Statistica Sinica Preprint No: SS-2024-0256	
Title	Asymptotic Theory for Linear Functionals of Kernel Ridge Regression
Manuscript ID	SS-2024-0256
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0256
Complete List of Authors	Rui Tuo and Lu Zou
Corresponding Authors	Rui Tuo
E-mails	ruituo@tamu.edu
Notice: Accepted author version.	

ASYMPTOTIC THEORY FOR LINEAR FUNCTIONALS OF KERNEL RIDGE REGRESSION

Rui Tuo and Lu Zou

Texas A&M University,

Shenzhen Polytechnic University

Abstract: An asymptotic theory is established for linear functionals of the predictive function given by kernel ridge regression, when the reproducing kernel Hilbert space is equivalent to a Sobolev space. The theory covers a wide variety of linear functionals, including point evaluations, evaluation of derivatives, L_2 inner products, etc. We establish the upper and lower bounds of the estimates and their asymptotic normality. We show the asymptotic normality of these estimators under mild conditions, which enables uncertainty quantification of a wide range of frequently used plug-in estimators. The theory also implies that the minimax L_∞ error of kernel ridge regression can be attained under $\lambda \sim n^{-1} \log n$.

Key words and phrases: kernel ridge regression, uncertainty quantification, plug-in estimators.

1. Introduction

Consider a nonparametric regression model

$$y_i = f(x_i) + e_i \tag{1.1}$$

with e_i's being independent and identically distributed random errors with mean zero and a finite variance σ^2. Here x_i's can be deterministic or random inputs independent of e_i's. Nonparametric regression aims to estimate f from data $(x_i, y_i), i = 1, \ldots, n$.

Kernel ridge regression (KRR) is defined as

$$\hat{f} := \underset{v \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - v(x_i))^2 + \lambda \|v\|_{\mathcal{H}}^2, \tag{1.2}$$

given data $(x_i, y_i)_{i=1}^n$, where $\mathcal{H}$ is the reproducing kernel Hilbert space generated by a kernel function K, and $\lambda > 0$ is called the smoothing parameter. We use the notation $\|\cdot\|_{\mathcal{H}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ to denote the norm and the inner product of $\mathcal{H}$, respectively. It is well known that $\hat{f}$ is a good estimator for f under mild conditions.

In many real-world problems, the quantity of interest is a linear functional of f, denoted by $l(f)$, such as an evaluation or a derivative of f at a pre-specified point, or an integral of f. Sometimes, the quantity of interest is nonlinear in f by itself, but is closely related to a linear functional. For instance, the maximizer of f is the zero point of the gradient function of f. Plug-in estimators are widely used in practice, that is, to estimate $l(f)$ by $l(\hat{f})$. This work aims at providing theoretical justification and a framework of uncertainty quantification for these plug-in estimators.

1.1 Problem of Interest and Overview of Our Results

In this work, we consider the asymptotic properties of a linear functional of $\hat{f} - f$ defined as general as

$$l(\hat{f} - f) := \langle \hat{f} - f, g \rangle_{\mathcal{H}}, \tag{1.3}$$

for some $g \in \mathcal{H}$. This includes many examples of practical interest, e.g., L_2 inner products $\int_\Omega (\hat{f} - f)(x)h(x)dx = \left\langle \hat{f} - f, \int_\Omega K(\cdot, x)h(x)dx \right\rangle_{\mathcal{H}}$, point evaluations $(\hat{f} - f)(x) = \left\langle \hat{f} - f, K(\cdot, x) \right\rangle_{\mathcal{H}}$, point evaluations of derivatives $\frac{\partial}{\partial x_i}(\hat{f} - f)(x) = \left\langle \hat{f} - f, \frac{\partial}{\partial x_i} K(\cdot, x) \right\rangle_{\mathcal{H}}$.

As we shall study theoretical properties as $n \to \infty$, the input and output data, the minimizer $\hat{f}$, and the tuning parameter λ should all naturally be dependent on n. In addition, unless otherwise specified, the true function f can depend on n as well. While keeping this fact in mind, we shall omit the subscript n for the sake of notational convenience throughout this article. Below is a summary of our major contributions.

1. We develop a new method to investigate the asymptotic properties of a single linear functional of the form $\langle \hat{f}, g \rangle_{\mathcal{H}}$ to answer the following questions: 1) How large is the bias and variance of $\langle \hat{f}, g \rangle_{\mathcal{H}}$ as an estimator of $\langle f, g \rangle_{\mathcal{H}}$; 2) What is an appropriate rate of λ to facilitate the estimation of $\langle f, g \rangle_{\mathcal{H}}$; and 3) Is $\langle \hat{f}, g \rangle_{\mathcal{H}}$ asymptotically normal?

   While our theory depicts a more general picture, we give Table 1 to highlight a few cases of particular practical interests. It can be seen that our theory gives the exact rate of convergence and the central limit theorem for these statistics under a wide range of λ. It also shows that $\lambda \sim n^{-1}$ balances the variance and the worst-case bias regardless of the specific linear functional.

Functional	Upper & lower rates		Range of λ	Central limit theorem
	Variance	Worst-case bias		
Point evaluation	$n^{-1}\lambda^{-\frac{d}{2m}}$	$\lambda^{\frac{1}{2}-\frac{d}{4m}}$	$\lambda = O(1)$ $\lambda^{-1} = O(n^{\frac{2m}{d}})$	Valid if $\lambda = o(1)$ and $\lambda^{-1} = o(n^{\frac{2m}{d}})$
Derivative evaluation	$n^{-1}\lambda^{-\frac{d+2\vert\alpha\vert}{2m}}$	$\lambda^{\frac{1}{2}-\frac{d+2\vert\alpha\vert}{4m}}$		
L_2 inner product	n^{-1}	No more than $\lambda^{\frac{1}{2}}$		

Table 1: Summary of asymptotic properties of linear functionals of practical interest, where d = input dimension, m = smoothness, $|\alpha|$ = total order of derivatives. Exact upper and lower rates of convergence are given, except for the worst-case bias for the L_2 inner product. Discussions regarding this matter is made in Section **??** of Supplementary Material.

2. Our asymptotic theory for linear functionals can be employed to find upper and lower bounds for uniform errors as well. In this work, we examine the global error of the KRR regression as well as the derivatives, in terms of $\sup_{x\in\Omega}|D^\alpha \hat{f}(x) - D^\alpha f(x)|$. An exact rate of convergence is given when the noise is normally distributed. We show that with $\lambda \sim n^{-1}\log n$, the resulting rate of convergence is $(n^{-1}\log n)^{\frac{1}{2}-\frac{d+2|\alpha|}{4m}}$, matching the known minimax rate in Nemirovski (2000). This result implies that λ reaches the L_∞-minimax rate differs from the one that reaches the L_2-minimax rate.

3. Our theory can be leveraged to cover some non-linear functionals that can be linearized asymptotically, such as $\max_{x\in\Omega} f(x)$.

The remainder of this article is organized as follows. We review the related work in Section 1.2. In Section 2, we introduce the bias-variance decomposition of the problem. The main results of our theory are presented in Section 3, in terms of the general theory of the upper and lower bounds and asymptotic normality. In Section 4, we present several examples to illustrate the scope of the proposed framework. In Section 5, we employ our theory to obtain some uniform error bounds for KRR and investigate a nonlinear problem to further demonstrate the applicability of our theory. Numerical studies and an analysis of real-world data are presented in Section 6. The Supplementary Materials provide a more in-depth review of the literature, other related results, detailed discussions of a key assumption, and all technical proofs.

1.2 Related Work

KRR was initially introduced in the context of spline models (Wahba, 1978) and support vector machines (Boser et al., 1992), due to its innate capacity to accommodate complex patterns and nonlinear relationships.

Error bounds for KRR. The minimax convergence rates for KRR in L_2 are well established in the existing literature; see, e.g., Caponnetto and De Vito (2007); Smale and Zhou (2007); Steinwart et al. (2009); Mendelson and Neeman (2010), among many others. Although there has been rich literature on the theoretical guarantees of KRR, theory on functionals of KRR estimators is scarce. The closely related work is Liu and Li (2023), which offers a non-asymptotic analysis of the plug-in KRR estimator for its partial mixed derivatives. This paper develops a general theory on rates of convergence and statistical inference covering a diverse set of linear functionals, which includes derivatives considered in Liu and Li (2023). Another series of work related to this paper delves into linear functional regression (Cai and Yuan, 2012; Yuan and Cai, 2010). Nevertheless, this literature often assumes the linear functional as the L_2 inner product of the input data with a slope function, and primarily focuses on the asymptotic properties of the slope function. Some linear functionals in terms of the L_2 inner product fall into the semiparametric regime, see Kosorok (2008); van de Geer (2000). Our theory also extends these results by weakening the requirements for the smoothness of the function in the L_2 inner product.

