Electronic Journal of Statistics
Vol. 14 (2020) 3070–3102
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1733

Nonparametric distributed learning under general designs

Meimei Liu and Zuofeng Shang*

Department of Statistics, Virginia Tech, Blacksburg, VA 24061
Department of Mathematical Sciences, NJIT, Newark, New Jersey 07102
e-mail: meimeiliu@vt.edu; zshang@njit.edu

Guang Cheng†

Department of Statistics, Purdue University, West Lafayette, IN 47906
e-mail: chengg@purdue.edu

Abstract: This paper focuses on the distributed learning in nonparametric regression framework. With sufficient computational resources, the efficiency of distributed algorithms improves as the number of machines increases. We aim to analyze how the number of machines affects statistical optimality. We establish an upper bound for the number of machines to achieve statistical minimax in two settings: nonparametric estimation and hypothesis testing. Our framework is *general* compared with existing work. We build a unified frame in distributed inference for various regression problems, including thin-plate splines and additive regression under random design: univariate, multivariate, and diverging-dimensional designs. The main tool to achieve this goal is a tight bound of an empirical process by introducing the Green function for equivalent kernels. Thorough numerical studies back theoretical findings.

AMS 2000 subject classifications: Primary 62G08; secondary 62G10.
Keywords and phrases: Computational limit, divide and conquer, kernel ridge regression, minimax optimality, nonparametric testing.

Received July 2019.

Contents

*Research sponsored by NSF DMS-1821157, and NSF DMS-1764280.
†This work was completed while Cheng was a member of Institute for Advanced Study, Princeton in the fall of 2019. Cheng would like to acknowledge hospitality of IAS, and also financial support from NSF DMS-1712907, DMS-1811812, DMS-1821183, Office of Naval Research, (ONR N00014-18-2759) and Adobe Data Science Fund.

1. Introduction

In a distributed computing environment, a common practice is to distribute a massive data set to multiple processors and then aggregate local results obtained from separate machines into global counterparts. Recently, researchers have made impressive progress in this modern Divide-and-Conquer (D&C) framework with different conquer strategies. Examples include median-of-means estimator proposed by [13], Bayesian aggregation considered by [18, 23, 20, 22], and simple averaging considered by [30] and [17].

Divide-and-Conquer often requires a growing number of machines to deal with an increasingly large data set. A fundamental question in distributed learning that statisticians are particularly interested in is how the number of machines affects statistical optimality? To address this question, [30] and [17] studied

the upper bounds for the number of machines s by analyzing statistical versus computational trade-off in D&C, where the number of deployed machines is treated as a simple proxy for computing cost. Consider a classical nonparametric regression setup, i.e., kernel ridge regression (KRR), [30] showed that, when s processors are employed with s in a suitable range, D&C method still preserves minimax optimal estimation. [17] derived *critical*, i.e., un-improvable, upper bounds for s to achieve either optimal estimation or optimal testing. The critical bound of processors for estimation in [17] significantly improves the one in [30] by polynomial order. Unfortunately, [17]'s results only focus on smoothing spline regression (a special case of KRR) with univariate even design and cannot be generalized to a random designed setting due to its technical limitation. However, in practice, data are usually generated with random designed, and multidimensional predictors. On the other hand, there is a lack of literature dealing with distributed nonparametric testing. To the best of our knowledge, [17] is the only reference but with the aforementioned model limitation.

In this paper, we consider distributed KRR in a general setup: design is random and multivariate. As our technical contribution, we characterize the upper bounds of s for achieving statistical optimality based on quantifying an empirical process. We show that a sharper concentration bound of the empirical process leads to a tighter upper bound of s. Efforts then have been devoted to a delicate bound of that empirical process. In the particular smoothing spline regression example, we establish a tight bound of the empirical process by introducing the Green function for equivalent kernels, leading to a polynomial order improvement of s compared with [30]. Our result is almost identical to the benchmark result in [17] (up to a logarithmic factor) for optimal estimation in smoothing spline, but under random design setting instead of the univariate evenly spaced design. Our theory can naturally handle various function spaces, including Sobolev space, Gaussian RKHS, or spaces of special structures such as additive functions, in a unified manner, as long as we can characterize the empirical process correspondingly.

The second contribution of this paper is to propose a Wald type test statistic for nonparametric testing in D&C regime. We derive the null limit distribution of the test statistics and characterize how the number of processors s affects minimax optimality of testing. The testing results are derived in a general framework that covers the aforementioned important function spaces. As an important byproduct, we obtain a minimax rate of testing for nonparametric additive models with a diverging number of components. Such rate is crucial in obtaining the upper bound of s for optimal testing and is of independent interest. Our results indicate an intrinsic difference in bounding s for optimal testing and estimation. For example, in smoothing spline, the upper bound of s for estimation is of the order $N^{2m/(2m+1)}/\log N$, while the one for testing is of the order $N^{(4m-3)(4m+1)}/\log N$.

The remainder of the article is organized as follows. In Section 2, we introduce background on reproducing kernel Hilbert space, describe the distributed kernel ridge regression and the nonparametric hypothesis testing. In Section 3, we establish minimax optimal estimation and testing for distributed KRR along

with some concrete examples. Section 4 contains thorough numerical studies. In Section 5, we conclude with a discussion. We defer the main proofs in Appendix.

2. Background and distributed kernel ridge regression

We begin by introducing some background on reproducing kernel Hilbert space (RKHS), and our nonparametric testing formulation under the distributed kernel ridge regression.

2.1. Nonparametric regression in reproducing kernel Hilbert spaces

Suppose that data $\{(Y_i, X_i) : i = 1, \ldots, N\}$ are *i.i.d* generated from the following regression model

$$Y_i = f(X_i) + \epsilon_i, \ \ i = 1, \ldots, N, \tag{2.1}$$

where ϵ_i are random errors with $E(\epsilon_i) = 0$, $E(\epsilon_i^2|X_i) = \sigma^2(X_i) > 0$, the covariates $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$ follows a distribution $\pi(x)$, and $Y_i \in \mathbb{R}$ is a real-valued response. Here, $d \geq 1$ is either fixed or diverging with N, and f is unknown.

Throughout we assume that $f \in \mathcal{H}$, where $\mathcal{H} \subset L_\pi^2(\mathcal{X})$ is a reproducing kernel Hilbert space (RKHS) associated with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and a reproducing kernel function $R(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. By Mercer's Theorem, R has the following spectral expansion

$$R(x, x') = \sum_{i=1}^{\infty} \mu_i \varphi_i(x) \varphi_i(x'), \quad x, x' \in \mathcal{X},$$

where $\mu_1 \geq \mu_2 \geq \cdots \geq 0$ is a sequence of eigenvalues and $\{\varphi_i\}_{i=1}^{\infty}$ form a basis in $L_\pi^2(\mathcal{X})$. Moreover, for any $i, j \in \mathbb{N}$,

$$\langle \varphi_i, \varphi_j \rangle_{L_\pi^2(\mathcal{X})} = \delta_{ij} \qquad \text{and} \qquad \langle \varphi_i, \varphi_j \rangle_{\mathcal{H}} = \delta_{ij}/\mu_i,$$

where δ_{ij} is Kronecker's δ.

We introduce an embedded norm $\|\cdot\|$ in $\mathcal{H}$ by combining the L_2 norm and $\|\cdot\|_{\mathcal{H}}$ norm to facilitate our statistical inference theory. For $f, g \in \mathcal{H}$, define

$$\langle f, g \rangle = V(f, g) + \lambda \langle f, g \rangle_{\mathcal{H}}, \tag{2.2}$$

where $V(f, g) = E\{f(X)g(X)\}$ and $\lambda > 0$ is the penalization parameter. Clearly, $\langle \cdot, \cdot \rangle$ defines an inner product on $\mathcal{H}$. As shown in [16], $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is also an RKHS with reproducing kernel function $K(\cdot, \cdot)$ satisfying the reproducing property

$$\langle f, K_x(\cdot) \rangle = f(x), \ \ \text{for all } f \in \mathcal{H},$$

where $K_x(\cdot) = K(x, \cdot)$ for $x \in \mathcal{X}$.

For any $f \in \mathcal{H}$, we can express the function in terms of the Fourier expansion as $f = \sum_{\nu \geq 1} V(f, \varphi_\nu)\varphi_\nu$. Therefore,

$$\langle f, \varphi_\nu \rangle = \sum_{i \geq 1} V(f, \varphi_i)\langle \varphi_i, \varphi_\nu \rangle = V(f, \varphi_\nu)(1 + \lambda/\mu_\nu). \tag{2.3}$$

Replacing f with K_x in (2.3), we have $V(K_x, \varphi_\nu) = \frac{\langle K_x, \varphi_\nu \rangle}{1+\lambda/\mu_\nu} = \frac{\varphi_\nu(x)}{1+\lambda/\mu_\nu}$. Then for any $x, y \in \mathcal{X}$, $K(x, y)$ has an explicit eigen-expansion expressed as

$$K(x, y) = \sum_{\nu \geq 1} V(K_x, \varphi_\nu)\varphi_\nu(y) = \sum_{\nu \geq 1} \frac{\varphi_\nu(x)\varphi_\nu(y)}{1 + \lambda/\mu_\nu}.$$

2.2. Distributed kernel ridge regression

To estimate f, we consider the kernel ridge regression (KRR) in a divide-and-conquer (D&C) regime. First, randomly divide the N samples into s subsamples. Let I_j denote the set of indices of the observations from subsample j for $j = 1, \ldots, s$. For simplicity, suppose $|I_j| = n$, i.e., all subsamples are of equal sizes. Hence, the total sample size is $N = ns$. Then, we estimate f based on the jth subsample through the following KRR method:

$$\widehat{f}_j = \operatorname*{argmin}_{f \in \mathcal{H}} \ell_{j,\lambda}(f) \equiv \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{2n} \sum_{i \in I_j} (Y_i - f(X_i))^2 + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2, \quad j = 1, \ldots, s,$$

where $\lambda > 0$ is the penalization parameter. The D&C estimator of f is defined as the average of $\widehat{f}_j$'s, that is, $\bar{f} = \sum_{j=1}^{s} \widehat{f}_j/s$. In Section 3.2, we characterize the upper bounds of s for $\bar{f}$ to be a minimax estimator.

For nonparametric inference, we focus on testing whether the nonparametric function in (2.1) is equal to some known function. That is, we consider the hypothesis testing problem

$$H_0 : f = f_0, \text{ vs. } H_1 : f \in \mathcal{H} \backslash \{f_0\},$$

where f_0 is an arbitrarily known hypothesized function. In general, testing $f = f_0$ is equivalent to testing $f_* = f - f_0 = 0$. Therefore, without loss of generality, we focus on the hypothesis testing

$$H_0 : f = 0, \text{ vs. } H_1 : f \in \mathcal{H} \backslash \{0\}. \tag{2.4}$$

Based on $\bar{f}$, we propose a Wald-type statistic

$$T_{N,\lambda} = \|\bar{f}\|^2, \tag{2.5}$$

where $\|\cdot\|$ is the embedded norm defined in (2.2). Intuitively, a large value of $T_{N,\lambda}$ tends to reject H_0. In Section 3.3, we will derive the null limit distribution of $T_{N,\lambda}$, and explicitly show how the number of processors s affects the minimax optimality of testing.

3. Main results

In this section, we derive some general results relating to $\bar{f}$ and $T_{N,\lambda}$. We first introduce some regularity assumptions.

3.1. Assumptions

We assume the design density is bounded, and the error ϵ has a finite fourth moment. Such assumption is commonly used in literature [3].

Assumption 3.1. *There exists a constant $c_\pi > 0$ such that for all $x \in \mathcal{X}$, $0 \le \pi(x), \sigma^2(x) \le c_\pi$.*

Assumption 3.2. *There exists a positive constant τ such that $E\{\epsilon^4|X\} < \tau$ almost surely.*

For any function f, define its supremum norm as $\|f\|_{\sup} = \sup_{x\in\mathcal{X}} |f(x)|$. We further assume the eigenfunctions $\{\varphi_\nu\}_{\nu=1}^\infty$ are uniformly bounded on $\mathcal{X}$, and the eigenvalues $\{\mu_\nu\}_{\nu=1}^\infty$ satisfy certain tail sum property.

Assumption 3.3. $c_\varphi := \sup_{j\ge1} \|\varphi_j\|_{\sup} < \infty$ *and* $\sup_{k\ge1} \frac{\sum_{\nu=k+1}^\infty \mu_i}{k\mu_k} < \infty$.

The uniform boundedness condition of eigenfunctions holds for various kernels including univariate periodic kernel, 2-dimensional Gaussian kernel, multivariate additive kernel; see [9], [12] and reference therein. The tail sum property can also be verified in various RKHS, and is deferred to Appendix.

Define $h = (\sum_{\nu\ge1} \frac{1}{1+\lambda/\mu_\nu})^{-1}$. h^{-1} is known as the effective dimension measuring the capacity of $\mathcal{H}$, and has been widely studied in [1], [11], [29]. In fact, there exists an explicit relationship between h and λ. For example, for the polynomial decaying kernels with $\mu_\nu \asymp \nu^{-2m}$, simple calculation shows that $h \asymp \lambda^{1/(2m)}$. We provide concrete examples to illustrate such connection in Section 3.4.

