - \overline{G}_{\infty,\lambda}\|_{\Sigma_F}$

First noting that $\delta R_\lambda(\overline{G}_{\infty,\lambda}) = 0$ by definition then

$$\delta\widehat{R}_\lambda(\overline{G}_{\infty,\lambda}) = \delta\widehat{R}_\lambda(\overline{G}_{\infty,\lambda}) - \delta R_\lambda(\overline{G}_{\infty,\lambda}) = \delta\widehat{R}(\overline{G}_{\infty,\lambda}) - \delta R(\overline{G}_{\infty,\lambda}). \tag{205}$$

So for any $H \in \mathcal{G}$ we can write

$$\begin{aligned}\Big(\delta\widehat{R}_\lambda(\overline{G}_{\infty,\lambda})[H]\Big)^2 &= \Big(\delta\widehat{R}(\overline{G}_{\infty,\lambda})[H] - \delta R(\overline{G}_{\infty,\lambda})[H]\Big)^2 \\ &= 4\Big(\frac{1}{n}\sum_{i=1}^{n}\langle U_i - \overline{G}_{\infty,\lambda}(F_i), H(F_i)\rangle_{L^2(D_\mathcal{Y})} - \mathbb{E}\Big[\langle U - \overline{G}_{\infty,\lambda}(F), H(F)\rangle_{L^2(D_\mathcal{Y})}\Big]\Big)^2.\end{aligned} \tag{206}$$

Using the fact that $U_i = \mathcal{T}^*(F_i) + \epsilon_i$ we get by Jensen's and Young's inequality

$$\begin{aligned}&\Big(\delta\widehat{R}_\lambda(\overline{G}_{\infty,\lambda})[H]\Big)^2 \\ &\le 4\Big(\frac{1}{n}\sum_{i=1}^{n}\langle(\overline{G}_{\infty,\lambda} - G_\mathcal{G})(F_i), H(F_i)\rangle_{L^2(D_\mathcal{Y})} - \mathbb{E}\Big[\langle(\overline{G}_{\infty,\lambda} - G_\mathcal{G})(F), H(F)\rangle_{L^2(D_\mathcal{Y})}\Big]\Big)^2 \\ &\quad + 4\Big(\frac{1}{n}\sum_{i=1}^{n}\langle G_\mathcal{G}(F_i) - \mathcal{T}^*(F_i), H(F_i)\rangle_{L^2(D_\mathcal{Y})} - \mathbb{E}\Big[\langle G_\mathcal{G}(F) - \mathcal{T}^*(F), H(F)\rangle_{L^2(D_\mathcal{Y})}\Big]\Big)^2 \\ &\quad + 4\Big(\frac{1}{n}\sum_{i=1}^{n}\langle\varepsilon_i, H(F_i)\rangle_{L^2(D_\mathcal{Y})} - \mathbb{E}[\langle\varepsilon, H(F)\rangle_{L^2(D_\mathcal{Y})}]\Big)^2 \\ &= 4\Big(\frac{1}{n}\sum_{i=1}^{n}\langle(\overline{G}_{\infty,\lambda} - G_\mathcal{G})(F_i), H(F_i)\rangle_{L^2(D_\mathcal{Y})} - \mathbb{E}\Big[\langle(\overline{G}_{\infty,\lambda} - G_\mathcal{G})(F), H(F)\rangle_{L^2(D_\mathcal{Y})}\Big]\Big)^2 \\ &\quad + 4\Big(\frac{1}{n}\sum_{i=1}^{n}\langle G_\mathcal{G}(F_i) - \mathcal{T}^*(F_i), H(F_i)\rangle_{L^2(D_\mathcal{Y})}\Big)^2 + 4\Big(\frac{1}{n}\sum_{i=1}^{n}\langle\varepsilon_i, H(F_i)\rangle_{L^2(D_\mathcal{Y})}\Big)^2\end{aligned} \tag{207}$$

where the second line uses the fact that $\mathbb{E}[\langle\varepsilon, H(F)\rangle_{L^2(D_\mathcal{X})}] = 0$ by independence and $\mathbb{E}[\langle G_\mathcal{G}(F) - \mathcal{T}^*(F), H(F)\rangle_{L^2(D_\mathcal{X})}] = \mathbb{E}[\langle G_\mathcal{G}(F) - U, H(F)\rangle_{L^2(D_\mathcal{X})}] + \mathbb{E}[\langle\varepsilon, H(F)\rangle_{L^2(D_\mathcal{X})}] = 0$. Combining

Lemma 17 with Remark 18, the ψ_1-norm for the first term above is

$$\begin{aligned}\|\langle(\overline{G}_{\infty,\lambda}-G_{\mathcal{G}})(F),H(F)\rangle_{L^2(D_{\mathcal{Y}})}\|_{\psi_1} &\lesssim \|(\overline{G}_{\infty,\lambda}-G_{\mathcal{G}})(F)\|_{\psi_2,(\overline{G}_{\infty,\lambda}-G_{\mathcal{G}})\Gamma_F(\overline{G}_{\infty,\lambda}-G_{\mathcal{G}})^T}\|H(F)\|_{\psi_2,H\Gamma_F H^T}\\ &\quad\cdot \mathsf{Tr}((\overline{G}_{\infty,\lambda}-G_{\mathcal{G}})\Gamma_F(\overline{G}_{\infty,\lambda}-G_{\mathcal{G}})^T)^{\frac{1}{2}}\,\mathsf{Tr}(H\Gamma_F H^T)^{\frac{1}{2}}\\ &\le \|F\|^2_{\psi_2,\Gamma_F}\|\overline{G}_{\infty,\lambda}-G_{\mathcal{G}}\|_{\Gamma_F}\|H\|_{\Gamma_F}\lesssim\|\overline{G}_{\infty,\lambda}-G_{\mathcal{G}}\|_{\Gamma_F}\|H\|_{\Gamma_F}\end{aligned}\tag{208}$$

and similarly for the third term above

$$\begin{aligned}\|\langle\epsilon,H(F)\rangle_{L^2(D_{\mathcal{X}})}\|_{\psi_1}&\lesssim\|\varepsilon\|_{\psi_2,\Gamma_\varepsilon}\|H(F)\|_{\psi_2,H\Gamma_F H^T}\,\mathsf{Tr}(\Gamma_\varepsilon)^{\frac{1}{2}}\,\mathsf{Tr}(H\Gamma_F H^T)^{\frac{1}{2}}\\ &\le\|\varepsilon\|_{\psi_2,\Gamma_\varepsilon}\|F\|_{\psi_2,\Gamma_F}\,\mathsf{Tr}(\Gamma_\varepsilon)^{\frac{1}{2}}\|H\|_{\Gamma_F}\lesssim\|H\|_{\Gamma_F}.\end{aligned}\tag{209}$$