Statistical inference for KRR. Another approach uses KRR for statistical inference, often investigating Gaussian approximation for KRR and its variants. Starting with Huang (2003), which established pointwise asymptotic normality for the polynomial B-spline estimator, several works have studied constructing uniform confidence bands assuming the objective function lies in an RKHS; see (Shang and Cheng, 2013; Cheng and Shang, 2013; Zhao et al., 2021). The uniform asymptotic inference results in this literature rely on expressing the KRR estimator through an orthonormal basis. Our result yields pointwise asymptotic normality for KRR under weaker conditions. Furthermore, we demonstrate that many other linear functionals of KRR also exhibit asymptotic normality under both fixed and random designs. The existing literature on statistical inference for KRR has mainly focused on regression functions. The relevant work in this area is Liu et al. (2023), which introduced a plug-in KRR estimator to estimate derivatives of a smoothing spline ANOVA model and provided convergence rates and asymptotic normality. Their estimation and inference theorem relies on the tensor structure and the equivalent kernel technique (Messer and Goldstein, 1993; Silverman, 1984). However, this method cannot be directly applied to non-tensor product structures like the Matérn kernels. Instead, we do not assume a tensor product structure and our analysis also covers derivatives of more general orders. A more detailed discussion of related literature is deferred to the Supplementary Material.

2. Bias and Variance

For simplicity, we introduce the following notation. For any $A = (a_1, \ldots, a_m)^T$ and $B = (b_1, \ldots, b_l)^T$, denote $K(A, B) = (K(a_i, b_j))_{ij}$. Denote $X = (x_1, \ldots, x_n)^T$ and $Y = (y_1, \ldots, y_n)^T$. Then the representer's theorem (Schölkopf et al., 2001; Wahba, 1990) provides an explicit expression of $\hat{f}$ in (1.2) as $\hat{f}(x) = K(x, X)(K(X, X) + \lambda nI)^{-1}Y$. Thus, we have $\langle \hat{f}, g\rangle_{\mathcal{H}} = g^T(X)(K(X, X) + \lambda nI)^{-1}Y$, where $g^T(X) = (g(x_1), \ldots, g(x_n))$. Now split $Y = F + E =: (f(x_1), \ldots, f(x_n))^T + (e_1, \ldots, e_n)^T$. Then

$$\langle \hat{f}, g\rangle_{\mathcal{H}} = g^T(X)(K(X, X) + \lambda nI)^{-1}F + g^T(X)(K(X, X) + \lambda nI)^{-1}E.$$

Let $\mathbb{E}_E$ and Var_E be the expectation and variance operators with respect to E, respectively. Note that X is independent of E, if X is random at all. Taking expectation or variance with respect to E will leave X as is. We call the quantity in (2.4) the *bias*, denoted as BIAS:

$$\mathrm{BIAS} := \mathbb{E}_E\langle \hat{f} - f, g\rangle_{\mathcal{H}} = g^T(X)(K(X, X) + \lambda nI)^{-1}F - \langle f, g\rangle_{\mathcal{H}}. \tag{2.4}$$

We call (2.5) the *variance term*.

$$\langle \hat{f} - \mathbb{E}_E \hat{f}, g \rangle_{\mathcal{H}} = g^T(X)(K(X,X) + \lambda n I)^{-1} E. \tag{2.5}$$

The term (2.6) is called the *variance*, denoted as VAR:

$$\mathrm{VAR} := \mathrm{Var}_E \langle \hat{f} - f, g \rangle_{\mathcal{H}} = \sigma^2 g^T(X)(K(X,X) + \lambda n I)^{-2} g(X). \tag{2.6}$$

A primary objective of this study is to quantify BIAS and VAR as the sample size tends to infinity. It is important to note that, unlike VAR, BIAS is dependent on the underlying true function f. Sometimes, we want to emphasize this dependency by denoting the bias as BIAS_f, when the interest lies in understanding the lower bounds of the *worst case bias* over the RKHS unit ball, defined as $\sup_{\|f\|_{\mathcal{H}} \leq 1} |\,\mathrm{BIAS}_f\,|$. To analyze the bias and variance, this work introduces an innovative tool called *noiseless kernel ridge regression*, which is detailed in Section **??** of the Supplementary Materials.

3. Main Results

In this section, we will present three types of major theoretical results: the upper bounds in Section 3.2, the lower bounds in Section 3.3, and the asymptotic normality results in Section 3.5. First, we introduce a set of assumptions in Section 3.1.

3.1 Assumptions

While the proposed techniques can be applied in other settings, in this work, we only consider the situations when $\mathcal{H}$ is equivalent to a (fractional) Sobolev space (see Section **??** of the Supplementary Materials), leading to Assumption 1.

Assumption 1. The input domain Ω is a convex and compact subset of $\mathbb{R}^d$ with a non-empty interior. In addition, $\mathcal{H}$ is equal to a (fractional) Sobolev space with order m (satisfying $m > d/2$), denoted by H^m, with equivalent norms.

The condition $m > d/2$ is to ensure that H^m is embedded into the space of continuous functions, according to the Sobolev embedding theorem. This embedding is necessary because otherwise, the point evaluation $f(x)$ is mathematically not well-defined. The spaces $\mathcal{H}$ and H^m are equivalent if K is an isotropic Matérn kernel with smoothness $\nu = m - d/2$, under the regularity conditions for Ω in Assumption 1; see (Wendland, 2004).

Now we formally introduce the smoothness requirement of g. The intuition behind Assumption 2 is that g has to be smoother than the baseline smoothness of $\mathcal{H}$. More discussion is deferred to Sections **??**-**??** in the Supplementary Material.

Assumption 2. There exist constants $C_g > 0$ and $\delta \in (0, 1]$, such that for each $v \in \mathcal{H}$,

$$|\langle g, v\rangle_{\mathcal{H}}| \leq C_g \|v\|_{L_2}^{\delta} \|v\|_{\mathcal{H}}^{1-\delta}. \tag{3.7}$$

Note that (3.7) is always true if $\delta = 0$, by plugging in $C_g = \|g\|_{\mathcal{H}}$, which imposes no extra conditions. This is why we need $\delta > 0$. As $\|\cdot\|_{\mathcal{H}}$ is stronger than $\|\cdot\|_{L_2}$, a larger δ fulfilling Assumption 2 can imply that Assumption 2 is also true for a smaller δ. As we will see later, the larger δ is, we can expect the more improvements in the rates of convergence. In Section 4, we will give the corresponding δ value for each of the aforementioned linear functionals.

We also need regularity conditions for the input sites. In this work, the design points can be either random or fixed, provided that Assumption 3 holds.

Assumption 3. If X is random, X is independent of E. Besides, there exists $C_1 > 0$, and for each $\epsilon > 0$, there exists $C_\epsilon > 0$, both independent of n and X, such that $\mathbb{P}(\Xi_\epsilon) \geq 1 - \epsilon$, where Ξ_ϵ denotes the event

$$\|v\|_{L_2} \leq \max\left\{C_1\|v\|_n, C_\epsilon n^{-m/d}\|v\|_{\mathcal{H}}\right\}, \tag{3.8}$$

$$\|v\|_n \leq \max\left\{C_1\|v\|_{L_2}, C_\epsilon n^{-m/d}\|v\|_{\mathcal{H}}\right\}. \tag{3.9}$$

for all $v \in \mathcal{H}$.

In Section **??** of the Supplementary Material, we give some sufficient conditions for Assumption 3. Specifically, Assumption 3 holds for 1) *random designs* whose points are independent and identically distributed samples from a probability density bounded away from zero and infinity, and 2) *fixed designs* that are quasi-uniform.

It is worth noting that in Assumption 3, the probability is taken with regard to the randomness of X, and in case X is deterministic, the norm inequalities (3.8) and (3.9) should hold unconditionally. To obtain the improved rates and the upper bounds, condition

(3.8) alone suffices. The lower bounds and the asymptotic normality will also need condition (3.9).

Connecting the $\|\cdot\|_n$ and the $\|\cdot\|_{L_2}$ norms is crucial in the theory of a variety of nonparametric regression methods; see Huang (2003); van de Geer (2000) for example. In Assumption 3, the event Ξ_ϵ serves as a set of high probability such that $\|\cdot\|_n$ and $\|\cdot\|_{L_2}$ are comparable. Lemma 3.1 shows a simple but important consequence of Assumption 3.

Lemma 3.1. *With Assumption 3 and the conditions $\sigma^2 \neq 0$ and $g \neq 0$, we have* $\mathrm{VAR} \neq 0$ *with probability tending to one, as $n \to \infty$.*

3.2 Upper Bounds

We shall use the following notation for asymptotic orders. For (possibly random) sequences $a_n, b_n > 0$, we denote $a_n \lesssim b_n$ if a_n/b_n is bounded in probability; denote $a_n \gtrsim b_n$ if $b_n \lesssim a_n$; and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$.

Theorem 3.1. *Suppose $\lambda \gtrsim n^{-2m/d}$. Under Assumptions 1-3, we have*

$$|\,\mathrm{BIAS}\,| = O_{\mathbb{P}}(\lambda^{\frac{\delta}{2}}\|f\|_{\mathcal{H}}), \tag{3.10}$$

$$\mathrm{VAR} = O_{\mathbb{P}}(\sigma^2 n^{-1}\lambda^{\delta-1}). \tag{3.11}$$

3.3 Lower Bounds

It is not surprising that VAR should have a lower bound, in view of the classic statistical theory such as the Cramér-Rao lower bound. Here we would like to pursue a lower bound as close as possible to the upper bound in Theorem 3.1.

