Nonparametric Testing under Randomized Sketching

Meimei Liu, Zuofeng Shang, Yun Yang, and Guang Cheng

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. An adaptive testing procedure is further established without prior knowledge of regularity. One technical contribution is to establish upper bounds for a range of tail sums of empirical kernel eigenvalues. Simulations and real data analysis are conducted to support our theory.

Index Terms—Computational limit, kernel ridge regression, minimax optimality, nonparametric testing, random sketch.

1 Introduction

A Number of computationally efficient statistical methods have been proposed for analyzing massive data sets. Examples include divide-and-conquer approaches [1]–[4]; low-rank approximations: random projection methods [5]–[8], subsampling methods [9]–[11], Nyström approximations [12], [13]; and online learning methods [14]–[16].

An interesting question arising from these new methods is the minimum computational cost required for obtaining statistically satisfactory solutions. This might be viewed as a type of "computational limit" from a statistical perspective. Such an issue has been addressed in certain situations. For divide-and-conquer approaches, [4] derived a *sharp* upper bound for the number of distributed computing units in the smoothing spline setup, while [17] 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. For example, [18] 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. Recently, [7] proposed the randomly sketched kernel ridge regression (KRR) estimator and studied the minimax optimal nonparametric estimation under random projection. To our knowledge, the computational limit for random projection methods remains unknown in nonparametric models.

There are two purposes in this paper: (i) develop an optimal nonparametric testing procedure based on random projection; (ii) explore its computational limit in the kernel ridge regression setup. We remark that classical nonparametric testing methods, e.g., the locally most powerful test, the generalized/penalized likelihood ratio test and the distance-based test [19]–[23], may not be directly applied to big data due to their high computational costs.

Specifically, we consider the following nonparametric model

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

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 random noise following Normal distribution with mean zero, variance σ^2. The regression function f belongs to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$. The hypothesis of interest is

$$H_0 : f = f_0 \ \text{ v.s. } H_1 : f \in \mathcal{H} \setminus \{f_0\}, \tag{2}$$

where f_0 is a hypothesized function. Testing optimality has been well studied in literature. [24] [25] established the minimax testing rate for Gaussian sequence model on the Sobolev class or the Besov class. Recently, [26] established the minimax nonparametric testing rate under a general eigen-decaying framework including the polynomial decay and exponential decay kernels. In practice, testing in (2) has wide applications. One motivating example is in the signal detection in cognitive radio and other wireless applications, it is assumed under the null hypothesis that the signal is completely specified, e.g., that no signal is present. Another example is testing the adequacy of a parametric linear model in nonparametric regression, where the null hypothesis assumes f_0 has a linear structure; see [27], [28].

Focusing on the testing in (2), we construct a distance-based test statistic $T_{n,\lambda} = \|\widehat{f}_R - f_0\|_n^2$, where $\widehat{f}_R$ is a randomly sketched KRR estimator, and $\|\cdot\|_n$ is the empirical norm. The sketched KRR estimator $\widehat{f}_R$ ([7]) enjoys both theoretical support and computational efficiency, especially

- *M. Liu is with the Department of Statistics, Virginia Tech, Blacksburg, VA 24060. E-mail: meimeiliu@vt.edu.*
- *Z. Shang is with the Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, New Jersey 07102. E-mail: zshang@njit.edu.*
- *Y. Yang is with the Department of Statistics, University of Illinois Urbana-Champaign, Champaign, IL 61820. E-mail: yy84@illinois.edu.*
- *G. Cheng is with the Department of Statistics, Purdue University, West Lafayette, IN 47906. E-mail:chengg@purdue.edu. Guang Cheng is a member of Institute for Advanced Study, Princeton and visiting Fellow of SAMSI for the Deep Learning Program in the Fall of 2019; he would like to thank both Institutes for their hospitality.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3063223, IEEE Transactions on Pattern Analysis and Machine Intelligence

compared with the classic KRR estimator $\widehat{f}_n$ ([29]) defined as

$$\widehat{f}_n := \underset{f\in\mathcal{H}}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\|f\|_{\mathcal{H}}^2 \right\}, \qquad (3)$$

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. Solving (3) is an n-dimensional quadratic program, which involves the computational cost and storage occupation of orders $\mathcal{O}(n^3)$ and $\mathcal{O}(n^2)$, respectively. Instead, $\widehat{f}_R$ is achieved by randomly projecting the row and column subspaces of the n-dimensional kernel matrix to an s-dimensional subspace, reducing (3) to an s-dimensional quadratic program, with time and storage costs as $\mathcal{O}(s^3)$ and $\mathcal{O}(s^2)$ under $s(\ll n)$ random projections; see Section 2 for detailed algorithm. The pre-processing step in computing 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 [30]), which can be further reduced to $O(n^2(\log s)/t)$ by using t clusters in a parallel fashion. After $\widehat{f}_R$ is obtained, $T_{n,\lambda}$ can be computed in a parallel fashion. Hence, s can be viewed as a simple proxy for computing and storage costs.