For the second term above we apply Remark 18. Using Assumption 2 we can let $M'=\|G_{\mathcal{G}}\|_{\text{op}}+M$ such that,

$$\begin{aligned}\|\langle G_{\mathcal{G}}(F)-\mathcal{T}^*(F),H(F)\rangle_{L^2(D_{\mathcal{Y}})}\|_{\psi_1}&\le\Big\|\|G_{\mathcal{G}}(F)-\mathcal{T}^*(F)\|^2_{L^2(D_{\mathcal{Y}})}\Big\|^{\frac{1}{2}}_{\psi_1}\Big\|\|H(F)\|^2_{L^2(D_{\mathcal{Y}})}\Big\|^{\frac{1}{2}}_{\psi_1}\\ &\le\Big\|M'\|F\|_{L^2(D_{\mathcal{X}})}+c\Big\|_{\psi_2}\Big\|\|H(F)\|^2_{L^2(D_{\mathcal{Y}})}\Big\|^{\frac{1}{2}}_{\psi_1}\\ &\le\Big(M'\Big\|\|F\|^2_{L^2(D_{\mathcal{X}})}\Big\|^{\frac{1}{2}}_{\psi_1}+c'\Big)\Big\|\|H(F)\|^2_{L^2(D_{\mathcal{Y}})}\Big\|^{\frac{1}{2}}_{\psi_1}\end{aligned}\tag{210}$$

and by an application of Chen and Yang (2021, Appendix A.2, Lemma A.4) we get that

$$\begin{aligned}\|\langle G_{\mathcal{G}}(F)-\mathcal{T}^*(F),H(F)\rangle_{L^2(D_{\mathcal{X}})}\|_{\psi_1}&\le\Big(M'\|F\|_{\psi_2,\Gamma_F}\,\mathsf{Tr}(\Gamma_F)^{\frac{1}{2}}+c'\Big)\|H(F)\|_{\psi_2,H\Gamma_F H^T}\,\mathsf{Tr}(H\Gamma_F H^T)^{\frac{1}{2}}\\ &\le\Big(M'\|F\|_{\psi_2,\Gamma_F}\,\mathsf{Tr}(\Gamma_F)^{\frac{1}{2}}+c'\Big)\|F\|_{\psi_2,\Gamma_F}\|H\|_{\Gamma_F}\\ &\lesssim\max(1,\|G_{\mathcal{G}}\|_{\text{op}}+M)\|H\|_{\Gamma_F}\\ &\lesssim\max(1,\|G_{\mathcal{G}}\|_{\text{op}})\|H\|_{\Gamma_F}.\end{aligned}\tag{211}$$

Hence by Bernstein's inequality,

$$\begin{aligned}&\Big|\frac{1}{n}\sum_{i=1}^n\langle(\overline{G}_{\infty,\lambda}-G_{\mathcal{G}})(F_i),H(F_i)\rangle_{L^2(D_{\mathcal{Y}})}-\mathbb{E}\Big[\langle(\overline{G}_{\infty,\lambda}-G_{\mathcal{G}})(F),H(F)\rangle_{L^2(D_{\mathcal{Y}})}\Big]\Big|\\ &\qquad\qquad\lesssim\|\overline{G}_{\infty,\lambda}-G_{\mathcal{G}}\|_{\Gamma_F}\|H\|_{\Gamma_F}\Big(\sqrt{\frac{\log(1/\delta)}{n}}\vee\frac{\log(1/\delta)}{n}\Big)\\ &\Big|\frac{1}{n}\sum_{i=1}^n\langle G_{\mathcal{G}}(F_i)-\mathcal{T}^*(F_i),H(F_i)\rangle_{L^2(D_{\mathcal{Y}})}\Big|\lesssim\max(1,\|G_{\mathcal{G}}\|_{\text{op}})\|H\|_{\Gamma_F}\Big(\sqrt{\frac{\log(1/\delta)}{n}}\vee\frac{\log(1/\delta)}{n}\Big)\\ &\Big|\frac{1}{n}\sum_{i=1}^n\langle\varepsilon_i,H(F_i)\rangle_{L^2(D_{\mathcal{Y}})}\Big|\lesssim\|H\|_{\Gamma_F}\Big(\sqrt{\frac{\log(1/\delta)}{n}}\vee\frac{\log(1/\delta)}{n}\Big)\end{aligned}\tag{212}$$

with probability at least $1 - \delta$. Note that we can write the first bound above more generally for all $G, H \in \mathcal{G}$ as

$$\begin{aligned}
\Big| \langle ((\widehat{\Sigma}_F - \Sigma_F) \otimes I) G, H \rangle_{L^2(D_{\mathcal{X}} \times D_{\mathcal{Y}})} \Big| &= \Big| \frac{1}{n} \sum_{i=1}^{n} \langle G(F_i), H(F_i) \rangle_{L^2(D_{\mathcal{Y}})} - \mathbb{E}\Big[\langle G(F), H(F) \rangle_{L^2(D_{\mathcal{Y}})} \Big] \Big| \\
&\lesssim \|G\|_{\Gamma_F} \|H\|_{\Gamma_F} \Big(\sqrt{\frac{\log(1/\delta)}{n}} \vee \frac{\log(1/\delta)}{n} \Big) \\
&\lesssim \|G\|_{\Sigma_F} \|H\|_{\Sigma_F} \Big(\sqrt{\frac{\log(1/\delta)}{n}} \vee \frac{\log(1/\delta)}{n} \Big)
\end{aligned} \tag{213}$$

where the last line follows since $\|G\|_{\Gamma_F} \lesssim \|G\|_{\Sigma_F}$ by Assumption 1.