Note that the upper bounds of the rate of convergence depend on the best δ value that ensures Assumption 2. Intuitively, a lower bound should rely on a δ value that disallows for (3.7) in Assumption 2. To elaborate on the condition to be introduced, we first present an equivalent statement of Assumption 2. For notational simplicity, we use the convention $\frac{0}{0} = 0$ throughout this article.

Proposition 3.1. *Under Assumption 1, given $g \in \mathcal{H}$ and $\delta \in (0, 1]$, $\sup_{v \in \mathcal{H}} \frac{|\langle g, v\rangle_{\mathcal{H}}|}{\|v\|_{L_2}^{\delta} \|v\|_{\mathcal{H}}^{1-\delta}}$ is finite if and only if for each $R > 0$,*

$$\sup_{\|v\|_{\mathcal{H}} \leq R\|v\|_{L_2}} \frac{|\langle g, v\rangle_{\mathcal{H}}|}{\|v\|_{L_2}} \leq CR^{1-\delta}, \tag{3.12}$$

for some constant $C > 0$ independent of R.

Our lower bounds rely on the reversed direction of the inequality (3.12), showing in Assumption 4.

Assumption 4. For some $\tau \in (0, 1]$, there exist constants $C_0 > 0$ and $R_0 > 0$ such that $\sup_{\|v\|_{\mathcal{H}} \leq R\|v\|_{L_2}} \frac{\langle g, v\rangle_{\mathcal{H}}}{\|v\|_{L_2}} > C_0 R^{1-\tau}$, for each $R \geq R_0$.

It is worth noting that Assumption 4 implies that $g \neq 0$. In view of Proposition 3.1, if Assumptions 2 and 4 are both true, we clearly have $\delta \leq \tau$. As opposed to Assumption 2, a smaller τ fulfilling Assumption 4 can imply that Assumption 4 is also true for a larger τ. The case $\tau = 1$ is trivially true provided that $g \neq 0$, for $R_0 = \|g\|_{\mathcal{H}}/\|g\|_{L_2}$ and $C_0 = \|g\|_{\mathcal{H}}^2/\|g\|_{L_2}$.

It is not hard to imagine that τ plays an important role in characterizing our lower bound of the rate of convergence in Theorem 3.2.

Theorem 3.2. *Suppose Assumptions 1-4 hold. Then for each $\epsilon > 0$, there exist constants $A_1, A_2, A_3 > 0$ depending only on $C_0, C_1, C_g, C_\epsilon, R_0, \delta$, and τ, such that, on the event Ξ_ϵ introduced in Assumption 3, for any n and λ satisfying $A_1 n^{-2m/d} \leq \lambda \leq A_2$, we have* $\mathrm{VAR} \geq A_3\sigma^2 n^{-1}\lambda^{\frac{\delta(\tau-1)}{\tau}}$.

The trivial case $\tau = 1$ leads to a "parametric-rate" lower bound $\mathrm{VAR} \gtrsim \sigma^2 n^{-1}$, which is not surprising. Besides, it is particularly interesting when $\delta = \tau$, as the lower rate in Theorem 3.2 coincides with the upper rate in Theorem 3.1. This leads to Theorem 3.3. We will show in Section 4 that $\delta = \tau$ is indeed true for many examples of practical interest.

Theorem 3.3. *Suppose $g \in \mathcal{H}$ satisfies Assumptions 2 and 4 with $\delta = \tau$. Besides, Assumptions 1 and 3 hold. Then for each $\epsilon > 0$, there exist constants $A_1, A_2, A_3, A_4 > 0$ depending only on $C_0, C_1, C_g, C_\epsilon, R_0, \delta$, and τ, such that, on the event Ξ_ϵ introduced in Assumption 3, for any n and λ satisfying $A_1 n^{-2m/d} \leq \lambda \leq A_2$, we have $A_3\sigma^2 n^{-1}\lambda^{\delta-1} \leq \mathrm{VAR} \leq A_4\sigma^2 n^{-1}\lambda^{\delta-1}$.*

Now we consider the bias term. First, we note that the bias depends on the underlying true function f. If $f \equiv 0$, we can clearly see $\mathrm{BIAS} = 0$. A more meaningful study of the lower bounds for bias is to consider the *worst-case bias*. To define a worst-case bias, we imagine the application of KRR to a family of models having the form of equation (1.1), but with different f. Nevertheless, the same g and parameter λ are used for each model. For each f, denote the corresponding bias by BIAS_f. we Theorem 3.4 provides a lower bound for the worst-case bias over the unit ball of $\mathcal{H}$.

Theorem 3.4. *Suppose Assumptions 1-4 hold. Then for each $\epsilon > 0$, there exist constants $A_1, A_2, A_3 > 0$ depending only on $C_0, C_1, C_g, C_\epsilon, R_0, \delta$, and τ, such that, on the event Ξ_ϵ introduced in Assumption 3, for any n and λ satisfying $A_1 n^{-2m/d} \leq \lambda \leq A_2$, we have*

$$\sup_{\|f\|_{\mathcal{H}} \leq 1} |\operatorname{BIAS}_f| \geq \begin{cases} A_3 \lambda^{\frac{2\tau - 2\delta + \delta^2 - \delta^2 \tau}{2\tau(1-\delta)}} & \text{if } \delta < 1 \\ A_3 \lambda & \text{if } \delta = 1 \end{cases}; \tag{3.13}$$

and in particular, if $\delta = \tau < 1$,

$$A_3 \lambda^{\frac{\delta}{2}} \leq \sup_{\|f\|_{\mathcal{H}} \leq 1} |\operatorname{BIAS}_f| \leq A_4 \lambda^{\frac{\delta}{2}}, \tag{3.14}$$

for some A_4 depending only on $C_0, C_1, C_g, C_\epsilon, R_0$, and δ.

Remark 1. There is a sharp transition in the lower bounds (3.13) between the case $\delta < 1$ and $\delta = 1$, showing completely different rates of convergence. Despite the weird appearance, this gap in the rate of convergence is genuine! When $\delta = 1$, there exists a *semiparametric effect* that may significantly boost the rate of convergence of the bias so that $\sup_{\|f\|_{\mathcal{H}} \leq 1} |\operatorname{BIAS}_f|$ can become much smaller than the lower bound suggested in (3.14). It is implied in the literature concerning the semiparametric properties of KRR (e.g., Mammen and van de Geer (1997); Tuo and Wu (2015); van de Geer (2000)) that there exist cases with $\delta = 1$, such that $\operatorname{BIAS} = o(n^{-1/2})$ whenever $n^{-1} \lesssim \lambda = o(n^{-1/2})$, which definitely violates (3.14). The semiparametric effect improves the bias rate of convergence through a mechanism different from what we have discussed. Further investigations in Section **??** of the Supplementary Materials also show that the lower bound (**??**) for $\delta = 1$ cannot be improved in general.

3.4 Discussion on the choice of λ

In view of Theorems 3.1, 3.3 and 3.4 we may choose $\lambda \asymp n^{-1}$ to balance the worst-case bias and the variance when $\delta = \tau < 1$. For $\delta = 1$, the variance becomes $O(n^{-1})$, the parametric rate, regardless of the choice of λ. From Theorem 3.1, a suitable choice of λ in this case would be $n^{-2m/d} \lesssim \lambda \lesssim n^{-1}$. Note that this differs from $\lambda \asymp n^{-\frac{2m}{2m+d}}$, the optimal order of magnitude of λ for $\|\hat{f} - f\|_{L_2}$ to reach the minimax rate of convergence (Stone, 1980). Of course, we would also expect that the actual $|\operatorname{BIAS}_f|$ for a specific f can be much smaller than the worst-case bias.

Theorem 3.5 shows that BIAS decays faster than the rate indicated by Theorem 3.1 for fixed f.

Theorem 3.5. *If f is fixed across all n and $\lambda = o(1)$, under the conditions of Theorem 3.1,* $|\operatorname{BIAS}| = o_{\mathbb{P}}(\lambda^{\delta/2})$.

More explicit improved rates for BIAS are given in Section **??** of the Supplementary Materials under extra smoothness conditions of f. In view of these results, when $\lambda \asymp n^{-1}$ is used, the bias will become negligible compared with the variance term. This, however, may not be disadvantageous when the statistical inference is of interest. We will see in Section 3.5 that the variance term is asymptotically normal. In this case, an asymptotically negligible bias enables us to construct an asymptotically unbiased confidence interval.

3.5 Asymptotic Normality

In this section, we provide sufficient conditions under which the statistic $\langle \hat{f}, g\rangle_{\mathcal{H}}$ is asymptotically normal. Because the bias is nonrandom given X, we only consider the asymptotic distribution of the variance term $g^T(X)(K(X,X)+\lambda nI)^{-1}E$. We use the notion "$\xrightarrow{\mathcal{L}}$" to denote the convergence in distribution.