In this paper, we reveal a phase transition phenomenon in terms of s to guarantee the testing optimality. Specifically, a sharp lower bound for s is established: when s is above this bound, $T_{n,\lambda}$ is minimax optimal; otherwise, minimax optimality becomes impossible even when the best possible λ is chosen. We next illustrate more subtle details using the following Figure 1, where the strength of the weakest detectable signals (SWDS) is characterized given any s and λ. In general, we require $s > s_\lambda$ for any λ, where s_λ is determined by kernel eigenvalues and λ. An important observation is that the smallest SWDS can be achieved at $\lambda = \lambda^*$ and $s > s_{\lambda^*} := s^*$ (note that when $s \ll s^*$, our testing procedure under a proper λ is still powerful as long as SWDS becomes sufficiently large). Both λ^* and s^* have precise orders in specific situations. For example, in an m-order polynomial decay kernel, the smallest SWDS achieves the minimax optimal rate $n^{-\frac{2m}{4m+1}}$ established in [24], [25] when $\lambda^* = n^{-\frac{4m}{4m+1}}$ and $s^* = n^{\frac{2}{4m+1}}$. As a by-product, we also derive a sharp lower bound for s for obtaining the minimax optimal estimation. Our results hold for a general class of random projection matrix, such as the sub-Gaussian matrix or certain data-dependent matrix.

It is worth mentioning that the construction of $T_{n,\lambda}$ crucially relies on the regularity of $\mathcal{H}$, which is often unavailable in practice. Hence, we propose an adaptive test statistic based on the maximum of a sequence of (standardized) non-adaptive test statistics corresponding to various regularities. Based on a recent Gaussian approximation result in [31], we prove that the null limit distribution is an extreme value distribution.

The proofs of main results rely on the behavior of the tail sum of empirical kernel eigenvalues. One technical contribution of this work is to derive upper bounds for a range of tail sums such that nonparametric estimation and testing can now be analyzed in a unified framework. This is obtained by flexibly adjusting the size of the function class associated with the Rademacher average in the local Rademacher complexity theory ([32]).

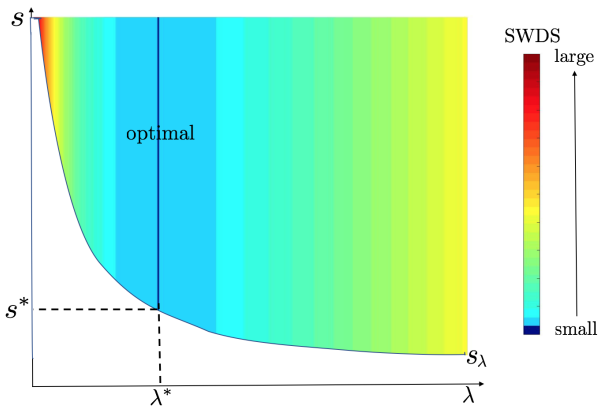


Fig. 1. Phase transition in (λ, s) for signal detection. The horizontal axis is the smoothing parameter λ, and the vertical axis is the projection dimension s. The shade indicates the values of SWDS: dark red corresponds to greater values of SWDS than light blue. The vertical line labeled by "optimal"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

In simulation studies, we find that the size and power of the proposed non-adaptive and adaptive test statistics are both satisfactory. In particular, the power cannot be further improved as the number of random projections grows beyond some threshold, as predicted by our theory. For an illustration purpose, we also demonstrate that when $n = 2^{12}$, conducting testing based on $\widehat{f}_R$ only takes 3.2 seconds in comparison with 42 seconds based on $\widehat{f}_n$. In practice, the smoothing parameter λ can be directly selected via generalized cross validation. We would like to point out that this is an advantage of the random projection method over the divide-and-conquer method [1] in estimation, where the selection of the smoothing parameter is nontrivial; see [33].

The rest of this paper is organized as follows. Section 2 introduces kernel ridge regression together with its approximation based on random projection. Our main results are presented in Section 3: Section 3.1 introduces one primary assumption on random projection; Sections 3.2 and 3.3 study testing consistency and power behaviors in terms of the projection dimension s and the smoothing parameter λ, with specific situations considered in Section 3.5; Section 3.6 proves the lower bound on s given in Section 3.5 to be sharp. An adaptive testing procedure is developed in Section 4. Section 5 includes numerical studies based on simulated and real data sets. All technical details are deferred to the Appendix.

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. $Pf^2 \equiv Ef(X)^2$, $\|f\|_n^2 \equiv P_n f^2 \equiv \frac{1}{n}\sum_{i=1}^n f(X_i)^2$. For a matrix $A \in \mathbb{R}^{m\times n}$, its operator norm is defined as $\|A\|_{\text{op}} = \max_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Ax\|_2}{\|x\|_2}$. 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 $\|X\|_{\psi_2} = \inf\{t > 0 : E\exp(X^2/t^2) \leq 2\}$. We will use c, c_1, c_2, C to denote

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3063223, IEEE Transactions on Pattern Analysis and Machine Intelligence

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), where the covariances x_i are samples of X from a distribution P_X with domain $\mathcal{X}$. Let $\mathcal{X}$ be a closed subset of $\mathbb{R}^a$, a is fixed,P_X a strictly positive Borel measure on $\mathcal{X}$. We recall that a Borel measure P_X on X is said to be strictly positive if the measure of every nonempty open subset in X is positive, an example being the Lebesgue measure in $\mathbb{R}^a$.

