Proceedings of Machine Learning Research vol 99:1–33, 2019 32nd Annual Conference on Learning Theory

Sharp Theoretical Analysis for Nonparametric Testing under Random Projection

Meimei Liu MEIMEI.LIU@DUKE.EDU
Department of Statistical Science, Duke University

Zuofeng Shang ZUOFENGSHANG@GMAIL.COM
Department of Mathematical Sciences, IUPUI & NJIT

Guang Cheng CHENGG@PURDUE.EDU
Department of Statistics, Purdue University

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

A common challenge in nonparametric inference is its high computational complexity when data volume is large. In this paper, we develop computationally efficient nonparametric testing by employing a random projection strategy. In the specific kernel ridge regression setup, a simple distance-based test statistic is proposed. Notably, we derive the minimum number of random projections that is sufficient for achieving testing optimality in terms of the minimax rate. As a by-product, the lower bound of projection dimension for minimax optimal estimation derived in [40] is proven to be sharp. One technical contribution is to establish upper bounds for a range of tail sums of empirical kernel eigenvalues.

Keywords: Computational limit, kernel ridge regression, minimax optimality, nonparametric testing, random projection.

1. Introduction

Computationally efficient statistical methods have been proposed for analyzing massive data sets. Examples include divide-and-conquer ([41; 19; 10; 32]); random projection ([25; 23; 40; 17]); subsampling ([20; 24; 2]); Nyström approximations ([14; 27]); and online learning ([5; 29; 12]). In particular, [40] studied minimax optimal estimation of kernel ridge regression (KRR) under random projections, which dramatically reduces computational and storage costs compared to ordinary KRR. An interesting question arising from this new method is the minimal computational cost required for obtaining statistically optimal results. This might be viewed as a type of "computational limit" from statistical perspective. This paper targets on this problem. We will develop computational limits for the projected KRR estimation and a relating hypothesis testing procedure. Our testing procedure has potential applications in massive data. Traditional nonparametric testing methods such as locally most powerful test, generalized/penalized likelihood ratio test and distance-based test [11; 22; 13; 31; 3] may not apply to massive data due to their high computational costs.

We consider the following nonparametric model

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \cdots, n, \tag{1.1}$$

where $x_i \in \mathcal{X} \subseteq \mathbb{R}^a$ for a fixed $a \geq 1$ are i.i.d. random design points, and ϵ_i are i.i.d. random noise with mean zero and variance σ^2. The regression function f is assumed to belong to a reproducing

kernel Hilbert space (RKHS) $\mathcal{H}$. Traditional kernel ridge regression (KRR) method estimates f through the following penalized least squares:

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

where $\|f\|_{\mathcal{H}}^2 = \langle f, f\rangle_{\mathcal{H}}$ with $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ the inner product of $\mathcal{H}$, $\lambda > 0$ is a smoothing parameter. When n is large, (1.2) is computationally complex due to the enormous dimension of the kernel matrix; computational and storage costs of $\widehat{f}_n$ are of orders $\mathcal{O}(n^3)$ and $\mathcal{O}(n^2)$, respectively. Recently, [40] proposed a projected KRR estimator, denoted $\widehat{f}_R$, in which KRR is fitted based on a randomly projected kernel matrix rather than the original one. Their method successfully reduces computational and storage costs to $\mathcal{O}(s^3)$ and $\mathcal{O}(s^2)$, respectively, when $s(\ll n)$ random projections are being used. The problem of computational limit amounts to characterizing the minimal choice of s such that the projected KRR method maintains statistical optimality.

1.1. Our Contributions

We consider the following nonparametric testing problem

$$H_0 : f = f_0 \ \text{ vs. } H_1 : f \in \mathcal{H}\setminus\{f_0\}, \tag{1.3}$$

where f_0 is a hypothesized function. We construct a test statistic $T_{n,\lambda} = \|\widehat{f}_R - f_0\|_n^2$, i.e., the squared empirical distance between $\widehat{f}_R$ and f_0, and derive a lower bound for s, denoted s^*, such that $T_{n,\lambda}$ achieves optimal testing rate. As a by-product, we also derive a lower bound for s, denoted $s^\dagger$, such that $\widehat{f}_R$ achieves optimal estimation rate. Table 1 summarizes the values of of $s^\dagger$ and s^* under both polynomially decaying kernel (PDK) and exponentially decaying kernel (EDK). We further prove the *sharpness* of $s^\dagger$ and s^* in the following sense:

(1) if $s = o(s^\dagger)$, there exist s random projections (satisfying Assumption A3) such that $\widehat{f}_R$ is sub-optimal for some f_0;

(2) if $s = o(s^*)$, there exist s random projections (satisfying Assumption A3) such that $T_{n,\lambda}$ fails to achieve high power even though the local alternatives are separated from some f_0 by optimal testing rate.

	$s^\dagger$	s^*
m-order PDK	$n^{\frac{1}{2m+1}}$	$n^{\frac{2}{4m+1}}$
p-order EDK	$(\log n)^{1/p}$	$(\log n)^{1/p}$

Table 1: *Values of $s^\dagger$ and s^* for PDK and EDK. Results summarized from Section 4.4.*

We illustrate our main findings in Figure 1.1, where the strength of the weakest detectable signals (SWDS (see Section 4.3 for detailed definition)) is characterized given any choice of s and λ. In general, we require $s \geq s_\lambda$ for any $\lambda > 0$, where s_λ is the number of kernel eigenvalues above λ. An important observation is that the smallest SWDS can be achieved at $\lambda = \lambda^*$ and $s \geq s_{\lambda^*} := s^*$, where λ^* represents an optimal choice of λ for testing. Even when $s \ll s^*$, our testing procedure under a proper λ still demonstrates some power as long as SWDS becomes sufficiently large.

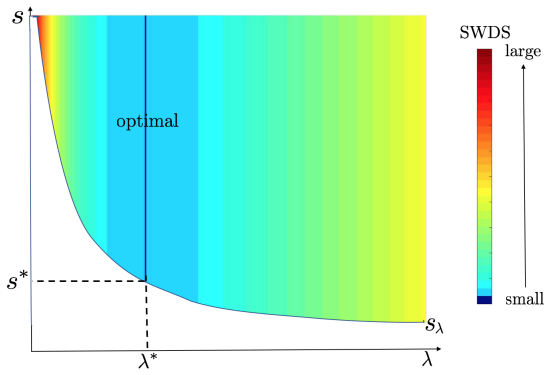


Figure 1: Phase transition in (λ, s) for signal detection. The horizontal axis is the smoothing parameter λ, and the vertical axis is the projection dimension s. The shade indicates the values of SWDS: dark red corresponds to greater values of SWDS than light blue. The vertical line labeled by "optimal" indicates the choices of λ that achieve the smallest SWDS.

One technical contribution of this paper is the generalized local Radermacher complexity (Section 3) which allows a unified treatment for nonparametric estimation and testing. The classic local Radermacher complexity developed by [4] is a special case (with unit variance-to-bias ratio) which only works for nonparametric estimation. This new technique is obtained by flexibly adjusting the size of the function class defining the Rademacher average. Our results hold for a general class of random projection matrix, such as the sub-Gaussian matrix or certain data-dependent matrix. The procedure can be generalized for composite hypothesis testing (see Section 4.2).

1.2. Related Literature

Computational limits have been addressed in other situations. For divide-and-conquer, [32] derived a *sharp* upper bound for the number of distributed computing units in smoothing splines, while [36] estimated the quantile regression process under an additional *sharp* lower bound on the number of quantile levels. For random projection methods, the literature nonetheless only focused on parametric cases such as compressed sensing, see [9]. For example, [9] showed that the minimum number of random projections is $s \log n$ for signal recovery, where n is the number of measurements and s is the number of nonzero components in the true signal. Relevant results in nonparametric setting are still missing.

Notation: Denote δ_{jk} the Kronecker delta: $\delta_{jk} = 1$ if $j = k$ and $\delta_{jk} = 0$ if $j \neq k$. For positive sequences a_n and b_n, put $a_n \lesssim b_n$ if there exists a constant $c > 0$ such that $a_n \leq cb_n$ for all $n \in \mathbb{N}$; $a_n \gtrsim b_n$ if there exists a constant $c > 0$ such that $a_n \leq cb_n$. Put $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. Frequently, we use $a_n \lesssim b_n$ and $a_n = \mathcal{O}(b_n)$ interchangeably. $Pf^2 \equiv \mathrm{E}\, f(X)^2$, $\|f\|_n^2 \equiv P_n f^2 \equiv \frac{1}{n}\sum_{i=1}^n f(X_i)^2$. For a matrix $A \in \mathbb{R}^{m\times n}$, its operator norm is defined as $\|A\|_{\mathrm{op}} = \max_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Ax\|_2}{\|x\|_2}$. A random variable X is said to be sub-Gaussian if there exists a constant $\sigma^2 > 0$ such that for any $t \geq 0$, $\mathrm{P}[|X| \geq t] \leq 2\exp(-t^2/(2\sigma^2))$. The sub-Gaussian norm of X is defined as $\|X\|_{\psi_2} = \inf\{t > 0 : \mathrm{E}\exp(X^2/t^2) \leq 2\}$. We will use c, c_1, c_2, C to denote generic absolute constants, whose values may vary from line to line.

2. Kernel Ridge Regression via Random Projection

In this section, we review kernel ridge regression and its variant based on random projection. Suppose that we have n i.i.d. observations $\{(x_i, y_i)\}_{i=1}^n$ from (1.1). Throughout assume that $f \in \mathcal{H}$, where $\mathcal{H} \subset L^2(P_X)$ is a reproducing kernel Hilbert space (RKHS) associated with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and a reproducing kernel function $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. By Mercer's theorem, K has the following spectral expansion:

$$K(x, x') = \sum_{i=1}^{\infty} \mu_i \phi_i(x) \phi_i(x'), \quad x, x' \in \mathcal{X}, \tag{2.1}$$

where $\mu_1 \geq \mu_2 \geq \cdots \geq 0$ is a sequence of ordered eigenvalues and the eigenfunctions $\{\phi_i\}_{i=1}^{\infty}$ form a basis in $L^2(P_X)$. Moreover, for any $i, j \in \mathbb{N}$,

$$\langle \phi_i, \phi_j \rangle_{L^2(P_X)} = \delta_{ij} \qquad \text{and} \qquad \langle \phi_i, \phi_j \rangle_{\mathcal{H}} = \delta_{ij}/\mu_i.$$

Throughout this paper, assume that ϕ_i's are uniformly bounded, a common condition in literature, e.g., [16], and μ_i's satisfy certain tail sum property.

Assumption A1 $c_K := \sup_{i \geq 1} \|\phi_i\|_{\sup} < \infty$ *and* $\sup_{k \geq 1} \frac{\sum_{i=k+1}^{\infty} \mu_i}{k \mu_k} < \infty$.

Assumption A1 is satisfied in two types of commonly used kernels, categorized by the eigenvalue decay rates. The first is $\mu_i \asymp i^{-2m}$ for a constant $m > 0$, called as polynomial decay kernel (PDK) of order m. The second is $\mu_i \asymp \exp(-\gamma i^p)$ for constants $\gamma, p > 0$, called as exponential decay kernel (EDK) of order p. Verification of Assumption A1 is deferred to Section B.7 in the Appendix. Examples of PDK include kernels of Sobolev space and periodic Sobolev space (see [37]). Examples of EDK include Gaussian kernel $K(x_1, x_2) = \exp(-(x_1 - x_2)^2/2)$ (see [30]).

Recall the KRR estimator $\widehat{f}_n$ from (1.2). By representer theorem, it has an expression $\widehat{f}_n(\cdot) = \sum_{i=1}^n \widehat{\omega}_i K(\cdot, x_i)$, where $\widehat{\omega} = (\widehat{\omega}_1, \ldots, \widehat{\omega}_n)^\top$ is a real vector determined by

$$\begin{aligned} \widehat{\omega} &= \operatorname*{argmin}_{\omega \in \mathbb{R}^n} \left\{ \omega^\top \boldsymbol{K}^2 \omega - \frac{2}{n} \omega^\top \boldsymbol{K} \boldsymbol{y} + \lambda \omega^\top \boldsymbol{K} \omega \right\} \\ &= \frac{1}{n} (\boldsymbol{K} + \lambda I)^{-1} \boldsymbol{y}, \end{aligned} \tag{2.2}$$

$\boldsymbol{y} = (y_1, \cdots, y_n)^\top$, $\boldsymbol{K} = [n^{-1} K(x_i, x_j)]_{1 \leq i,j \leq n}$, and $I \in \mathbb{R}^{n \times n}$ is identity. This standard procedure requires storing $(\boldsymbol{K}^2, \boldsymbol{K}, \boldsymbol{K}\boldsymbol{y})$ and inverting $\boldsymbol{K} + \lambda I$, which requires $\mathcal{O}(n^2)$ memory usage and $\mathcal{O}(n^3)$ floating operations.

The above computational and storage constraints become severe for a large sample size, and thus motivate the random projection approach proposed by [40]. Specifically, ω in (2.2) is substituted with $S^\top \beta$, where $\beta \in \mathbb{R}^s$ and S is an $s \times n$ real-valued random matrix; see Section 4.1. Then, β is solved as:

$$\begin{aligned} \widehat{\beta} &= \operatorname*{argmin}_{\beta \in \mathbb{R}^s} \left\{ \beta^\top (S\boldsymbol{K})(\boldsymbol{K} S^\top) \beta - \frac{2}{n} \beta^\top S \boldsymbol{K} \boldsymbol{y} + \lambda \beta^\top S \boldsymbol{K} S^\top \beta \right\}, \\ &= \frac{1}{n} (S\boldsymbol{K}^2 S^\top + \lambda S \boldsymbol{K} S)^{-1} S \boldsymbol{K} \boldsymbol{y}. \end{aligned} \tag{2.3}$$

Hence, the resulting estimator of f becomes

$$\widehat{f}_R(\cdot) = \sum_{i=1}^{n} (S^\top \widehat{\beta})_i K(\cdot, x_i), \tag{2.4}$$

which requires computing and storing $(S\boldsymbol{K}^2 S^\top, S\boldsymbol{K}S^\top, S\boldsymbol{K}\boldsymbol{y})$, along with inverting an $s \times s$ matrix. The cost in the pre-processing step to compute the kernel approximation normally takes $O(sn^2)$, and can be easily reduced to $O(n^2(\log s))$ for suitably chosen random matrices (see Ailon and Chazelle [1]), which can be further reduced to $O(n^2(\log s)/t)$ by using t clusters in a parallel fashion. Furthermore, the memory usage and floating operations are reduced to $\mathcal{O}(s^2)$ and $\mathcal{O}(s^3)$, respectively, when $s = o(n)$. On the other hand, s cannot be too small in order to maintain sufficient data information for achieving statistical optimality. Critical lower bounds for s will be derived in Section 4.5.

3. Tail Sum of Empirical Eigenvalues

An accurate upper bound for the tail sum of empirical eigenvalues is needed for studying nonparametric testing and estimation. However, this bound was often *assumed* to hold in the kernel learning literature, e.g., [7; 39]. And, the application of concentration inequalities of individual eigenvalues ([33; 8]) only provides a loose bound due to accumulative errors. Recently, the local Rademacher complexity (LRC) theory ([4]) was employed by [40] to derive a more accurate upper bound that is useful in studying nonparametric estimation. However, this upper bound no longer works for testing problems, due to the improper size of the function class defining Rademacher average.

In this section, we establish upper bounds, i.e., Lemma 3.1, for a range of tail sums of empirical eigenvalues that can be applied to both nonparametric estimation and testing. This result may be of independent interest. Consider the singular value decomposition $\boldsymbol{K} = UDU^\top$, where $UU^\top = I_n$ and $D = \text{diag}(\widehat{\mu}_1, \widehat{\mu}_2, \ldots, \widehat{\mu}_n)$ with $\widehat{\mu}_1 \geq \widehat{\mu}_2 \geq \cdots \widehat{\mu}_n \geq 0$. For any $\lambda > 0$, define $\widehat{s}_\lambda$ (or s_λ) to be the number of $\widehat{\mu}_i$'s (or μ_i's) greater than λ, i.e.,

$$\widehat{s}_\lambda = \text{argmin}\{i : \widehat{\mu}_i \leq \lambda\} - 1, \quad s_\lambda = \text{argmin}\{i : \mu_i \leq \lambda\} - 1. \tag{3.1}$$

We have the following assumption on the population eigenvalues through s_λ.

Assumption A2 *s_λ diverges as $\lambda \to 0$.*

Assumption A2 is satisfied in various classes of kernels, including PDK and EDK introduced in Section 4.4.

For a range of λ, Lemma 3.1 below provides an upper bound for the tail sum of $\widehat{\mu}_i$ in terms of population quantities s_λ and μ_{s_λ}, with known orders.

Lemma 3.1 *If $1/n < \lambda \to 0$, then with probability at least $1 - 4e^{-s_\lambda}$, $\sum_{i=\widehat{s}_{\lambda+1}}^{n} \widehat{\mu}_i \leq C s_\lambda \mu_{s_\lambda}$, where $C > 0$ is an absolute constant.*

Clearly, Lemma 3.1 is a sample analog to the tail sum assumption for μ_i in Assumption A1. Lemma 3.1 is crucial in the verification of "K-satisfiable" property of random projection matrices introduced in Section 4. The proof of Lemma 3.1 is based on an adaptation of the classical LRC theory as explained below.

In Section 4.4, it will be shown that λ and s_λ/n correspond to (squared-)bias and variance of $\widehat{f}_R$, respectively. We then define the variance-to-bias ratio as

$$\kappa_\lambda = \frac{s_\lambda}{n\lambda}, \tag{3.2}$$

for any $\lambda > 0$. Consider a bundle of function classes indexed by κ_λ:

$$\mathcal{F}_\lambda = \{f \in \mathcal{H} : f \text{ maps } \mathcal{X} \text{ to } [-1,1], \|f\|_{\mathcal{H}}^2 \le \kappa_\lambda\}, \lambda > 0.$$

To characterize the complexity of $\mathcal{F}_\lambda$, we introduce a generalized version of local Rademacher complexity function:

$$\Psi_\lambda(r) = \mathrm{E}\Big\{ \sup_{\substack{f\in\mathcal{F}_\lambda \\ Pf^2\le r}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)\Big\}, r \ge 0, \tag{3.3}$$

where $\sigma_1, \ldots, \sigma_n$ are independent Rademacher random variables, i.e., $\mathrm{P}(\sigma_i = 1) = \mathrm{P}(\sigma_i = -1) = 1/2$. Let $\widehat{\Psi}_\lambda(\cdot)$ be an empirical version of $\Psi_\lambda(\cdot)$ defined as

$$\widehat{\Psi}_\lambda(r) = \mathrm{E}\Big\{ \sup_{\substack{f\in\mathcal{F}_\lambda \\ P_n f^2\le r}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)\Big| x_1, \cdots, x_n\Big\}, r \ge 0. \tag{3.4}$$

When $\kappa_\lambda \asymp 1$, $\Psi_\lambda(\cdot)$ and $\widehat{\Psi}_\lambda(\cdot)$ become the original LRC functions introduced in [4]. Note that $\kappa_\lambda \asymp 1$ actually corresponds to the optimal bias vs. variance trade-off required for estimation. Rather, a different type of trade-off is needed for optimal testing as revealed by [18; 31], which corresponds to a different choice of κ_λ in $\mathcal{F}_\lambda$ as demonstrated later in Section 4.4.