Finally, since $\|\overline{G}_{\infty,\lambda} - G_{\mathcal{G}}\|_{\Sigma_F}^2 \leq \lambda J(G_{\mathcal{G}})$ by the determinstic error derivations we conclude from (212) that uniformly over all $H \in \mathcal{G}$,

$$\begin{aligned}
\Big(\delta \widehat{R}_\lambda(\overline{G}_{\infty,\lambda})[H] \Big)^2 &\leq \max\Big(1, \|G_{\mathcal{G}}\|_{\text{op}}, \lambda J(G_{\mathcal{G}})\Big) \frac{\log(1/\delta)}{n} \|H\|_{\Gamma_F}^2 \\
&\leq \max\Big(1, \|G_{\mathcal{G}}\|_{\text{op}}, \lambda J(G_{\mathcal{G}})\Big) \frac{\log(1/\delta)}{n} \|H\|_{\Sigma_F}^2
\end{aligned} \tag{214}$$

with probability $1 - \delta$ where the last inequality again follows from Assumption 1. For shorthand, let us define $\kappa(G_{\mathcal{G}}) = \max\Big(1, \|G_{\mathcal{G}}\|_{\text{op}}, \lambda J(G_{\mathcal{G}})\Big)$. Now we can write

$$\begin{aligned}
\|\tilde{G} - \overline{G}_{\infty,\lambda}\|_{\Sigma_F}^2 &= \Big\| \nabla^2 R_\lambda^{-1} \Big(\nabla \widehat{R}_\lambda(\overline{G}_{\infty,\lambda}) \Big) \Big\|_{\Sigma_F}^2 = \frac{1}{4} \Big\| \sum_{k=1}^{\infty} \nu_k^{-1} (1 + \lambda \gamma_k^{-1})^{-1} \Big(\nu_k \delta \widehat{R}_\lambda(\overline{G}_{\infty,\lambda})[\Omega_k] \Omega_k \Big) \Big\|_{\Sigma_F}^2 \\
&= \frac{1}{4} \sum_{k=1}^{\infty} (1 + \lambda \gamma_k^{-1})^{-2} \Big(\delta \widehat{R}_\lambda(\overline{G}_{\infty,\lambda})[\Omega_k] \Big)^2 \\
&\lesssim \kappa(G_{\mathcal{G}}) \frac{\log(1/\delta)}{n} \sum_{k=1}^{\infty} (1 + \lambda \gamma_k^{-1})^{-2} \|\Omega_k\|_{\Sigma_F}^2 \\
&= \kappa(G_{\mathcal{G}}) \frac{\log(1/\delta)}{n} \sum_{k=1}^{\infty} (1 + \lambda \gamma_k^{-1})^{-2} \\
&\lesssim \kappa(G_{\mathcal{G}}) \frac{\log(1/\delta)}{n} \sum_{k=1}^{\infty} (1 + \lambda k^r)^{-2} \\
&\asymp \kappa(G_{\mathcal{G}}) \frac{\log(1/\delta)}{n} \int_1^\infty (1 + \lambda x^r)^{-2} dx \\
&= \kappa(G_{\mathcal{G}}) \frac{\log(1/\delta)}{n} \lambda^{-\frac{1}{r}} \int_{\lambda^{\frac{1}{r}}}^\infty (1 + x^r)^{-2} dx
\end{aligned} \tag{215}$$

Noting that $\int_{\lambda^{\frac{1}{r}}}^\infty (1 + x^r)^{-2} dx \leq \int_{\lambda^{\frac{1}{r}}}^1 (1 + x^r)^{-2} dx + \int_1^\infty x^{-2r} dx \leq 1 + \frac{1}{2r-1} - \lambda^{\frac{1}{r}}$ for all $r > \frac{1}{2}$ we have that

$$\|\tilde{G} - \overline{G}_{\infty,\lambda}\|_{\Sigma_F}^2 \lesssim \kappa(G_{\mathcal{G}}) \frac{\log(1/\delta)}{n} \lambda^{-\frac{1}{r}}. \tag{216}$$

2. Bounding $\|\widehat{G}_{n,\lambda} - \tilde{G}\|_{\Sigma_F}$

Now we move on to bounding $\|\widehat{G}_{n,\lambda} - \tilde{G}\|_{\Sigma_F}$. First clearly $\nabla \widehat{R}_\lambda(\widehat{G}_{n,\lambda}) = 0$ by first-order optimality. Since $\widehat{R}_\lambda(G)$ is quadratic then we can in fact write a Taylor series expansion

$$\nabla \widehat{R}_\lambda(\widehat{G}_{n,\lambda}) = \nabla \widehat{R}_\lambda(\overline{G}_{\infty,\lambda}) + \nabla^2 \widehat{R}_\lambda(\widehat{G}_{n,\lambda} - \overline{G}_{\infty,\lambda}) = 0. \tag{217}$$

Also, since $\tilde{G} = \overline{G}_{\infty,\lambda} - \nabla^2 R_\lambda^{-1}\Big(\nabla \widehat{R}_\lambda(\overline{G}_{\infty,\lambda})\Big)$ then

$$\nabla^2 R_\lambda(\tilde{G} - \overline{G}_{\infty,\lambda}) = -\nabla \widehat{R}_\lambda(\overline{G}_{\infty,\lambda}). \tag{218}$$

To conclude, we know that

$$\nabla \widehat{R}_\lambda(\overline{G}_{\infty,\lambda}) = \nabla^2 R_\lambda(\overline{G}_{\infty,\lambda} - \tilde{G}) = \nabla^2 \widehat{R}_\lambda(\overline{G}_{\infty,\lambda} - \widehat{G}_{n,\lambda}). \tag{219}$$

Hence we can write

$$\begin{aligned}\nabla^2 R_\lambda(\widehat{G}_{n,\lambda} - \tilde{G}) &= \nabla^2 R_\lambda(\widehat{G}_{n,\lambda} - \overline{G}_{\infty,\lambda}) + \nabla^2 R_\lambda(\overline{G}_{\infty,\lambda} - \tilde{G}) \\ &= \nabla^2 R_\lambda(\widehat{G}_{n,\lambda} - \overline{G}_{\infty,\lambda}) - \nabla^2 \widehat{R}_\lambda(\widehat{G}_{n,\lambda} - \overline{G}_{\infty,\lambda}) \\ &= \nabla^2 R(\widehat{G}_{n,\lambda} - \overline{G}_{\infty,\lambda}) - \nabla^2 \widehat{R}(\widehat{G}_{n,\lambda} - \overline{G}_{\infty,\lambda})\end{aligned} \tag{220}$$

which proves that

$$\widehat{G}_{n,\lambda} - \tilde{G} = \nabla^2 R_\lambda^{-1}\Big(\nabla^2 R(\widehat{G}_{n,\lambda} - \overline{G}_{\infty,\lambda}) - \nabla^2 \widehat{R}(\widehat{G}_{n,\lambda} - \overline{G}_{\infty,\lambda})\Big). \tag{221}$$