Theorem 3.6. *Suppose $\sigma^2 \in (0,\infty)$ is independent of n, and $g \neq 0$. The design points X are either deterministic, or random but independent of the random error E. Under Assumptions 1-4, we have the central limit theorem*

$$\frac{1}{\sqrt{\mathrm{VAR}}} g^T(X)(K(X,X)+\lambda nI)^{-1}E \xrightarrow{\mathcal{L}} N(0,1), \ \ \textit{as } n \to \infty, \tag{3.15}$$

provided that $\lambda = o(1)$ and

$$\lambda^{-1} = o\left(n^{\frac{2m}{d+2m(1-\delta/\tau)}}\right). \tag{3.16}$$

In particular, if $\delta = \tau$, (3.16) becomes $\lambda^{-1} = o(n^{\frac{2m}{d}})$.

Theorem 3.6 conveys two important messages. First, $\lambda \asymp n^{-\frac{2m}{2m+d}}$, the optimal order of magnitude of λ to reach the minimax rate of $\|\hat{f} - f\|_{L_2}$, always entails the asymptotic normality of the variance term. Second, if $\delta = \tau$, the variance term enjoys asymptotic normality for almost all choices of λ under the assumption of Theorem 3.1.

The asymptotic normality (3.15) can be used to construct an asymptotic confidence interval for the “biased true value” $\mathbb{E}_E\langle \hat{f}, g\rangle_{\mathcal{H}}$. In practice, more interest lies in building confidence intervals for the true value $\langle f, g\rangle_{\mathcal{H}}$. This can be done if the bias is asymptotically negligible compared with the variance term. In view of Theorem 3.5, when $\delta = \tau$, $\text{BIAS}^2/\text{VAR} \xrightarrow{p} 0$ as $n \to \infty$, under the choice $\lambda \asymp n^{-1}$. Suppose $\hat{\sigma}^2$ is a consistent estimate of σ^2, such as $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{f}(x_i))^2$. Then we can estimate VAR with $\widehat{\text{VAR}} = \hat{\sigma}^2 g^T(X)(K(X,X)+\lambda n I)^{-2} g(X)$. So the suggested $1-\alpha$ confidence interval for $\langle f, g\rangle_{\mathcal{H}}$ is $\left[\langle \hat{f}, g\rangle_{\mathcal{H}} - z_{\alpha/2}\sqrt{\widehat{\text{VAR}}}, \langle \hat{f}, g\rangle_{\mathcal{H}} + z_{\alpha/2}\sqrt{\widehat{\text{VAR}}},\right]$, where $z_{\alpha/2}$ denotes the $\alpha/2$ upper quantile of the standard normal distribution.

4. Examples

In this section, we present several examples to demonstrate the breadth of the proposed framework, including special cases of practical interest.

4.1 Point Evaluations

Consider the point evaluation $l(f) = f(x_0)$ for some $x_0 \in \Omega$. We have

$$\text{VAR} = \sigma^2 K(x_0, X)(K(X,X)+\lambda n I)^{-2} K(X, x_0). \tag{4.17}$$

We use the interpolation inequality (Theorem 3.8 of Adams and Fournier (2003); also see

Brezis and Mironescu (2019) for non-integer m)

$$\|v\|_{L_\infty} \le A\|v\|_{L_2}^{1-\frac{d}{2m}}\|v\|_{H^m}^{\frac{d}{2m}}, \tag{4.18}$$

which holds for all $v \in H^m$ and some constant $A > 0$, provided that $m > d/2$. Because $f(x_0) \le \|f\|_{L_\infty}$, the interpolation inequality implies that Assumption 2 is true with $\delta = 1 - \frac{d}{2m}$. On the other hand, it can also be shown that $\tau = 1 - \frac{d}{2m}$ if x_0 is an interior point of Ω. Hence, we have the following result.

Theorem 4.1. *Suppose Assumptions 1 and 3 are true. Suppose $\lambda = o(1)$ and $\lambda^{-1} = o(n^{\frac{2m}{d}})$. Let x_0 be an interior point of Ω and* VAR *be as in (4.17). Then, we have*

1. VAR $\asymp \sigma^2 n^{-1}\lambda^{-\frac{d}{2m}}$.

2. $\sup_{\|f\|_{\mathcal{H}}\le 1}\left|\mathbb{E}_E\hat{f}(x_0) - f(x_0)\right| \asymp \lambda^{\frac{1}{2}-\frac{d}{4m}}$.

3. *Regarding σ^2 as a positive constant, under the optimal order $\lambda \asymp n^{-1}$,*

$$\sup_{\|f\|_{\mathcal{H}}\le 1} |\hat{f}(x_0) - f(x_0)| \asymp n^{-\frac{1}{2}+\frac{d}{4m}}.$$

4. *In addition, if $\sigma^2 > 0$ and $\lambda = o(n^{-1})$,* $(\mathrm{VAR})^{-\frac{1}{2}}(\hat{f}(x_0) - f(x_0)) \xrightarrow{\mathscr{L}} N(0,1)$.

Remark 2. For point evaluations of KRR, Shang and Cheng (2013); Zhao et al. (2021) obtained the rate of convergence and the asymptotic normality of the variance term, using a device called the functional Bahadur representation (Shang, 2010). The results presented in this work are under broader situations and weaker conditions: both random and deterministic designs are allowed, with wider ranges for λ and m, and there is no uniform boundedness requirement for the eigenfunctions of the kernel. Besides, we give the order of magnitude of the worst-case bias together with the best order of magnitude of λ.

4.2 Derivatives

Let $\alpha = (\alpha_1, \ldots, \alpha_d)^T \in \mathbb{N}^d$ be a multi-index and $|\alpha| = \alpha_1 + \cdots + \alpha_d$. Denote $D^\alpha f = \frac{\partial^{|\alpha|}}{\partial \chi_1^{\alpha_1} \cdots \partial \chi_d^{\alpha_d}} f$ with $x =: (\chi_1, \ldots, \chi_d)^T$. Note that the zeroth order derivative stands for the identity mapping. (Thus, the point evaluation is a special case here.) The goal is to study the asymptotic properties of $D^\alpha \hat{f}(x_0)$ for $x_0 \in \Omega$, as an estimator of $D^\alpha f(x_0)$. First, we have

$$\mathrm{VAR} = \sigma^2 D^\alpha K(x_0, X)(K(X,X) + \lambda n I)^{-2} D^\alpha K(X, x_0), \tag{4.19}$$

where $D^\alpha K$ stands for the α-th derivative of K with respect to the first argument (or the second argument, as K is symmetric.) The Sobolev embedding theorem asserts that the linear operator $l(f) = D^\alpha f(x_0)$ is bounded provided that $m > d/2 + |\alpha|$. A different version of the interpolation inequality says that

$$\|D^\alpha v\|_{L_\infty} \leq A \|v\|_{L_2}^{1 - \frac{d+2|\alpha|}{2m}} \|v\|_{H^m}^{\frac{d+2|\alpha|}{2m}}, \tag{4.20}$$

some constant $A > 0$, provided that $m > d/2 + |\alpha|$. This shows $\delta = 1 - \frac{d+2|\alpha|}{2m}$. Similarly, we have $\tau = 1 - \frac{d+2|\alpha|}{2m}$ for each interior point $x_0 \in \Omega$, giving the following result.

Theorem 4.2. *Suppose Assumptions 1 and 3 are true, and* $m > d/2+|\alpha|$. *Suppose* $\lambda = o(1)$ *and* $\lambda^{-1} = o(n^{\frac{2m}{d}})$. *Let* x_0 *be an interior point of* Ω *and* VAR *be as in (4.19). Then, we have*

1. $\mathrm{VAR} \asymp \sigma^2 n^{-1}\lambda^{-\frac{d+2|\alpha|}{2m}}$.
2. $\sup_{\|f\|_{\mathcal{H}}\leq 1}\left|\mathbb{E}_E D^\alpha \hat{f}(x_0) - D^\alpha f(x_0)\right| \asymp \lambda^{\frac{1}{2}-\frac{d+2|\alpha|}{4m}}$.
3. *Regarding* σ^2 *as a positive constant, under the optimal order* $\lambda \asymp n^{-1}$,
$$\sup_{\|f\|_{\mathcal{H}}\leq 1}|D^\alpha \hat{f}(x_0) - D^\alpha f(x_0)| \asymp n^{-\frac{1}{2}+\frac{d+2|\alpha|}{4m}}.$$
4. *In addition, if* $\sigma^2 > 0$ *and* $\lambda = o(n^{-1})$, $(\mathrm{VAR})^{-\frac{1}{2}}(D^\alpha \hat{f}(x_0) - D^\alpha f(x_0)) \xrightarrow{\mathscr{L}} N(0,1)$.

Frequently, it is imperative to establish a multivariate central limit theorem for the variance term concerning various locations or partial derivatives. For example, the joint asymptotic normality of the gradient is needed in the example introduced in Section 5.2.