Lemma 3.2 says that both Ψ_λ and $\widehat{\Psi}_\lambda$ possess unique (positive) fixed points. This fixed point property is crucial in proving Lemma 3.1. Interestingly, we find that the fixed points turn out to be proportional to the estimation variance asymptotically.

Lemma 3.2 *There exist uniquely positive r_λ and $\widehat{r}_\lambda$ such that $\Psi_\lambda(r_\lambda) = r_\lambda$ and $\widehat{\Psi}_\lambda(\widehat{r}_\lambda) = \widehat{r}_\lambda$. Furthermore, if $\lambda > 1/n$, then $r_\lambda \asymp s_\lambda/n$, and there exists an absolute constant $c > 0$ such that, with probability at least $1 - e^{-cs_\lambda}$, $\widehat{r}_\lambda \asymp s_\lambda/n$.*

We are now ready to sketch the proof of Lemma 3.1. Detailed proofs are deferred to Appendix A.1. First, note that

$$\sum_{i=\widehat{s}_\lambda+1}^n \widehat{\mu}_i = \sum_{i=\widehat{s}_\lambda+1}^n \min\{\lambda, \widehat{\mu}_i\} \le \sum_{i=1}^n \min\{\lambda, \widehat{\mu}_i\}.$$

By Lemma 3.2, we have $\widehat{r}_\lambda/\kappa_\lambda \asymp \lambda$ with high probability. Then,

$$\sum_{i=1}^n \min\left\{\lambda, \widehat{\mu}_i\right\} \asymp \sum_{i=1}^n \min\left\{\frac{\widehat{r}_\lambda}{\kappa_\lambda}, \widehat{\mu}_i\right\} \asymp \frac{n}{\kappa_\lambda}\widehat{\Psi}_\lambda(\widehat{r}_\lambda)^2 = \frac{n\widehat{r}_\lambda^2}{\kappa_\lambda} \asymp \lambda s_\lambda \le s_\lambda \mu_{s_\lambda},$$

where the second step is by Lemma B.1 that

$$\widehat{\Psi}_\lambda(\widehat{r}_\lambda) \asymp \sqrt{\frac{\kappa_\lambda}{n}\sum_{i=1}^n \min\{\frac{\widehat{r}_\lambda}{\kappa_\lambda}, \widehat{\mu}_i\}},$$

the third step follows from the fixed point property stated in Lemma 3.2, and the last step follows from the definition of μ_{s_λ} given in (3.1).

4. Main Results

Consider the nonparametric testing problem (1.3). For convenience, assume $f_0 = 0$, i.e., we will test

$$H_0 : f = 0 \quad \text{vs.} \quad H_1 : f \in \mathcal{H} \setminus \{0\}. \tag{4.1}$$

In general, testing $f = f_0$ (for an arbitrary known f_0) is equivalent to testing $f_* \equiv f - f_0 = 0$. So, (4.1) has no loss of generality. Based on $\widehat{f}_R$, we propose the following distance-based test statistic:

$$T_{n,\lambda} = \|\widehat{f}_R\|_n^2. \tag{4.2}$$

In the subsequent sections, we will derive the null limit distribution of $T_{n,\lambda}$ (Theorems 4.2 and 4.4), and further provide a *sufficient and necessary* condition in terms of s such that $T_{n,\lambda}$ is minimax optimal (Section 4.5). As a byproduct, we derive a critical bound in terms of s such that $\widehat{f}_R$ is minimax optimal. Proof of such results rely on an exact analysis on the kernel and projection matrices which requires an accurate estimate of the tail sum of the empirical eigenvalues by Lemma 3.1. Our results hold for a general choice of projection matrix (see Section 4.1 for discussion).

4.1. Choice of Projection Matrix

Recall the singular value decomposition $\boldsymbol{K} = UDU^\top$. Put $U = (U_1, U_2)$ with U_1 consisting of the first $\widehat{s}_\lambda$ columns of U and U_2 consisting of the rest $n - \widehat{s}_\lambda$ columns; $D = \mathrm{diag}(D_1, D_2)$, with $D_1 = \mathrm{diag}(\widehat{\mu}_1, \ldots, \widehat{\mu}_{\widehat{s}_\lambda})$, $D_2 = \mathrm{diag}(\widehat{\mu}_{\widehat{s}_\lambda+1}, \ldots, \widehat{\mu}_n)$.

The following definition of "$\boldsymbol{K}$-satisfiability" describes a class of matrices that preserve the principal components of the kernel matrix.

Definition 1 *($\boldsymbol{K}$-satisfiability) A matrix $S \in \mathbb{R}^{s\times n}$ is said to be $\boldsymbol{K}$-satisfiable if there exists a constant $c > 0$ such that*

$$\|(SU_1)^\top SU_1 - I_{\widehat{s}_\lambda}\|_{op} \le 1/2, \quad \|SU_2 D_2^{1/2}\|_{op} \le c\lambda^{1/2}.$$

By Definition 1, a $\boldsymbol{K}$-satisfiable S will make $(SU_1)^\top SU_1$ "nearly" identity as well as down-weight the tail eigenvalues. Such a matrix will be able to extract the principle information from the kernel matrix. A special case of the above "K-satisfiability" condition was studied in [40] by fixing λ as the optimal estimation rate. However, by choosing a range of λ as threshold to select the leading eigenvalues, our general form of "K-satisfiability" condition allows us to study estimation and testing in a unified framework.

Besides, we need the following definition to simplify the statement of our assumptions.

Definition 2 *An event $\mathcal{E}$ is said to be of (a, b)-type for $a, b \in (0, \infty]$, if $P(P(\mathcal{E}|x_1, \cdots, x_n) \ge 1 - \exp(-a)) \ge 1 - \exp(-b)$.*

Definition 2 describes events whose probabilities have exponential type lower bounds. It is easy to see that, if $\mathcal{E}$ is of (a, b)-type, then $\mathrm{P}(\mathcal{E}) \ge (1 - \exp(-a))(1 - \exp(-b))$. In particular, $\mathcal{E}$ is of (∞, ∞)-type if and only if $\mathcal{E}$ occurs almost surely.

Throughout the rest of this paper, assume the following condition on S.

Assumption A3

(a) $s \geq ds_\lambda$ *for a sufficiently large constant* $d > 0$.

(b) *There exist* $c_1, c_2 \in (0, \infty]$ *such that the event "*S *is* $\boldsymbol{K}$*-satisfiable" is of* $(c_1 s, c_2 s_\lambda)$*-type.*

Assumption A3 (a) requires a sufficient amount of random projections to preserve data information. Assumption A3 (b) requires S to be $\boldsymbol{K}$-satisfiable with high probability which holds in a broad range of situations such as matrix of sub-Gaussian entries (Example 1) and certain data dependent matrix (Example 2).

Example 1 *Let* S *be an* $s \times n$ *random matrix of entries* $S_{ij}/\sqrt{s}$, $i = 1, \ldots, s$, $j = 1, \ldots, n$, *where* S_{ij} *are independent (not necessarily identically distributed) sub-Gaussian variables. Examples of such sub-Gaussian variables include Gaussian variables, bounded variables such as Bernoulli, multinomial, uniform, variables with strongly log-concave density (see [28]), or mixtures of sub-Gaussian variables. The following lemma shows that Assumption A3 (b) holds in all these situations.*

Lemma 4.1 *Let* $S_{ij} : 1 \leq i \leq s,\ 1 \leq j \leq n$ *be independent sub-Gaussian of mean zero and variance one, and* $\lambda \in (1/n, 1)$. *If* $s \geq ds_\lambda$ *for a sufficiently large constant* d, *then Assumption A3 (b) holds for* $S = [S_{ij}/\sqrt{s}]_{1 \leq i \leq s, 1 \leq j \leq n}$.

In the proof of Lemma 4.1, the operator norm of $SU_2\sqrt{D_2}$ is concentrated on the empirical tail sums of eigenvalues, which can be further bounded by the population version based on Lemma 3.1.

Example 2 *Let* $S = U_s^\top$, *where* U_s *is an* $n \times s$ *matrix consisting of the first* s *columns of* U. *Then it trivially holds that, almost surely,* $(SU_1)^\top SU_1 = I_{\widehat{s}_\lambda}$ *and* $\|SU_2 D_2^{1/2}\|_{op} = 0$, *i.e., Assumption A3 (b) holds.*

The eigen-decomposition in Example 2 is as burdensome as computing the matrix inverse, which is not preferred in practice. The purpose of this example is to illustrate one situation that satisfies Assumption A3, also useful for deriving computational limits (see Section 4.5).

4.2. Testing Consistency

In this section, we derive the null limit distribution of (standardized) $T_{n,\lambda}$ as standard Gaussian, and then extend our result to the case of composite hypothesis testing.

Theorem 4.2 *Suppose that* $\lambda \to 0$ *and* $s \to \infty$ *as* $n \to \infty$. *Suppose Assumption A2 is satisfied. Then under* H_0, *we have*

$$\frac{T_{n,\lambda} - \mu_{n,\lambda}}{\sigma_{n,\lambda}} \xrightarrow{d} N(0,1), \quad \text{as } n \to \infty.$$

Here, $\mu_{n,\lambda} := \mathrm{E}_{H_0}\{T_{n,\lambda}|\boldsymbol{x}, S\} = \mathrm{tr}(\Delta^2)/n$, $\sigma^2_{n,\lambda} := \mathrm{Var}_{H_0}\{T_{n,\lambda}|\boldsymbol{x}, S\} = 2\,\mathrm{tr}(\Delta^4)/n^2$ *with* $\boldsymbol{x} = (x_1, \cdots, x_n)$ *and* $\Delta = \boldsymbol{K}S^\top(S\boldsymbol{K}^2S^\top + \lambda S\boldsymbol{K}S^\top)^{-1}S\boldsymbol{K}$.

Theorem 4.2 holds once s diverges (no matter how slowly). Theorem 4.2 implies the following testing rule at significance level α:

$$\phi_{n,\lambda} = I(|T_{n,\lambda} - \mu_{n,\lambda}| \geq z_{1-\alpha/2}\sigma_{n,\lambda}) \tag{4.3}$$

where $z_{1-\alpha/2}$ is the $100 \times (1-\alpha/2)$th percentile of $N(0,1)$.

As an important consequence of Theorem 4.2, we comment that the optimal estimation rate in [40] can also be obtained as a by-product; see the following Corollary 4.3 with proof in B.5.

Corollary 4.3 *Suppose that $1/n < \lambda < 1$ and Assumption A1-A3 holds. Then with probability approaching one, it holds that*

$$\|\widehat{f}_R - f_0\|_n^2 \leq C r_{n,\lambda},$$

where $r_{n,\lambda} = \lambda + \mu_{n,\lambda}$ and C is an absolute constant.

From Corollary 4.3, the best upper bound can be obtained through balancing λ and $\mu_{n,\lambda}$. Denote $\lambda^\dagger$ the optimizer. This in turn provides a lower bound $s^\dagger$ for s according to (3.1), i.e., $s^\dagger = s_{\lambda^\dagger}$. In Section 4.4, we will show that the upper bound under $\lambda^\dagger$ is minimax optimal, and further provide explicit orders for $s^\dagger$ in concrete settings.

In practice, it is often of interest to test certain structure of f, e.g., linearity,

$$H_0^{\text{linear}} : f \in \mathcal{L}(\mathcal{X}) \text{ vs. } H_1^{\text{linear}} : f \notin \mathcal{L}(\mathcal{X}),$$

where $\mathcal{L}(\mathcal{X})$ is the class of linear functions over $\mathcal{X} \subseteq \mathbb{R}^a$. Testing H_0^{linear} can be easily converted into simple hypothesis testing problem as follows. Suppose that $f_0(x) = \beta_0 + \beta_1^\top x$ is the "true" function under H_0^{linear}. The corresponding MLE is $\widehat{f}_0(x) = \widehat{\beta}_0 + \widehat{\beta}_1^\top x$, where $\widehat{\beta} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\boldsymbol{y} \equiv (\widehat{\beta}_0, \widehat{\beta}_1)$. By defining $f^* = f - \widehat{f}_0$, it amounts to testing $f^* = 0$. Correspondingly, we define $\boldsymbol{y}^* = \boldsymbol{y} - \widehat{\boldsymbol{y}}_0$, where $\widehat{\boldsymbol{y}}_0 = (\widehat{f}_0(x_1), \ldots, \widehat{f}_0(x_n))^\top = H\boldsymbol{y}$ and $H = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}$. This leads to the randomly projected KRR estimator $\widehat{f}_R^*$ and $T_{n,\lambda}^* = \|\widehat{f}_R^*\|_n^2$, whose null limit distribution is given in the following theorem.

Theorem 4.4 *Suppose that $\lambda \to 0$ and $s \to \infty$ as $n \to \infty$. Suppose Assumption A2 is satisfied. Under H_0^{linear}, we have*

$$\frac{T_{n,\lambda}^* - \mu_{n,\lambda}^*}{\sigma_{n,\lambda}^*} \xrightarrow{d} N(0,1), \quad \text{as } n \to \infty,$$

where $\mu_{n,\lambda}^* = \mathrm{E}_{H_0^{linear}}\{T_{n,\lambda}^* | \boldsymbol{x}, S\} = \mathrm{tr}((I-H)\Delta^2(I-H))/n$ and $\{\sigma_{n,\lambda}^*\}^2 = \mathrm{Var}_{H_0^{linear}}(T_{n,\lambda}^* | \boldsymbol{x}, S) = 2\,\mathrm{tr}((I-H)\Delta^4(I-H))/n^2$.

Clearly, our testing procedure and theory can be easily generalized to polynomial testing such as $H_0^{\text{poly}} : f$ is polynomial of order q.

4.3. Power Analysis

In this section, we investigate the power of $T_{n,\lambda}$ under a sequence of local alternatives. The following result shows that $T_{n,\lambda}$ can achieve high power provided that s diverges fast enough *and* the local alternative is separated from the null by at least an amount of $d_{n,\lambda}$. Here we call $d_{n,\lambda}$ as the weakest detectable signals (SWDS) or separation rate.

Theorem 4.5 *Suppose that $1/n < \lambda \to 0$ as $n \to \infty$, Assumption A1-A2 are satisfied, and Assumption A3 holds for $c_1, c_2 \in (0, \infty]$. Then for any $\varepsilon > 0$, there exist positive constants C_ε and N_ε such that, with probability greater than $1 - e^{-c_1 s} - e^{-c_2 s_\lambda}$,*

$$\inf_{n \geq N_\varepsilon} \inf_{\substack{f \in \mathcal{B} \\ \|f\|_n \geq C_\varepsilon d_{n,\lambda}}} P_f(\phi_{n,\lambda} = 1 | \boldsymbol{x}, S) \geq 1 - \varepsilon,$$

where $d_{n,\lambda} := \sqrt{\lambda + \sigma_{n,\lambda}}$ and $\mathcal{B} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq C\}$ for a constant C and $P_f(\cdot | \boldsymbol{x}, S)$ is the conditional probability measure under f given $\boldsymbol{x}, S$.

In view of Theorem 4.5, to maximize the power of $T_{n,\lambda}$, one needs to minimize $d_{n,\lambda} = \sqrt{\lambda + \sigma_{n,\lambda}}$ through balancing λ and $\sigma_{n,\lambda}$. Denote λ^* the optimizer. The lower bound s^* for s is obtained via (3.1), i.e., $s^* = s_{\lambda^*}$. The explicit forms of λ^* and s^* varies for different reproducing kernels, and lead to specific optimal testing rate, depending on their eigendecay rate.

4.4. Examples

Next, we derive the lower bounds for s to achieve optimal estimation and testing in two featured examples: PDK and EDK, based on the main results obtained in Corollary 4.3 and Theorem 4.5. It is easy to check that Assumption A1 and A2 hold for these two examples; see Section B.7.

Theorem 4.6 *For the two kinds of eigenvalue decaying rates, suppose Assumption A3 holds, we have the following optimal estimation and testing rates by properly choosing the tuning parameters and the lower bound of projection dimension:*

- **Polynomially decaying kernel (with $\mu_i \asymp i^{-2m}$)**
  - *When $\lambda \asymp n^{-\frac{2m}{2m+1}}$ and $s \gtrsim n^{\frac{1}{2m+1}}$ with $m > 3/2$, $\|\widehat{f}_R - f_0\|_n^2 = \mathcal{O}_P(n^{-\frac{2m}{2m+1}})$.*
  - *When $\lambda \asymp n^{-\frac{4m}{4m+1}}$ and $s \gtrsim n^{\frac{2}{4m+1}}$ with $m > 3/2$, $T_{n,\lambda}$ achieves the minimax optimal rate of testing $n^{-\frac{2m}{4m+1}}$.*
- **Exponentionally decaying kernel (with $\mu_i \asymp \exp(-\gamma i^p)$)**
  - *When $\lambda \asymp (\log n)^{1/p} n^{-1}$ and $s \gtrsim (\log n)^{1/p}$, $\|\widehat{f}_R - f_0\|_n^2 = \mathcal{O}_P(n^{-1}(\log n)^{1/p})$.*
  - *When $\lambda \asymp (\log n)^{1/(2p)} n^{-1}$ and $s \gtrsim (\log n)^{1/p}$, $T_{n,\lambda}$ achieves the minimax optimal rate of testing $n^{-\frac{1}{2}}(\log n)^{\frac{1}{4p}}$.*

Based on Lemma 3.1, Lemma A.1 characterized the order of $\mu_{n,\lambda}$ and $\sigma_{n,\lambda}$ as $\mu_{n,\lambda} \asymp s_\lambda/n$ and $\sigma_{n,\lambda}^2 \asymp s_\lambda/n^2$ with high probability. It follows from Corollary 4.3 that $\widehat{f}_R$ has the convergence rate $r_{n,\lambda} = \lambda + s_\lambda/n$. On the other hand, λ is the bias of $\widehat{f}_R$ by Lemma B.3, and s_λ/n is the variance of $\widehat{f}_R$ by (**??**) and Lemma A.1. Hence, the optimal estimation rate $r_{n,\lambda}^\dagger$ ([40]) is achieved as follows

$$r_{n,\lambda}^\dagger = \operatorname{argmin}\Big\{\lambda \,:\, \lambda > s_\lambda/n\Big\}.$$

To find the lower bound for s in achieving optimal testing, by Theorem 4.5, $d_{n,\lambda} \asymp \sqrt{\lambda + \sqrt{s_\lambda}/n}$, then the optimal separation rate d_n^* ([18], [38]) can be achieved by another type of trade-off, i.e., the bias of $\widehat{f}_R$ v.s. the standard derivation of $T_{n,\lambda}$, as follows

$$d_n^{*\,2} = \operatorname{argmin}\Big\{\lambda \,:\, \lambda > \sqrt{s_\lambda}/n\Big\}.$$

For PDK example, Theorem 4.6 can be directly achieved based on (3.1), $\mu_i \asymp i^{-2m}$, and $s_\lambda \asymp \lambda^{-\frac{1}{2m}}$. The results for EDK can be achieved similarly.