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 $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. K is a symmetric positive definite kernel on X satisfying: for any finite set of points $\{x_i\}_{i=1}^n$ in X and real numbers $\{a_i\}_{i=1}^n$ that $\sum_{i,j=1}^n a_i a_j K(x_i, x_j) \geq 0$. Assume further that K is a continuous function on $\mathcal{X} \times \mathcal{X}$ and $\int_{\mathcal{X}} \int_{\mathcal{X}} K(x,t) dP_X(x) dP_X(t) < \infty$. Then by Mercer's theorem, K has the following spectral expansion:

$$K(x, x') = \sum_{i=1}^{\infty} \mu_i \phi_i(x) \phi_i(x'), \quad x, x' \in \mathcal{X}, \tag{4}$$

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 $L^2(P_X)$. We refer the reader to the standard sources [34], [35] for more details on RKHSs and their properties.

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., [36], 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. Examples of kernels in this class include the m^{th} order Sobolev spaces for some fixed integer $m \geq 1$ with Lebesgue measure on a bounded domain; see [35].

The second is $\mu_i \asymp \exp(-\gamma i^p)$ for constants $\gamma, p > 0$, called as exponential decay kernel (EDK) of order p. Examples of EDK include the Gaussian kernel, which for the Lebesgue measure satisfies such a bound with $p = 1$ (compact domain) or $p = 2$ (real line); see [37]. Verification of Assumption A1 with concrete examples is deferred to Section S.5.3 in Supplementary.

Recall the KRR estimator $\widehat{f}_n$ from (3). 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

$$\begin{aligned} \widehat{\omega} &= \operatorname*{argmin}_{\omega \in \mathbb{R}^n} \Big\{ \omega^\top \mathbf{K}^2 \omega - \frac{2}{n} \omega^\top \mathbf{K} \mathbf{y} + \lambda \omega^\top \mathbf{K} \omega \Big\} \\ &= \frac{1}{n} (\mathbf{K} + \lambda I)^{-1} \mathbf{y}, \end{aligned} \tag{5}$$

$\mathbf{y} = (y_1, \cdots, y_n)^\top$, $\mathbf{K} = [n^{-1} K(x_i, x_j)]_{1 \leq i,j \leq n}$, and $I \in \mathbb{R}^{n \times n}$ is identity. This standard procedure requires storing $(\mathbf{K}^2, \mathbf{K}, \mathbf{K}\mathbf{y})$ and inverting $\mathbf{K} + \lambda I$, 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 [7]. Specifically, ω in (5) 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 3.1. Then, β is solved as:

$$\begin{aligned} \widehat{\beta} =& \operatorname*{argmin}_{\beta \in \mathbb{R}^s} \Big\{ \beta^\top (S\mathbf{K})(\mathbf{K} S^\top) \beta - \frac{2}{n} \beta^\top S \mathbf{K} \mathbf{y} + \lambda \beta^\top S \mathbf{K} S^\top \beta \Big\} \\ =& \frac{1}{n} (S\mathbf{K}^2 S^\top + \lambda S \mathbf{K} S)^{-1} S \mathbf{K} \mathbf{y}. \end{aligned} \tag{6}$$

Hence, the resulting estimator of f becomes

$$\widehat{f}_R(\cdot) = \sum_{i=1}^n (S^\top \widehat{\beta})_i K(\cdot, x_i), \tag{7}$$

which requires computing and storing $(S\mathbf{K}^2 S^\top, S\mathbf{K}S^\top, S\mathbf{K}\mathbf{y})$, 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 [30]), 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 3.6.

3 Main Results

Consider the nonparametric testing problem (2). For convenience, assume $f_0 = 0$, i.e., we will test

$$H_0 : f = 0 \quad \text{vs.} \quad H_1 : f \in \mathcal{H} \setminus \{0\}. \tag{8}$$

In general, testing $f = f_0$ (for an arbitrary known f_0) is equivalent to testing $f_* \equiv f - f_0 = 0$. So, (8) has no loss of generality. Based on $\widehat{f}_R$, we propose the following distance-based test statistic:

$$T_{n,\lambda} = \|\widehat{f}_R\|_n^2. \tag{9}$$

In the subsequent sections, we will derive the null limit distribution of $T_{n,\lambda}$ (Theorems 3.3), and further provide a sufficient and necessary condition in terms of s such that $T_{n,\lambda}$ is minimax optimal (Section 3.6). 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 which will be discussed in Section 3.1.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3063223, IEEE Transactions on Pattern Analysis and Machine Intelligence

