ASYMPTOTIC THEORY FOR LINEAR FUNCTIONALS OF KERNEL RIDGE REGRESSION

Rui Tuo and Lu Zou

Texas A&M University,

Shenzhen Polytechnic University

Supplementary Material

In this supplementary material, we provide the technical details of our theoretical results. An additional literature review is available in Section S1. Section S2 provides additional convergence results and discussion to complement the findings presented in the main article. Section S3 offers preliminary information on function spaces. In Section S4, we present the equivalent conditions for a key assumption. Section S5 contains the supporting lemmas and the proofs of the theorems in our main article. Section S6 contains additional figures of the numerical results.

Contents

S1 Additional Related Literature 3

S2 Additional Convergence Results and Discussion 6

S2.1 Supporting Lemmas for Bias and Variance in Section 2 6

S2.2 Comments on Assumption 2 9

S2.3 Further Improvements in Bias 10

S2.4 Discussion on the Semiparametric Effect 11

S2.5 Expressions in terms of the Eigensystem 13

S2.6 Uniform Bounds 15

S2.7 A Nonlinear Problem 17

S3 Review of Sobolev Spaces 18

S4 Sufficient Conditions of Assumption 3 in Section 3 19

S4.1 Random Designs 20

S4.2 Fixed Designs 26

S5 Supporting Lemmas and Technical Details 27

S5.1 Proofs for Section 2 28
S5.2 Supporting Lemmas for Upper Bound in Section 3.2 . 29
S5.3 Proof for Section 3.1 30
S5.4 Proof for Upper Bound in Section 3.2 31
S5.5 Supporting Lemmas for Lower Bound in Section 3.3 . 32
S5.6 Proofs for Lower Bound in Section 3.3 33
S5.7 Proofs for Improved Results on BIAS 37
S5.8 Supporting Lemmas for Asymptotic Normality in Section 3.5 38
S5.9 Proofs for Asymptotic Normality in Section 3.5 . . . 39
S5.10 Proofs for Examples in Section 4 41
S5.11 Proofs for Uniform Bound in Section 5.1 50
S5.12 Proof for Nonlinear Problem in Section 5.2 56
S5.13 Proof of Theorem S2.2 60
S6 Additional Figures for Numerical Results 61

S1 Additional Related Literature

KRR is a prevailing technique in machine learning and statistical modeling, demonstrating extensive utility across diverse areas, including predictive modeling (Ciliberto et al., 2020; Pourkamali-Anaraki et al., 2020),

classification (Cortes et al., 2012; Zien and Ong, 2007), generative modeling (Dziugaite et al., 2015; Li et al., 2019), and statistical inference. In statistical inference areas, KRR finds specific applications in tasks such as two-sample testing, independence testing (Bach and Jordan, 2002; Gretton et al., 2005, 2012), and causal inference (Singh et al., 2019, 2020).

Error bounds for KRR. The minimax convergence rates for KRR in L_2 are thoroughly documented in the current literature. More recently, Fischer and Steinwart (2020) extended these rates to Sobolev norms without requiring the regression function to be contained in the hypothesis space. For more recent work on the convergence rate for KRR, please refer to Yang et al. (2017); Wang (2021); Cui et al. (2021); Marteau-Ferey et al. (2019); Talwai et al. (2022); Zhang et al. (2023). In recent years, there has been significant interest in characterizing the learning curve for KRR, which captures the magnitude of the generalization error as it fluctuates in response to regularization parameters. Several works (e.g., Bordelon et al. (2020); Cui et al. (2021)) depicted the learning curve of KRR under the Gaussian design. Subsequently, these results were extended to a more general random design; see Loureiro et al. (2021). It has been discovered in practice and reported in the literature (Bauer et al., 2007; Gerfo et al., 2008) that incorporating extra smoothness and refining the qualifications of the

algorithm could yield a higher convergence rate for KRR. Recent research, including works by Dicker et al. (2017); Li et al. (2022); Lian et al. (2021); Lin et al. (2020); Tuo et al. (2020), further explores strategies for achieving this improved convergence rate.

Another line of research relevant to this paper explores linear functional regression, as detailed in (Li and Zhu, 2020; Lv et al., 2023; Sun et al., 2018). These studies focus on the linear functional defined as the L_2 inner product of the input data with a slope function. Recently, Koltchinskii and Li (2023) demonstrated the asymptotic normality of smooth functionals with plug-in estimators, which relies on the assumption that the plug-in estimator can be well approximated by a normal random variable. For further literature on functional linear regression with special structures, please refer to Cui et al. (2020); Balasubramanian et al. (2022).

Statistical inference for KRR. Another approach uses KRR for statistical inference, often investigating Gaussian approximation for KRR and its variants. More recently, Singh and Vijaykumar (2023) proposed a uniform confidence band for KRR, which also provided the pointwise asymptotic normality for KRR as a byproduct. In econometric literature, exploring the linear functional form includes investigating other nonparametric regression estimators like B-spline and wavelet models. Cattaneo and Farrell

(2013) provided the uniform Bahadur representation for linear functionals of local polynomial partitioning estimators. These results are contingent upon Hölder conditions for both the underlying function and its derivatives. In a related context, (Belloni et al., 2015; Chen and Christensen, 2015) offered similar theoretical results under more general conditions.

S2 Additional Convergence Results and Discussion

This section provides additional convergence results and discussion that supplement the findings presented in the main article.

S2.1 Supporting Lemmas for Bias and Variance in Section 2

In this part we introduce a major auxiliary problem that plays a central role in our theory.

The first goal of this work is to quantify the bias and variance. It turns out that these quantities are intimately related to an auxiliary problem, called the *noiseless kernel ridge regression*.

Definition 1. Given KRR problem (1.2) and function $g \in \mathcal{H}$, the associated noiseless KRR problem is defined as

$$\hat{g} = \underset{v \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{g(x_i) - v(x_i)\}^2 + \lambda \|v\|_{\mathcal{H}}^2, \tag{S2.1}$$

where λ takes the same value as in (1.2).

Note that the target function of the noiseless KRR is g, not f. The following lemma establishes the relationship between the bias and variance, and the noiseless KRR. For notational simplicity, we will denote $\|v\|_n^2 = \frac{1}{n}\sum_{i=1}^n v^2(x_i)$ for any v.

Lemma S2.1. *The following formulas are true:*

$$\begin{aligned} |\,\mathrm{BIAS}\,| &= |\langle \hat{g} - g, f\rangle_{\mathcal{H}}| \le \|\hat{g} - g\|_{\mathcal{H}}\|f\|_{\mathcal{H}}, \qquad \text{(S2.2)} \\ \mathrm{VAR} &= \sigma^2 n^{-1}\lambda^{-2}\|\hat{g} - g\|_n^2. \end{aligned}$$

It is worth noting that Lemma S2.1 does not postulate any assumptions on the input points X. These points can be arbitrary: either deterministic or random.

To make Lemma S2.1 useful, it is critical to establish the rates of convergence of $\|\hat{g}-g\|_n$ and $\|\hat{g}-g\|_{\mathcal{H}}$. Under a standard theory (in the sense of a minimax rate of convergence), we can only have $\|\hat{g} - g\|_{\mathcal{H}} = O(1)$, which is insufficient. The key here is: if g is "smoother" than the baseline smoothness of $\mathcal{H}$, $\|\hat{g}-g\|_n$ and $\|\hat{g}-g\|_{\mathcal{H}}$ may decay faster than their minimax rates. Such a result is called an *improved rate of convergence*. Improved rates are widely available for methodologies with a variational or optimization-based formulation, such as finite element methods (Brenner and Scott, 2008) and

radial basis function approximation (Wendland, 2004). In statistics, it was also discovered long ago that extra smoothness and boundary conditions could yield a higher convergence rate for smoothing splines (Wahba and Wold, 1975). Such extra conditions are referred to as the source conditions in the machine learning literature Bauer et al. (2007); Li et al. (2022); Rastogi and Sampath (2017). Recent advances have demonstrated the general ideas to pursue an improved convergence rate for KRR (Dicker et al., 2017; Fischer and Steinwart, 2020; Guo et al., 2017; Lin et al., 2017; Tuo et al., 2020). In this work, we will adopt the approach of Tuo et al. (2020) to derive the improved rates, which leads to results in terms of both the $\|\cdot\|_n$ and $\|\cdot\|_{\mathcal{H}}$ norms.

We also highlight that the Cauchy-Schwarz inequality used in (S2.2) is sharp: the equality holds if f is a multiple of $\hat{g} - g$. This implies that $\|\hat{g} - g\|_{\mathcal{H}}$ is the *worst-case bias* over the unit ball of $\mathcal{H}$. To be more precise, when referring to the 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, and then we immediately have Corollary S2.1.

Corollary S2.1. $\sup_{\|f\|_{\mathcal{H}} \leq 1} |\,\text{BIAS}_f\,| = \|\hat{g} - g\|_{\mathcal{H}}$.

S2.2 Comments on Assumption 2

Assumption 2 is a critical condition to ensure an improved rate of convergence for $\hat{g} - g$, by imposing an extra smoothness condition on g. Technically, Assumption 2 holds if g lies in a function space $\mathcal{G}$ such that the dual space of $\mathcal{G}$ (with respect to the inner product of $\mathcal{H}$), denoted as $\mathcal{G}^*$, is an intermediate space between L_2 and $\mathcal{H}$, i.e., $L_2 \supset \mathcal{G}^* \supset \mathcal{H} \supset \mathcal{G}$. In this case,

$$\langle g, v \rangle_{\mathcal{H}} \leq \|g\|_{\mathcal{G}} \|v\|_{\mathcal{G}^*}.$$

If an "interpolation inequality" with the form

$$\|v\|_{\mathcal{G}^*} \leq C \|v\|_{L_2}^{\delta} \|v\|_{\mathcal{H}}^{1-\delta} \tag{S2.3}$$

holds for some $\delta \in (0, 1]$, Assumption 2 is valid. In general, an interpolation inequality is an inequality of the form $\|v\|_1 \leq C\|v\|_2^{1-\theta}\|v\|_3^{\theta}$ for $0 < \theta < 1$, which describes the relative strength of the norms $\|\cdot\|_1, \|\cdot\|_2$ and $\|\cdot\|_3$. For example, the following inequality, which follows simply from Hölder's inequality, links three L_p norms:

$$\|v\|_{L_{p_\theta}} \leq \|v\|_{L_{p_0}}^{1-\theta} \|v\|_{L_{p_1}}^{\theta}, \tag{S2.4}$$

where the indices $1 \leq p_0 < p_1 \leq \infty$ and $0 < \theta < 1$ satisfy

$$\frac{1}{p_\theta} = \frac{1-\theta}{p_0} + \frac{\theta}{p_1}. \tag{S2.5}$$

In view of (S2.4)-(S2.5), we can regard space L_{p_θ} as an "interpolation" of spaces L_{p_0} and L_{p_1}, and this is where its name derives from. In Section 4.1, we use the interpolation inequality (4.18) that links the L_2, L_∞ and H^m norms. A related field from functional analysis is referred to as "interpolation theory" (e.g., the Riesz-Thorin theorem). An interpolation inequality is usually a consequence of the corresponding interpolation theory.

Besides using interpolation inequalities, Assumption 2 can be verified directly when a series expansion is applied for g. See Proposition S2.1 in Section S2.5.

S2.3 Further Improvements in Bias

In case f also possesses an extra smoothness, the bias upper bound in Theorem 3.1 can be further improved. Assumption S2.1 is analogous to Assumption 2.

Assumption S2.1. There exist constants $C_f > 0$ and $\gamma \in (0, 1]$, such that for each $v \in \mathcal{H}$,

$$|\langle f, v\rangle_{\mathcal{H}}| \leq C_f \|v\|_{L_2}^{\gamma} \|v\|_{\mathcal{H}}^{1-\gamma}. \tag{S2.6}$$

Theorem S2.1. *Under the conditions and notation of Theorem 3.1 in addition to Assumption S2.1, we have* $|\,\mathrm{BIAS}\,| = O_{\mathbb{P}}(C_f \lambda^{\frac{\gamma+\delta}{2}})$.

In view of Corollary S2.1, in the presence of Assumption S2.1, the best

order of magnitude of λ to balance the bias and the variance is $\lambda \asymp n^{-\frac{1}{\gamma+1}}$. In particular, if $\gamma = 1$, one can choose $\lambda \asymp n^{-\frac{1}{2}}$. However, as γ is unknown in practice, it is difficult to take advantage of this improved rate in statistical inference.

S2.4 Discussion on the Semiparametric Effect

As shown in Proposition 4.1, when $\delta = 1$, there exists h such that $\langle g, v\rangle_{\mathcal{H}} = \langle h, v\rangle_{L_2}$ for each $v \in \mathcal{H}$. In this section, we will discuss the known results from the standard semiparametric statistical theory through the lens of the proposed approach. In the literature, it is often assumed that the input points x_i's are independent and identical random samples. Denote the probability density function of x_1 by p_X. With the techniques articulated in Mammen and van de Geer (1997); Tuo and Wu (2015), one can prove

$$\sqrt{n}\int_\Omega (\hat{f} - f)(x)h(x)dx \xrightarrow{\mathscr{L}} N\left(0, \sigma^2 \int_\Omega h^2(x)/p_X(x)dx\right), \tag{S2.7}$$

under $\lambda = o_p(n^{-1/2})$ in addition to some other conditions. Among these conditions, the most important one to our attention is

$$h/p_X \in \mathcal{H}. \tag{S2.8}$$

The objective of this part is to further understand (S2.7) together with the condition (S2.8). Clearly, (S2.7) implies that BIAS $= o_{\mathbb{P}}(n^{-1/2})$, which

cannot be obtained by simply applying Theorem 3.1 under the condition $\lambda = o(n^{-1/2})$. This implies that further improvement in the rate of convergence emerges.

To explain the actual reason, we should take the perspective of numerical integration. Define

$$\mathscr{E} := \int_\Omega (\mathbb{E}_E \hat{f} - f)(x)h(x)dx - \frac{1}{n}\sum_{i=1}^n (\mathbb{E}_E \hat{f} - f)(x_i)\frac{h(x_i)}{p_X(x_i)}, \qquad \text{(S2.9)}$$

the error of approximating the integral $\int_\Omega (\hat{f} - f)(x)h(x)dx$ with the summation $n^{-1}\sum_{i=1}^n (\hat{f} - f)(x_i)h(x_i)/p_X(x_i)$. Under our setting, x_i's are not necessarily random, and p_X can be any function of our choice with the goal of making $|\mathscr{E}|$ small.

We will first show that the second term in (S2.9) is small if $h/p_X \in \mathcal{H}$.

Theorem S2.2. *If $h/p_X \in \mathcal{H}$,*

$$\left| \frac{1}{n}\sum_{i=1}^n (\mathbb{E}_E \hat{f} - f)(x_i)\frac{h(x_i)}{p_X(x_i)} \right| \le \lambda \|f\|_{\mathcal{H}} \|h/p_X\|_{\mathcal{H}}.$$

Thus, regarding $\|f\|_{\mathcal{H}}$ and $\|h/p_X\|_{\mathcal{H}}$ as constants,

$$\begin{aligned} |\,\mathrm{BIAS}\,| &= \left| \int_\Omega (\mathbb{E}_E \hat{f} - f)(x)h(x)dx \right| \le |\mathscr{E}| + \left| \frac{1}{n}\sum_{i=1}^n (\mathbb{E}_E \hat{f} - f)(x_i)\frac{h(x_i)}{p_X(x_i)} \right| \\ &= |\mathscr{E}| + O(\lambda). \end{aligned}$$

In case x_i's are indeed independent copies with density p_X, the standard empirical process theory (van de Geer, 2000) can show that $|\mathscr{E}| = o_{\mathbb{P}}(n^{-1/2})$,

provided that $\int_\Omega (\mathbb{E}_E \hat{f} - f)^2(x) p_X(x) dx = o_{\mathbb{P}}(1)$. Hence we have recovered the results from the semiparametric statistical literature. If the input points X are carefully chosen, the integration error can be much smaller than that from a Monte Carlo sampling. For example, when $\Omega = [0,1]$ and X are evenly distributed, choosing $p_X = 1$, then $|\mathscr{E}|$ can be as small as $O(n^{-2})$. In this situation, $|\mathscr{E}|$ can be smaller than $O(\lambda)$ when λ is not too small, which implies that the lower bound in Theorem 3.4 can be reached.