Define $Pf = E\{f(X)\}$, $P_j f = n^{-1}\sum_{i\in I_j} f(X_i)$ and

$$\xi_j = \sup_{\substack{f,g\in\mathcal{H}\\ \|f\|=\|g\|=1}} |P_j fg - Pfg|,\ 1 \le j \le s.$$

ξ_j is the supremum of the empirical processes indexed by the class $\mathcal{H}\cdot\mathcal{H} := \{f\cdot g : f, g \in \mathcal{H}\}$ based on subsample j. The quantity $\max_{1\le j\le s} \xi_j$ plays a vital role in determining the critical upper bound of s to guarantee statistical optimality. As shown in our main theorems in Section 3.2 and 3.3 later, a sharper bound of ξ_j directly leads to an improved upper bound of s. The following Assumption 3.4 provides a concentration bound for ξ_j, and says that ξ_j are uniformly bounded by $\sqrt{\frac{\log^b N}{nh^a}}$, a, b are constants that are specified in various kernels. Verification of Assumption 3.4 is deferred to Section 3.4 in concrete settings based on empirical processes methods, where the values of a, b will be explicitly specified.

Assumption 3.4. *There exist nonnegative constants a, b such that*

$$\max_{1\le j\le s} \xi_j = O_P\left(\sqrt{\frac{\log^b N}{nh^a}}\right).$$

3.2. Minimax optimal estimation

We are ready to establish the minimax property for the distributed KRR estimator with the assumptions stated in place.

Let $\mathbf{X}_j = \{X_i : i \in I_j\}$ and $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_s\}$. Suppose that (2.1) holds under $f = f_0$. Let $\mathcal{P}_\lambda$ be a self-adjoint operator from $\mathcal{H}$ to itself such that $\langle \mathcal{P}_\lambda f, g\rangle = \lambda\langle f, g\rangle_{\mathcal{H}}$ for all $f, g \in \mathcal{H}$. Then for any $f \in \mathcal{H}$, $\|f\|^2 = E\{f^2(X)\} + \langle \mathcal{P}_\lambda f, f\rangle$. The existence of $\mathcal{P}_\lambda$ follows by [16, Proposition 2.1].

In the following Lemma 3.1, we obtain an uniform error bound for $\widehat{f}_j$'s ($j = 1, \ldots, s$). Then a general error bound for $\bar{f}$ can be achieved by aggregating local estimators.

Lemma 3.1. *Suppose Assumptions 3.1, 3.3, 3.4 are satisfied and* $\log^b N = o(nh^a)$ *with a, b given in Assumption 3.4. Then with probability approaching one, for any* $1 \le j \le s$,

$$E\{\|\widehat{f}_j - E\{\widehat{f}_j|\mathbf{X}_j\} - \frac{1}{n}\sum_{i\in I_j}\epsilon_i K_{X_i}\|^2|\mathbf{X}_j\} \le \frac{4c_\pi c_\varphi^2 \xi_j^2}{nh}, \tag{3.1}$$

$$\|E\{\widehat{f}_j|\mathbf{X}_j\} - f_0 + \mathcal{P}_\lambda f_0\| \le 2\xi_j \lambda^{1/2}\|f_0\|_{\mathcal{H}} \tag{3.2}$$

In Lemma 3.1, we decompose the deviation from $\widehat{f}_j$ to f_0 as two terms representing bias and variance, that is, $\widehat{f}_j - f_0 = E\{\widehat{f}_j|\mathbf{X}_j\} - f_0 + \widehat{f}_j - E\{\widehat{f}_j|\mathbf{X}_j\}$. Equation (3.1) quantifies the variance of $\widehat{f}_j$ via the leading term $\frac{1}{n}\sum_{i\in I_j}\epsilon_i K_{X_i}$ and a higher order remainder term involving ξ_j. Equation (3.2) represents the bias of $\widehat{f}_j$ as the dominating term $\mathcal{P}_\lambda f_0$ and a higher order reminder as a function of ξ_j for any $1 \le j \le s$.

Lemma 3.1 immediately leads to the result on $\bar{f}$ via triangle inequality. Specifically, (3.1) and (3.2) lead to the following (3.3) in Theorem 3.1, which, together with the rates of $\sum_{i=1}^N \epsilon_i K_{X_i}$ and $\mathcal{P}_\lambda f_0$ in Lemma A.1, leads to (3.4).

Theorem 3.1. *If the conditions in Lemma 3.1 hold, then with probability approaching one,*

$$E\{\|\bar{f} - \frac{1}{N}\sum_{i=1}^N \epsilon_i K_{X_i} - f_0 + \mathcal{P}_\lambda f_0\|^2|\mathbf{X}\} \le 4\left(\frac{c_\pi c_\varphi^2}{Nh} + \lambda\|f_0\|_{\mathcal{H}}^2\right)\max_{1\le j\le s}\xi_j^2 \tag{3.3}$$

$$E\{\|\bar{f} - f_0\|^2|\mathbf{X}\} \le \frac{4c_\pi c_\varphi^2}{Nh} + 8\lambda\|f_0\|_{\mathcal{H}}^2. \tag{3.4}$$

Theorem 3.1 is a general result that holds for many commonly used kernels. The upper bound of s is implied by the key condition $\max_{1\le j\le s}\xi_j = o(1)$, that is, $\log^b N = o(nh^a)$ according to Assumption 3.4 with $n = N/s$. Then Equation (3.4) in Theorem 3.1 states that as long as s is dominated by $Nh^a/\log^b N$, the conditional mean squared errors can be upper bounded by the variance term $(Nh)^{-1}$ and the squared bias term $\lambda\|f_0\|_{\mathcal{H}}^2$. Minimax optimal estimation can be obtained through the particular λ^* that satisfies such bias-variance trade-off; see [1], [26]. Since h is a function of λ, denote $h^* = (\sum_{\nu\ge1}\frac{1}{1+\lambda^*/\mu_\nu})^{-1}$, we claim $Nh^{*a}/\log^b N$ as the upper bound of s to achieve optimal estimation. Section 3.4 further illustrates concrete, and interpretable guarantees on the conditional mean squared errors to particular kernels with a, b specified accordingly.

We build a connection between the upper bound of s and the performance of $\bar{f}$ through the uniform bound of the empirical process ξ_j. A tighter upper bound of s can be achieved by a sharper concentration bound of $\max_{1\le j\le s}\xi_j$. Therefore, in Section 3.4, efforts are devoted to a tight bound of $\max_{1\le j\le s}\xi_j$ based on various empirical process methods. For instance, in smoothing spline regression stated in Section 3.4.1, we provide a sharp concentration bound of ξ_j with $a = b = 1$ holds in Assumption 3.4 based on [3]. Consequently, we achieve an upper bound for s almost identical to the critical one obtained in [17] (up to a logarithmic factor), and improve [17]'s sharp result in the sense of removing the fixed univariate design assumption.

3.3. Minimax optimal testing

In this section, we study the nonparametric distributed inference based on $\bar{f}$. Consider the hypothesis testing (2.4), we first derive the asymptotic distribution of the Wald-type test statistics $T_{N,\lambda} := \|\bar{f}\|^2$ and further investigate its power behavior. For simplicity, assume that $\sigma^2(x) \equiv \sigma^2$ is known. Otherwise, we can replace σ^2 by its consistent estimator to fulfill our procedure.

To prove the testing consistency, we show that $T_{N,\lambda} = \|\frac{1}{N}\sum_{i=1}^N \epsilon_i K_{X_i}\|^2 +$ *remainder*; detailed proof is deferred to Appendix. It is feasible to characterize the asymptotic behavior of $\|\frac{1}{N}\sum_{i=1}^N \epsilon_i K_{X_i}\|^2$ thanks to the explicit expression of the embedded kernel K_{X_i}.

Define $W(N) = \sum_{1\le i<k\le N} W_{ik}$ with $W_{ik} = 2\epsilon_i\epsilon_k K(X_i, X_k)$, and let $\sigma^2(N) = \mathrm{Var}\{W(N)\}$. Denote the empirical kernel matrix as $\mathbf{K} = [K(X_i, X_j)]_{i,j=1}^N$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)^T$. In the following Lemma 3.2, we characterize the asymptotic behavior of $\|\frac{1}{N}\sum_{i=1}^N \epsilon_i K_{X_i}\|^2$.

Lemma 3.2. *Suppose Assumptions 3.1, 3.2, 3.3, 3.4 are all satisfied, and* $N \to \infty$, $h = o(1)$, $Nh^2 \to \infty$. *Then it holds that*

$$\boldsymbol{\epsilon}'\mathbf{K}\boldsymbol{\epsilon} = \sigma^2 Nh^{-1} + W(N) + O_P(\sqrt{Nh^{-2}}). \tag{3.5}$$

Furthermore, as $N \to \infty$, $\frac{W(N)}{\sigma(N)} \xrightarrow{d} N(0,1)$, *where* $\sigma^2(N) = 2\sigma^4 N(N-1)\sum_{\nu\ge1}\frac{1}{(1+\lambda/\mu_\nu)^2} \asymp N^2h^{-1}$.

The following Theorem 3.2 shows that $T_{N,\lambda}$ is asymptotically normal under H_0, provided that s satisfies a key condition $\log^b N = o(nh^{a+1})$, where a, b are determined through the uniform bound of ξ_j in Assumption 3.4.

Theorem 3.2. *Suppose Assumptions 3.1 to 3.4 hold, and as $N \to \infty$, $h = o(1)$, $Nh^2 \to \infty$, and $\log^b N = o(nh^{a+1})$. Then, as $N \to \infty$,*

$$\frac{N^2}{\sigma(N)}\left(T_{N,\lambda} - \frac{\sigma^2}{Nh}\right) \xrightarrow{d} N(0,1).$$

By Theorem 3.2, we can define an asymptotic testing rule with $(1-\alpha)$ significance level as follows:

$$\psi_{N,\lambda} = I\left(|T_{N,\lambda} - \sigma^2/(Nh)| \geq z_{1-\alpha/2}\sigma(N)/N^2\right),$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)\times 100$ percentile of standard normal distribution.

Intuitively, the smaller $\|f\|^2$ is, the harder it is to distinguish the alternative hypothesis from the null. The power performance can be evaluated by the minimax rate of testing (MRT) that is defined as the minimal distance between the null and the alternative hypotheses such that valid testing is possible ([5], [7]). In the following, we show that the distributed test statistic $T_{N,\lambda}$ can achieve minimax rate of testing, provided that the number of divisions s belongs to a suitable range.

For any $f \in \mathcal{H}$, define the separation rate

$$d_{N,\lambda} = \underbrace{\lambda^{1/2}\|f\|_{\mathcal{H}}}_{\text{Bias of } \bar{f}} + \underbrace{(Nh^{1/2})^{-1/2}}_{\text{Standard deviation of } T_{N,\lambda}}. \tag{3.6}$$

The separation rate $d_{N,\lambda}$ is used to measure the distance between the null and the alternative hypotheses. The following Theorem 3.3 shows that, if the alternative signal f is separated from zero by an order $d_{N,\lambda}$, then the proposed test statistic asymptotically achieves high power. It is sufficient to minimize the separation rate $d_{N,\lambda}$ to achieve optimal testing. We show that the minimax rate of testing can be achieved by selecting λ to balance the trade-off between the bias of $\bar{f}$ and the standard derivation of $T_{N,\lambda}$; see [8], [24].

Theorem 3.3. *If the conditions in Theorem 3.2 hold, then for any $\varepsilon > 0$, there exist C_ε and N_ε s.t.*

$$\inf_{\|f\| \geq C_\varepsilon d_{N,\lambda}} P_f(\psi_{N,\lambda} = 1) \geq 1 - \varepsilon, \quad \textit{for any } N \geq N_\varepsilon.$$

In Section 3.4, we develop upper bounds for s in various concrete examples based on the above general theorems. Our results indicate that there has an intrinsic difference in bounding s for optimal testing and estimation. The rationale behind this phenomenon is that, different from the classical "bias-variance" trade-off in the optimal nonparametric estimation; the optimal nonparametric testing can be achieved by another type of trade-off between the squared bias of the estimator and the standard deviation of the test statistic, leading to a different number of processors s.

3.4. Examples

In this section, we derive upper bounds for s in four featured examples to achieve optimal estimation and testing, based on the general results obtained in Sections 3.2 and 3.3. Our examples cover the settings of univariate, multivariate, and diverging-dimensional designs.

3.4.1. Example 1: Smoothing spline regression

Suppose $\mathcal{H} = \{f \in S^m(\mathbb{I}) : \|f\|_{\mathcal{H}} \leq C\}$ for a constant $C > 0$, where $S^m(\mathbb{I})$ is the mth order Sobolev space on $\mathbb{I} \equiv [0,1]$, i.e.,

$$S^m(\mathbb{I}) = \big\{ f \in L^2(\mathbb{I}) |\, f^{(j)} \text{ are abs. cont. for } j = 0, 1, \ldots, m-1, \\ \text{and } \int_{\mathbb{I}} |f^{(m)}(x)|^2 dx < \infty \big\},$$

and $\|f\|_{\mathcal{H}} = \int_{\mathbb{I}} |f^{(m)}(x)|^2 dx$. Then model (2.1) becomes the usual smoothing spline regression. In addition to Assumption 3.1, we assume that

$$c_\pi^{-1} \leq \pi(x) \leq c_\pi, \text{ for any } x \in \mathbb{I}. \tag{3.7}$$

We call the design satisfying (3.7) as quasi-uniform, a common assumption on many statistical problems; see [3]. Quasi-uniform assumption excludes cases where design density is (nearly) zero at certain data points, which may cause estimation inaccuracy at those points.