3.1 Choice of Projection Matrix

Consider the singular value decomposition $\mathbf{K} = UDU^\top$, where $U = (U_1, U_2)$ 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, $D = \text{diag}(D_1, D_2)$ with $D_1 = \text{diag}(\widehat{\mu}_1, \ldots, \widehat{\mu}_{\widehat{s}_\lambda})$ and $D_2 = \text{diag}(\widehat{\mu}_{\widehat{s}_\lambda+1}, \ldots, \widehat{\mu}_n)$. Here $\{\widehat{\mu}_i\}_{i=1}^n$ are empirical eigenvalues in decreasing order. For any $\lambda > 0$, $\widehat{s}_\lambda$ (or s_λ) is defined to be the number of $\widehat{\mu}_i$'s (or μ_i's) greater than λ, i.e.,

$$\widehat{s}_\lambda = \text{argmin}\{i : \widehat{\mu}_i \le \lambda\} - 1, \quad s_\lambda = \text{argmin}\{i : \mu_i \le \lambda\} - 1. \tag{10}$$

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 3.5. In the Supplementary S.5.3, we verify Assumption A2 with two concrete examples:

An accurate upper bound for the tail sum of empirical eigenvalues $\sum_{i=\widehat{s}_{\lambda+1}}^n \widehat{\mu}_i$ is needed for studying nonparametric testing and estimation. However, this bound was often *assumed* to hold in the kernel learning literature, e.g., [38], [39]. The application of concentration inequalities of individual eigenvalues ([40], [41]) only provides a very loose bound due to accumulative errors. Recently, the local Rademacher complexity theory ([32]) was employed by [7] 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. We establish the upper bounds, i.e., Lemma 3.1, for a range of tail sums of empirical eigenvalues in terms of population quantities s_λ and μ_{s_λ}, with known orders. This result can be applied to both nonparametric estimation and testing, and may be of independent interest.

Lemma 3.1. If $1/n < \lambda \to 0$, then with probability at least $1 - 4e^{-s_\lambda}$, $\sum_{i=\widehat{s}_{\lambda+1}}^n \widehat{\mu}_i \lesssim s_\lambda \mu_{s_\lambda}$.

Clearly, Lemma 3.1 is a sample analog to the tail sum assumption for μ_i in Assumption A1. The proof of Lemma 3.1 is based on the development of a generalized function class and its associated local Rademacher complexity theory as explained in Appendix 7.1.

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

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

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

By Definition 1, a $\mathbf{K}$-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 [7] 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 which will make the statement of our assumptions more precise and concise.

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

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 $\text{P}(\mathcal{E}) \ge (1 - \exp(-a))(1 - \exp(-b))$. 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) $s \ge qs_\lambda$ for a sufficiently large constant $q > 0$.

(b) There exist $c_1, c_2 \in (0, \infty]$ such that the event "S is $\mathbf{K}$-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 $\mathbf{K}$-satisfiable with high probability which holds in a broad range of situations such as matrix of sub-Gaussian entries (Example 3) and certain data dependent matrix (Example 4).

Example 3. 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 [42]), or mixtures of sub-Gaussian variables. The following lemma shows that Assumption A3 (b) holds in all these situations.

Lemma 3.2. Let $S_{ij} : 1 \le i \le s$, $1 \le j \le n$ be independent sub-Gaussian of mean zero and variance one, and $\lambda \in (1/n, 1)$. If $s \ge qs_\lambda$ for a sufficiently large constant q, then Assumption A3 (b) holds for $S = [S_{ij}/\sqrt{s}]_{1\le i\le s, 1\le j\le n}$.

Example 4. 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 $\|SU_2 D_2^{1/2}\|_{\text{op}} = 0$, i.e., Assumption A3 (b) holds.

The proof of Lemma 3.2 relies on the bound of the tail sums of empirical eigenvalues in Lemma 3.1. We point out that obtaining the eigen-decomposition in Example 4 is as burdensome as computing the matrix inverse, which is not preferred in practice. Rather, the purpose of this example is to directly illustrate one situation that Assumption A3 can be satisfied.

3.2 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 3.3. 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, $\mu_{n,\lambda} := E_{H_0}\{T_{n,\lambda}|\mathbf{x}, S\} = tr(\Delta^2)/n$, $\sigma^2_{n,\lambda} := \mathrm{Var}_{H_0}\{T_{n,\lambda}|\mathbf{x}, S\} = 2tr(\Delta^4)/n^2$ with $\mathbf{x} = (x_1, \cdots, x_n)$ and $\Delta = \mathbf{K}S^\top(S\mathbf{K}^2S^\top + \lambda S\mathbf{K}S^\top)^{-1}S\mathbf{K}$.