It is worth emphasizing that $\lambda^\dagger, s^\dagger$ are different from λ^*, s^*, indicating a fundamental difference between estimation and testing. A more explicit reason for such a difference in minimax rate is due to two different types of trade-off, as illustrated in Figure 2. Table 1 summarizes our findings of this section.

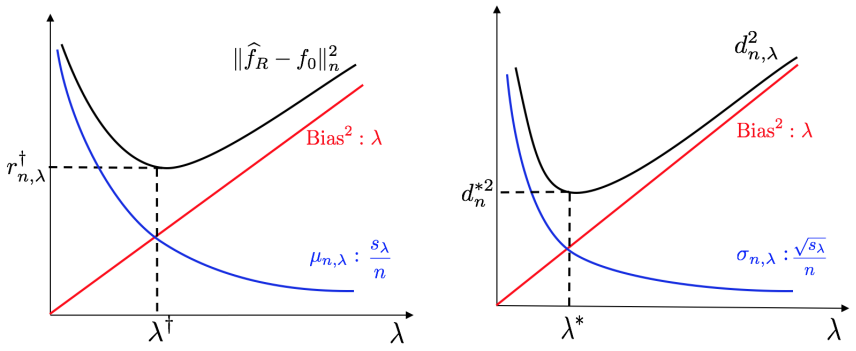


Figure 2: Trade-offs for achieving (a) optimal estimation rate; (b) optimal testing rate.

4.5. Sharpness of $s^\dagger$ and s^*

In this section, we will show that s^* and $s^\dagger$ derived in PDK and EDK are actually sharp. For technical convenience, define

$$\delta_n = \begin{cases} n^{-\frac{2}{2m-1}}, & K \text{ is PDK} \\ (\log n)^{-2/p}, & K \text{ is EDK} \end{cases}$$

Our first result is about the sharpness of $s^\dagger$. Theorem 4.7 shows that when $s \ll s^\dagger$, there exists a true function f such that $\|\widehat{f}_R - f\|_n^2$ is substantially slower than the optimal estimation rate.

Theorem 4.7 *Suppose $s = o(s^\dagger)$. Then there exists an $s \times n$ random matrix S satisfying Assumption A3, such that with probability greater than $1 - e^{-cn\delta_n} - e^{-c_1 s} - e^{-c_2 s_\lambda}$, it holds that*

$$\sup_{f \in \mathcal{B}} \|\widehat{f}_R - f\|_n^2 \gg r^\dagger_{n,\lambda},$$

where c is a constant independent of n, and $c_1, c_2 \in (0, \infty]$ are given in Assumption A3 (b).

We sketch the constructive proof as follows; detailed proof can be found in Appendix A.6. Note that

$$\|\widehat{f}_R - f\|_n^2 = \|\operatorname{E}_\epsilon \widehat{f}_R - f\|_n^2 + \|\widehat{f}_R - \operatorname{E}_\epsilon \widehat{f}_R\|_n^2 + \frac{2}{n}(\widehat{f}_R - \operatorname{E}_\epsilon \widehat{f}_R)^\top (\operatorname{E}_\epsilon \widehat{f}_R - f_0) \equiv T_1 + T_2 + T_3,$$

where $T_1 \equiv \|\operatorname{E}_\epsilon \widehat{f}_R - f\|_n^2 = \|U^T \boldsymbol{K} S^\top (S\boldsymbol{K}^2 S^\top + \lambda S \boldsymbol{K} S^\top)^{-1} S\boldsymbol{K} f - U^T f\|_n^2$. Let $S = U_s$, where U_s is the first s columns of U. Let $f(\cdot) = \sum_{i=1}^n K(x_i, \cdot) w_i$ with $w = (w_1, \cdots, w_n)^\top = U\alpha$, $\alpha \in \mathbb{R}^n$ satisfying

$$\alpha_i^2 = \begin{cases} \frac{1}{n}\frac{C}{s}\widehat{\mu}_i^{-1}, & \text{for } \ i = s+1, \cdots, 2s; \\ 0, & \text{otherwise.} \end{cases} \tag{4.4}$$

Then $\|f\|_{\mathcal{H}}^2 = n\alpha^\top D\alpha = n\sum_{i=s+1}^{2s} \alpha_i^2 \widehat{\mu}_i = C$, and

$$\|\operatorname{E}_\epsilon \widehat{f}_R - f\|_n^2 = n\sum_{i=1}^{s} \alpha_i^2 (\widehat{\mu}_i^2 (\widehat{\mu}_i + \lambda)^{-1} - \widehat{\mu}_i)^2 + n \sum_{i=s+1}^{n} \alpha_i^2 \widehat{\mu}_i^2 = \frac{1}{s} \sum_{i=s+1}^{2s} \widehat{\mu}_i \geq \widehat{\mu}_{2s} \geq \frac{1}{2}\mu_{2s} \gg \lambda^\dagger.$$

The last inequality holds with probability greater than $1 - e^{-n\delta_n}$ by Lemma B.2. Furthermore, it can be shown that $T_1 = \mathcal{O}_P(\lambda)$, $T_2 = o_p(T_1)$, and $T_3 = o_P(T_1)$. Then with probability at least $1 - e^{-n\delta_n}$,

$$\sup_{f_0 \in \mathcal{B}} \|\widehat{f}_R - f_0\|_n^2 \gtrsim \sup_{f_0 \in \mathcal{B}} \|\operatorname{E}_\epsilon \widehat{f}_R - f_0\|_n^2 \gg C\lambda^\dagger.$$

Our second result is about the sharpness of s^*. Theorem 4.8 shows that when $s \ll s^*$, there exists a local alternative f that is not detectable by $T_{n,\lambda}$ even when it is separated from zero by d_n^*. In this case, the asymptotic testing power is actually smaller than the nominal level α.

Theorem 4.8 *Suppose $s = o(s^*)$. Then there exists an $s \times n$ projection matrix S satisfying Assumption A3 and a positive nonrandom sequence $\beta_{n,\lambda}$ satisfying $\lim_{n\to\infty} \beta_{n,\lambda} = \infty$ such that, with probability at least $1 - e^{-cn\delta_n} - e^{-c_1 s} - e^{-c_2 s_\lambda}$,*

$$\limsup_{n\to\infty} \inf_{\substack{f \in \mathcal{B} \\ \|f\|_n \geq \beta_{n,\lambda} d_n^*}} P_f(\phi_{n,\lambda} = 1 | \boldsymbol{x}, S) \leq \alpha,$$

where c is a constant independent of n, and $c_1, c_2 \in (0, \infty]$ are given in Assumption A3 (b). Recall $1 - \alpha$ is the significance level.

The proof of Theorem 4.8 is similar as that of Theorem 4.7, except that a different true function is constructed as $f(\cdot) = \sum_{i=1}^n K(x_i, \cdot) w_i$ with $w = (w_1, \cdots, w_n)^\top = U\alpha$, where $\alpha \in \mathbb{R}^n$ satisfies

$$\alpha_i^2 = \begin{cases} \frac{1}{n} \frac{C}{s-1} \widehat{\mu}_{gs+k}^{-1} & \text{for } \ i = (gs + k) \quad k = 1, 2, \cdots, s-1; \\ 0 & \text{otherwise,} \end{cases} \tag{4.5}$$

with $g \geq 1$ as an integer satisfying $(g + 1)s \ll s^*$; see Appendix A.7 for detailed proof.

In view of Theorems 4.5 and 4.8, we observe a subtle phase transition phenomenon for testing the existence of signals as shown in Figure 1.1. The precise order of s^* in specific situations can be found in Table 1.

We remark that Theorem 4.8 can hold for any K-satisfiable random matrix S, by constructing the true function f with $\|f\|_{\mathcal{H}} \leq C$ and $\|f\|_n^2 \geq \widehat{\mu}_{s+1}$, such that under f, $\frac{T_{n,\lambda} - \mu_{n,\lambda}}{\sigma_{n,\lambda}}$ is asymptotically standard normal. This implies that the power of the test will converge to α. Specifically, we construct such f with an expression $\sum_{i=1}^n K(x_i, \cdot) w_i$, where w_i is selected from the orthogonal complement of a subsapce properly generated by S and K. When $s \ll s^*$, we can show that $\widehat{\mu}_{s+1} \gg d_n^{*2}$ with high probability, which implies that even if the norm of f is greater than d_n^*, our test still cannot achieve high power. Theorem 4.7 can be generalized in a similar manner for any S satisfying Assumption A3.

References

[1] Nir Ailon and Bernard Chazelle. The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on computing*, 39(1):302–322, 2009.

[2] Ahmed Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. *Advances in Neural Information Processing Systems*, pages 775–783, 2015.

[3] Theodore W Anderson and Donald A Darling. Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2):193–212, 1952.

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

[5] Albert Bifet, Geoff Holmes, Richard Kirkby, and Bernhard Pfahringer. Moa: Massive online analysis. *Journal of Machine Learning Research*, 11(May):1601–1604, 2010.

[6] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. *Advanced Lectures on Machine Learning*, pages 169–207, 2004.

[7] Christos Boutsidis and Alex Gittens. Improved matrix algorithms via the subsampled randomized hadamard transform. *SIAM Journal on Matrix Analysis and Applications*, 34(3): 1301–1340, 2013.

[8] Mikio L Braun. Accurate error bounds for the eigenvalues of the kernel matrix. *Journal of Machine Learning Research*, 7(Nov):2303–2328, 2006.

[9] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[10] Xiangyu Chang, Shaobo Lin, and Dingxuan Zhou. Distributed semi-supervised learning with kernel ridge regression. *Journal of Machine Learning Research*, 18(46):1–22, 2017.

[11] Dennis Cox, Eunmee Koh, Grace Wahba, and Brian S Yandell. Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *The Annals of Statistics*, 16(1):113–119, 1988.

[12] Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

[13] Jianqing Fan, Chunming Zhang, and Jian Zhang. Generalized likelihood ratio statistics and wilks phenomenon. *The Annals of statistics*, 29(1):153–193, 2001.

[14] Alex Gittens and Michael W Mahoney. Revisiting the nyström method for improved large-scale machine learning. *Journal of Machine Learning Research*, 28(3):567–575, 2013.

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

[16] Wensheng Guo. Inference in smoothing spline analysis of variance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):887–898, 2002.

[17] Cheng Huang and Xiaoming Huo. A statistically and numerically efficient independence test based on random projections and distance covariance. *arXiv preprint arXiv:1701.06054*, 2017.

[18] Yuri I Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. i, ii, iii. *Mathematical Methods of Statistics*, 2(2):85–114, 1993.

[19] Michael I Jordan and RI Jacobs. Supervised learning and divide-and-conquer: A statistical approach. *Proceedings of the Tenth International Conference on Machine Learning*, pages 159–166, 2014.

[20] YoungJu Kim and Chong Gu. Smoothing spline gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):337–356, 2004.

[21] Daniel A Klain. Invariant valuations on star-shaped sets. *Advances in Mathematics*, 125(1): 95–113, 1997.

[22] Anna Liu and Yuedong Wang. Hypothesis testing in smoothing spline models. *Journal of Statistical Computation and Simulation*, 74(8):581–597, 2004.

[23] Miles Lopes, Laurent Jacob, and Martin J Wainwright. A more powerful two-sample test in high dimensions using random projection. *Advances in Neural Information Processing Systems*, pages 1206–1214, 2011.

[24] Ping Ma, Nan Zhang, Jianhua Z Huang, and Wenxuan Zhong. Adaptive basis selection for exponential family smoothing splines with application in joint modeling of multiple sequencing samples. *Statistica Sinica*, 27:1757–1777, 2017.

[25] Cameron Musco and Christopher Musco. Recursive sampling for the nyström method. *Advances in Neural Information Processing Systems*, pages 3836–3848, 2017.

[26] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(82):1–9, 2013.

[27] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, pages 1657–1665, 2015.

[28] Adrien Saumard and Jon A Wellner. Log-concavity and strong log-concavity: a review. *Statistics surveys*, 8:45, 2014.

[29] Elizabeth D Schifano, Jing Wu, Chun Wang, Jun Yan, and MingHui Chen. Online updating of statistical inference in the big data setting. *Technometrics*, 58(3):393–403, 2016.

[30] Bernhard Schölkopf, Christopher JC Burges, and Alexander J Smola. *Advances in kernel methods: support vector learning*. MIT press, 1999.

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

[32] Zuofeng Shang and Guang Cheng. Computational limits of a distributed algorithm for smoothing spline. *Journal of Machine Learning Research*, 18(108):1–37, 2017.

[33] John Shawe-Taylor, Christopher KI Williams, Nello Cristianini, and Jaz Kandola. On the eigenspectrum of the gram matrix and the generalization error of kernel-pca. *IEEE Transactions on Information Theory*, 51(7):2510–2522, 2005.

[34] 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.

[35] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[36] Stanislav Volgushev, Shih-Kang Chao, and Guang Cheng. Distributed inference for quantile regression processes. *arXiv preprint arXiv:1701.06088*, 2017.

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

[38] Yuting Wei and Martin J Wainwright. The local geometry of testing in ellipses: Tight control via localized kolmogorov widths. *arXiv preprint arXiv:1712.00711*, 2017.

[39] Yuting Wei, Fanny Yang, and Martin J Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. In *Advances in Neural Information Processing Systems*, pages 6067–6077, 2017.

[40] Yun Yang, Mert Pilanci, and Martin J Wainwright. Randomized sketches for kernels: Fast and optimal non-parametric regression. *The Annals of Statistics*, 456:991 – 1023, 2017.

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

Appendix A. Proof of Main Results

In this section, we present main proofs of Lemma 3.1, Theorem 4.2, Theorem 4.4, Theorem 4.5, Theorem 4.6, Theorem 4.7 and Theorem 4.8 in the main text.

A.1. Proof of Lemma 3.1

Proof By Lemma 3.2, there exist fixed points r_λ and $\widehat{r}_\lambda$ for Ψ_λ and $\widehat{\Psi}_\lambda$, respectively. Plugging these fixed points into (B.1) and (B.2) in Lemma B.1, we have

$$r_\lambda \asymp \sqrt{\frac{1}{n}\sum_{i=1}^{\infty} \kappa_\lambda \min\Big\{\frac{r_\lambda}{\kappa_\lambda}, \mu_i\Big\}}, \tag{A.1}$$

$$\widehat{r}_\lambda \asymp \sqrt{\frac{1}{n}\sum_{i=1}^{n} \kappa_\lambda \min\Big\{\frac{\widehat{r}_\lambda}{\kappa_\lambda}, \widehat{\mu}_i\Big\} + \frac{c_1\delta}{n}}. \tag{A.2}$$

Lemma 3.2 further shows that $r_\lambda \asymp s_\lambda/n$, and $r_\lambda/\kappa_\lambda \asymp \lambda$; for the empirical version, let $\delta = s_\lambda$, then with probability at least $1 - 4e^{-s_\lambda}$, $\widehat{r}_\lambda \asymp s_\lambda/n$ leads to $\widehat{r}_\lambda/\kappa_\lambda \asymp \lambda$. Recall that $\widehat{s}_\lambda = \operatorname{argmin}\{i : \widehat{\mu}_i \le \lambda\} - 1$. Then by (A.2), with probability at least $1 - 4e^{-s_\lambda}$,

$$\frac{1}{n}\sum_{i=\widehat{s}_\lambda+1}^{n} \widehat{\mu}_i \le \frac{1}{n}\sum_{i=1}^{n} \min\big\{\lambda, \widehat{\mu}_i\big\} \asymp \frac{1}{n}\sum_{i=1}^{n} \min\Big\{\frac{\widehat{r}_\lambda}{\kappa_\lambda}, \widehat{\mu}_i\Big\} \lesssim \widehat{r}_\lambda^2/\kappa_\lambda \asymp \lambda s_\lambda/n,$$

where the last step is by $\kappa_\lambda = \frac{s_\lambda}{n\lambda}$, and $\widehat{r}_\lambda \asymp s_\lambda/n$. Therefore,

$$\sum_{i=\widehat{s}_\lambda+1}^{n} \widehat{\mu}_i \lesssim \lambda s_\lambda \leq s_\lambda \mu_{s_\lambda},$$

based on the definition (3.1) that $\lambda < \mu_{s_\lambda}$. ■

A.2. Proof of Theorem 4.2

Proof Let $\Delta = \boldsymbol{K}S^\top(S\boldsymbol{K}^2S^\top + \lambda S\boldsymbol{K}S^\top)^{-1}S\boldsymbol{K}$, under the null hypothesis, $T_{n,\lambda} = \frac{1}{n}\epsilon^\top\Delta^2\epsilon$. We first derive the testing consistency of $T_{n,\lambda}$ conditional on $\boldsymbol{x} = (x_1, \cdots, x_n)$ and S. By the Gaussian assumption of ϵ, we have $\mu_{n,\lambda} \equiv \mathrm{E}\left(T_{n,\lambda}|\boldsymbol{x}, S\right) = \frac{\mathrm{tr}(\Delta^2)}{n}$ and $\sigma^2_{n,\lambda} \equiv \mathrm{Var}\left(T_{n,\lambda}|\boldsymbol{x}, S\right) = 2\,\mathrm{tr}(\Delta^4)/n^2$. Define $U = \frac{T_{n,\lambda}-\mu_{n,\lambda}}{\sigma_{n,\lambda}}$, then for any $t \in (-1/2, 1/2)$, we have

$$\begin{aligned}
&\log \mathrm{E}_\epsilon\left(\exp(itU)\right)\\
=&\log \mathrm{E}_\epsilon\left(\exp(it\epsilon^\top\Delta^2\epsilon/(n\sigma_{n,\lambda}))\right) - it\mu_{n,\lambda}/(n\sigma_{n,\lambda})\\
=&-\frac{1}{2}\log\det(I_n - 2it\Delta^2/(n\sigma_{n,\lambda})) - it\mu_{n,\lambda}/(n\sigma_{n,\lambda})\\
=&it\cdot\mathrm{tr}(\Delta^2)/(n\sigma_{n,\lambda}) - t^2\,\mathrm{tr}(\Delta^4)/(n^2\sigma^2_{n,\lambda}) + \mathcal{O}(t^3\,\mathrm{tr}(\Delta^6)/(n^3\sigma^3_{n,\lambda})) - it\mu_{n,\lambda}/(n\sigma_{n,\lambda})\\
=&-t^2/2 + \mathcal{O}(t^3\,\mathrm{tr}(\Delta^6)/(n^3\sigma^3_{n,\lambda})),
\end{aligned}$$

where $i = \sqrt{-1}$, E_ϵ is the expectation with respect to ϵ, and I_n is $n \times n$ identity matrix. Therefore, to prove the normality of U, we need to show $\mathrm{tr}(\Delta^6)/(n^3\sigma^3_{n,\lambda}) = o(1)$. Note that

$$\frac{\mathrm{tr}(\Delta^6)}{(n^3\sigma^3_{n,\lambda})} \asymp \frac{\mathrm{tr}(\Delta^6)}{\mathrm{tr}(\Delta^4)}\cdot\frac{1}{\sqrt{\mathrm{tr}(\Delta^4))}}$$

where $\mathrm{tr}(\Delta^6) = \mathrm{tr}\left((I+\lambda(S\boldsymbol{K}^2S^\top)^{-1}S\boldsymbol{K}S^\top)^{-6}\right)$ and $\mathrm{tr}(\Delta^4) = \mathrm{tr}\left((I+\lambda(S\boldsymbol{K}^2S^\top)^{-1}S\boldsymbol{K}S^\top)^{-4}\right)$. Since $\mathrm{tr}(\Delta^6)/\mathrm{tr}(\Delta^4) < 1$, it is sufficient to prove $\frac{1}{\mathrm{tr}(\Delta^4)} = o(1)$ as $n \to \infty$.