It is known that when $m > 1/2$, $S^m(\mathbb{I})$ is an RKHS under the inner product $\langle \cdot, \cdot \rangle$; see [16], [4]. Meanwhile, Assumption 3.3 holds with kernel eigenvalues $\mu_\nu \asymp \nu^{-2m}$, $\nu \geq 1$. Hence, Proposition A.1 holds with $h \asymp \lambda^{1/(2m)}$. We next provide a sharp concentration inequality to bound ξ_j.

Proposition 3.1. *Under (3.7), there exist universal positive constants c_1, c_2, c_3 such that for any $1 \leq j \leq s$,*

$$P\left(\xi_j \geq t\right) \leq 2n \exp\left(-\frac{nht^2}{c_1 + c_2 t}\right), \text{ for all } t \geq c_3 (nh)^{-1}.$$

The proof of Proposition 3.1 is based on the technical tool that applying Green function for equivalent kernels; see [3, Corollary 5.41]. An immediate consequence of Proposition 3.1 is that Assumption 3.4 holds with $a = b = 1$. Then based on Theorem 3.1 and Theorem 3.3, we have the following results.

Corollary 3.1. *Suppose that $\mathcal{H} = S^m(\mathbb{I})$, (3.7) holds, and Assumptions 3.1, 3.2 hold.*

(a) If $m > 1/2$, $s = o(N^{2m/(2m+1)}/\log N)$ and $\lambda \asymp N^{-2m/(2m+1)}$, then $\|\bar{f} - f_0\| = O_P(N^{-m/(2m+1)})$.

(b) If $m > 3/4$, $s = o(N^{(4m-3)/(4m+1)}/\log N)$ and $\lambda \asymp N^{-4m/(4m+1)}$, then the Wald-type test achieves minimax rate of testing $N^{-2m/(4m+1)}$.

It is known that the estimation rate $N^{-m/(2m+1)}$ is minimax-optimal; see [21]. Furthermore, the testing rate $N^{-2m/(4m+1)}$ is also minimax optimal, in the sense of [8]. It is worth noting that the upper bound for $s = o(N^{2m/(2m+1)}/\log N)$ matches (upto a logarithmic factor) the critical one by [17] in evenly spaced design, which is substantially larger than the one obtained by [30], i.e., $s = o(N^{(2m-1)/(2m+1)}/\log N)$; see Table 1 for the comparison.

TABLE 1
Comparison of upper bounds of s to achieve minimax estimation.

	Zhang et al. [30]	Shang et al. [17]	Our approach
smoothing spline regression	$s \lesssim N^{\frac{2m-1}{2m+1}}/\log N$ sharpness of s ✗	$s \lesssim N^{\frac{2m}{2m+1}}$ sharpness of s ✓	$s = o(N^{\frac{2m}{2m+1}}/\log N)$ sharpness of s ✓

3.4.2. Example 2: Nonparametric additive regression

Consider the function space

$$\mathcal{H} = \{f(x_1, \ldots, x_d) = \sum_{k=1}^{d} f_k(x_k) : \ f_k \in S^m(\mathbb{I}), \ \|f_k\|_{\mathcal{H}} \le C \text{ for } k = 1, \ldots, d\},$$

where $C > 0$ is a constant. That is, any $f \in \mathcal{H}$ has an additive decomposition of f_k's. Here, d is either fixed or slowly diverging. Such additive model has been well studied in many literatures; see [21], [10], [15], [28] among others. For $x = (x_1, \cdots, x_d) \in \mathcal{X}$, suppose x_i, x_j are independent for $i \neq j \in \{1, \cdots, d\}$ and each x_i satisfies (3.7). For identifiability, assume $E\{f_k(x_k)\} = 0$ for all $1 \le k \le d$. For $f = \sum_{k=1}^{d} f_k$ and $g = \sum_{k=1}^{d} g_k$, define

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{k=1}^{d} \langle f_k, g_k\rangle_{\mathcal{H}} = \sum_{k=1}^{d} \int_{\mathbb{I}} f_k^{(m)}(x) g_k^{(m)}(x) dx, \quad \text{and}$$

$$V(f, g) = \sum_{k=1}^{d} V_k(f_k, g_k) \equiv \sum_{k=1}^{d} E\{f_k(X_k) g_k(X_k)\}.$$

It is easy to verify that $\mathcal{H}$ is an RKHS under $\langle \cdot, \cdot \rangle$ defined in (2.2). Lemma 3.3 below summarizes the properties for $\mathcal{H}$ with d additive components.

Lemma 3.3. *(a) There exist eigenfunctions φ_ν and eigenvalues μ_ν satisfying Assumption 3.3.*
(b) It holds that $\sum_{\nu \ge 1}(1 + \lambda/\mu_\nu)^{-1} := h^{-1} \asymp d\lambda^{-1/(2m)}$, and $\sum_{\nu \ge 1}(1 + \lambda/\mu_\nu)^{-2} \asymp h^{-1}$ accordingly.
(c) For $f \in \mathcal{H}$, $\|\mathcal{P}_\lambda f\|^2 \le cd\lambda$, where c is a bounded constant.
(d) Assumption 3.4 holds with $a = b = 1$.

Lemma 3.3 (d) establishes a concentration inequality of ξ_j for the additive model, such that $\max_{1 \le j \le s} = O_P(\sqrt{\frac{\log N}{nh}})$. The proof is based on the extension

of the Green function techniques [3] to diverging dimensional setting; see Lemma A.2 in Appendix.

Combining Lemma 3.3, Theorems 3.1, 3.2 and 3.3, we have the following result.

Corollary 3.2. *(a) Suppose Assumptions 3.1, 3.2 hold. If* $m > 1/2$, $d = o(N^{\frac{2m}{2m+1}}/\log N)$, $s = o(d^{-1}N^{\frac{2m}{2m+1}}/\log N)$, $\lambda \asymp N^{-\frac{2m}{2m+1}}$, *then* $\|\bar{f} - f_0\| = O_P(d^{1/2}N^{-\frac{m}{2m+1}})$.

(b) Suppose Assumptions 3.1, 3.2 hold.

If $m > 3/4$, $d = o(N^{\frac{4m-3}{4(2m+1)}}(\log N)^{-\frac{4m+1}{4(2m+1)}})$, $s = o(d^{-\frac{4(2m+1)}{4m+1}}N^{\frac{4m-3}{4m+1}}/\log N)$, *and* $\lambda \asymp d^{-\frac{2m}{4m+1}}N^{-\frac{4m}{4m+1}}$, *then the Wald-type test achieves minimax rate of testing with* $d^{\frac{2m+1}{2(4m+1)}}N^{-\frac{2m}{4m+1}}$.

Remark 3.1. *It was shown by [15] that* $d^{1/2}N^{-\frac{m}{2m+1}}$ *is the minimax estimation rate in nonparametric additive model. Corollary 3.2 (a) provides an upper bound for* s *such that* $\bar{f}$ *achieves this rate. Meanwhile, Corollary 3.2 (b) provides a different upper bound for* s *such that our Wald-type test achieves minimax rate of testing* $d^{\frac{2m+1}{2(4m+1)}}N^{-\frac{2m}{4m+1}}$. *It should be emphasized that such a minimax rate of testing is a new result in literature, which is of independent interest. The proof is based on a local geometry approach recently developed by [24]. When* $d = 1$, *all results in this section reduce to Example 1 on univariate smoothing splines.*

3.4.3. Example 3: Gaussian RKHS regression

Suppose $\mathcal{H}$ is an RKHS generated by the Gaussian kernel $K(x, x') = \exp(-c\|x - x'\|_2^2), x, x' \in \mathbb{R}^d$, where $c, d > 0$ are constants, $\|\cdot\|_2$ is the Euclidean norm. Here we consider $d = 1, 2$. Then Assumption 3.3 holds with $\mu_\nu \asymp [(\sqrt{5}-1)/2]^{-(2\nu+1)}$, $\nu \geq 1$; see [19]. It can be shown that $h \asymp (-\log\lambda)^{-1/2}$ holds. To verify Assumption 3.4, we need the following lemma.

Lemma 3.4. *For Gaussian RKHS, Assumption 3.4 holds with* $a = 2$, $b = d+2$.

Following Theorem 3.1, Theorems 3.2 and 3.3, we get the following consequence.

Corollary 3.3. *Suppose that* $\mathcal{H}$ *is a Gaussian RKHS and Assumptions 3.1 and 3.2 hold.*

(a) If $s = o(N/\log^{d+3}(N))$ *and* $\lambda \asymp N^{-1}\sqrt{\log N}$, *then*

$$\|\bar{f} - f_0\| = O_P(N^{-1/2}\log^{1/4} N).$$

(b) If $s = o(N/\log^{d+3.5} N)$ *and* $\lambda \asymp N^{-1}\log^{1/4} N$, *then the Wald-type test achieves minimax rate of testing* $N^{-1/2}\log^{1/8} N$.

Corollary 3.3 shows that in Gaussian RKHS, as the sample size in the local machine is greater than a logarithmic order of N, one can obtain both optimal estimation and testing. This conclusion is consistent with the upper bound obtained by [30] for optimal estimation, which is of a different logarithmic factor.

In fact, the effective dimension h^{-1} for Gaussian RKHS is of order $\log N$, which is used to measure the space complexity.

3.4.4. Example 4: Thin-Plate spline regression

Consider the mth order Sobolev space on $\mathbb{I}^d$, i.e., $\mathcal{H} = S^m(\mathbb{I}^d)$, with $d = 2$ being fixed. It is known that Assumption 3.3 holds with $\mu_\nu \asymp \nu^{-2m/d}$; see [6]. Hence $h \asymp \lambda^{d/(2m)}$. The following lemma verifies Assumption 3.4.

Lemma 3.5. *For thin-plate splines, Assumption 3.4 holds with* $a = 3 - d/(2m)$, $b = 1$.

Following Theorem 3.1, Theorem 3.2 and Theorem 3.3, we have the following result.

Corollary 3.4. *Suppose* $f \in S^m(\mathbb{I}^d)$ *with* $d = 2$, *Assumption 3.1 and Assumption 3.2 hold.*

(a) If $s = o(N^{\frac{(2m-d)^2}{2m(2m+d)}} / \log N)$ *and* $\lambda \asymp N^{-\frac{2m}{2m+d}}$, *then*

$$\|\bar{f} - f_0\| = O_P(N^{-m/(2m+d)}).$$

(b) If $s = o(N^{\frac{4m^2-7dm+d^2}{(4m+d)m}} / \log N)$ *and* $\lambda \asymp N^{-\frac{4m}{4m+d}}$, *then the Wald-type test achieves minimax rate of testing* $N^{-2m/(4m+d)}$.

Corollary 3.4 demonstrates upper bounds on s. These upper bounds are smaller compared with Corollary 3.1 in the univariate case, since the proof technique in bounding the empirical process ξ_j here is not as sharp as the Green function technique used in Proposition 3.1 for the univariate example.

4. Simulation

In this section, we examined the performance of our proposed estimation and testing procedures versus various choices of the number of machines in two examples based on simulated datasets.

4.1. Smoothing spline regression

The data were generated from the following regression model

$$Y_i = c * (0.6 \sin(1.5\pi X_i)) + \epsilon_i, \quad i = 1, \cdots, N, \tag{4.1}$$

where $X_i \overset{iid}{\sim} \text{Unif}[0,1]$, $\epsilon_i \overset{iid}{\sim} N(0,1)$ and c is a constant. Cubic spline (i.e., $m = 2$ in Section 3.4.1) was employed for estimating the regression function. To display the impact of the number of divisions s on statistical performance, we set sample sizes $N = 2^l$ for $9 \le l \le 13$ and chose $s = N^\rho$ for $0.1 \le \rho \le 0.8$. To examine the estimation procedure, we generated data from model (4.1) with $c = 1$.

Mean squared errors (MSE) were reported based on 100 independent replicated experiments. The left panel of Figure 1 summarizes the results. Specifically, it displays that the MSE increases as s does so; while the MSE increases suddenly when $\rho \approx 0.7$, where $\rho \equiv \log(s)/\log(N)$. Recall that the theoretical upper bound for s, is $N^{0.8}$; see Corollary 3.1. Hence, estimation performance becomes worse near this theoretical boundary.

Next consider the hypothesis testing problem $H_0 : f = 0$. To examine the proposed Wald test, we generated data from model (4.1) at both $c = 0, 1$; $c = 0$ used for examining the size of the test, and $c = 1$ used for examining the power of the test. The significance level was chosen as 0.05. Both size and power were calculated as the proportions of rejections based on 500 independent replications. The middle and right panels of Figure 1 summarize the results. Specifically, the right panel shows that the size approaches the nominal level 0.05 under various choices of (s, N), confirming the validity of the Wald test. The middle panel displays that the power increases when ρ decreases; the power maintains at 100% when $\rho \leq 0.5$ and $N \geq 4096$. Whereas the power quickly drops to zero when $\rho \geq 0.6$. This result is consistent with our theoretical finding. Recall that the theoretical upper bound for s is $N^{0.56}$; see Corollary 3.1. The numerical results also reveal that the upper bound of s to achieve optimal testing is smaller than the one required for optimal estimation.