Now denoting $\widehat{G}_{n,\lambda} = \sum_{k=1}^\infty \widehat{b}_k \Omega_k$ and $\overline{G}_{\infty,\lambda} = \sum_{k=1}^\infty \overline{b}_k \Omega_k$ we get by Cauchy–Schwarz and (213) that

$$\begin{aligned}&\|\widehat{G}_{n,\lambda} - \tilde{G}\|_{\Sigma_F}^2 \\ &\quad = \frac{1}{4}\Big\| \sum_{k=1}^\infty \nu_k^{-1}(1+\lambda\gamma_k^{-1})^{-1}\Big(\nu_k \delta \widehat{R}^2(\widehat{G}_{n,\lambda} - \overline{G}_{\infty,\lambda})[\Omega_k]\Omega_k - \nu_k \delta R^2(\widehat{G}_{n,\lambda} - \overline{G}_{\infty,\lambda})[\Omega_k]\Omega_k\Big)\Big\|_{\Sigma_F}^2 \\ &\quad = \frac{1}{4}\sum_{k=1}^\infty (1+\lambda\gamma_k^{-1})^{-2}\Big(\sum_{j=1}^\infty (\widehat{b}_j - \overline{b}_j)\langle((\widehat{\Sigma}_F - \Sigma_F)\otimes I)\Omega_j, \Omega_k\rangle_{L^2(D_\mathcal{X}\times D_\mathcal{Y})}\Big)^2 \\ &\quad \lesssim \frac{\log(1/\delta)}{n}\sum_{k=1}^\infty (1+\lambda\gamma_k^{-1})^{-2}\sum_{j=1}^\infty (\widehat{b}_j - \overline{b}_j)^2 \|\Omega_j\|_{\Sigma_F}^2 \|\Omega_k\|_{\Sigma_F}^2 \\ &\quad \leq \frac{\log(1/\delta)}{n}\sum_{k=1}^\infty (1+\lambda\gamma_k^{-1})^{-2}\sum_{j=1}^\infty (\widehat{b}_j - \overline{b}_j)^2 \\ &\quad \leq \frac{\log(1/\delta)}{n}\lambda^{-\frac{1}{r}}\|\widehat{G}_{n,\lambda} - \overline{G}_{\infty,\lambda}\|_{\Sigma_F}^2\end{aligned} \tag{222}$$

for all $r > \frac{1}{2}$ with probability at least $1-\delta$.

3. Combining both bounds

By the triangle inequality we know that

$$\|\tilde{G}-\overline{G}_{\infty,\lambda}\|_{\Sigma_F} \geq \|\widehat{G}_{n,\lambda}-\overline{G}_{\infty,\lambda}\|_{\Sigma_F}-\|\widehat{G}_{n,\lambda}-\tilde{G}\|_{\Sigma_F} \geq \Big(1-C\lambda^{-\frac{1}{2r}}\sqrt{\frac{\log(1/\delta)}{n}}\Big)\|\widehat{G}_{n,\lambda}-\overline{G}_{\infty,\lambda}\|_{\Sigma_F} \tag{223}$$

for some absolute constant $C > 0$. Hence by (216), if $\frac{\log(1/\delta)}{n}\lambda^{-\frac{1}{r}} \lesssim 1$ then for sufficiently large n,

$$\|\widehat{G}_{n,\lambda} - \overline{G}_{\infty,\lambda}\|_{\Sigma_F} \leq \Big(1 - C\lambda^{-\frac{1}{2r}}\sqrt{\frac{\log(1/\delta)}{n}}\Big)^{-1}\|\tilde{G} - \overline{G}_{\infty,\lambda}\|_{\Sigma_F} \lesssim \sqrt{\kappa(G_{\mathcal{G}})}\lambda^{-\frac{1}{2r}}\sqrt{\frac{\log(1/\delta)}{n}}. \tag{224}$$

where $\kappa(G_{\mathcal{G}}) = \max\Big(1, \|G_{\mathcal{G}}\|_{\text{op}}, \lambda J(G_{\mathcal{G}})\Big)$. So finally, squaring both sides proves that

$$\|\widehat{G}_{n,\lambda} - \overline{G}_{\infty,\lambda}\|_{\Sigma_F}^2 \lesssim \max\Big(1, \|G_{\mathcal{G}}\|_{\text{op}}, \lambda J(G_{\mathcal{G}})\Big)\frac{\log(1/\delta)}{n}\lambda^{-\frac{1}{r}} \tag{225}$$

with probability at least $1 - \delta$. ■

Appendix G. Enforcing Symmetries & Invariances in RKHSs

Here we described how to transform an RKHS so that functions in this Hilbert space satisfy constraints such as coordinate symmetries, time causality, and time invariance.

G.1 Coordinate Symmetries

Suppose we have an RKHS of functions $\mathcal{H} \in L^2(D \times D)$ and we would like to transform this space such that every function $f \in \mathcal{H}$ is symmetric in its coordinates

$$f(x, y) = f(y, x), \quad \forall x, y \in D. \tag{226}$$

We assume here that our RKHS $\mathcal{H}$ has a continuous, square-integrable, and positive semidefinite kernel $K(D^2 \times D^2) \to \mathbb{R}$ that satisfies the symmetry property

$$K(x, y, \xi, \eta) = K(y, x, \eta, \xi), \quad \forall x, y, \xi, \eta \in D. \tag{227}$$

As shorthand, for any $f \in L^2(D \times D)$ we define $f^T \in L^2(D \times D)$ given by $f^T(x, y) = f(y, x)$. Now we can state the following theorem.

Theorem 19 *If the kernel $K : D^4 \to \mathbb{R}$ of $\mathcal{H}$ satisfies the symmetry property $K(x, y, \xi, \eta) = K(y, x, \eta, \xi)$ then we know that $f^T \in \mathcal{H}$ for any $f \in \mathcal{H}$. Furthermore, we can define the symmetrized RKHS*

$$\mathcal{S} := \Big\{ \frac{f + f^T}{2} : f \in \mathcal{H} \Big\} = \Big\{ f = f^T : f \in \mathcal{H} \Big\} \tag{228}$$

with inner product inherited from $\mathcal{H}$ whose reproducing kernel takes the form

$$K_{symm}(x, y, \xi, \eta) = \frac{1}{4}\Big[K(x, y, \xi, \eta) + K(x, y, \eta, \xi) + K(y, x, \xi, \eta) + K(y, x, \eta, \xi)\Big]. \tag{229}$$

We give the proof of this result below.

Proof We assume that K is continuous, square integrable, and positive semidefinite so it is a Mercer kernel with the decomposition

$$K(x, y, \xi, \eta) = \sum_{k=1}^{\infty} \lambda_k \psi_k(x, y) \psi_k(\xi, \eta) \tag{230}$$

where λ_k are the eigenvalues and $\psi_k \in L^2(D_{\mathcal{X}} \times D_{\mathcal{Y}})$ are the L^2 orthonormal eigenfunctions.