S2.5 Expressions in terms of the Eigensystem

Suppose Assumption 1 is true. Let $\rho_1 \geq \rho_2 \geq \cdots$ and $\eta_1, \eta_2, \ldots$ be the eigenvalues and L_2-normalized eigenfunctions of the integral operator $L(v) = \int_\Omega K(\cdot, x) v(x) dx$. In this case, we have the representation

$$\left\| \sum_{i=1}^{\infty} c_i \eta_i \right\|_{\mathcal{H}}^2 = \sum_{i=1}^{\infty} \frac{c_i^2}{\rho_i}, \tag{S2.10}$$

for any $c_i \in \mathbb{R}$ such that the right side of (S2.10) is convergent. On the other hand, $\mathcal{H}$ is equal to all functions in the form of (S2.10) with a finite norm. Proposition S2.1 links the series presentation of functions with Assumption 2.

Proposition S2.1. *Under Assumption 1, suppose* $w = \sum_{i=1}^{n} c_i \eta_i \in \mathcal{H}$

satisfies

$$\|w\|_{\mathcal{H},\kappa}^2 := \sum_{i=1}^{\infty} \frac{c_i^2}{\rho^{1+\kappa}} < \infty, \tag{S2.11}$$

for some $\kappa \in (0, 1]$. *Then for any* $v \in \mathcal{H}$,

$$|\langle w, v\rangle_{\mathcal{H}}| \leq \|w\|_{\mathcal{H},\kappa}\|v\|_{L_2}^{\kappa}\|v\|_{\mathcal{H}}^{1-\kappa}.$$

Remark 1. A condition equivalent to (S2.11) was considered in Dicker et al. (2017) to pursue an improved rate. When $\kappa = 1$, $\langle w, \cdot\rangle_{\mathcal{H}}$ is equal to an L_2 inner product; see Proposition 4.1 and Tuo et al. (2020); Wendland (2004).

Corollary S2.2 is an immediate consequence of Proposition S2.1 and Theorem 3.1.

Corollary S2.2. *Under the conditions of Theorem 3.1 and Proposition S2.1, we have*

$$\begin{aligned} |\,\mathrm{BIAS}\,| &= O_{\mathbb{P}}(\lambda^{\frac{\kappa}{2}}\|f\|_{\mathcal{H}}), \\ \mathrm{VAR} &= O_{\mathbb{P}}(\sigma^2 n^{-1}\lambda^{\kappa-1}). \end{aligned}$$

In addition, if $\kappa > \frac{d}{2m}$, $\sigma^2 > 0$, *and* $\lambda = o(n^{-1})$, *then*

$$(\mathrm{VAR})^{-\frac{1}{2}}\langle \hat{f} - f, g\rangle_{\mathcal{H}} \xrightarrow{\mathscr{L}} N(0, 1). \tag{S2.12}$$

Because κ can be arbitrarily small, (S2.11) is a relatively weak condition given that $w \in \mathcal{H}$. Thus we can say that improved rates are generally

available for "most" functions in $\mathcal{H}$. A relevant conclusion is that the further improved rate for the bias term in Section S2.3 are also commonly available. Specifically, by Corollary S2.1, if $\|f\|_{\mathcal{H},\kappa} < \infty$, we have BIAS $= O_{\mathbb{P}}(\lambda^{\frac{\delta+\kappa}{2}})$.

The asymptotic normality (S2.12) requires $\kappa > \frac{d}{2m}$, because we use $\tau = 1$ in Assumption 4. We conjecture that this cannot be improved in general, when the magnitude of the coefficients c_i's of g in (S2.10) fluctuates wildly.

S2.6 Uniform Bounds

Recall that the goal is to determine the rate of convergence of the *uniform bias* and the uniform variance term.

As we remarked in Section 3.2, the event Ξ_ϵ is independent of g. Therefore, the uniform bias is simply the largest bias on Ξ_ϵ, which, together with the interpolation inequality (4.20), leads to Corollary S2.3.

Corollary S2.3. *Suppose Assumptions 1 and 3 are true. In addition,* $m > d/2 + |\alpha|$ *and* $\lambda \gtrsim n^{-2m/d}$. *Then we have*

$$\sup_{x\in\Omega} \left| \mathbb{E}_E D^\alpha \hat{f}(x) - D^\alpha f(x) \right| = O_{\mathbb{P}}(\lambda^{\frac{1}{2} - \frac{d+2|\alpha|}{4m}} \|f\|_{\mathcal{H}}). \tag{S2.13}$$

The uniform variance term is a supremum of a stochastic process, which seemingly depends on the random noise's tail property. Theorem S2.3 shows that, if the random noise has a sub-Gaussian tail, the uniform variance term

has almost the same order of magnitude as the pointwise variance term, except for a logarithmic factor.

Theorem S2.3. *Suppose Assumptions 1 and 3 are met,* $m > d/2 + |\alpha|$ *and* $n^{-2m/d} \lesssim \lambda \leq 1$. *In addition, if the random error satisfies* $\mathbb{E}\exp\{\vartheta e_1\} \leq \exp\{\vartheta^2\varsigma^2/2\}$ *for all* $\vartheta \in \mathbb{R}$ *and some* $\varsigma^2 > 0$, *we have*

$$\sup_{x\in\Omega}\left|D^\alpha \hat{f}(x) - \mathbb{E}_E D^\alpha \hat{f}(x)\right| = O_{\mathbb{P}}\left(\varsigma n^{-\frac{1}{2}}\lambda^{-\frac{d+2|\alpha|}{4m}}\sqrt{\log\left(\frac{C}{\lambda}\right)}\right), \quad \text{(S2.14)}$$

for some $C > 1$ *independent of* ς^2, λ, *and* n.

Comparing Theorem S2.3 with the pointwise bound given by Theorem 4.2, it can be seen that the uniform bound is inflated only by a logarithmic factor $\sqrt{\log(C/\lambda)}$. This factor cannot be improved in general, as shown in the lower bound in Theorem S2.4 under the assumption that the noise follows a normal distribution.

Theorem S2.4. *Suppose Assumptions 1 and 3 are met,* $m > d/2 + |\alpha|$ *and* $n^{-2m/d} \lesssim \lambda \leq 1$. *In addition, if the random error follows a normal distribution, i.e.,* $e_1 \sim N(0, \sigma^2)$ *with* $\sigma^2 > 0$, *we have*

$$\mathbb{P}\left(\sup_{x\in\Omega}\left|D^\alpha \hat{f}(x) - \mathbb{E}_E D^\alpha \hat{f}(x)\right| \geq C_1\sigma n^{-\frac{1}{2}}\lambda^{-\frac{d+2|\alpha|}{4m}}\sqrt{\log\left(\frac{C_2}{\lambda}\right)}\right) \geq \frac{1}{2}$$

for some constants $C_1 > 0, C_2 > 1$ *independent of* σ^2, λ *and* n.

S2.7 A Nonlinear Problem

Consider the nonlinear functionals $\min_{x\in\Omega} f(x)$ and $\operatorname{argmin}_{x\in\Omega} f(x)$. By plugging in the KRR estimator $\hat{f}$, we obtain intuitive estimators of $\min_{x\in\Omega} f(x)$ and $\operatorname{argmin}_{x\in\Omega} f(x)$ as

$$\hat{f}_{\min} := \min_{x\in\Omega} \hat{f}(x) \text{ and } \hat{x}_{\min} := \operatorname*{argmin}_{x\in\Omega} \hat{f}(x),$$

respectively. The goal is to study the asymptotic properties of these estimators. In order to linearize the problem, we make some regularity assumptions.

Assumption S2.2. The function f has a unique minimizer $x_{\min}$. Besides, $x_{\min}$ is an interior point of Ω, and f is twice differentiable at $x_{\min}$ with a positive definite Hessian matrix $H := \frac{\partial^2 f}{\partial x \partial x^T}(x_{\min})$.

Here we provide a rigorous result following the intuition provided in the main article. For simplicity, we only show the result under the optimal choice of the tuning parameter $\lambda \asymp n^{-1}$, which yields the best rate of convergence. The results are given in Theorem S2.5.

Theorem S2.5. *Suppose Assumptions 1, 3, and S2.2 are true,* $\sigma^2 > 0$, *and* $m > 2 + d/2$. *The covariance matrix* COV *and its estimate* $\widehat{\mathrm{COV}}$ *are defined by (5.28) and (5.30), respectively. Then under the optimal choice of the tuning parameter* $\lambda \asymp n^{-1}$, *we have*

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

S3 Review of Sobolev Spaces

Let $\Omega \subset \mathbb{R}^d$ be a domain. For a non-negative integer k, the Sobolev space $H^k(\Omega)$ is defined as the closure of sufficiently smooth functions over the norm

$$\|f\|^2_{H^k(\Omega)} = \sum_{|\alpha| \le k} \|D^\alpha f\|^2_{L_2(\Omega)}.$$

To define $H^m(\Omega)$ for non-integer $m = k + s$ for $k \in \mathbb{N}$ and $s \in (0, 1)$, there is a direct approach using the Sobolev–Slobodeckij semi-norm

$$|f|^2_{W^{k+s}(\Omega)} := \sum_{|\alpha|=k} \int_\Omega \int_\Omega \frac{|D^\alpha f(x) - D^\alpha f(y)|^2}{\|x - y\|^{d+2s}} dxdy, \tag{S3.15}$$

and an equivalent norm of $H^{k+s}(\Omega)$ is given by

$$\|f\|^2_{H^{k+s}(\Omega)} := \|f\|^2_{H^k(\Omega)} + |f|^2_{W^{k+s}(\Omega)}.$$

For notational simplicity, we omit the domain Ω in notation like $H^m(\Omega)$ and $\|\cdot\|_{H^m(\Omega)}$ if Ω is the experimental region of the KRR problem of our main interest.

A reproducing kernel Hilbert space $\mathcal{H}$ is a Hilbert space of continuous

functions over a domain Ω, satisfying the reproducing property

$$f(x) = \langle f, K(\cdot, x)\rangle_{\mathcal{H}}, \tag{S3.16}$$

for each $f \in \mathcal{H}$ and $x \in \Omega$. Here $K(\cdot, \cdot)$ is a positive semi-definite function called the reproducing kernel. Stationary kernels, i.e., $K(x, y) = \Phi(x - y)$ for some $\Phi : \mathbb{R}^d \mapsto \mathbb{R}$, are commonly used. When the Fourier transform of Φ, denoted as $\tilde{\Phi}$, satisfies

$$c_1(1 + \|\omega\|^2)^{-m} \leq \tilde{\Phi}(\omega) \leq c_2(1 + \|\omega\|^2)^{-m}, \tag{S3.17}$$

for $m > d/2$, some constants $0 < c_1 < c_2$, and all $\omega \in \mathbb{R}^d$, and Ω has a Lipschitz boundary, then $\mathcal{H} = H^m$ with equivalent norms; see Wendland (2004). A prominent example of kernels satisfying (S3.17) is the Matérn correlation family Stein (1999) with smoothness $\nu = m - d/2$, defined as

$$\Phi(x; \nu, \phi) := \frac{1}{\Gamma(\nu)2^{\nu-1}}(2\sqrt{\nu}\phi\|x\|)^{\nu} K_{\nu}(2\sqrt{\nu}\phi\|x\|),$$

where $\phi, \nu > 0$, and K_{ν} is the modified Bessel function of the second kind.

S4 Sufficient Conditions of Assumption 3 in Section 3

In this section, we provide the equivalent conditions for Assumption 3 under both random designs and fixed designs.

By Assumption 1, $\mathcal{H}$ and H^m are equivalent. Therefore, we can replace the $\mathcal{H}$-norm by the H^m-norm in Assumption 3, with possibly different C_1 and C_ϵ, to obtain an equivalent assumption. In this section, we shall show sufficient conditions for this equivalent assumption.

S4.1 Random Designs

The goal of this section is to prove Theorems S4.1 and S4.2 under the random design Assumptions S4.1 and S4.2, respectively. These results refine Lemma 5.16 of van de Geer (2000).

Assumption S4.1. The input sites $x_1, \ldots, x_n$ are independent and identically distributed random variables over Ω with density function $\mu(\cdot)$. In addition, $\inf_{x\in\Omega} \mu(x) = \mu_0 > 0$.

Theorem S4.1. *Suppose Ω satisfies the conditions in Assumption 1. Under Assumption S4.1, there exist constants $C_1, C_2, C_3, C_4 > 0$ independent of n and X, such that for each $t \geq 1$,*

$$\mathbb{P}\left(\|v\|_{L_2} \leq \max\left\{C_1\|v\|_n, C_2 t n^{-m/d}\|v\|_{H^m}\right\} \text{ for all } v \in H^m\right) \geq 1 - C_3 \exp\{-C_4 t^{d/m}\}, \quad \text{(S4.18)}$$

Assumption S4.2. The input sites $x_1, \ldots, x_n$ are independent and identically distributed random variables over Ω with density function $\mu(\cdot)$. In

addition, $\sup_{x\in\Omega}\mu(x) < \infty$.

Theorem S4.2. *Suppose Ω satisfies the conditions in Assumption 1. Under Assumption S4.2, there exist constants $C_1, C_2, C_3, C_4 > 0$ independent of n and X, such that for each $t \geq 1$,*

$$\mathbb{P}\left(\|v\|_n \leq \max\left\{C_1\|v\|_{L_2}, C_2 t n^{-m/d}\|v\|_{H^m}\right\} \text{ for all } v \in H^m\right) \geq 1 - C_3\exp\{-C_4 t^{d/m}\}. \quad \text{(S4.19)}$$

There are three major steps to prove Theorem S4.1. Here we call $\|v\|_{L_2} \leq C_1\|v\|_n + C_2 t n^{-m/d}\|v\|_{H^m}$ the "norm inequality" for simplicity.

1. Use a Bernstein inequality to show that the norm inequality is true with a high probability for each fixed v. This is given by Lemma S4.1.
2. Apply a "peeling device" (van de Geer, 2000) with regard to the L_2 norm, and show that the norm inequality is true with a high probability for all v satisfying $\|v\|_{L_2} \geq r$ and $\|v\|_{H^m} \leq R$ with fixed (r, R). This is given by Lemma S4.2.
3. Use a normalization argument to show (S4.18). The proof is given at the end of this section.

Theorem S4.2 can be proved in a similar fashion, starting from the opposite side of the Bernstein inequality. Hence, we omit the proof of Theorem S4.2.

Lemma S4.1. *Suppose v is continuous over Ω. Under Assumption S4.1, we have*

$$\mathbb{P}(\|v\|_{L_2} \leq \sqrt{2/\mu_0}\|v\|_n) \geq 1 - \exp\left(-\frac{3n\mu_0\|v\|_{L_2}^2}{28\|v\|_{L_\infty}^2}\right).$$

Proof. We first note that $|v^2(x_1) - \mathbb{E}v^2(x_1)| \leq \max\{v^2(x_1), \mathbb{E}v^2(x_1)\} \leq \|v\|_{L_\infty}^2$, and

$$\mathbb{E}[v^2(x_1) - \mathbb{E}v^2(x_1)]^2 \leq \mathbb{E}v^4(x_1) \leq \|v\|_{L_\infty}^2 \mathbb{E}v^2(x_1).$$

Let $t = 0.5$; we use the Bernstein inequality to obtain

$$\begin{aligned}
\mathbb{P}(\|v\|_n^2 - \mathbb{E}v^2(x_1) \leq -t\mathbb{E}v^2(x_1)) &\leq \exp\left\{-\frac{nt^2(\mathbb{E}v^2(x_1))^2/2}{\|v\|_{L_\infty}^2\mathbb{E}v^2(x_1) + \|v\|_{L_\infty}^2 t\mathbb{E}v^2(x_1)/3}\right\} \\
&= \exp\left\{-\frac{3nt^2\mathbb{E}v^2(x_1)}{2\|v\|_{L_\infty}^2(t+3)}\right\} \\
&= \exp\left\{-\frac{3n\mathbb{E}v^2(x_1)}{28\|v\|_{L_\infty}^2}\right\},
\end{aligned}$$

which, together with the property

$$\mathbb{E}v^2(x_1) = \int_\Omega v^2(x)\mu(x)dx \geq \mu_0\|v\|_{L_2}^2,$$

yields the desired result. □

The above lemma works only for a specific v. To get a bound uniform for a range of v, we need to consider the covering number.

Definition 2 ($d_{\mathcal{V}}$-covering number)**.** Let $\mathcal{V}$ be a set of functions over Ω, and $d_{\mathcal{V}}(\cdot,\cdot)$ be a semi-metric over $\mathcal{V}$. Define $N(\epsilon, \mathcal{V}, d_{\mathcal{V}})$ the smallest integer

N, such that there exist functions (also referred to as *centers*) $v_1, \ldots, v_N$ satisfying $\sup_{v \in \mathcal{V}} \min_{1 \le i \le N} d_{\mathcal{V}}(v, v_i) \le \epsilon$. In particular, for the case $\mathcal{V} \subset L_\infty$, we denote $N(\epsilon, \mathcal{V}, \|\cdot\|_{L_\infty})$ as $N(\epsilon, \mathcal{V})$ for simplicity.