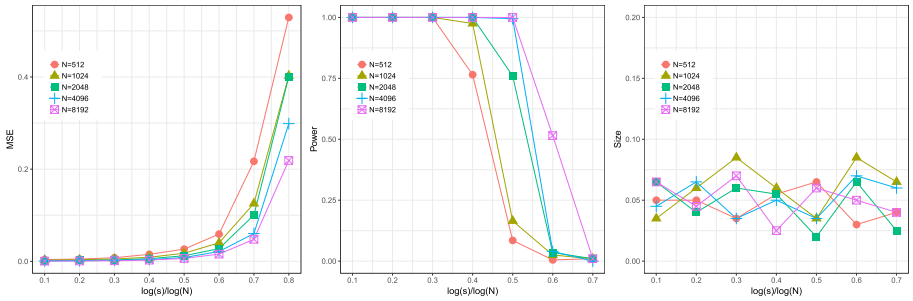

FIG 1. *Smoothing Spline Regression. (a) MSE of $\bar{f}$ versus $\rho \equiv \log(s)/\log(N)$. (b) Power of the Wald test versus ρ. (c) Size of the Wald test versus ρ.*

4.2. Nonparametric additive regression

We generated data from the following nonparametric model of two additive components

$$Y_i = c * f(X_{i1}, X_{i2}) + \epsilon_i, \ \ i = 1, \cdots, N, \tag{4.2}$$

where $f(x_1, x_2) = 0.4 \sin(1.5\pi x_1) + 0.1(0.5 - x_2)^3$, and $X_{i1}, X_{i2} \overset{iid}{\sim} \text{Unif}[0, 1]$, $\epsilon_i \overset{iid}{\sim} N(0, 1)$, and c is a constant. To examine the estimation procedure, we generated data from (4.2) with $c = 1$. To examine the testing procedure, we generated data at $c = 0, 1$. N, s were chosen to be the same as the smoothing

spline example in Section 4. Results are summarized in Figure 2. The interpretations are again similar to Figure 1, only with a slightly different asymptotic trend. Specifically, the MSE suddenly increases at $\rho \approx 0.6$, and the power quickly approaches one at $\rho \approx 0.5$. The sizes are around the nominal level 0.05 for all cases.

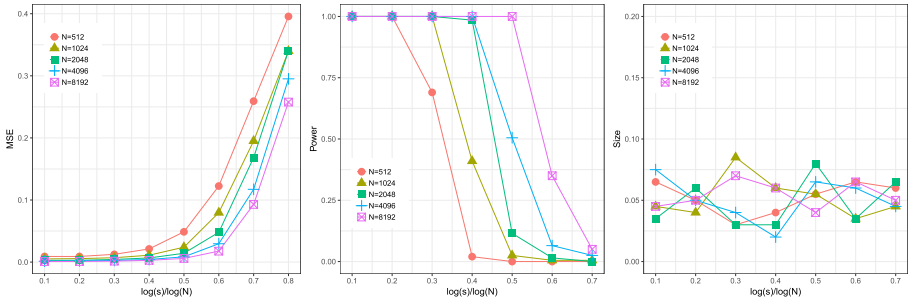

FIG 2. *Additive Regression Model. (a) MSE of $\bar{f}$ versus $\rho \equiv \log(s)/\log(N)$. (b) Power of the Wald test versus ρ. (c) Size of the Wald test versus ρ.*

5. Conclusion

This paper offers theoretical insights on how to allocate data in parallel computing for KRR in both estimation and testing procedures. In comparison with [30] and [17], our work provides a general and unified treatment of such problems in modern diverging-dimension or big data settings. Furthermore, using the green function for equivalent kernels to provide a sharp concentration bound on the empirical processes related to s, we have improved the upper bound of the number of machines in smoothing spline regression by [30] from $N^{(2m-1)/(2m+1)}/\log N$ to $N^{2m/(2m+1)}/\log N$ for optimal estimation, which is proven un-improvable in [17] (up to a logarithmic factor). In the end, we would like to point out that our theory is useful in designing a distributed version of generalized cross validation method that is developed to choose tuning parameter λ and the number of machines s; see [25].

Appendix A: Proofs of main results

A.1. Notation table

A.2. Some preliminary results

Lemma A.1. *(a) For any $x, y \in \mathcal{X}$, $K(x,y) \le c_\varphi^2 h^{-1}$.*
(b) For any $f \in \mathcal{H}$, $\|\mathcal{P}_\lambda f\| \le \lambda^{1/2}\|f\|_{\mathcal{H}}$.

TABLE 2
A table that lists all useful notation and their meanings.

N	sample size
Y	response
X	covariate
ϵ	random error
$\mathcal{H}$	reproducing kernel Hilbert space (RKHS)
$\pi(x)$	density distribution
d	dimension of covariate
$\langle\cdot,\cdot\rangle_{\mathcal{H}}$, $\|\cdot\|_{\mathcal{H}}$	the inner product and norm under $\mathcal{H}$
$R(\cdot,\cdot)$	kernel function under the norm $\|\cdot\|_{\mathcal{H}}$
μ_i	eigenvalue
φ	eigenfunction
$\langle\cdot,\cdot\rangle_{L^2_\pi(\mathcal{X})}$	L_2 inner product
$\langle\cdot,\cdot\rangle$, $\|\cdot\|$	embedded inner product and norm
$V(\cdot,\cdot)$	L_2 inner product
$K(\cdot,\cdot)$	kernel function equipped with $\|\cdot\|$
$K_x(\cdot)$	$= K(x,\cdot)$
s	number of division
I_j	the set of indices of the observation from subsample j
n	the subsample size
$\widehat{f}_j$	the estimate of f based on subsample j
λ	penalization parameter
$\bar{f}$	D&C estimator
$T_{N,\lambda}$	test statistic
$\|\cdot\|_{\sup}$	the supremum norm
h^{-1}	$= \sum_{\nu\geq 1} \frac{1}{1+\lambda/\mu_\nu}$
ξ_j	$= \sup_{\substack{f,g\in\mathcal{H}\\ \|f\|=\|g\|=1}} \|P_j fg - Pfg\|$
$\mathcal{P}_\lambda$	self-adjoint operator satisfies $\langle\mathcal{P}_\lambda f, g\rangle = \lambda\langle f, g\rangle_{\mathcal{H}}$
$\mathbf{K}$	empirical kernel matrix
$S^m(\mathbb{I})$	the mth order Sobolev space on $\mathbb{I} \equiv [0,1]$

Proof. (a)

$$K(x,y) = \sum_{\nu\geq 1} \frac{\varphi_\nu(x)\varphi_\nu(y)}{1+\lambda/\mu_\nu} \leq c_\varphi^2 h^{-1},$$

where the last inequality is by Assumption 3.3 and the definition of h^{-1}.
(b)

$$\begin{aligned}\|\mathcal{P}_\lambda f\| &= \sup_{g\in\mathcal{H},\|g\|\leq 1} \langle\mathcal{P}_\lambda f, g\rangle = \sup_{g\in\mathcal{H},\|g\|\leq 1} \lambda\langle f, g\rangle_{\mathcal{H}} \\ &\leq \sup_{g\in\mathcal{H},\|g\|\leq 1} \lambda^{1/2}\|f\|_{\mathcal{H}}\lambda^{1/2}\|g\|_{\mathcal{H}} \leq \lambda^{1/2}\|f\|_{\mathcal{H}}.\end{aligned}$$

□

Another quantity of interest is the series $\sum_{\nu\geq 1}(1+\lambda/\mu_\nu)^{-2}$, which represents the variance term of the test statistics that will be analyzed in Theorem 3.2. In the following Proposition A.1, we show that such variance term has the same order of the effective dimension.

Proposition A.1. *Suppose Assumption 3.3 holds. For any* $\lambda > 0$, $\sum_{\nu\geq 1}(1+\lambda/\mu_\nu)^{-2} \asymp h^{-1}$.

A.3. Proofs in Section 3.2

Our theoretical analysis relies on a set of Fréchet derivatives to be specified below: for $j = 1, 2, \ldots, s$, the Fréchet derivative of $\ell_{j,\lambda}$ can be identified as: for any $f, f_1, f_2 \in \mathcal{H}$,

$$\begin{aligned} D\ell_{j,\lambda}(f)f_1 &= -\frac{1}{n}\sum_{i\in I_j}(Y_i - f(X_i))\langle K_{X_i}, f_1\rangle + \langle \mathcal{P}_\lambda f, f_1\rangle := \langle S_{j,\lambda}(f), f_1\rangle, \\ DS_{j,\lambda}(f)f_1 f_2 &= \frac{1}{n}\sum_{i\in I_j} f_2(X_i)\langle K_{X_i}, f_1\rangle + \langle \mathcal{P}_\lambda f_2, f_1\rangle = \langle DS_{j,\lambda}(f)f_2, f_1\rangle, \\ D^2 S_{j,\lambda}(f) &\equiv 0. \end{aligned}$$

More specifically,

$$\begin{aligned} S_{j,\lambda}(f) &= -\frac{1}{n}\sum_{i\in I_j}(Y_i - f(X_i))K_{X_i} + \mathcal{P}_\lambda f, \\ DS_{j,\lambda}(f)g &= \frac{1}{n}\sum_{i\in I_j} g(X_i)K_{X_i} + \mathcal{P}_\lambda g. \end{aligned}$$

Define $S_\lambda(f) = E\{S_{j,\lambda}(f)\}$, hence, $DS_\lambda(f) = E\{DS_{j,\lambda}(f)\}$. It follows from [16] that

$$\langle DS_\lambda(f)f_1, f_2\rangle = \langle f_1, f_2\rangle$$

for any $f, f_1, f_2 \in \mathcal{H}$ which leads to $DS_\lambda(f) = id$.

A.3.1. Proof of Lemma 3.1

Proof. Throughout the proof, let $\widetilde{f}_j = E\{\widehat{f}_j|\mathbf{X}_j\}$. It is easy to see that

$$\begin{aligned} 0 &= S_{j,\lambda}(\widehat{f}_j) = -\frac{1}{n}\sum_{i\in I_j}(Y_i - \widehat{f}_j(X_i))K_{X_i} + \mathcal{P}_\lambda \widehat{f}_j, \\ 0 &= \frac{1}{n}\sum_{i\in I_j}(\widetilde{f}_j(X_i) - f_0(X_i))K_{X_i} + \mathcal{P}_\lambda \widetilde{f}_j. \end{aligned}$$

Subtracting the two equations one gets that

$$\frac{1}{n}\sum_{i\in I_j}(\widehat{f}_j - \widetilde{f}_j)(X_i)K_{X_i} + \mathcal{P}_\lambda(\widehat{f}_j - \widetilde{f}_j) = \frac{1}{n}\sum_{i\in I_j}\epsilon_i K_{X_i}. \tag{A.1}$$

Equation (A.1) shows that

$$\widehat{f}_j - \widetilde{f}_j = \operatorname*{argmin}_{f\in\mathcal{H}} \ell^\star_{j,\lambda}(f) \equiv \operatorname*{argmin}_{f\in\mathcal{H}} \frac{1}{2n}\sum_{i\in I_j}(\epsilon_i - f(X_i))^2 + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}}.$$

Let $e_j = \frac{1}{n}\sum_{i\in I_j}\epsilon_i K_{X_i}$ and $\varepsilon_j = \widehat{f}_j - \widetilde{f}_j$. Then consider Taylor's expansion

$$\begin{aligned}
\ell^\star_{j,\lambda}(e_j) - \ell^\star_{j,\lambda}(\varepsilon_j) &= \frac{1}{2}D^2\ell^\star_{j,\lambda}(\varepsilon_j)(e_j-\varepsilon_j)(e_j-\varepsilon_j)\\
&= \frac{1}{2}P_j(e_j-\varepsilon_j)^2 + \frac{1}{2}\langle\mathcal{P}_\lambda(e_j-\varepsilon_j), e_j-\varepsilon_j\rangle,\\
\ell^\star_{j,\lambda}(\varepsilon_j) - \ell^\star_{j,\lambda}(e_j) &= D\ell^\star_{j,\lambda}(e_j)(\varepsilon_j-e_j) + \frac{1}{2}D^2\ell^\star_{j,\lambda}(e_j)(\varepsilon_j-e_j)(\varepsilon_j-e_j)\\
&= (P_j-P)(e_j(\varepsilon_j-e_j)) + \frac{1}{2}P_j(\varepsilon_j-e_j)^2\\
&\quad +\frac{1}{2}\langle\mathcal{P}_\lambda(\varepsilon_j-e_j), \varepsilon_j-e_j\rangle.
\end{aligned}$$

Adding the two equations one obtains that

$$P_j(\varepsilon_j-e_j)^2 + \langle\mathcal{P}_\lambda(\varepsilon_j-e_j), \varepsilon_j-e_j\rangle + (P_j-P)(e_j(\varepsilon_j-e_j)) = 0.$$