First we prove that if $f \in \mathcal{H}$ then $f^T \in \mathcal{H}$. A consequence of Mercer's theorem is that $\mathcal{H}$ can be characterized as

$$\mathcal{H} = \Big\{ f \in L^2(D \times D) \Big| \sum_{k=1}^{\infty} \frac{\langle f, \psi_k \rangle^2_{L^2(D \times D)}}{\lambda_k} \Big\}. \tag{231}$$

Since K satisfies the symmetry $K(x, y, \xi, \eta) = K(y, x, \eta, \xi)$, it is immediate that

$$K(x, y, \xi, \eta) = \sum_{k=1}^{\infty} \lambda_k \psi_k(y, x) \psi_k(\eta, \xi) \tag{232}$$

so we can also write

$$\mathcal{H} = \Big\{ f \in L^2(D \times D) \Big| \sum_{k=1}^{\infty} \frac{\langle f, \psi_k^T \rangle^2_{L^2(D\times D)}}{\lambda_k} \Big\}. \tag{233}$$

For any $f \in \mathcal{H}$ we know that

$$\sum_{k=1}^{\infty} \frac{\langle f, \psi_k \rangle^2_{L^2(D\times D)}}{\lambda_k} < \infty \tag{234}$$

which implies that

$$\sum_{k=1}^{\infty} \frac{\langle f^T, \psi_k^T \rangle^2_{L^2(D\times D)}}{\lambda_k} < \infty \tag{235}$$

proving that $f^T \in \mathcal{H}$.

Now define the symmetrized set of functions

$$\mathcal{S} := \Big\{ \frac{f + f^T}{2} : f \in \mathcal{H} \Big\} \subseteq \mathcal{H}. \tag{236}$$

Let's prove that $\mathcal{S}$ as a set of functions is equal to

$$\mathcal{S}' = \Big\{ f = f^T : f \in \mathcal{H} \Big\}. \tag{237}$$

This follows immediately since for any $f \in S$ we can check that $f \in \mathcal{H}$ and $f = f^T$ so $f \in \mathcal{S}'$. In the other direction, for any $f \in \mathcal{S}'$ by definition $f \in \mathcal{H}$ and $f = f^T$ so $\frac{f+f^T}{2} = f$ implying that $f \in \mathcal{S}$. Since $\mathcal{S} = \mathcal{S}'$ is a closed subset of $\mathcal{H}$ we can naturally equip it with the inner product from $\mathcal{H}$, proving that it is a Hilbert space.

Finally, we need to show that $\mathcal{S}$ is an RKHS with kernel

$$K_{\text{symm}}(x, y, \xi, \eta) = \frac{1}{4}\Big[K(x, y, \xi, \eta) + K(x, y, \eta, \xi) + K(y, x, \xi, \eta) + K(y, x, \eta, \xi) \Big]. \tag{238}$$

Remember that we can interpret the reproducing kernels K and K_{symm} as maps $K : \mathcal{H} \to \mathcal{H}$ and $K_{\text{symm}} : \mathcal{S} \to \mathcal{S}$ by

$$\begin{aligned} K(f)(x, y) &= \int_D \int_D K(x, y, \xi, \eta) f(\xi, \eta) \mathrm{d}\xi \mathrm{d}\eta \\ K_{\text{symm}}(f)(x, y) &= \int_D \int_D K_{\text{symm}}(x, y, \xi, \eta) f(\xi, \eta) \mathrm{d}\xi \mathrm{d}\eta. \end{aligned} \tag{239}$$

First note that K_{symm} is continuous, square integrable, symmetric, and positive semidefinite. This last condition can be checked by noting that $K_{\text{symm}}(f) = K(f)$ for all $f \in \mathcal{S}$ which implies that K_{symm} inherits the positive semidefiniteness of $\mathcal{H}$. Furthermore, $K_{\text{symm}(x_0,y_0)} \in \mathcal{S}$ for all $x_0, y_0 \in D$ since $K_{(x_0,y_0)}, K_{(y_0,x_0)} \in \mathcal{H}$. Here the notation $K_{(x_0,y_0)}$ means we are centering the kernel at a point $(x_0, y_0) \in D \times D$ to get a function $K(x, y, x_0, y_0)$.

We conclude by showing that K_{symm} satisfies the reproducing property on $\mathcal{S}$. First note that the symmetry property $K(x, y, \xi, \eta) = K(y, x, \eta, \xi)$ gives us that $K(\mathcal{S}) \subseteq \mathcal{S}$ and $K(\mathcal{A}) \subseteq K(\mathcal{A})$ for the space of antisymmetric functions

$$\mathcal{A} := \Big\{ \frac{f - f^T}{2} : f \in \mathcal{H} \Big\} = \Big\{ f = -f^T : f \in \mathcal{H} \Big\}. \tag{240}$$

This proves that $K(\mathcal{S}) = \mathcal{S}$ and $K(\mathcal{A}) = \mathcal{A}$. Hence, for any $f \in \mathcal{S}$ we know that $K^{-1}(f) \in \mathcal{S}$ so

$$\begin{aligned}\langle K_{\text{symm}(x,y)}, f\rangle_{\mathcal{S}} &:= \langle K_{\text{symm}(x,y)}, f\rangle_{\mathcal{H}} = \langle K_{\text{symm}(x,y)}, K^{-1}(f)\rangle_{L^2(D\times D)} \\ &= \langle K_{\text{symm}(x,y)}, K^{-1}(f)\rangle_{L^2(D\times D)} = K_{\text{symm}}(K^{-1}(f))(x,y) \\ &= K(K^{-1}(f))(x,y) = f(x,y)\end{aligned} \tag{241}$$

where the equality from the second to the third line follows since $K_{\text{symm}} = K$ on $\mathcal{S}$. This proves that $K_{\text{symm}} : (D \times D)^2 \to \mathbb{R}$ is the reproducing kernel for $\mathcal{S}$ with the inner product inherited from $\mathcal{H}$. ■

Theorem 19 can naturally be extended to coordinate symmetries (permutations) of n variables by iterating the statement over pairs of coordinates at a time.

G.2 Flip ($\mathbb{Z}_2$) Symmetries

Another important class of symmetries are flips along a coordinate axis. Suppose we have an RKHS of functions $\mathcal{H} \in L^2([-a,a])$ and we would like to transform it so that every $f \in \mathcal{H}$ satisfies

$$f(x) = f(-x), \quad \forall x \in [-a,a]. \tag{242}$$

where $a > 0$ can be finite or infinite.

Let's assume that our initial RKHS $\mathcal{H}$ has a continuous, square-integrable, and positive semidefinite kernel $K([-a,a]^2) \to \mathbb{R}$ that satisfies the symmetry property

$$K(x,\xi) = K(-x,-\xi), \quad \forall x,\xi \in [-a,a]. \tag{243}$$

Through a nearly identical argument to the case of coordinate symmetries, we can show the following result. As shorthand, for any $f \in L^2([-a,a])$ we define $f^- \in L^2(-[a,a])$ given by $f^-(x) = f(-x)$.