Specifically, given locations $z_1,\ldots,z_{d_0}\in\Omega$ and multi-indices $\alpha_1,\ldots,\alpha_{d_0}\in\mathbb{N}^d$ for some $d_0\in\mathbb{N}_+$. Then the variance term of $D_i^\alpha \hat{f}(z_i)$ is $D^{\alpha_i}K(z_i,X)(K+\lambda nI)^{-1}E$. Thus the $d_0\times d_0$ covariance matrix of the vector of the variance terms is

$$\mathrm{COV} := \left(\sigma^2 D^{\alpha_i}K(z_i,X)(K+\lambda nI)^{-2}D^{\alpha_j}(X,z_j)\right)_{i,j}. \tag{4.21}$$

Theorem 4.3 shows a multivariate central limit theorem for the variance term when α_i's are homogeneous, in the sense that $|\alpha_1| = \cdots = |\alpha_{d_0}|$.

Theorem 4.3. *Suppose Assumption 1 is true. The covariance matrix* COV *defined in (4.21) is invertible with probability tending to one, provided that the pairs* $(\alpha_1, z_1), \ldots, (\alpha_{d_0}, z_{d_0})$ *are distinct and* $\sigma^2 > 0$*. In addition, if Assumption 3 is true,* $|\alpha_1| = \cdots = |\alpha_{d_0}| = k$*,* $m > k+d/2$*, and* z_i*'s are interior points of* Ω*, let* $\lambda = o(1)$ *and* $\lambda^{-1} = o(n^{\frac{2m}{d}})$*, then we have*

$$\mathrm{COV}^{-\frac{1}{2}} \begin{pmatrix} D^{\alpha_1} K(z_1, X) \\ \vdots \\ D^{\alpha_{d_0}} K(z_{d_0}, X) \end{pmatrix} (K + \lambda n I)^{-1} E \xrightarrow{\mathscr{L}} N(0, I),$$

4.3 L_2 Inner Products

As shown in Proposition 4.1, if $\delta = 1$, the linear functional $\langle g, \cdot \rangle_{\mathcal{H}}$ must be an L_2 inner product.

Proposition 4.1. *Suppose Assumption 1 holds. If* $g \in \mathcal{H}$ *satisfies Assumption 2 with* $\delta = 1$*, under Assumption 1, there exists a unique* $h \in L_2$*, such that* $\langle g, v \rangle_{\mathcal{H}} = \langle h, v \rangle_{L_2}$ *for each* $v \in \mathcal{H}$*.*

Let $l(f) = \int_\Omega f(x)h(x)dx$. We have

$$\mathrm{VAR} = \int_\Omega \int_\Omega h(s) K(s, X)(K(X, X) + \lambda n I)^{-2} K(X, t) h(t) ds dt, \tag{4.22}$$

Set $\delta = \tau = 1$. Corollary 4.1 follows immediately.

Corollary 4.1. *Suppose Assumptions 1 and 3 are true. Suppose* $\lambda = o(1)$ *and* $\lambda^{-1} = o(n^{\frac{2m}{d}})$*.*

Let VAR *be as in (4.22). Then, we have*

1. $\mathrm{VAR} \asymp \sigma^2 n^{-1}$.

2. $|\int_\Omega(\hat{f}-f)(x)h(x)dx| = O_{\mathbb{P}}(\lambda^{\frac{1}{2}}\|f\|_{\mathcal{H}} + \sigma n^{-\frac{1}{2}})$.

3. *In addition, if* $\sigma^2 > 0$ *and* $\lambda = o(n^{-1})$, $(\mathrm{VAR})^{-\frac{1}{2}}\int_\Omega(\hat{f}-f)(x)h(x)dx \xrightarrow{\mathscr{L}} N(0,1)$.

Remark 3. Tuo and Wu (2015) considered the L_2 inner product and demonstrated its impact on the calibration of computer models. The techniques adopted in Tuo and Wu (2015) were available in much earlier literature to study the parametric part of smoothing splines and partial linear models. All these results show a root-n rate of convergence and the asymptotic normality. The existing approach cannot deal with general $h \in L_2$, but under extra smoothness conditions of h, the theory gives the rate of convergence $O_{\mathbb{P}}(\lambda\|f\|_{\mathcal{H}} + \sigma n^{-1/2})$; see Section **??** of the Supplementary Materials for further discussion.

4.4 Expressions in terms of the Eigensystem

A more abstract, but potentially general statement starts with an equivalent representation of $\mathcal{H}$ (Wendland, 2004). The discussion is deferred to Section **??** of the Supplementary Materials.

5. Other applications of the linear functional theory

Our theory of the linear functionals of KRR can be leveraged to handle other problems. Two prominent cases would be: 1) supremum over a set of linear functionals, e.g., the uniform error, and 2) nonlinear functionals that can be linearized asymptotically, e.g., the maximum point of a function. In this section, we outline our findings. The full technical details are deferred to Sections **??** and **??** of the Supplementary Materials.

5.1 Uniform Bounds

The methodology introduced in Section 3 can be extended to study the uniform errors in terms of $\sup_{g\in\mathscr{G}}|\langle\hat{f}-f,g\rangle_{\mathcal{H}}|$. We are particularly interested in the uniform error of the partial derivatives, i.e.,

$$\sup_{x\in\Omega}\left|D^{\alpha}\hat{f}(x)-D^{\alpha}f(x)\right|, \tag{5.23}$$

for some $\alpha\in\mathbb{N}^d$. Note that (5.23) includes the L_∞ error by setting $\alpha=0$. Following the idea in Section 2, we break (5.23) into two terms.

$$(5.23)\leq\sup_{x\in\Omega}\left|\mathbb{E}_E D^{\alpha}\hat{f}(x)-D^{\alpha}f(x)\right|+\sup_{x\in\Omega}\left|D^{\alpha}\hat{f}(x)-\mathbb{E}_E D^{\alpha}\hat{f}(x)\right|. \tag{5.24}$$

With some abuse of terminology, we call the first term in (5.24) the *uniform bias* and the second term the uniform variance term.

Our analysis shows the upper bound for the uniform bias

$$\text{uniform bias}=O_{\mathbb{P}}(\lambda^{\frac{1}{2}-\frac{d+2|\alpha|}{4m}}\|f\|_{\mathcal{H}}), \tag{5.25}$$

which is attainable in the worst-case scenario. The magnitude of the variance term would depend on the random noise's tail property. When the noise has a sub-Gaussian tail, i.e., $\mathbb{E}\exp\{\vartheta e_1\} \le \exp\{\vartheta^2\varsigma^2/2\}$ for all $\vartheta \in \mathbb{R}$ and some $\varsigma^2 > 0$, we have the bound

$$\text{uniform variance term} = O_{\mathbb{P}}\left(\varsigma n^{-\frac{1}{2}}\lambda^{-\frac{d+2|\alpha|}{4m}}\sqrt{\log\left(\frac{C}{\lambda}\right)}\right). \tag{5.26}$$

Compared with the pointwise bound given by Theorem 4.2, (5.26) is inflated only by a logarithmic factor $\sqrt{\log(C/\lambda)}$. This factor cannot be improved in general, as the bound is shown to be sharp when the noise follows a normal distribution.

The bias and variance terms in (5.25) and (5.26) can be balanced by choosing $\lambda \sim n^{-1}\log n$ which is independent of m, d, and α, and the resulting rate of convergence is

$$\sup_{x\in\Omega}\left|D^\alpha\hat{f}(x) - D^\alpha f(x)\right| = O_{\mathbb{P}}\left((n^{-1}\log n)^{\frac{1}{2}-\frac{d+2|\alpha|}{4m}}\right). \tag{5.27}$$

Remark 4. The rate of convergence shown in (5.27) matches the classic L_∞ minimax rate. Nemirovski (2000) demonstrates that, under grid-based designs, the lower bounds for the minimax risk under the L_∞ norm of $D^\alpha\hat{f}(x) - D^\alpha f(x)$ in a unit ball of a Sobolev space with smoothness m, as stated in Theorem 2.1.1, is $(n/\log n)^{\frac{1}{2}-\frac{2|\alpha|+d}{4m}}$.