Uniformly for j, it holds that

$$\begin{aligned}
|(P_j-P)(e_j(\varepsilon_j-e_j))| &\le \xi_j\|e_j\|\cdot\|\varepsilon_j-e_j\|,\\
P_j(\varepsilon_j-e_j)^2 + \langle\mathcal{P}_\lambda(\varepsilon_j-e_j), (\varepsilon_j-e_j)\rangle &\ge (1-\xi_j)\|\varepsilon_j-e_j\|^2.
\end{aligned}$$

Combining the two inequalities one gets that

$$(1-\xi_j)\|\varepsilon_j-e_j\|^2 \le \xi_j\|e_j\|\cdot\|\varepsilon_j-e_j\|.$$

Taking expectations conditional on $\mathbf{X}_j$ on both sides and noting that ξ_j is $\sigma(\mathbf{X}_j)$-measurable, one gets that

$$\begin{aligned}
(1-\xi_j)E\{\|\varepsilon_j-e_j\|^2|\mathbf{X}_j\} \le& \xi_j E\{\|e_j\|\cdot\|\varepsilon_j-e_j\||\mathbf{X}_j\}\\
\le& \xi_j E\{\|e_j\|^2|\mathbf{X}_j\}^{1/2}E\{\|\varepsilon_j-e_j\|^2|\mathbf{X}_j\}^{1/2}.
\end{aligned}$$

By assumption $\log^b N = o(nh^a)$ and Assumption 3.4, $\max_{1\le j\le s}\xi_j = o_P(1)$, i.e., with probability approaching one $\max_{1\le j\le s}\xi_j \le 1/2$, hence,

$$\begin{aligned}
E\{\|\varepsilon_j-e_j\|^2|\mathbf{X}_j\} &\le 4\xi_j^2 E\{\|e_j\|^2|\mathbf{X}_j\}\\
&= \frac{4\xi_j^2}{n^2}\sum_{i,i'\in I_j}E\{\epsilon_i\epsilon_{i'}K(X_i,X_{i'})|\mathbf{X}_j\}\\
&= \frac{4\xi_j^2}{n^2}\sum_{i\in I_j}\sigma^2(X_i)K(X_i,X_i)\\
&\le \frac{4c_\pi c_\varphi^2\xi_j^2}{nh}, \qquad \text{(A.2)}
\end{aligned}$$

where the last inequality follows from Assumption 3.1 and Lemma A.1 that $K(x,x)\le c_\varphi^2 h^{-1}$. This proves (3.1).

By (A.2), it is easy to derive

$$E\{\|\widehat{f}_j - \widetilde{f}_j\|^2 | \mathbf{X}_j\} \leq \frac{4c_\pi c_\varphi^2}{nh}. \tag{A.3}$$

Now we look at $\|\widetilde{f}_j - f_0^\star\|$, where $f_0^\star = (id - \mathcal{P}_\lambda) f_0$. Note that $\widetilde{f}_j$ is the minimizer of the following problem

$$\widetilde{f}_j = \operatorname*{argmin}_{f \in \mathcal{H}} \widetilde{\ell}_{j,\lambda}(f) \equiv \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{2n} \sum_{i \in I_j} (f_0(X_i) - f(X_i))^2 + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2.$$

We use a similar strategy for handling part (3.1). Note that

$$\begin{aligned}
\widetilde{\ell}_{j,\lambda}(f_0^\star) - \widetilde{\ell}_{j,\lambda}(\widetilde{f}_j) &= \frac{1}{2} D^2 \widetilde{\ell}_{j,\lambda}(\widetilde{f}_j)(f_0^\star - \widetilde{f}_j)(f_0^\star - \widetilde{f}_j) \\
&= \frac{1}{2} P_j (f_0^\star - \widetilde{f}_j)^2 + \frac{1}{2} \langle \mathcal{P}_\lambda (f_0^\star - \widetilde{f}_j), f_0^\star - \widetilde{f}_j \rangle, \\
\widetilde{\ell}_{j,\lambda}(\widetilde{f}_j) - \widetilde{\ell}_{j,\lambda}(f_0^\star) &= P_j (f_0^\star - f_0)(\widetilde{f}_j - f_0^\star) + \langle \mathcal{P}_\lambda f_0^\star, \widetilde{f}_j - f_0^\star \rangle \\
&\quad + \frac{1}{2} P_j (\widetilde{f}_j - f_0^\star)^2 + \frac{1}{2} \langle \mathcal{P}_\lambda (\widetilde{f}_j - f_0^\star), \widetilde{f}_j - f_0^\star \rangle.
\end{aligned}$$

Adding the two equations, one gets that

$$\begin{aligned}
& P_j (\widetilde{f}_j - f_0^\star)^2 + \langle \mathcal{P}_\lambda (\widetilde{f}_j - f_0^\star), \widetilde{f}_j - f_0^\star \rangle \\
= \; & P_j (f_0 - f_0^\star)(\widetilde{f}_j - f_0^\star) - \langle \mathcal{P}_\lambda f_0^\star, \widetilde{f}_j - f_0^\star \rangle \\
= \; & (P_j - P)(f_0 - f_0^\star)(\widetilde{f}_j - f_0^\star) + P(f_0 - f_0^\star)(\widetilde{f}_j - f_0^\star) - \langle \mathcal{P}_\lambda f_0^\star, \widetilde{f}_j - f_0^\star \rangle \\
= \; & (P_j - P)(f_0 - f_0^\star)(\widetilde{f}_j - f_0^\star) + \langle f_0 - f_0^\star, \widetilde{f}_j - f_0^\star \rangle \\
& - \langle \mathcal{P}_\lambda (f_0 - f_0^\star), \widetilde{f}_j - f_0^\star \rangle - \langle \mathcal{P}_\lambda f_0^\star, \widetilde{f}_j - f_0^\star \rangle \\
= \; & (P_j - P)(f_0 - f_0^\star)(\widetilde{f}_j - f_0^\star) + \langle f_0 - f_0^\star - \mathcal{P}_\lambda (f_0 - f_0^\star) - \mathcal{P}_\lambda f_0^\star, \widetilde{f}_j - f_0^\star \rangle \\
= \; & (P_j - P)(f_0 - f_0^\star)(\widetilde{f}_j - f_0^\star).
\end{aligned}$$

Therefore,

$$\begin{aligned}
(1 - \xi_j) \|\widetilde{f}_j - f_0^\star\|^2 \leq & \xi_j \|f_0 - f_0^\star\| \times \|\widetilde{f}_j - f_0^\star\| = \xi_j \|\mathcal{P}_\lambda f_0\| \times \|\widetilde{f}_j - f_0^\star\| \\
\leq & C \xi_j \lambda^{1/2} \|f_0\|_{\mathcal{H}} \|\widetilde{f}_j - f_0^\star\|,
\end{aligned}$$

implying that, with probability approaching one, for any $1 \leq j \leq s$, $\|\widetilde{f}_j - f_0^\star\| \leq 2C \xi_j \lambda^{1/2} \|f_0\|_{\mathcal{H}}$. This proves (3.2). □

A.3.2. Proof of Theorem 3.1

Proof. Recall $f_0^\star = (id - \mathcal{P}_\lambda) f_0$ and $\widetilde{f}_j = E(\widehat{f}_j | \mathbf{X}_j)$. Also notice that $\frac{1}{N} \sum_{i=1}^N \epsilon_i K_{X_i} = \frac{1}{s} \sum_{j=1}^s e_j$. By direct calculations and Lemma 3.1, we have

with probability approaching one,

$$\begin{aligned} & E\{\|\bar{f} - f_0^\star - \frac{1}{N}\sum_{i=1}^{N}\epsilon_i K_{X_i}\|^2|\mathbf{X}\} \\ = & \frac{1}{s^2}\sum_{j=1}^{s} E\{\|\widehat{f}_j - \widetilde{f}_j - e_j\|^2|\mathbf{X}_j\} + \frac{1}{s^2}\|\sum_{j=1}^{s}(\widetilde{f}_j - f_0^\star)\|^2 \\ \leq & 4\left(\frac{c_\pi c_\varphi^2}{Nh} + \lambda\|f_0\|_{\mathcal{H}}^2\right)\max_{1\leq j\leq s}\xi_j^2. \end{aligned}$$

This proves (3.3). The result (3.4) immediately follows by the assumption that $\max_{1\leq j\leq s}\xi_j^2 = o_P(1)$. □

A.4. Proofs in Section 3.3

A.4.1. Proof of Lemma 3.2

Proof. It is easy to see that

$$\boldsymbol{\epsilon}'\mathbf{K}\boldsymbol{\epsilon} = \sum_{i=1}^{N}\epsilon_i^2 K(X_i, X_i) + W(N).$$

Since

$$Var\left(\sum_{i=1}^{N}\epsilon_i^2 K(X_i, X_i)\right) \leq NE\{\epsilon_i^4 K(X_i, X_i)^2\} \leq \tau c_\varphi^4 N h^{-2},$$

where the last "≤" follows by Assumption 3.2 and Lemma A.1 that $K(x, x) \leq c_\varphi^2 h^{-1}$, we get that

$$\begin{aligned} \sum_{i=1}^{N}\epsilon_i^2 K(X_i, X_i) &= E\{\sum_{i=1}^{N}\epsilon_i^2 K(X_i, X_i)\} + O_P\left(\sqrt{c_\varphi^4 N h^{-2}}\right) \\ &= \sigma^2 N h^{-1} + O_P(\sqrt{c_\varphi^4 N h^{-2}}). \end{aligned}$$

Next we prove asymptotic normality of $W(N)$. Note $\sigma^2(N) = E\{W(N)^2\}$. Let G_I, G_{II}, G_{IV} be defined as

$$\begin{aligned} G_I &= \sum_{1\leq i<t\leq n} E\{W_{it}^4\}, \\ G_{II} &= \sum_{1\leq i<t<k\leq n}(E\{W_{it}^2 W_{ik}^2\} + E\{W_{ti}^2 W_{tk}^2\} + E\{W_{ki}^2 W_{kt}^2\}) \\ G_{IV} &= \sum_{1\leq i<t<k<l\leq n}(E\{W_{it}W_{ik}W_{lt}W_{lk}\} + E\{W_{it}W_{il}W_{kt}W_{kl}\} \\ &\quad + E\{W_{ik}W_{il}W_{tk}W_{tl}\}). \end{aligned}$$

Since $K(x,x) \le c_\varphi^2 h^{-1}$, we have $G_I = O(N^2 h^{-4})$ and $G_{II} = O(N^3 h^{-4})$. It can also be shown that for pairwise distinct i,k,t,l,

$$\begin{aligned} & E\{W_{ik}W_{il}W_{tk}W_{tl}\} \\ = \;& 2^4 E\{\epsilon_i^2\epsilon_k^2\epsilon_t^2\epsilon_l^2 K(X_i,X_k)K(X_i,X_l)K(X_t,X_k)K(X_t,X_l)\} \\ = \;& 2^4\sigma^8 \sum_{\nu=1}^{\infty} \frac{1}{(1+\lambda/\mu_\nu)^4} = O(h^{-1}), \end{aligned}$$

which implies that $G_{IV} = O(N^4 h^{-1})$. In the mean time, a straight algebra leads to that

$$\begin{aligned} \sigma^2(N) &= 4\sigma^4 \binom{N}{2} \sum_{\nu=1}^{\infty} \frac{1}{(1+\lambda/\mu_\nu)^2} \\ &= 2\sigma^4 N(N-1) \sum_{\nu\ge 1} \frac{1}{(1+\lambda/\mu_\nu)^2} \asymp N^2 h^{-1}, \end{aligned}$$

where the last conclusion follows by Proposition A.1. Thanks to the conditions $h \to 0$, $Nh^2 \to \infty$, G_I, G_{II} and G_{IV} are all of order $o(\sigma^4(N))$. Then it follows by [2] that as $N \to \infty$,

$$\frac{W(N)}{\sigma(N)} \xrightarrow{d} N(0,1).$$

The above limit leads to that $W(N) = O_P(Nh^{-1/2})$. □

A.4.2. Proof of Theorem 3.2

Proof. The proof is based on Lemma 3.2. Under $f_0 = 0$, it follows from Corollary 3.1 and Assumption 3.4 that

$$E\{\|\bar{f} - \frac{1}{N}\sum_{i=1}^{N} \epsilon_i K_{X_i}\|^2 | \mathbf{X}\} = O_P\left(\frac{c_\varphi^2 \log^b N}{Nh^{1+a}}\right),$$

leading to

$$\|\bar{f} - \frac{1}{N}\sum_{i=1}^{N} \epsilon_i K_{X_i}\|^2 = O_P\left(\frac{c_\varphi^2 \log^b N}{Nh^{1+a}}\right).$$

Following the proof of Lemma 3.1 and the trivial fact $\widehat{f}_j = 0$ when $f_0 = 0$, we have for any $1 \le j \le s$,

$$E\{\|\widehat{f}_j - e_j\|^2 | \mathbf{X}_j\} \le \frac{4c_\pi c_\varphi^2 \xi_j^2}{nh}, \quad E\{\|e_j\|^2 | \mathbf{X}_j\} \le \frac{c_\pi c_\varphi^2}{nh}, \quad \text{a.s.} \tag{A.4}$$