The following result can be found in Edmunds and Triebel (2008). Define $H^m(R) = \{v \in H^m : \|v\|_{H^m} \le R\}$ for $R \ge 0$.

Proposition S4.1. *Suppose Ω satisfies the conditions in Assumption 1. There exists a constant $A > 0$ depending only on Ω, m, d, such that all $r > 0$,*

$$\log N(r, H^m(R)) \le A(R/r)^{d/m}. \tag{S4.20}$$

Lemma S4.2. *Suppose Ω satisfies the conditions in Assumption 1. Fix $R > 0$. For any $r > 0$ satisfying*

$$\sqrt{n}(r/R)^{d/m} \ge 1, \tag{S4.21}$$

under Assumption S4.1, there exists constants C_1, C_2, C_3, C_4 depending only on Ω, m, d and μ_0, such that

$$\begin{aligned}\mathbb{P}\left(\|v\|_{L_2} \le C_1\|v\|_n \text{ for all } v \in H^m(R) \text{ with } \|v\|_{L_2} \ge C_2 r\right) \\ \ge 1 - C_3 \exp\{-C_4 n(r/R)^{d/m}\}\end{aligned}$$

Proof. The proof proceeds by applying a peeling device. Let $|\Omega|$ be the volume of Ω, and $\mathcal{V}_s := \{v \in \mathcal{V}(R) : (s-1)|\Omega|^{1/2} r \le \|v\|_{L_2} \le s|\Omega|^{1/2} r\}$

for $s = 1, 2, \ldots$. By the definition of the covering number, we have $N(r, \mathcal{V})$ centers. For $v \in \mathcal{V}_s$, denote its associated center as $\operatorname{ctr} v$. Note that

$$(s-2)|\Omega|^{1/2}r \le \|v\|_{L_2} - |\Omega|^{1/2}r \le \|\operatorname{ctr} v\|_{L_2} \le \|v\|_{L_2} + |\Omega|^{1/2}r \le (s+1)|\Omega|^{1/2}r.$$

Define event

$$E_v := \{\|\operatorname{ctr} v\|_{L_2} \le \sqrt{2/\mu_0}\|\operatorname{ctr} v\|_n\}.$$

Then on E_v, we have

$$\begin{aligned}
\|v\|_{L_2} &\le s|\Omega|^{1/2}r \le \frac{s}{s-2}\|\operatorname{ctr} v\|_{L_2} \le \frac{s\sqrt{2/\mu}}{s-2}\|\operatorname{ctr} v\|_n \\
&\le \frac{s\sqrt{2/\mu_0}}{s-2}(\|v\|_n + r) \le \frac{s\sqrt{2/\mu_0}}{s-2}\left(\|v\|_n + \frac{|\Omega|^{-1/2}}{s-1}\|v\|_{L_2}\right).
\end{aligned}$$

Then for $s > 2|\Omega|^{-1/2} + 1$, we have $\|v\|_{L_2} \le 4\sqrt{2/\mu_0}\|v\|_n$. This proves $E_v \subset \{\|v\|_{L_2} \le 4\sqrt{2/\mu_0}\|v\|_n\}$. Therefore, by Lemma S4.1, we have

$$\begin{aligned}
&\mathbb{P}(\|v\|_{L_2} > 4\sqrt{2/\mu_0}\|v\|_n) \le \mathbb{P}(E_v^c) \le \exp\left(-\frac{3n\mu_0\|v\|_{L_2}^2}{28\|v\|_{L_\infty}^2}\right) \\
\le\ & \exp\left(-\frac{3n\mu_0\|v\|_{L_2}^2}{28C\|v\|_{L_2}^{2-d/m}\|v\|_{H^m}^{d/m}}\right) = \exp\left(-\frac{3n\mu_0}{28C}\frac{\|v\|_{L_2}^{d/m}}{\|v\|_{H^m}^{d/m}}\right) \\
\le\ & \exp\left\{-\frac{3n\mu_0}{28C}\left(\frac{(s-1)|\Omega|^{1/2}r}{R}\right)^{d/m}\right\},
\end{aligned}$$

where the third inequality follows from the interpolation inequality (4.2). Choose S_0 large enough such that we also have $3\mu_0(S_0-1)^{d/m}|\Omega|^{d/(2m)}/(28C) > (A+1)S_0^{d/(2m)}$, for A defined in (S4.20). Now we arrive at

$$\mathbb{P}\left(\bigcup_{v \in \cup_{s \ge S_0}\mathcal{V}_s}\left\{\|v\|_{L_2} > 4\sqrt{2/\mu_0}\|v\|_n\right\}\right)$$

$$
\begin{aligned}
&\leq \sum_{s\geq S_0} \mathbb{P}\left(\bigcup_{v\in\mathcal{V}_s}\left\{\|v\|_{L_2} > 4\sqrt{2/\mu_0}\|v\|_n\right\}\right) \\
&\leq \sum_{s\geq S_0} \mathbb{P}\left(\bigcup_{v\in\mathcal{V}_s} E_v^c\right) \\
&\leq \sum_{s\geq S_0} \exp\left\{\log N(r, H^m(R)) - \frac{3n\mu_0}{28C}\left(\frac{(s-1)|\Omega|^{1/2}r}{R}\right)^{d/m}\right\} \\
&\leq \sum_{s\geq S_0} \exp\left\{A(R/r)^{d/m} - (A+1)s^{d/(2m)}n(r/R)^{d/m}\right\} \\
&\leq \sum_{s\geq S_0} \exp\left\{-s^{d/(2m)}n(r/R)^{d/m}\right\},
\end{aligned}
$$

where the last inequality follows from (S4.21). This completes the proof. □

Now we are ready to prove (S4.18).

Proof of Theorem S4.1. Because $\|0\|_{L_2} \leq \max\left\{A_1\|0\|_n, A_2 n^{-m/d}\|0\|_{H^m}\right\}$ is certainly true, we only need to consider the $v \neq 0$ case. In this case,

$$
\|v\|_{L_2} \leq \max\left\{A_1\|v\|_n, A_2 n^{-m/d}\|v\|_{H^m}\right\} \tag{S4.22}
$$

is equivalent to

$$
\left\|\frac{v}{\|v\|_{H^m}}\right\|_{L_2} \leq \max\left\{A_1\left\|\frac{v}{\|v\|_{H^m}}\right\|_n, A_2 n^{-m/d}\left\|\frac{v}{\|v\|_{H^m}}\right\|_{H^m}\right\}.
$$

This implies that we only need to show (S4.22) for v with $\|v\|_{H^m} = 1$. Now we invoke Lemma S4.2 with $R = 1$ and $r = tn^{-m/d}$ for $t \geq 1$, which fulfills the condition (S4.21). Let C_1, C_2, C_3, C_4 be constants suggested by Lemma S4.2. We consider two cases.

Case 1). If $\|v\|_{L_2} < C_2 r = C_2 t n^{-m/d}$, then $\|v\|_{L_2} < C_2 t n^{-m/d} \|v\|_{H^m}$ is automatically true.

Case 2). If $\|v\|_{L_2} \geq C_2 r$. Lemma 4 implies that on an event Ξ independent of v, we have $\|v\|_{L_2} \leq C_1 \|v\|_n$ and $\mathbb{P}(\Xi) \geq 1 - C_3 \exp\left(-t^{d/m}\right)$.

Combining the above two cases, we get

$$\begin{aligned} \mathbb{P}\left(\|v\|_{L_2} \leq \max\left\{C_1\|v\|_n, C_2 t n^{-m/d}\|v\|_{H^m}\right\} \text{ for all } v \in H^m\right) \\ \geq 1 - C_3 \exp\left(-C_4 t^{d/m}\right) \end{aligned}$$

which completes the proof. □

Remark 2. Theorem S4.1 improves Lemma 5.16 of van de Geer (2000). As we examine the proof, it can be seen that this improvement is mainly due to the use of the interpolation inequality (4.18).

S4.2 Fixed Designs

For fixed designs, we assume they are *quasi-uniform*, defined as below. For a set of design points $X = \{x_1, x_2, ..., x_n\} \subset \Omega$, define the fill distance as

$$h_{X,\Omega} = \sup_{x \in \Omega} \inf_{x_j \in X} ||x - x_j||,$$

and the separation radius as

$$q_X = \min_{1 \leq j \neq k \leq n} ||x_j - x_k||/2.$$

A set of input points X is said to be *quasi-uniform* in Ω if

$$h_{X,\Omega}/q_X \leq A,$$

for some $A > 0$ independent of n.

Suppose Ω satisfies the conditions in Assumption 1. For quasi-uniform designs, Assumption 3 is a consequence of Theorems 3.3 and 3.4 of Utreras (1988). Of course, here we do not need a probabilistic statement, and (3.8)-(3.9) can be simplified to:

$$\|v\|_{L_2} \leq \max\left\{C_1\|v\|_n, C_2 n^{-m/d}\|v\|_{H^m}\right\},$$
$$\|v\|_n \leq \max\left\{C_1\|v\|_{L_2}, C_2 n^{-m/d}\|v\|_{H^m}\right\}.$$

for all $v \in H^m$ and constants C_1, C_2 depending only on Ω, d, m and the quasi-uniform constant A.

S5 Supporting Lemmas and Technical Details

In this section, we present supporting lemmas used in our main theorems and proofs of the main theorems and lemmas presented in the main article.

S5.1 Proofs for Section 2

Proof of Lemma S2.1. Denote $(u_1, \ldots, u_n)^T := \{K(X,X) + \lambda nI\}^{-1} g(X)$.

Use the representation

$$\hat{g}(\cdot) = K(\cdot, X)\{K(X,X) + \lambda nI\}^{-1} g(X), \tag{S5.23}$$

we have

$$\begin{aligned}
\text{BIAS} &= g^T(X)\{K(X,X) + \lambda nI\}^{-1} F - \langle f, g\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n} u_i f(x_i) - \langle f, g\rangle_{\mathcal{H}} \\
&= \left\langle f, \sum_{i=1}^{n} u_i K(\cdot, x_i) \right\rangle_{\mathcal{H}} - \langle f, g\rangle_{\mathcal{H}} \\
&= \left\langle f, g^T(X)\{K(X,X) + \lambda nI\}^{-1} K(X,\cdot) - g \right\rangle_{\mathcal{H}} \\
&= \langle f, \hat{g} - g\rangle_{\mathcal{H}} \\
&\leq \|f\|_{\mathcal{H}} \|\hat{g} - g\|_{\mathcal{H}},
\end{aligned}$$

where the third equality follows from the reproducing property (S3.16). This proves the bias part.

For the variance part, it suffices to note from (S5.23) that

$$\hat{g}(x_i) - g(x_i) = -\lambda n u_i. \tag{S5.24}$$

Therefore

$$\frac{1}{n}\sum_{i=1}^{n} \{\hat{g}(x_i) - g(x_i)\}^2 = n\lambda^2 g^T(X)\{K(X,X) + \lambda nI\}^{-2} g(X),$$

which proves the variance part. □

S5.2 Supporting Lemmas for Upper Bound in Section 3.2

The following lemma states the results on the improved rates for the noiseless KRR.

Lemma S5.1. *Under Assumptions 1-3, on the event Ξ_ϵ introduced in Assumption 3, we have*

$$\begin{cases} \|\hat{g}-g\|_n \le 2C_g C_1^\delta \lambda^{\frac{1+\delta}{2}} \\ \|\hat{g}-g\|_{\mathcal{H}} \le 2C_g C_1^\delta \lambda^{\frac{\delta}{2}} \end{cases}, \quad \textit{if } C_1\lambda^{\frac{1}{2}} \ge C_\epsilon n^{-\frac{m}{d}} \ \textbf{\textit{(smoothing regime)}},$$

$$\begin{cases} \|\hat{g}-g\|_n \le 2C_g C_\epsilon^\delta \lambda^{\frac{1}{2}} n^{-\frac{\delta m}{d}} \\ \|\hat{g}-g\|_{\mathcal{H}} \le 2C_g C_\epsilon^\delta n^{-\frac{\delta m}{d}} \end{cases}, \quad \textit{if } C_1\lambda^{\frac{1}{2}} < C_\epsilon n^{-\frac{m}{d}} \ \textbf{\textit{(interpolation regime)}}.$$

Lemma S5.1 shows that, depending on the choice of λ, there are two types of upper bounds. Given ϵ, we say that λ lies in the *interpolation regime* if $\lambda < (C_\epsilon/C_1)^2 n^{-2m/d}$; and otherwise, say that λ lies in the *smoothing regime*. In the interpolation regime, $\hat{g}$ behaves similarly as the kernel interpolant, i.e., the KRR estimator with $\lambda = 0$. Specifically, we have seen that as λ decreases, $\|\hat{g}-g\|_{\mathcal{H}}$ also decreases as $O_{\mathbb{P}}(\lambda^{\delta/2})$ until λ enters the interpolation regime, and thereafter $\|\hat{g}-g\|_{\mathcal{H}}$ stays as $O_{\mathbb{P}}(n^{-\delta m/d})$. This is not surprising, as $O_{\mathbb{P}}(n^{-\delta m/d})$ is the limit of this estimation: it is the rate of convergence of the kernel interpolants under the same conditions; see Tuo and Bhattacharya (2023); Wendland (2004).

It is important to note that the event Ξ_ϵ, introduced in Assumption 3, is independent of the target function g. In other words, the inequalities in Lemma S5.1 hold *simultaneously* for all g satisfying Assumption 2. This property enables us to quantify the uniform errors in terms of $\sup_{g\in\mathcal{G}} |\langle \hat{f} - f, g\rangle_{\mathcal{H}}|$. Further details will be given in Section 5.1 in the main article.

It is also worth noting that, Lemma S5.1 concerns noiseless KRR, which has only the bias but no variance. So there is no downside to using a small λ. In the presence of random noise, however, it is of no practical interest to choose λ inside the interpolation regime (say, $\lambda = o(n^{-2m/d})$), because doing so will result in way too large variances! Therefore, hereafter we only consider results in the smoothing regime in an asymptotic sense, i.e., $\lambda^{-1} = O(n^{2m/d})$, for simplicity.

Theorem 3.1 is an immediate consequence of Lemmas S2.1 and S5.1.

S5.3 Proof for Section 3.1

Proof of Lemma 3.1. By the definition

$$\text{VAR} = \sigma^2 g^T(X)\{K(X,X) + \lambda n I\}^{-2} g(X),$$

together with the condition $\sigma^2 > 0$ and that $\{K(X,X)+\lambda nI\}^{-2}$ is positive definite, VAR $= 0$ if and only if $g(X) = 0$, which implies $\|g\|_n = 0$. For any $\epsilon > 0$, let C_ϵ and Ξ_ϵ be defined in Assumption 3. Because $\|g\|_{L_2} \neq 0$,

for sufficiently large n, we have $\|g\|_{L_2} > C_\epsilon n^{-m/d}\|g\|_{\mathcal{H}}$. Then by (3.8), on the event Ξ_ϵ, $\|g\|_{L_2} \leq C_1\|g\|_n$. This shows that VAR $\neq 0$ on Ξ_ϵ, and the desired result follows. □

S5.4 Proof for Upper Bound in Section 3.2

Proof of Lemma S5.1. By the definition of noiseless KRR, we have the basic inequality

$$\|\hat{g} - g\|_n^2 + \lambda\|\hat{g}\|_{\mathcal{H}}^2 \leq \|g - g\|_n^2 + \lambda\|g\|_{\mathcal{H}}^2 = \lambda\|g\|_{\mathcal{H}}^2,$$

which is equivalent to

$$\|\hat{g} - g\|_n^2 + \lambda\|\hat{g} - g\|_{\mathcal{H}}^2 \leq 2\lambda\langle g, \hat{g} - g\rangle_{\mathcal{H}}.$$

Plugging in Assumptions 2-3, on Ξ_ϵ, we have

$$\begin{aligned} &\|\hat{g} - g\|_n^2 + \lambda\|\hat{g} - g\|_{\mathcal{H}}^2 \leq 2\lambda C_g\|\hat{g} - g\|_{L_2}^{\delta}\|\hat{g} - g\|_{\mathcal{H}}^{1-\delta}. \\ \leq\ & 2\lambda C_g \max\left\{C_1^{\delta}\|\hat{g} - g\|_n^{\delta}\|\hat{g} - g\|_{\mathcal{H}}^{1-\delta}, C_\epsilon^{\delta} n^{-\delta m/d}\|\hat{g} - g\|_{\mathcal{H}}\right\}, \end{aligned}$$

which can be broken down into two cases.