5.2 A Nonlinear Problem

Although this work primarily focuses on linear functionals of f, the results can help study certain nonlinear functionals if they can be linearized. In this section, we consider the nonlinear functionals $\min_{x\in\Omega} f(x)$ and $\operatorname{argmin}_{x\in\Omega} f(x)$. Consider the plug-in estimators of $\min_{x\in\Omega} f(x)$ and $\operatorname{argmin}_{x\in\Omega} f(x)$, defined as $\hat{f}_{\min} := \min_{x\in\Omega} \hat{f}(x)$ and $\hat{x}_{\min} := \operatorname{argmin}_{x\in\Omega} \hat{f}(x)$, respectively. To linearize $\hat{x}_{\min} - x_{\min}$, intuitively, we use a Taylor expansion argument $0 = \frac{\partial \hat{f}}{\partial x}(\hat{x}_{\min}) \approx \frac{\partial \hat{f}}{\partial x}(x_{\min}) + \frac{\partial^2 \hat{f}}{\partial x \partial x^T}(x_{\min})(\hat{x}_{\min} - x_{\min})$, which implies $\hat{x}_{\min} - x_{\min} \approx -H^{-1}\frac{\partial \hat{f}}{\partial x}(x_{\min})$. This inspires us to consider the linear functional $l(\hat{f} - f) = \frac{\partial(\hat{f}-f)}{\partial x}(x_{\min})$. The covariance matrix of the variance term is

$$\mathrm{COV} = \sigma^2 \frac{\partial K}{\partial x}(x_{\min}, X)(K(X,X) + \lambda n I)^{-2} \frac{\partial K}{\partial x^T}(X, x_{\min}). \tag{5.28}$$

Because both H and COV contain unknown parameters, we consider estimators

$$\hat{H} := \frac{\partial^2 \hat{f}}{\partial x \partial x^T}(\hat{x}_{\min}), \tag{5.29}$$

$$\widehat{\mathrm{COV}} := \hat{\sigma}^2 \frac{\partial K}{\partial x}(\hat{x}_{\min}, X)(K(X,X) + \lambda n I)^{-2} \frac{\partial K}{\partial x^T}(X, \hat{x}_{\min}), \tag{5.30}$$

where $\hat{\sigma}^2$ is a consistent estimator of σ^2. Under the optimal tuning parameter $\lambda \asymp n^{-1}$, we show that

1. $\|\hat{x}_{\min} - x_{\min}\| = O_{\mathbb{P}}(n^{-\frac{1}{2}+\frac{d+2}{4m}})$, $f(\hat{x}_{\min}) - f(x_{\min}) = O_{\mathbb{P}}(n^{-1+\frac{d+2}{2m}})$;

2. $\widehat{\mathrm{COV}}^{-\frac{1}{2}} \hat{H}(\hat{x}_{\min} - x_{\min}) \xrightarrow{\mathscr{L}} N(0, I)$.

6. Numerical Studies

In this section, we conduct numerical studies to examine both the pointwise asymptotic confidence interval (CI) for the estimated optimal point $\hat{x}_{\text{min}}$ and the finite-sample coverage probability of the proposed derivative estimator. We begin by evaluating the performance of the proposed estimator for estimating the optimal point using both a toy example and real data, focusing on the accuracy of the pointwise CIs for $\hat{x}_{\text{min}}$. Next, we compare the finite-sample coverage probability of the proposed derivative estimator with several alternative methods in a toy example. The results provide numerical evidence supporting the theoretical asymptotic properties of the proposed estimator.

6.1 Asymptotic Confidence Interval for Optimal Point

We conduct numerical studies to examine the pointwise asymptotic CI for the estimated optimal point $\hat{x}_{min}$ in the objective function. Three test regression functions are considered:

1. $f_1(x) = 1.8[\beta_{10,5}(x) + \beta_{7,7}(x) + \beta_{5,10}(x)]$,

2. $f_2(x) = 2.4\beta_{30,17}(x) + 2.8\beta_{4,11}(x)$,

3. $f_3(x) = \frac{7}{5}\beta_{15,30}(x) + 8\sin(32\pi x - \frac{4\pi}{3}) - 6\cos(16\pi x) - \frac{1}{5}\cos(64\pi x)$,

where $\beta_{a,b}(x)$ stands for the density function of a Beta(a, b) distribution. In all cases, we generate independent and identically distributed input data X from the uniform distribution over $[0, 1]$. The response y is given by model (1.1) after adding an independent and identi-

	Coverage Probability under Normal Noise with $\alpha = 0.05$					
	f_1		f_2		f_3	
n	$\sigma = 0.5$	$\sigma = 5$	$\sigma = 0.5$	$\sigma = 5$	$\sigma = 0.5$	$\sigma = 5$
100	0.9031	0.8010	0.8452	0.5872	0.5968	0.5978
300	0.9317	0.8304	0.9178	0.7665	0.8386	0.6223
500	0.9533	0.8821	0.9398	0.8415	0.9118	0.8344
1000	0.9543	0.9412	0.9577	0.9205	0.9441	0.8898
1500	0.9573	0.9532	0.9470	0.9389	0.9407	0.9382

Table 2: Estimated Coverage Probability for Normal Distributed Noise.

cally distributed noise. Two types of noise distributions are used: the normal distribution with a variance of 3 and the student's t-distribution with degrees of freedom $\nu = 3$. Each distribution type is used under the mean zero and two different variance (σ^2) levels.

In all simulation experiments, we choose the Matérn kernel with $\nu = 3$ and choose both its hyperparameters and the regularization parameter λ, where λ is set near the order of $O(n^{-1})$, through cross-validation. We then construct CIs for each $\hat{x}_{\min}$ at a 95% nominal level following the result in Section 5.2. The coverage probability (CP) is estimated as the proportion of the CIs that cover the true value in a total of 800 replications. In addition, we present the Q-Q plots of the test statistics $\hat{x}_{min}$ to visualize their empirical distributions versus the normal distributions. The test functions are plotted as solid curves in Figure **??** in the supplementary material. As shown in the plots, all three test functions are smooth, but have an increasing number of local optimal points.

Tables 2 and 3 summarize the CP of our asymptotic CI over 800 replications. Tables 2 and 3 imply that in the first two cases, the proposed asymptotic confidence intervals provide