Theorem 3.3 holds once s diverges (no matter how slowly). Theorem 3.3 implies the following testing rule at significance level α:

$$\phi_{n,\lambda} = I(|T_{n,\lambda} - \mu_{n,\lambda}| \geq z_{1-\alpha/2}\sigma_{n,\lambda}) \tag{11}$$

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

As an important consequence of Theorem 3.3, we comment that the optimal estimation rate in [7] can also be obtained as a by-product, that is

$$\|\widehat{f}_R - f_0\|_n^2 = O_P(r^2_{n,\lambda}), \tag{12}$$

where $r^2_{n,\lambda} = \lambda + \mu_{n,\lambda}$. The proof of (12) is sketched as follows. Suppose that $f_0 \in \mathcal{H}$ is the "true" function in (1). Note that $\|\widehat{f}_R - f_0\|_n^2$ has a trivial upper bound

$$\|\widehat{f}_R - f_0\|_n^2 \leq 2\|\widehat{f}_R - E_\epsilon\widehat{f}_R\|_n^2 + 2\|E_\epsilon\widehat{f}_R - f_0\|_n^2, \tag{13}$$

where E_ϵ is the expectation w.r.t. $\epsilon = (\epsilon_1, \cdots, \epsilon_n)$ with ϵ_i the random noise in the regression function defined in (1). By direct examinations, it can be shown that $\|\widehat{f}_R - E_\epsilon\widehat{f}_R\|_n^2 = \epsilon^\top\Delta^2\epsilon/n$, hence, $E_\epsilon\|\widehat{f}_R - E_\epsilon\widehat{f}_R\|_n^2 = tr(\Delta^2)/n = \mu_{n,\lambda}$. This leads to $\|\widehat{f}_R - E_\epsilon\widehat{f}_R\|_n^2 = O_P(\mu_{n,\lambda})$. Meanwhile, it follows from Lemma S.1 below that $\|E_\epsilon\widehat{f}_R - f_0\|_n^2 = O_P(\lambda)$. This completes the proof of (12). The above discussions are summarized in the following corollary.

Corollary 3.4. Suppose that $1/n < \lambda < 1$ and Assumption A1-A3 holds. Then with probability approaching one, it holds that

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

where $r^2_{n,\lambda} = \lambda + \mu_{n,\lambda}$ and C is an absolute constant.

From Corollary 3.4, 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 (10), i.e., $s^\dagger = s_{\lambda^\dagger}$. In Section 3.5, we will show that the upper bound under $\lambda^\dagger$ is minimax optimal, and further provide explicit orders for $s^\dagger$ in concrete settings.

3.3 Power Analysis

In this section, we investigate the power of $T_{n,\lambda}$ under a sequence of local alternatives by assuming $f \in \mathcal{B} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq C\}$ for some existing constant C.

For any generic 0-1 valued testing rule ϕ where $\phi = 0$ if H_0 is preferred and 1 otherwise, define the total error $\mathrm{Err}(\phi, d_n)$ of ϕ under a separation rate $d_n > 0$ as

$$\mathrm{Err}(\phi, d_n) = E(\phi \mid H_0 \text{ is true}) + \sup_{\|f\|_n \geq d_n} E(1-\phi \mid H_1 \text{ is true}). \tag{14}$$

Notice $E(\phi \mid H_0 \text{ is true})$ is the probability of making a type I error and $E(1-\phi \mid H_1 \text{ is true})$ is the probability of making a type II error, and the total error represents the maximum possible type I error and type II error. The separation rate d_n is used to measure the distance between the null and the alternative hypotheses. Intuitively, the smaller d_n is, the harder it is to distinguish the alternative hypothesis from the null. For any $\varepsilon \in (0,1)$, define the minimax separation rate $d_n^*(\varepsilon)$ as

$$d_n^*(\varepsilon) = \inf\{d_n > 0 : \inf_\phi \mathrm{Err}(\phi, d_n) \leq \varepsilon\}, \tag{15}$$

where the infimum in (15) is taken over all 0-1 valued testing rules based on samples $((x_1, y_1), \ldots, (x_n, y_n))$. $d_n^*(\varepsilon)$ characterizes the smallest separation between the null and local alternatives such that there exists a testing approach with a total error of at most ε. [24] [25] established the minimax separation rate and revealed its difference with optimal estimation rate. Their work are derived based on Gaussian sequence model with focus on the Sobolev class or the Besov class with polynomial decaying eigenvalues. Recently, [26] established the minimax nonparametric testing rate under a general eigen-decaying framework including the polynomial decay and exponential decay kernels. Next, we show that our proposed Wald-type test can achieve the minimax separation rate under appropriate λ and s.

For any $f \in \mathcal{H}$, define the squared separation rate $d^2_{n,\lambda}$ as

$$d^2_{n,\lambda} = \underbrace{\lambda}_{\text{Bias of } \widehat{f}_R} + \underbrace{\sigma_{n,\lambda}}_{\text{Standard deviation of } T_{n,\lambda}}. \tag{16}$$

In the following Theorem 3.5, we show 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 $d_{n,\lambda}$. It is sufficient to minimize the separation rate $d_{n,\lambda}$ to achieve optimal testing. We show that our test can achieve the minimax rate of testing by selecting λ to balance the trade-off between the bias of $\widehat{f}_R$ and the standard derivation of $T_{n,\lambda}$ shown in (16).

Theorem 3.5. 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 $1 - e^{-c_1 s} - e^{-c_2 s_\lambda}$,

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

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

In view of Theorem 3.5, 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 (10), 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.