Case 1): $\|\hat{g} - g\|_n^2 + \lambda\|\hat{g} - g\|_{\mathcal{H}}^2 \leq 2\lambda C_g C_1^{\delta}\|\hat{g} - g\|_n^{\delta}\|\hat{g} - g\|_{\mathcal{H}}^{1-\delta}$, which implies

$$\begin{cases} \|\hat{g} - g\|_n^2 \leq 2\lambda C_g C_1^{\delta}\|\hat{g} - g\|_n^{\delta}\|\hat{g} - g\|_{\mathcal{H}}^{1-\delta}, \\ \lambda\|\hat{g} - g\|_{\mathcal{H}}^2 \leq 2\lambda C_g C_1^{\delta}\|\hat{g} - g\|_n^{\delta}\|\hat{g} - g\|_{\mathcal{H}}^{1-\delta}. \end{cases} \tag{S5.25}$$

The above system can be solved with elementary algebra. The solution is

$$\begin{cases} \|\hat{g} - g\|_n \leq 2C_g C_1^{\delta} \lambda^{\frac{1+\delta}{2}}, \\ \|\hat{g} - g\|_{\mathcal{H}} \leq 2C_g C_1^{\delta} \lambda^{\frac{\delta}{2}}. \end{cases} \tag{S5.26}$$

Case 2): $\|\hat{g} - g\|_n^2 + \lambda \|\hat{g} - g\|_{\mathcal{H}}^2 \leq 2\lambda C_g C_\epsilon^{\delta} n^{-\delta m/d} \|\hat{g} - g\|_{\mathcal{H}}$, which implies

$$\begin{cases} \|\hat{g} - g\|_n^2 \leq 2\lambda C_g C_\epsilon^{\delta} n^{-\delta m/d} \|\hat{g} - g\|_{\mathcal{H}}, \\ \lambda \|\hat{g} - g\|_{\mathcal{H}}^2 \leq 2\lambda C_g C_\epsilon^{\delta} n^{-\delta m/d} \|\hat{g} - g\|_{\mathcal{H}}. \end{cases}$$

The solution is

$$\begin{cases} \|\hat{g} - g\|_n \leq 2C_g C_\epsilon^{\delta} \lambda^{\frac{1}{2}} n^{-\frac{\delta m}{d}}, \\ \|\hat{g} - g\|_{\mathcal{H}} \leq 2C_g C_\epsilon^{\delta} n^{-\frac{\delta m}{d}}. \end{cases} \tag{S5.27}$$

Clearly, if $C_1^{\delta} \lambda^{\delta/2} \geq C_\epsilon^{\delta} n^{-\delta m/d}$, (S5.27) is implied by (S5.26); otherwise, (S5.26) is implied by (S5.27). This completes the proof. □

S5.5 Supporting Lemmas for Lower Bound in Section 3.3

In view of Lemma S2.1, we have analogous lower bounds for the noiseless KRR.

Lemma S5.2. *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*

$$\begin{aligned} \|\hat{g} - g\|_n &\geq A_3 \lambda^{\frac{\delta\tau - \delta + 2\tau}{2\tau}}, \qquad \text{(S5.28)} \\ \|\hat{g} - g\|_{\mathcal{H}} &\geq \begin{cases} A_3 \lambda^{\frac{2\tau - 2\delta + \delta^2 - \delta^2\tau}{2\tau(1-\delta)}} & \text{if } \delta < 1 \\ A_3 \lambda & \text{if } \delta = 1 \end{cases}. \qquad \text{(S5.29)} \end{aligned}$$

In particular, if $\delta = \tau$*, we have*

$$\begin{aligned} \|\hat{g} - g\|_n &\geq A_3 \lambda^{\frac{1+\delta}{2}}, \\ \|\hat{g} - g\|_{\mathcal{H}} &\geq \begin{cases} A_3 \lambda^{\frac{\delta}{2}} & \text{if } \delta < 1 \\ A_3 \lambda & \text{if } \delta = 1 \end{cases}. \qquad \text{(S5.30)} \end{aligned}$$

When $\delta = \tau < 1$, the noiseless KRR's convergence rate is completely known.

S5.6 Proofs for Lower Bound in Section 3.3

Proof of Proposition 3.1. Suppose $\sup_{v \in \mathcal{H}} \frac{|\langle g, v \rangle_{\mathcal{H}}|}{\|v\|_{L_2}^{\delta} \|v\|_{\mathcal{H}}^{1-\delta}} = A$. Then for each $R > 0$,

$$\begin{aligned} \sup_{\|v\|_{\mathcal{H}} \leq R\|v\|_{L_2}} \frac{|\langle g, v \rangle_{\mathcal{H}}|}{\|v\|_{L_2}} &\leq \sup_{\|v\|_{\mathcal{H}} \leq R\|v\|_{L_2}} \frac{A\|v\|_{L_2}^{\delta} \|v\|_{\mathcal{H}}^{1-\delta}}{\|v\|_{L_2}} \\ &= A \sup_{\|v\|_{\mathcal{H}} \leq R\|v\|_{L_2}} \frac{\|v\|_{\mathcal{H}}^{1-\delta}}{\|v\|_{L_2}^{1-\delta}} \\ &\leq C R^{1-\delta}. \end{aligned}$$

Conversely, suppose $\sup_{\|v\|_{\mathcal{H}} \leq R\|v\|_{L_2}} \frac{|\langle g, v \rangle_{\mathcal{H}}|}{\|v\|_{L_2}} \leq C R^{1-\delta}$ for each $R > 0$. First we note that, under Assumption 1, $\|\cdot\|_{\mathcal{H}}$ is stronger than $\|\cdot\|_{L_2}$, which

means $\|v\|_{L_2}/\|v\|_{\mathcal{H}} \leq A_1$ for all $v \in \mathcal{H}$. Then for each $v \in \mathcal{H}$ satisfying $\|v\|_{\mathcal{H}} \leq \|v\|_{L_2}$, we have

$$\frac{|\langle g, v\rangle_{\mathcal{H}}|}{\|v\|_{L_2}^{\delta}\|v\|_{\mathcal{H}}^{1-\delta}} = \frac{|\langle g, v\rangle_{\mathcal{H}}|}{\|v\|_{L_2}} \cdot \frac{\|v\|_{L_2}^{1-\delta}}{\|v\|_{\mathcal{H}}^{1-\delta}} \leq CA_1^{1-\delta}. \tag{S5.31}$$

Next, for each $i = 1, 2, \ldots$ and $v \in \mathcal{H}$ satisfying $2^{i-1} \leq \|v\|_{\mathcal{H}}/\|v\|_{L_2} \leq 2^i$,

$$\begin{aligned} \frac{|\langle g, v\rangle_{\mathcal{H}}|}{\|v\|_{L_2}^{\delta}\|v\|_{\mathcal{H}}^{1-\delta}} &= \frac{|\langle g, v\rangle_{\mathcal{H}}|}{\|v\|_{L_2}} \cdot \frac{\|v\|_{L_2}^{1-\delta}}{\|v\|_{\mathcal{H}}^{1-\delta}} \leq C(2^i)^{1-\delta} \cdot (2^{1-i})^{1-\delta} \\ &= 2^{1-\delta}C. \end{aligned} \tag{S5.32}$$

Combining (S5.31) and (S5.32) leads to

$$\sup_{v\in\mathcal{H}} \frac{|\langle g, v\rangle_{\mathcal{H}}|}{\|v\|_{L_2}^{\delta}\|v\|_{\mathcal{H}}^{1-\delta}} \leq \max\{2^{1-\delta}, A_1^{1-\delta}\}C < +\infty.$$

This completes the proof. □

Proof of Theorem 3.2. Note that $\text{VAR} = \sigma^2 g^T(X)\{K(X,X)+\lambda nI\}^{-2}g(X)$. By Cauchy-Schwarz inequality, we have

$$\begin{aligned} [\mathbf{v}\{K(X,X) + \lambda nI\}^{-1}g(X)]^2 &\leq \mathbf{v}^T\mathbf{v} \cdot g^T(X) \cdot \\ &\quad \{K(X,X) + \lambda nI\}^{-2}g(X), \end{aligned} \tag{S5.33}$$

for each $\mathbf{v} \in \mathbb{R}^n$. Now take $\mathbf{v} = v(X)$ for some $v \in \mathcal{H}$. By (S5.23),

$$\mathbf{v}\{K(X,X) + \lambda nI\}^{-1}g(X) = \langle v, \hat{g}\rangle_{\mathcal{H}};$$

thus (S5.33) implies

$$g^T(X)\{K(X,X) + \lambda nI\}^{-2}g(X) \geq \sup_{v\in\mathcal{H}} \frac{\langle v, \hat{g}\rangle_{\mathcal{H}}^2}{n\|v\|_n^2}. \tag{S5.34}$$

(Actually, the equality holds as $v(X)$ can go over the entire $\mathbb{R}^n$.) In view of (S5.34), the strategy is to bound $\langle v, \hat{g}\rangle_{\mathcal{H}}/\|v\|_n$ from below with a carefully chosen v.

To proceed, we need to get rid of the annoying $\|\cdot\|_n$ norm and the KRR estimator. This can be done by invoking the bounds in Assumption 3 and Lemma S5.1, which state that on the event Ξ_ϵ,

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

and

$$|\langle v, \hat{g} - g\rangle_{\mathcal{H}}| \le \|v\|_{\mathcal{H}}\|\hat{g} - g\|_{\mathcal{H}} \le 2C_g C_1^{\delta}\lambda^{\frac{\delta}{2}}\|v\|_{\mathcal{H}}.$$

Therefore, for any v satisfying

$$C_\epsilon n^{-m/d}\|v\|_{\mathcal{H}}/\|v\|_{L_2} \le C_1, \tag{S5.35}$$

we have

$$\begin{aligned}\frac{\langle v, \hat{g}\rangle_{\mathcal{H}}}{\|v\|_n} &\ge \frac{\langle v, \hat{g}\rangle_{\mathcal{H}}}{C_1\|v\|_{L_2}} = \frac{\langle v, g\rangle_{\mathcal{H}}}{C_1\|v\|_{L_2}} + \frac{\langle v, \hat{g} - g\rangle_{\mathcal{H}}}{C_1\|v\|_{L_2}} \\ &\ge \frac{\langle v, g\rangle_{\mathcal{H}}}{C_1\|v\|_{L_2}} - 2C_g C_1^{\delta-1}\lambda^{\frac{\delta}{2}}\frac{\|v\|_{\mathcal{H}}}{\|v\|_{L_2}}. \end{aligned} \tag{S5.36}$$

Assumption 4 implies that for each $R \ge R_0$, there exists $v \in \mathcal{H}$ such that $\|v\|_{\mathcal{H}}/\|v\|_{L_2} \le R$ and $R^{1-\tau}\langle v, g\rangle_{\mathcal{H}}/\|v\|_{L_2} > C_0$. Using this specific v in (S5.36) leads to

$$\frac{\langle v, \hat{g}\rangle_{\mathcal{H}}}{\|v\|_n} \ge \frac{C_0 R^{1-\tau}}{C_1} - 2C_g C_1^{\delta-1}\lambda^{\frac{\delta}{2}}R. \tag{S5.37}$$

Our goal is to make the right-hand side of (S5.37) no less than $\frac{C_0 R^{1-\tau}}{2C_1}$, which requires taking R no more than $4^{-1/\tau} C_0^{1/\tau} C_1^{-\delta/\tau} C_g^{-1/\tau} \lambda^{-\delta/(2\tau)}$. Clearly, we can find suitable constants A_1, A_2 depends only on $C_0, C_1, C_g, C_\epsilon, R_0, \delta, \tau$, such that for each λ satisfying

$$A_1 n^{-\frac{2m}{d}} \leq \lambda \leq A_2,$$

we have

$$R_0 \leq 4^{-1/\tau} C_0^{1/\tau} C_1^{-\delta/\tau} C_g^{-1/\tau} \lambda^{-\delta/(2\tau)} \leq \frac{C_1 n^{\frac{m}{d}}}{C_\epsilon}. \tag{S5.38}$$

This implies that the choice $R = 4^{-1/\tau} C_0^{1/\tau} C_1^{-\delta/\tau} C_g^{-1/\tau} \lambda^{-\delta/(2\tau)}$ fulfills the conditions $R \geq R_0$ and (S5.35), and leads to

$$\frac{\langle v, \hat{g} \rangle_{\mathcal{H}}}{\|v\|_n} \geq \frac{C_0 R^{1-\tau}}{2C_1},$$

which, together with (S5.34), completes the proof. □

Proof of Lemma S5.2. The relationship between $\|\hat{g} - g\|_n$ and VAR in Lemma S2.1 and Theorem 3.2 lead to (S5.28) immediately. To bound $\|\hat{g} - g\|_{\mathcal{H}}$ from below, we first make a possible adjustment of A_1 from that given by Theorem 3.2, so that λ lies in the smoothing regime defined in Lemma S5.1. Then we resort to the first inequality in (S5.25) from the proof of Lemma S5.1, which states

$$\|\hat{g} - g\|_n^2 \leq 2\lambda C_g C_1^\delta \|\hat{g} - g\|_n^\delta \|\hat{g} - g\|_{\mathcal{H}}^{1-\delta}. \tag{S5.39}$$

When $\delta < 1$, we substitute the lower bound of $\|\hat{g} - g\|_n$ to (S5.39), and arrive at the first part of (S5.29) by elementary algebraic calculations. For $\delta = 1$, we note that Assumption 2 is also true for any $\delta' < 1$. We then invoke the first part of (S5.29), which we just proved, by substituting $\delta \leftarrow \delta'$ and $\tau \leftarrow 1$. The resulting lower bound is $A_3\lambda$, regardless of the choice of δ'. □

Combining Lemmas S5.1, S5.2 and Corollary S2.1, we obtain Theorem 3.4.

S5.7 Proofs for Improved Results on BIAS

Proof of Theorem 3.5. We shall use the eigensystem representation of the RKHS norm in this proof. We follow the notation introduced in Section S2.5 and denote $f = \sum_{i=1}^{\infty} c_i \eta_i$ and $\hat{g} - g = \sum_{i=1}^{\infty} a_i \eta_i$. By Lemma S2.1,

$$|\,\mathrm{BIAS}\,| = |\langle \hat{g} - g, f \rangle_{\mathcal{H}}| = \left| \sum_{i=1}^{\infty} \frac{a_i c_i}{\rho_i} \right|. \tag{S5.40}$$

Basic calculus suggests that we can find an infinite series which converges slower than $\sum_{i=1}^{\infty} c_i^2/\rho_i$. In other words, there exists a sequence $\xi_i \downarrow 0$, such that $\sum_{i=1}^{\infty} \frac{c_i^2}{\rho_i \xi_i} < \infty$. Now we apply the Cauchy-Schwarz inequality to (S5.40) to find

$$|\,\mathrm{BIAS}\,| \leq \left(\sum_{i=1}^{\infty} \frac{a_i^2 \xi_i}{\rho_i} \right)^{1/2} \left(\frac{c_i^2}{\rho_i \xi_i} \right)^{1/2}.$$

Now it suffices to prove that $\sum_{i=1}^{\infty} a_i^2 \xi_i / \rho_i = o_{\mathbb{P}}(\lambda^{\delta})$. Because $\xi_i \downarrow 0$, for any $\epsilon > 0$, there exists N such that $\xi_i < \epsilon$ for all $i > N$. We now write

$$\begin{aligned}
\sum_{i=1}^{\infty} \frac{a_i^2 \xi_i}{\rho_i} &= \left(\sum_{i=1}^{N} + \sum_{i=N+1}^{\infty} \right) \frac{a_i^2 \xi_i}{\rho_i} \\
&\leq \left(\max_{1 \leq i \leq N} \frac{\xi_i}{\rho_i} \right) \sum_{i=1}^{\infty} a_i^2 + \epsilon \sum_{i=1}^{\infty} \frac{a_i^2}{\rho_i} \\
&= \left(\max_{1 \leq i \leq N} \frac{\xi_i}{\rho_i} \right) \|\hat{g} - g\|_{L_2}^2 + \epsilon \|\hat{g} - g\|_{\mathcal{H}}^2.
\end{aligned}$$

Employing Assumption 3 and Theorem 3.1, together with the condition $\lambda = o(1)$, on Ξ_{ϵ}, the above equation is no more than

$$\begin{aligned}
&\left(\max_{1 \leq i \leq N} \frac{\xi_i}{\rho_i} \right) \max \left\{ C_1^2 \|\hat{g} - g\|_n^2, C_{\epsilon}^2 n^{-m/d} \|\hat{g} - g\|_{\mathcal{H}}^2 \right\} + \epsilon \|\hat{g} - g\|_{\mathcal{H}}^2 \\
= &\left(\max_{1 \leq i \leq N} \frac{\xi_i}{\rho_i} \right) o(\lambda^{\delta}) + \epsilon O(\lambda^{\delta}).
\end{aligned}$$

This proves $\sum_{i=1}^{\infty} a_i^2 \xi_i / \rho_i = o_{\mathbb{P}}(\lambda^{\delta})$ as ϵ is arbitrary. □

Proof of Theorem S2.1. The desired result follows from using (S2.6) instead of the Cauchy-Schwarz inequality in (S2.2), together with Lemma S5.1. □

S5.8 Supporting Lemmas for Asymptotic Normality in Section 3.5

The following Lemma S5.3 is a consequence of the Lindeberg central limit theorem, and it is the key lemma for proving the asymptotic normality of our estimator. We use the notion "$\xrightarrow{\mathscr{L}}$" to denote the convergence in

distribution.