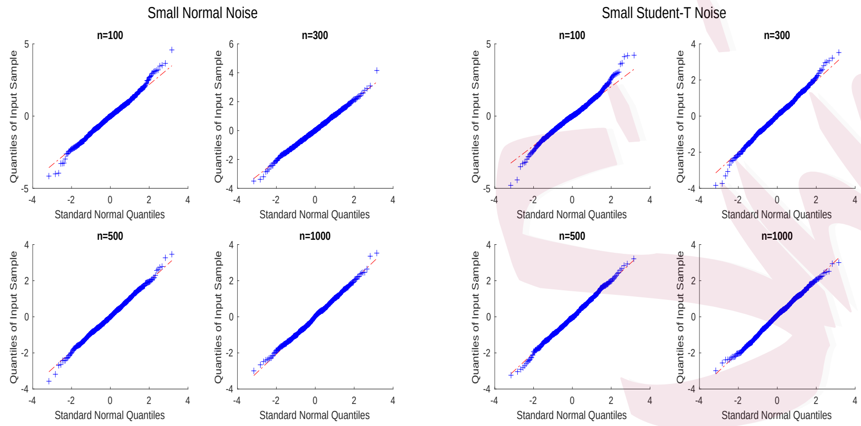

Figure 1: Results for Test Function f_1 with low-level noise $\sigma = 0.5$

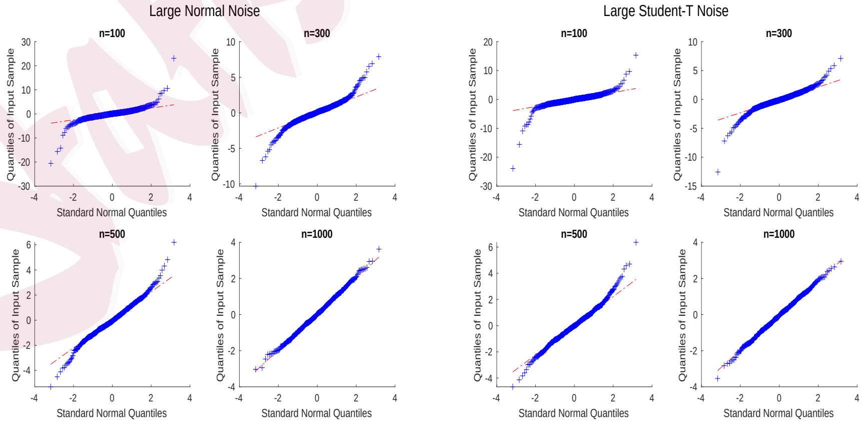

Figure 2: Results for Test Function f_1 with high-level noise $\sigma = 5$

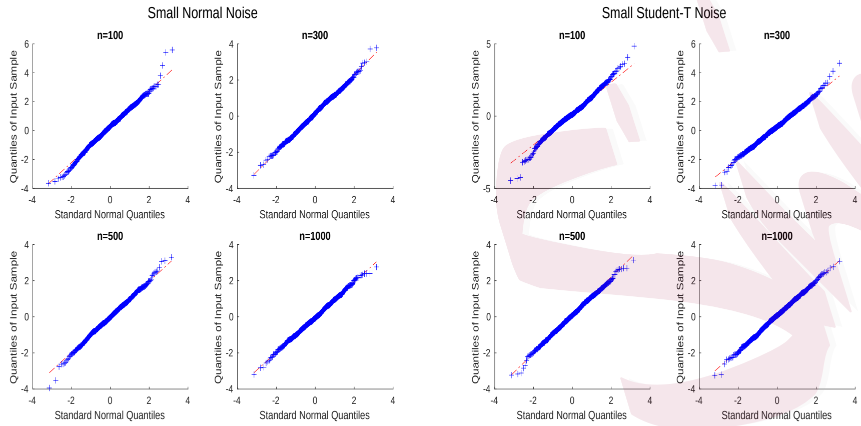

Figure 3: Results for Test Function f_2 with low-level noise $\sigma = 0.5$

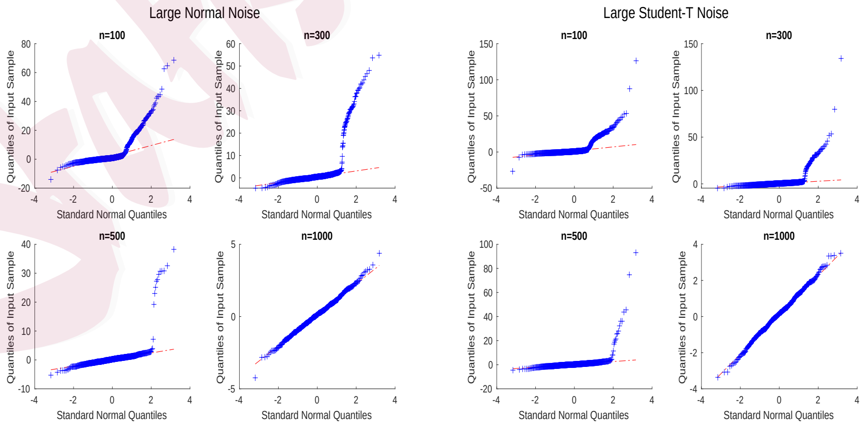

Figure 4: Results for Test Function f_2 with high-level noise $\sigma = 5$

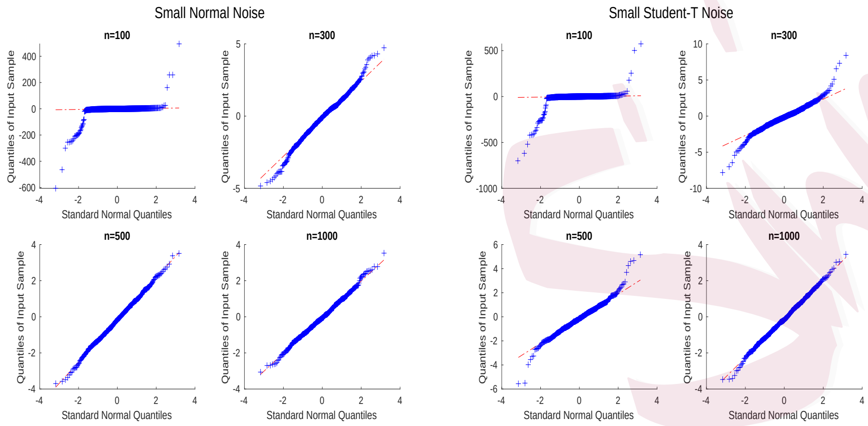

Figure 5: Results for Test Function f_3 with low-level noise $\sigma = 0.5$

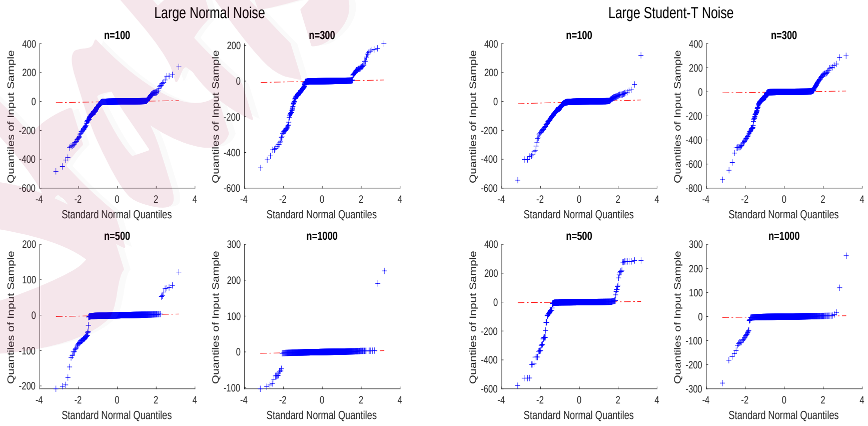

Figure 6: Results for Test Function f_3 with high-level noise $\sigma = 5$

	Coverage Probability under t_3 Noise with $\alpha = 0.05$					
	f_1		f_2		f_3	
n	$\sigma = 0.5$	$\sigma = 5$	$\sigma = 0.5$	$\sigma = 5$	$\sigma = 0.5$	$\sigma = 5$
100	0.9005	0.8101	0.8801	0.6006	0.5114	0.5578
300	0.9329	0.8412	0.9217	0.7912	0.8359	0.5976
500	0.9532	0.8897	0.9470	0.8584	0.9295	0.7716
1000	0.9402	0.9509	0.9501	0.9142	0.9310	0.8475
1500	0.9472	0.9417	0.9629	0.9401	0.9389	0.9293

Table 3: Estimated Coverage Probability for Student's-t Distributed Noise.

decent coverage rates (i.e., close to the nominal level 95%) for both functions, regardless of the type of the error distribution. For Case 3, we suffer from the under-coverage problem in high noise scenarios, KRR cannot accurately reconstruct the function and thus pinpoint the global minimum point. But such a problem is mitigated when the sample size is sufficiently large: when $n = 1500$, the proposed asymptotic CI has a CP close to 0.95.

Figures 1-6 present the Q-Q plots of the aforementioned statistics over the replications. As shown in Figures 1 and 3, when the error variance is small, the distribution of statistical quantities corresponding to two different error distributions is close to the normal distribution even under small sample sizes. However, in Case 3 with small noise, the statistical values associated with the normal distribution error closely align with the normal distribution under small sample sizes, in contrast to those associated with the t-distribution error. Nevertheless, as sample size increases, the statistics corresponding to both error distributions progressively approach the normal distribution. When the error variance is relatively large, as observed in Figures 2, 4, and 6, the Q-Q plots for both types of error distribution exhibit an S-shape, indicating that the statistics' distribution has heavier tails than the normal distribution, especially with a sample of less than 500. In particular, as demonstrated in Figure 6, the statistics with both the t-distributed errors and normally distributed errors severely deviate from a normal distribution even under a sample size of 1000. As said before, this deviation is mainly due to the large *uniform estimation errors*, so we cannot correctly pinpoint which local optimal is the global optimal. Nevertheless, as exhibited in Table 2 and Table 3, the coverage rates of the test statistics associated with a normal distribution are slightly better than those with t-distributed errors across all sample sizes. In view of the different simulation results led by the noise distribution, these results support our hypothesis in Remark 4 that the uniform rate of convergence of KRR depends on the tail property of the random noise.

In summary, the simulation results show that the asymptotic confidence interval for the optimal point generally aligns with our asymptotic analysis. The CP uniformly approaches the desired confidence level as the sample size grows, showing the validity of the intervals. In addition, the resulting confidence intervals are not sensitive to the error distribution.

6.2 Real Data Analysis

Event-related potentials (ERPs) are electroencephalogram (EEG) signals recorded in response to external stimuli, and the amplitude and latency of their characteristic waveform components are well known to reflect sensory and cognitive processes. For our real-data analysis, we use a publicly available ERP dataset (`http://dsenturk.bol.ucla.edu/supplements.html`) consisting of recordings from a single participant diagnosed with autism spectrum disorder (ASD) under one electrode and one experimental condition. The dataset contains 72 trials, each with 250 time points. Our study targets two well-established ERP components—N1, typically occurring between 100 and 250, and P3, between 190 and 370—both of which have been extensively investigated for their links to sensory and cognitive function. To capture both components, we restrict the analysis to the [100,370]. We then apply our method to construct confidence intervals for the optimal point of these component latencies, providing a calibrated assessment of their estimation uncertainty.

The aim is to estimate the optimal maximum values of the ERP signal, specifically the peak latencies of the N1 and P3 components, within the time window [100, 370]. Since EEG signals are inherently noisy, neuroscientists traditionally average the signals across trials to obtain a grand average ERP waveform. This averaged waveform is then used to estimate the amplitude and latency of the ERP components. The optimal points are estimated based on these averaged waveforms. In the supplementary material, Figure **??** plots the 72 individual ERP trial waveforms together with their grand average, with two vertical lines indicating the time window used as the search region for estimating the optimal point.

Figure 7 displays the Q–Q plot of the optimal point estimates for the real ERP data, showing close agreement between the empirical and theoretical quantiles. The empirical coverage rate of the 95% confidence intervals is 0.948, consistent with the nominal level and indicating that the intervals effectively capture the true optimal points.

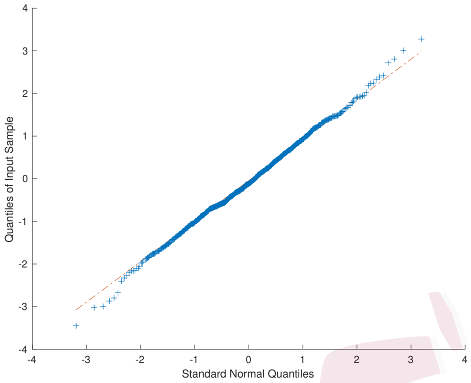

Figure 7: Q-Q Plot of Optimal Point Estimations for Real ERP Data

6.3 Comparison with existing Methods for Derivative Estimator

We consider two regression functions:

1. $f_4(x) = 5\exp\left(-2(1-2x)^2\right)(1-2x)$, with $x \in [0,1]$.