3.4 Parametric versus nonparametric fits

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

$$H_0^{\text{linear}} : f \in \mathcal{L}(\mathcal{X}) \text{ vs. } H_1^{\text{linear}} : f \notin \mathcal{L}(\mathcal{X}), \tag{17}$$

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. Intuitively, if the parametric structure

is right, then its residuals should be patternless and independent of input features. So we can apply non-parametric smoothing to the parametric residuals and see if their pattern is approximately zero everywhere; see Section 4.2 in [25].

Denote $\widehat{f}_n^*$ as the least square estimator of f, $\widehat{f}_R$ as the randomly projected KRR estimator, and f_0 is a hypothesized "true" parameter with unknown value. Write

$$\begin{aligned} T_{n,\lambda}^* =& \|\widehat{f}_n^* - \widehat{f}_R\|_n^2 = \|\widehat{f}_n^* - f_0 + f_0 - \widehat{f}_R\|_n^2 \\ =& \|\widehat{f}_n^* - f_0\|_n^2 + \|f_0 - \widehat{f}_R\|_n^2 + \frac{2}{n}\sum_{i=1}^{n}(\widehat{f}_n^* - f_0)(f_0 - \widehat{f}_R) \\ =& T_{n,\lambda}^{(1*)} + T_{n,\lambda}^{(2*)} + T_{n,\lambda}^{(3*)} \end{aligned} \tag{18}$$

It can be shown that $T_{n,\lambda}^{(1*)} = O_P(n^{-1})$ by conventional parametric theory ([43]), and accordingly $T_{n,\lambda}^{(3*)} = o(T_{n,\lambda}^{(2*)})$ by Cauchy-Schwarz inequality. The dominate term in (18) is $T_{n,\lambda}^{(2*)}$. It turns out that $T_{n,\lambda}^{(2*)}$ is exactly the test for testing H_0^{linear}. Therefore, we have that $T_{n,\lambda}^*$ has the same limiting distribution with $T_{n,\lambda}^{(2*)}$ under H_0^{linear}. Applying Theorem 3.3, we have the following Corollary for the null asymptotic distribution of $T_{n,\lambda}^*$.

Corollary 3.6. Suppose that $\lambda \to 0$ and $s \to \infty$ as $n \to \infty$. Suppose Assumption A2 is satisfied. Then under H_0^{linear}, we have

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

Here, $\mu_{n,\lambda}^* := \mathrm{E}_{H_0}\{T_{n,\lambda}^* | x, S\} = \mathrm{tr}(\Delta^2)/n$, $\sigma_{n,\lambda}^{*2} := \mathrm{Var}_{H_0}\{T_{n,\lambda}^* | x, S\} = 2\mathrm{tr}(\Delta^4)/n^2$

To characterize the power of the composite hypothesis testing problem in (17), we define the test error under a separation rate $d_n > 0$,

$$\begin{aligned} \mathrm{Err}(\phi^*, d_n) =& E(\phi^* \mid H_0^{linear} \text{ is true}) \\ &+ \sup_{\|f - P_{\mathcal{L}(\mathcal{X})}f\|_n \geq d_n} E(1 - \phi^* \mid H_1 \text{ is true}). \end{aligned}$$

where ϕ^* is any desion rule for hypothesis testing problem in (17) and $P_{\mathcal{L}}(\mathcal{X})f$ is the projection of f on $\mathcal{L}(\mathcal{X})$. Notice that $f - P_{\mathcal{L}}(\mathcal{X})f = 0$ under the null hypothesis; the magnitude of $||f - P_{\mathcal{L}}(\mathcal{X})f||_n^2$ characterize how far the f is deviated from a linear function. Since the plugin estimate $\widehat{f}_n^*$ approaches $P_{\mathcal{L}(\mathcal{X})}f$ with $1/n$ rate, the separation rate for the decision rule for $\phi_{n,\lambda}^* = I(|T_{n,\lambda}^* - \mu_{n,\lambda}| \geq z_{1-\alpha/2}\sigma_{n,\lambda})$ is the same as $\phi_{n,\lambda}$ given in Theorem 3.5.

3.5 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 3.4 and Theorem 3.5. It is easy to check that Assumption A1 and A2 hold for these two examples; see Section S.5.3 in Supplementary.

Theorem 3.7. For the two kinds of eigenvalue decaying rates, suppose Assumption A3 holds. Suppose that $1/n < \lambda \to 0$ as $n \to \infty$, then with probability approaches 1, it holds that $\mu_{n,\lambda} \asymp s_\lambda/n$ and $\sigma_{n,\lambda}^2 \asymp s_\lambda/n^2$.

Furthermore, 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 $s \gtrsim (\log n)^{1/p}$ [1], $T_{n,\lambda}$ achieves the minimax optimal rate of testing $n^{-\frac{1}{2}}(\log n)^{\frac{1}{4p}}$.

The derivation of optimal estimation and testing rates is attributed to the accurate characterization of $\mu_{n,\lambda}$ and $\sigma_{n,\lambda}$ by employing Lemma 3.1 to bound the tail sums of empirical eigenvalues.