Lemma S5.3. *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. If*

$$n^{-1/2} \frac{\|\hat{g} - g\|_{L_\infty}}{\|\hat{g} - g\|_n} \xrightarrow{p} 0, \text{ as } n \to \infty, \tag{S5.41}$$

then we have the central limit theorem

$$\frac{1}{\sqrt{\text{VAR}}} g^T(X)(K(X, X) + \lambda n I)^{-1} E \xrightarrow{\mathscr{L}} N(0, 1), \text{ as } n \to \infty. \tag{S5.42}$$

We can verify (S5.41) provided that we have an upper bound of $\|\hat{g} - g\|_{L_\infty}$ and a lower bound of $\|\hat{g} - g\|_n$. The final result is given in Theorem 3.6.

S5.9 Proofs for Asymptotic Normality in Section 3.5

Proof of Lemma S5.3. For clarity, we shall reinstate the subscript n for each term depending on n in this proof. For instance, we will denote X by X_n to emphasize its dependence on n. Again, we define $(u_{1n}, \ldots, u_{nn})^T := (K(X_n, X_n) + \lambda_n n I_n)^{-1} g(X_n)$. Then

$$g^T(X_n)(K(X_n, X_n) + \lambda_n n I_n)^{-1} E_n = \sum_{i=1}^{n} u_{in} e_i.$$

First, we regard X_n as a fixed sequence, i.e., we set conditioning on X_n if the design is random. Then u_{in}'s are fixed. Then we shall have the central

limit theorem

$$\frac{1}{\sigma\sqrt{\sum_{i=1}^n u_{in}^2}}\sum_{i=1}^n u_{in}e_i \xrightarrow{d} N(0,1),$$

provided that the Lindeberg condition

$$\lim_{n\to\infty}\frac{1}{\sigma^2\sum_{i=1}^n u_{in}^2}\sum_{i=1}^n \mathbb{E}\left[u_{in}^2 e_i^2 \mathbf{1}_{\{u_{in}^2 e_i^2 > \epsilon^2\sigma^2\sum_{j=1}^n u_{jn}^2\}}\right] = 0 \tag{S5.43}$$

is fulfilled. It is easily seen that a sufficient condition of (S5.43) is (see also Lemma 3.1 of Huang (2003), and the proof of Theorem 4.3)

$$\max_{1\le i\le n} u_{in}^2 / \sum_{i=1}^n u_{in}^2 \to 0, \text{ as } n\to\infty. \tag{S5.44}$$

This is equivalent to, by (S5.24),

$$\frac{\max_{1\le i\le n}(\hat{g}_n - g)^2(x_{in})}{\sum_{i=1}^n(\hat{g}_n - g)^2(x_{in})} \to 0, \text{ as } n\to\infty, \tag{S5.45}$$

which is ensured by (S5.41), except that (S5.41) converges only in probability. In fact, the convergence in probability in the Lindeberg condition still leads to the central limit theorem, because X is independent of E; see, e.g., Theorem 1 on page 171 of Pollard (1984). □

Proof of Theorem 3.6. The interpolation inequality (4.18), together with Assumption 1, implies that

$$\|\hat{g} - g\|_{L_\infty} \le C\|\hat{g} - g\|_{L_2}^{1-\frac{d}{2m}}\|\hat{g} - g\|_{\mathcal{H}}^{\frac{d}{2m}}.$$

Invoke Assumption 3, we know on Ξ_ϵ,

$$\|\hat{g} - g\|_{L_\infty} \le C\max\left\{C_1\|\hat{g} - g\|_n^{1-\frac{d}{2m}}\|\hat{g} - g\|_{\mathcal{H}}^{\frac{d}{2m}}, C_\epsilon n^{-\frac{m}{d}\left(1-\frac{d}{2m}\right)}\|\hat{g} - g\|_{\mathcal{H}}\right\}$$

$$\begin{aligned} &\leq 2CC_gC_1 \max\left\{C_1(\lambda^{\frac{1+\delta}{2}})^{1-\frac{d}{2m}}\lambda^{\frac{\delta d}{4m}}, C_\epsilon n^{-\frac{m}{d}\left(1-\frac{d}{2m}\right)}\lambda^{\frac{\delta}{2}}\right\} \\ &= 2CC_gC_1 \max\{C_1\lambda^{\frac{1+\delta}{2}-\frac{d}{4m}}, n^{-\frac{2m-d}{2d}}\lambda^{\frac{\delta}{2}}\} \\ &= O(\lambda^{\frac{1+\delta}{2}-\frac{d}{4m}}), \end{aligned} \tag{S5.46}$$

where the second inequality follows from Lemma S5.1, and the last equality follows from the condition $\lambda^{-1} = O(n^{2m/d})$. Combining the above upper bound of $\|\hat{g}-g\|_{L_\infty}$ with the lower bound given in Lemma S5.2 and condition (3.16) leads to (S5.41). Then we invoke Lemma S5.3 to arrive at the desired result. □

S5.10 Proofs for Examples in Section 4

To prove Theorems 4.1 and 4.2, it suffices to show that $\delta = \tau$ in both cases, which is implied by Proposition S5.1.

Proposition S5.1. *Let α be a multi-index. Suppose $m > d/2 + |\alpha|$. For each interior point of Ω, denoted as x_0, there exist $A_1, A_2 > 0$, such that*

$$\sup_{\|v\|_{H^m} \leq R\|v\|_{L_2}} \frac{D^\alpha v(x_0)}{\|v\|_{L_2}} \geq A_1 R^{\frac{d+2|\alpha|}{2m}},$$

for all $R \geq A_2$.

Proof. Choose a function $B(x)$ such that $B(x) \in C^\infty(\mathbb{R}^d)$ and $B(x)$ is supported in the unit ball of $\mathbb{R}^d$. The function $D^\alpha B$ must be nonzero at some point. Without loss of generality, we assume that $D^\alpha B(0) = 1$,

because if otherwise, we can translate, dilate, and rescale B to make this happen. Define $w(x) = B((x - x_0)/\rho)$ for $\rho \in (0, 1)$. For each multi-index β, the chain rule implies

$$D^\beta w(x) = \rho^{-|\beta|} D^\beta B((x - x_0)/\rho). \tag{S5.47}$$

Thus for each sufficiently small ρ such that w is supported in Ω, we have

$$\int_\Omega \{D^\beta w(x)\}^2 dx \le \rho^{-2|\beta|+d} \int_{\mathbb{R}^d} \{D^\beta B(x)\}^2 dx,$$

which implies that $\|w\|_{H^k} \le \rho^{-(k-d/2)} \|B\|_{H^k(\mathbb{R}^d)}$ for integer k. In particular, we have

$$\|w\|_{L_2} = \rho^{d/2} \|B\|_{L_2(\mathbb{R}^d)}.$$

If m is not an integer, direct calculations show the Sobolev-Slobodeckij semi-norm in (S3.15) satisfies

$$|w|_{W^m} = \rho^{-(m-d/2)} |B|_{W^m(\mathbb{R}^d)},$$

which, again, implies $\|w\|_{H^m} \le \rho^{-(m-d/2)} \|B\|_{H^m(\mathbb{R}^d)}$. Besides, (S5.47) also shows $D^\alpha w(x_0) = \rho^{-|\alpha|}$. In summary, for sufficiently small ρ, we have

$$\sup_{\frac{\|v\|_{H^m}}{\|v\|_{L_2}} \le \rho^{-m} \frac{\|B\|_{H^m(\mathbb{R}^d)}}{\|B\|_{L_2(\mathbb{R}^d)}}} \frac{D^\alpha v(x_0)}{\|v\|_{L_2}} \ge \frac{D^\alpha w(x_0)}{\|w\|_{L_2}} = \frac{\rho^{-|\alpha|-d/2}}{\|B\|_{L_2(\mathbb{R}^d)}},$$

which leads to the desired result by replacing $R = \rho^{-m}$. □

Proof of Theorem 4.3. Without loss of generality, we assume that $\sigma^2 = 1$. First, we prove that COV is invertible with probability tending to one.

It suffices to prove that the smallest eigenvalue of COV, denoted by $\underline{\lambda}$, is positive. Note that

$$\underline{\lambda} = \min_{\|\mathbf{a}\|=1} \mathbf{a}^T \,\mathrm{COV}\, \mathbf{a}, \tag{S5.48}$$

where $\mathbf{a} = (a_1, \ldots, a_{d_0})^T$. By the definition of COV in (4.5),

$$\begin{aligned} \mathbf{a}^T \,\mathrm{COV}\, \mathbf{a} &= \left\{ \sum_{i=1}^{d_0} a_i D^{\alpha_i} K(z_i, X) \right\} (K + \lambda n I)^{-2} \left\{ \sum_{i=1}^{d_0} a_i D^{\alpha_j} K(X, z_j) \right\} \\ &= g_a^T(X)(K + \lambda n I)^{-2} g_a(X), \end{aligned} \tag{S5.49}$$

for $g_a = \sum_{i=1}^{d_0} a_i D^{\alpha_j} K(\cdot, z_j)$. Because (α_i, z_i)'s are distinct, $D^{\alpha_j} K(\cdot, z_j)$'s are linearly independent, and therefore $g_a \neq 0$ for any $\|a\| = 1$. Because $\|g\|_{L_2}/\|g_a\|_{\mathcal{H}}$, as a function of a, is continuous over the unit sphere $\{a : \|a\| = 1\}$, $\|g\|_{L_2}/\|g_a\|_{\mathcal{H}}$ has an attainable infimum, denoted as $\underline{r} > 0$. Now for any $\epsilon > 0$, let C_ϵ and Ξ_ϵ be defined as in Assumption 3. Then for $n > (C_\epsilon/\underline{r})^{d/m}$, $\|g_a\|_{L_2} > C_\epsilon n^{-m/d} \|g_a\|_{\mathcal{H}}$ for all $\|a\| = 1$. Then by (3.8), on the event Ξ_ϵ, $\|g_a\|_{L_2} \leq C_1 \|g_a\|_n$ for all $\|a\| = 1$. This shows that $\mathbf{a}^T \,\mathrm{COV}\, \mathbf{a} \neq 0$ for all $\|a\| = 1$, and implies that COV is invertible.

Now assume that α_i's are homogenous and denote $k := |\alpha_i|$. Let us establish a lower bound of $\underline{\lambda}_n$ for the future use. By (S5.48) and (S5.49), it suffices to find a lower bound of the variance term of $\langle \hat{f} - f, g_a \rangle_{\mathcal{H}}$. In order to invoke Theorem 3.2, we need to verify Assumption 4 for g_a. The idea is similar to the proof of Proposition S5.1 but with more involved details.

Without loss of generality, we assume that $a_i \neq 0$ for each i.

First, we group the triads (a_i, α_i, z_i)'s based on the value of z_i: each group has a common z_i, and different groups have different z_i. Denote the groups by $\mathcal{G}_1, \ldots, \mathcal{G}_J$. Again, each group consists of triads (a_i, α, z_i) with the same z_i value. Then the linear functional $\langle g_a, \cdot \rangle_{\mathcal{H}}$ can be rewritten as

$$\langle g_a, v \rangle_{\mathcal{H}} = \sum_{j=1}^{J} \sum_{(a_i, \alpha_i, z_i) \in \mathcal{G}_j} a_i D^{\alpha_i} v(z_i). \tag{S5.50}$$

The goal is to construct v under the condition $\|v\|_{\mathcal{H}} / \|v\|_{L_2} \leq R$, such that $\langle g_a, v \rangle_{\mathcal{H}} / \|v\|_{L_2}$ reaches the optimal order of magnitude. For a moment, suppose that, for each $j = 1, \ldots, J$, we can find a function $B_j \in C^{\infty}(\mathbb{R}^d)$, such that

$$\sum_{(a_i, \alpha_i, z_i) \in \mathcal{G}_j} a_i D^{\alpha_i} B_j(0) \geq \frac{1}{2} \sum_{(a_i, \alpha_i, z_i) \in \mathcal{G}_j} a_i^2, \tag{S5.51}$$

and B_j is supported in the unit ball of $\mathbb{R}^d$. Now define $v_j(x) := B_j((x - z_j)/\rho)$ for $\rho \in (0, 1)$ and $v := \sum_{j=1}^{J} v_j$. Clearly, if ρ is sufficiently small, v_j's have disjoint supports, and thus

$$\|v\|_{L_2}^2 = \sum_{i=1}^{J} \|v_j\|_{L_2}^2, \quad \|v\|_{H^m}^2 = \sum_{i=1}^{J} \|v_j\|_{H^m}^2.$$

By the calculations in the proof of Proposition S5.1, we have

$$\begin{aligned} D^{\alpha_i} v_j(z_i) &= \rho^{-k} D^{\alpha_i} B_j(0), \\ \|v\|_{L_2} &= \rho^{d/2} \left\{ \sum_{i=1}^{J} \|B_j\|_{L_2}^2 \right\}^{1/2} =: \rho^{d/2} A_1, \end{aligned}$$

$$\|v\|_{H^m} \leq \rho^{-(m-d/2)} \left(\sum_{i=1}^{J} \|B_j\|_{H^m}^2 \right)^{1/2} := \rho^{-(m-d/2)} A_2.$$

On the other hand, we have

$$\begin{aligned} \frac{\langle g_a, v \rangle_{\mathcal{H}}}{\|v\|_{L_2}} &= \frac{\sum_{j=1}^{J} \sum_{(a_i,\alpha_i,z_i)\in\mathcal{G}_j} a_i D^{\alpha_i} v_j(z_i)}{\rho^{d/2} A_1} \\ &= \frac{\rho^{-k} \sum_{j=1}^{J} \sum_{(a_i,\alpha_i,z_i)\in\mathcal{G}_j} a_i D^{\alpha_i} B_j(0)}{\rho^{d/2} A_1} \\ &\geq \frac{1}{2} \rho^{-k-d/2} \sum_{i=1}^{n} a_i^2 / A_1 = \frac{1}{2A_1} \rho^{-k-d/2}. \end{aligned}$$

Setting $R = \rho^m$ implies Assumption 4 with $\tau = 1 - \frac{2k+d}{2m}$.

Now we prove the existence of B_j's subject to (S5.51) and the compact supportedness condition. A simple configuration that fulfills (S5.51) is to ensure

$$D^{\alpha_i} B_j(0) = a_i, \text{ whenever } (a_i, \alpha_i, z_i) \in \mathcal{G}_j. \tag{S5.52}$$

Building a function B_j subject to (S5.52) can be done by a multivariate Hermite interpolation. For example, we can use kring Morris et al. (1993); Zongmin (1992) with a Gaussian kernel to produce a function in $C^\infty(\mathbb{R}^d)$ that satisfies (S5.52). Denote such a function by B_{j1}. To introduce the compact supportedness, define

$$B_{j2}(x) := \begin{cases} B_{j1}(x) & \text{if } \|x\| \leq 1/2, \\ 0 & \text{otherwise.} \end{cases}$$

Then we smooth B_{j2} via a convolution. Choose $\varphi \in C^\infty(\mathbb{R}^d)$ supported in the unit ball of $\mathbb{R}^d$ with $\int_{\mathbb{R}^d} \varphi(x)dx = 1$. Let $\varphi_\rho := \rho^{-d}\varphi(\cdot/\rho)$, and $B_{j3}(x;\rho) = \int_{\mathbb{R}^d} B_{j2}(x-t)\varphi_\rho(t)dt$ for small ρ. Then $B_{j3}(\cdot;\rho) \in C^\infty(\mathbb{R}^d)$, $B_{j3}(\cdot;\rho)$ is supported in the unit ball of $\mathbb{R}^d$, and $\lim_{\rho\downarrow 0} D^{\alpha_i} B_{j3}(0;\rho) = D^{\alpha_i} B_{j1}(0)$. Therefore, we can set $B_j = B_{j3}(\cdot;\rho)$ for sufficiently small ρ such that (S5.52) is satisfied.