Let $(S\boldsymbol{K}^2S^\top)^{-1}S\boldsymbol{K}S^\top = P\Lambda P^{-1}$, where Λ is an $s \times s$ diagonal matrix, then

$$\mathrm{tr}(\Delta^4) = \mathrm{tr}\left((I+\lambda\Lambda)^{-4}\right) = \sum_{i=1}^{s}(1+\lambda\Lambda_i)^{-4},$$

with Λ_i as the ith diagonal element in Λ. Next we show $\lambda\Lambda$ has at least $\min\{s, \widehat{s}_\lambda\}$ bounded eigenvalues. Notice that $(S\boldsymbol{K}^2S^\top)^{-1}S\boldsymbol{K}S^\top$ has the same non-zero eigenvalues as $\boldsymbol{K}^{1/2}S^\top(S\boldsymbol{K}^2S^\top)^{-1}S\boldsymbol{K}^{1/2}$. For $\boldsymbol{K} = UDU^\top$, let $U = (U_s, U_{n-s})$, $D = (D_s, D_{n-s})$ with $D_s = \mathrm{diag}\{\widehat{\mu}_1, \cdots, \widehat{\mu}_s\}$, $D_{n-s} = \mathrm{diag}\{\widehat{\mu}_{s+1}, \cdots, \widehat{\mu}_n\}$. Let $\widetilde{S}_1 = SU_s$, $\widetilde{S}_2 = SU_{n-s}$, $\boldsymbol{K}^{1/2}S^\top(S\boldsymbol{K}^2S^\top)^{-1}S\boldsymbol{K}^{1/2}$ can be rewritten as the block matrix:

$$\boldsymbol{K}^{1/2}S^\top(S\boldsymbol{K}^2S^\top)^{-1}S\boldsymbol{K}^{1/2} = \begin{pmatrix} D_s^{1/2}\widetilde{S}_1^\top \\ D_{n-s}^{1/2}\widetilde{S}_2^\top \end{pmatrix}(S\boldsymbol{K}^2S^\top)^{-1}\begin{pmatrix} \widetilde{S}_1D_s^{1/2} & \widetilde{S}_2D_{n-s}^{1/2} \end{pmatrix} = \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix},$$

where $A_1 = D_s^{1/2}\widetilde{S}_1^\top(S\boldsymbol{K}^2S^\top)^{-1}\widetilde{S}_1D_s^{1/2}$. By Lemma B.5 of the eigenvalue interlacing for principal submatrices theorem, we only need to prove λA_1 has at least $\min\{s, s_\lambda\}$ bounded eigenvalues. Using Binomial Inverse Theorem,

$$(S\boldsymbol{K}^2S^\top)^{-1} = (\widetilde{S}_1D_s^2\widetilde{S}_1^T)^{-1} - (\widetilde{S}_1D_s^2\widetilde{S}_1^T)^{-1}\Gamma(\widetilde{S}_1D_s^2\widetilde{S}_1^T)^{-1}, \tag{A.3}$$

where Γ is a symmetric matrix defined as

$$\Gamma = \widetilde{S}_2D_{n-s}^2\widetilde{S}_2^\top\left(\widetilde{S}_2D_{n-s}^2\widetilde{S}_2^\top + \widetilde{S}_2D_{n-s}^2\widetilde{S}_2^\top(\widetilde{S}_1D_1^2\widetilde{S}_1^T)^{-1}\widetilde{S}_2D_{n-s}^2\widetilde{S}_2^\top\right)^{-1}\widetilde{S}_2D_{n-s}^2\widetilde{S}_2^\top.$$

Plugging (A.3) into A_1, we have

$$D_s^{1/2}\widetilde{S}_1^\top(SK^2S^\top)^{-1}\widetilde{S}_1D_s^{1/2} = D_s^{-1} - H,$$

where H is a semi-positive matrix. Based on Lemma B.6 of Weyl's inequality, the i^{th} eigenvalue of D_s^{-1} is greater than the i^{th} eigenvalue of A_1. Recall $\widehat{s}_\lambda = \operatorname{argmin}\{i : \widehat{\mu}_i \le \lambda\} - 1$, we have $\lambda/\widehat{\mu}_i \le 1$ for $i = 1, \cdots, \widehat{s}_\lambda$. Hence, there exist at least $\min\{s, \widehat{s}_\lambda\}$ bounded eigenvalues for λA_1. Finally, we have

$$\operatorname{tr}(\Delta^4) \ge C\min\{s, \widehat{s}_\lambda\}, \quad \text{where } C \text{ is some constant.} \tag{A.4}$$

When $n \to \infty$ and $\lambda \to 0$, we have $s \to \infty$ and $s_\lambda \to \infty$ by Assumption A2. On the other hand, by Lemma 3.2, with probability at least $1 - e^{-cs_\lambda}$, $s_\lambda \asymp \widehat{s}_\lambda$ for $1/n < \lambda < 1$. Therefore, $\widehat{s}_\lambda \to \infty$ as $n \to \infty$ and $\lambda \to 0$. Then $\mathrm{E}_\epsilon(e^{itU}) \longrightarrow e^{-\frac{t^2}{2}}$ with probability approaches 1 as $n \to \infty$ and $\lambda \to 0$.

We next consider $\mathrm{E}_{\boldsymbol{x},S}\,\mathrm{E}_\epsilon(e^{itU})$ by taking expectation w.r.t $\boldsymbol{x}, S$ on $\mathrm{E}_\epsilon(e^{itU})$. We claim $\mathrm{E}_{\boldsymbol{x},S}\,\mathrm{E}_\epsilon(e^{itU}) \longrightarrow e^{-\frac{t^2}{2}}$ for $t \in (-\frac{1}{2}, \frac{1}{2})$. If not, there exists a subsequence of r.v $\{\boldsymbol{x}_{n_k}, S_{n'_k}\}$, such that for $\forall\varepsilon > 0$, $|\,\mathrm{E}_{\boldsymbol{x}_{n_k},S_{n_k}}\,\mathrm{E}_\epsilon\, e^{itU} - e^{-\frac{t^2}{2}}| > \varepsilon$. On the other hand, since $\mathrm{E}_\epsilon\, e^{itU(\boldsymbol{x}_{n_k},S_{n_k})} \xrightarrow{P} e^{-\frac{t^2}{2}}$, which is bounded, there exists a sub-sub sequence $\{\boldsymbol{x}_{n_{k_l}}, S_{n_{k'_l}}\}$, such that

$$\mathrm{E}_\epsilon\, e^{itU(\boldsymbol{x}_{n_{k_l}},S_{n_{k'_l}})} \xrightarrow{a.s} e^{-\frac{t^2}{2}}.$$

Thus by dominate convergence theorem, $\mathrm{E}_{\boldsymbol{x}_{n_{k_l}},S_{n_{k'_l}}}\,\mathrm{E}_\epsilon\, e^{itU} \longrightarrow e^{-\frac{t^2}{2}}$, which is a contradiction. Therefore, we have $U = \frac{T_{n,\lambda}-\mu_{n,\lambda}}{\sigma_{n,\lambda}}$ asymptotically converges to a standard normal distribution. ■

A.3. Proof of Theorem 4.4

Proof Under H_0^{linear}, it can be shown that

$$T^*_{n,\lambda} = \frac{1}{n}\epsilon'(I-H)\Delta^2(I-H)\epsilon$$

with $\mu^*_{n,\lambda} = \operatorname{tr}\left((I-H)\Delta^2(I-H)\right)/n$ and $\sigma^*_{n,\lambda} = \sqrt{2\operatorname{tr}\left((I-H)\Delta^4(I-H)\right)}/n$.

Similar to Theorem 4.2, we only need to prove $\operatorname{tr}\left((I-H)\Delta^2(I-H)\right) \to \infty$ as $n \to \infty$. Notice that $\operatorname{tr}\left((I-H)\Delta^2(I-H)\right) = \operatorname{tr}\left(\Delta^2(I-H)\right) = \operatorname{tr}(\Delta^2) - \operatorname{tr}(\Delta^2H)$. For $\operatorname{tr}(\Delta^2H)$, we have $\operatorname{tr}(\Delta^2H) = \operatorname{tr}(\Delta^2H^2) = \operatorname{tr}(H\Delta^2H)$. Since $rank(H\Delta^2H) \le a+1$, and $\lambda_{\max}(H\Delta H) \le 1$, therefore $\operatorname{tr}(\Delta^2H) \le a+1$. Recall a is the dimension of x. Finally we have $\operatorname{tr}\left((I-H)\Delta^2(I-H)\right) \ge \min\{s, \widehat{s}_\lambda\} - a - 1$. The last step is based on the proof of Theorem 4.2. ■

A.4. Proof of Theorem 4.5

We prove the the testing is minimax optimal as stated in Theorem 4.5.

Proof

$$\begin{aligned} n\|\widehat{f}_R\|_n^2 =& n\|\boldsymbol{K}S^\top(S\boldsymbol{K}^2S^\top+\lambda S\boldsymbol{K}S^\top)^{-1}S\boldsymbol{K}f+\boldsymbol{K}S^\top(S\boldsymbol{K}^2S^\top\lambda S\boldsymbol{K}S^\top)^{-1}S\boldsymbol{K}\epsilon\|_n^2 \\ =& n\|\operatorname{E}_\epsilon \widehat{f}_R\|_n^2+n\|\boldsymbol{K}S^\top(S\boldsymbol{K}^2S^\top+\lambda S\boldsymbol{K}S^\top)^{-1}S\boldsymbol{K}\epsilon\|_n^2 \\ &+2(\operatorname{E}_\epsilon \widehat{\boldsymbol{f}}_R)^\top\boldsymbol{K}S^\top(S\boldsymbol{K}^2S^\top+\lambda S\boldsymbol{K}S^\top)^{-1}S\boldsymbol{K}\epsilon \\ :=& T_1+T_2+T_3. \end{aligned} \tag{A.5}$$

Lemma B.3 shows that $\|f-\operatorname{E}_\epsilon \widehat{f}_R\|_n^2\le C\lambda$ with probability $1-e^{-c_1 s}-e^{-c_2 s_\lambda}$. Set $C'=\sqrt{2C}$. Given the separation rate $\|f\|_n^2\ge C'^2 d_{n,\lambda}^2=2C(\lambda+\sigma_{n,\lambda})$, we have

$$T_1=n\|\operatorname{E}_\epsilon \widehat{f}_R\|_n^2\ge\frac{n}{2}\|f\|_n^2-n\|f-\operatorname{E}_\epsilon \widehat{f}_R\|_n^2\ge nC(\lambda+\sigma_{n,\lambda})-nC\lambda\ge nC\sigma_{n,\lambda}$$

with probability at least $1-e^{-c_1 s}-e^{-c_2 s_\lambda}$, where c_1, c_2 is specified in Assumption A3.

Next, notice that $T_3=(\operatorname{E}_\epsilon \widehat{\boldsymbol{f}}_R)^\top\Delta\epsilon$. Consider $\eta^\top\Delta^2\eta$, where $\eta=(\eta_1,\cdots,\eta_n)\in\mathbb{R}^n$ is an arbitrary vector. Since $\eta^\top\Delta^2\eta\le\lambda_{\max}(\Delta^2)\eta^\top\eta$, where Δ^2 has the same non-zero eigenvalue as

$$\widetilde{\Delta}^2=(S\boldsymbol{K}^2S^\top+\lambda S\boldsymbol{K}S^\top)^{-1}S\boldsymbol{K}^2S^\top(S\boldsymbol{K}^2S^\top+\lambda S\boldsymbol{K}S^\top)^{-1}S\boldsymbol{K}^2S^\top=(I+\lambda(S\boldsymbol{K}^2S^\top)^{-1}S\boldsymbol{K}S)^{-2},$$

then we have $\|\widetilde{\Delta}^2\|_{\text{op}}\le 1$, and $\lambda_{\max}(\Delta^2)\le 1$. Therefore,

$$\operatorname{E}_\epsilon T_3^2=(\operatorname{E}_\epsilon \widehat{\boldsymbol{f}}_R)^\top\Delta^2(\operatorname{E}_\epsilon \widehat{\boldsymbol{f}}_R)\le(\operatorname{E}_\epsilon \widehat{\boldsymbol{f}}_R)^\top(\operatorname{E}_\epsilon \widehat{\boldsymbol{f}}_R)=T_1,$$

then

$$\operatorname{P}\Big(|T_3|\ge\varepsilon^{-\frac{1}{2}}T_1^{1/2}\Big|\boldsymbol{x},S\Big)\le\frac{\operatorname{E}_\epsilon T_3^2}{\varepsilon^{-1}T_1}\le\varepsilon$$

Define $\mathcal{E}_1=\{T_1\ge Cn\sigma_{n,\lambda}\}$, $\mathcal{E}_2=\{\frac{T_2/n-\mu_{n,\lambda}}{\sigma_{n,\lambda}}\le C_\epsilon\}$, where C_ε satisfies $\operatorname{P}(\mathcal{E}_2)\ge 1-\varepsilon$, $\mathcal{E}_3=\{T_3\ge-\varepsilon^{-1/2}T_1^{1/2}\}$. Finally, with probability at least $1-e^{-c_1 s}-e^{-c_2 s_\lambda}$,

$$\begin{aligned} &\operatorname{P}_f\Big(\frac{\frac{1}{n}(T_1+T_2+T_3)-\mu_{n,\lambda}}{\sigma_{n,\lambda}}\ge z_{1-\alpha/2}\Big|\boldsymbol{x},S\Big) \\ \ge&\operatorname{P}_f\Big(\frac{T_1+T_2}{n\sigma_{n,\lambda}}+\frac{\frac{1}{n}T_2-\mu_{n,\lambda}}{\sigma_{n,\lambda}}\ge z_{1-\alpha/2},\,\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\Big|\boldsymbol{x},S\Big) \\ \ge&\operatorname{P}_f\Big(\frac{T_1(1-\varepsilon^{-1/2}T_1^{-1/2})}{n\sigma_{n,\lambda}}-C_\varepsilon\ge z_{1-\alpha/2},\,\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\Big|\boldsymbol{x},S\Big) \\ \ge&\operatorname{P}_f\Big(C(1-\frac{1}{\sqrt{Cn\sigma_{n,\lambda}\varepsilon}})-C_\varepsilon\ge z_{1-\alpha/2},\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\Big) \\ =&\operatorname{P}_f(\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3)\ge 1-3\varepsilon \end{aligned}$$

The second to the last equality is achieved by choosing C to satisfy

$$\frac{1}{\sqrt{Cn\sigma_{n,\lambda}\varepsilon}}<\frac{1}{2}\quad\text{and}\quad\frac{1}{2}C-C_\varepsilon\ge z_{1-\alpha/2}.$$

■

A.5. Proof of Theorem 4.6

To prove Theorem 4.6, we only need to prove the orders of $\mu_{n,\lambda}$ and $\sigma^2_{n,\lambda}$. Suppose that $\mathcal{H}$ is generated by an m-order PDK or γ-order EDK. The following Lemma characterizes the orders of $\mu_{n,\lambda}$ and $\sigma^2_{n,\lambda}$ for PDK and EDK.

Lemma A.1

(a) *(PDK) Suppose that $1/n < \lambda \to 0$ as $n \to \infty$. Meanwhile, Assumption A3 holds with $c_1, c_2 \in (0, \infty]$ and $m > 3/2$. Then with probability at least $1 - e^{-c_m n^{(2m-3)/(2m-1)}} - e^{-c_1 s} - e^{-c_2 s_\lambda}$, it holds that $\mu_{n,\lambda} \asymp s_\lambda/n$ and $\sigma^2_{n,\lambda} \asymp s_\lambda/n^2$, where c_m is an absolute constant depending on m only.*

(b) *(EDK) Suppose that $\mathcal{H}$ is generated by EDK with $\gamma > 0$, $p \geq 1$. Suppose that Assumption A3 holds with $c_1, c_2 \in (0, \infty]$. Then with probability at least $1 - e^{-c_{\gamma,p} n (\log n)^{-2/p}} - e^{-c_1 s} - e^{-c_2 s_\lambda}$, it holds that $\mu_{n,\lambda} \asymp s_\lambda/n$ and $\sigma^2_{n,\lambda} \asymp s_\lambda/n^2$, where $c_{\gamma,p}$ is an absolute constant depending on γ, p.*

Proof We first analyze the orders of $\mu_{n,\lambda}$ and $\sigma_{n,\lambda}$ for PDK, i.e., Lemma A.1 (a). Recall in (A.4), we proved that $\mathrm{tr}(\Delta) \gtrsim \min\{s, \widehat{s}_\lambda\}$. Next we show with probability approaching 1,

$$\mathrm{tr}(\Delta^4) \leq \mathrm{tr}(\Delta^2) \leq \mathrm{tr}(\Delta) \lesssim \widehat{s}_\lambda. \tag{A.6}$$

On the other hand, when $\lambda \geq 1/n$, by Lemma B.2 (a), with probability at least $1 - e^{-c_m n^{(2m-3)/(2m-1)}}$, $\widehat{s}_\lambda \asymp s_\lambda$. Combining (A.4) with (A.6), we have $\sigma^2_{n,\lambda} \asymp s_\lambda/n^2$ and $\mu_{n,\lambda} \asymp s_\lambda/n$ with probability approaching 1.