Theorem 20 *If the kernel $K : [-a,a]^2 \to \mathbb{R}$ of $\mathcal{H}$ satisfies the symmetry property $K(x,\xi) = K(-x,-\xi)$ then we know that $f^- \in \mathcal{H}$ for any $f \in \mathcal{H}$. Furthermore, we can define the symmetrized RKHS*

$$\mathcal{F} := \left\{\frac{f + f^-}{2} : f \in \mathcal{H}\right\} = \left\{f = f^- : f \in \mathcal{H}\right\} \tag{244}$$

with inner product inherited from $\mathcal{H}$ whose reproducing kernel takes the form

$$K_{\textit{flip}}(x,\xi) = \frac{1}{4}\Big[K(x,\xi) + K(x,-\xi) + K(-x,\xi) + K(-x,-\xi)\Big]. \tag{245}$$

G.3 Enforcing Time Causality

Here we discuss how to transform an RKHS so that functions in this space satisfy a time constraint known as causality. We begin with an RKHS of functions $\mathcal{H} \in L^2([a,b] \times [a,b])$ where $[a,b]$ denotes an interval of time and may generally have open, closed, finite, or infinite endpoints. We aim to transform our RKHS so that for every $f \in \mathcal{H}$ we have that

$$f(s,t) = \mathbf{1}_{t \geq s} f(s,t). \tag{246}$$

The new function $\mathbf{1}_{t\geq s}f(s,t)$ is time causal because taking an input signal $p(s)$ and integrating it against the s-coordinate

$$q(t) = \int_a^b \mathbf{1}_{t\geq s}f(s,t)p(s)\mathrm{d}s \tag{247}$$

generates an output signal $q(t)$ where the influence of the output at time t only depends on the perturbation $p(s)$ at previous times $s \leq t$. Hence, the causal order of time is respected by the linear filter $\mathbf{1}_{t\geq s}f(s,t)$.

Multiplication of functions in an RKHS by a fixed positive weighting function is a standard procedure described in Saitoh and Sawano (2016, Section 2.3.4, Corollary 2.5). The difference here is that $\mathbf{1}_{t\geq s}$ is not strictly positive at all points (s,t). Hence, applying this weighting to all functions in $\mathcal{H}$ does not automatically produce a new RKHS that inherits the original inner product of $\mathcal{H}$.

We proceed to construct this RKHS in the following way. Using Theorem 19 we symmetrize $\mathcal{H}$ in the time coordinate to obtain the RKHS $\mathcal{S}$ whose elements satisfy $f(s,t) = f(t,s)$. In order to do this, we implicitly assume that our kernel $K : [a,b]^4 \to \mathbb{R}$ for the RKHS $\mathcal{H}$ satisfies the property $K(s,t,\sigma,\tau) = K(t,s,\tau,\sigma)$. Then we know that $\mathcal{S} \subseteq \mathcal{H}$ is an RKHS with inner product inherited from $\mathcal{H}$ and reproducing kernel

$$K_{\text{symm}}(t,s,\tau,\sigma) = \frac{1}{4}\Big[K(s,t,\sigma,\tau) + K(s,t,\tau,\sigma) + K(t,s,\sigma,\tau) + K(t,s,\tau,\sigma)\Big]. \tag{248}$$

Now define the causal RKHS

$$\mathcal{H}_{\text{causal}} := \Big\{\mathbf{1}_{\{t\geq s\}}f : f \in \mathcal{S}\Big\} \tag{249}$$

with the inherited inner product

$$\langle \overline{f}, \overline{g}\rangle_{\mathcal{H}_{\text{causal}}} := \Big\langle f, g\Big\rangle_{\mathcal{S}}. \tag{250}$$

for all $\overline{f} = \mathbf{1}_{\{t\geq s\}}f$ and $\overline{g} = \mathbf{1}_{\{t\geq s\}}g \in \mathcal{H}_{\text{causal}}$ where $f, g \in \mathcal{S}$. Equipping $\mathcal{H}_{\text{causal}}$ with this inner product implies that it is isometrically isomorphic to the time-symmetrized RKHS $\mathcal{S}$.

It is not hard to check that $\mathcal{H}_{\text{causal}}$ has the reproducing kernel

$$K_{\text{causal}} = \mathbf{1}_{\{t\geq s\}}\mathbf{1}_{\{\tau\geq\sigma\}}K_{\text{symm}} \tag{251}$$

since it is symmetric, positive semidefinite, and it satisfies the reproducing property

$$\begin{aligned}\langle K_{\text{causal}(s,t)}, \overline{f}\rangle_{\mathcal{H}_{\text{causal}}} &= \mathbf{1}_{\{t\geq s\}}\langle K_{\text{symm}_{(s,t)}}, f\rangle_{\mathcal{S}} \\ &= \mathbf{1}_{\{t\geq s\}}f(s,t) = \overline{f}(s,t)\end{aligned} \tag{252}$$

for any $\overline{f} \in \mathcal{H}_{\text{causal}}$ where $\overline{f} = \mathbf{1}_{\{t\geq s\}}f$ for some $f \in \mathcal{S}$. Here again the notation $K_{\text{symm}_{(s,t)}}$ denotes the kernel centered at a given point $K_{\text{symm}}(\cdot,\cdot,s,t)$.

In summary, to create an RKHS of causal functions from some initial RKHS $\mathcal{H}$, we first symmetrize it in time to obtain the RKHS $\mathcal{S} \subseteq \mathcal{H}$ and then for each $f \in \mathcal{S}$ we set $f(s,t) = 0$ for all $s > t$ to obtain the RKHS $\mathcal{H}_{\text{causal}}$ along with the formula for its reproducing kernel K_{causal}.

In exactly the same way, we can define the space of *anticausal functions*

$$\mathcal{H}_{\text{anticausal}} := \Big\{\mathbf{1}_{\{t\geq s\}}f : f \in \mathcal{S}\Big\} \tag{253}$$

where functions in this space satisfy

$$f(s,t) = 0 \quad \text{for} \quad t > s. \tag{254}$$

Such functions are less common but occur for example in systems with prescribed terminal conditions at time $t = T$.