To verify Assumption 2, we use the interpolation inequality (4.20) to show that

$$\begin{aligned}\left|\sum_{i=1}^{d_0} a_i D^{\alpha_i} g(z_i)\right| &\leq \left(\sum_{i=1}^{d_0} a_i^2\right)^{1/2} \left\{\sum_{i=1}^{d_0} [D^{\alpha_i} g(z_i)]^2\right\}^{1/2} \\ &\leq d_0^{\frac{1}{2}} A \|v\|_{L_2}^{1-\frac{2k+d}{2m}} \|v\|_{H^m}^{\frac{2k+d}{2m}},\end{aligned}$$

where A is given in (4.20). Thus we have verified Assumption 2 with $\delta = 1 - \frac{2k+d}{2m}$. Since $\delta = \tau$, we are ready to invoke Corollary 3.3 to obtain that $\mathbf{a}^T \operatorname{COV} \mathbf{a} \geq A_3 n^{-1} \lambda^{-\frac{2k+d}{2m}}$, for some $A_3 > 0$. Note that the constants we established for Assumptions 2 and 4 are independent of $\mathbf{a}$. Thus A_3 is also independent of $\mathbf{a}$, which implies that on the event Ξ_ϵ,

$$\underline{\lambda} \geq A_3 n^{-1} \lambda^{-\frac{2k+d}{2m}}. \tag{S5.53}$$

Next, we move to the central limit theorem. We shall use the notation similar to the proof of Lemma S5.3, by reinstating the subscript n. Again, we assume that X_n is fixed, which is equivalent to conditioning on X_n.

Denote $\left(u_{1n}^{(i)}, \ldots, u_{nn}^{(i)}\right)^T := \{K(X_n, X_n) + \lambda_n n I_n\}^{-1} D^{\alpha_i} K(X_n, z_i)$. Define

$$\mathbf{u}_{n,i} := \left(u_{in}^{(1)}, \ldots, u_{in}^{(d_0)}\right)^T.$$

Then

$$\begin{pmatrix} D^{\alpha_1} K(z_1, X) \\ \vdots \\ D^{\alpha_{d_0}} K(z_{d_0}, X) \end{pmatrix} \{K(X_n, X_n) + \lambda_n n I\}^{-1} E = \sum_{i=1}^{n} \mathbf{u}_{n,i} e_i.$$

We now use a version of the multivariate Lindeberg central limit theorem Hansen (2022), which ensures the desired result provided that

$$\lim_{n\to\infty} \frac{1}{\underline{\lambda}_n} \sum_{i=1}^{n} \mathbb{E}\left[\|\mathbf{u}_{n,i}\|^2 e_i^2 \mathbf{1}_{\{\|\mathbf{u}_{n,i}\|^2 e_i^2 \geq \varepsilon^2 \underline{\lambda}_n\}}\right] = 0, \tag{S5.54}$$

for each $\varepsilon > 0$, where $\underline{\lambda}_n$ denotes the minimum eigenvalue of COV_n. In view of (S5.53), on the event $\Xi_{\epsilon,n}$

$$\begin{aligned} & \frac{1}{\underline{\lambda}_n} \sum_{i=1}^{n} \mathbb{E}\left[\|\mathbf{u}_{n,i}\|^2 e_i^2 \mathbf{1}_{\{\|\mathbf{u}_{n,i}\|^2 e_i^2 \geq \varepsilon^2 \underline{\lambda}_n\}}\right] \\ \leq \; & A_3^{-1} n \lambda_n^{\frac{2k+d}{2m}} \sum_{i=1}^{n} \mathbb{E}\left[\|\mathbf{u}_{n,i}\|^2 e_i^2 \mathbf{1}_{\left\{\|\mathbf{u}_{n,i}\|^2 e_i^2 \geq \varepsilon^2 A_3 n^{-1} \lambda_n^{-\frac{2k+d}{2m}}\right\}}\right]. \end{aligned} \tag{S5.55}$$

On the other hand, let $g_j = D^{\alpha_j} K(z_j, X)$ for $j = 1, \ldots, d_0$. Then, on $\Xi_{\epsilon,n}$, we have

$$\begin{aligned} \|\mathbf{u}_{n,i}\|^2 &= \sum_{j=1}^{d_0} \left(u_{in}^{(j)}\right)^2 = \lambda_n^{-2} n^{-2} \sum_{j=1}^{d_0} \{\hat{g}_{jn}(x_{in}) - g_j(x_{in})\}^2 \\ &\leq \lambda_n^{-2} n^{-2} \sum_{j=1}^{d_0} \|\hat{g}_{jn} - g_j\|_{L_\infty}^2 \end{aligned}$$

$$\leq \quad A_4 n^{-2} \lambda_n^{-\frac{k+d}{m}}, \tag{S5.56}$$

where the second equality follows from (S5.24); and the last inequality follows from (S5.46) and $A_4 > 0$ is a constant independent of n and λ_n. Combining (S5.55) and (S5.56), we obtain that on $\Xi_{\epsilon,n}$,

$$\begin{aligned}
& \frac{1}{\underline{\lambda}_n} \sum_{i=1}^{n} \mathbb{E}\left[\|\mathbf{u}_{n,i}\|^2 e_i^2 \mathbf{1}_{\{\|\mathbf{u}_{n,i}\|^2 e_i^2 \geq \varepsilon^2 \underline{\lambda}_n\}} \right] \\
\leq \quad & A_3^{-1} n \lambda_n^{\frac{2k+d}{2m}} \sum_{i=1}^{n} \mathbb{E}\left[\|\mathbf{u}_{n,i}\|^2 e_i^2 \mathbf{1}_{\left\{ e_i^2 \geq \varepsilon^2 A_3 A_4^{-1} n \lambda_n^{\frac{d}{2m}} \right\}} \right] \\
= \quad & A_3^{-1} n \lambda_n^{\frac{2k+d}{2m}} \mathbb{E}\left[e_1^2 \mathbf{1}_{\left\{ e_1^2 \geq \varepsilon^2 A_3 A_4^{-1} n \lambda_n^{\frac{d}{2m}} \right\}} \right] \sum_{i=1}^{n} \|\mathbf{u}_{n,i}\|^2.
\end{aligned} \tag{S5.57}$$

Note that on $\Xi_{\epsilon,n}$,

$$\begin{aligned}
\sum_{i=1}^{n} \|\mathbf{u}_{n,i}\|^2 \quad &= \quad \sum_{j=1}^{d_0} \sum_{i=1}^{n} \left(u_{in}^{(j)} \right)^2 = \sum_{j=1}^{d_0} \lambda_n^{-2} n^{-2} \sum_{i=1}^{n} \{ \hat{g}_{jn}(x_{in}) - g(x_{in}) \}^2 \\
&= \quad \lambda_n^{-2} n^{-1} \sum_{j=1}^{d_0} \|\hat{g}_{jn} - g_j\|_n^2 \\
&\leq \quad A_5 n^{-1} \lambda_n^{-\frac{2k+d}{2m}},
\end{aligned} \tag{S5.58}$$

where the second equality follows from (S5.24); and the inequality follows from Lemma S5.1 and $A_5 > 0$ is a constant independent of n and λ_n. Combining (S5.57) and (S5.58) yields that, on $\Xi_{\epsilon,n}$,

$$\frac{1}{\underline{\lambda}_n} \sum_{i=1}^{n} \mathbb{E}\left[\|\mathbf{u}_{n,i}\|^2 e_i^2 \mathbf{1}_{\{\|\mathbf{u}_{n,i}\|^2 e_i^2 \geq \varepsilon^2 \underline{\lambda}_n\}} \right] \leq A_3^{-1} A_5 \mathbb{E}\left[e_1^2 \mathbf{1}_{\left\{ e_1^2 \geq \varepsilon^2 A_3 A_4^{-1} n \lambda_n^{\frac{d}{2m}} \right\}} \right],$$

which tends to zero due to the condition $\lim_{n\to\infty} n\lambda_n^{\frac{d}{2m}} = \infty$ and the dominated convergence theorem.

Hence we have proven the Lindeberg condition, where the convergence is in probability. It can be argued that, similar to that in the proof of Lemma S5.3, such a condition still ensures the central limit theorem. □

Proof of Proposition 4.1. Consider the linear functional $l(v) : \mathcal{H} \mapsto \mathbb{R}$ with $l(v) = \langle v, g\rangle_{\mathcal{H}}$. Assumption 1 ensures that $\mathcal{H}$ is dense in L_2, which, together with the condition

$$\sup_{v\in\mathcal{H}} \frac{l(v)}{\|v\|_{L_2}} < \infty,$$

implies that l can be continuously and uniquely extended to L_2. Then the Riesz representation theorem asserts there exists a unique $h \in L_2$, such that $l(v) = \langle v, h\rangle_{L_2}$. □

Proof of Proposition S2.1. Under Assumption 1, L_2 is dense in $\mathcal{H}$. Consequently, we know that 1) $\rho_i > 0$ for each i, and 2) $\{\eta_i\}_{i=1}^{\infty}$ forms an orthonormal basis of L_2. Let $v = \sum_{i=1}^{s} a_i\eta_i$. Theorem 10.29 of Wendland (2004) shows the representation of the RKHS inner product as

$$\left\langle \sum_{i=1}^{\infty} a_i\eta_i, \sum_{i=1}^{\infty} c_i\eta_i \right\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} \frac{a_i c_i}{\rho_i}.$$

This implies

$$\begin{aligned} |\langle g, v\rangle_{\mathcal{H}}| = \left|\sum_{i=1}^{\infty} \frac{c_i a_i}{\rho_i}\right| &\leq \left(\sum_{i=1}^{\infty} \frac{c_i^2}{\rho_i^{1+\kappa}}\right)^{1/2} \left(\sum_{i=1}^{\infty} \frac{a_i^2}{\rho_i^{1-\kappa}}\right)^{1/2} \\ &= \|g\|_{\mathcal{H},\kappa} \left\{\sum_{i=1}^{\infty} (a_i^2)^{\kappa} \left(\frac{a_i^2}{\rho_i}\right)^{1-\kappa}\right\}^{1/2} \end{aligned}$$

$$\begin{aligned} &\leq \|g\|_{\mathcal{H},\kappa} \left(\sum_{i=1}^{\infty} a_i^2 \right)^{\frac{\kappa}{2}} \left(\sum_{i=1}^{\infty} \frac{a_i^2}{\rho_i} \right)^{\frac{1-\kappa}{2}} \\ &= \|g\|_{\mathcal{H},\kappa} \|v\|_{L_2}^{\kappa} \|v\|_{\mathcal{H}}^{1-\kappa}, \end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality, the second inequality follows from the Hölder's inequality with $(p, q) = (\frac{1}{\kappa}, \frac{1}{1-\kappa})$.

□

S5.11 Proofs for Uniform Bound in Section 5.1

Proof of Theorem S2.3. The main idea is to invoke Dudley's theorem (Vaart and Wellner, 2000; Vershynin, 2018), which states that a zero-mean sub-Gaussian process with respect to a semi-metric d_Z, i.e., a stochastic process $Z(x)$ satisfying $\mathbb{E} \exp\left[\theta\{Z(x_1) - Z(x_2)\}\right] \leq \exp\{\vartheta^2 d_Z^2(x_1, x_2)/2\}$ for all possible ϑ, x_1, x_2, is subject to the following uniform bound

$$\mathbb{E}\left[\sup_{t \in \mathcal{T}} |Z(t)|\right] \leq \mathbb{E}|Z(t_0)| + A \int_0^D \sqrt{\log N(\epsilon, \mathcal{T}, d_Z)} d\epsilon, \tag{S5.59}$$

for any $t_0 \in \mathcal{T}$, where D is the d_Z-diameter of $\mathcal{T}$, and A is a universal constant.

Denote $g_x(\cdot) = D^\alpha K(\cdot, x)$. Then by (S5.24), we have

$$D^\alpha \hat{f}(x) - D^\alpha f(x) = -\lambda^{-1} n^{-1} \sum_{i=1}^{n} (\hat{g}_x - g_x)(x_i) e_i := Z(x). \tag{S5.60}$$

Because e_i is ς^2-sub-Gaussian, conditional on X, $Z(x)$ is a zero-mean

sub-Gaussian process with respect to the semi-metric

$$d_\Omega(x_1, x_2) = \varsigma\lambda^{-1}n^{-\frac{1}{2}}\|\hat{g}_{x_1} - g_{x_1} - \hat{g}_{x_2} + g_{x_2}\|_n. \tag{S5.61}$$

Then we can find an upper bound of $d_\Omega(x_1, x_2)$ by using the triangle inequality

$$d_\Omega(x_1, x_2) \leq \varsigma\lambda^{-1}n^{-\frac{1}{2}}\left(\|\hat{g}_{x_1} - g_{x_1}\|_n + \|\hat{g}_{x_2} + g_{x_2}\|_n\right).$$

Thus, by Lemma S5.1, on the event Ξ_ϵ defined in Assumption 3, we have

$$d_\Omega(x_1, x_2) \leq C_\Omega\varsigma\lambda^{-1}n^{-\frac{1}{2}}\lambda^{\frac{1+1-\frac{d+2|\alpha|}{2m}}{2}} = C_\Omega\varsigma n^{-\frac{1}{2}}\lambda^{-\frac{d+2|\alpha|}{4m}},$$

for some constant $C_\Omega > 0$. On the other hand, let $g_{x_1,x_2} := g_{x_1} - g_{x_2}$. Because KRR is linear, we have $\hat{g}_{x_1,x_2} = \hat{g}_{x_1} - \hat{g}_{x_2}$, and thus

$$d_\Omega(x_1, x_2) = \varsigma\lambda^{-1}n^{-\frac{1}{2}}\|\hat{g}_{x_1,x_2} - g_{x_1,x_2}\|_n. \tag{S5.62}$$

Now we verify Assumption 2 for g_{x_1,x_2}. Note that, for $v \in H^m$

$$\langle g_{x_1,x_2}, v\rangle_{\mathcal{H}} = D^\alpha v(x_1) - D^\alpha v(x_2).$$

Clearly, $D^\alpha v \in H^{m-|\alpha|}$. Noting $m > d/2 + |\alpha|$, we can find $m > m' > d/2 + |\alpha|$. Because $m' - |\alpha| > d/2$, the Sobolev embedding theorem (see, e.g., Theorem 4.47 of Demengel (2012)) claims the embedding relationship $H^{m'-|\alpha|} \hookrightarrow C^{0,\tau}$ for $\tau := \min(m' - |\alpha| - d/2, 1)$, where $C^{0,\tau}$ denotes the Hölder space with the norm

$$\|f\|_{C^{0,\tau}} := \sup_{x\neq x'}\frac{|f(x) - f(x')|}{\|x - x'\|^\tau}.$$

Thus,

$$\langle g_{x_1,x_2}, v\rangle_{\mathcal{H}} \leq \|D^\alpha v\|_{C^{0,\tau}}\|x_1-x_2\|^\tau \leq \|D^\alpha v\|_{H^{m'-|\alpha|}}\|x_1-x_2\|^\tau \leq \|v\|_{H^{m'}}\|x_1-x_2\|^\tau.$$

Next we use the interpolation inequality

$$\|v\|_{H^{m'}} \leq A\|v\|_{L_2}^{1-\frac{m'}{m}}\|v\|_{H^m}^{\frac{m'}{m}}.$$

This implies Assumption 2 for $C_{g_{x_1,x_2}} = A\|x_1 - x_2\|^\tau$ and $\delta = 1 - \frac{m'}{m}$. Thus, by Lemma S5.1 and (S5.62), on the event Ξ_ϵ defined in Assumption 3, we find

$$d_\Omega(x_1, x_2) \leq C\varsigma\lambda^{-1}n^{-\frac{1}{2}}\lambda^{\frac{1+1-\frac{m'}{m}}{2}}\|x_1 - x_2\|^\tau = C_1\varsigma n^{-\frac{1}{2}}\lambda^{-\frac{m'}{2m}}\|x_1 - x_2\|^\tau,$$

for some $C_1 > 0$. Using the fact that Ω is a d-dimensional bounded region, we obtain that

$$N(\varepsilon, \Omega, d_\Omega) \leq N\left(\left\{\epsilon/(C_1\varsigma n^{-\frac{1}{2}}\lambda^{-\frac{m'}{2m}})\right\}^{1/\tau}, \Omega, \|\cdot\|\right).$$

Thus, by Lemma 4.1 of Pollard,

$$\log N(\varepsilon, \Omega, d_\Omega) \leq d\log\left\{16D_\Omega\left(\frac{C_1\varsigma n^{-\frac{1}{2}}\lambda^{-\frac{m'}{2m}}}{\epsilon}\right)^{1/\tau} + 1\right\},$$

where D_Ω is the Euclidean diameter of Ω.