Note that $\mathrm{tr}(\Delta) = \mathrm{tr}(\widetilde{\Delta})$, where $\widetilde{\Delta} = DU^\top (S\boldsymbol{K}^2 S^\top + \lambda S \boldsymbol{K} S^\top)^{-1} SUD$. $\widetilde{\Delta}$ can be written as

$$\widetilde{\Delta} = \begin{pmatrix} \widetilde{\Delta}_1 & \widetilde{\Delta}_2 \\ \widetilde{\Delta}_3 & \widetilde{\Delta}_4 \end{pmatrix}$$

with $\widetilde{\Delta}_1 = D_s \widetilde{S}_1^\top (S\boldsymbol{K}^2 S^\top + \lambda S \boldsymbol{K} S^\top)^{-1} \widetilde{S}_1 D_s$, and $\widetilde{\Delta}_4 = D_{n-s} \widetilde{S}_2^\top (S\boldsymbol{K}^2 S^\top + \lambda S \boldsymbol{K} S^\top)^{-1} \widetilde{S}_2 D_{n-s}$. Here $D_s = \mathrm{diag}\{\widehat{\mu}_1, \cdots, \widehat{\mu}_s\}$, $D_{n-s} = \mathrm{diag}\{\widehat{\mu}_{s+1}, \cdots, \widehat{\mu}_n\}$, $\widetilde{S}_1 = SU_s$, and $\widetilde{S}_2 = SU_{n-s}$, where U_s is the first s column of U and U_{n-s} is the last $n-s$ column of U.

Let $\Lambda = \mathrm{diag}\{\Lambda_1, \Lambda_2\} = \mathrm{diag}\{D_s^2 + \lambda D_s, D_{n-s}^2 + \lambda D_{n-s}\}$. Then $\widetilde{\Delta}_1$ can be expressed as

$$\begin{aligned} \widetilde{\Delta}_1 &= D_s \widetilde{S}_1^\top (\widetilde{S}_1 \Lambda_1 \widetilde{S}_1^\top + \widetilde{S}_2 \Lambda_2 \widetilde{S}_2^\top)^{-1} \widetilde{S}_1 D_s \\ &= D_s^2 \Lambda_1^{-1} - D_s \widetilde{S}_1^T (\widetilde{S}_1 \Lambda_1 \widetilde{S}_1^\top)^{-1} \big((\widetilde{S}_2 \Lambda_2 \widetilde{S}_2^\top)^{-1} + (\widetilde{S}_1 \Lambda_1 \widetilde{S}_1^\top)^{-1} \big)^{-1} (\widetilde{S}_1 \Lambda_1 \widetilde{S}_1^\top)^{-1} \widetilde{S}_1 D_s. \end{aligned}$$

Therefore,

$$\mathrm{tr}\left(D_1 \widetilde{S}_1^\top (S\boldsymbol{K}^2 S^\top + \lambda S \boldsymbol{K} S^\top)^{-1} \widetilde{S}_1 D_1 \right) \leq \mathrm{tr}(D_1^2 \Lambda_1^{-1}) \leq s_\lambda. \tag{A.7}$$

The last inequality is deduced by the following step

$$\mathrm{tr}(D_1^2 \Lambda_1^{-1}) = \sum_{i=1}^{\widehat{s}_\lambda} \frac{\widehat{\mu}_i}{\widehat{\mu}_i + \lambda} + \sum_{i=\widehat{s}_\lambda + 1}^{s} \frac{\widehat{\mu}_i}{\widehat{\mu}_i + \lambda} \leq \widehat{s}_\lambda + \frac{1}{\lambda} \sum_{i=\widehat{s}_\lambda + 1}^{s} \widehat{\mu}_i \leq 2 s_\lambda$$

with probability at least $1-e^{-c_m n^{(2m-3)/(2m-1)}}-e^{-c_1 s}-e^{-c_2 s_\lambda}$ by Lemma B.2(a) and Lemma 3.1.

Next, we consider $\widetilde{\Delta}_4$. Note that

$$\begin{aligned}
&\operatorname{tr}\left(D_{n-s}\widetilde{S}_2^\top(S\boldsymbol{K}^2S^\top+\lambda S\boldsymbol{K}S^\top)^{-1}\widetilde{S}_2D_{n-s}\right)\\
\le&\operatorname{tr}\left(D_{n-s}\widetilde{S}_2^\top(\lambda S\boldsymbol{K}S^\top)^{-1}\widetilde{S}_2D_{n-s}\right)\\
\le&\frac{\widehat{\mu}_s}{\lambda}\operatorname{tr}\left(\left((\widetilde{S}_2D_{n-s}\widetilde{S}_2^\top)^{-1}\widetilde{S}_1D_s\widetilde{S}_1^\top+I_{s\times s}\right)^{-1}\right)\le s\widehat{\mu}_s/\lambda. \qquad \text{(A.8)}
\end{aligned}$$

We show $s\widehat{\mu}_s/(\lambda s_\lambda)\le C$ with probability approaching 1, where C is some absolute constant. Then by (A.8), $\operatorname{tr}(\widetilde{\Delta}_4)\lesssim s_\lambda$. If $ds_\lambda\le s\le n^{1/(2m)}$, then by Lemma B.2(a), we have $\mu_s/2\le\widehat{\mu}_s\le 3\mu_s/2$ with probability at least $1-e^{-c_m n^{(2m-3)/(2m-1)}}$, where $m>3/2$. Then

$$\frac{s\widehat{\mu}_s}{\lambda s_\lambda}\le\frac{3s\mu_s}{2\lambda s_\lambda}\lesssim\frac{s^{1-2m}}{s_\lambda^{1-2m}}=\mathcal{O}(1).$$

If $s\gg n^{1/(2m)}$, based on the proof of Lemma B.2, in (B.3) and (B.4),

$$\mathrm{P}\big(|\widehat{\mu}_i-\mu_i|\ge\mu_i\widetilde{\varepsilon}+r^{1-2m}\big)\le 1-\exp\big(-c_m'n\widetilde{\varepsilon}^2/r^2\big).$$

Let $\widetilde{\varepsilon}=n^{-\frac{2m-1}{2m}}s^{2m-1}$ and $r=n^{\frac{1}{2m}}s^{\frac{1}{2m-1}}$, then

$$s\widehat{\mu}_s\le s\mu_s\widetilde{\varepsilon}+sr^{1-2m}=n^{-\frac{2m-1}{2m}},$$

with probability at least $1-e^{-c'n^{(2m-3)/(2m-1)}}$. The probability is obtained by calculating $n\widetilde{\varepsilon}^2/r^2$. Based on the assumption $1/n\le\lambda\le 1$, $\lambda s_\lambda\asymp\lambda^{1-1/(2m)}\ge n^{-\frac{2m-1}{2m}}$. Finally we have $\frac{s\widehat{\mu}_s}{\lambda s_\lambda}\le C$, i.e., $s\widehat{\mu}_s/\lambda\le Cs_\lambda$, where C is some bounded constant.

Combining with (A.7), we have

$$\operatorname{tr}(\Delta)=\operatorname{tr}(\widetilde{\Delta})\le\operatorname{tr}(\widetilde{\Delta}_1)+\operatorname{tr}(\widetilde{\Delta}_4)\lesssim s_\lambda,$$

with probability at least $1-e^{-c'n^{(2m-3)/(2m-1)}}-e^{-c_1 s}-e^{-c_2 s_\lambda}$, where c' is a constant only depending on m and $c_1,c_2>0$ are defined in Assumption A3.

Next, we prove Lemma A.1 (b). Following the same notation and strategy in the proof of Lemma A.1 a, (A.7) also holds for EDK, with probability at least $1-e^{-c_{\gamma,p}n(\log n)^{-2/p}}-e^{-c_1 s}-e^{-c_2 s_\lambda}$ by Lemma B.2 (a) and Lemma 3.1. For EDK, (A.8) also holds. Next we will prove that $\operatorname{tr}(\widetilde{\Delta}_2)\le s\widehat{\mu}_s/\lambda\le Cs_\lambda$, where C is an absolute constant. If $ds_\lambda\le s\lesssim n^{1/2-\varepsilon}\triangleq n^a$ for any $0<\varepsilon<1/2$, then by Lemma B.2 (b), $\widehat{\mu}_s\le\frac{3}{2}\mu_s$. Therefore

$$\frac{s\widehat{\mu}_s}{\lambda s_\lambda}\lesssim\frac{s\mu_s}{s_\lambda\mu_{s_\lambda}}\le 1,$$

with probability at least $1-e^{-c_{\gamma,p}n(\log n)^{-2/p}}$, where the last inequality is by the fact that $s\mu_s\asymp se^{-\gamma s^p}$ is decreasing w.r.t s when $\gamma ps^p-1>0$. When $s\ge n^{1/2}$, by Lemma B.2 (b),

$$\frac{s\widehat{\mu}_s}{\lambda s_\lambda}\le\frac{s^2\mu_s}{\lambda s_\lambda}\lesssim\frac{ns^2}{s_\lambda}e^{-\gamma s^p}=o(1),$$

with probability at least $1-e^{-n}$. Thus, we achieve $\operatorname{tr}\left(D_{n-s}\widetilde{S}_2^\top(S\boldsymbol{K}^2S^\top+\lambda S\boldsymbol{K}S^\top)^{-1}\widetilde{S}_2D_{n-s}\right)\le s_\lambda$ with probability at least $1-e^{-c_{\gamma,p}n(\log n)^{-2/p}}$. Combining with (A.4) and (A.7), we have $\operatorname{tr}(\Delta)\lesssim s_\lambda$ with probability at least $1-e^{-c_{\gamma,p}n(\log n)^{-2/p}}-e^{-c_1 s}-e^{-c_2 s_\lambda}$. ■

A.6. Proof of Theorem 4.7

Proof Notice that

$$\begin{aligned}\|\widehat{f}_R - f_0\|_n^2 &= \|\,\mathrm{E}_\epsilon\,\widehat{f}_R - f_0\|_n^2 + \|\widehat{f}_R - \mathrm{E}_\epsilon\,\widehat{f}_R\|_n^2 + \frac{2}{n}\big(\widehat{f}_R - \mathrm{E}_\epsilon\,\widehat{f}_R\big)^\top\big(\mathrm{E}_\epsilon\,\widehat{f}_R - f_0\big)\\ &\equiv T_1 + T_2 + T_3.\end{aligned}$$

We first consider T_1 as follows:

$$\begin{aligned}\|\,\mathrm{E}_\epsilon\,\widehat{f}_R - f_0\|_n^2 &= \|\boldsymbol{K}S^\top(S\boldsymbol{K}^2S^\top + \lambda S\boldsymbol{K}S^\top)^{-1}S\boldsymbol{K}f_0 - f_0\|_n^2\\ &= \|U^T\boldsymbol{K}S^\top(S\boldsymbol{K}^2S^\top + \lambda S\boldsymbol{K}S^\top)^{-1}S\boldsymbol{K}f_0 - U^Tf_0\|_n^2.\end{aligned}$$

Let $S = U_s$, where U_s is the first s columns of U. Let $f_0(\cdot) = \sum_{i=1}^n K(x_i,\cdot)w_i$ with $w = (w_1,\cdots,w_n)^\top = U\alpha$, where $\alpha \in \mathbb{R}^n$ satisfies

$$\alpha_i^2 = \begin{cases}\frac{1}{n}\frac{C}{s}\widehat{\mu}_i^{-1}, & \text{for } i = s+1,\cdots,2s;\\ 0, & \text{otherwise.}\end{cases} \tag{A.9}$$

Then $\|f_0\|_{\mathcal{H}}^2 = n\alpha^\top D\alpha = n\sum_{i=s+1}^{2s}\alpha_i^2\widehat{\mu}_i = C$, and

$$\begin{aligned}\|\,\mathrm{E}_\epsilon\,\widehat{f}_R - f_0\|_n^2 &= n\sum_{i=1}^{s}\alpha_i^2(\widehat{\mu}_i^2(\widehat{\mu}_i+\lambda)^{-1} - \widehat{\mu}_i)^2 + n\sum_{i=s+1}^{n}\alpha_i^2\widehat{\mu}_i^2\\ &= \frac{1}{s}\sum_{i=s+1}^{2s}\widehat{\mu}_i \geq \widehat{\mu}_{2s} \geq \frac{1}{2}\mu_{2s} \gg \lambda^\dagger.\end{aligned} \tag{A.10}$$

The last inequality holds with probability greater than $1-e^{-n\delta_n}$ by Lemma B.2. On the other hand, there always exists $\lambda \gg \lambda^\dagger$, such that the corresponding $s_\lambda = s/d$. Then by (A.10), with probability greater than $1-e^{-n\delta_n}$,

$$\|\,\mathrm{E}_\epsilon\,\widehat{f}_R - f_0\|_n^2 \leq \frac{3}{2}\mu_s \leq \frac{3}{2}\mu_{s_\lambda} \asymp \lambda,$$

i.e., $T_1 = \mathcal{O}_P(\lambda)$, based on the definition (3.1) for s_λ.

Furthermore, we have $T_2 = \mathcal{O}_P(\mu_{n,\lambda}) = \mathcal{O}_P(\frac{s_\lambda}{n})$ by the proof of Corollary 4.3. Note that $\lambda^\dagger$ satisfies $\lambda^\dagger \asymp \frac{s_\lambda^\dagger}{n}$. Then for $\lambda \gg \lambda^\dagger$, we have $T_2 = o_p(T_1)$. Therefore, $T_3 = o_P(T_1)$ due to Cauchy-Schwarz inequality $T_3 \leq T_1^{1/2}T_2^{1/2}$. Finally, with probability at least $1-e^{-n\delta_n}$,

$$\sup_{f_0\in\mathcal{B}}\|\widehat{f}_R - f_0\|_n^2 \gtrsim \sup_{f_0\in\mathcal{B}}\|\,\mathrm{E}_\epsilon\,\widehat{f}_R - f_0\|_n^2 \geq C\mu_{2s} \gg C\lambda^\dagger.$$

The last step is based on the definition of $\lambda^\dagger$ and the fact that $2s \ll s^\dagger$. ■

A.7. Proof of Theorem 4.8

Proof Without loss of generality, here we consider $H_0 : f = f_0$ with $f_0 = 0$. We construct the true $f(\cdot) = \sum_{i=1}^n K(x_i,\cdot)w_i$ with $w = (w_1,\cdots,w_n)^\top = U\alpha$, where $\alpha \in \mathbb{R}^n$ satisfies

$$\alpha_i^2 = \begin{cases} \frac{1}{n}\frac{C}{s-1}\widehat{\mu}_{gs+k}^{-1} & \text{for } i = (gs+k) \quad k = 1, 2, \cdots, s-1; \\ 0 & \text{otherwise.} \end{cases} \tag{A.11}$$

Choose $g \geq 1$ to be an integer satisfying $(g+1)s \ll s^*$. By definition,

$$\|f\|_{\mathcal{H}}^2 = n\alpha^\top D\alpha = n\sum_{k=1}^{s-1} \alpha_{gs+k}^2 \widehat{\mu}_{gs+k} = C$$

and

$$\|f\|_n^2 = n\alpha^\top D^2\alpha = n\sum_{k=1}^{s-1} \alpha_{gs+k}^2 \widehat{\mu}_{gs+k}^2 = \frac{C}{s-1}\sum_{k=1}^{s-1} \widehat{\mu}_{gs+k}.$$

Then by Lemma B.2, with probability at least $1 - e^{-n\delta_n}$,

$$\|f\|_n^2 \geq \frac{C}{2}\mu_{gs+s} = \beta_{n,\lambda}^2 d^{*2},$$

where $\beta_{n,\lambda}^2 = \frac{C}{2}\mu_{(g+1)s}/\mu_{s^*}$, and $\beta_{n,\lambda}^2 \to \infty$ as $n \to \infty$, $d^{*2} = \lambda^* \asymp \mu_{s^*}$.