Therefore, by Cauchy-Schwartz inequality,

$$E\{|\langle \widehat{f}_j - e_j, e_j\rangle| \big| \mathbf{X}_j\} \le \sqrt{E\{\|\widehat{f}_j - e_j\|^2 | \mathbf{X}_j\} E\{\|e_j\|^2 | \mathbf{X}_j\}} \le \frac{2c_\pi c_\varphi^2}{nh}\xi_j,$$

and hence,

$$E\left\{\sum_{j=1}^{s}|\langle\widehat{f}_j-e_j,e_j\rangle|\Big|\mathbf{X}\right\}\leq\frac{2c_\pi sc_\varphi^2}{nh}\max_{1\leq j\leq s}\xi_j.$$

By Assumption 3.4, the above leads to that

$$\sum_{j=1}^{s}\langle\widehat{f}_j-e_j,e_j\rangle=O_P\left(\frac{sc_\varphi^2}{nh}\sqrt{\frac{\log^b N}{nh^a}}\right).$$

Meanwhile, it holds that

$$\sum_{j\neq l}\langle\widehat{f}_j-e_j,e_l\rangle=\sum_{j<l}\langle\widehat{f}_j-e_j,e_l\rangle+\sum_{j>l}\langle\widehat{f}_j-e_j,e_l\rangle\equiv R_1+R_2,$$

with

$$R_1=O_P\left(\frac{sc_\varphi^2}{nh}\sqrt{\frac{\log^b N}{nh^a}}\right),\quad R_2=O_P\left(\frac{sc_\varphi^2}{nh}\sqrt{\frac{\log^b N}{nh^a}}\right).$$

To see this, note that

$$\begin{aligned}E\{R_1^2|\mathbf{X}\}&=\sum_{j<l}E\{|\langle\widehat{f}_j-e_j,e_l\rangle|^2|\mathbf{X}\}\\&\leq\sum_{j<l}E\{\|\widehat{f}_j-e_j\|^2\|e_l\|^2|\mathbf{X}\}\\&=\sum_{j<l}E\{\|\widehat{f}_j-e_j\|^2|\mathbf{X}_j\}E\{\|e_l\|^2|\mathbf{X}_l\}\\&\leq\binom{s}{2}\frac{4c_\pi^2c_\varphi^4}{n^2h^2}\max_{1\leq j\leq s}\xi_j^2,\end{aligned}$$

where the last inequality is based on (A.4). Similar result holds for R_2. Hence, by Lemma 3.2 and direct algebra, we get that

$$\begin{aligned}T_{N,\lambda}&=N^{-2}\boldsymbol{\epsilon}'\mathbf{K}\boldsymbol{\epsilon}+\frac{2}{s^2}\sum_{j,l=1}^{s}\langle\widehat{f}_j-e_j,e_l\rangle+\|\bar{f}-\frac{1}{N}\sum_{i=1}^{N}\epsilon_iK_{X_i}\|^2\\&=N^{-2}\boldsymbol{\epsilon}'\mathbf{K}\boldsymbol{\epsilon}+\frac{2}{s^2}\sum_{j=1}^{s}\langle\widehat{f}_j-e_j,e_j\rangle+\frac{2}{s^2}(R_1+R_2)+\|\bar{f}-\frac{1}{N}\sum_{i=1}^{N}\epsilon_iK_{X_i}\|^2\\&=\frac{\sigma^2}{Nh}+\frac{W(N)}{N^2}+O_P\left(\frac{c_\varphi^2}{N^{3/2}h}\right)+O_P\left(\frac{c_\varphi^2}{Nh}\sqrt{\frac{\log^b N}{nh^a}}\right)+O_P\left(\frac{c_\varphi^2\log^b N}{Nnh^{1+a}}\right)\\&=\frac{\sigma^2}{Nh}+\frac{W(N)}{N^2}+O_P\left(\frac{c_\varphi^2}{N^{3/2}h}\right)+O_P\left(\frac{c_\varphi^2}{Nh}\sqrt{\frac{\log^b N}{nh^a}}\right).\end{aligned}$$

The last equality follows from the condition $\log^b N = o(nh^{a+1})$. Therefore, by $c_\varphi^4/(Nh) = o(1)$, $Nh \to \infty$ (from $Nh^2 \to \infty$ and $h \to 0$), condition $\log^b N = o(nh^{a+1})$ and $\sigma^2(N) \asymp N^2 h^{-1}$ (Lemma 3.2), as $N \to \infty$,

$$\begin{aligned}\frac{N^2}{\sigma(N)}\left(T_{N,\lambda} - \frac{\sigma^2}{Nh}\right) &= \frac{W(N)}{\sigma(N)} + O_P\left(\frac{c_\varphi^2}{\sqrt{Nh}} + c_\varphi^2\sqrt{\frac{\log^b N}{nh^{a+1}}}\right)\\ &= \frac{W(N)}{\sigma(N)} + o_P(1) \xrightarrow{d} N(0,1).\end{aligned}$$

Proof is completed. □

A.4.3. Proof of Theorem 3.3

Proof. For any $f \in \mathcal{H}$, define $R_f = \bar{f} - N^{-1}\sum_{i=1}^N \epsilon_i K_{X_i} - f + \mathcal{P}_\lambda f$. By direct examinations, it holds that

$$\begin{aligned}&\|\bar{f}\|^2 - \sigma^2/(Nh)\\ =\ &\|R_f + \frac{1}{N}\sum_{i=1}^N \epsilon_i K_{X_i} + f - \mathcal{P}_\lambda f\|^2 - \sigma^2/(Nh)\\ \geq\ &\left\{\boldsymbol{\epsilon}'\mathbf{K}\boldsymbol{\epsilon}/N^2 - \sigma^2/(Nh)\right\} + \|f - \mathcal{P}_\lambda f\|^2 - \frac{2}{N}\sum_{i=1}^N \epsilon_i(f - \mathcal{P}_\lambda f)(X_i)\\ &+\frac{2}{N}\sum_{i=1}^N \epsilon_i R_f(X_i) - 2\langle f - \mathcal{P}_\lambda f, R_f\rangle\\ \equiv\ &T_1 + T_2 + T_3 + T_4 + T_5.\end{aligned}$$

It follows by (3.5), Theorem 3.1, Assumption 3.4 that, uniformly for $f \in \mathcal{H}$,

$$\begin{aligned}&T_1 = W(N)/N^2 + O_P((N^{3/2}h)^{-1}), \quad \text{(by (3.5))}\\ &P_f\left(|T_3| \geq \sigma\|f - \mathcal{P}_\lambda f\|/(\varepsilon\sqrt{N})\right) \leq \varepsilon^2, \quad \text{for arbitrary } \varepsilon > 0\\ &T_4 = O_P(b_{N,\lambda}/\sqrt{Nh}), \quad \text{(by Theorem 3.1, Assumption 3.4 and (3.5))}\\ &T_5 = \|f - \mathcal{P}_\lambda f\| \times O_P(b_{N,\lambda}), \quad \text{(by Theorem 3.1 and Assumption 3.4)}\end{aligned}$$

Note that $\|\mathcal{P}_\lambda f\| \leq \lambda^{1/2}\|f\|_{\mathcal{H}}$ for any $f \in \mathcal{H}$. Therefore, to achieve high power, i.e., power is at least $1 - \varepsilon$, one needs to choose a large N_ε and C_ε s.t. $N \geq N_\varepsilon$ and

$$\begin{aligned}&\|f\| \geq C_\varepsilon/\sqrt{Nh^{1/2}}, \quad \|f\| \geq C_\varepsilon/\sqrt{N}, \quad \|f\| \geq C_\varepsilon\sqrt{b_{N,\lambda}/\sqrt{Nh}},\\ &\|f\| \geq C_\varepsilon b_{N,\lambda}, \quad \|f\| \geq C_\varepsilon\lambda^{1/2}\|f\|_{\mathcal{H}}.\end{aligned}$$

Proof is completed. □

A.5. Proofs in Section 3.4

A.5.1. Proof of Lemma 3.3 (a)

Proof. For each $\nu \geq 1$, there exist $p \in \mathbb{N}$ and $1 \leq k \leq d$, such that $\nu = pd + k$. Suppose $x = (x_1, \cdots, x_d)$, then for each x_k, there exists $(\varphi_p^{(k)}, \mu_p^{(k)})$ and $(\varphi_{p'}^{(k)}, \mu_{p'}^{(k)})$ satisfying $V_k(\varphi_p^{(k)}, \varphi_{p'}^{(k)}) = \delta_{pp'}$ and $\int_{\mathbb{I}} \varphi_p^{(k)}(x)\varphi_{p'}^{(k)}(x)dx = \delta_{pp'}/\mu_p^{(k)}$. In fact, the eigenfunctions φ_ν and eigenvalues μ_ν can be constructed by an ordered sequence of $\varphi_p^{(k)}, \mu_p^{(k)}$ as $\varphi_\nu(x) = \varphi_p^{(k)}(x_k)$ and $\mu_\nu = \mu_p^{(k)}$.

Next, we verify such construction of eigenfunctions φ_ν and eigenvalues μ_ν satisfy Assumption 3.3. When $\nu \neq \mu$, then there exist p_1, q_1, p_2, q_2, such that $\nu = p_1 d + q_1, \mu = p_2 d + q_2$, then

$$\begin{aligned} &V(\varphi_{p_1 d+q_1}, \varphi_{p_2 d+q_2}) \\ &\quad = V(\varphi_{p_1}^{q_1}(x_{q_1}), \varphi_{p_2}^{q_2}(x_{q_2})) \\ &\quad = \begin{cases} 0 & p_1 \neq p_2, q_1 = q_2 \\ V_{q_1}(\varphi_{p_1}^{q_1}(x_{q_1}), 0) + V_{q_2}(0, \varphi_{p_2}^{q_2}(x_{q_2})) = 0 & q_1 \neq q_2 \end{cases} \end{aligned}$$

On the other hand,

$$\langle \varphi_\nu, \varphi_\mu \rangle_{\mathcal{H}} = \langle \varphi_{p_1}^{q_1}, \varphi_{p_2}^{q_2} \rangle_{\mathcal{H}} = \begin{cases} 1/\mu_{p_1}^{q_1} = 1/\mu_\nu & p_1 = p_2, q_1 = q_2 \\ 0 & \nu \neq \mu \end{cases}$$

For any $f \in \mathcal{H}$,

$$\begin{aligned} f(x_1, \cdots, x_d) = f_1(x_1) + \cdots + f_d(x_d) &= \sum_{k=1}^{d} \sum_{\nu=1}^{\infty} V_k(f_k, \varphi_\nu^{(k)}) \varphi_\nu^{(k)}(x_k) \\ &= \sum_{k=1}^{d} \sum_{\nu=1}^{\infty} V(f, \varphi_\nu^{(k)}) \varphi_\nu^{(k)}(x_k) = \sum_{\nu=1}^{\infty} V(f, \varphi_\nu) \varphi_\nu(x) \end{aligned}$$

□

A.5.2. Proof of Lemma 3.3 (b)

Proof. It is easy to see that

$$\sum_{\nu \geq 1} (1 + \lambda/\mu_\nu)^{-1} = \sum_{q=1}^{d} \sum_{p \geq 1} (1 + \lambda/\mu_p^{(k)})^{-1} \asymp d\lambda^{-1/(2m)} := h^{-1}.$$

□

A.5.3. Proof of Lemma 3.3 (c)

Proof. Notice that $\|f\|_{\mathcal{H}}^2 \leq \sum_{i=1}^{d} \|f_k\|_{\mathcal{H}}^2 \leq Cd$, then by Lemma A.1 (b), $\|\mathcal{P}_\lambda f\|^2 \leq \lambda \|f\|_{\mathcal{H}}^2 \leq Cd\lambda$. □

Next, we prove Lemma 3.3 (d). To prove Lemma 3.3 (d), it is sufficient to prove the following Lemma A.2.

Lemma A.2. *Under (3.7), there exist universal positive constants c_1, c_2, c_3 such that for any $1 \le j \le s$,*

$$P\left(\xi_j \ge t\right) \le 2n \exp\left(-\frac{nht^2}{c_1 + c_2 t}\right), \text{ for all } t \ge c_3 (nh)^{-1},$$

where $h^{-1} \asymp d\lambda^{-1/(2m)}$.

The proof of Lemma 3.3 is based on the green function for equivalent kernel technique in [3], see Section B.3 for details.

A.5.4. Proof of Lemma 3.4

Proof. For $p, \delta > 0$, define $\mathcal{G}(p) = \{f \in \mathcal{H} : \|f\|_{\sup} \le 1, \|f\|_{\mathcal{H}} \le p\}$ and the corresponding entropy integral

$$J(p, \delta) = \int_0^\delta \psi_2^{-1}\left(D(\varepsilon, \mathcal{G}(p), \|\cdot\|_{\sup})\right) d\varepsilon + \delta \psi_2^{-1}\left(D(\delta, \mathcal{G}(p), \|\cdot\|_{\sup})^2\right), \quad \text{(A.5)}$$

where $\psi_2(s) = \exp(s^2) - 1$ and $D(\varepsilon, \mathcal{G}(p), \|\cdot\|_{\sup})$ is the ε-packing number of $\mathcal{G}(p)$ in terms of $\|\cdot\|_{\sup}$-metric. In what follows, we particularly choose $p = c_K^{-1}(h/\lambda)^{1/2}$, where $c_K \equiv \sup_{g \in \mathcal{H}} h^{1/2}\|g\|_{\sup}/\|g\|$ is finite, according to [27].

