Journal of Machine Learning Research 24 (2023) 1-37 Submitted 9/21; Revised 7/23; Published 7/23

On the Estimation of Derivatives Using Plug-in Kernel Ridge Regression Estimators

Zejian Liu ZEJIAN.LIU@RICE.EDU
Department of Statistics
Rice University
Houston, TX 77005, USA

Meng Li MENG@RICE.EDU
Department of Statistics
Rice University
Houston, TX 77005, USA

Editor: Garvesh Raskutti

Abstract

We study the problem of estimating the derivatives of a regression function, which has a wide range of applications as a key nonparametric functional of unknown functions. Standard analysis may be tailored to specific derivative orders, and parameter tuning remains a daunting challenge particularly for high-order derivatives. In this article, we propose a simple plug-in kernel ridge regression (KRR) estimator in nonparametric regression with random design that is broadly applicable for multi-dimensional support and arbitrary mixed-partial derivatives. We provide a non-asymptotic analysis to study the behavior of the proposed estimator in a unified manner that encompasses the regression function and its derivatives, leading to two error bounds for a general class of kernels under the strong L_∞ norm. In a concrete example specialized to kernels with polynomially decaying eigenvalues, the proposed estimator recovers the minimax optimal rate up to a logarithmic factor for estimating derivatives of functions in Hölder and Sobolev classes. Interestingly, the proposed estimator achieves the optimal rate of convergence with the same choice of tuning parameter for any order of derivatives. Hence, the proposed estimator enjoys a *plug-in property* for derivatives in that it automatically adapts to the order of derivatives to be estimated, enabling easy tuning in practice. Our simulation studies show favorable finite sample performance of the proposed method relative to several existing methods and corroborate the theoretical findings on its minimax optimality.

Keywords: Derivative estimation, kernel ridge regression, plug-in property

1. Introduction

Estimating the derivatives of the regression function has a wide range of applications in many areas, such as cosmology (Holsclaw et al., 2013), spatial process models (Banerjee et al., 2003), and shape-constrained function estimation that builds on the derivative process or virtual derivative observations (Riihimäki and Vehtari, 2010; Wang and Berger, 2016). Furthermore, derivative estimation may improve the computational efficiency for nonlinear dynamic system identification (Solak et al., 2003), while serving as a vital tool in detecting local extrema (Song et al., 2006; Li et al., 2021) and efficient modeling of functional data (Dai et al., 2018).

Existing methods for estimating derivatives of regression functions include smoothing spline, local polynomial regression, and difference-based methods. Local polynomial regression and smoothing spline base the estimation of derivatives on estimating the regression function. Key smoothing parameters in these two methods often depend on the order of the derivative being estimated and are difficult to choose in practice (Wahba and Wang, 1990; Charnigo et al., 2011). Difference-based methods require boundary correction, hindering theoretical studies such as those establishing uniform convergence rates. Moreover, existing methods are either restricted to the fixed design setting, or only applicable to one-dimensional support and low-order derivatives. The goal of this article is to develop a unified framework for derivative estimation that achieves broadened applicability and enables simple optimal parameter tuning with theoretical guarantees.

In this paper, we propose a simple plug-in kernel ridge regression (KRR) estimator for the derivatives of the regression function and develop a non-asymptotic framework that provides theoretical support for general kernels. We consider a random design setting with multi-dimensional support, and derive convergence rates for partial mixed derivatives of arbitrary order under the strong L_∞ norm. In a concrete example where the regression function belongs to a Hölder or Sobolev class, we show that the proposed estimator is nearly minimax optimal with the same choice of tuning parameters for *any* order of derivatives to be estimated. Hence, unlike methods such as smoothing spline, the proposed estimator remarkably adapts to the order of derivatives and achieves the so-called *plug-in property* (Bickel and Ritov, 2003) for derivative estimation. This leads to immediate insight for parameter tuning, which along with the closed-form expression substantially facilitates the implementation of the proposed method in broad settings.

Kernel ridge regression (Wahba, 1990; Györfi et al., 2006; Cucker and Zhou, 2007), also known as regularized least squares, is a popular technique in supervised learning and has been widely used in an immense variety of areas, including computer vision (Cheng et al., 2016), speech recognition (Chen et al., 2016), forecasting (Exterkate et al., 2016), and biomedical fields (Mohapatra et al., 2016).

There has been a rich literature on the theoretical guarantees of KRR (Cucker and Smale, 2002; Zhang, 2005; Caponnetto and De Vito, 2007; Steinwart et al., 2009; Mendelson and Neeman, 2010). However, theory on nonparametric functionals of KRR estimators such as derivatives is comparatively underdeveloped. We contribute to the growing literature of KRR by developing non-asymptotic analysis for derivatives of arbitrary order, with added focus on its generality to encompass a large class of kernels and the strong L_∞ norm.

1.1 Related work

One popular method to estimate function derivatives is to differentiate estimates of the regression functions. For example, smoothing spline produces derivative estimation by differentiating the spline basis. Stone (1985); Zhou and Wolfe (2000) studied theoretical properties of smoothing spline, including the L_2 minimax optimal convergence rate. Local polynomial regression is another standard method, which relies on local polynomial fitting obtained by Taylor expansion; Fan and Gijbels (1996); Delecroix and Rosa (1996) provided asymptotic normality and strong uniform consistency for local polynomial regression, respectively. However, the smoothing parameter in these methods typically depends on the order of the derivative and is usually difficult to choose in practice (Wahba and Wang, 1990; Wang and Lin, 2015). The implementation of these methods is also specific to one particular derivative order. Yatracos (1989) related error bounds for general plug-in derivative estimators to those for the original estimators, concluding that derivative estimation is more challenging than the estimation of the original function. This line of research was recently expanded by Yatracos (2019), focusing on mixing density estimation instead of nonparametric regression.

Difference-based methods have attracted increasing attention. These methods create a new noisy dataset with derivatives as the mean, followed by nonparametric smoothing (Müller et al., 1987; Härdle, 1990). Along this line, Charnigo et al. (2011); De Brabanter et al. (2013) proposed an empirical derivative estimator with improved variance and established pointwise consistency. Wang and Lin (2015) derived an asymptotic L_2 convergence rate for estimating the first derivative. However, they required the true regression function to be five times differentiable, which is a very strict assumption. Liu and Brabanter (2018); Liu and De Brabanter (2020) extended the difference-based estimator to random design. Wang et al. (2019) adopted L_1 regression instead of least squares regression, improving the robustness to outliers and heavy-tailed errors. However, difference-based methods typically aim at estimating the first or second derivative of regression functions with one-dimensional support, and have limited developments for high-order derivatives or multi-dimensional cases. Moreover, difference-based estimators require boundary correction in general, necessitating a separate treatment when studying their behavior at the boundaries.

In this article, we provide a non-asymptotic analysis for the proposed plug-in KRR estimator. In the literature of KRR for nonparametric regression, Cucker and Smale (2002) provided a non-asymptotic upper bound under the L_2 norm, utilizing the covering number of an open subset of the reproducing kernel Hilbert space (RKHS). Smale and Zhou (2005, 2007) replaced the covering number technique by the method of integral operators and obtained tighter bounds, but assumed the outputs to be uniformly bounded above, excluding the Gaussian error. This assumption was later relaxed by moment conditions (Wang and Zhou, 2011; Guo and Zhou, 2013). However, they did not particularly focus on learning rates for derivatives of KRR estimators. The optimality of the induced learning rates and parameter tuning that is of great practical relevance, in the presence of varying derivative orders, have not been studied in the literature. Note that derivatives with kernel methods have been considered in different settings, including nonparametric sparse regression (Rosasco et al., 2013) and semi-supervised learning (Cabannes et al., 2021).

1.2 Contributions

Our contributions can be summarized as follows:

(1) We propose a plug-in KRR estimator for derivatives of arbitrary order. The proposed estimator is analytically given and applicable for multi-dimensional support and sub-Gaussian error, enabling fast computation and broad practicability. We allow the derivative order to be zero throughout the article and thus unify the study of the regression function and its derivatives.

(2) We provide two non-asymptotic error bounds for the proposed plug-in KRR estimators under the L_∞ norm: the $\tilde{\mathbb{H}}$-bound for Mercer kernels with uniformly bounded eigenfunctions (Section 3.2) and the $\mathbb{H}$-bound for all Mercer kernels (Section 3.3). To the best of our knowledge, learning rates for derivatives of KRR estimators have not been addressed in the literature. Our analysis rests on an operator-theoretic approach, equivalent kernels, and the Hanson-Wright inequality, encompassing a general class of kernel functions and the strong L_∞ norm.

(3) In a concrete example where the kernel has polynomially decaying eigenvalues and the regression function belongs to a Hölder or Sobolev class, we show that our general analysis recovers the nearly minimax optimal L_2 convergence rate, suggesting the sharpness of the established bounds (Section 4). Given the smoothness level of the regression function, the rate-optimal estimation is achieved under the same choice of the regularization parameter that does not depend on the derivative order. Therefore, the proposed estimator enjoys a remarkable plug-in property that it automatically adapts to the order of the derivative to be estimated, leading to easy tuning in practice.

1.3 Notation

Let $\mathbb{N}$ be the set of all positive integers and write $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. We let $C(\mathcal{X})$ denote the space of continuous functions. For a multi-index $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d) \in \mathbb{N}_0^d$, we write $|\boldsymbol{\beta}| = \beta_1 + \cdots + \beta_d$ and $\partial^{\boldsymbol{\beta}} = \partial_{x_1}^{\beta_1} \cdots \partial_{x_d}^{\beta_d}$. For any $m \in \mathbb{N}$, let $C^m(\mathcal{X})$ stand for the space of all functions possessing continuous mixed partial derivatives up to order m, *i.e.*, $C^m(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} | \partial^{\boldsymbol{\beta}} f \in C(\mathcal{X}) \text{ for all } \boldsymbol{\beta} \in \mathbb{N}_0^d \text{ with } |\boldsymbol{\beta}| \leq m\}$. Let $C(\mathcal{X}, \mathcal{X})$ denote the space of continuous bivariate functions and $C^{2m}(\mathcal{X}, \mathcal{X}) = \{K : \mathcal{X} \times \mathcal{X} \to \mathbb{R} | \partial^{\boldsymbol{\beta},\boldsymbol{\beta}} K \in C(\mathcal{X}, \mathcal{X}) \text{ for all } \boldsymbol{\beta} \in \mathbb{N}_0^d \text{ with } |\boldsymbol{\beta}| \leq m\}$ denote the space of m-times continuously differentiable bivariate functions, where $\partial^{\boldsymbol{\beta},\boldsymbol{\beta}} K(\boldsymbol{x}, \boldsymbol{x}') = \partial_{\boldsymbol{x}}^{\boldsymbol{\beta}} \partial_{\boldsymbol{x}'}^{\boldsymbol{\beta}} K(\boldsymbol{x}, \boldsymbol{x}')$. For any $f : \mathcal{X} \to \mathbb{R}$, let $\|f\|_\infty$ be the L_∞ norm. For two sequences a_n and b_n, we write $a_n \lesssim b_n$ if $a_n \leq C b_n$ for a universal constant $C > 0$, and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

2. Plug-in KRR estimator for function derivatives

Suppose that we have n iid observations $\{X_i, y_i\}_{i=1}^n$ from an unknown data generating probability $\mathbb{P}_0$ on $\mathcal{X} \times \mathbb{R}$, where $\mathcal{X} \subset \mathbb{R}^d$ is a compact metric space for $d \geq 1$. Denote the marginal distribution on $\mathcal{X}$ by $\mathbb{P}_X$ with Lebesgue density p_X. Let $L_{p_X}^2(\mathcal{X})$ be the L_2 space with respect to the measure $\mathbb{P}_X$, with the L_2 norm $\|f\|_2 = (\int_{\mathcal{X}} f^2 d\mathbb{P}_X)^{1/2}$ and the inner product $\langle f, g \rangle_2 = (\int_{\mathcal{X}} fg d\mathbb{P}_X)^{1/2}$. The regression model is given by

$$y_i = f_0(X_i) + \varepsilon_i, \tag{1}$$

where the random error ε_i is sub-Gaussian with mean zero and variance proxy σ^2, *i.e.*, $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[e^{t\varepsilon_i}] \leq e^{\sigma^2 t^2/2}$ for any $t \in \mathbb{R}$. Given a multi-index $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d) \in \mathbb{N}_0^d$, our goal is to estimate $\partial^{\boldsymbol{\beta}} f_0$, the mixed partial derivative of the regression function, assuming its existence.

Let $X = (X_1^T, \ldots, X_n^T)^T \in \mathbb{R}^{n \times d}$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$. Let $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel, *i.e.*, a continuous, symmetric, and positive definite bivariate function. In this article, we propose the following closed-form plug-in KRR estimator for $\partial^{\boldsymbol{\beta}} f_0$, assuming differentiability of K:

$$\widehat{\partial^{\boldsymbol{\beta}} f_0}(\boldsymbol{x}) =: \partial^{\boldsymbol{\beta}} \hat{f}_n(\boldsymbol{x}) = [\partial^{\boldsymbol{\beta}} K(\boldsymbol{x}, X)][K(X,X) + n\lambda \boldsymbol{I}_n]^{-1}\boldsymbol{y}, \tag{2}$$

where $\partial^{\boldsymbol{\beta}} K(\boldsymbol{x}, X) = (\partial_{x_i}^{\beta_i} K(x_i, X_j))_{1\le i\le d, 1\le j\le n}$ is a d by n matrix, $K(X,X)$ is the n by n matrix $(K(X_i, X_j))_{1\le i,j\le n}$, and $\lambda > 0$ is a regularization parameter that possibly depends on the sample size n. Here $\hat{f}_n(\boldsymbol{x}) = \widehat{\partial^{\boldsymbol{0}} f_0}(\boldsymbol{x})$ is the classical KRR estimator for the regression function f_0. It is well known that $\hat{f}_n$ is also the solution to the following optimization problem:

$$\hat{f}_n = \operatorname*{arg\,min}_{f\in\mathbb{H}} \left\{ \frac{1}{n}\sum_{i=1}^{n} (y_i - f(X_i))^2 + \lambda \|f\|_{\mathbb{H}}^2 \right\}, \tag{3}$$

where $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ is the reproducing kernel Hilbert space (RKHS) induced by the kernel K.

The closed-form expression in (2) enables fast calculation for any order of derivatives. The proposed estimator is applicable for d-dimensional support with $d \ge 1$. Taking one-dimensional support $\mathcal{X} \subset \mathbb{R}$ as a special case, the plug-in KRR estimator for $f_0^{(m)}$ with $m \in \mathbb{N}_0$ is

$$\hat{f}_n^{(m)}(x) = [\partial^m K(x, X)][K(X,X) + n\lambda \boldsymbol{I}_n]^{-1}\boldsymbol{y},$$

where $\partial^m K(x, X) = (\partial_x^m K(x, X_j))_{1\le j\le n}$ is a 1 by n vector. In addition, (2) enables convenient inference. For example, the proposed estimator at any $\boldsymbol{x}$ is normally distributed for Gaussian error $\varepsilon_i \sim N(0, \sigma^2)$ with variance given by

$$\sigma^2 [\partial^{\boldsymbol{\beta}} K(\boldsymbol{x}, X)][K(X,X) + n\lambda \boldsymbol{I}_n]^{-2}[\partial^{\boldsymbol{\beta}} K(\boldsymbol{x}, X)]^T.$$

3. Non-asymptotic analysis

We take an operator-theoretic approach to study non-asymptotic properties of the plug-in KRR estimator, which characterizes behaviors of the proposed estimator, provides insights for choosing regularization parameters, and leads to asymptotic optimality.

3.1 Preliminaries

We begin with reviewing preliminaries and introducing commonly used notation (Smale and Zhou, 2005, 2007; Steinwart et al., 2009). For any $f \in L^2_{p_X}(\mathcal{X})$, we introduce an integral operator $L_K : L^2_{p_X}(\mathcal{X}) \to \mathbb{H} \subset L^2_{p_X}(\mathcal{X})$ defined by

$$L_K(f)(\boldsymbol{x}) = \int_{\mathcal{X}} K(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}') d\mathbb{P}_X(\boldsymbol{x}'), \quad \boldsymbol{x} \in \mathcal{X}.$$

Since L_K is compact, positive definite, and self-adjoint on $L^2_{p_X}(\mathcal{X})$ (*i.e.*, as an operator mapping $L^2_{p_X}(\mathcal{X})$ to $L^2_{p_X}(\mathcal{X})$), by the spectral theorem (see, *e.g.*, Theorem A.5.13 in Steinwart and Christmann (2008)), there exist countable pairs of eigenvalues and eigenfunctions $(\mu_i, \psi_i)_{i\in\mathbb{N}} \subset (0,\infty) \times L^2_{p_X}(\mathcal{X})$ of L_K such that

$$L_K \psi_i = \mu_i \psi_i, \quad i \in \mathbb{N},$$

where $\{\psi_i\}_{i=1}^{\infty}$ form an orthonormal basis of $L^2_{p_X}(\mathcal{X})$ and $\mu_1 \geq \mu_2 \geq \cdots > 0$ with $\lim_{i\to\infty} \mu_i = 0$. By Mercer's Theorem, we have that for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$,

$$K(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{\infty} \mu_i \psi_i(\boldsymbol{x}) \psi_i(\boldsymbol{x}'),$$

where the convergence is absolute and uniform. It follows that $\mathbb{H}$ can be characterized by a series representation

$$\mathbb{H} = \left\{ f \in L^2_{p_X}(\mathcal{X}) : \|f\|_{\mathbb{H}}^2 = \sum_{i=1}^{\infty} \frac{f_i^2}{\mu_i} < \infty \right\},$$

where $f_i = \langle f, \psi_i \rangle_2$. The corresponding inner product is given by $\langle f, g \rangle_{\mathbb{H}} = \sum_{i=1}^{\infty} f_i g_i / \mu_i$ for any $f = \sum_{i=1}^{\infty} f_i \psi_i$ and $g = \sum_{i=1}^{\infty} g_i \psi_i$ in $\mathbb{H}$.