Plugging in $\mu_{n,\lambda} \asymp s_\lambda/n$ to Corollary 3.4, we have $\widehat{f}_R$ with the convergence rate $r_{n,\lambda}^2 = \lambda + s_\lambda/n$. As shown in (13), λ is the squared bias of $\widehat{f}_R$, and s_λ/n quantifies the variance of $\widehat{f}_R$. Hence, the optimal estimation rate $r_{n,\lambda}^{\dagger 2}$ is achieved via the bias-variance tradeoff as follows

$$r_{n,\lambda}^{\dagger 2} = \mathrm{argmin}\Big\{\lambda : \lambda > s_\lambda/n\Big\}.$$

We denote the choice of λ and lower bound of s to achieve the optimal estimation rate as $\lambda^\dagger$ and $s^\dagger$ respectively.

To find the lower bound for s in achieving optimal testing, by Theorem 3.5, $d_{n,\lambda}^2 \asymp \lambda + \sqrt{s_\lambda}/n$, then the optimal separation rate d_n^* can be achieved by another type of trade-off, i.e., the squared bias of $\widehat{f}_R$ v.s. the standard derivation of $T_{n,\lambda}$, as follows

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

We use λ^* and s^* to represent the optimal λ and s_λ to achieve $d_n^{*\,2}$.

Take PDK as an example, plugging in $\mu_i \asymp i^{-2m}$ to the definition of s_λ in (10), we have $s_\lambda \asymp \lambda^{-\frac{1}{2m}}$. Then Theorem 3.7 can be directly achieved based on the above two types of tradeoffs in estimation and testing. The results for EDK can be achieved similarly. We conclude our findings of this section in the following Table 1.

It is worth emphasizing that $\lambda^\dagger, s^\dagger$ are different from λ^*, s^* due to different types of trade-off discussed above, indicating a fundamental difference between estimation and testing ([24], [25]). Figure 2 summarizes the two different types of trade-off to achieve the minimax rate in estimation and testing.

In the following Theorem 3.8, we show that the upper bound between $\widehat{f}_R$ and $\widehat{f}_n$ can fall below the statistical error $\|\widehat{f}_n - f_0\|_n$ by further increasing the projection dimension.

1. In fact, for EDK, $s^* \asymp (\log n - \frac{1}{2p}\log\log n)^{1/p}$. For simplicity, we keep the main term $s^* \asymp (\log n)^{1/p}$.

Estimation			
	$\lambda^\dagger$	$s^\dagger$	$r^{\dagger 2}_{n,\lambda}$
PDK	$n^{-\frac{2m}{2m+1}}$	$n^{\frac{1}{2m+1}}$	$n^{-\frac{2m}{2m+1}}$
EDK	$(\log n)^{1/p} n^{-1}$	$(\log n)^{1/p}$	$(\log n)^{1/p} n^{-1}$
Testing			
	λ^*	s^*	d^{*2}_n
PDK	$n^{-\frac{4m}{4m+1}}$	$n^{\frac{2}{4m+1}}$	$n^{-\frac{4m}{4m+1}}$
EDK	$(\log n)^{1/2p} n^{-1}$	$(\log n)^{1/p}$	$(\log n)^{1/(2p)} n^{-1}$

TABLE 1
Lower bound of s and choice of λ for optimal estimation or testing in PDK and EDK

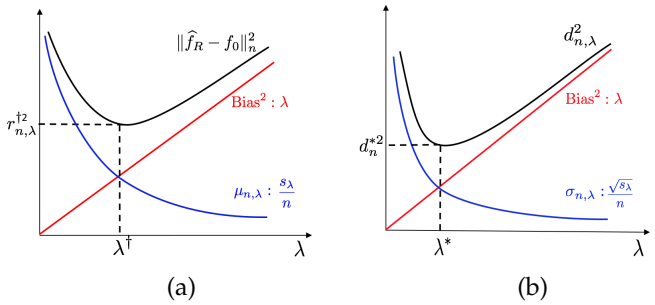


Fig. 2. Trade-offs for achieving (a) optimal estimation rate; (b) optimal testing rate.

This result improves upon existing approximation error bound in [7] that is of the same order as the statistic error. When $s > s^\dagger$, the difference between $\widehat{f}_R$ and $\widehat{f}_n$ is ignorable compared with the difference between $\widehat{f}_n$ and f_0.

Theorem 3.8. Let $\delta^\sharp > 0$ satisfying $\delta^\sharp \le \lambda$. Define $s^\sharp = \operatorname{argmin}\{j : \widehat{\mu}_j < \delta^\sharp\}$. Suppose $s \ge cs^\sharp$, then with probability approaching 1,

$$\|\widehat{f}_R - \widehat{f}_n\|_n^2 \le \delta^\sharp + \frac{s^\sharp \delta^\sharp}{n\lambda} \le \lambda + \frac{s_\lambda}{n}. \tag{19}$$

Furthermore, when $\delta^\sharp \ll \lambda$ and $\lambda \to 0$ as $n \to \infty$,

$$\|\widehat{f}_R - \widehat{f}_n\|_n^2 = o_P(\lambda + \frac{s_\lambda}{n}).$$

3.6 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 3.9 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. Our proof is constructive in the sense we construct the above true function as $\sum_{i=1}^n K(x_i, \cdot) w_i$ with w_i being selected from the orthogonal complement of a subsapce properly generated by S and K; see Appendix 7.2 for details.