Define $\psi_i(g) = c_k^{-1} h^{1/2} g(X_i)$ and $Z_j(g) = n^{-1/2} \sum_{i \in I_j} [\psi_i(g) K_{X_i} - E\{\psi_i(g) K_{X_i}\}]$. Following [27, Lemma 6.1], for any $1 \le j \le s$, for any $t \ge 0$,

$$P\left(\sup_{g \in \mathcal{G}(p)} \|Z_j(g)\| \ge t\right) \le 2 \exp\left(-\frac{t^2}{C^2 J(p,1)^2}\right), \quad \text{(A.6)}$$

for an absolute constant $C > 0$. Since $\|f\| = 1$ implies that $c_K^{-1} h^{1/2} f \in \mathcal{G}(p)$. Then it can be shown that

$$\sqrt{n}\xi_j \le c_K^2 h^{-1} \sup_{g \in \mathcal{G}(p)} \|Z_j(g)\|, \quad j = 1, \ldots, s.$$

Following (A.6) we have

$$P\left(\sqrt{n} \max_{1 \le j \le s} \xi_j \ge t\right) \le 2s \exp\left(-\frac{c_K^{-4} h^2 t^2}{C^2 J(p,1)^2}\right),$$

which implies that

$$\sqrt{n} \max_{1 \le j \le s} \xi_j = O_P\left(\sqrt{\frac{\log N}{h^2}} J(p,1)\right). \quad \text{(A.7)}$$

It follows by [31, Proposition 1] that $J(p,1) = O\left([\log(h/\lambda)]^{(d+1)/2}\right) = O\left([\log N]^{(d+1)/2}\right)$. Then

$$\max_{1 \le j \le s} \xi_j = O_P\left(\sqrt{\frac{\log^{d+2} N}{nh^2}}\right).$$

That is, Assumption 3.4 holds with $a = 2$ and $b = d + 2$. Proof completed. □

A.5.5. Proof of Lemma 3.5

Proof.

$$\begin{aligned} J(p,1) &\le \int_0^1 \sqrt{\log D(\varepsilon, \mathcal{G}, \|\cdot\|_{\sup})}\, d\varepsilon + \sqrt{\log D(1, \mathcal{G}, \|\cdot\|_{\sup})} \\ &\le \int_0^1 \sqrt{\left(\frac{p}{\varepsilon}\right)^{\frac{d}{m}} + 1}\, d\varepsilon + \sqrt{2} p^{\frac{d}{2m}} \\ &\le c_d' p^{d/(2m)} \end{aligned}$$

where the penultimate step is based on [14]. Therefore, $J(p,1) = O(p^{\frac{d}{2m}})$, where $p = (h/\lambda)^{1/2}$. From e.q. (A.7), we have

$$\max_{1 \le j \le s} \xi_j = O_P\left(\sqrt{\frac{\log N}{n h^{3-d/(2m)}}}\right) \qquad \square$$

Appendix B: Some technical proofs and auxiliary lemmas

B.1. Proof of Proposition 3.1

Proof. Define

$$s_\lambda = \operatorname{argmin}\{j : \mu_j \le \lambda\} - 1,$$

that is, s_λ is the number of eigenvalues that are greater than λ. Then the effective dimension can be written as

$$h^{-1} = \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda} = \sum_{j=1}^{s_\lambda} \frac{\mu_j}{\mu_j + \lambda} + \sum_{j=s_\lambda+1}^{\infty} \frac{\mu_j}{\mu_j + \lambda}.$$

Note that $\sum_{j=1}^{s_\lambda} \mu_j/(\mu_j + \lambda) \le s_\lambda$, then we have

$$s_\lambda \le h^{-1} \le s_\lambda + \sum_{j=s_\lambda+1}^{\infty} \frac{\mu_j}{\mu_j + \lambda} \le s_\lambda + \frac{1}{\lambda} \sum_{j=s_\lambda+1}^{\infty} \mu_j. \tag{B.1}$$

By Assumption 3.3, we have $\sum_{j=s_\lambda+1}^{\infty} \mu_j \le C s_\lambda \mu_{s_\lambda} \le s_\lambda \lambda$. Therefore, by (B.1), we have $h^{-1} \asymp s_\lambda$. Next we show $\sum_{\nu \ge 1} (1 + \lambda/\mu_\nu)^{-2} \asymp h^{-1}$.

Note that

$$\sum_{\nu \ge 1} (1 + \lambda/\mu_\nu)^{-2} = \sum_{j=1}^{\infty} \frac{\mu_j^2}{(\mu_j + \lambda)^2} = \sum_{j=1}^{s_\lambda} \left(\frac{\mu_j}{\mu_j + \lambda}\right)^2 + \sum_{j=s_\lambda+1}^{\infty} \left(\frac{\mu_j}{\mu_j + \lambda}\right)^2,$$

similar to (B.1), we have

$$s_\lambda \le \sum_{\nu \ge 1} (1 + \lambda/\mu_\nu)^{-2} \le s_\lambda + \sum_{j=s_\lambda+1}^{\infty} \left(\frac{\mu_j}{\mu_j + \lambda}\right)^2 \le s_\lambda + \frac{1}{\lambda^2} \sum_{j=s_\lambda+1}^{\infty} \mu_j^2.$$

Since $\frac{1}{\lambda^2} \sum_{j=s_\lambda+1}^{\infty} \mu_j^2 \le \frac{\mu_{s_\lambda+1}}{\lambda^2} \sum_{j=s_\lambda+1}^{\infty} \mu_j \le \frac{1}{\lambda} \sum_{j=s_\lambda+1}^{\infty} \mu_j \le s_\lambda$. Then we have $\sum_{\nu \ge 1} (1 + \lambda/\mu_\nu)^{-2} \asymp s_\lambda$. Based on the previous conclusion that $h^{-1} \asymp s_lambda$, we finally get $\sum_{\nu \ge 1} (1 + \lambda/\mu_\nu)^{-2} \asymp h^{-1}$. $\square$

B.2. Verification of Assumption 3.3

Let us verify Assumption 3.3 in polynomially decaying kernels (PDK) and exponentially decaying kernels (EDK).

First consider PDK with $\mu_i \asymp i^{-2m}$ for a constant $m > 1/2$ which includes kernels of Sobolev space and Besov Space. An m-th order Sobolev space, denoted $\mathcal{H}^m([0,1])$, is defined as

$$\mathcal{H}^m([0,1]) = \{f : [0,1] \to \mathbb{R} | f^{(j)} \text{ is abs. cont for } j = 0, 1, \cdots, m-1, \text{ and } f^m \in L_2([0,1])\}.$$

An m-order periodic Sobolev space, denoted $H_0^m(\mathbb{I})$, is a proper subspace of $\mathcal{H}^m([0,1])$ whose element fulfills an additional constraint $g^{(j)}(0) = g^{(j)}(1)$ for $j = 0, \ldots, m-1$. The basis functions φ_i's of $H_0^m(\mathbb{I})$ are

$$\varphi_i(z) = \begin{cases} \sigma, & i = 0, \\ \sqrt{2}\sigma \cos(2\pi k z), & i = 2k, k = 1, 2, \ldots, \\ \sqrt{2}\sigma \sin(2\pi k z), & i = 2k-1, k = 1, 2, \ldots. \end{cases}$$

The corresponding eigenvalues are $\mu_{2k} = \mu_{2k-1} = \sigma^2 (2\pi k)^{-2m}$ for $k \geq 1$ and $\mu_0 = \infty$. In this case, $\sup_{i \geq 1} \|\varphi\|_{\sup} < \infty$. For any $k \geq 1$,

$$\sum_{i=k+1}^{\infty} \mu_i \lesssim \int_k^{\infty} x^{-2m} dx = \frac{k^{1-2m}}{2m-1} \lesssim \frac{k\mu_k}{2m-1}.$$

Therefore, there exists a constant $C < \infty$, such that

$$\sup_{k \geq 1} \frac{\sum_{i=k+1}^{\infty} \mu_i}{k\mu_k} = C < \infty.$$

Hence, Assumption 3.3 holds true.

Next, let us consider EDK with $\mu_i \asymp \exp(-\gamma i^p)$ for constants $\gamma > 0$ and $p > 0$. Gaussian kernel $K(x, x') = \exp\left(-(x-x')^2/\sigma^2\right)$ is an EDK of order $p = 2$, with eigenvalues $\mu_i \asymp \exp(-\pi i^2)$ as $i \to \infty$, and the corresponding eigenfunctions

$$\varphi_i(x) = (\sqrt{5}/4)^{1/4} (2^{i-1} i!)^{-1/2} e^{-(\sqrt{5}-1)x^2/4} H_i((\sqrt{5}/2)^{1/2} x),$$

where $H_i(\cdot)$ is the i-th Hermite polynomial; see [19] for more details. Then $\sup_{i \geq 1} \|\varphi_i\|_{\sup} < \infty$ trivially holds. For any $k \geq 1$,

$$\sum_{i=k+1}^{\infty} \mu_i \lesssim \int_k^{\infty} e^{-\gamma x^p} dx = \frac{1}{\gamma p k^{p-1}} e^{-\gamma k^p} - \int_k^{\infty} \frac{p-1}{\gamma p x^p} e^{-\gamma x^p} dx \leq \frac{1}{\gamma p k^{p-1}} e^{-\gamma k^p}.$$

Therefore,

$$\sup_{k \geq 1} \frac{\sum_{i=k+1}^{\infty} \mu_i}{k\mu_k} < \infty.$$

Hence, Assumption 3.3 holds.

B.3. Proof of Lemma A.2

To prove Lemma A.2, based on Lemma 3.5 and Lemma 3.4 in Chapter 21 in [3], we only need to bound

$$\|\frac{1}{n}\sum_{i=1}^{n} K_h(X_i,\cdot) - \mathrm{E}[K_h(X_i),\cdot]\|_\infty,$$
$$\text{and} \quad \|\frac{1}{n}\sum_{i=1}^{n} hK_h'(X_i,\cdot) - h\,\mathrm{E}[K_h'(X_i),\cdot]\|_\infty.$$

To prove this, we use the Green function technic tool that replace the kernels $K_h(x,y)$ with the one-sided exponential family $g_h(x-y)$, defined by

$$g(x) = \exp(-x)1_{\{x\geq 0\}}$$
$$g_h(x) = h^{-1}g(h^{-1}x), \quad x\in\mathbb{R}.$$

Lemma B.1. *Assume that the family $K_h = \sum_{j=1}^{d} K_{h_0,j}$ with $K_{h_0,j}$, $0 < h_0 \leq 1$ is convolution-like. Then there exists a constant c, such that, for all h, $0 < h \leq 1$, and for every strictly positive design $X_1, X_2, \cdots, X_n \in (0,1]^d$,*

$$\|\frac{1}{n}\sum_{i=1}^{n} K_h(X_i,\cdot)\|_\infty \leq c\|\frac{1}{n}\sum_{i=1}^{n} g_h(X_i-\cdot)\|_\infty.$$

Proof. For $t = (t_1,\cdots,t_d)\in[0,1]^d$ and $x = (x_1,\cdots,x_d)\in[0,1]^d$, let $S_{nh}(t) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i,t)$, and $s^{nh}(t) = \frac{1}{n}\sum_{i=1}^{n} g_h(X_i-t)$. For $j = 1,\cdots,d$, $K_{h,j}$ satisfies

$$\begin{aligned}&K_{h_0,j}(t_j,x_j)\\ =&h_0 g_{h,j}(x)K_{h_0,j}(t_j,0) + \int_0^1 g_{h_0,j}(x_j-z_j)\{h_0K_{h_0,j}'(t_j,z_j) + K_{h_0,j}(t_j,z_j)\}dz_j,\end{aligned}$$

where $h_0 = dh$. Note that $K_{h_0,j}$, $h_0K_{h_0,j}'$ are all convolutional-like, then $|h_0K_{h_0,j}'(t_j,z_j)| \leq ch_0^{-1}$ and $|K_{h_0,j}(t_j,z_j)| \leq ch_0^{-1}$. Therefore,

$$\begin{aligned}&\int_0^1 g_{h_0,j}(x_j-z_j)\{h_0K_{h_0,j}'(t_j,z_j) + K_{h_0,j}(t_j,z_j)\}dz_j\\ &\leq 2c\cdot h_0^{-1}\int_0^1 g_{h_0,j}(x_j-z_j)dz_j\\ &= 2c\cdot h_0^{-2}\int_0^1 e^{-h_0^{-1}(x_j-z_j)}dz_j\\ &\leq 2c\cdot\big(g_{h_0,j}(x_j) - g_{h_0,j}(x_j-1)\big) \leq 2c\cdot g_{h_0,j}(x_j).\end{aligned}$$

Then, we have $K_{h_0,j}(t_j,x_j) \leq h_0\cdot g_{h_0,j}(x)K_{h_0,j}(t_j,0) + c\cdot g_{h_0,j}(x_j)$.

$$K_h(x,t) = \sum_{j=1}^{d} K_{h_0,j}(t_j,x_j) \leq h_0\sum_{j=1}^{d} g_{h_0,j}(x_j)K_{h_0,j}(t_j,0) + c\sum_{j=1}^{d} g_{h_0,j}(x_j)$$

$$\leq c_1 \sum_{j=1}^{d} g_{h_0,j}(x_j) + c \sum_{j=1}^{d} g_{h_0,j}(x_j) \leq c' \sum_{j=1}^{d} g_{h_0,j}(x_j) = c' g_h(x),$$

where $c_1 = \max\{h_0 K_{h_0,1}(t_1, 0), \cdots, h_0 K_{h_0,d}(t_d, 0)\}$ is a bounded constant by the convolution-like assumption. Let $X_i = x$ and substitute the formula above into the expression for $S_{nh}(t)$ and $s^{nh}(t)$, this gives $S_{nh}(t) \leq c' s^{nh}(0)$. Therefore, $\|S_{nh}\|_\infty \leq c' |s^{nh}(0)| \leq \|s^{nh}\|_\infty$. The last inequality is due to the fact that all X_i are strictly positive, then $s^{nh}(t)$ is continuous at $t = 0$, and so $s^{nh}(0) \leq \|s^{nh}\|_\infty$. □

Let P_n be the empirical distribution function of the design $X_1, X_2, \cdots, X_n$, and let P_0 be the design distribution function. Here $P_0 = \pi(x)$. Define

$$[g_h \circledast (dP_n - dP_0)](t) = \int_{[0,1]^d} g_h(x - t)(dP_n(x) - dP_0(x)),$$

then based on Lemma B.1, we only need to show the following results to prove Lemma A.2.