Therefore, the integral in Dudley's theorem (S5.59) has the upper bound

$$\int_0^{C_\Omega n^{-\frac{1}{2}}\lambda^{-\frac{d+2|\alpha|}{4m}}} \sqrt{\log N(\varepsilon, \Omega, d_\Omega)}d\varepsilon$$

$$
\begin{aligned}
&\lesssim \int_0^{C_\Omega \varsigma n^{-\frac{1}{2}} \lambda^{-\frac{d+2|\alpha|}{4m}}} \sqrt{\log\left\{16 D_\Omega \left(\frac{C_1 \varsigma n^{-\frac{1}{2}} \lambda^{-\frac{m'}{2m}}}{\epsilon}\right)^{1/\tau} + 1\right\}} d\varepsilon \\
&= \varsigma n^{-\frac{1}{2}} \lambda^{-\frac{d+2|\alpha|}{4m}} \int_0^{C_\Omega} \sqrt{\log\left\{16 D_\Omega \left(\frac{C_1 \lambda^{\frac{d+2|\alpha|}{4m} - \frac{m'}{2m}}}{\epsilon}\right)^{1/\tau} + 1\right\}} d\varepsilon \\
&\lesssim \varsigma n^{-\frac{1}{2}} \lambda^{-\frac{d+2|\alpha|}{4m}} \sqrt{\log\left(\frac{C}{\lambda}\right)},
\end{aligned}
$$

for some $C > 0$; where the last inequality is based on algebraic calculations similar to (33)-(36) in Tuo and Wang (2020). This term would dominate the first term of (S5.59), which is given by Theorem 4.2. Hence, we prove the desired result as $\mathbb{P}(\Xi_\epsilon)$ tends to one as $n \to \infty$. □

Proof of Theorem S2.4. Let $Z(x)$ be the same as (S5.60). We can see that conditional on X, $Z(x)$ is a zero-mean Gaussian process with the natural distance

$$
d_\Omega(x_1, x_2) := \left[\mathbb{E}_E\{Z(x_1) - Z(x_2)\}^2\right]^{1/2} = \sigma \lambda^{-1} n^{-\frac{1}{2}} \|\hat{g}_{x_1} - g_{x_1} - \hat{g}_{x_2} + g_{x_2}\|_n.
$$

Now the idea is to invoke the Sudakov's lower bound (Vershynin, 2018), which states that

$$
\mathbb{E}_E \left[\sup_{x \in \Omega} |Z(x)|\right] \gtrsim \sup_{\varepsilon > 0} \varepsilon \sqrt{\log N(\varepsilon, \Omega, d_\Omega)}. \tag{S5.63}
$$

The boundary effect of Ω may cause some problems to our proof. So we define Ω' as a subset of Ω such that each $x \in \Omega'$ is distant from the boundary

of Ω by at least η in the Euclidean distance, where η is sufficiently small such that Ω' contains an open set. Because $\sup_{x\in\Omega}|Z(x)| \geq \sup_{x\in\Omega'}|Z(x)|$, we will work only on a lower bound of $\sup_{x\in\Omega'}|Z(x)|$.

Let $C_2 > 0$ be a constant to be determined, and let $M := \lceil (C_2/\lambda)^{\frac{d}{2m}} \rceil$. In view of the lower bound for the covering number of a Euclidean compact set (Pollard, 1990), when $M \geq 2$, for any M points $\{\xi_1, \ldots, \xi_M\} \subset \Omega'$, there exists two points, say $\{\xi_1, \xi_2\}$ without loss of generality, such that $\|\xi_1 - \xi_2\| \leq CM^{-1/d}$ for some constant C depending only on Ω'. Because $\lambda \to 0$, we shall assume that $CM^{-1/d} < 2C(\lambda/C_2)^{1/(2m)} < \eta$ without loss of generality.

Now let us consider $d_\Omega(\xi_1, \xi_2)$. The goal is to show that

$$d_\Omega(\xi_1, \xi_2) \gtrsim \sigma n^{-\frac{1}{2}} \lambda^{-\frac{d+2|\alpha|}{4m}}. \tag{S5.64}$$

If (S5.64) is true, we essentially prove that for $\varepsilon \leq C_1 \sigma n^{-\frac{1}{2}} \lambda^{-\frac{d+2|\alpha|}{4m}}$ for some constant $C_1 > 0$, we have $N(\varepsilon, \Omega', d_\Omega) \geq M \geq (C_2/\lambda)^{d/2m}$, which, together with (S5.63), imples

$$\mathbb{E}_E \left[\sup_{x\in\Omega'} |Z(x)| \right] \gtrsim C_1 \sigma n^{-\frac{1}{2}} \lambda^{-\frac{d+2|\alpha|}{4m}} \sqrt{\frac{d}{2m} \log\left(\frac{C_2}{\lambda}\right)}. \tag{S5.65}$$

To prove (S5.64), we use Lemma S5.2. We have to be mindful that the constants in Lemma S5.2 may depend on n as ξ_1, ξ_2 are dependent on n. This necessities a closer look at the proof of Theorem 3.2, which Lemma S5.2

primarily relies on. First, we note that the interpolation inequality (4.4) gives $\delta = 1 - \frac{d+2|\alpha|}{2m}$, independent of n. The constants from Assumption 3 are also constants. However, constants in Assumption 4 should be examined carefully. Our goal is to ensure Assumption 4 with $\tau = \delta = 1 - \frac{d+2|\alpha|}{2m}$. To this end, we consider the function constructed in Proposition S5.1. In the proof of Proposition S5.1, we have constructed a function ϕ_ρ for each $\rho < CM^{-1/d} < \eta$ that satisfies the following properties:

1. $\phi_\rho(\xi) = 0$ if $\|\xi - \xi_1\| \geq \rho$;
2. $D^\alpha \phi_\rho(\xi_1) = 1$;
3. $\|\phi_\rho\|_{L_2} = C_3 \rho^{d/2}$ for some constant $C_3 > 0$ depending only on m.
4. $\|\phi_\rho\|_{H^m} / \|\phi_\rho\|_{L_2} \leq C_4 \rho^{-m}$ for some $C_4 > 0$ depending only on m.

Hence, we have

$$\langle \phi_\rho, g_{x_1} - g_{x_2} \rangle_{\mathcal{H}} = \phi_\rho(x_1) - \phi_\rho(x_2) = 1,$$

whenever $\rho < CM^{-1/d}$. This ensures Assumption 4 for $g = g_{x_1} - g_{x_2}$ with $\tau = \delta$ independent of n, $C_0 = C_3$ independent of n, $R_0 = C'(CM^{-1/d})^{-m} \geq 2^{-m} C^{-m} C' (C_2/\lambda)^{1/2}$. Because only R_0 depends on n (or λ), by examining the proof of Theorem 3.2, we can see that we only need to ensure (S5.38), that is,

$$2^{-m} C^{-m} C' (C_2/\lambda)^{1/2} \leq 4^{-1/\delta} C_0^{-1/\delta} C_1^{-1} C_g^{-1/\delta} \lambda^{-1/2}, \tag{S5.66}$$

for the validity of Theorem 3.2 and consequently, Lemma S5.2. and (S5.66) can be ensured provided that C_2 is sufficiently small. Now we are ready to use Lemma S5.2, which states that under the event Ξ_ϵ, (S5.64) is true. Therefore we have proven (S5.65), under Ξ_ϵ.

Because ϵ can be chosen arbitrarily small, the desired result is a direct consequence of the above statement together with the concentration inequality of Gaussian processes (Vershynin, 2018). □

S5.12 Proof for Nonlinear Problem in Section 5.2

Proof of Theorem S2.5. It is well known that,

$$\sup_{x\in\Omega} |D^\alpha \hat{f}(x) - D^\alpha f(x)| = o_{\mathbb{P}}(1), \tag{S5.67}$$

for all $\alpha \in \mathbb{N}^d$ with $|\alpha| = 0, 1, 2$ under the condition $\lambda \sim n^{-1}$; see van de Geer (2000). The uniform convergence of $\hat{f} - f$ implies the consistency of $\hat{f}_{\min}$ and $\hat{x}_{\min}$.

Next, we study the rates of convergence of the estimators. Because $x_{\min}$ and $\hat{x}_{\min}$ minimize f and $\hat{f}$, respectively, we have

$$\begin{aligned} 0 &= \frac{\partial \hat{f}}{\partial x}(\hat{x}_{\min}) = \frac{\partial \hat{f}}{\partial x}(x_{\min}) + \frac{\partial^2 \hat{f}}{\partial x \partial x^T}(x^*)(\hat{x}_{\min} - x_{\min}) \\ &= \frac{\partial(\hat{f} - f)}{\partial x}(x_{\min}) + \frac{\partial^2 \hat{f}}{\partial x \partial x^T}(x^*)(\hat{x}_{\min} - x_{\min}) \end{aligned} \tag{S5.68}$$

where x^* lies between $x_{\min}$ and $\hat{x}_{\min}$. The consistency of $\hat{x}_{\min}$ and (S5.67)

implies that $\frac{\partial^2 \hat{f}}{\partial x \partial x^T}(x^*)$ converges weakly to H, which, together with the condition that H is positive definite, implies that $\frac{\partial^2 \hat{f}}{\partial x \partial x^T}(x^*)$ is invertible with probability tending to one. Therefore, (S5.68) implies

$$\hat{x}_{\min} - x_{\min} = -\left\{ \frac{\partial^2 \hat{f}}{\partial x \partial x^T}(x^*) \right\}^{-1} \frac{\partial (\hat{f} - f)}{\partial x}(x_{\min}). \tag{S5.69}$$

By Theorem 4.2, under the optimal choice $\lambda \asymp n^{-1}$, $\|\frac{\partial (\hat{f}-f)}{\partial x}(x_{\min})\| = O_{\mathbb{P}}(n^{-\frac{1}{2}+\frac{d+2}{4m}})$. Thus $\|\hat{x}_{\min} - x_{\min}\| = O_{\mathbb{P}}(n^{-\frac{1}{2}+\frac{d+2}{4m}})$.

To show the rate of convergence of $f(\hat{x}_{\min}) - f(x_{\min})$, we use the Taylor expansion of f at $x_{\min}$ to obtain

$$f(\hat{x}_{\min}) - f(x_{\min}) = (\hat{x}_{\min} - x_{\min})^T \frac{\partial^2 f}{\partial x \partial x^T}(x_*)(\hat{x}_{\min} - x_{\min}),$$

for some x_* lying between $x_{\min}$ and $\hat{x}_{\min}$. Again, we have that $\frac{\partial^2 f}{\partial x \partial x^T}(x_*)$ converges to H weakly, and therefore $f(\hat{x}_{\min}) - f(x_{\min}) = O_{\mathbb{P}}(n^{-1+\frac{d+2}{2m}})$.

By (S5.69) and Theorems 3.5 and 4.3, we have

$$\mathrm{COV}^{-\frac{1}{2}} \frac{\partial^2 \hat{f}}{\partial x \partial x^T}(x^*) \left(\hat{x}_{\min} - x_{\min}\right) \xrightarrow{\mathscr{L}} N(0, I). \tag{S5.70}$$

To prove the desired result, it suffices to show that

$$\left\{ \frac{\partial^2 \hat{f}}{\partial x \partial x^T}(\hat{x}_{\min}) \right\}^{-1} \widehat{\mathrm{COV}}^{\frac{1}{2}} \, \mathrm{COV}^{-\frac{1}{2}} \frac{\partial^2 \hat{f}}{\partial x \partial x^T}(x^*) \xrightarrow{p} I. \tag{S5.71}$$

Define

$$\mathscr{C}(\cdot) := \frac{\partial K}{\partial x}(\cdot, X)\{K(X,X) + \lambda n I\}^{-2} \frac{\partial K}{\partial x^T}(X, \cdot).$$

Because both $\frac{\partial^2 \hat{f}}{\partial x \partial x^T}(\hat{x}_{\min})$ and $\frac{\partial^2 \hat{f}}{\partial x \partial x^T}(x^*)$ converges to H weakly and $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$, (S5.71) is equivalent to

$$\{\mathscr{C}(\hat{x}_{\min})\}^{\frac{1}{2}} \{\mathscr{C}(x_{\min})\}^{-\frac{1}{2}} \xrightarrow{p} I. \tag{S5.72}$$

We shall use the operator norm over $\mathbb{R}^{d \times d}$, given by

$$\|M\|_{op} := \sup_{\mathbf{v} \in \mathbb{R}^d} \frac{\|M\mathbf{v}\|}{\|\mathbf{v}\|},$$

which is equal to the greatest absolute eigenvalue of M. By the sub-multiplicativity of the operator norm,

$$\begin{aligned} & \left\| \{\mathscr{C}(\hat{x}_{\min})\}^{\frac{1}{2}} \{\mathscr{C}(x_{\min})\}^{-\frac{1}{2}} - I \right\|_{op} \\ = & \left\| \left[\{\mathscr{C}(\hat{x}_{\min})\}^{\frac{1}{2}} - \{\mathscr{C}(x_{\min})\}^{\frac{1}{2}} \right] \{\mathscr{C}(x_{\min})\}^{-\frac{1}{2}} \right\|_{op} \\ \leq & \left\| \{\mathscr{C}(\hat{x}_{\min})\}^{\frac{1}{2}} - \{\mathscr{C}(x_{\min})\}^{\frac{1}{2}} \right\|_{op} \left\| \{\mathscr{C}(x_{\min})\}^{-\frac{1}{2}} \right\|_{op}. \end{aligned}$$

Now let $\mathbf{a}$ be a unit eigenvector of $\{\mathscr{C}(\hat{x}_{\min})\}^{\frac{1}{2}} - \{\mathscr{C}(x_{\min})\}^{\frac{1}{2}}$ corresponding to an eigenvalue λ_0 such that $|\lambda_0| = \|\{\mathscr{C}(\hat{x}_{\min})\}^{\frac{1}{2}} - \{\mathscr{C}(x_{\min})\}^{\frac{1}{2}}\|_{op}$. Then, we have

$$\begin{aligned} & \mathbf{a}^T \{\mathscr{C}(\hat{x}_{\min}) - \mathscr{C}(x_{\min})\} \mathbf{a} = \mathbf{a}^T \{\mathscr{C}(\hat{x}_{\min})\}^{\frac{1}{2}} \left[\{\mathscr{C}(\hat{x}_{\min})\}^{\frac{1}{2}} - \{\mathscr{C}(x_{\min})\}^{\frac{1}{2}} \right] \mathbf{a} \\ & + \mathbf{a}^T \left[\{\mathscr{C}(\hat{x}_{\min})\}^{\frac{1}{2}} - \{\mathscr{C}(x_{\min})\}^{\frac{1}{2}} \right] \{\mathscr{C}(x_{\min})\}^{\frac{1}{2}} \mathbf{a} \\ = & \lambda_0 \mathbf{a}^T \left[\{\mathscr{C}(\hat{x}_{\min})\}^{\frac{1}{2}} + [\mathscr{C}(x_{\min})\}^{\frac{1}{2}} \right] \mathbf{a}. \end{aligned}$$

Therefore,

$$\left\|\{\mathscr{C}(\hat{x}_{\min})\}^{\frac{1}{2}} - \{\mathscr{C}(x_{\min})\}^{\frac{1}{2}}\right\|_{op} = \frac{\left|\mathbf{a}^T\{\mathscr{C}(\hat{x}_{\min}) - \mathscr{C}(x_{\min})\}\mathbf{a}\right|}{\mathbf{a}^T\left[\{\mathscr{C}(\hat{x}_{\min})\}^{\frac{1}{2}} + \{\mathscr{C}(x_{\min})\}^{\frac{1}{2}}\right]\mathbf{a}}.$$

Denote $\mathbf{a} = (a_1, \ldots, a_d)^T$, and

$$\mathbf{g}(x) := \sum_{i=1}^{d} a_i \frac{\partial K}{\partial \chi_i}(x, X), \text{ and } h(x) := \mathbf{g}(x)\{K(X,X) + \lambda n I\}^{-2}\mathbf{g}^T(x).$$

Then by the mean value theorem, there exists $\tilde{x}$ between $\hat{x}_{\min}$ and $x_{\min}$, such that

$$\begin{aligned} &\left|\mathbf{a}^T\{\mathscr{C}(\hat{x}_{\min}) - \mathscr{C}(x_{\min})\}\mathbf{a}\right| = |h(\hat{x}_{\min}) - h(x_{\min})| \\ = &\left|\frac{\partial h}{\partial x^T}(\tilde{x})(\hat{x}_{\min} - x_{\min})\right| \le \left\|\frac{\partial h}{\partial x^T}(\tilde{x})\right\| \|\hat{x}_{\min} - x_{\min}\| \end{aligned}$$