We then define the sample analog $L_{K,X} : \mathbb{H} \to \mathbb{H}$ by

$$L_{K,X}(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i) K_{X_i}, \tag{4}$$

where $K_{\boldsymbol{x}}(\cdot) := K(\boldsymbol{x}, \cdot)$. It is easy to see $L_{K,X}$ is also a compact, positive definite, self-adjoint operator because for any $f, g \in \mathbb{H}$, we have

$$\langle f, L_{K,X} g \rangle_{\mathbb{H}} = \frac{1}{n} \sum_{i=1}^{n} f(X_i) g(X_i) = \langle L_{K,X} f, g \rangle_{\mathbb{H}}$$

and $\langle f, L_{K,X} f \rangle_{\mathbb{H}} \geq 0$. Thus, the eigenvalues of $L_{K,X}$ are all non-negative, which implies

$$\|(L_{K,X} + \lambda I)^{-1} f\|_{\mathbb{H}} \leq \frac{1}{\lambda} \|f\|_{\mathbb{H}}, \tag{5}$$

for any $f \in \mathbb{H}$. We remark that the operator L_K can also be defined on $\mathbb{H}$ and so does $L_{K,X}$ on the space of all bounded functions; we use the same notation when they are defined on different domains.

We further consider a proximate function of f_0 in $\mathbb{H}$

$$f_\lambda = (L_K + \lambda I)^{-1} L_K f_0,$$

where I is the identity operator. The function f_λ is chosen this way to minimize the population counterpart of (3), *i.e.*,

$$f_\lambda = \underset{f \in \mathbb{H}}{\arg\min} \left\{ \|f - f_0\|_2^2 + \lambda \|f\|_{\mathbb{H}}^2 \right\}. \tag{6}$$

We next present two non-asymptotic bounds for Mercer kernels with uniformly bounded eigenfunctions and general Mercer kernels, respectively, which provide the basis for more specific rate calculation and may be of independent interest for the KRR community.

3.2 $\tilde{\mathbb{H}}$-bound for Mercer kernels with uniformly bounded eigenfunctions

The proximate function f_λ can be obtained using another integral operator $L_{\tilde{K}}$ through $f_\lambda = L_{\tilde{K}} f_0$, where $\tilde{K}$ is the so-called equivalent kernel (Rasmussen and Williams, 2006; Sollich and Williams, 2005). Compared to K, the equivalent kernel $\tilde{K}$ has the same eigenfunctions but its eigenvalues are altered to $\nu_i = \mu_i/(\lambda + \mu_i)$ for $i \in \mathbb{N}$, *i.e.*,

$$\tilde{K}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{\infty} \nu_i \psi_i(\boldsymbol{x}) \psi_i(\boldsymbol{x}').$$

Let $\tilde{\mathbb{H}}$ be the RKHS induced by $\tilde{K}$, which is equivalent to $\mathbb{H}$ as a functional space, but with a different inner product

$$\langle f, g \rangle_{\tilde{\mathbb{H}}} = \langle f, g \rangle_2 + \lambda \langle f, g \rangle_{\mathbb{H}}.$$

Let the corresponding RKHS norm be $\| \cdot \|_{\tilde{\mathbb{H}}}$. Note that $\tilde{K}$ is also a Mercer kernel; thus, all preliminaries in Section 3.1 hold for $\tilde{K}$. For example, in view of (4), we can similarly define the sample analog by

$$L_{\tilde{K},X}(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \tilde{K}_{X_i},$$

which is compact, positive definite, and self-adjoint. Note that throughout the article, the tilde notation such as $\tilde{K}$ and $\tilde{\mathbb{H}}$ indicates dependence on λ, although λ may not be explicitly spelled out in the notation with exceptions of $\tilde{\kappa}_\lambda^2$ and $\tilde{\kappa}_{\boldsymbol{\beta},\lambda}^2$ defined later.

We introduce the following assumption on the differentiability of the regression function and the covariance kernel.

Assumption A *There exists an $m \in \mathbb{N}_0$ such that $f_0 \in C^m(\mathcal{X})$ and $K \in C^{2m}(\mathcal{X}, \mathcal{X})$.*

Under Assumption A, we can estimate $\partial^{\boldsymbol{\beta}} f_0$ by $\partial^{\boldsymbol{\beta}} \hat{f}_n$ for any $|\boldsymbol{\beta}| \leq m$. Note that we do not necessarily require that f_0 and K have exactly the same smoothness level. Indeed, if there is a mismatch between the differentiability of f_0 and K such that $f_0 \in C^{m_1}(\mathcal{X})$ and $K \in C^{2m_2}(\mathcal{X}, \mathcal{X})$, we can take $m = \min\{m_1, m_2\}$ to satisfy Assumption A. From the perspective of kernel selection, this assumption indicates that for estimating $\partial^{\boldsymbol{\beta}} f_0$, we should choose a kernel $K \in C^{2m}(\mathcal{X}, \mathcal{X})$ such that $m \geq |\boldsymbol{\beta}|$. Such kernels are widely available. For example, it is satisfied by a general class of kernels with polynomially decaying eigenvalues (given in Definition 8) under mild conditions. In particular, the Matérn kernel is known to be $2m$ times differentiable if and only if the smoothness parameter $\nu > m$ (Stein, 1999, Chapter 2.7). The squared exponential kernel as a limit case of Matérn kernel satisfies Assumption A for any m. Assumption A directly implies that $\psi_i \in C^m(\mathcal{X})$.

Define $\tilde{\kappa}_\lambda^2 := \sup_{\boldsymbol{x} \in \mathcal{X}} \tilde{K}(\boldsymbol{x}, \boldsymbol{x})$. It is easy to see $\tilde{\kappa}_\lambda^2 \leq C_\psi^2 \sum_{i=1}^{\infty} \nu_i \lesssim \sum_{i=1}^{\infty} \mu_i/(\lambda + \mu_i)$, where the last expression is the effective dimension (Zhang, 2005) of the kernel K with respect to $L_{p_X}^2(\mathcal{X})$. We also define high-order analogies of $\tilde{\kappa}_\lambda^2$ for any multi-index $\boldsymbol{\beta} \in \mathbb{N}_0^d$ and $|\boldsymbol{\beta}| \leq m$:

$$\tilde{\kappa}_{\boldsymbol{\beta},\lambda}^2 := \sup_{\boldsymbol{x} \in \mathcal{X}} \partial^{\boldsymbol{\beta},\boldsymbol{\beta}} \tilde{K}(\boldsymbol{x}, \boldsymbol{x}) = \sup_{\boldsymbol{x} \in \mathcal{X}} \sum_{i=1}^{\infty} \frac{\mu_i}{\lambda + \mu_i} \{\partial^{\boldsymbol{\beta}} \psi_i(\boldsymbol{x})\}^2. \tag{7}$$

Note that $\tilde{\kappa}_\lambda^2 = \tilde{\kappa}_{\boldsymbol{\beta},\lambda}^2$ with $\boldsymbol{\beta} = \mathbf{0}$. In general, $\tilde{\kappa}_{\boldsymbol{\beta},\lambda}^2$ is determined by the decay rate of the eigenvalues of K, the derivatives of the eigenfunctions, and the regularization parameter λ. For given $\lambda > 0$, there holds

$$\tilde{\kappa}_{\boldsymbol{\beta},\lambda}^2 \leq \lambda^{-1} \sup_{\boldsymbol{x}\in\mathcal{X}} \sum_{i=1}^{\infty} \mu_i \{\partial^{\boldsymbol{\beta}} \psi_i(\boldsymbol{x})\}^2 = \lambda^{-1} \sup_{\boldsymbol{x}\in\mathcal{X}} \partial^{\boldsymbol{\beta},\boldsymbol{\beta}} K(\boldsymbol{x},\boldsymbol{x}) < \infty,$$

where the boundedness of $\sup_{\boldsymbol{x}\in\mathcal{X}} \partial^{\boldsymbol{\beta},\boldsymbol{\beta}} K(\boldsymbol{x},\boldsymbol{x})$ in the last step is due to Assumption A as $\partial^{\boldsymbol{\beta},\boldsymbol{\beta}} K(\boldsymbol{x},\boldsymbol{x})$ is a continuous bivariate function on a compact support $\mathcal{X}$.

The following assumption on the eigenfunctions pertains to the equivalent kernel technique considered in this section; the error bound established in Section 3.3 does not require such an assumption.

Assumption B *There exists a constant $C_\psi > 0$ such that $\|\psi_i\|_\infty \leq C_\psi$ for all $i \in \mathbb{N}$.*

The following lemma indicates that functions in the RKHS inherit the differentiability of the kernel, and the L_∞ norm of the derivative is upper bounded by the RKHS norm of the function.

Lemma 1 *Under Assumption A, $f \in C^m(\mathcal{X})$ for any $f \in \tilde{\mathbb{H}}$. Moreover, for any $\boldsymbol{\beta} \in \mathbb{N}_0^d$, $|\boldsymbol{\beta}| \leq m$, we have $\|\partial^{\boldsymbol{\beta}} f\|_\infty \leq \tilde{\kappa}_{\boldsymbol{\beta},\lambda} \|f\|_{\tilde{\mathbb{H}}}$ for any $f \in \tilde{\mathbb{H}}$.*

Theorem 2 provides error bounds for Mercer kernels with bounded eigenfunctions.

Theorem 2 ($\tilde{\mathbb{H}}$-bound) *Under Assumptions A and B, for any $\boldsymbol{\beta} \in \mathbb{N}_0^d$ with $|\boldsymbol{\beta}| \leq m$ and $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that*

$$\begin{aligned}\|\partial^{\boldsymbol{\beta}} \hat{f}_n - \partial^{\boldsymbol{\beta}} f_0\|_\infty \leq \|\partial^{\boldsymbol{\beta}} f_\lambda - \partial^{\boldsymbol{\beta}} f_0\|_\infty &+ \frac{\tilde{\kappa}_{\boldsymbol{\beta},\lambda} \tilde{\kappa}_\lambda^{-1} C(n,\tilde{\kappa}_\lambda)}{1 - C(n,\tilde{\kappa}_\lambda)} \|f_\lambda - f_0\|_\infty \\ &+ \frac{1}{1 - C(n,\tilde{\kappa}_\lambda)} \frac{C_1 \tilde{\kappa}_{\boldsymbol{\beta},\lambda} \tilde{\kappa}_\lambda \sigma \sqrt{\log(3/\delta)}}{\sqrt{n}},\end{aligned}$$

where $C_1 > 0$ does not depend on K or n, and $C(n,\tilde{\kappa}_\lambda) = \frac{\tilde{\kappa}_\lambda^2 \sqrt{\log(3/\delta)}}{\sqrt{n}} \left(4 + \frac{4\tilde{\kappa}_\lambda \sqrt{\log(3/\delta)}}{3\sqrt{n}}\right)$.

3.3 $\mathbb{H}$-bound for general Mercer kernels

The $\tilde{\mathbb{H}}$-bound established in the preceding section relies on the crucial Assumption B that the eigenfunctions are uniformly bounded, which does not necessarily hold for all Mercer kernels. For example, Zhou (2002) constructed a C^∞ kernel that does not satisfy this assumption. To thoroughly study the learning rate in a more general setting, we provide another error bound under the RKHS norm $\|\cdot\|_{\mathbb{H}}$ for any Mercer kernel, referred to as the $\mathbb{H}$-bound.

Let $f_{X,\lambda}$ be the noiseless counterpart of $\hat{f}_n$ by replacing noisy data with their means given by the true regression function, *i.e.*,

$$f_{X,\lambda} := K(\cdot, X)[K(X,X) + n\lambda \boldsymbol{I}_n]^{-1} f_0(X),$$

where $f_0(X) := (f_0(X_1), \dots, f_0(X_n))^T$. An equivalent operator-based representation akin to (4) gives $f_{X,\lambda} = (L_{K,X} + \lambda I)^{-1} L_{K,X} f_0$.

The following lemma is a parallel version of Lemma 1 but applies to K and the $\|\cdot\|_{\mathbb{H}}$ norm. While $\tilde{\kappa}^2_{\boldsymbol{\beta},\lambda}$ relies on the regularization parameter λ, the characterization of using the $\|\cdot\|_{\mathbb{H}}$ norm only involves the single parameter $\kappa^2_{\boldsymbol{\beta}}$ of K, where $\kappa^2_{\boldsymbol{\beta}} := \sup_{\boldsymbol{x}\in\mathcal{X}} \partial^{\boldsymbol{\beta},\boldsymbol{\beta}} K(\boldsymbol{x},\boldsymbol{x}) < \infty$ and $\kappa^2 := \kappa^2_{\mathbf{0}}$.

Lemma 3 *Under Assumption A, $f \in C^m(\mathcal{X})$ for any $f \in \mathbb{H}$. Moreover, for any $\boldsymbol{\beta} \in \mathbb{N}_0^d$, $|\boldsymbol{\beta}| \le m$, we have $\|\partial^{\boldsymbol{\beta}} f\|_\infty \le \kappa_{\boldsymbol{\beta}} \|f\|_{\mathbb{H}}$ for any $f \in \mathbb{H}$.*

Invoking Lemma 3 and decomposing $\hat{f}_n - f_\lambda = (\hat{f}_n - f_{X,\lambda}) + (f_{X,\lambda} - f_\lambda)$ yield convergence rates of the mixed partial derivatives of $\hat{f}_n$ under the L_∞ norm.

Theorem 4 ($\mathbb{H}$-bound) *Under Assumption A, for any $\boldsymbol{\beta} \in \mathbb{N}_0^d$ with $|\boldsymbol{\beta}| \le m$ and $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that*

$$\|\partial^{\boldsymbol{\beta}} \hat{f}_n - \partial^{\boldsymbol{\beta}} f_0\|_\infty \le \|\partial^{\boldsymbol{\beta}} f_\lambda - \partial^{\boldsymbol{\beta}} f_0\|_\infty + \frac{\kappa_{\boldsymbol{\beta}} \kappa \|f_0\|_\infty \sqrt{\log(9/\delta)}}{\sqrt{n}\lambda} \left(10 + \frac{4\kappa\sqrt{\log(9/\delta)}}{3\sqrt{n}\lambda} \right) + \frac{C_2 \kappa_{\boldsymbol{\beta}} \kappa \sigma \sqrt{\log(3/\delta)}}{\sqrt{n}\lambda},$$

where $C_2 > 0$ does not depend on K or n.

We have established two non-asymptotic error bounds in Theorem 2 and Theorem 4 under the L_∞ norm. A few remarks are in order:

Remark Both $\tilde{\mathbb{H}}$-bound and $\mathbb{H}$-bound have three terms. This three-term structure stems from applying the triangle inequality to the decomposition $\partial^{\boldsymbol{\beta}} \hat{f}_n - \partial^{\boldsymbol{\beta}} f_0 = (\partial^{\boldsymbol{\beta}} \hat{f}_\lambda - \partial^{\boldsymbol{\beta}} f_0) + (\partial^{\boldsymbol{\beta}} \hat{f}_n - \partial^{\boldsymbol{\beta}} f_\lambda)$, with $\|\partial^{\boldsymbol{\beta}} f_\lambda - \partial^{\boldsymbol{\beta}} f_0\|_\infty$ being the first term, and $\partial^{\boldsymbol{\beta}} \hat{f}_n - \partial^{\boldsymbol{\beta}} f_\lambda$ bounded by the second and third terms combined. $\tilde{\mathbb{H}}$-bound and $\mathbb{H}$-bound use different approaches to bound $\partial^{\boldsymbol{\beta}} \hat{f}_n - \partial^{\boldsymbol{\beta}} f_\lambda$. As the name suggests, $\tilde{\mathbb{H}}$-bound employs the $\|\cdot\|_{\tilde{\mathbb{H}}}$ norm of $\partial^{\boldsymbol{\beta}} \hat{f}_n - \partial^{\boldsymbol{\beta}} f_\lambda$, while $\mathbb{H}$-bound uses the $\|\cdot\|_{\mathbb{H}}$ norm. When the observations are noiseless, *i.e.*, $\boldsymbol{y} = f_0(X)$ in (3), the two bounds can be simplified by letting $\sigma = 0$, zeroing out the third term in both bounds. Indeed, all subsequent error bounds imply a noise-free version by substituting $\sigma = 0$; we do not present them separately due to space constraints.