Let $S = U_s$, where U_s is the first s columns of U. Then S satisfies Assumption A3 with $c_1 = c_2 = +\infty$, i.e., the K-satisfiability holds almost surely. $SU = (\widetilde{S}_1\ \widetilde{S}_2)$, where $\widetilde{S}_1 = SU_s = I_{s\times s}$ and $\widetilde{S}_2 = SU_{n-s} = \mathbf{0}$. Recall in eq. (A.5) for $nT_{n,\lambda}$. Plugging S and f into T_1, we have

$$\begin{aligned} T_1 &= \boldsymbol{f}^\top \boldsymbol{K} S^\top (S\boldsymbol{K}^2 S^\top + \lambda S\boldsymbol{K}S^\top)^{-1} S\boldsymbol{K}^2 S^\top (S\boldsymbol{K}^2 S^\top + \lambda S\boldsymbol{K}S^\top)^{-1} S\boldsymbol{K}\boldsymbol{f} \\ &= \boldsymbol{f}^\top U D\widetilde{S}^\top (\widetilde{S}D^2\widetilde{S}^\top + \lambda\widetilde{S}D\widetilde{S}^\top)^{-1}\widetilde{S}D^2\widetilde{S}^\top (\widetilde{S}d\widetilde{S}^\top + \lambda\widetilde{S}D\widetilde{S}^\top)^{-1}\widetilde{S}DU^\top\boldsymbol{f} \\ &= n\sum_{i=1}^{s} \alpha_i^2 \frac{\widehat{\mu}_i^4}{(\widehat{\mu}_i + \lambda)^2} = 0, \end{aligned}$$

where the last step is by the construction of α that $\alpha_1 = \alpha_s = 0$. $T_1 = 0 \ll n\sigma_{n,\lambda}$. Furthermore, $|T_3| = T_1^{1/2} O_{P_f}(1) = o_{P_f}(n\sigma_{n,\lambda})$. Therefore

$$\begin{aligned} \frac{T_{n,\lambda} - \mu_{n,\lambda}}{\sigma_{n,\lambda}} &= \frac{T_1 + T_3}{n\sigma_{n,\lambda}} + \frac{T_2/n - \mu_{n,\lambda}}{\sigma_{n,\lambda}} \\ &= \frac{T_2/n - \mu_{n,\lambda}}{\sigma_{n,\lambda}} + o_{P_f}(\sigma_{n,\lambda}) \\ &\xrightarrow{d} N(0,1). \end{aligned}$$

Then we have, as $n \to \infty$, with probability at least $1 - e^{-cn\delta_n} - e^{-c_1 s} - e^{-c_2 s_\lambda}$,

$$\inf_{f\in\mathcal{B}, \|f\|_n \geq \beta_{n,\lambda} d^*} \mathrm{P}_f\big(\phi_{n,\lambda} = 1 | \boldsymbol{x}, S\big) \leq \mathrm{P}_f\big(\phi_{n,\lambda} = 1 | \boldsymbol{x}, S\big) \to \alpha.$$

■

Appendix B. Additional Proofs and Technical Lemmas

B.1. Some key lemmas

We first show that $\Psi_\lambda(r)$ and $\widehat{\Psi}_\lambda(r)$ have an asymptotically equivalent expression in terms of μ_i's ($\widehat{\mu}_i$'s) for a wide-ranging α, where $\widehat{\mu}_i$'s are the eigenvalues of $\boldsymbol{K}$ (in a descending order $\widehat{\mu}_1 \geq \widehat{\mu}_2 \geq \cdots \geq \widehat{\mu}_n \geq 0$). Recall that μ_i's are eigenvalues of the kernel function K; see (2.1). $\Psi_\lambda(r)$ and $\widehat{\Psi}_\lambda(r)$ are defined in (3.3) and (3.4), respectively.

Lemma B.1

(a) *Suppose $\mu_1 > 1/n$. For any $\lambda > 1/n$, it holds that*

$$\Psi_\lambda(r) \asymp \sqrt{\frac{1}{n}\sum_{i=1}^{\infty} \kappa_\lambda \min\{\frac{r}{\kappa_\lambda}, \mu_i\}}. \tag{B.1}$$

(b) *For any $\lambda > 0$, it holds that*

$$\widehat{\Psi}_\lambda(r) \asymp \sqrt{\frac{1}{n}\sum_{i=1}^{n} \kappa_\lambda \min\{\frac{r}{\kappa_\lambda}, \widehat{\mu}_i\}}. \tag{B.2}$$

Proof We first prove (B.2). Define $\boldsymbol{x} = (x_1, \cdots, x_n)$. Let $(K, \langle\cdot,\cdot\rangle)$ be a RKHS $\mathcal{H}$. Any $f \in \mathcal{H}$ can be presented as $f(\cdot) = \sum_{i=1}^n c_i K(x_i, \cdot) + \xi(\cdot)$ with $\xi(\cdot) \perp \text{span}\{K(x_1,\cdot), \cdots, K(x_n,\cdot)\}$. Therefore $f(x_j) = \sum_{i=1}^n c_i K(x_i, x_j)$ for $j = 1, \cdots, n$. Let $\boldsymbol{f} = (f(x_1), \cdots, f(x_n))^\top$, we have $\boldsymbol{f} = n\boldsymbol{K}\boldsymbol{c}$, where $\boldsymbol{K}$ is the kernel matrix, $\boldsymbol{c} = (c_1, \cdots, c_n)^\top$. Let $\boldsymbol{K} = \Phi\Phi^\top$ with $\Phi = UD^{1/2}$. Then $n\boldsymbol{K}\boldsymbol{c} = n\Phi\Phi^\top\boldsymbol{c} = \sqrt{n}\Phi\beta$ with $\sqrt{n}\Phi^\top\boldsymbol{c} \equiv \beta$. Then we have

$$\sum_{i=1}^n f^2(x_i) = n^2\boldsymbol{c}^\top\boldsymbol{K}^2\boldsymbol{c} = n\beta^\top\Phi^\top\Phi\beta = n\beta^\top D^{1/2}U^\top UD^{1/2}\beta = n\sum_{i=1}^n \beta_i^2\widehat{\mu}_i.$$

Note that $P_n f^2 = \frac{1}{n}\sum_{i=1}^n f(x_i)^2 \leq r$ is equivalent to $\sum_{i=1}^n \beta_i^2\widehat{\mu}_i \leq r$. Therefore,

$$\sup_{\substack{f\in\mathcal{F}_\lambda \\ P_n f^2 \leq r}} |\sum_{i=1}^n \sigma_i f(x_i)|^2 = \sup_{\substack{f\in\mathcal{F}_\lambda \\ P_n f^2 \leq r}} n^2|\sigma^\top\boldsymbol{K}\boldsymbol{c}|^2 = \sup_{\substack{\beta\in\mathbb{R}^n, \sum_{i=1}^n \beta_i^2 \leq \kappa_\lambda \\ \sum_{i=1}^n \beta_i^2\widehat{\mu}_i \leq r}} n|\sigma^\top\Phi\beta|^2$$

Define $\widehat{F}_\lambda = \{\beta \in \mathbb{R}^n | \sum_{i=1}^n \beta_i^2 \leq \kappa_\lambda, \sum_{i=1}^n \beta_i^2\widehat{\mu}_i \leq r\}$, and $\widetilde{F}_\lambda = \{\beta \in \mathbb{R}^n | \sum_{i=1}^n \widehat{d}_i\beta_i^2 \leq 1\}$, where $\widehat{d}_i = (\kappa_\lambda \min\{1, r/(\kappa_\lambda\widehat{\mu}_i)\})^{-1}$. So $\widetilde{F}_\lambda \subseteq \widehat{F}_\lambda \subseteq \sqrt{2}\widetilde{F}_\lambda$. Let $\Lambda = \text{diag}\{\widehat{d}_1, \cdots, \widehat{d}_n\}$. Then we have

$$\begin{aligned}
&\mathrm{E}\Big(\sup_{\substack{\beta\in\mathbb{R}^n, \sum_{i=1}^n \beta_i^2 \leq \kappa_\lambda \\ \sum_{i=1}^n \beta_i^2\widehat{\mu}_i \leq r}} |\sigma^\top\Phi\beta|^2 \Big| \boldsymbol{x}\Big) \asymp \mathrm{E}\Big(\sup_{\beta\in\widetilde{F}_\lambda} |\sigma^\top\Phi\Lambda^{-1/2}\Lambda^{1/2}\beta|^2 \Big| \boldsymbol{x}\Big) \\
=&\mathrm{E}\Big(\sup_{\substack{d\in\mathbb{R}^n \\ d'd\leq 1}} |\sigma^\top\Phi\Lambda^{-1/2}d|^2 \Big| \boldsymbol{x}\Big) = \mathrm{E}\Big(\|\sigma^\top\Phi\Lambda^{-1/2}\|_2^2 \Big| \boldsymbol{x}\Big) = \mathrm{E}\Big(\sigma^\top\Phi\Lambda^{-1}\Phi^\top\sigma \Big| \boldsymbol{x}\Big)
\end{aligned}$$

Note that

$$\mathrm{E}\left(\sigma^\top \Phi\Lambda^{-1}\Phi^\top\sigma\middle|\boldsymbol{x}\right) = \mathrm{tr}\left(\Phi\Lambda^{-1}\Phi^\top\right) = \mathrm{tr}\left(\Lambda^{-1}\mathrm{diag}\{\widehat{\mu}_1,\cdots,\widehat{\mu}_n\}\right) = \sum_{i=1}^{n}\frac{\widehat{\mu}_i}{\widehat{\widehat{d}}_i}$$

Therefore, by Kahane-Khintchine inequality, we have

$$\widehat{\Psi}_\lambda(r) \asymp \sqrt{\frac{1}{n}\sum_{i=1}^{n}\widehat{\mu}_i/\widehat{d}_i} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\kappa_\lambda \min\{\widehat{\mu}_i, \frac{r}{\kappa_\lambda}\}}.$$

Similarily, we can achieve (B.1). ■

B.2. Properties of eigenvalues

Lemma B.2

(a) *Suppose that K has eigenvalues satisfying $\mu_i \asymp i^{-2m}$ with $m > 3/2$. Then for $i = 1,\cdots,n^{1/(2m)}$,*

$$P\left(|\widehat{\mu}_i - \mu_i| \le \frac{1}{2}\mu_i\right) \ge 1 - e^{-c_m n i^{-4m/(2m-1)}}.$$

where c_m is an universal constant depending only on m.

(b) *Suppose that K has eigenvalues satisfying $\mu_i \asymp \exp(-\gamma i^p)$ with $\gamma > 0$, $p \ge 1$. Then for $i = o(n^{1/2})$,*

$$P\big(|\widehat{\mu}_i - \mu_i| \le \frac{1}{2}\mu_i\big) \ge 1 - e^{-c_{\gamma,p} n i^{-2}},$$

where $c_{\gamma,p}$ is an universal constant depending only on γ and p.
For $i = O(n^{1/2})$, we have

$$P\left(|\widehat{\mu}_i - \mu_i| \le i\mu_i\right) \ge 1 - e^{-c'_{\gamma,p} n},$$

where $c'_{\gamma,p}$ is an universal constant depending only on γ and p.

Proof We apply the proof of Theorem 3 in [8] to deduce our results. Recall in Theorem 3 of [8], for $1 \le i \le n$, $1 \le r \le n$,

$$|\widehat{\mu}_i - \mu_i| \le \mu_i\|C_n^r\| + \mu_r + \Lambda_{>r}, \tag{B.3}$$

where $\Lambda_{>r} = \sum_{i=r+1}^{\infty}\mu_i$, and $\|C_n^r\|$ satisfies

$$\mathrm{P}\Big(\|C_n^r\| \ge \widetilde{\varepsilon}\Big) \le r(r+1)e^{-\frac{n\widetilde{\varepsilon}^2}{2M^4 r^2}}, \tag{B.4}$$

based on Lemma 7 in [8]. M is an absolute constant here.

First we prove Lemma B.2 (a). Consider the polynomial decaying kernel with $\mu_i \asymp i^{-2m}$. Notice that

$$\Lambda_{>r} \asymp \sum_{i=r+1}^{\infty} i^{-2m} \le \int_r^\infty x^{-2m}dx = \frac{r^{1-2m}}{2m-1}$$

Consider $i = 1, \cdots, n^{\frac{1}{2m}}$, let

$$\frac{r^{1-2m}}{2m-1} = \frac{1}{4}\mu_i,$$

then $r = a_m i^{2m/(2m-1)}$, where a_m is a constant only depends on m. Let $\widetilde{\varepsilon} = \frac{1}{4}$. Plugging r into (B.4), we have

$$\mathrm{P}\Big(\|C_n^r\| \geq \frac{1}{4}\Big) \leq e^{-c_m n i^{-4m/(2m-1)}},$$

where $c_m = (64M^4 a_m^2)^{-1}$ is an universal constant depends on m. Then we obtain that

$$\mathrm{P}\Big(|\widehat{\mu}_i - \mu_i| \leq \frac{1}{2}\mu_i\Big) \geq 1 - e^{-c_m n i^{-4m/(2m-1)}}.$$

Next, we prove Lemma B.2 (b). Consider the exponential decaying kernel with $\mu_i \asymp e^{-\gamma i^p}$. For $1 \leq r \leq n$, when $p = 1$, then

$$\Lambda_{>r} \asymp \sum_{i=r+1}^{\infty} e^{-\gamma i} \leq \int_r^{\infty} e^{-\gamma x} dx = \frac{e^{-\gamma r}}{\gamma};$$

when $p \geq 2$, using integration by parts, we have

$$\begin{aligned}\Lambda_{>r} \asymp \sum_{i=r+1}^{\infty} e^{-\gamma i^p} &\leq \int_r^{\infty} e^{-\gamma x^p} dx \\ &= \frac{1}{\gamma p r^{p-1}} e^{-\gamma r^p} - \int_r^{\infty} \frac{p-1}{\gamma p x^p} e^{-\gamma x^p} dx \leq a_{\gamma,p} e^{-\gamma r^p}.\end{aligned}$$

For $i = o(n^{1/2})$, let $\mu_r + \Lambda_{>r} \leq (1 + a_{\gamma,p}) e^{-\gamma r^p} = \frac{1}{4}\mu_i$, we have $r = b_{\gamma,p} i$, where $b_{\gamma,p}$ is a constant only depends on γ, p. Then plugging $\widetilde{\epsilon} = \frac{1}{4}$ and r into (B.4), we have

$$\mathrm{P}\Big(\|C_n^r\| \geq \frac{1}{4}\Big) \leq e^{-c_{\gamma,p} n i^{-2}}.$$

where $c_{\gamma,p} = (64M^4 b_{\gamma,p}^2)^{-1}$ is an absolute constant only depends on γ, p. Finally, by (B.3), we have

$$\mathrm{P}\Big(|\widehat{\mu}_i - \mu_i| \leq \frac{1}{2}\mu_i\Big) \geq 1 - e^{-c_{\gamma,p} n i^{-2}}.$$

When $i \geq n^{1/2}$, we do not need a very tight bound. Let $\widetilde{\varepsilon} = i$, $r = i$, then we have

$$\mathrm{P}\Big(|\widehat{\mu}_i - \mu_i| \leq i\mu_i\Big) \geq 1 - e^{-c'_{\gamma,p} n},$$

where $c'_{\gamma,p}$ is an absolute constant only depends on γ, p. ■

B.3. Proof of Lemma 3.2

Proof We first observe that $\mathcal{F}_\lambda = \text{star}(\mathcal{F}_\lambda, 0)$, the star-hull of $\mathcal{F}_\lambda$ at zero. Note that the supremum in the definitions of Ψ_λ and $\widehat{\Psi}_\lambda$ is based on "quadratic type" constraints $Pf^2 \le r$ and $P_n f^2 \le r$. Then following [6], Ψ_λ and $\widehat{\Psi}_\lambda$ are both sub-root functions, and thus have unique nonzero fixed points. Also refer to [21] for the definitions of star-hull and sub-root functions. Define

$$\widehat{\Psi}'_\lambda(r) = \widehat{\Psi}_\lambda(r) + \frac{c_1\delta}{n} = \mathrm{E}\Big\{ \sup_{\substack{f\in\mathcal{F}_\lambda \\ P_n f^2 \le r}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i) \Big| \boldsymbol{x} \Big\} + \frac{c_1\delta}{n},$$

where c_1 is a constant. Then by Theorem 4.2 in [4], the fixed points r_λ and $\widehat{r}_\lambda$ of $\Psi_\lambda(r)$ and $\widehat{\Psi}'_\lambda(r)$ satisfy: $r_\lambda \asymp \widehat{r}_\lambda$ with probability at least $1 - 4e^{-\delta}$, provided that $r_\lambda \ge c_1\delta/n$.

Let $r = r_\lambda$ in Lemma B.1, we have

$$r_\lambda \asymp \sqrt{\frac{1}{n}\sum_{i=1}^\infty \kappa_\lambda \min\big\{\frac{r_\lambda}{\kappa_\lambda}, \mu_i\big\}}. \tag{B.5}$$

Define $\eta_\lambda = \text{argmin}\{i : \mu_i \le r_\lambda/\kappa_\lambda\} - 1$, then (B.5) implies

$$\frac{nr_\lambda^2}{\kappa_\lambda} \asymp \eta_\lambda \frac{r_\lambda}{\kappa_\lambda} + \sum_{i=\eta_\lambda+1}^\infty \mu_i. \tag{B.6}$$

Note that

$$\frac{\sum_{i=k+1}^\infty \mu_i}{k\mu_{k+1}} = \frac{1}{k} + \frac{k+1}{k}\cdot\frac{\sum_{i=k+2}^\infty \mu_i}{(k+1)\mu_{k+1}} \le 1 + 2C,$$

where $C = \sup_{k\ge1} \frac{\sum_{i=k+1}^\infty \mu_i}{k\mu_k} < \infty$ by Assumption A1. Therefore, $\sum_{i=\eta_\lambda+1}^\infty \mu_i \lesssim \eta_\lambda\mu_{\eta_\lambda+1} \le \eta_\lambda\frac{r_\lambda}{\kappa_\lambda}$. Then by (B.6), we have $\frac{nr_\lambda^2}{\kappa_\lambda} \asymp \eta_\lambda\frac{r_\lambda}{\kappa_\lambda}$, i.e., $r_\lambda \asymp \frac{\eta_\lambda}{n}$. Note

$$\mu_{\eta_\lambda} > \frac{r_\lambda}{\kappa_\lambda} \asymp \frac{\eta_\lambda}{n\kappa_\lambda} = \frac{\lambda\eta_\lambda}{s_\lambda},$$

and recall $s_\lambda = \text{argmin}\{i : \mu_i \le \lambda\} - 1$ which implies $\mu_{s_\lambda+1} \le \lambda < \mu_{s_\lambda}$. Then,

$$\frac{\mu_{\eta_\lambda}}{\eta_\lambda} \gtrsim \frac{\lambda}{s_\lambda} \ge \frac{\mu_{s_\lambda+1}}{s_\lambda} \ge \frac{\mu_{s_\lambda+1}}{s_\lambda+1}. \tag{B.7}$$

Note that μ_k/k is a decreasing function of k, we thus have $\eta_\lambda \lesssim s_\lambda + 1$ by (B.7), i.e., $\eta_\lambda \lesssim s_\lambda$. On the other hand,

$$\mu_{\eta_\lambda+1} \le \frac{r_\lambda}{\kappa_\lambda} \asymp \frac{\lambda\eta_\lambda}{s_\lambda} \lesssim \frac{\mu_{s_\lambda}(\eta_\lambda+1)}{s_\lambda},$$

i.e., $\frac{\mu_{\eta_\lambda+1}}{\eta_\lambda+1} < \frac{\mu_{s_\lambda}}{s_\lambda}$, and we have $\eta_\lambda + 1 \gtrsim s_\lambda$, i.e., $\eta_\lambda \gtrsim s_\lambda$. Therefore, $\eta_\lambda \asymp s_\lambda$. Then, we achieve that $r_\lambda \asymp \frac{s_\lambda}{n}$. Suppose there exists an constant c_2, such that $r_\lambda \ge c_2\frac{s_\lambda}{n}$, let $\delta = c_2 s_\lambda/c_1$, then with probability greater than $1 - e^{-cs_\lambda}$, $\widehat{r}_\lambda \asymp r_\lambda \asymp s_\lambda/n$, where $c = \frac{c_2}{2c_1}$. ■

B.4. Bound of the bias term in $r_{n,\lambda}$ in Corollary 4.3

In this section, we show that the bound of the bias term in $r_{n,\lambda}$ in Corollary 4.3.