By Cauchy-Schwarz inequality,

$$\begin{aligned} \left\|\frac{\partial h}{\partial x^T}(\tilde{x})\right\| &= \left\|2\frac{\partial \mathbf{g}}{\partial x}(\tilde{x})\{K(X,X) + \lambda n I\}^{-2}\mathbf{g}^T(\tilde{x})\right\| \\ &\le 2\left\|\frac{\partial \mathbf{g}}{\partial x}(\tilde{x})\{K(X,X) + \lambda n I\}^{-2}\frac{\partial \mathbf{g}^T}{\partial x}(\tilde{x})\right\|_{op}^{1/2} \cdot \\ &\quad \left[\mathbf{g}(\tilde{x})\{K(X,X) + \lambda n I\}^{-2}\mathbf{g}^T(\tilde{x})\right]^{1/2} \end{aligned}$$

In the proof of Theorem 4.3, we proved the upper and lower bounds of the maximum and the minimum eigenvalues of the covariance matrices. Note that these bounds do not depend on the choice of x, and thus are also true for $\hat{x}_{\min}$ and $\tilde{x}$. Specifically, we have

$$\left\|\frac{\partial \mathbf{g}}{\partial x}(\tilde{x})\{K(X,X) + \lambda n I\}^{-2}\frac{\partial \mathbf{g}^T}{\partial x}(\tilde{x})\right\|_{op} \lesssim n^{-\frac{2m-4-d}{2m}}.$$

$$\lambda_{\min}(\mathscr{C}(\hat{x}_{\min})) \gtrsim n^{-\frac{2m-2-d}{2m}}.$$

$$\lambda_{\min}(\mathscr{C}(x_{\min})) \gtrsim n^{-\frac{2m-2-d}{2m}}.$$

Hence, we obtain

$$\left\| \{\mathscr{C}(\hat{x}_{\min})\}^{\frac{1}{2}} \{\mathscr{C}(x_{\min})\}^{-\frac{1}{2}} - I \right\|_{op} \lesssim n^{-\frac{2m-3-d}{2m}} n^{-\frac{2m-2-d}{4m}} n^{\frac{2m-2-d}{2m}} = n^{-\frac{2m-4-d}{4m}} \to 0,$$

as $n \to \infty$. This completes the proof. □

S5.13 Proof of Theorem S2.2

First, note that similar to (S5.24),

$$(f - \mathbb{E}_E \hat{f})(X) = \lambda n \{K(X,X) + \lambda n I\}^{-1} f(X).$$

Therefore, by Cauchy-Schwarz inequality,

$$\begin{aligned} &\left| \frac{1}{n} \sum_{i=1}^{n} (\mathbb{E}_E \hat{f} - f)(x_i)(h/p_X)(x_i) \right| \\ &= \left| \lambda f^T(X) \{K(X,X) + \lambda n I\}^{-1} (h/p_X)(X) \right| \\ &\leq \lambda \left[f^T(X) \{K(X,X) + \lambda n I\}^{-1} f(X) \right]^{1/2} \cdot \\ &\quad \left[(h/p_X)^T(X) \{K(X,X) + \lambda n I\}^{-1} (h/p_X)(X) \right]^{1/2}. \end{aligned} \tag{S5.73}$$

The two factors on the right hand of the above inequality are related to the noiseless KRR. For notational clarity, we denote the noiseless KRR for f as $\mathscr{A} f$, i.e.,

$$\mathscr{A} f := \operatorname*{argmin}_{v \in \mathcal{H}} \|f - v\|_n^2 + \lambda \|v\|_{\mathcal{H}}^2. \tag{S5.74}$$

Using the solution $\mathscr{A}f = K(\cdot, X)\{K(X,X) + \lambda nI\}^{-1}f(X)$ and by direct calculations, we have

$$\begin{aligned} & \|f - \mathscr{A}f\|_n^2 + \lambda\|\mathscr{A}f\|_{\mathcal{H}}^2 \\ = \; & \lambda^2 n f^T(X)(K(X,X) + \lambda nI)^{-2}f(X) + \\ & \lambda f^T(X)\{K(X,X) + \lambda nI\}^{-1}K(X,X)\{K(X,X) + \lambda nI\}^{-1}f(X) \\ = \; & \lambda f^T(X)\{K(X,X) + \lambda nI\}^{-1}f(X). \end{aligned} \tag{S5.75}$$

Also, (S5.74) implies that

$$\|f - \mathscr{A}f\|_n^2 + \lambda\|\mathscr{A}f\|_{\mathcal{H}}^2 \leq \lambda\|f\|_{\mathcal{H}}^2,$$

which, together with (S5.75), leads to

$$f^T(X)\{K(X,X) + \lambda nI\}^{-1}f(X) \leq \|f\|_{\mathcal{H}}^2. \tag{S5.76}$$

Similarly, we have

$$(h/p_X)^T(X)\{K(X,X) + \lambda nI\}^{-1}(h/p_X)(X) \leq \|h/p_X\|_{\mathcal{H}}^2. \tag{S5.77}$$

Combining (S5.73), (S5.76) and (S5.77) yields the desired result.

S6 Additional Figures for Numerical Results

This section presents additional experimental results that complement the main content.

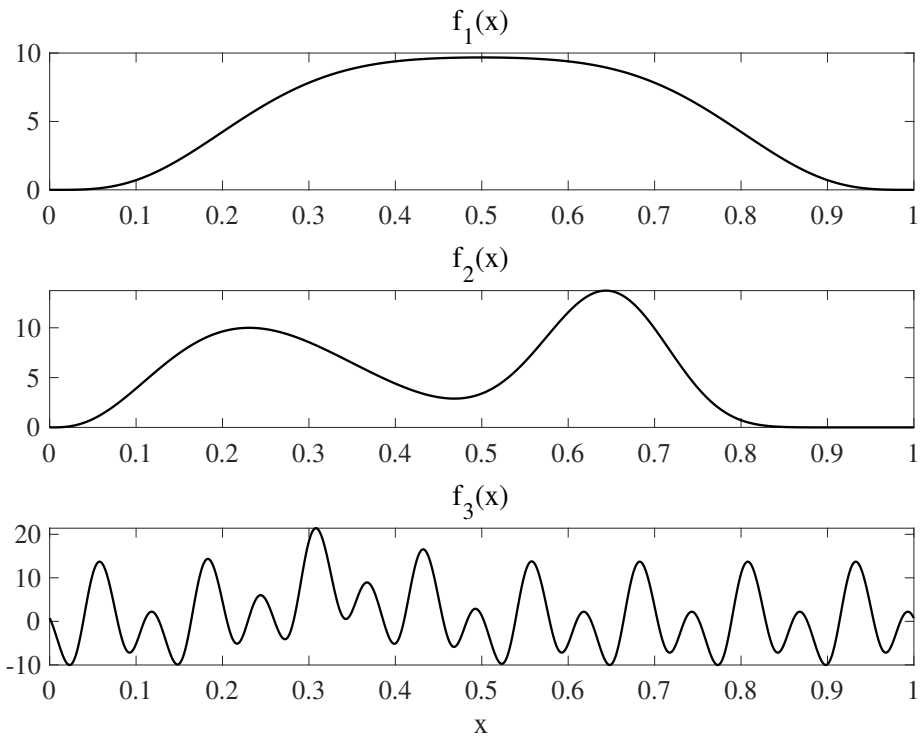

Figure 1: Plots of Three Test Functions f_1, f_2, f_3

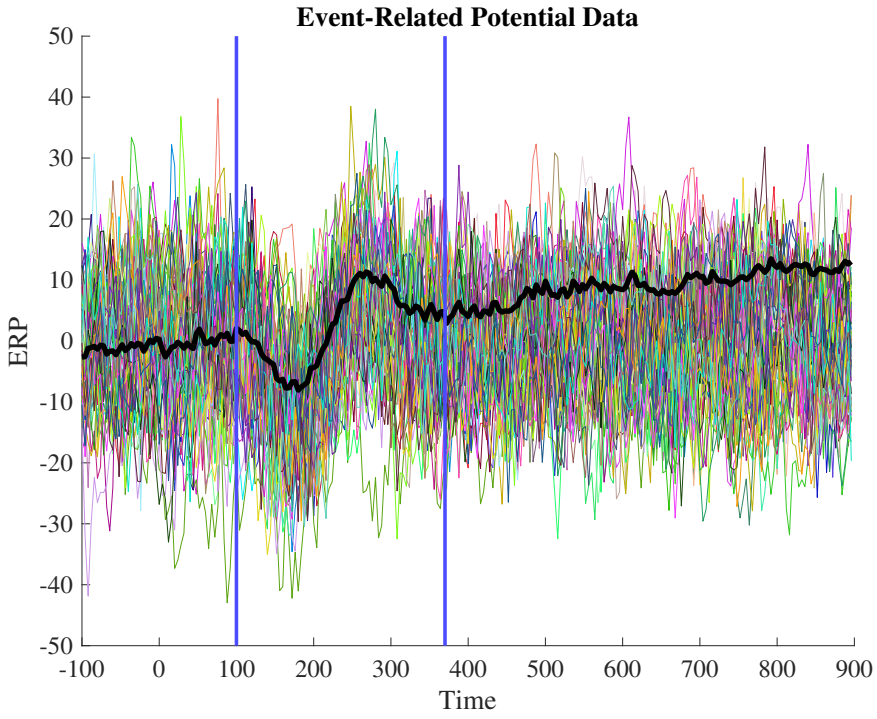

Figure 2: The ERP data consists of 72 individual time series, each representing a single trial, along with the grand average time course computed from all trials. The time window between the two vertical lines indicates the search region.

Figure 1 depicts the test functions used in the experiments in Subsection 6.1, providing a visual reference for the simulation settings.

Figure 2 displays the ERP dataset, consisting of 72 single-trial waveforms and their grand average. The two vertical lines indicate the search window used to estimate the optimal point in our real-data analysis.

Bibliography

Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48.

Balasubramanian, K., Müller, H.-G., and Sriperumbudur, B. K. (2022). Unified rkhs methodology and analysis for functional linear and single-index models. *arXiv preprint arXiv:2206.03975*.

Bauer, F., Pereverzev, S., and Rosasco, L. (2007). On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366.

Bordelon, B., Canatar, A., and Pehlevan, C. (2020). Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR.

Brenner, S. C. and Scott, L. R. (2008). *The Mathematical Theory of Finite Element Methods*, volume 3. Springer.

Cattaneo, M. D. and Farrell, M. H. (2013). Optimal convergence rates,

bahadur representation, and asymptotic normality of partitioning estimators. *Journal of Econometrics*, 174(2):127–143.

Chen, X. and Christensen, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465.

Ciliberto, C., Rosasco, L., and Rudi, A. (2020). A general framework for consistent structured prediction with implicit loss embeddings. *The Journal of Machine Learning Research*, 21(1):3852–3918.

Cortes, C., Mohri, M., and Rostamizadeh, A. (2012). Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13(Mar):795–828.

Cui, H., Loureiro, B., Krzakala, F., and Zdeborová, L. (2021). Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143.

Cui, X., Lin, H., and Lian, H. (2020). Partially functional linear regression in reproducing kernel hilbert spaces. *Computational Statistics & Data Analysis*, 150:106978.

Demengel, F. (2012). *Functional Spaces for the Theory of Elliptic Partial Differential Equations*. Springer.

Dicker, L. H., Foster, D. P., and Hsu, D. (2017). Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators. *Electronic Journal of Statistics*, 11:1022–1047.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 258–267.

Edmunds, D. E. and Triebel, H. (2008). *Function Spaces, Entropy Numbers, Differential Operators*, volume 120. Cambridge University Press.

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.

Gerfo, L. L., Rosasco, L., Odone, F., Vito, E. D., and Verri, A. (2008). Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola,

A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.

Guo, Z.-C., Lin, S.-B., and Zhou, D.-X. (2017). Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7):074009.

Hansen, B. (2022). *Econometrics*. Princeton University Press.

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

Koltchinskii, V. and Li, M. (2023). Functional estimation in high-dimensional and infinite-dimensional models. *arXiv preprint arXiv:2310.16129*.

Li, C.-L., Chang, W.-C., Mroueh, Y., Yang, Y., and Poczos, B. (2019). Implicit kernel learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2007–2016. PMLR.

Li, T. and Zhu, Z. (2020). Inference for generalized partial functional linear regression. *Statistica Sinica*, 30(3):1379–1397.

Li, Y., Zhang, H., and Lin, Q. (2022). On the saturation effect of kernel ridge regression. In *The Eleventh International Conference on Learning Representations*.

Lian, H., Liu, J., and Fan, Z. (2021). Distributed learning for sketched kernel regression. *Neural Networks*, 143:368–376.

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

Lin, S.-B., Guo, X., and Zhou, D.-X. (2017). Distributed learning with regularized least squares. *The Journal of Machine Learning Research*, 18(1):3202–3232.

Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mezard, M., and Zdeborová, L. (2021). Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151.

Lv, S., He, X., and Wang, J. (2023). Kernel-based estimation for partially functional linear model: Minimax rates and randomized sketches. *Journal of Machine Learning Research*, 24(55):1–38.

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

Marteau-Ferey, U., Ostrovskii, D., Bach, F., and Rudi, A. (2019). Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In *Conference on learning theory*, pages 2294–2340. PMLR.

Morris, M. D., Mitchell, T. J., and Ylvisaker, D. (1993). Bayesian design and analysis of computer experiments: Use of derivatives in surface prediction. *Technometrics*, 35(3):243–255.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer Science & Business Media.

Pollard, D. (1990). Empirical processes: Theory and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pages i–86. JSTOR.

Pourkamali-Anaraki, F., Hariri-Ardebili, M. A., and Morawiec, L. (2020). Kernel ridge regression using importance sampling with application to seismic response prediction. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 511–518. IEEE.

Rastogi, A. and Sampath, S. (2017). Optimal rates for the regularized

learning algorithms under general source condition. *Frontiers in Applied Mathematics and Statistics*, 3:3.

Singh, R., Sahani, M., and Gretton, A. (2019). Kernel instrumental variable regression. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 4593–4605.

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

Singh, R., Xu, L., and Gretton, A. (2020). Kernel methods for causal functions: Dose, heterogeneous, and incremental response curves. *arXiv preprint arXiv:2010.04855*.

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

Sun, X., Du, P., Wang, X., and Ma, P. (2018). Optimal penalized function-on-function regression under a reproducing kernel hilbert space framework. *Journal of the American Statistical Association*, 113(524):1601–1611.

Talwai, P., Shameli, A., and Simchi-Levi, D. (2022). Sobolev norm learning rates for conditional mean embeddings. In *International conference on artificial intelligence and statistics*, pages 10422–10447. PMLR.

Tuo, R. and Bhattacharya, R. (2023). Privacy-aware Gaussian process regression. *arXiv preprint arXiv:2305.16541*.

Tuo, R. and Wang, W. (2020). Kriging prediction with isotropic Matérn correlations: robustness and experimental designs. *Journal of Machine Learning Research*, 21(187):1–38.

Tuo, R., Wang, Y., and Jeff Wu, C. (2020). On the improved rates of convergence for Matérn-type kernel ridge regression with application to calibration of computer models. *SIAM/ASA Journal on Uncertainty Quantification*, 8(4):1522–1547.

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

Utreras, F. I. (1988). Convergence rates for multivariate smoothing spline functions. *Journal of approximation theory*, 52(1):1–27.

Vaart, A. and Wellner, J. (2000). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.

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

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wahba, G. and Wold, S. (1975). Periodic splines for spectral density estimation: The use of cross validation for determining the degree of smoothing. *Communications in Statistics – Theory and Methods*, 4(2):125–141.

Wang, W. (2021). On the inference of applying gaussian process modeling to a deterministic function. *Electronic Journal of Statistics*, 15(2):5014–5066.

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

Yang, Y., Bhattacharya, A., and Pati, D. (2017). Frequentist coverage and sup-norm convergence rate in gaussian process regression. *arXiv preprint arXiv:1708.04753*.

Zhang, H., Li, Y., Lu, W., and Lin, Q. (2023). On the optimality of misspecified kernel ridge regression. *arXiv preprint arXiv:2305.07241*.

Zien, A. and Ong, C. S. (2007). Multiclass multiple kernel learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1191–1198.

Zongmin, W. (1992). Hermite-birkhoff interpolation of scattered data by radial basis functions. *Approximation Theory and its Applications*, 8(2):1–10.