Remark Theorem 2 and Theorem 4 are applicable for any derivative order $\boldsymbol{\beta} \in \mathbb{N}_0^d$ as long as $|\boldsymbol{\beta}| \le m$. For example, if $K \in C^2(\mathcal{X}, \mathcal{X})$, these two theorems give learning rates for each component (*e.g.*, the jth component) of the gradient $\nabla(\hat{f}_n - f_0)$ by setting the jth element of $\boldsymbol{\beta}$ to one and others to zero, in which case $\kappa^2_{\boldsymbol{\beta}} = \sup_{\boldsymbol{x}\in\mathcal{X}} \partial_{x_j} \partial_{x_j} K(\boldsymbol{x},\boldsymbol{x})$ for $1 \le j \le d$. Although our focus is on derivative estimation, setting $\boldsymbol{\beta} = \mathbf{0}$ in the two theorems also provides error bounds for estimating the regression function using the KRR estimator; hence, we unify the study of the regression function and its derivatives in this article. There are existing error bounds for the KRR estimator $\hat{f}_n$. For example, Theorem 6 in Smale and Zhou (2005) established an RKHS bound for $\hat{f}_n - f_\lambda$ under the bounded sampling setting (*i.e.*, the response y_i is bounded), and their rate $O(1/\sqrt{n}\lambda)$ is similar to our $\mathbb{H}$-bound with $\boldsymbol{\beta} = \mathbf{0}$; Yang et al. (2017) provided error bounds for $\hat{f}_n$ under the L_∞ norm for kernels with polynomially decaying eigenvalues.

$\mathbb{H}$-bound is applicable for any Mercer kernels, broadening the applicability of the proposed method. $\tilde{\mathbb{H}}$-bound relies on the additional Assumption B. When both $\tilde{\mathbb{H}}$-bound and $\mathbb{H}$-bound hold, the following result compares them by providing sufficient conditions under which $\mathbb{H}$-bound cannot be tighter than $\tilde{\mathbb{H}}$-bound.

Corollary 5 *Take $\delta = n^{-10}$ in both $\tilde{\mathbb{H}}$-bound and $\mathbb{H}$-bound. If $\|f_\lambda - f_0\|_\infty = o(1)$, $\tilde{\kappa}_\lambda^2 = o(\sqrt{n/\log n})$ and $\tilde{\kappa}_{\boldsymbol{\beta},\lambda}\tilde{\kappa}_\lambda = o(\lambda^{-1})$, then $\tilde{\mathbb{H}}$-bound is asymptotically less than $\mathbb{H}$-bound.*

The three conditions in Corollary 5 can be verified using special examples. For instance, considering the kernel and regression function in Theorem 14 and invoking Lemma 1, Lemma 11 and Lemma 13, $\|f_\lambda - f_0\|_\infty = o(1)$ and $\tilde{\kappa}_{\boldsymbol{\beta},\lambda}\tilde{\kappa}_\lambda = o(\lambda^{-1})$ automatically hold, whereas $\tilde{\kappa}_\lambda^2 = o(\sqrt{n/\log n})$ is equivalent to $\lambda \gtrsim (\log n/n)^\alpha$. In particular, this condition on λ is satisfied by the optimal value $(\log n/n)^{\frac{2\alpha}{2\alpha+1}}$ derived in Theorem 14.

Theorem 2 and Theorem 4 lead to convergence rates of the plug-in KRR estimator $\partial^{\boldsymbol{\beta}}\hat{f}_n$ by invoking augmenting estimates of $\partial^{\boldsymbol{\beta}} f_\lambda - \partial^{\boldsymbol{\beta}} f_0$. We conclude this section with a few relatively abstract examples for such estimates, and introduce another more concrete example in detail in the next section.

Theorem 6 *Suppose Assumption A holds and $\boldsymbol{\beta} \in \mathbb{N}_0^d$ with $|\boldsymbol{\beta}| \leq m$.*

(a) Under Assumption B, if $L_K^{-r} f_0 \in L^2_{p_X}(\mathcal{X})$ for some $1/2 < r \leq 1$, then it holds

$$\|\partial^{\boldsymbol{\beta}} f_\lambda - \partial^{\boldsymbol{\beta}} f_0\|_\infty \leq \kappa_{\boldsymbol{\beta}} \lambda^{r-1/2} \|L_K^{-r} f_0\|_2.$$

(b) Suppose that K assumes eigendecomposition with respect to the Fourier basis and $L_K^{-r} f_0 \in C^p(\mathcal{X})$ for some $0 < r \leq 1$ and $p > d + |\boldsymbol{\beta}|$. Then there exists $C_3 > 0$ such that

$$\|\partial^{\boldsymbol{\beta}} f_\lambda - \partial^{\boldsymbol{\beta}} f_0\|_\infty \leq C_3 \lambda^r \zeta(p - d + 1 - |\boldsymbol{\beta}|).$$

Remark The condition $L_K^{-r} f_0 \in L^2_{p_X}(\mathcal{X})$ is adopted from Smale and Zhou (2007), in which r can be understood as a smoothness parameter of f_0. When $r = 1/2$, the condition $L_K^{-1/2} f_0 \in L^2_{p_X}(\mathcal{X})$ is equivalent to $f_0 \in \mathbb{H}$. To see this, note that $\|L_K^{-1/2} f_0\|_2^2 = \|\sum_{i=1}^\infty f_i \psi_i/\sqrt{\mu_i}\|_2^2 = \sum_{i=1}^\infty f_i^2/\mu_i = \|f\|_{\mathbb{H}}^2$. Hence, part (a) of Theorem 6 provides a rate for $f_0 \in \mathbb{H}$, while part (b) allows a wider range of r and does not necessarily require $f_0 \in \mathbb{H}$. But part (b) requires the image $L_K^{-r} f_0 \in C^p(\mathcal{X})$. We next study a general setting where $L_K^{-r} f_0 \in L^2_{p_X}(\mathcal{X})$ for $0 < r \leq 1/2$.

We state the following assumption on an embedding property, which is also considered in Fischer and Steinwart (2020).

Assumption C $\|L_K^{q/2} f\|_\infty \leq A\|f\|_2$ *for $0 < q \leq 1$, some constant $A > 0$ and any $f \in L^2_{p_X}(\mathcal{X})$.*

The larger q is, the weaker the embedding property is. Assumption C always holds for $q = 1$ by noting that $\|L_K^{1/2} f\|_\infty \leq \kappa \|L_K^{1/2} f\|_{\mathbb{H}} = \kappa \|f\|_2$.

Theorem 7 *Under Assumptions A and C, if K assumes eigendecomposition with respect to the Fourier basis and $L_K^{-r} f_0 \in L^2_{p_X}(\mathcal{X})$ for some $0 < r \leq 1/2$, then there exists $C_4 > 0$ such that for any $\boldsymbol{\beta} \in \mathbb{N}_0^d$ with $|\boldsymbol{\beta}| \leq m$,*

$$\|\partial^{\boldsymbol{\beta}} f_\lambda - \partial^{\boldsymbol{\beta}} f_0\|_\infty \leq A C_4 \lambda^{r - q/2 - q|\boldsymbol{\beta}|} \|g^*\|_2,$$

where $g^* \in L^2_{p_X}(\mathcal{X})$ is determined by f_0, r and $\boldsymbol{\beta}$.

4. Nearly minimax optimal rate for Hölder and Sobolev class functions

In this section, we demonstrate the sharpness of the established bounds using a concrete example. In particular, we use kernels with polynomially decaying eigenvalues for K and consider $\mathcal{X} = [0,1]$ along with a uniform sampling process for p_X to ease presentation. We also assume that the eigenfunctions of the kernel $\{\psi_i\}_{i=1}^{\infty}$ are the Fourier basis:

$$\psi_1(x) = 1,\ \psi_{2i}(x) = \cos(2\pi i x),\ \psi_{2i+1} = \sin(2\pi i x), i \in \mathbb{N}, \tag{8}$$

which clearly satisfies Assumption B. We formalize the definition of such kernels as follows.

Definition 8 *A kernel with polynomially decaying eigenvalues $K_\alpha : [0,1] \times [0,1] \to \mathbb{R}$ assumes an eigendecomposition with respect to the Lebesgue measure μ such that the eigenvalues $\mu_i \asymp i^{-2\alpha}$ for some $\alpha > 0$ and the eigenfunctions $\{\psi_i\}_{i=1}^{\infty}$ are the Fourier basis functions.*

Examples of kernels with polynomially decaying eigenvalues include the Matérn kernel and Sobolev kernel (Wahba, 1990; Gu, 2013). For example, it is well known that the eigenvalues of Matérn kernel with parameter ν satisfy $\mu_i \asymp i^{-2(\nu+1/2)}$ for $i \in \mathbb{N}$. In the following, we consider two function classes, a Hölder class $H^\alpha[0,1]$ and a Sobolev class $S^\alpha[0,1]$, for the true regression function f_0.

Definition 9 *Let $\{\psi_i\}_{i=1}^{\infty}$ be the Fourier basis of $L^2_\mu[0,1]$ in (8). For any $\alpha > 0$, the Hölder class $H^\alpha[0,1]$ is a Hilbert space defined as*

$$H^\alpha[0,1] = \left\{ f \in L^2_\mu[0,1] : \|f\|^2_{H^\alpha[0,1]} = \sum_{i=1}^{\infty} i^\alpha |f_i| < \infty \right\},$$

where $f_i = \langle f, \psi_i \rangle_2$.

For any $f \in H^\alpha[0,1]$, f lies in the *α-smooth Hölder space*, *i.e.*, it has continuous derivatives up to order $\lfloor \alpha \rfloor$ and the $\lfloor \alpha \rfloor$th derivative is Lipschitz continuous of order $\alpha - \lfloor \alpha \rfloor$ (Yang et al., 2017). To see this, note that

$$\begin{aligned} |f^{\lfloor \alpha \rfloor}(x) - f^{\lfloor \alpha \rfloor}(x')| &= \left| \sum_{i=1}^{\infty} f_i (\psi_i^{\lfloor \alpha \rfloor}(x) - \psi_i^{\lfloor \alpha \rfloor}(x')) \right| \lesssim \sum_{i=1}^{\infty} |f_i| i^{\lfloor \alpha \rfloor} \\ &\lesssim \sum_{i=1}^{\infty} |f_i| i^{\lfloor \alpha \rfloor} (i|x - x'|)^{\alpha - \lfloor \alpha \rfloor} \lesssim |x - x'|^{\alpha - \lfloor \alpha \rfloor}. \end{aligned}$$

Definition 10 *Let $\{\psi_i\}_{i=1}^{\infty}$ be the Fourier basis of $L^2_\mu[0,1]$ in (8). For any $\alpha > 1/2$, the Sobolev class $S^\alpha[0,1]$ is a Hilbert space defined as*

$$S^\alpha[0,1] = \left\{ f \in L^2_\mu[0,1] : \|f\|^2_{S^\alpha[0,1]} = \sum_{i=1}^{\infty} i^{2\alpha} f_i^2 < \infty \right\},$$

where $f_i = \langle f, \psi_i \rangle_2$.

For $\alpha \in \mathbb{N}$, functions in $S^\alpha[0,1]$ belong to the *α-smooth Sobolev space* (cf. Theorem 7.11 in Wasserman (2006)), which consists of functions with absolutely continuous $\alpha-1$ derivatives and whose αth derivative has uniformly bounded L_2 norm and is also a Hilbert space.

It is easy to see that $H^\alpha[0,1] \subset S^\alpha[0,1]$ by definition. On the other hand, $S^{\alpha+\alpha_0}[0,1] \subset H^\alpha[0,1]$ for $\alpha_0 > 1/2$. Indeed, it holds that

$$i^\alpha |f_i| \leq i^{-2\alpha_0} + i^{2(\alpha+\alpha_0)} f_i^2$$

for all $i \in \mathbb{N}$. Hence, if $f \in S^{\alpha+\alpha_0}[0,1]$, we have

$$\sum_{i=1}^\infty i^\alpha |f_i| \leq \sum_{i=1}^\infty i^{-2\alpha_0} + \sum_{i=1}^\infty i^{2(\alpha+\alpha_0)} f_i^2 < \infty.$$

Considering the equivalent kernel $\tilde{K}_\alpha$ of K_α. Let the higher-order analog of the effective dimension for K_α be $\tilde{\kappa}^2_{\alpha,m,\lambda}$ for $m \in \mathbb{N}_0$, where the subscript α emphasizes the use of K_α compared to the general definition in (7). Similarly to the preceding section, all results in this section cover the regression function as a special case with $m = 0$. For example, we allow $m = 0$ in $\tilde{\kappa}_{\alpha,m,\lambda}$, which corresponds to $\tilde{\kappa}^2_{\alpha,0,\lambda} = \tilde{\kappa}^2_{\alpha,\lambda} = \sup_{x\in[0,1]} \tilde{K}_\alpha(x,x)$.

Lemma 11 provides the differentiability of K_α and the exact order of $\tilde{\kappa}_{\alpha,m,\lambda}$ with respect to λ.

Lemma 11 *If $\alpha > m + 1/2$ for $m \in \mathbb{N}_0$, then $K_\alpha \in C^{2m}([0,1]\times[0,1])$, $\tilde{K}_\alpha \in C^{2m}([0,1]\times[0,1])$ and $\tilde{\kappa}^2_{\alpha,m,\lambda} \asymp \lambda^{-\frac{2m+1}{2\alpha}}$.*

Thus, K_α satisfies Assumption A whenever $\alpha > m + 1/2$. The 1/2 gap between m and α appears to be smaller than those required by existing literature; for example, local polynomial and smoothing splines often require the regression function $f_0 \in C^{m+1}(\mathcal{X})$ when estimating the mth derivative (De Brabanter et al., 2013; Charnigo et al., 2011), and difference-based methods such as Wang and Lin (2015) assumed the true regression function to be five times differentiable when estimating the first derivative.

The next lemma studies the differentiability of functions in the RKHS $\mathbb{H}_\alpha$ induced by K_α. It turns out that the equivalent RKHS norm upper bounds the L_2 norm of the derivatives. Note that $\mathbb{H}_\alpha$ and $\tilde{\mathbb{H}}_\alpha$ consist of the same class of functions, thus sharing the same differentiability property.

Lemma 12 *If $\alpha > m + 1/2$ for $m \in \mathbb{N}_0$, then $\mathbb{H}_\alpha \subset C^m[0,1]$. Moreover, there exists a constant $C_m > 0$ that does not depend on λ such that $\|f^{(m)}\|_2 \leq C_m \tilde{\kappa}^{-1}_{\alpha,\lambda} \tilde{\kappa}_{\alpha,m,\lambda} \|f\|_{\tilde{\mathbb{H}}_\alpha}$ for any $f \in \mathbb{H}_\alpha$.*

As we have seen in Theorem 2 and Theorem 4, calculating the learning rate of $\hat{f}_n^{(m)} - f_0^{(m)}$ requires the rate of $f_\lambda^{(m)} - f_0^{(m)}$. We next provide the error bound for this quantity under the $\tilde{\mathbb{H}}_\alpha$ norm.

Lemma 13 *Suppose $f_0 \in H^\alpha[0,1]$ or $f_0 \in S^\alpha[0,1]$ for $\alpha > 1/2$. If the kernel is chosen to be K_α, then it holds that*

$$\|f_\lambda - f_0\|_{\tilde{\mathbb{H}}_\alpha} \lesssim \lambda^{\frac{1}{2}}.$$

When $m = 0$, the three lemmas above provide error bounds for estimating the regression function. In particular, Lemma 11 implies $\tilde{\kappa}^2_{\alpha,\lambda} \asymp \lambda^{-\frac{1}{2\alpha}}$, Lemma 12 gives $\|f\|_2 \leq \|f\|_{\tilde{\mathbb{H}}_\alpha}$, and Lemma 13 leads to $\|f_\lambda - f_0\|_{\tilde{\mathbb{H}}_\alpha} \lesssim \lambda^{\frac{1}{2}}$.

We are now in a position to present a non-asymptotic convergence rate of $\hat{f}_n^{(m)}$.

Theorem 14 *Suppose $f_0 \in H^\alpha[0,1]$ or $f_0 \in S^\alpha[0,1]$ for $\alpha > m + 1/2$ and $m \in \mathbb{N}_0$, and the kernel is chosen to be K_α. Then it holds with $\mathbb{P}_0^{(n)}$-probability at least $1 - n^{-10}$ that*

$$\|\hat{f}_n^{(m)} - f_0^{(m)}\|_2 \lesssim \left(\frac{\log n}{n}\right)^{\frac{\alpha-m}{2\alpha+1}},$$

with the corresponding choice of regularization parameter $\lambda \asymp (\log n/n)^{\frac{2\alpha}{2\alpha+1}}$.