Lemma B.3 *Suppose that* $1/n < \lambda < 1$ *and Assumption A3 holds with* $c_1, c_2 \in (0, \infty]$. *Then with probability greater than* $1 - e^{-c_1 s} - e^{-c_2 s_\lambda}$,

$$\| \mathrm{E}_\epsilon \, \widehat{f}_R - f_0 \|_n^2 \leq C\lambda,$$

where C *is a positive absolute constant.*

Proof Suppose the true function is f_0, then $y_i = f_0(x_i) + \epsilon_i$ for $i = 1, \cdots, n$. Notice that $\mathrm{E}_\epsilon \, \widehat{f}_R(\cdot) = \sum_{i=1}^n (S\beta^\dagger)_i K(\cdot, x_i)$ with $\beta^\dagger = \frac{1}{n}(S\boldsymbol{K}^2 S^\top + \lambda S\boldsymbol{K}S^\top)^{-1} S\boldsymbol{K}\boldsymbol{f}_0$, where $\boldsymbol{f}_0 = (f_0(x_1), \cdots, f_0(x_n))^\top$. It is in fact the solution of a noiseless version of quadratic program:

$$\beta^\dagger = \operatorname*{argmin}_{\beta \in \mathbb{R}^s} \Big\{ \frac{1}{n} \| \boldsymbol{f}_0 - n\boldsymbol{K}S^\top \beta \|_2^2 + n\lambda \beta^\top S\boldsymbol{K}S^\top \beta \Big\}. \tag{B.8}$$

To prove $\| \mathrm{E}_\epsilon \, \widehat{f}_R - f_0 \|_n^2 \leq C\lambda$ with probability approaching 1, we only need to find an $\widetilde{\beta}$, such that

$$\frac{1}{n} \| \boldsymbol{f}_0 - n\boldsymbol{K}S^\top \widetilde{\beta} \|_2^2 + n\lambda \widetilde{\beta}^\top S\boldsymbol{K}S^\top \widetilde{\beta} \leq C\lambda$$

with probability at least $1 - e^{-c_1 s} - e^{-c_2 s_\lambda}$, where c_1, c_2 are defined in Assumption A3. Note that $\boldsymbol{K} = UDU^\top$. Setting $\boldsymbol{z} = \frac{1}{\sqrt{n}} U^\top \boldsymbol{f}_0$, (B.8) is equivalent to (B.9) as follows:

$$\beta^\dagger = \operatorname*{argmin}_{\beta \in \mathbb{R}^s} \Big\{ \| \boldsymbol{z} - \sqrt{n} D\widetilde{S}^\top \beta \|_2^2 + n\lambda \beta^\top \widetilde{S}^\top D \widetilde{S} \beta \Big\}. \tag{B.9}$$

Let $\boldsymbol{z} = (z_1, z_2)^\top$, where $z_1 \in \mathbb{R}^{\widehat{s}_\lambda}$, and $z_2 \in \mathbb{R}^{n - \widehat{s}_\lambda}$. Correspondingly, divide D into D_1, D_2, where $D_1 = \mathrm{diag}\{\widehat{\mu}_1, \cdots, \widehat{\mu}_{\widehat{s}_\lambda}\}$ and $D_2 = \mathrm{diag}\{\widehat{\mu}_{\widehat{s}_\lambda + 1}, \cdots, \widehat{\mu}_n\}$. Denote $\widetilde{S} = SU = (\widetilde{S}_1, \widetilde{S}_2)$ with $\widetilde{S}_1 \in \mathbb{R}^{s \times s_\lambda}$ as the left block and $\widetilde{S}_2 \in \mathbb{R}^{s \times (n - s_\lambda)}$ as the right block. We construct an $\widetilde{\beta}$ as

$$\widetilde{\beta} = \frac{1}{\sqrt{n}} \widetilde{S}_1 (\widetilde{S}_1^\top \widetilde{S}_1)^{-1} D_1^{-1} z_1 \in \mathbb{R}^s.$$

Plugging $\widetilde{\beta}$ into (B.9), we see that

$$\| \boldsymbol{z} - \sqrt{n} D\widetilde{S}^\top \widetilde{\beta} \|_2^2 = \| z_1 - \sqrt{n} D_1 \widetilde{S}_1^\top \widetilde{\beta} \|_2^2 + \| z_2 - D_2 \widetilde{S}_2^\top \widetilde{S}_1 (\widetilde{S}_1^\top \widetilde{S}_1)^{-1} D_1^{-1} z_1 \|_2^2 = T_1^2 + T_2^2.$$

It is obvious that $T_1^2 = 0$, and next we analyze T_2.

Note that for any $f_0(\cdot) \in \mathcal{H}$, there exists an $n \times 1$ vector ω, such that $f_0(\cdot) = \sum_{i=1}^n K(\cdot, x_i)\omega_i + \xi(\cdot)$, where $\xi(\cdot) \in \mathcal{H}$ is orthogonal to the span of $\{K(\cdot, x_i), i = 1, \cdots, n\}$. Then, $\xi(x_j) = \langle \xi, K(\cdot, x_j) \rangle = 0$, and $f_0(x_j) = \sum_{i=1}^n K(x_i, x_j)\omega_i$. Therefore $\boldsymbol{f}_0 = n\boldsymbol{K}\omega$, where $\boldsymbol{K}$ is the empirical kernel matrix. Suppose $\| f_0 \|_{\mathcal{H}} \leq 1$, then

$$n\omega^\top \boldsymbol{K}\omega \leq \| f_0 \|_{\mathcal{H}} \leq 1 \Rightarrow n\omega^\top \boldsymbol{K}\boldsymbol{K}^{-1}\boldsymbol{K}\omega^\top \leq 1 \Rightarrow \frac{1}{n} \boldsymbol{f}_0{}^\top \boldsymbol{K}^{-1} \boldsymbol{f}_0 \leq 1 \Rightarrow \frac{1}{n} \boldsymbol{f}_0{}^\top U D^{-1} U^\top \boldsymbol{f}_0 \leq 1,$$

which leads to the ellipse constrain that $\|D^{-1/2}\boldsymbol{z}\|_2 \leq 1$, where $\boldsymbol{z} = \frac{1}{\sqrt{n}}U^\top \boldsymbol{f}_0$. Obviously, $\|D_1^{-1/2}z_1\|_2 \leq 1$, $\|D_2^{-1/2}z_2\|_2 \leq 1$. Notice that $\frac{1}{\lambda}\boldsymbol{f}^\top U_2U_2^\top \boldsymbol{f} \leq \boldsymbol{f}^\top U_2 D_2^{-1}U_2^\top \boldsymbol{f} < n$, then $\boldsymbol{f}^\top U_2U_2^\top \boldsymbol{f} \leq n\lambda$, and

$$T_2 \leq \|z_2\|_2 + \|\sqrt{D_2}\|_{\text{op}}\|\sqrt{D_2}\widetilde{S}_2^\top\|_{\text{op}}\|\widetilde{S}_1\|_{\text{op}}\|(\widetilde{S}_1^\top\widetilde{S}_1)^{-1}\|_{\text{op}}\|D_1^{-1/2}\|_{\text{op}}\|D_1^{-1/2}z_1\|_{\text{op}} \leq (1+3c)\sqrt{\lambda}$$

with probability at least $1-e^{-c_1 s}-e^{-c_2 s_\lambda}$ by Assumption A3. Therefore, we have $\|\boldsymbol{z}-D\widetilde{S}^\top\widetilde{\beta}\|_2^2 \leq c'\lambda$, where $c' = (1+3c)^2$. For the penalty term,

$$\begin{aligned} n\widetilde{\beta}^\top S\boldsymbol{K}S^\top\widetilde{\beta} \leq & z_1^\top D_1^{-1}z_1 + \|z_1^\top D_1^{-1/2}\|_2\|D_1^{-1/2}\|_{\text{op}}\|\widetilde{S}_2 D_2^{1/2}\|_{\text{op}}\|D_2^{1/2}\widetilde{S}^\top\|_{\text{op}}\|D_1^{-1/2}\|_{\text{op}}\|D_1^{-1/2}z_1\|_{\text{op}} \\ \leq & 1+c^2, \end{aligned}$$

where c is constant from the definition 1. Finally, we can claim that

$$\|\operatorname{E}_\epsilon \widehat{f}_R - f_0\|_n^2 \leq \|\boldsymbol{z}-\sqrt{n}D\widetilde{S}^\top\widetilde{\beta}\|_2^2 + n\lambda\widetilde{\beta}^\top\widetilde{S}^\top D\widetilde{S}\widetilde{\beta} \leq C\lambda$$

with probability at least $1-e^{-c_1 s}-e^{-c_2 s_\lambda}$, where c is some constant, $C = 2+10c^2+6c$ is an absolute constant. ■

B.5. Proof of Corollary 4.3

Proof Denote $\operatorname{E}_\epsilon \widehat{f}_R$ as the the expectation of $\widehat{f}_R$ w.r.t ϵ. Note that

$$\|\widehat{f}_R - f_0\|_n^2 \leq 2\|\widehat{f}_R - \operatorname{E}_\epsilon \widehat{f}_R\|_n^2 + 2\|\operatorname{E}_\epsilon \widehat{f}_R - f_0\|_n^2$$

and $\|\widehat{f}_R - \operatorname{E}_\epsilon \widehat{f}_R\|_n^2 = \frac{\epsilon^\top}{\sqrt{n}}\Delta^2\frac{\epsilon}{\sqrt{n}}$, where $\|\frac{\epsilon}{\sqrt{n}}\|_{\psi_2} \leq \frac{L}{\sqrt{n}}$ and $\|\Delta^2\|_{\text{op}} \leq 1$. Recall $\|\cdot\|_{\psi_2}$ is the sub-Gaussian norm. Here $\|\epsilon\|_{\psi_2} \leq L$, with L as an absolute constant. Then by Hanson-Wright concentration inequality ([26]) (stated in Lemma B.4), with probability greater than $1-e^{-c_1 s}-e^{-c_2 s_\lambda}$,

$$\begin{aligned} &\operatorname{P}\Big(\|\widehat{f}_R - \operatorname{E}_\epsilon \widehat{f}_R\|_n^2 - \operatorname{E}_\epsilon\|\widehat{f}_R - \operatorname{E}_\epsilon \widehat{f}_R\|_n^2 \geq \frac{\operatorname{tr}(\Delta^2)}{2n}\Big|\boldsymbol{x}, S\Big) \\ =&\operatorname{P}\Big(\frac{1}{n}\epsilon^\top\Delta^2\epsilon - \frac{\operatorname{tr}(\Delta^2)}{n} \geq \frac{\operatorname{tr}(\Delta^2)}{2n}\Big|\boldsymbol{x}, S\Big) \\ \leq &\exp\Big(-c\min\big(\frac{\operatorname{tr}^2(\Delta^2)}{4K^4\|\Delta^2\|_{\text{F}}^2}, \frac{\operatorname{tr}(\Delta^2)}{\|\Delta^2\|_{\text{op}}}\big)\Big) \leq \exp(-c\operatorname{tr}(\Delta^2)), \end{aligned}$$

where $\|\cdot\|_{\text{F}}$ is the Frobenius norm. The last inequality holds by the fact that $\|\Delta^2\|_{\text{F}}^2 \leq \|\Delta^2\|_{\text{op}}\operatorname{tr}(\Delta^2)$ and $\|\Delta^2\|_{\text{op}} \leq 1$. Lastly, by (A.4), $\operatorname{tr}(\Delta^2) \geq \min\{s, \widehat{s}_\lambda\} \geq \widehat{s}_\lambda$, which goes to $+\infty$ as $n \to \infty$, we have that, with probability approaching 1, $\|\widehat{f}_R - \operatorname{E}_\epsilon \widehat{f}_R\|_n^2 \leq \frac{3}{2}\mu_{n,\lambda}$.

Meanwhile, it follows from Lemma B.3 that $\|\operatorname{E}_\epsilon \widehat{f}_R - f_0\|_n^2 = O_P(\lambda)$. This completes the proof of Corollary 4.3. ■

B.6. Proof of Lemma 4.1

In this section, we first prove the following (B.10) and (B.11):

$$\mathrm{P}\Big(\frac{1}{\sqrt{2}} \le \lambda_{\min}(SU_1) \le \lambda_{\max}(SU_1) \le \sqrt{\frac{3}{2}}\Big|\boldsymbol{x}\Big) \ge 1 - \exp\big(-c_1 s\big) \tag{B.10}$$

almost surely, where $c_1 > 0$ is an absolute constant independent of n, s; $\lambda_{\min}(SU_1)$ ($\lambda_{\max}(SU_1)$) is the smallest (largest) singular value of SU_1.

$$\mathrm{P}\Big(\mathrm{P}\big(\|SU_2 D_2^{1/2}\|_{\mathrm{op}} \le c\lambda\big|\boldsymbol{x}\big) \ge 1 - e^{-c_1' s}\Big) \ge 1 - e^{-c_2 s_\lambda}, \tag{B.11}$$

where c and c_1' are constants independent of n, s and $c_2 = 1/2$. The result of Lemma 4.1 directly follow from (B.10) and (B.11).

Proof For $\boldsymbol{K} = UDU^\top$, let $U = (U_1, U_2)$ with $U_1 \in \mathbb{R}^{n\times \widehat{s}_\lambda}$, and $U_2 \in \mathbb{R}^{n\times(n-\widehat{s}_\lambda)}$. Recall $S = \frac{1}{\sqrt{s}} S^*$, where S^* is the random matrix with independent centered sub-Gaussian entries (with variance as one), then each row S_i^* is independent sub-Gaussian isotropic random vectors in $\mathbb{R}^n$, i.e., $\mathrm{E}\, S_i^* S_i^{*\top} = I_{n\times n}$, $i = 1, \cdots, s$. Let $SU_1 = \frac{1}{\sqrt{s}}(\eta_1, \cdots, \eta_s)^\top$, where $\eta_i \in \mathbb{R}^{\widehat{s}_\lambda \times 1}$ with each entry $\eta_{ij} = S_i^{*\top} U_{1(j)}$, $U_{1(j)}$ is the jth column of U_1, $j = 1, \cdots, \widehat{s}_\lambda$.

Firstly, conditional on $\boldsymbol{x}$, by the definition of sub-Gaussian random vector, each entry η_{ij} is sub-Gaussian, η_i and $\eta_j (i \ne j)$ are independent, and η_i is isotropic sub-Gaussian random vector due to the fact that $\mathrm{E}(\eta_i \eta_i^\top | \boldsymbol{x}) = U_1^\top (\mathrm{E}\, S_i^* S_i^{*\top}) U_1 = I_{\widehat{s}_\lambda \times \widehat{s}_\lambda}$. By Theorem 5.39 in [35], for any $t > 0$,

$$\begin{aligned}&\mathrm{P}\Big(\frac{\sqrt{s} - C\sqrt{s_\lambda} - t}{\sqrt{s}} \le \lambda_{\min}(SU_1) \le \lambda_{\max}(SU_1) \le \frac{\sqrt{s} + C\sqrt{s_\lambda} + t}{\sqrt{s}}\Big|\boldsymbol{x}\Big)\\ \ge& 1 - 2e^{-ct^2}\end{aligned}$$

almost surely. Let $t = \frac{\sqrt{s}}{5}$, and $d \ge (0.02C)^2$, we have

$$\mathrm{P}\Big(\frac{1}{\sqrt{2}} \le \lambda_{\min}(SU_1) \le \lambda_{\max}(SU_1) \le \sqrt{\frac{3}{2}}\Big|\boldsymbol{x}\Big) \ge 1 - 2e^{-cs/25} \tag{B.12}$$

almost surely. Here $C, c > 0$ only depend on the sub-Gaussian norm $L := \max_i \|\eta_i\|_{\psi_2}$ conditional on $\boldsymbol{x}$. Note that $\eta_i = S_i^{*\top} U_i$,

$$\|U_i^\top S_i^*\|_{\psi_2} = \sup_{\nu \in \mathcal{S}^{s_\lambda - 1}} \|\langle U_1^\top S_i^*, \nu\rangle\|_{\psi_2} = \sup_{\kappa \in \mathcal{S}^{n-1}} \|\kappa^\top S_i^*\|_{\psi_2} \quad \text{and}$$

$$\sup_{\kappa \in \mathcal{S}^{n-1}} \|\kappa^\top S_i^*\|_{\psi_2}^2 = \|\sum_{j=1}^n \kappa_j S_{ij}^*\|_{\psi_2}^2 \le C \sum_{j=1}^n \kappa_j^2 \|S_{ij}^*\|_{\psi_2}^2 \le C \max_{1\le j\le n} \|S_{ij}^*\|_{\psi_2}^2$$

Therefore, $L \le \max_{i,j} \|S_{ij}^*\|_{\psi_2}$, which is bounded. Lastly, we have

$$\mathrm{P}\Big(\mathrm{P}\Big(\frac{1}{\sqrt{2}} \le \lambda_{\min}(SU_1) \le \lambda_{\max}(SU_1) \le \sqrt{\frac{3}{2}}\Big|\boldsymbol{x}\Big) \ge 1 - 2e^{-cs/25}\Big) = 1.$$

Set $\tilde{c}_1 = c/32$. Then (B.10) has been proved.

Next, we prove (B.11). Define $\mathcal{A} = \{\boldsymbol{x} : \boldsymbol{x} \text{ satisfies } \sum_{i=\widehat{s}_\lambda+1}^{n} \widehat{\mu}_i \le C s_\lambda \mu_{s_\lambda}\}$. Then $\mathrm{P}(\boldsymbol{x} \in \mathcal{A}) \ge 1 - 4e^{-s_\lambda}$ by Lemma 3.1.