G.4 Enforcing Time Invariance

Another important constraint we may want functions in our RKHS to satisfy is invariance to time. A function $f(s,t)$ for $s,t \in [a,b]$ is time-invariant if it can be rewritten as the difference of its two coordinates

$$f(s,t) = f(t-s). \tag{255}$$

To build such a space of functions, we start with an RKHS of one-dimensional functions $\mathfrak{h} \subset L^2([-(b-a),(b-a)])$ with kernel $k : [-(b-a),(b-a)]^2 \to \mathbb{R}$. We then lift this RKHS to the space of two-dimensional functions

$$\mathcal{C}(\mathfrak{h}) := \Big\{ f : [a,b] \times [a,b] \to \mathbb{R},\ f(s,t) = h(t-s),\ \forall h \in \mathfrak{h} \Big\} \tag{256}$$

equipped with the inherited inner product for all $f_1, f_2 \in \mathcal{C}(\mathfrak{h})$,

$$\langle f_1, f_2 \rangle_{\mathcal{C}(\mathfrak{h})} = \langle h_1, h_2 \rangle_{\mathfrak{h}} \tag{257}$$

where $f_1(s,t) = h_1(t-s)$ and $f_2(s,t) = h_2(t-s)$ with $h_1, h_2 \in \mathfrak{h}$. It is easy to see that $\mathcal{C}(\mathfrak{h}) \subset L^2([a,b]^2)$ is isometrically isomorphic to $\mathfrak{h}$ and by properties of the reproducing kernel, we can check that $K : [a,b]^4 \to \mathbb{R}$ where

$$K(s,t,\sigma,\tau) = k(t-s, \tau-\sigma) \tag{258}$$

is the reproducing kernel of $\mathcal{C}(\mathfrak{h})$.

Computationally, time-invariant constraints are useful as they reduce the amount of information stored in a function (i.e. reduce it from a function of two variables s,t to a function of one variable $t-s$). As discussed in Section 3 our standard way of constructing a function $f(s,t) \in \mathcal{C}(\mathfrak{h})$ is to take equally-spaced grid points $s_j = t_j = \frac{b-a}{m-1}(j-1) + a$ for $j = 1, \ldots, m$ and a matrix of weights $W \in \mathbb{R}^{m \times m}$ where

$$f(s,t) = \sum_{i=1}^{m} \sum_{j=1}^{m} K(s,t,s_i,t_j) W_{ij}. \tag{259}$$

Using the convolutional form of K we can further write

$$\begin{aligned} f(s,t) = \sum_{i=1}^{m} \sum_{j=1}^{m} k(t-s, t_j - s_i) W_{ij} &= \sum_{i=1}^{m} \sum_{j=1}^{m} k\Big(t-s, \frac{b-a}{m-1}(j-i)\Big) W_{ij} \\ &= \sum_{i=-(m-1)}^{m-1} k\Big(t-s, \frac{b-a}{m-1} i\Big) w_i. \end{aligned} \tag{260}$$

where $w \in \mathbb{R}^{2m-1}$ is a new set of weights. Thus, making the physically relevant assumption that f lies in an RKHS of convolutional operators can significantly decrease computation time and memory if optimized as in the equation above.

References

Jonas Adler and Ozan Öktem. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems*, 33(12):124007, 2017.

Mohammed Al-Smadi and Omar Abu Arqub. Computational algorithm for solving Fredholm time-fractional partial integrodifferential equations of Dirichlet functions type with error estimates. *Applied Mathematics and Computation*, 342:280–294, 2019.

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

Omar Abu Arqub. Numerical solutions for the Robin time-fractional partial differential equations of heat and fluid flows based on the reproducing kernel algorithm. *International Journal of Numerical Methods for Heat & Fluid Flow*, 2018.

Omar Abu Arqub. Numerical simulation of time-fractional partial differential equations arising in fluid flows via reproducing kernel method. *International Journal of Numerical Methods for Heat & Fluid Flow*, 2019.

Omar Abu Arqub and Mohammed Al-Smadi. An adaptive numerical approach for the solutions of fractional advection–diffusion and dispersion equations in singular case under Riesz's derivative operator. *Physica A: Statistical Mechanics and its Applications*, 540:123257, 2020.

Kaijun Bao, Xu Qian, Ziyuan Liu, and Songhe Song. An operator learning approach via function-valued reproducing kernel Hilbert space for differential equations. *arXiv preprint arXiv:2202.09488*, 2022.

Leah Bar and Nir Sochen. Unsupervised deep learning algorithm for PDE-based forward and inverse problems. *arXiv preprint arXiv:1904.05417*, 2019.

Mario Bebendorf and Wolfgang Hackbusch. Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L∞-coefficients. *Numerische Mathematik*, 95(1):1–28, 2003.

Rosalie Bélanger-Rioux and Laurent Demanet. Compressed absorbing boundary conditions via matrix probing. *SIAM Journal on Numerical Analysis*, 53(5):2441–2471, 2015.

Mikhail Belkin. Approximation beats concentration? An approximation view on inference with smooth radial kernels. In *Conference On Learning Theory*, pages 1348–1361. PMLR, 2018.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Saakaar Bhatnagar, Yaser Afshar, Shaowu Pan, Karthik Duraisamy, and Shailendra Kaushik. Prediction of aerodynamic flow fields using convolutional neural networks. *Computational Mechanics*, 64(2):525–545, 2019.

Kaushik Bhattacharya, Bamdad Hosseini, Nikola B. Kovachki, and Andrew M. Stuart. Model reduction and neural networks for parametric PDEs. *The SMAI Journal of computational mathematics*, 7: 121–157, 2021. doi: 10.5802/smai-jcm.74. URL `https://smai-jcm.centre-mersenne.org/articles/10.5802/smai-jcm.74/`.

Nicolas Boullé and Alex Townsend. Learning elliptic partial differential equations with randomized linear algebra. *Foundations of Computational Mathematics*, pages 1–31, 2022.

Nicolas Boullé, Christopher J Earls, and Alex Townsend. Data-driven discovery of Green's functions with human-understandable deep learning. *Scientific reports*, 12(1):1–9, 2022a.

Nicolas Boullé, Seick Kim, Tianyi Shi, and Alex Townsend. Learning Green's functions associated with time-dependent partial differential equations. *Journal of Machine Learning Research*, 23 (218):1–34, 2022b.

Benjamin Charlier, Jean Feydy, Joan Alexis Glaunès, François-David Collin, and Ghislain Durif. Kernel operations on the GPU, with autodiff, without memory overflows. *Journal of Machine Learning Research*, 22(74):1–6, 2021. URL `http://jmlr.org/papers/v22/20-275.html`.

Jiong Chen, Florian Schäfer, Jin Huang, and Mathieu Desbrun. Multiscale Cholesky preconditioning for ill-conditioned problems. *ACM Trans. Graph.*, 40(4), jul 2021a. ISSN 0730-0301. doi: 10.1145/3450626.3459851. URL `https://doi.org/10.1145/3450626.3459851`.

Xiaohui Chen and Yun Yang. Hanson–Wright inequality in Hilbert spaces with application to k-means clustering for non-Euclidean data. *Bernoulli*, 27(1):586–614, 2021.