2. $f_5(x) = \sin(8.5x) + \cos(8.5x) + \log(2+x)$, with $x \in [-1,1]$.

Random design points from the uniform distributions over the designated intervals are used with sample size $n = 500$. The response y is given by model (1.1) after adding an independent and identically distributed Gaussian noise $\epsilon_i \sim N(0, 2^2)$.

We consider the first order derivative to accommodate competing methods, but note that the proposed method is readily available for any order. We construct a CI for each $\hat{f}'(x)$ with a 95% nominal level by applying Theorem 4.2. The CP is estimated as the proportion of the CIs that cover the true value in a total of 800 replications. For the plug-in KRR estimator, we adopt the same simulation setting as described in Section 6.1. We compare the plug-in KRR estimator with three other methods: local polynomial regression with degree $p = 4$ (R package `nprobust` in (Calonico et al., 2019), denoted as `locpol4` in the figures), smoothing spline (R package `lspartition` in (Cattaneo et al., 2020)) with higher-order-basis bias correction (denoted as `bspline1`) and with least squares bias correction (denoted as `bspline2`). For more details of the bias correction estimator, please refer to (Calonico et al., 2022).

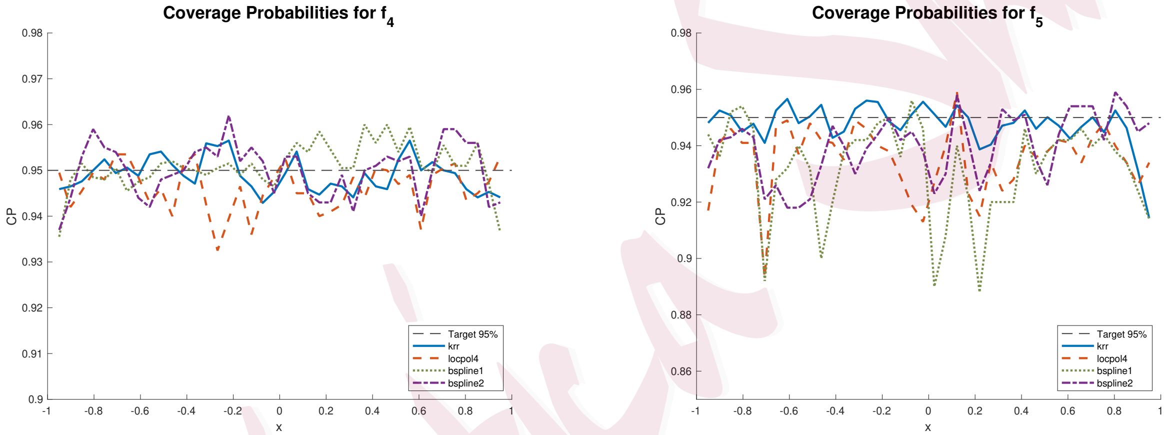

Figure 8: Estimated Coverage Probability for Derivative

Figure 8 presents the estimated coverage probabilities for f_4 (left) and f_5 (right) using the plug-in KRR estimator (`krr`), local polynomial regression with degree $p = 4$ (`locpol4`), smoothing spline with higher-order-basis bias correction (`bspline1`), and smoothing spline with least squares bias correction (`bspline2`). For f_4, all methods produce similar results across the domain, with coverage probabilities close to the nominal 95% level. For f_5, the proposed KRR method outperforms the alternative approaches over most of the domain, except near the left boundary where its coverage probability is slightly lower. For both functions, the KRR estimator exhibits relatively small fluctuations in coverage compared to other methods. Table 4 summarizes the average confidence interval widths for the derivative estimates across all target functions. The proposed KRR method yields the narrowest intervals in both cases, demonstrating superior estimation efficiency while maintaining nominal coverage. Overall, these results indicate that the proposed method maintains stable and accurate coverage across different target functions.

Method	f_4	f_5
`krr`	12.5803	11.3918
`locpol4`	17.2081	13.2785
`bspline1`	15.6488	12.0351
`bspline2`	16.8536	12.4222

Table 4: Average Lengths of the 95% Confidence Intervals for Each Method.

7. Discussion

In this paper, we develop an asymptotic theory for a variety of linear functionals of kernel ridge regression. Our theory encompasses both upper and lower bounds for the estimator's performance and its asymptotic normality under both deterministic and random designs. We also demonstrate that our asymptotic theory on linear functionals can be utilized to obtain results for uniform errors and certain non-linear problems.

This article is based on the assumption that the true function f resides within the RKHS ($\mathcal{H}$) associated with the kernel K. Our analysis can be extended to scenarios where the smoothness levels of $\mathcal{H}$ surpass those of the functional space in which the true function lies in (Fischer and Steinwart, 2020). Additionally, deriving sharp and uniform confidence bands for the estimator, presenting another interesting direction for future research. The challenge in constructing sharp and uniform confidence bands arises from the reliance of existing methods for constructing uniform confidence bands on expressing the KRR estimator through an orthonormal basis; see Shang and Cheng (2013); Singh and Vijaykumar (2023). Since linear functional estimators, such as derivatives, are typically non-orthogonal within this basis (Liu et al., 2023), existing testing procedures cannot be directly adapted to these estimators.

Supplementary Materials

The Supplementary Materials contain extra convergence results, details about the function spaces, a discussion of a key assumption, all technical proofs, an extended literature review, and additional figures from the numerical studies.

References

Adams, R. A. and Fournier, J. J. (2003). *Sobolev Spaces*, volume 140. Academic Press.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

Brezis, H. and Mironescu, P. (2019). Where Sobolev interacts with Gagliardo–Nirenberg. *Journal of Functional Analysis*, 277(8):2839–2864.

Cai, T. T. and Yuan, M. (2012). Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, 107(499):1201–1216.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2019). nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference. *Journal of Statistical Software*, 91:1–33.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2022). Coverage error optimal confidence intervals for local polynomial regression. *Bernoulli*, 28(4):2998–3022.

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

Cattaneo, M. D., Farrell, M. H., et al. (2020). lspartition: Partitioning-based least squares regression. *R Journal*, 12(1).

Cheng, G. and Shang, Z. (2013). Joint asymptotics for semi-nonparametric models under penalization. *arXiv preprint arXiv:1311.2628*.

Fischer, S. and Steinwart, I. (2020). Sobolev norm learning rates for regularized least-squares algorithms. *The Journal of Machine Learning Research*, 21(1):8464–8501.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics*, 31(5):1600–1635.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

Liu, R., Li, K., and Li, M. (2023). Estimation and hypothesis testing of derivatives in smoothing spline anova models. *arXiv preprint arXiv:2308.13905*.

Liu, Z. and Li, M. (2023). On the estimation of derivatives using plug-in kernel ridge regression estimators. *Journal of Machine Learning Research*, 24(266):1–37.

Mammen, E. and van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *The Annals of Statistics*, 25(3):1014–1035.

Mendelson, S. and Neeman, J. (2010). Regularization in kernel learning. *Ann. Statist.*, 38(1):526–565.

Messer, K. and Goldstein, L. (1993). A new class of kernels for nonparametric curve estimation. *The Annals of Statistics*, 21(1):179–195.

Nemirovski, A. (2000). Topics in non-parametric statistics. *Lecture Notes in Mathematics*, 1738:86–282.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer.

Shang, Z. (2010). Convergence rate and bahadur type representation of general smoothing spline m-estimates. *Electronic Journal of Statistics*, 4:1411–1442.

Shang, Z. and Cheng, G. (2013). Local and global asymptotic inference in smoothing spline models. *The Annals of Statistics*, 41(5):2608–2638.

Silverman, B. (1984). Spline smoothing: The equivalent variable kernel method. *Ann. Statist.*, 12(1):898–916.

Singh, R. and Vijaykumar, S. (2023). Kernel ridge regression inference with applications to preference data. *arXiv preprint arXiv:2302.06578v2*.

Smale, S. and Zhou, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172.

Steinwart, I., Hush, D. R., Scovel, C., et al. (2009). Optimal rates for regularized least squares regression. In *COLT*, pages 79–93.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–

1360.

Tuo, R. and Wu, C. F. J. (2015). Efficient calibration for imperfect computer models. *The Annals of Statistics*, 43(6):2331–2352.

van de Geer, S. A. (2000). *Empirical Processes in M-Estimation*, volume 6. Cambridge University Press.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 40(3):364–372.

Wahba, G. (1990). *Spline Models for Observational Data*, volume 59. SIAM.

Wendland, H. (2004). *Scattered Data Approximation*, volume 17. Cambridge University Press.

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

Zhao, S., Liu, R., and Shang, Z. (2021). Statistical inference on panel data models: A kernel ridge regression method. *Journal of Business & Economic Statistics*, 39(1):325–337.

Rui Tuo

Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX 77843, USA

E-mail: ruituo@tamu.edu

Lu Zou

School of Management, Shenzhen Polytechnic University, 7098 Liuxian Avenue, Nanshan District, Shenzhen, Guangdong, 518055, China

E-mail: luzou0330@szpu.edu.cn