Theorem 14 yields that the plug-in KRR estimator is minimax optimal up to a logarithmic factor for estimating $f_0^{(m)}$ with $f_0 \in H^\alpha[0,1]$ or $f_0 \in S^\alpha[0,1]$ under the L_2 norm. To see this, first consider $f_0 \in H^\alpha[0,1]$ and let $\epsilon_{\alpha,m,n}$ be the minimax optimal rate for estimating $f_0^{(m)}$. Note that $n^{-\frac{\alpha-m}{2\alpha+1}}$ is the optimal rate for estimating the mth derivative of α-smooth Hölder functions (Stone, 1982). Let $H^\alpha_{per}[0,1]$ denote the subset of α-smooth Hölder space that satisfies the periodic boundary condition $f^{(j)}(0) = f^{(j)}(1)$ for $0 \leq j \leq \lfloor \alpha \rfloor - 1$. It can be shown that the corresponding L_2 minimax rate for $H^\alpha_{per}[0,1]$ is also $n^{-\frac{\alpha-m}{2\alpha+1}}$. Suppose $f \in H^{\alpha'}_{per}[0,1]$ for $\alpha' > \alpha$, according to Theorem 12.20 in Gockenbach (2005) and Theorem 3.14 in Cuddy (2012), the Fourier coefficient of f satisfies $f_i = O(i^{-\alpha'-1})$. Consequently, $H^{\alpha'}_{per}[0,1] \subset H^\alpha[0,1]$, and hence $\epsilon_{\alpha,m,n} \geq n^{-\frac{\alpha'-m}{2\alpha'+1}}$. Letting $\alpha' \downarrow \alpha$, we have $\epsilon_{\alpha,m,n} \geq n^{-\frac{\alpha-m}{2\alpha+1}}$, indicating the near minimax optimality of the plug-in KRR estimator when the function class is $H^\alpha[0,1]$. The same minimax optimality extends to the function class $S^\alpha[0,1]$. When $m = 0$, it is known that the optimal rate for estimating $f_0 \in S^\alpha[0,1]$ is $n^{-\frac{\alpha}{2\alpha+1}}$ (cf. Theorem 7.32 in Wasserman (2006)). For a general derivative order $m \geq 0$, noting that $H^\alpha[0,1] \subset S^\alpha[0,1]$, the minimax rate for $S^\alpha[0,1]$ is no faster than that for $H^\alpha[0,1]$. Since the plug-in KRR estimator achieves the same convergence rate for both classes as shown in Theorem 14, we arrive at the conclusion that the proposed KRR estimator is also nearly minimax optimal when estimating $f_0^{(m)}$ with $f_0 \in S^\alpha[0,1]$.

Remark We can see that given the smoothness level of the regression function, the rate-optimal estimation of the derivatives shares the same choice of λ between various derivative orders. Thus, the plug-in KRR estimator is *adaptive* to the order of the derivative to be estimated. This indicates that the proposed estimator enjoys the so-called *plug-in property* (Bickel and Ritov, 2003), a phenomenon in which a rate-optimal nonparametric estimator also efficiently estimates some bounded linear functionals. Instead of bounded linear functionals, we establish the plug-in property for function derivatives.

Remark The adaptivity and plug-in property of the plug-in KRR estimator are in sharp contrast to some existing methods. The minimax optimal rate for a specific derivative order can be achieved by various methods in the literature but with caveats, including difference-based methods (Wang and Lin, 2015; Dai et al., 2016; Liu and De Brabanter, 2020), local polynomial regression (Fan and Gijbels, 1996; Delecroix and Rosa, 1996), and smoothing splines (Stone, 1985; Zhou and Wolfe, 2000). For difference-based methods, the optimality only applies to interior points, and boundary correction is required for both practical implementation and theoretical understanding. Difference-based methods are typically used to estimate the first two derivatives and require more tuning parameters for a higher derivative order. For local polynomial regression and smoothing splines, as pointed out by Wahba and Wang (1990) and Charnigo et al. (2011), the optimal choice of smoothing parameter depends on the order of the derivative. Hence, they do not enjoy the aforementioned plug-in property for function derivatives while the proposed plug-in KRR method does. In other words, for local polynomial regression and smoothing splines, when the estimator achieves the optimal rate of convergence for the regression function f_0, the plug-in derivative estimators with the same tuning parameter values will be sub-optimal, and *vice versa*. The lack of adaptivity to derivative orders in these existing methods renders parameter tuning challenging in the presence of varying derivative orders.

Remark There has been a rich literature on theoretical guarantees of KRR, but much of the focus has been on regression functions (Cucker and Smale, 2002; Zhang, 2005; Caponnetto and De Vito, 2007; Steinwart et al., 2009; Mendelson and Neeman, 2010; Yang et al., 2017). For example, Yang et al. (2017) derived error bounds for $\hat{f}_n$ when the regression function belongs to similar function classes. Our focus in this section is instead on derivative estimation, which requires different assumptions and techniques. The framework developed in this article for estimating function derivatives, including the adaptivity and plug-in property of the plug-in KRR estimators, might provide useful insights and results that could be helpful in building upon existing bounds on KRR estimators for derivatives.

5. Practical consideration

Estimating the derivatives of the regression function has a wide range of applications in many areas. In the regression setting considered in this article, function derivatives have direct real-world applications. For example, estimating function derivatives is directly useful in understanding the behavior of the hypothesized dark energy equation in cosmology Holsclaw et al. (2013), which is a function of the second derivative of the data process. In ocean sciences, the derivative function provides the rate of sea-level change at a particular time point (Cahill et al., 2015), offering insights into the evolution of dynamic sea-level rise over time. In more general settings, derivatives are frequently used in spatial process models (Banerjee et al., 2003) and shape-constrained regression that utilizes derivative processes (Riihimäki and Vehtari, 2010; Wang and Berger, 2016), and may improve the computational efficiency for nonlinear dynamic system identification (Solak et al., 2003), while serving as a tool in detecting local extrema (Song et al., 2006; Li et al., 2021) and efficient modeling of functional data (Dai et al., 2018). An important implication of our developed theory and methods is that KRR estimators as a common method for unknown regression functions can be used to infer derivatives of functions by the simple plug-in strategy, with easy tuning and explicit expressions.

In practice, one needs to choose the kernel and regularization parameter λ, and account for computational complexity.

Kernel selection. There is a rich menu for the covariance kernel (Rasmussen and Williams, 2006, Chapter 4), and below we introduce several choices with polynomially decaying eigenvalues. Such kernels are commonly used in the literature (Amini and Wainwright, 2012; Zhang et al., 2015; Yang et al., 2017). The Matérn kernel is given by

$$K_{\mathrm{Mat},\nu}(x,x') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}|x-x'|\right)^{\nu} B_{\nu}\left(\sqrt{2\nu}|x-x'|\right),$$

where $B_\nu(\cdot)$ is the modified Bessel function of the second kind with smoothness parameter ν to be determined. It is well known that the eigenvalues of the Matérn kernel satisfy that $\mu_i \asymp i^{-2(\nu+1/2)}$ for $i \in \mathbb{N}$. In practice, we select ν via leave-one-out cross validation and minimize the mean square error of the regression function. The Sobolev kernel is another class of kernels with polynomial decaying eigenvalues that underlie the Sobolev spaces with different orders of smoothness (Birman and Solomyak, 1967; Gu, 2013). In our numerical experiment we consider the second-order Sobolev kernel

$$K_{\mathrm{Sob}}(x,x') = 1 + xx' + \min\{x,x'\}^2(3\max\{x,x'\} - \min\{x,x'\})/6,$$

which generates an RKHS of Lipschitz functions with smoothness $\alpha = 2$. Other higher-order Sobolev kernels also exhibit polynomial eigendecay with larger smoothness levels. Choosing the covariance kernel can be largely assisted by domain knowledge in many fields as each kernel encodes various properties of its samples from the induced RKHS. For example, the squared exponential covariance kernel as the limiting case of the Matérn kernel with $\nu \to \infty$ is a popular choice in event-related potential analysis in neuroscience (Yu et al., 2023) thanks to the induced smooth functions that agree with domain knowledge, and similarly, Matérn kernels are more popular for less smooth functions such as in spatial process models (Banerjee et al., 2003). Shape constraints also point to specific kernels; for example, inference on periodic functions necessitates choosing kernels defined on spheres that encode periodicity (Li and Ghosal, 2017). One can also resort to cross validation as a data-driven solution to choose a kernel among multiple options using a model selection perspective.

Parameter tuning. For a given kernel and under normal assumptions, we estimate error variance σ^2 by its maximum marginal likelihood estimator (MMLE)

$$\hat{\sigma}_n^2 := \lambda \boldsymbol{y}^T[K(X,X) + n\lambda \boldsymbol{I}_n]^{-1}\boldsymbol{y}$$

and choose the regularization parameter λ by maximizing the marginal likelihood

$$\boldsymbol{y} \mid X \sim N(0, \hat{\sigma}_n^2(n\lambda)^{-1}K(X,X) + \hat{\sigma}_n^2\boldsymbol{I}_n).$$

These parameters are used for *any* order of derivatives in view of the order adaptive property of the proposed method; in contrast, optimal parameter tuning in competing methods may vary with the derivative order and deviate from the one chosen for estimating the regression function. We will assess the finite-sample performance of plug-in KRR estimators with this choice of λ in the next section.

The above method for parameter tuning is also known as the empirical Bayes approach. We advocate for this approach because of its empirical success in the Bayesian literature, and an established equivalence link between Bayesian and non-Bayesian frameworks that allows us to transfer concepts from the Bayesian regime to kernel ridge regression (Liu and Li, 2020). In other settings, we have noticed that the MMLE of λ tends to adapt to the unknown smoothness level of the underlying function when paired with an oversmooth kernel. That is, the MMLE of λ often leads to excellent performance when the kernel's smoothness level is equal to or greater than that of the regression function. This suggests that an alternative effective strategy for estimating smooth functions with unknown smoothness and their derivatives could involve deploying an oversmooth kernel, such as the squared exponential kernel, and choosing λ via the MMLE. The use of oversmooth kernels, including the squared exponential kernel, is consistent with existing literature such as Bach (2013). A formal investigation of this practically appealing adaptivity feature of the MMLE is an interesting future work, and part of our efforts in this direction will be reported in Liu and Li (2022).

Computational complexity. The proposed method has analytical forms for any order of derivatives, facilitating fast implementation. The average total running time of the proposed method is 0.31 when $n = 100$ and 0.97 seconds when $n = 500$ in R on a PC with 2.3 GHz 8-Core Intel Core i9 CPU. Computing the eigendecomposition of $K(X, X)$ typically takes $O(n^3)$ times, but this is a one-time cost as we can store the eigendecomposition of $K(X, X)$ to speed up the calculation of $[K(X, X)+n\lambda\boldsymbol{I}_n]^{-1}$ for any given λ. The subsequent estimation process consists of two steps. First, we use limited-memory bound constrained "BFGS" in the "optim" function in R to find the optimal λ. This tuning step has complexity $O(kn^2)$, where k is the number of iterations, and is finished within an average of 0.28 seconds when $n = 100$ and 0.56 seconds when $n = 500$. The following step is to calculate the plug-in KRR estimate given the optimal λ, which has complexity $O(n^2)$.

6. Simulation

In this section, we assess the finite sample performance of the plug-in KRR estimator relative to several methods and provide numerical evidence of its agreement with the minimax optimal rate.

6.1 Comparison with existing methods

We consider two regression functions: $f_{01}(x) = \exp\{-4(1 - 2x)^2\}(1 - 2x)$ and $f_{02}(x) = \sin(8x) + \cos(8x) + \log(4/3 + x)$ for $x \in [0, 1]$, with random design $X_i \sim \text{Unif}[0, 1]$ and sample size $n = 500$. We generate the response $\boldsymbol{y}$ following Model (1) by adding Gaussian error $\varepsilon_i \sim N(0, 0.2^2)$ to f_{01} and f_{02}. We consider up to the third derivative to accommodate competing methods, but note that the proposed method is readily available for any order. We also conduct simulations under fixed design; the comparison is similar, and the results are deferred to the Supplementary Material.

For the proposed method, we use the second-order Sobolev kernel and Matérn kernel given in Section 5. We compare the plug-in KRR estimator with three other methods: local polynomial regression with degree $p = 2$ (R package 'locpol' in Cabrera (2018)), penalized smoothing spline (R package 'pspline' in Ripley (2017)) and locally weighted least squares regression (coded as 'LowLSR') proposed by Wang et al. (2019). For local polynomial regression, we use the Gaussian kernel and select the bandwidth via cross validation. For smoothing spline, we use cubic penalized smoothing spline with other parameters set to the default values. When implementing LowLSR, we set the number of difference quotients k

to 50 for the first derivative and increase it to 100 for the second derivative, leading to 400 and 300 non-boundary points, respectively. We remark that it is not easy for LowLSR to estimate high-order derivatives, and we only use it to estimate the first two derivatives.

We conduct a Monte Carlo study with 100 repetitions. We evaluate each estimator except LowLSR at 100 equally spaced points in $[0, 1]$, and calculate the root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{100}\sum_{t=0}^{99}\{\hat{s}(t/99) - s(t/99)\}^2},$$

where $\hat{s}$ is the estimated function and s the true function ($f_{01}^{(m)}$ or $f_{02}^{(m)}$ for $k = 1, 2, 3$). Since LowLSR does not allow evaluation at boundary points or points different from the observed X_i, we compute the RMSE at every other 5 points from the sorted X_i that are away from the boundaries, resulting in 80 and 60 testing points for first and second derivative estimation, respectively.

Figure 1 presents the boxplot of RMSEs for estimating f'_{01} and f'_{02} for each method. We can see that for f'_{02} KRR with Matérn kernel achieves the lowest median RMSE among all methods, while KRR with Sobolev kernel is comparable to penalized smoothing spline and outperforms the other two methods. For f'_{01}, we observe a similar result with the relative position switched between the Sobolev kernel and Matérn kernel. For both functions, LowLSR exhibits the highest median and the most variability of RMSE; this might be partly because LowLSR is designed for the fixed design setting. We consider a fixed design simulation in the Supplementary Material, in which LowLSR improves but still gives considerably larger RMSE than the better method of the two KRR estimators. Overall, the plug-in KRR estimator with Matén kernel leads to the best RMSE for f'_{02}, and gives close results to the leading approach KRR with Sobolev kernel for f'_{01}.

Figure 2 presents the boxplot of RMSEs for each method when estimating f''_{01} and f''_{02}. KRR with Matérn kernel achieves the lowest median RMSE among all methods.

Figure 3 displays the boxplot of RMSEs for estimating f'''_{01} and f'''_{02}. We can see that the performance of penalized smoothing spline is significantly worsened with high variability and the largest median RMSE, indicating the challenge when estimating high-order derivatives. The two KRR methods continue to give the best RMSEs, confirming our theory that the proposed estimator is adaptive to the derivative order. Comparing these two plug-in KRR estimators suggests the Matén kernel leads to either similar or better RMSE, and appears to be the recommended choice under our simulation settings.

Figure 4 displays the result from one random run in the Monte Carlo study for estimating the first derivatives. It can be seen that locpol and LowLSR do not perform well for estimating f'_{01}, while all methods produce relatively satisfactory results for estimating f'_{02}. KRR with either kernel estimates f'_{01} fairly well, while the Sobolev kernel slightly underperforms Matérn kernel when estimating the boundaries of f'_{02}.

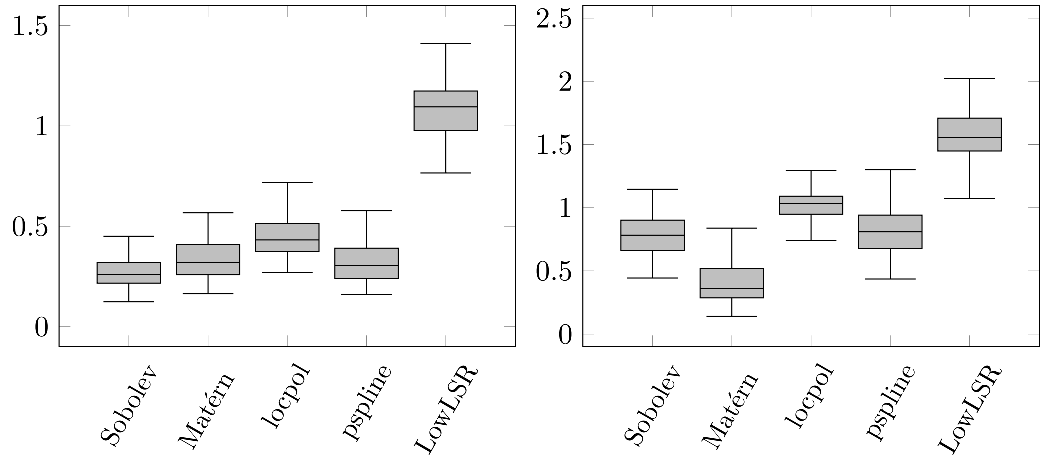

Figure 1: Boxplots of RMSEs: f'_{01} (left) and f'_{02} (right).

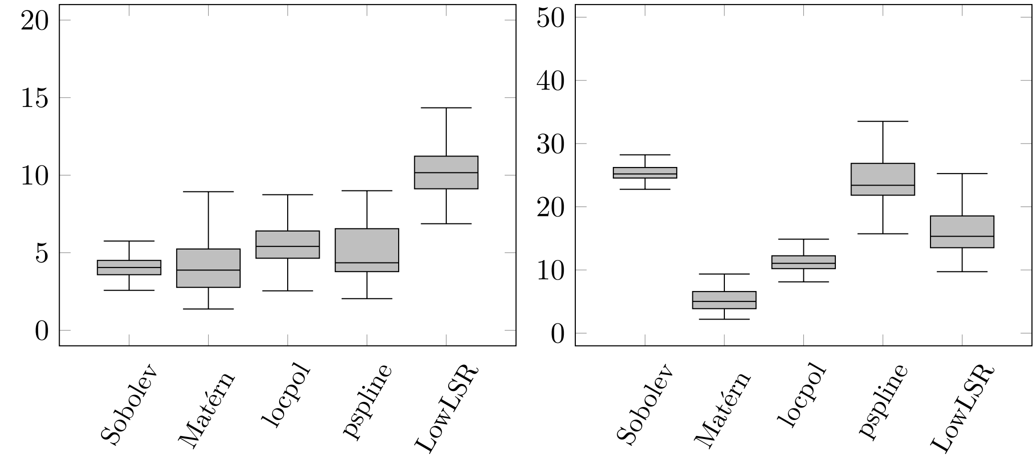

Figure 2: Boxplots of RMSEs: f''_{01} (left) and f''_{02} (right).