Lemma B.2. *For all $x = (x_1, \cdots, x_d) \in [0, 1]^d, t > 0$,*

$$\mathbb{P}\Big[|[g_h \circledast (dP_n - dP_0)](x)| > t \Big] \leq 2 \exp\Big\{ - \frac{nht^2}{w_2 + 2/3t} \Big\}, \tag{B.2}$$

where w_2 is an upper bound on the density $P_0(x)$.

Proof. Consider for fixed x, $\frac{1}{n}\sum_{i=1}^{n} g_h(X_i - x) = \sum_{k=1}^{d} \sum_{i=1}^{n} \theta_{ik}$, with $\theta_{ik} = \frac{1}{n} g_{h_0,k}(x_{i,k} - x_k)$. Then θ_{ik} $(i = 1, \cdots, n; k = 1, \cdots, d)$ are i.i.d. and $|\theta_{ik}| \leq (nh_0)^{-1}$, where $h_0 = d^{-1}h$. For the variance $\mathrm{Var}(\theta_{ik})$,

$$\begin{aligned}
\mathrm{Var}(\theta_{ik}) =& \frac{1}{n^2} \big\{ [g^2_{h_0,k} \circledast dP_0](x_k) - ([g_{h_0,k} \circledast dP_0](x))^2 \big\} \\
\leq& \frac{1}{n^2} \big[g^2_{h_0,k} \circledast dP_0 \big](x_k) \\
=& n^{-2} \int_0^1 h_0^{-2} e^{-2h_0^{-1}(X_{ik} - x_k)} dP_0(x_k) \\
\leq& \frac{1}{2} w_2 n^{-2} h_0^{-1}.
\end{aligned}$$

Therefore, $V := \sum_{i=1}^{n} \sum_{k=1}^{d} \mathrm{Var}(\theta_{ik}) \leq \frac{1}{2} w_2 n^{-1} h^{-1}$. Then by Bernstein's inequality, (B.2) has been proved. □

Lemma B.3. *For all $j = 1, \cdots, n$,*

$$\mathbb{P}\{[g_h \circledast (dP_n - dP_0)](X_j) > t\} \leq 2 \exp\{ - \frac{1/4nht^2}{w_2 + 2/3t} \},$$

provided $t \geq 2(1 + w_2)(nh)^{-1}$, where w_2 is an upper bound on the density.

Proof. Consider $j = n$. Note that

$$\begin{aligned}[g_h \circledast dP_n](X_n) &= \frac{1}{n} g_h(0) + \frac{1}{n}\sum_{i=1}^{n-1} g_h(X_i - X_n) \\ &= \frac{1}{n}\sum_{k=1}^{d} g_{h_0,k}(0) + \frac{1}{n}\sum_{i=1}^{n-1} g_h(X_i - X_n) \\ &= d(nh_0)^{-1} + \frac{n-1}{n}[g_h \circledast dP_{n-1}](X_n),\end{aligned}$$

so that its expectation, conditional on X_n, equals

$$\mathbb{E}\big[[g_h \circledast dP_n](X_n)|X_n\big] = (nh)^{-1} + \frac{n-1}{n}[g_h \circledast dP_0](X_n).$$

Then $\mathbb{P}\big[|[g_h \circledast (dP_{n-1} - dP_0)](X_n)| > t|X_n\big] \le 2\exp\{-\frac{(n-1)ht^2}{w_2 + 2/3t}\}$. Note that this upper bound does not involve X_n, it follows that

$$\mathbb{P}\big[|[g_h \circledast (dP_{n-1} - dP_0)](X_n) > t|\big] = \mathbb{E}\Big[\mathbb{P}\big[|[g_h \circledast (dP_{n-1} - dP_0)](X_n)| > t|X_n\big]\Big]$$

has the same bound. Finally, note that

$$[g_h \circledast (dP_n - dP_0)](X_n) = \varepsilon_{nh} + \frac{n-1}{n}[g_h \circledast (dP_{n-1} - dP_0)](X_n),$$

where $|\varepsilon_{nh}| = |(nh)^{-1} - \frac{1}{n}[g_h \circledast dP_0](X_n)| \le (nh)^{-1} + (nh)^{-1}w_2 \le c_2(nh)^{-1}$. Therefore,

$$\begin{aligned}&\mathbb{P}\Big\{\big[|g_h \circledast (dP_n - dP_0)\big](X_n)| > t\Big\} \\ \le&\mathbb{P}\Big\{|\big[g_h \circledast (dP_{n-1} - dP_0)\big](X_n)| > \frac{n}{n-1}(t - c_2(nh)^{-1})\Big\} \\ \le&2\exp\Big\{-\frac{nh(t - c_2(nh)^{-1})^2}{w_2 + 2/3(t - c_2(nh)^{-1})}\Big\}.\end{aligned}$$

□

B.4. Proof of Corollary 3.2

Note that for any $x, y \in [0,1]^d$, by Lemma A1, we have $K(x,y) \le c_\varphi^2 h^{-1}$, where $h^{-1} \asymp d\lambda^{-1/(2m)}$, and $\|\mathcal{P}_\lambda f\|^2 \le \lambda\|f\|_{\mathcal{H}}^2 \le Cd\lambda$, then Corollary 3.2 can be easily achieved by applying Theorem 3.1 and Theorem 3.3.

Next, we show that $d^*_{N,\lambda,d} = d^{\frac{2m+1}{2(4m+1)}}N^{-\frac{2m}{4m+1}}$ is the minimax testing rate. Consider the model

$$\widetilde{y} = \theta + w, \tag{B.3}$$

where $\theta \in \mathbb{R}^n$ satisfies the ellipse constraint $\sum_{j=1}^n \frac{\theta_j^2}{\mu_j} \le d$, where $\mu_1 \ge \mu_2 \ge \cdots \ge 0$, and the noise vector w is zero-mean with variance $\frac{\sigma^2}{n}$. Note that model

(2.1) is equivalent to model (B.3) (see Example 3 in [24] for details), thus we only need to prove the minimax testing rate under model (B.3) for the testing problem $\theta = 0$ with $\mu_j \asymp \lceil \frac{i}{d} \rceil^{-2m}$.

Let $m_u(\delta;\varepsilon) := \operatorname{argmax}_{1\le k\le d}\{d\mu_k \ge \frac{1}{2}\delta^2\}$, and $m_l(\delta;\varepsilon) := \operatorname{argmax}_{1\le k\le d}\{d\mu_{k+1} \ge \frac{9}{16}\delta^2\}$. Then by Corollary 1 in [24], we have

$$\sup\{\delta|\delta \le \frac{1}{4}\sigma^2\frac{\sqrt{m_l(\delta;\varepsilon)}}{\delta}\} \le d^*_{N,\lambda,d} \le \inf\{\delta|\delta \ge c\sigma^2\frac{\sqrt{m_u(\delta;\varepsilon)}}{\delta}\}.$$

Let δ^* satisfies $\delta^2 \asymp \sqrt{m_l(\delta;\varepsilon)} \asymp \sqrt{m_u(\delta;\varepsilon)}$, we have

$$\delta^* = d^*_{N,\lambda,d} \asymp d^{\frac{2m+1}{2(4m+1)}} N^{-\frac{2m}{4m+1}}.$$

References

[1] Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. MR2166554

[2] Peter de Jong. A central limit theorem for generalized quadratic forms. *Probability Theory and Related Fields*, 75(2):261–277, 1987. MR0885466

[3] PPB Eggermont and VN LaRiccia. *Maximum penalized likelihood estimation*, volume 2. Springer, 2009. MR2817245

[4] Jianqing Fan, Chunming Zhang, and Jian Zhang. Generalized likelihood ratio statistics and wilks phenomenon. *The Annals of Statistics*, pages 153–193, 2001. MR1833962

[5] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2016. MR3588285

[6] Chong Gu. *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media, 2013. MR3025869

[7] Yuri Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science & Business Media, 2012. MR1991446

[8] Yuri I Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. i, ii, iii. *Math. Methods Statist*, 2(2):85–114, 1993. MR1257978

[9] Junwei Lu, Guang Cheng, and Han Liu. Nonparametric heterogeneity testing for massive data. *arXiv preprint* arXiv:1601.06212, 2016.

[10] Lukas Meier, Sara Van de Geer, Peter Bühlmann, et al. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009. MR2572443

[11] Shahar Mendelson. Geometric parameters of kernel machines. In *International Conference on Computational Learning Theory*, pages 29–43. Springer, 2002. MR2040403

[12] Ha Quang Minh, Partha Niyogi, and Yuan Yao. Mercer's theorem, feature maps, and smoothing. In *International Conference on Computational Learning Theory*, pages 154–168. Springer, 2006. MR2280604

[13] Stanislav Minsker, et al. Distributed statistical estimation and rates of convergence in normal approximation. *Electronic Journal of Statistics*, 13(2):5213–5252, 2019. MR4043072

[14] Tomaso Poggio and Christian R Shelton. On the mathematical foundations of learning. *American Mathematical Society*, 39(1):1–49, 2002.

[15] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13(Feb):389–427, 2012. MR2913704

[16] Zuofeng Shang and Guang Cheng. Local and global asymptotic inference in smoothing spline models. *The Annals of Statistics*, 41(5):2608–2638, 2013. MR3161439

[17] Zuofeng Shang and Guang Cheng. Computational limits of a distributed algorithm for smoothing spline. *The Journal of Machine Learning Research*, 18(1):3809–3845, 2017. MR3725447

[18] Zuofeng Shang, Botao Hao, and Guang Cheng. Nonparametric bayesian aggregation for massive data. *Journal of Machine Learning Research*, 20(140):1–81, 2019. MR4030154

[19] Peter Sollich and Christopher KI Williams. Understanding gaussian process regression using the equivalent kernel. In *Deterministic and statistical methods in machine learning*, pages 211–228. Springer, 2005.

[20] Sanvesh Srivastava, Cheng Li, and David B Dunson. Scalable bayes via barycenter in wasserstein space. *The Journal of Machine Learning Research*, 19(1):312–346, 2018. MR3862415

[21] Charles J Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, pages 689–705, 1985. MR0790566

[22] Botond Szabó and Harry van Zanten. An asymptotic analysis of distributed nonparametric methods. *Journal of Machine Learning Research*, 20(87):1–30, 2019. MR3960941

[23] Sebastian Weber, Andrew Gelman, Daniel Lee, Michael Betancourt, Aki Vehtari, and Amy Racine-Poon. Bayesian aggregation of average data: An application in drug development. *The Annals of Applied Statistics*, 12(3):1583–1604, 2018. MR3852689

[24] Yuting Wei and Martin J Wainwright. The local geometry of testing in ellipses: Tight control via localized kolmogorov widths. *IEEE Transactions on Information Theory*, 2020.

[25] Ganggang Xu, Zuofeng Shang, and Guang Cheng. Optimal tuning for divide-and-conquer kernel ridge regression with massive data. In *International Conference on Machine Learning*, volume 80, pages 5483–5491. PMLR, 2018.

[26] Yun Yang, Mert Pilanci, Martin J Wainwright, et al. Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics*, 45(3):991–1023, 2017. MR3662446

[27] Yun Yang, Zuofeng Shang, and Guang Cheng. Non-asymptotic theory for nonparametric testing. *In Conference on Learning Theory, to appear*, 2020.

[28] Ming Yuan, Ding-Xuan Zhou, et al. Minimax optimal rates of estimation in high dimensional additive models. *The Annals of Statistics*, 44(6):2564–

2593, 2016. MR3576554

[29] Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005. MR2175849

[30] Yuchen Zhang, John C Duchi, and Martin J Wainwright. Divide and conquer kernel ridge regression. In *Conference on Learning Theory*, pages 592–617, 2013. MR3450540

[31] Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002. MR1928805