Yifan Chen, Bamdad Hosseini, Houman Owhadi, and Andrew M Stuart. Solving and learning nonlinear PDEs with Gaussian processes. *Journal of Computational Physics*, 447:110668, 2021b.

Jiawei Chiu and Laurent Demanet. Matrix probing and its conditioning. *SIAM Journal on Numerical Analysis*, 50(1):171–193, 2012.

Dennis D Cox and Finbarr O'Sullivan. Asymptotic analysis of penalized likelihood and related estimators. *The Annals of Statistics*, pages 1676–1695, 1990.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

Maarten V de Hoop, Nikola B Kovachki, Nicholas H Nelsen, and Andrew M Stuart. Convergence rates for learning linear operators from noisy data. *arXiv preprint arXiv:2108.12515*, 2021.

Laurent Demanet, Pierre-David Létourneau, Nicolas Boumal, Henri Calandra, Jiawei Chiu, and Stanley Snelson. Matrix probing: a randomized preconditioner for the wave-equation Hessian. *Applied and Computational Harmonic Analysis*, 32(2):155–168, 2012.

Emilien Dupont, Yee Whye Teh, and Arnaud Doucet. Generative models as distributions of functions. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pages 2989–3015. PMLR, 2022. URL `https://proceedings.mlr.press/v151/dupont22a.html`.

Weinan E and Bing Yu. The deep Ritz method: a deep learning-based numerical algorithm for solving variational problems. *Communications in Mathematics and Statistics*, 6(1):1–12, 2018.

Lawrence C Evans. Partial differential equations. *Graduate studies in mathematics*, 19(2), 1998.

Gregory E Fasshauer. *Meshfree approximation methods with MATLAB*, volume 6. World Scientific, 2007.

Bengt Fornberg and Natasha Flyer. Solving PDEs with radial basis functions. *Acta Numerica*, 24: 215–258, 2015.

Marc G Genton. Classes of kernels for machine learning: a statistics perspective. *Journal of machine learning research*, 2(Dec):299–312, 2001.

Craig R Gin, Daniel E Shea, Steven L Brunton, and J Nathan Kutz. DeepGreen: Deep learning of Green's functions for nonlinear boundary value problems. *Scientific reports*, 11(1):1–14, 2021.

Xiaoxiao Guo, Wei Li, and Francesco Iorio. Convolutional neural networks for steady flow approximation. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 481–490, 2016.

Viktor Holubec, Klaus Kroy, and Stefano Steffenoni. Physically consistent numerical solver for time-dependent Fokker–Planck equations. *Physical Review E*, 99(3):032117, 2019.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, Alain Rakotomamonjy, and Julien Audiffren. Operator-valued kernels for learning from functional response data. *The Journal of Machine Learning Research*, 17(1):613–666, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

Randall J LeVeque. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*. SIAM, 2007.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Graph kernel network for partial differential equations. *arXiv preprint arXiv:2003.03485*, 2020.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=c8P9NQVtmnO`.

Lin Lin, Jianfeng Lu, and Lexing Ying. Fast construction of hierarchical matrix representation from matrix–vector multiplication. *Journal of Computational Physics*, 230(10):4071–4087, 2011.

Lu Lu, Pengzhan Jin, and George Em Karniadakis. DeepONet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv preprint arXiv:1910.03193*, 2019.

Charles A Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural computation*, 17(1):177–204, 2005.

Nicholas H Nelsen and Andrew M Stuart. The random feature model for input-output maps between Banach spaces. *SIAM Journal on Scientific Computing*, 43(5):A3212–A3243, 2021.

Houman Owhadi. Bayesian numerical homogenization. *Multiscale Modeling & Simulation*, 13(3): 812–828, 2015.

Houman Owhadi. Multigrid with rough coefficients and multiresolution operator decomposition from hierarchical information games. *SIAM Review*, 59(1):99–149, 2017.

Houman Owhadi and Clint Scovel. *Operator-Adapted Wavelets, Fast Solvers, and Numerical Homogenization: From a Game Theoretic Approach to Numerical Approximation and Algorithm Design*, volume 35. Cambridge University Press, 2019.

Houman Owhadi, Clint Scovel, and Florian Schäfer. Statistical numerical approximation. *Notices of the AMS*, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

Vern I Paulsen and Mrinal Raghupathi. *An introduction to the theory of reproducing kernel Hilbert spaces*, volume 152. Cambridge university press, 2016.

Andrei D Polyanin and Vladimir E Nazaikinskii. *Handbook of linear partial differential equations for engineers and scientists*. CRC press, 2015.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

J. O. Ramsay and C. J. Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):539–561, 1991. doi: https://doi.org/10.1111/j.2517-6161.1991.tb01844.x. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1991.tb01844.x`.

James O Ramsay. Functional data analysis. *Encyclopedia of Statistical Sciences*, 4, 2004.

James O Ramsay and Bernard W Silverman. *Applied functional data analysis: methods and case studies*. Springer, 2007.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL `https://openreview.net/forum?id=ryQu7f-RZ`.

Albert Reuther, Jeremy Kepner, Chansup Byun, Siddharth Samsi, William Arcand, David Bestor, Bill Bergeron, Vijay Gadepally, Michael Houle, Matthew Hubbell, Michael Jones, Anna Klein, Lauren Milechin, Julia Mullen, Andrew Prout, Antonio Rosa, Charles Yee, and Peter Michaleas. Interactive supercomputing on 40,000 cores for machine learning and data analysis. In *2018 IEEE High Performance extreme Computing Conference (HPEC)*, pages 1–6. IEEE, 2018.

Saburou Saitoh and Yoshihiro Sawano. *Theory of reproducing kernels and applications*. Springer, 2016.

Florian Schäfer, Matthias Katzfuss, and Houman Owhadi. Sparse Cholesky factorization by Kullback–Leibler minimization. *SIAM Journal on Scientific Computing*, 43(3):A2019–A2046, 2021a.

Florian Schäfer, Timothy John Sullivan, and Houman Owhadi. Compression, inversion, and approximate pca of dense kernel matrices at near-linear computational complexity. *Multiscale Modeling & Simulation*, 19(2):688–730, 2021b.

Lloyd N Trefethen. *Spectral methods in MATLAB*. SIAM, 2000.

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

Ming Yuan, T Tony Cai, et al. A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6):3412–3444, 2010.

Yinhao Zhu and Nicholas Zabaras. Bayesian deep convolutional encoder–decoder networks for surrogate modeling and uncertainty quantification. *Journal of Computational Physics*, 366: 415–447, 2018.

Olgierd Cecil Zienkiewicz, Robert Leroy Taylor, Perumal Nithiarasu, and JZ Zhu. *The finite element method*, volume 3. McGraw-hill London, 1977.