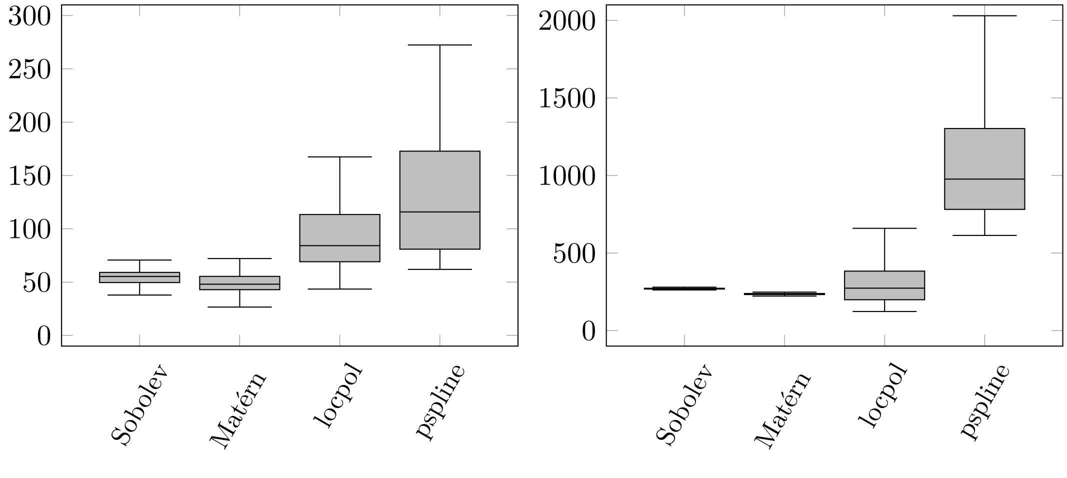

Figure 3: Boxplots of RMSEs: f'''_{01} (left) and f'''_{02} (right). LowLSR is not applicable to estimate the third derivative.

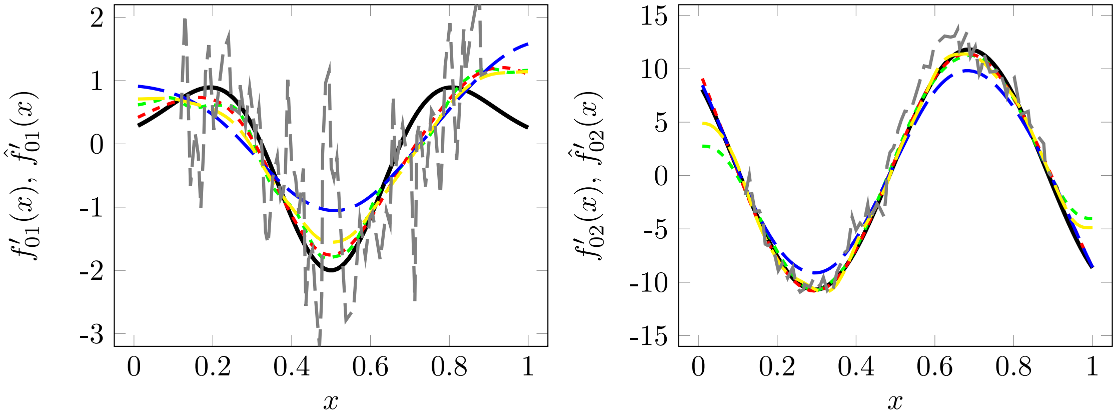

Figure 4: Estimates of f'_{01} (left) and f'_{02} (right) in one simulation: true derivative (full line); KRR with Sobolev kernel (green dash), Matérn kernel (red dash); locpol (blue long dash), spline (yellow long dash) and LowLSR (grey long dash).

6.2 Finite-sample comparison with minimax bounds

We next perform experiments as proof of concepts that the derived upper bound is observed in practice. To this end, we examine how the empirical error scales with the sample size, with λ selected by maximizing its marginal likelihood. We consider the true regression function $f_0(x) = \sqrt{2}\sum_{i=1}^{\infty} i^{-5} \sin i \cos[(i-1/2)\pi x]$ for $x \in [0,1]$, which belongs to $H^{\alpha}[0,1]$ with $\alpha = 4$. Hence, we use a Matérn kernel with $\nu = 3.5$. We simulate n_i observations from the regression model (1) with $\varepsilon_i \sim N(0, 0.1)$ and $X_i \sim \text{Unif}[0,1]$. The sample size n_i varies from 10 to 500 such that $\log(n_i)$'s are 100 equally spaced points in $[\log(10), \log(500)]$. We replicate the simulation 100 times for each sample size n_i. We then compute the average RMSE error_i of the 100 replications as an estimate of the L_2 error $\|\hat{f}'_{n_i} - f'_0\|_2$. The minimax optimal rate for estimating f'_0 is $n^{-1/3}$ under the L_2 norm (Stone, 1982). If our plug-in estimator is able to achieve this optimal rate, then the scatterplot of $(\log(n_i), \log(\text{error}_i))$ should come close to forming a straight line $\log(\text{error}_i) = -\frac{1}{3}\log(n_i) + \text{constant}$.

The left panel of Figure 5 plots $\log(\text{error}_i)$ versus $\log(n_i)$. The reference line in red has slope $-1/3$; its intercept is determined by least square fitting with fixed slope $-1/3$, which is $\sum_{i=1}^{100}\{\log(\text{error}_i) + \frac{1}{3}\log(n_i)\}/100$. We can see that the points are distributed around the line, suggesting that the estimation error of our plug-in KRR estimator agrees with the theoretical minimax rate. To investigate the effect of sample size more dynamically, the right panel of Figure 5 shows the rolling least square slopes with moving windows of 40 observations, *i.e.*, the kth slope in the plot is obtained by linear regression using data $\{(\log(n_i), \log(\text{error}_i)) : k \le i \le k+39\}$ for $1 \le k \le 61$. The slopes are close to the reference line (in red) that represents the minimax rate $-1/3$ for all the sample sizes under consideration, and we do not observe a phase transition phenomenon from these results.

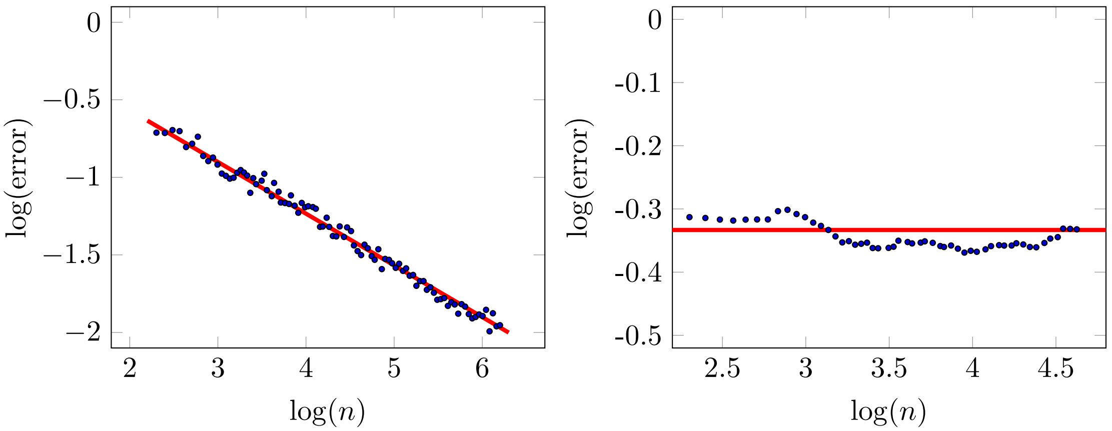

Figure 5: Log-log plots for the plug-in KRR estimator. Left panel: Scatterplot of $(\log(n_i), \log(\text{error}_i))$. Right panel: Slopes from rolling linear regression with moving windows of 40 observations. The reference lines in red are $y = -x/3+\text{constant}$ (left panel) and $y = -1/3$ (right panel), both representing the minimax rate.

We repeat the same experiment for local polynomial regression with degree $p = 2$, where the Gaussian kernel is used and the bandwidth is selected via cross validation. The results are shown in Figure 6. It can be seen that compared with the plug-in KRR estimator, local polynomial has larger errors across different sample sizes, and the error decreases at a rate slower than the optimal rate. This is not surprising as local polynomial regression lacks adaptivity to derivative orders, and we acknowledge that its performance might be improved had the tuning parameter been chosen that is better suited for the first derivative of the regression function.

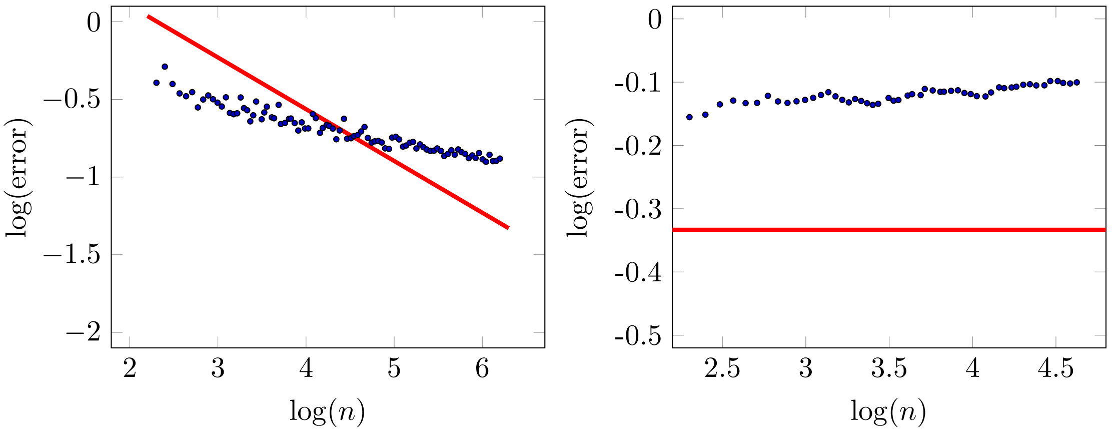

Figure 6: Log-log plots for local polynomial regression. Left panel: Scatterplot of $(\log(n_i), \log(\text{error}_i))$. Right panel: Slopes from rolling linear regression with moving windows of 40 observations. The reference lines in red are $y = -x/3+\text{constant}$ (left panel) and $y = -1/3$ (right panel), both representing the minimax rate.

7. Discussion

In this paper, we propose a plug-in kernel ridge regression estimator for estimating mixed-partial derivatives of a nonparametric regression function. The proposed estimator is analytically given and applicable for multi-dimensional support and sub-Gaussian error, enabling fast computation, broad practicability, and convenient inference. We study non-asymptotic behaviors, L_∞ convergence rates, and minimax optimality of the proposed estimator. Our analysis shows that the proposed method automatically adapts to the order of derivatives to be estimated, leading to easy tuning in practice. Simulations confirm the established minimax optimality and suggest favorable performance of the proposed estimator compared to existing methods under both random and fixed designs.

The present article is based on the commonly used iid error assumption with sub-Gaussian distributions; extension to heterogeneous or dependent error is beyond our scope but is an interesting future topic. In addition, while our theory including Theorems 2, 4, 6(b), and 7 accommodates multivariate functions and does not require the regression function f_0 to reside within the RKHS, the minimax optimality in the considered special examples is established for univariate functions and functions in the RKHS only. Future work could expand upon our results to consider multivariate and high-dimensional functions; challenges in this area include defining a practically useful function space (possibly with dimension-dependent smoothness levels) for interesting derivative estimations, as well as designing an appropriate kernel and parameter tuning methods that ensure rate optimality. Finally, we have focused on kernel ridge regression estimators in terms of plug-in properties for derivatives, and it is interesting to consider other related algorithms and loss functions, for example, spectral filtering based on the work of Lin et al. (2020) and self-concordant losses based on the work of Marteau-Ferey et al. (2019).

Acknowledgments

We thank Ding-Xuan Zhou for helpful discussions, and WenWu Wang for providing R code to implement LowLSR in the simulation section. We thank the Action Editor and two anonymous referees for constructive comments that helped to improve the paper. This research is partly supported by the grant DMS-2015569 and DMS/NIGMS-2153704 from the National Science Foundation.

Appendix A. Proofs

This section contains proofs of all results. We shall make use of the following Lemma 15 repeatedly in the sequel, which provides an error bound for $L_{K,X} - L_K$ under the $\|\cdot\|_{\mathbb{H}}$ norm. The proof of Lemma 15 mainly relies on the McDiarmid inequality and its Bernstein form, which can be found in Smale and Zhou (2005).

Lemma 15 (Lemma 3 in Smale and Zhou (2005)) *For any Mercer kernel K, bounded $f \in L^2_{p_X}(\mathcal{X})$ and $0 < \delta < 1$, with probability at least $1-\delta$, there holds*

$$\|L_{K,X}(f) - L_K(f)\|_{\mathbb{H}} = \left\| \frac{1}{n}\sum_{i=1}^n f(x_i)K_{x_i} - L_K f \right\|_{\mathbb{H}} \leq \frac{4\kappa\|f\|_\infty}{3n}\log(1/\delta) + \frac{\kappa\|f\|_2}{\sqrt{n}}(1+\sqrt{8\log(1/\delta)}).$$

A.1 Proofs in Section 3.2

Proof [Proof of Lemma 1] The proof can be found in Corollary 4.36 in Steinwart and Christmann (2008) or Theorem 4.7 in Ferreira and Menegatto (2012). ■

Proof [Proof of Theorem 2] Letting $\Delta f = \hat{f}_n - f_\lambda$, we have

$$L_{\tilde{K},X}(\Delta f) - L_{\tilde{K}}(\Delta f) = L_{\tilde{K},X}(\hat{f}_n) - L_{\tilde{K},X}(f_\lambda) - L_{\tilde{K}}(\hat{f}_n) + L_{\tilde{K}}(f_\lambda).$$

Noting that $L_{\tilde{K},X}(\boldsymbol{y} - \hat{f}_n) = \hat{f}_n - L_{\tilde{K}}(\hat{f}_n)$ and $L_{\tilde{K}}(f_0) = f_\lambda$, the preceding display becomes

$$L_{\tilde{K},X}\boldsymbol{y} - \hat{f}_n - L_{\tilde{K},X}(f_\lambda) + L_{\tilde{K}}(f_\lambda) = L_{\tilde{K},X}(\boldsymbol{y} - f_\lambda) - \Delta f - L_{\tilde{K}}(f_0 - f_\lambda).$$

Consequently,

$$\begin{aligned}\|\Delta f\|_{\tilde{\mathbb{H}}} &\leq \|L_{\tilde{K},X}(\Delta f) - L_{\tilde{K}}(\Delta f)\|_{\tilde{\mathbb{H}}} + \|L_{\tilde{K},X}(\boldsymbol{y} - f_\lambda) - L_{\tilde{K}}(f_0 - f_\lambda)\|_{\tilde{\mathbb{H}}} \\ &\leq \|L_{\tilde{K},X}(\Delta f) - L_{\tilde{K}}(\Delta f)\|_{\tilde{\mathbb{H}}} + \|L_{\tilde{K},X}(f_0 - f_\lambda) - L_{\tilde{K}}(f_0 - f_\lambda)\|_{\tilde{\mathbb{H}}} + \|L_{\tilde{K},X}\boldsymbol{w}\|_{\tilde{\mathbb{H}}},\end{aligned}$$

where $\boldsymbol{w} = \boldsymbol{y} - f_0(X)$ follows a multivariate Gaussian distribution with zero mean and variance $\sigma^2\boldsymbol{I}_n$. Let $\Omega = [\tilde{K}(X_i, X_j)]_{i,j=1}^n$, which implies that $\|L_{\tilde{K},X}\boldsymbol{w}\|^2_{\tilde{\mathbb{H}}} = n^{-2}\boldsymbol{w}^T\Omega\boldsymbol{w}$. Note that

$$\mathrm{tr}(\Omega) \leq \sum_{i=1}^n \tilde{K}(X_i, X_i) \leq n\tilde{\kappa}^2_\lambda \quad \text{and} \quad \mathrm{tr}(\Omega^2) = \sum_{i,j=1}^n \tilde{K}(X_i, X_j)^2 \leq n^2\tilde{\kappa}^4_\lambda.$$

According to the Hanson-Wright inequality (Rudelson and Vershynin, 2013), we have with probability at least $1 - 2e^{-ct^2}$ that

$$\boldsymbol{w}^T\Omega\boldsymbol{w} \leq \sigma^2\mathrm{tr}(\Omega) + 2\sigma^2\sqrt{\mathrm{tr}(\Omega^2)}(t + t^2) \leq 2\sigma^2 n\tilde{\kappa}^2_\lambda(t+1)^2,$$

for any $t > 0$ and $c > 0$ that does not depend on K or n. Therefore, with probability $1-\delta$, there holds

$$\|L_{\tilde{K},X}\boldsymbol{w}\|_{\tilde{\mathbb{H}}} \leq \frac{\sqrt{2}\tilde{\kappa}_\lambda\sigma}{\sqrt{n}}\left(1 + \sqrt{2c^{-1}\log(1/\delta)}\right).$$