Since $(SU_2 D_2^{1/2})^\top (SU_2 D_2^{1/2})$ has the same non-zero eigenvalues as $SU_2 D_2 U_2^\top S^\top$, it is equivalent to prove $\lambda_{\max}(SU_2 D_2 U_2^\top S^\top) \lesssim \lambda$, where $\lambda_{\max}(\cdot)$ refers to the maximum singular value. For every $\nu \in \mathcal{S}^{s-1}$, $\nu = \kappa + w$, where κ belongs to the $1/2$-net $\mathcal{N} = \{\mu_1, \cdots, \mu_M\}$ of the set $\mathcal{S}^{s-1}$, here $M \le e^{2s}$; and $\|w\| \le 1/2$, where $\|\cdot\|$ is the Euclidean norm. Then

$$\begin{aligned}
\|SU_2 D_2 U_2^\top S^\top\|_{\mathrm{op}} &= \sup_{\|\nu\|=1, \nu \in \mathcal{S}^{s-1}} \|SU_2 D_2 U_2^\top S^\top \nu\| \\
&\le \sup_{\kappa \in \mathcal{S}^{s-1}} \|SU_2 D_2 U_2^\top S^\top \kappa\| + \sup_{w \in \mathcal{S}^{s-1}} \|SU_2 D_2 U_2^\top S^\top w\| \\
&\le \max_{\kappa \in \mathcal{S}^{s-1}} \|SU_2 D_2 U_2^\top S^\top \kappa\| + \frac{1}{2} \|SU_2 D_2 U_2^\top S^\top\|_{\mathrm{op}},
\end{aligned}$$

therefore

$$\begin{aligned}
\|SU_2 D_2 U_2^\top S^\top\|_{\mathrm{op}} &\le 2 \max_{\kappa \in \mathcal{S}^{s-1}} \|SU_2 D_2 U_2^\top S^\top \kappa\| = 2 \max_{\kappa \in \mathcal{S}^{s-1}} |\langle SU_2 D_2 U_2^\top S^\top \kappa, \kappa \rangle| \\
&= 2 \max_{\kappa \in \mathcal{S}^{s-1}} |\kappa' S U_2 D_2 U_2^\top S^\top \kappa| = \frac{2}{s} \max_{\kappa \in \mathcal{S}^{s-1}} |\kappa' S^* U_2 D_2 U_2^\top S^{*\top} \kappa|,
\end{aligned}$$

where the last equality is by the definition that $S = S^*/\sqrt{s}$.

Note that $\eta = S^{*\top} \kappa \in \mathbb{R}^n$ is a sub-Gaussian vector, and $\eta_i = \sum_{j=1}^{s} \kappa_j S^*_{ji}$ is independent with η_j for $i, j \in \{1, \cdots, n\}, i \ne j$; also, $\mathrm{E}(\eta_i) = 0$,

$$\mathrm{Var}\left(\sum_{j=1}^{s} \kappa_j S^*_{ji}\right) = \sum_{j=1}^{s} \kappa_j^2 \,\mathrm{Var}(S^*_{ji}) \le C,$$

where the last inequality is due to the fact that $\sum_{j=1}^{s} \kappa_j^2 = 1$. Let $Q = \frac{1}{s} U_2 D_2 U_2^\top$. By Hanson-Wright inequality (stated in Lemma B.4), we have

$$\mathrm{P}\big(|\eta' Q \eta - \mathrm{tr}(Q)| \ge t \big| \boldsymbol{x} \in \mathcal{A}\big) \le 2e^{-c \min\{\frac{t^2}{K^4 \|Q\|_{\mathrm{F}}^2}, \frac{t}{K^2 \|Q\|_{\mathrm{op}}}\}}. \tag{B.13}$$

Conditional on $\boldsymbol{x} \in \mathcal{A}$, $\mathrm{tr}(Q) = \frac{1}{s}\mathrm{tr}(D_2 U_2^\top U_2) = \frac{1}{s}\sum_{i=\widehat{s}_\lambda+1}^{n} \widehat{\mu}_i \le \frac{s_\lambda \lambda}{s} \le L\lambda$, where L is some absolute constant by the assumption that $s \ge d s_\lambda$. The penultimate inequality is based on Lemma 3.1 and the definition of s_λ in eq. (3.1). Also, note that

$$\|Q\|_{\mathrm{F}}^2 = \frac{1}{s^2}\mathrm{tr}(D_2^2) \le \frac{1}{s^2} \widehat{\mu}_{s_\lambda+1} \sum_{i=\widehat{s}_\lambda+1}^{n} \widehat{\mu}_i \le \frac{s_\lambda}{s^2} \widehat{\mu}_{s_\lambda+1} \lambda$$

the last step is based on Lemma 3.1 for $\boldsymbol{x} \in \mathcal{A}$. Let $t = L\lambda/2$, then

$$\frac{t^2}{\|Q\|_{\mathrm{F}}^2} \ge \frac{\lambda s}{\widehat{\mu}_{s_\lambda+1}} = \frac{\lambda}{\|Q\|_{\mathrm{op}}}.$$

Therefore, (B.13) can be further stated as

$$\mathrm{P}(|\eta' Q \eta - \mathrm{tr}(Q)| \ge L\lambda/2 | \boldsymbol{x} \in \mathcal{A}) \le 2e^{-cL\lambda s/(2\widehat{\mu}_{s_\lambda+1})} \le 2e^{-c' L s},$$

where the last inequality is by the definition of $\widehat{\mu}_{s_\lambda+1}$. Finally, taking union bound over all $\mu \in \mathcal{N}$, we have

$$\begin{aligned}&\mathrm{P}\Big(\mathrm{P}\big(\|SU_2D_2U_2^\top S^\top\|_{\mathrm{op}} \le 3L\lambda|\boldsymbol{x}\big) \ge 1 - e^{-(c'L-2)s}\Big)\\ \ge&\mathrm{P}\Big(\mathrm{P}\big(\|SU_2D_2U_2^\top S^\top\|_{\mathrm{op}} \le 3L\lambda|\boldsymbol{x} \in \mathcal{A}\big) \ge 1 - e^{-(c'L-2)s}\Big) \cdot \mathrm{P}\Big(\boldsymbol{x} \in \mathcal{A}\Big) \ge 1 - 4e^{-s_\lambda}.\end{aligned}$$

Let $c = 3L$, $c_1' = c'L - 2 > 0$ and $c_2 = 1/2$. Then, we have proved (B.11). Finally, taking $c_1 = \min\{\tilde{c}_1, c_1'\}$, where $\tilde{c}_1$ refers to (B.10) and c_1' refers to (B.11), Lemma 4.1 have been proved. ■

B.7. Verification of Assumption A1

Let us verify Assumption A1 in PDK and 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

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

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

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

The corresponding eigenvalues are $\mu_{2k} = \mu_{2k-1} = \sigma^2(2\pi k)^{-2m}$ for $k \ge 1$ and $\mu_0 = \infty$. In this case, $\sup_{i\ge1}\|\phi\|_{\sup} < \infty$. For any $k \ge 1$,

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

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

$$\sup_{k\ge1}\frac{\sum_{i=k+1}^{\infty}\mu_i}{k\mu_k} = C < \infty.$$

Hence, Assumption A1 holds true. Verification of Assumption A1 on the eigenfunctions for Sobolev space kernel can be found in [31].

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

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

where $H_i(\cdot)$ is the i-th Hermite polynomial; see [34] for more details. Then $\sup_{i\geq 1}\|\phi_i\|_{\sup} < \infty$ 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 A1 holds.

B.8. Auxiliary lemmas

Lemma B.4 *(Hanson-Wright inequality [26]) Let $X = (X_1, \cdots, X_n) \in \mathbb{R}^n$ be a random vector with independent components X_i which satisfy $\mathrm{E}\, X_i = 0$ and $\|X_i\|_{\psi_2} \leq K$. Let A be an $n \times n$ matrix. Then, for every $t \geq 0$,*

$$P\Big(|X^\top AX - \mathrm{E}\, X^\top AX| > t\Big) \leq 2\exp\Big(-c\min\big(\frac{t^2}{K^4\|A\|_{HS}^2}, \frac{t}{K^2||A||}\big)\Big)$$

Here $\|A\|_{HS}$ is the Hilbert-Schmidt (or Frobenius) norm of A.

Lemma B.5 *(Eigenvalue interlacing theorem) Suppose $A \in \mathbb{R}^{n\times n}$ is symmetric. Let $B \in \mathbb{R}^{m\times m}$ with $m < n$ be a principal submatrix (obtained by deleting both $i-$th row and $i-$th column for some values of i). Suppose A has eigenvalues $\lambda_1 \leq \cdots \lambda_n$ and B has eigenvalues $\beta_1 \leq \cdots \leq \beta_m$. Then*

$$\lambda_k \leq \beta_k \leq \lambda_{k+n-m} \qquad \text{for } k = 1, \cdots, m.$$

And if $m = n-1$,

$$\lambda_1 \leq \beta_1 \leq \lambda_2 \leq \beta_2 \leq \cdots \leq \beta_{n-1} \leq \lambda_n.$$

Lemma B.6 *(Weyl's inequality) Let M, H and P are $n \times n$ Hermitian matrices with $M = H + P$, where M has eigenvalues $\mu_1 \geq \cdots \geq \mu_n$, and H has eigenvalues $\nu_1 \geq \cdots \geq \nu_n$, and P has eigenvalues $\rho_1 geq \cdots \rho_n$. Then the following inequalities hold for $i = 1, \cdots, n$:*

$$\nu_i + \rho_n \leq \mu_i \leq \nu_i + \rho_1$$

If P is positive definite, then this implies

$$\mu_i > \nu_i, \quad \forall i = 1, \cdots, n.$$

Appendix C. Simulation Study

In this section, we examine the performance of the proposed testing procedure through simulation studies in Sections C.1 and C.2.

C.1. Simulation Study I: PDK

Data were generated from the regression model (1.1) with $f(x) = c(3\beta_{30,17}(x) + 2\beta_{3,11}(x))$, where $\beta_{a,b}$ is the density function for Beta(a, b), $x_i \overset{iid}{\sim}$ Unif$[0, 1]$, $\epsilon_i \overset{iid}{\sim} N(0, 1)$ and c is a constant. To fit the model, we consider the periodic Sobolev kernel with eigenvalues $\mu_{2i} = \mu_{2i-1} = (2\pi i)^{-2m}$ for $i \geq 1$; see [15] for details. Set $n = 2^9, 2^{10}, 2^{11}, 2^{12}$, and $H_0 : f = 0$. The significance level was chosen as 0.05 and the Gaussian random projection matrix was applied in this setting.

We examined the empirical performance of the distance-based test $T_{n,\lambda}$ The projection dimension s was chosen as $2n^\gamma$ for $\gamma = 1/(4m+1), 2/(4m+1), 3/(4m+1)$, with $m = 2$ corresponding to cubic splines. The smoothing parameter λ was chosen by a projection version of Wahba's GCV score ([37]) based on $\widehat{f}_R$. Specifically, our new GCV score is defined as follows

$$V_S(\lambda) = \frac{\frac{1}{n}\boldsymbol{Y}^\top (I - A_S(\lambda))^2 \boldsymbol{Y}}{\left(\frac{1}{n}\operatorname{tr}(I - A_S(\lambda))\right)^2}, \ \lambda > 0,$$

where $A_S = \boldsymbol{K}S^\top (S\boldsymbol{K}^2 S^\top + \lambda S\boldsymbol{K}S^\top)^{-1} S\boldsymbol{K}$ is a projection version of the classical smoothing matrix. Our new GCV score enjoys much computational advantage than the classical one in our empirical study.

Empirical size was evaluated at $c = 0$, and power was evaluated at $c = 0.01, 0.02, 0.03$. Both size and power were calculated based on 500 independent replications. Figure 3 (a) shows that the size of our testing rule approach the nominal level 0.05 under various choices of (s, n), demonstrating the validity of the proposed testing procedure. Figure 3 $(b), (c), (d)$ displays the power of $T_{n,\lambda}$. Under various choices of c and γ, it is not surprising to see from Figure 3 (b), (c), and (d) that the power approaches one as n or c increases. Rather, a key observation is that the power cannot be further improved as γ grows beyond the critical point $2/(4m+1)$ when $c \geq 0.02$. This is consistent with our theoretical result; see Theorem 4.8.

C.2. Simulation Study II: EDK

In this section, we consider a multivariate case and test $H_0 : f = 0$. Data were generated from

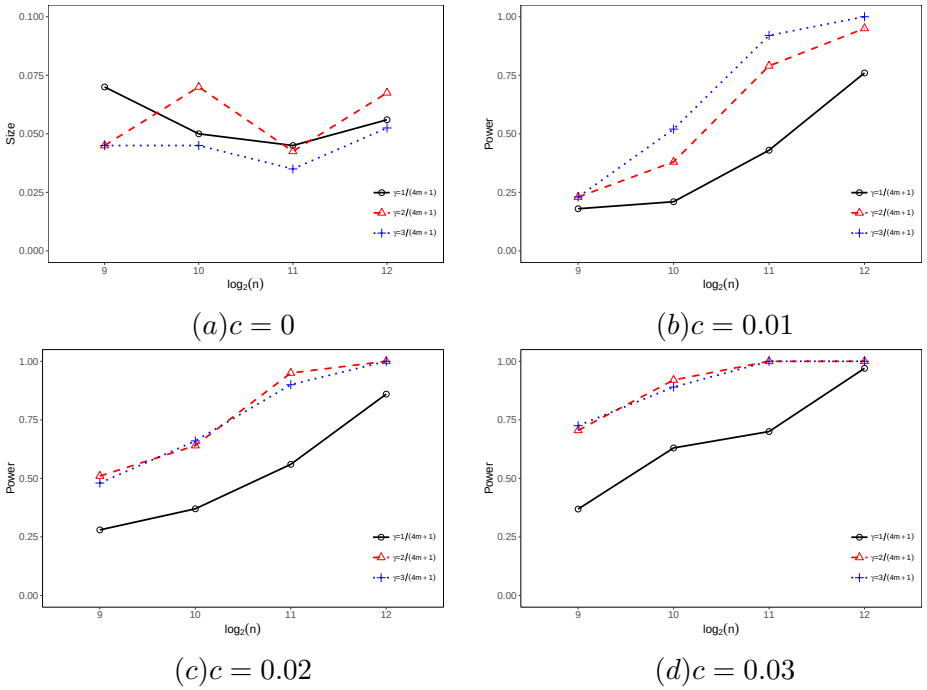
$$y_i = c(x_{i1}^2 + 2x_{i1}x_{i2} + 4x_{i1}x_{i2}x_{i3}) + \epsilon_i, \ i = 1, \cdots, n,$$

where (x_{i1}, x_{i2}, x_{i3}) follows from $N(\mu, I_3)$ with $\mu = (0, 0, 0)$, $\epsilon_i \sim N(0, 1)$, and $c \in \{0, 0.05, 0.1, 0.15\}$. Specifically, we chose the Gaussian kernel

$$K(\mathbf{x}, \mathbf{x}') = e^{-\frac{1}{2}\sum_{i=1}^3 (x_i - x_i')^2}.$$

We considered sample sizes $n = 2^9$ to $n = 2^{12}$ and sketch dimensions $s = 1.2\log(n), 1.2(\log n)^{3/2}$, $1.2(\log n)^2$. For each pair (n, s), experiments were independently repeated 500 times for calculating the size and power.

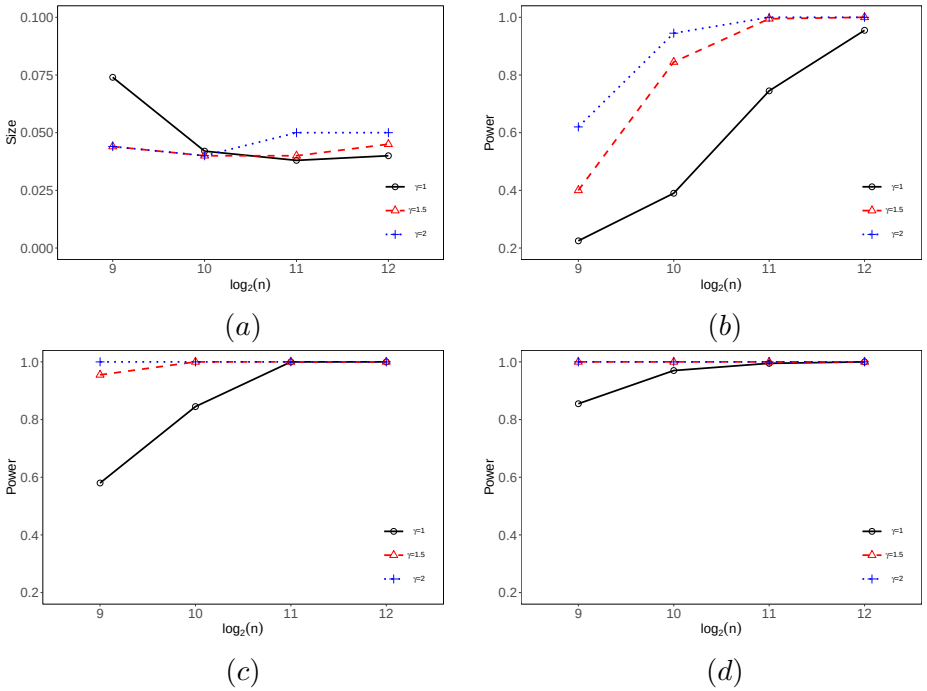
Interpretations for Figure 4 about the size and power are similar to those for Figures 3. Interestingly, we observe that the power increases dramatically as γ increases from 1 to 1.5, while becomes stable near one as $\gamma \geq 1.5$. This is consistent with Theorem 4.6. Figure 5 demonstrates the significant computational advantage of our test statistics (corresponding to $\gamma < 1$ for PDK, and $\gamma \leq 2$ for EDK) over the testing procedure based on standard KRR.



$(a) c = 0$ $(b) c = 0.01$

$(c) c = 0.02$ $(d) c = 0.03$

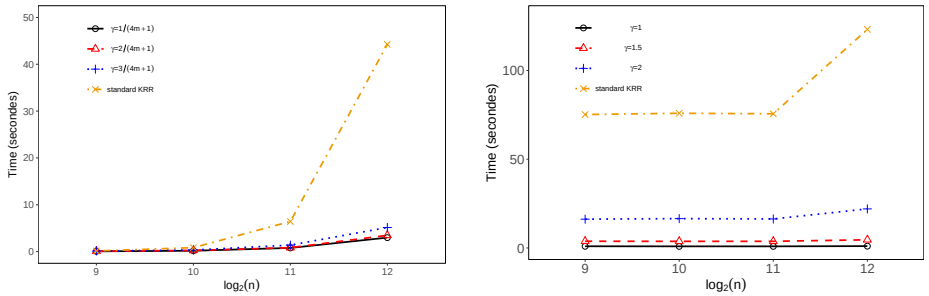
Figure 3: Size and power for $T_{n,\lambda}$ with projection dimension and signal strength varies.



(a) (b)

(c) (d)

Figure 4: Size and power for $T_{n,\lambda}$ with varying projection dimensions. Signal strength $c = 0$ for (a); $c = 0.05$ for (b); $c = 0.1$ for (c); $c = 0.15$ for (d).



(a) (b)

Figure 5: Computing time for our testing rule with varying projection dimensions: (a) is polynomially decay kernels; (b) is exponentially decay kernels.