Applying Lemma 15 with $\tilde{K}$ twice separately to Δf and $f_0 - f_\lambda$, with probability at least $1 - 3\delta$, we have

$$\begin{aligned}\|\Delta f\|_{\tilde{\mathbb{H}}} \leq \frac{4\tilde{\kappa}_\lambda(\|\Delta f\|_\infty + \|f_\lambda - f_0\|_\infty)}{3n}\log(1/\delta) + \frac{\tilde{\kappa}_\lambda(\|\Delta f\|_2 + \|f_\lambda - f_0\|_2)}{\sqrt{n}}\left(1 + \sqrt{8\log(1/\delta)}\right) \\ + \frac{\sqrt{2}\tilde{\kappa}_\lambda\sigma}{\sqrt{n}}\left(1 + \sqrt{2c^{-1}\log(1/\delta)}\right).\end{aligned}$$

Note that $\|f\|_2 \leq \|f\|_\infty$ for any $f \in L^2_{p_X}(\mathcal{X})$. Consider any $\delta \in (0, 1/3)$ such that $\log(1/\delta) > \log 3 > 1$. Then the upper bound in the preceding inequality becomes

$$\frac{\tilde{\kappa}_\lambda\sqrt{\log(1/\delta)}}{\sqrt{n}}\left(4 + \frac{4\tilde{\kappa}_\lambda\sqrt{\log(1/\delta)}}{3\sqrt{n}}\right)(\|\Delta f\|_\infty + \|f_\lambda - f_0\|_\infty) + \frac{C_1\tilde{\kappa}_\lambda\sigma\sqrt{\log(1/\delta)}}{\sqrt{n}},$$

where $C_1 > 0$ is a universal constant that does not depend on K or n. Therefore, with probability at least $1 - \delta$ for any $\delta \in (0, 1)$, we have

$$\|\Delta f\|_{\tilde{\mathbb{H}}} \leq \frac{\tilde{\kappa}_\lambda\sqrt{\log(3/\delta)}}{\sqrt{n}}\left(4 + \frac{4\tilde{\kappa}_\lambda\sqrt{\log(3/\delta)}}{3\sqrt{n}}\right)(\|\Delta f\|_\infty + \|f_\lambda - f_0\|_\infty) + \frac{C_1\tilde{\kappa}_\lambda\sigma\sqrt{\log(3/\delta)}}{\sqrt{n}}.$$

By Lemma 1 we obtain that with probability at least $1 - \delta$,

$$\begin{aligned}\|\Delta f\|_{\tilde{\mathbb{H}}} &\leq \frac{\tilde{\kappa}_\lambda\sqrt{\log(3/\delta)}}{\sqrt{n}}\left(4 + \frac{4\tilde{\kappa}_\lambda\sqrt{\log(3/\delta)}}{3\sqrt{n}}\right)(\tilde{\kappa}_\lambda\|\Delta f\|_{\tilde{\mathbb{H}}} + \|f_\lambda - f_0\|_\infty) + \frac{C_1\tilde{\kappa}_\lambda\sigma\sqrt{\log(3/\delta)}}{\sqrt{n}} \\ &= C(n, \tilde{\kappa}_\lambda)\|\Delta f\|_{\tilde{\mathbb{H}}} + \tilde{\kappa}_\lambda^{-1}C(n, \tilde{\kappa}_\lambda)\|f_\lambda - f_0\|_\infty + \frac{C_1\tilde{\kappa}_\lambda\sigma\sqrt{\log(3/\delta)}}{\sqrt{n}}, \qquad (9)\end{aligned}$$

where

$$C(n, \tilde{\kappa}_\lambda) = \frac{\tilde{\kappa}_\lambda^2\sqrt{\log(3/\delta)}}{\sqrt{n}}\left(4 + \frac{4\tilde{\kappa}_\lambda\sqrt{\log(3/\delta)}}{3\sqrt{n}}\right).$$

The proof is completed by applying Lemma 1 and the triangle inequality.

■

A.2 Proofs in Section 3.3

Proof [Proof of Lemma 3] This lemma is a variant of Lemma 1 but uses the $\|\cdot\|_{\mathbb{H}}$ instead of $\|\cdot\|_{\tilde{\mathbb{H}}}$ norm. The arguments used in proving Lemma 1 go verbatim. ■

Lemma 16 *For any bounded $f \in L^2_{p_X}(\mathcal{X})$, let*

$$\begin{aligned}E(K, X, f) &:= (L_{K,X} + \lambda I)^{-1}L_{K,X}f - (L_K + \lambda I)^{-1}L_K f \\ &= K(\cdot, X)[K(X, X) + n\lambda\boldsymbol{I}_n]^{-1}f(X) - (L_K + \lambda I)^{-1}L_K f, \qquad (10)\end{aligned}$$

For any $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that

$$\|E(K,X,f)\|_{\mathbb{H}} \le \frac{\kappa\|f\|_\infty\sqrt{\log(3/\delta)}}{\sqrt{n}\lambda}\left(10+\frac{4\kappa\sqrt{\log(3/\delta)}}{3\sqrt{n}\lambda}\right).$$

Proof [Proof of Lemma 16] We introduce an intermediate quantity $(L_{K,X}+\lambda I)^{-1}L_K f$ and decompose $E(K,X,f) = (\tilde{L}_{K,X}f - (L_{K,X}+\lambda I)^{-1}L_K f) + ((L_{K,X}+\lambda I)^{-1}L_K f - \tilde{L}_K f)$. We will calculate error bounds for both terms by applying Lemma 15 twice. First we have

$$\begin{aligned}
&\|\tilde{L}_{K,X}f - (L_{K,X}+\lambda I)^{-1}L_K f\|_{\mathbb{H}}\\
=\ &\|(L_{K,X}+\lambda I)^{-1}(L_{K,X}f - L_K f)\|_{\mathbb{H}}\\
\le\ &\frac{1}{\lambda}\left\|L_{K,X}f - L_K f\right\|_{\mathbb{H}},
\end{aligned}$$

where the last inequality is due to (5) in the main paper. Applying Lemma 15, then with probability at least $1-\delta$, we have

$$\|\tilde{L}_{K,X}f - (L_{K,X}+\lambda I)^{-1}L_K f\|_{\mathbb{H}} \le \frac{4\kappa\|f\|_\infty}{3n\lambda}\log(1/\delta) + \frac{\kappa\|f\|_2}{\sqrt{n}\lambda}(1+\sqrt{8\log(1/\delta)}). \tag{11}$$

On the other hand, we have

$$\begin{aligned}
&\|(L_{K,X}+\lambda I)^{-1}L_K f - \tilde{L}_K f\|_{\mathbb{H}}\\
=\ &\|(L_{K,X}+\lambda I)^{-1}(L_K+\lambda I)\tilde{L}_K f - (L_{K,X}+\lambda I)^{-1}(L_{K,X}+\lambda I)\tilde{L}_K f\|_{\mathbb{H}}\\
=\ &\|(L_{K,X}+\lambda I)^{-1}(L_K\tilde{L}_K f - L_{K,X}\tilde{L}_K f)\|_{\mathbb{H}}\\
\le\ &\frac{1}{\lambda}\|L_K\tilde{L}_K f - L_{K,X}\tilde{L}_K f\|_{\mathbb{H}}.
\end{aligned}$$

Applying Lemma 15 to $\tilde{L}_K f$ gives

$$\|(L_{K,X}+\lambda I)^{-1}L_K f - \tilde{L}_K f\|_{\mathbb{H}} \le \frac{4\kappa\|\tilde{L}_K f\|_\infty}{3n\lambda}\log(1/\delta) + \frac{\kappa\|\tilde{L}_K f\|_2}{\sqrt{n}\lambda}(1+\sqrt{8\log(1/\delta)}).$$

Letting $f=0$ in (6) gives

$$\|f_\lambda - f_0\|_2^2 + \lambda\|f_\lambda\|_{\mathbb{H}}^2 \le \|f_0\|_2^2,$$

which yields

$$\|f_\lambda\|_2 \le \sqrt{2}\|f_0\|_2 \quad \text{and} \quad \|f_\lambda\|_{\mathbb{H}} \le \lambda^{-1/2}\|f_0\|_2. \tag{12}$$

By Lemma 3 we have $\|\tilde{L}_K f\|_\infty \le \kappa\|\tilde{L}_K f\|_{\mathbb{H}}$. This together with (12) gives

$$\|(L_{K,X}+\lambda I)^{-1}L_K f - \tilde{L}_K f\|_{\mathbb{H}} \le \frac{4\kappa^2\|f\|_2/\sqrt{\lambda}}{3n\lambda}\log(1/\delta) + \frac{\sqrt{2}\kappa\|f\|_2}{\sqrt{n}\lambda}(1+\sqrt{8\log(1/\delta)}). \tag{13}$$

Again consider any $\delta \in (0,1/3)$ such that $\log(1/\delta) > \log 3 > 1$. Then, the two bounds in equations (11) and (13) become

$$\frac{4\kappa\|f\|_\infty}{3n\lambda}\log(1/\delta) + \frac{4\kappa\|f\|_\infty}{\sqrt{n}\lambda}\sqrt{\log(1/\delta)}, \quad \frac{4\kappa^2\|f\|_\infty/\sqrt{\lambda}}{3n\lambda}\log(1/\delta) + \frac{6\kappa\|f\|_\infty}{\sqrt{n}\lambda}\sqrt{\log(1/\delta)},$$

respectively. Consequently, with probability at least $1-2\delta > 1-3\delta$, we have

$$\begin{aligned}\|E(K,X,f)\|_{\mathbb{H}} = \|\tilde{L}_{K,X}f - \tilde{L}_K f\|_{\mathbb{H}} &\leq \frac{\kappa\|f\|_\infty}{\sqrt{n}\lambda}\left(10\sqrt{\log(1/\delta)} + \frac{4}{3\sqrt{n}}\log(1/\delta) + \frac{4\kappa}{3\sqrt{n}\lambda}\log(1/\delta)\right) \\ &\leq \frac{\kappa\|f\|_\infty\sqrt{\log(1/\delta)}}{\sqrt{n}\lambda}\left(10 + \frac{4\kappa\sqrt{\log(1/\delta)}}{3\sqrt{n\lambda}}\right).\end{aligned}$$

Therefore, with probability at least $1-\delta$ for any $\delta \in (0,1)$, we have

$$\|E(K,X,f)\|_{\mathbb{H}} \leq \frac{\kappa\|f\|_\infty\sqrt{\log(3/\delta)}}{\sqrt{n}\lambda}\left(10 + \frac{4\kappa\sqrt{\log(3/\delta)}}{3\sqrt{n\lambda}}\right).$$

■

Proof [Proof of Theorem 4] Substituting $f = f_0$ into $E(K,X,f)$ defined in (10) yields $E(K,X,f_0) = f_{X,\lambda} - f_\lambda$. By Lemma 16, we have with probability at least $1-\delta$ that

$$\|f_{X,\lambda} - f_\lambda\|_{\mathbb{H}} \leq \frac{\kappa\|f_0\|_\infty\sqrt{\log(3/\delta)}}{\sqrt{n}\lambda}\left(10 + \frac{4\kappa\sqrt{\log(3/\delta)}}{3\sqrt{n\lambda}}\right). \tag{14}$$

Note that

$$\hat{f}_n - f_{X,\lambda} = K(\cdot,X)[K(X,X) + n\lambda\boldsymbol{I}_n]^{-1}\boldsymbol{w} = K(\cdot,X)[K(X,X)/n + \lambda\boldsymbol{I}_n]^{-1}\boldsymbol{w}/n,$$

where $\boldsymbol{w} = \boldsymbol{y} - f_0(X)$ follows a multivariate Gaussian distribution with zero mean and variance $\sigma^2\boldsymbol{I}_n$. Thus,

$$\begin{aligned}\|\hat{f}_n - f_{X,\lambda}\|_{\mathbb{H}}^2 &= \frac{1}{n^2}\boldsymbol{w}^T[K(X,X)/n + \lambda\boldsymbol{I}_n]^{-1}K(X,X)[K(X,X)/n + \lambda\boldsymbol{I}_n]^{-1}\boldsymbol{w} \\ &\leq \frac{1}{n^2}\kappa^2\boldsymbol{w}^T\Sigma\boldsymbol{w},\end{aligned}$$

where $\Sigma = [K(X,X)/n + \lambda\boldsymbol{I}_n]^{-2}$. Since $K(X,X)/n$ is non-negative definite, all eigenvalues of $K(X,X)/n + \lambda\boldsymbol{I}_n$ are bounded below by λ, which leads to

$$\mathrm{tr}(\Sigma) \leq n\lambda^{-2} \quad \text{and} \quad \mathrm{tr}(\Sigma^2) \leq n^2\lambda^{-4}.$$

According to the Hanson-Wright inequality (Rudelson and Vershynin, 2013), we have with probability at least $1-2e^{-ct^2}$ that

$$\boldsymbol{w}^T\Sigma\boldsymbol{w} \leq \sigma^2\mathrm{tr}(\Sigma) + 2\sigma^2\sqrt{\mathrm{tr}(\Sigma^2)}(t+t^2) \leq 2\sigma^2 n\lambda^{-2}(t+1)^2,$$

for any $t > 0$. Therefore, with probability $1-\delta$, there holds

$$\|\hat{f}_n - f_{X,\lambda}\|_{\mathbb{H}} \leq \frac{\sqrt{2}\kappa\sigma}{\sqrt{n}\lambda}\left(1 + \sqrt{2c^{-1}\log(1/\delta)}\right) \leq \frac{C_2\kappa\sigma\sqrt{\log(1/\delta)}}{\sqrt{n}\lambda}, \tag{15}$$

where we consider any $\delta \in (0, 1/3)$ such that $\log(1/\delta) > \log 3 > 1$ and $C_2 > 0$ is a universal constant that does not depend on K or n. Combining (14) and (15), it holds that with probability at least $1 - 2\delta > 1 - 3\delta$,

$$\|\hat{f}_n - f_\lambda\|_{\mathbb{H}} \leq \frac{\kappa\|f_0\|_\infty\sqrt{\log(3/\delta)}}{\sqrt{n}\lambda}\left(10 + \frac{4\kappa\sqrt{\log(3/\delta)}}{3\sqrt{n\lambda}}\right) + \frac{C_2\kappa\sigma\sqrt{\log(1/\delta)}}{\sqrt{n}\lambda}.$$

Hence, for any $\delta \in (0, 1)$, it holds with probability at least $1 - \delta$ that

$$\|\hat{f}_n - f_\lambda\|_{\mathbb{H}} \leq \frac{\kappa\|f_0\|_\infty\sqrt{\log(9/\delta)}}{\sqrt{n}\lambda}\left(10 + \frac{4\kappa\sqrt{\log(9/\delta)}}{3\sqrt{n\lambda}}\right) + \frac{C_2\kappa\sigma\sqrt{\log(3/\delta)}}{\sqrt{n}\lambda}.$$

The proof is completed by applying Lemma 3 and the triangle inequality. ■

Proof [Proof of Corollary 5] We first simplify $\tilde{\mathbb{H}}$-bound. With $\delta = n^{-10}$, we have

$$C(n, \tilde{\kappa}_\lambda) = \frac{\tilde{\kappa}_\lambda^2\sqrt{10\log(3n)}}{\sqrt{n}}\left(4 + \frac{4\tilde{\kappa}_\lambda\sqrt{10\log(3n)}}{3\sqrt{n}}\right).$$

The condition $\tilde{\kappa}_\lambda^2 = o(\sqrt{n/\log n})$ yields that $C(n, \tilde{\kappa}_\lambda) = o(1)$ and further $\tilde{\kappa}_{\boldsymbol{\beta},\lambda}\tilde{\kappa}_\lambda^{-1}C(n, \tilde{\kappa}_\lambda) \lesssim \tilde{\kappa}_{\boldsymbol{\beta},\lambda}\tilde{\kappa}_\lambda\sqrt{\log n/n}$. Noting that $\|f_\lambda - f_0\|_\infty = o(1)$, the second term in $\tilde{\mathbb{H}}$-bound is bounded by the third term. Hence, $\tilde{\mathbb{H}}$-bound becomes

$$\|\partial^{\boldsymbol{\beta}} f_\lambda - \partial^{\boldsymbol{\beta}} f_0\|_\infty + \frac{C_1'\tilde{\kappa}_{\boldsymbol{\beta},\lambda}\tilde{\kappa}_\lambda\sigma\sqrt{10\log(3n)}}{\sqrt{n}}.$$

With $\delta = n^{-10}$, $\mathbb{H}$-bound becomes

$$\|\partial^{\boldsymbol{\beta}} f_\lambda - \partial^{\boldsymbol{\beta}} f_0\|_\infty + \frac{\kappa_{\boldsymbol{\beta}}\kappa\|f_0\|_\infty\sqrt{10\log(9n)}}{\sqrt{n}\lambda}\left(10 + \frac{4\kappa\sqrt{10\log(9n)}}{3\sqrt{n\lambda}}\right) + \frac{C_2\kappa_{\boldsymbol{\beta}}\kappa\sigma\sqrt{10\log(3n)}}{\sqrt{n}\lambda}.$$

Comparing the preceding display with $\mathbb{H}$-bound, we can see that if $\tilde{\kappa}_{\boldsymbol{\beta},\lambda}\tilde{\kappa}_\lambda = o(\lambda^{-1})$, $\tilde{\mathbb{H}}$-bound is asymptotically less than $\mathbb{H}$-bound.

■

Proof [Proof of Theorem 6] We first prove (a). Rewrite f_0 as $f_0 = L_K^r g$ for some $g = L_K^{-r} f_0 \in L^2_{p_X}(\mathcal{X})$ and thus $f_i = \mu_i^r g_i$. Representing the function g by $g = \sum_{i=1}^\infty g_i\psi_i$, we have

$$f_\lambda - f_0 = -\sum_{i=1}^\infty \frac{\lambda}{\mu_i + \lambda}\mu_i^r g_i\psi_i.$$

When $\frac{1}{2} < r \leq 1$, we have

$$\begin{aligned}\|f_\lambda - f_0\|_{\mathbb{H}}^2 &= \sum_{i=1}^\infty \left(\frac{\lambda}{\mu_i + \lambda}\mu_i^r g_i\right)^2 / \mu_i \\ &= \lambda^{2r-1}\sum_{i=1}^\infty \left(\frac{\lambda}{\mu_i + \lambda}\right)^{3-2r}\left(\frac{\mu_i}{\mu_i + \lambda}\right)^{2r-1} g_i^2 \\ &\leq \lambda^{2r-1}\|L_K^{-r} f_0\|_2^2.\end{aligned}$$

The proof is completed by applying Lemma 3.

For (b), we have $\partial^{\boldsymbol{\beta}} f_\lambda - \partial^{\boldsymbol{\beta}} f_0 = -\sum_{i=1}^\infty \frac{\lambda}{\mu_i+\lambda}\mu_i^r g_i \partial^{\boldsymbol{\beta}}\psi_i$, where $\{\psi_i\}_{i=1}^\infty$ is the Fourier basis, *i.e.*, $\psi_1(\boldsymbol{x}) = 1, \psi_{2i}(\boldsymbol{x}) = \cos(2\pi \boldsymbol{I}_i \cdot \boldsymbol{x}), \psi_{2i+1} = \sin(2\pi \boldsymbol{I}_i \cdot \boldsymbol{x})$; here $\boldsymbol{I}_i \in \mathbb{N}_0^d$ are ordered multi-indexes. It follows that

$$\begin{aligned}\|\partial^{\boldsymbol{\beta}} f_\lambda - \partial^{\boldsymbol{\beta}} f_0\|_\infty &\le \sum_{i=1}^\infty \frac{\lambda}{\mu_i+\lambda}\mu_i^r |g_i||\partial^{\boldsymbol{\beta}}\psi_i| \\ &= \lambda^r \sum_{i=1}^\infty \left(\frac{\lambda}{\mu_i+\lambda}\right)^{1-r}\left(\frac{\mu_i}{\mu_i+\lambda}\right)^r |g_i||\partial^{\boldsymbol{\beta}}\psi_i| \le \lambda^r \sum_{i=1}^\infty |g_i||\partial^{\boldsymbol{\beta}}\psi_i|.\end{aligned}$$

Since $g \in C^p(\mathcal{X})$, the Fourier coefficients satisfy $|g_i| \lesssim \binom{i+d}{d-1} i^{-p} \lesssim i^{d-p-1}$. Moreover, $|\partial^{\boldsymbol{\beta}}\psi_i| \lesssim i(i-1)\cdots(i-|\boldsymbol{\beta}|+1) \lesssim i^{|\boldsymbol{\beta}|}$. Therefore,

$$\|\partial^{\boldsymbol{\beta}} f_\lambda - \partial^{\boldsymbol{\beta}} f_0\|_\infty \le C_3 \lambda^r \sum_{i=1}^\infty i^{d-p-1+|\boldsymbol{\beta}|} = C_3 \lambda^r \zeta(p-d+1+|\boldsymbol{\beta}|).$$

■

Proof [Proof of Theorem 7] We write $f_0 = L_K^r g$ for some $g = L_K^{-r} f_0 \in L^2_{p_X}(\mathcal{X})$ and thus $f_i = \mu_i^r g_i$. Representing the function g by $g = \sum_{i=1}^\infty g_i \psi_i$, then we have

$$f_\lambda - f_0 = -\sum_{i=1}^\infty \frac{\lambda}{\mu_i+\lambda}\mu_i^r g_i \psi_i.$$

We have $\partial^{\boldsymbol{\beta}} f_\lambda - \partial^{\boldsymbol{\beta}} f_0 = -\sum_{i=1}^\infty \frac{\lambda}{\mu_i+\lambda}\mu_i^r g_i \partial^{\boldsymbol{\beta}}\psi_i$, where $\{\psi_i\}_{i=1}^\infty$ is the Fourier basis. Define $\{g_i^*\}_{i=1}^\infty$ such that $g_i \partial^{\boldsymbol{\beta}}\psi_i = g_i^* i^{\boldsymbol{\beta}}\psi_i$ and let $g^* = \sum_{i=1}^\infty g_i^* \psi_i \in L^2_{p_X}(\mathcal{X})$.

According to Lemma 10 in Fischer and Steinwart (2020), Assumption C implies that the eigenvalues decay with a polynomial upper bound of order $1/q$, *i.e.*, there exists a constant $C_4 > 0$ such that for all $i \in \mathbb{N}$,

$$\mu_i \le C_4 i^{-1/q},$$

which implies that $i^{|\boldsymbol{\beta}|} \le C_4^{q|\boldsymbol{\beta}|}\mu_i^{-q|\boldsymbol{\beta}|} = C_4 \mu_i^{-q|\boldsymbol{\beta}|}$. Thus,

$$\begin{aligned}\|\partial^{\boldsymbol{\beta}} f_\lambda - \partial^{\boldsymbol{\beta}} f_0\|_\infty &\le \left\|\sum_{i=1}^\infty \frac{\lambda}{\mu_i+\lambda}\mu_i^r g_i \partial^{\boldsymbol{\beta}}\psi_i\right\|_\infty \\ &\le \lambda \sup_{i\ge 1} \frac{\mu_i^{r-q/2} i^{|\boldsymbol{\beta}|}}{\mu_i+\lambda} \left\|\sum_{i=1}^\infty \mu_i^{q/2} g_i^* \psi_i\right\|_\infty \\ &\le C_4 \lambda \sup_{i\ge 1} \frac{\mu_i^{r-q/2-q|\boldsymbol{\beta}|}}{\mu_i+\lambda} \left\|L_K^{q/2} g^*\right\|_\infty \\ &\le A C_4 \lambda^{r-q/2-q|\boldsymbol{\beta}|}\|g^*\|_2,\end{aligned}$$

where the last inequality follows from Lemma 25 in Fischer and Steinwart (2020) and Assumption C. ■

A.3 Proofs in Section 4

Proof [Proof of Lemma 11] When $\alpha > m + 1/2$, we have

$$\partial^{m,m} K_\alpha(x, x') = \sum_{i=1}^{\infty} \mu_i \psi_i^{(m)}(x)\psi_i^{(m)}(x') \lesssim \sum_{i=1}^{\infty} i^{-2\alpha} i^{2m} < \infty.$$

Thus, $K_\alpha \in C^{2m}([0,1] \times [0,1])$.

Recall the definition of $\tilde{\kappa}^2_{m,\lambda}$ in (7). It follows that for any $m \in \mathbb{N}_0$,

$$\begin{aligned}\tilde{\kappa}^2_{\alpha,m,\lambda} &= \sup_{x\in[0,1]} \sum_{i=1}^{\infty} \frac{\mu_i}{\lambda + \mu_i} \psi_i^{(m)}(x)^2 \\ &\lesssim \sum_{i=1}^{\infty} \frac{i^{2m}}{1 + \lambda i^{2\alpha}} \le \int_0^\infty \frac{(x+1)^{2m} dx}{1 + \lambda x^{2\alpha}} \asymp \lambda^{-\frac{2m+1}{2\alpha}},\end{aligned}$$

where the last step holds for $\alpha > m + \frac{1}{2}$. On the other hand, we have

$$\begin{aligned}\tilde{\kappa}^2_{\alpha,m,\lambda} &\gtrsim \sum_{i=1}^{\infty} \left[\frac{(2i)^{2m}}{1 + \lambda(2i)^{2\alpha}} \cos(2\pi i x)^2 + \frac{(2i+1)^{2m}}{1 + \lambda(2i+1)^{2\alpha}} \sin(2\pi i x)^2 \right] \\ &\ge \sum_{i=1}^{\infty} \min\left\{ \frac{(2i)^{2m}}{1 + \lambda(2i)^{2\alpha}}, \frac{(2i+1)^{2m}}{1 + \lambda(2i+1)^{2\alpha}} \right\} \\ &\ge \frac{1}{2} \sum_{i=1}^{\infty} \frac{i^{2m}}{1 + \lambda i^{2\alpha}} \asymp \lambda^{-\frac{2m+1}{2\alpha}},\end{aligned}$$

where we also need $\alpha > m + \frac{1}{2}$. The differentiability of $\tilde{K}_\alpha$ directly follows from the boundedness of $\tilde{\kappa}^2_{\alpha,m,\lambda}$ for any fixed λ. ■

Proof [Proof of Lemma 12] In view of Lemma 1 and Lemma 11, we can see that $f \in C^m[0,1]$ for any $f \in \tilde{\mathbb{H}}_\alpha$. This is also true for $f \in \mathbb{H}_\alpha$ since $\mathbb{H}_\alpha$ and $\tilde{\mathbb{H}}_\alpha$ contain the same functions.

Now we prove the norm inequality. Let $f = \sum_{i=1}^{\infty} f_i \psi_i$, where $\{\psi_i\}_{i=1}^{\infty}$ is the Fourier basis. Then, $\|f^{(m)}\|_2^2 \asymp \sum_{i=1}^{\infty} (f_i i^m)^2$ for any $m \in \mathbb{N}_0$. It is equivalent to showing that

$$\tilde{\kappa}^2_{\alpha,\lambda} \cdot \sum_{i=1}^{\infty} f_i^2 i^{2m} \le C_m \tilde{\kappa}^2_{\alpha,m,\lambda} \cdot \sum_{i=1}^{\infty} f_i^2 \frac{\lambda + \mu_i}{\mu_i},$$

for some $C_m > 0$. Hence, it suffices to show that for any $i \in \mathbb{N}$,

$$\tilde{\kappa}^2_{\alpha,\lambda} \cdot f_i^2 i^{2m} \le C_m \tilde{\kappa}^2_{\alpha,m,\lambda} \cdot f_i^2 \frac{\lambda + \mu_i}{\mu_i}.$$

In view of Lemma 11, we have $\tilde{\kappa}^2_{\alpha,m,\lambda} \asymp \lambda^{-\frac{2m+1}{2\alpha}}$, which also leads to $\tilde{\kappa}^2_{\alpha,\lambda} \asymp \lambda^{-\frac{1}{2\alpha}}$ when taking $m = 0$. Since $\mu_i \asymp i^{-2\alpha}$, it is sufficient to show that

$$\lambda^{\frac{m}{\alpha}} i^{2m} \le C_m (1 + \lambda i^{2\alpha}),$$

for some constant $C_m > 0$. The above equation trivially holds for $C_m = 1$ if $\lambda^{\frac{m}{\alpha}} i^{2m} \le 1$. If $\lambda^{\frac{m}{\alpha}} i^{2m} \ge 1$, then $\lambda^{\frac{m}{\alpha}} i^{2m} \le (\lambda^{\frac{m}{\alpha}} i^{2m})^{\frac{\alpha}{m}} = \lambda i^{2\alpha}$ since $m < \alpha$. Taking $C_m = 1$ completes the proof. ■

Proof [Proof of Lemma 13] Let $f_0 = \sum_{i=1}^{\infty} f_i \psi_i$. Then,

$$f_\lambda - f_0 = -\sum_{i=1}^{\infty} \frac{\lambda}{\lambda + \mu_i} f_i \psi_i.$$

Note that

$$\begin{aligned}\|f_\lambda - f_0\|_{\tilde{\mathbb{H}}_\alpha}^2 &= \sum_{i=1}^{\infty} \left(\frac{\lambda}{\lambda + \mu_i} f_i \right)^2 \Big/ \frac{\mu_i}{\lambda + \mu_i} = \lambda \sum_{i=1}^{\infty} \frac{\lambda}{\lambda + \mu_i} \frac{f_i^2}{\mu_i} \\ &\lesssim \lambda \sum_{i=1}^{\infty} i^{2\alpha} f_i^2 \le \lambda \left(\sum_{i=1}^{\infty} i^{\alpha} |f_i| \right)^2 \lesssim \lambda.\end{aligned}$$

Therefore, for any $f_0 \in H^\alpha[0,1]$ or $f_0 \in S^\alpha[0,1]$, we have $\|f_\lambda - f_0\|_{\tilde{\mathbb{H}}_\alpha} \lesssim \lambda^{\frac{1}{2}}$. ■

Proof [Proof of Theorem 14] Applying Lemma 12 to (9) and taking $\delta = n^{-10}$ yields with $\mathbb{P}_0^{(n)}$-probability at least $1 - n^{-10}$ that

$$\begin{aligned}\|\hat{f}_n^{(m)} - f_\lambda^{(m)}\|_2 &\le \tilde{\kappa}_{\alpha,\lambda}^{-1} \tilde{\kappa}_{\alpha,m,\lambda} \|\hat{f}_n - f_\lambda\|_{\tilde{\mathbb{H}}_\alpha} \\ &\le \frac{\tilde{\kappa}_{\alpha,m,\lambda} \tilde{\kappa}_{\alpha,\lambda}^{-2} C(n, \tilde{\kappa}_{\alpha,\lambda})}{1 - C(n, \tilde{\kappa}_{\alpha,\lambda})} \|f_\lambda - f_0\|_\infty + \frac{1}{1 - C(n, \tilde{\kappa}_{\alpha,\lambda})} \frac{C_1 \tilde{\kappa}_{\alpha,m,\lambda} \sigma \sqrt{10 \log(3n)}}{\sqrt{n}}.\end{aligned}$$

By choosing λ such that $\tilde{\kappa}_{\alpha,\lambda}^2 = o(\sqrt{n/\log n})$, we have $\tilde{\kappa}_{\alpha,\lambda}^{-2} C(n, \tilde{\kappa}_{\alpha,\lambda}) \asymp \sqrt{\log n / n}$ and $C(n, \tilde{\kappa}_{\alpha,\lambda}) \le 1/2$ for sufficiently large n. We arrive at

$$\|\hat{f}_n^{(m)} - f_\lambda^{(m)}\|_2 \lesssim \frac{2\tilde{\kappa}_{\alpha,m,\lambda} \sqrt{\log n}}{\sqrt{n}} \|f_\lambda - f_0\|_\infty + \frac{2C_1 \tilde{\kappa}_{\alpha,m,\lambda} \sigma \sqrt{10 \log(3n)}}{\sqrt{n}} \lesssim \tilde{\kappa}_{\alpha,m,\lambda} \sqrt{\frac{\log n}{n}},$$

given that $\|f_\lambda - f_0\|_\infty = o(1)$. Hence,

$$\|\hat{f}_n^{(m)} - f_0^{(m)}\|_2 \lesssim \|f_\lambda^{(m)} - f_0^{(m)}\|_2 + \tilde{\kappa}_{\alpha,m,\lambda} \sqrt{\frac{\log n}{n}}. \tag{16}$$

In view of Lemma 1 and Lemma 13, we can see that for any $f_0 \in H^\alpha[0,1]$ or $f_0 \in S^\alpha[0,1]$, $\|f_\lambda - f_0\|_\infty \le \tilde{\kappa}_{\alpha,\lambda} \|f_\lambda - f_0\|_{\tilde{\mathbb{H}}_\alpha} \lesssim \lambda^{\frac{1}{2} - \frac{1}{4\alpha}} = o(1)$. Invoking Lemma 12, we have

$$\|f_\lambda^{(m)} - f_0^{(m)}\|_2 \lesssim \tilde{\kappa}_{\alpha,\lambda}^{-1} \tilde{\kappa}_{\alpha,m,\lambda} \|f_\lambda - f_0\|_{\tilde{\mathbb{H}}_\alpha} \lesssim \lambda^{\frac{1}{4\alpha}} \lambda^{-\frac{2m+1}{4\alpha}} \lambda^{\frac{1}{2}} = \lambda^{\frac{1}{2} - \frac{m}{2\alpha}}.$$

It follows from (16) that

$$\|\hat{f}_n^{(m)} - f_0^{(m)}\|_2 \lesssim \lambda^{\frac{1}{2} - \frac{m}{2\alpha}} + \lambda^{-\frac{2m+1}{4\alpha}} \sqrt{\frac{\log n}{n}}.$$

The upper bound in the preceding display is minimized when $\lambda \asymp (\log n/n)^{\frac{2\alpha}{2\alpha+1}}$, which satisfies $\tilde{\kappa}^2_{\alpha,\lambda} \asymp (n/\log n)^{\frac{1}{2\alpha+1}} = o(\sqrt{n/\log n})$. The optimal rate is derived by substituting λ. This completes the proof. ■

Appendix B. Additional simulation results under fixed design

We conduct a Monte Carlo study in the fixed design setting. We consider the same functions and sample size $n = 500$ as in Section 6.1 of the main paper and choose fixed design points $X_i = i/500$ for $i = 1, \ldots, 500$. We run 100 repetitions and evaluate each estimator according to RMSE as in Section 6.1.

Figure 7–9 display the boxplots of RMSEs for estimating up to the third derivative in the fixed design setting. It can be seen that LowLSR performs slightly better than in the random design setting, especially when estimating the derivatives of f_{02}. Still, the two KRR methods compare favorably to the benchmarks, and the Matérn kernel gives the best median RMSEs for almost all cases, with one exception in the left panel of Figure 7 where it ties with the proposed estimator with Sobolev kernel. This is consistent with our observations made in the random design setting.

Figure 10 shows the result from one random run in our Monte Carlo study for estimating the first derivatives in the fixed design setting. We observe similar trends as in the random design setting in Figure 4. For example, the estimation of LowLSR appears undesirably wiggly when estimating f'_{01}. These results suggest that the proposed plug-in KRR estimator continues to work well in the fixed design setting.

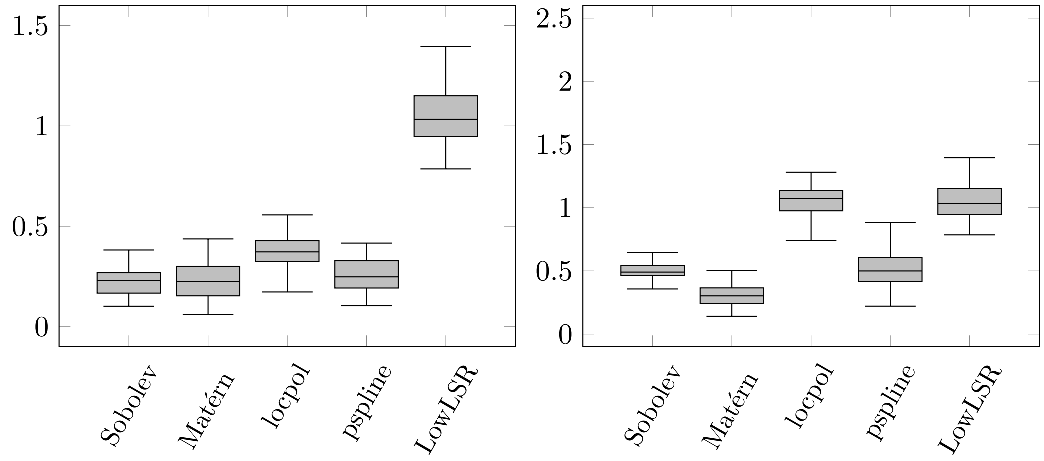

Figure 7: Boxplots of RMSEs for estimating f'_{01} (left) and f'_{02} (right) in fixed design setting.

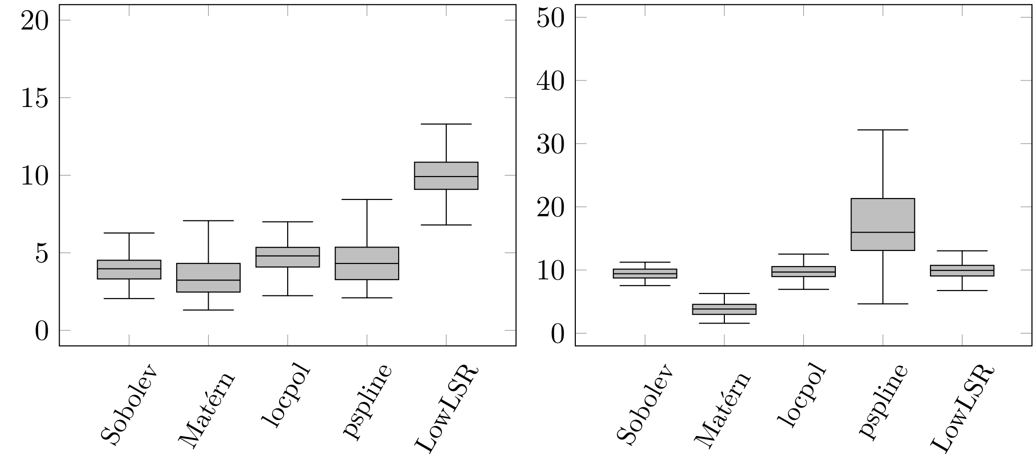

Figure 8: Boxplots of RMSEs for estimating f''_{01} (left) and f''_{02} (right) in fixed design setting.

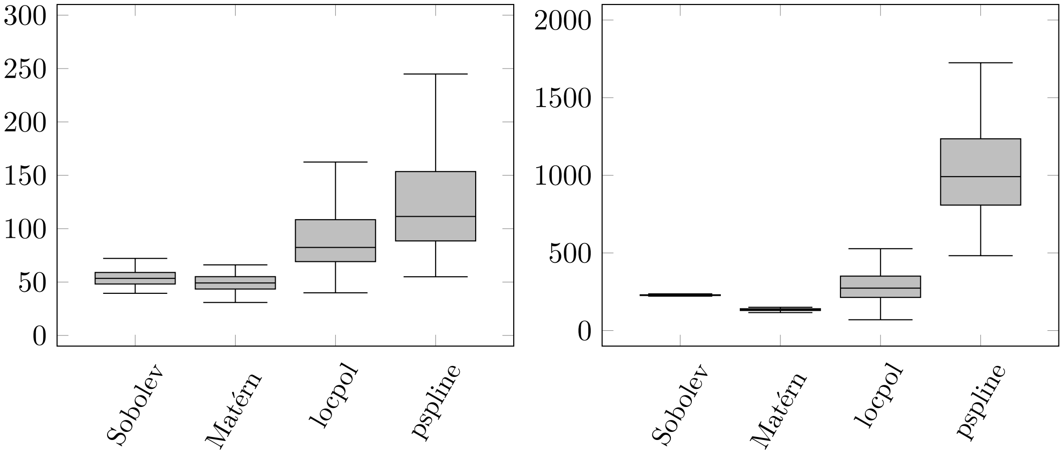

Figure 9: Boxplots of RMSEs for estimating f'''_{01} (left) and f'''_{02} (right) in fixed design setting. LowLSR is not applicable to estimate the third derivative.

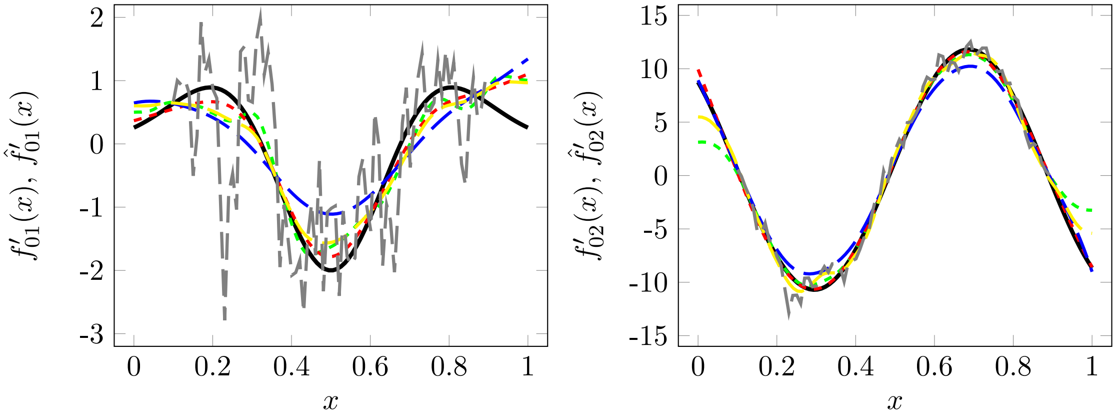

Figure 10: One random run in the Monte Carlo study for estimating f'_{01} (left) and f'_{02} (right) under fixed design: true derivative (full line), KRR with Sobolev kernel (green dashed line), Matérn kernel (red dashed line), locpol (blue long dash), spline (yellow long dash) and LowLSR (grey long dash).

References

Arash A Amini and Martin J Wainwright. Sampled forms of functional PCA in reproducing kernel Hilbert spaces. *The Annals of Statistics*, 40(5):2483–2510, 2012.

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

Sudipto Banerjee, Alan E Gelfand, and CF Sirmans. Directional rates of change under spatial process models. *Journal of the American Statistical Association*, 98(464):946–954, 2003.

Peter J Bickel and Ya'acov Ritov. Nonparametric estimators which can be "plugged-in". *The Annals of Statistics*, 31(4):1033–1053, 2003.

Mikhail Shlemovich Birman and Mikhail Zakharovich Solomyak. Piecewise-polynomial approximations of functions of the classes W_p^α. *Matematicheskii Sbornik*, 115(3):331–355, 1967.

Vivien Cabannes, Loucas Pillaud-Vivien, Francis Bach, and Alessandro Rudi. Overcoming the curse of dimensionality with Laplacian regularization in semi-supervised learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Jorge Luis Ojeda Cabrera. *locpol: Kernel local polynomial regression*, 2018. URL `https://CRAN.R-project.org/package=locpol`. R package version 0.7-0.

Niamh Cahill, Andrew C Kemp, Benjamin P Horton, and Andrew C Parnell. Modeling sea-level change using errors-in-variables integrated Gaussian processes. *The Annals of Applied Statistics*, 9(2):547–571, 2015.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Richard Charnigo, Benjamin Hall, and Cidambi Srinivasan. A generalized C_p criterion for derivative estimation. *Technometrics*, 53(3):238–253, 2011.

Jie Chen, Lingfei Wu, Kartik Audhkhasi, Brian Kingsbury, and Bhuvana Ramabhadrari. Efficient one-vs-one kernel ridge regression for speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2454–2458, 2016.

Yu Cheng, Zhigang Jin, Tao Gao, Hongcai Chen, and Nikola Kasabov. An improved collaborative representation based classification with regularized least square (CRC–RLS) method for robust face recognition. *Neurocomputing*, 215:250–259, 2016.

Felipe Cucker and Steve Smale. Best choices for regularization parameters in learning theory: On the bias-variance problem. *Foundations of Computational Mathematics*, 2(4): 413–428, 2002.

Felipe Cucker and Ding-Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*, volume 24. Cambridge University Press, 2007.

Kevin Stephen Stotter Cuddy. Convergence of Fourier series. 2012. URL `http://math.uchicago.edu/~may/REU2012/REUPapers/Cuddy.pdf`.

Wenlin Dai, Tiejun Tong, and Marc G Genton. Optimal estimation of derivatives in nonparametric regression. *Journal of Machine Learning Research*, 17(1):5700–5724, 2016.

Xiongtao Dai, Hans-Georg Müller, and Wenwen Tao. Derivative principal component analysis for representing the time dynamics of longitudinal and functional data. *Statistica Sinica*, 28(3):1583–1609, 2018.

Kris De Brabanter, Jos De Brabanter, Bart De Moor, and Irene Gijbels. Derivative estimation with local polynomial fitting. *Journal of Machine Learning Research*, 14(1):281–301, 2013.

M Delecroix and AC Rosa. Nonparametric estimation of a regression function and its derivatives under an ergodic hypothesis. *Journal of Nonparametric Statistics*, 6(4):367–382, 1996.

Peter Exterkate, Patrick JF Groenen, Christiaan Heij, and Dick van Dijk. Nonlinear forecasting with many predictors using kernel ridge regression. *International Journal of Forecasting*, 32(3):736–753, 2016.

Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*, volume 66. CRC Press, 1996.

José C. Ferreira and Valdir A. Menegatto. Reproducing properties of differentiable Mercer-like kernels. *Mathematische Nachrichten*, 285(8-9):959–973, 2012. ISSN 0025584X. doi: 10.1002/mana.201100072.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(205):1–38, 2020.

Mark S Gockenbach. *Partial Differential Equations: Analytical and Numerical Methods*, volume 122. Siam, 2005.

Chong Gu. *Smoothing Spline ANOVA Models*, volume 297. Springer Science & Business Media, 2013.

Zheng-Chu Guo and Ding-Xuan Zhou. Concentration estimates for learning with unbounded sampling. *Advances in Computational Mathematics*, 38(1):207–223, 2013.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Science & Business Media, 2006.

Tracy Holsclaw, Bruno Sansó, Herbert KH Lee, Katrin Heitmann, Salman Habib, David Higdon, and Ujjaini Alam. Gaussian process modeling of derivative curves. *Technometrics*, 55(1):57–67, 2013.

Wolfgang Härdle. *Applied Nonparametric Regression*. Number 19 in Econometric Society Monographs. Cambridge University Press, 1990.

Meng Li and Subhashis Ghosal. Bayesian detection of image boundaries. *The Annals of Statistics*, 45(5):2190–2217, 2017.

Meng Li, Zejian Liu, Cheng-Han Yu, and Marina Vannucci. Semiparametric Bayesian inference for local extrema of functions in the presence of noise. *arXiv preprint arXiv:2103.10606*, 2021.

Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020.

Yu Liu and Kris De Brabanter. Derivative estimation in random design. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3449–3458, 2018.

Yu Liu and Kris De Brabanter. Smoothed nonparametric derivative estimation using weighted difference quotients. *Journal of Machine Learning Research*, 21(65):1–45, 2020.

Zejian Liu and Meng Li. Equivalence of convergence rates of posterior distributions and Bayes estimators for functions and nonparametric functionals. *arXiv preprint arXiv:2011.13967*, 2020.

Zejian Liu and Meng Li. Optimal plug-in Gaussian processes for modelling derivatives. *arXiv preprint arXiv:2210.11626*, 2022.

Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In *Conference on Learning Theory*, pages 2294–2340. PMLR, 2019.

Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565, 2010.

P Mohapatra, Sreejit Chakravarty, and PK Dash. Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system. *Swarm and Evolutionary Computation*, 28:144–160, 2016.

Hans-Georg Müller, Ulrich Stadtmüller, and Thomas Schmitt. Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika*, 74(4):743–749, 1987.

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

Jaakko Riihimäki and Aki Vehtari. Gaussian processes with monotonicity information. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 645–652, 2010.

Brian Ripley. *pspline: Penalized smoothing splines*, 2017. URL `https://CRAN.R-project.org/package=pspline`. R package version 1.0-18.

Lorenzo Rosasco, Silvia Villa, Sofia Mosci, Matteo Santoro, and Alessandro Verri. Nonparametric sparsity and regularization. *Journal of Machine Learning Research*, 14(1): 1665–1714, 2013.

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

Steve Smale and Ding-Xuan Zhou. Shannon sampling II: Connections to learning theory. *Applied and Computational Harmonic Analysis*, 19(3):285–302, 2005.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

Ercan Solak, Roderick Murray-Smith, William E Leithead, Douglas J Leith, and Carl E Rasmussen. Derivative observations in Gaussian process models of dynamic systems. In *Advances in Neural Information Processing Systems*, pages 1057–1064, 2003.

Peter Sollich and Christopher Williams. Using the equivalent kernel to understand Gaussian process regression. In *Advances in Neural Information Processing Systems*, pages 1313–1320, 2005.

Peter X-K Song, Xin Gao, Rui Liu, and Wen Le. Nonparametric inference for local extrema with application to oligonucleotide microarray data in yeast genome. *Biometrics*, 62(2): 545–554, 2006.

Michael L Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media, 1999.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

Ingo Steinwart, Don R Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 79–93, 2009.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

Charles J Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705, 1985.

Grace Wahba. *Spline Models for Observational Data*, volume 59. Siam, 1990.

Grace Wahba and Yonghua Wang. When is the optimal regularization parameter insensitive to the choice of the loss function? *Communications in Statistics-Theory and Methods*, 19 (5):1685–1700, 1990.

Cheng Wang and Ding-Xuan Zhou. Optimal learning rates for least squares regularized regression with unbounded sampling. *Journal of Complexity*, 27(1):55–67, 2011.

WenWu Wang and Lu Lin. Derivative estimation based on difference sequence via locally weighted least squares regression. *Journal of Machine Learning Research*, 16(1):2617–2641, 2015.

WenWu Wang, Ping Yu, Lu Lin, and Tiejun Tong. Robust estimation of derivatives using locally weighted least absolute deviation regression. *Journal of Machine Learning Research*, 20(1):2157–2205, 2019.

Xiaojing Wang and James O Berger. Estimating shape constrained functions using Gaussian processes. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1–25, 2016.

Larry Wasserman. *All of Nonparametric Statistics.* Springer Science & Business Media, 2006.

Yun Yang, Anirban Bhattacharya, and Debdeep Pati. Frequentist coverage and sup-norm convergence rate in Gaussian process regression. *arXiv preprint arXiv:1708.04753*, 2017.

Yannis G Yatracos. On the estimation of the derivatives of a function with the derivatives of an estimate. *Journal of Multivariate Analysis*, 28(1):172–175, 1989.

Yannis G Yatracos. Plug-in L_2-upper error bounds in deconvolution, for a mixing density estimate in R^d and for its derivatives, via the L_1-error for the mixture. *Statistics*, 53(6): 1251–1268, 2019.

Cheng-Han Yu, Meng Li, Colin Noe, Simon Fischer-Baum, and Marina Vannucci. Bayesian inference for stationary points in Gaussian process regression models for event-related potentials analysis. *Biometrics*, 79(2):629–641, 2023.

Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005.

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16(1):3299–3340, 2015.

Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3): 739–767, 2002.

Shanggang Zhou and Douglas A Wolfe. On derivative estimation in spline regression. *Statistica Sinica*, 10:93–108, 2000.