Theorem 3.9. Suppose $s = o(s^\dagger)$. Then for any $s \times n$ random matrix S satisfying Assumption A3, with probability greater than $1 - e^{-cn\delta_n}$, it holds that

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

where c is a constant independent of n.

Our second result is about the sharpness of s^*. Theorem 3.10 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 α. The proof of Theorem 3.10 is similar as that of Theorem 3.9, except that a different true function is constructed; see Appendix 7.3 for details.

Theorem 3.10. Suppose $s = o(s^*)$. Then for any $s \times n$ 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 $1 - e^{-cn\delta_n}$,

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

where c is a constant independent of n. Recall $1 - \alpha$ is the significance level.

In view of Theorems 3.5 and 3.10, we observe a subtle phase transition phenomenon for testing signals as shown in Figure 1.

4 Adaptive Testing

In this section, we focus on the case of PDK as a leading example, and construct an adaptive testing procedure that does not require any exact prior knowledge on m except for $m \ge 2$. The adaptive procedure is proven to achieve the minimax rate of testing established by [44] (up to an iterative-logarithmic term).

Consider an RKHS generated by a PDK of order $m_* \ge 2$, i.e., $\mathcal{H} = \mathcal{H}_{m_*}$. To reflect the role of m, we modify all previous notation by adding a subscript m. For example, let $K_m(\cdot,\cdot)$ be the reproducing kernel function associated with $\mathcal{H}_m$, and $\mathbf{K}_m = \frac{1}{n}[K_m(x_i, x_j)]_{1\le i,j\le n}$ be the corresponding empirical kernel matrix. Let S_m be an $s_m \times n$ projection matrix. We will construct the corresponding $\widehat{f}_{R,m}(\cdot)$ based on (7) under S_m and λ_m. Here

$$\lambda_m = cn^{-4m/(4m+1)} (\log\log n)^{2m/(4m+1)},$$

and the corresponding projection dimension s_m is an integer satisfying

$$s_m \ge qn^{2/(4m+1)} (\log\log n)^{-1/(4m+1)}, \tag{20}$$

where $q > 0$ is a sufficiently large constant.

Given each m, the sketched KRR estimator $\widehat{f}_{R,m} = \Delta_m \mathbf{y}$, where $\Delta_m = \mathbf{K}_m S_m^\top (S_m \mathbf{K}_m^2 S_m^\top + \lambda_m S_m \mathbf{K}_m S_m^\top)^{-1} S_m \mathbf{K}_m$. The test statistic is defined as

$$T_{n,m} \equiv \|\widehat{f}_{R,m}\|_n^2 = \frac{1}{n} \mathbf{y}^\top \Delta_m^2 \mathbf{y}. \tag{21}$$

Denote $m_n \asymp (\log n)^{d_0}$ for a constant $d_0 \in (0, 1/2)$. Based on $T_{n,m}$, our adaptive testing procedure is constructed as follows.

Step 1. For any $2 \le m \le m_n \to \infty$, standardize $T_{n,m}$ as

$$\tau_m = \frac{nT_{n,m} - tr(\Delta_m^2)}{\sqrt{2tr(\Delta_m^4)}}.$$

Step 2. Calculate $\tau_n^* = \max_{1\le m\le m_n} \tau_m$.

Step 3. Find $\tau_{n,m_n} = B_n(\tau_n^* - B_n)$, where B_n[2] satisfies

$$2\pi B_n^2 \exp(B_n^2) = m_n^2. \tag{22}$$

By allowing $m_n \to \infty$, the unknown m_* will be eventually covered over a sequence of test statistics. Under the null hypothesis (8), $T_{n,m} = \frac{1}{n}\epsilon^\top \Delta_m^2 \epsilon$, and thus τ_m is of a standardized quadratic form. Then, τ_n^* is the maxima of a sequence of *dependent* τ_m's. Based on a recent Gaussian approximation result in [31], i.e., Lemma S.4 (stated in Supplementary), we prove in the following Theorem 4.1 that the null limit distribution of τ_{n,m_n} is some extreme value distribution.

Theorem 4.1. Suppose that $m_n \asymp (\log n)^{d_0}$ for a constant $d_0 \in (0, 1/2)$, and, for $2 \le m \le m_n$, S_m satisfies Assumption A3 (b) with projection dimension s_m. Then, under H_0 in (8), for any $\alpha \in (0,1)$, it holds that

$$\mathrm{P}(\tau_{n,m_n} \le c_\alpha) \to 1 - \alpha, \quad \text{as } n \to \infty,$$

where $c_\alpha = -\log(-\log(1-\alpha))$.

Our next result states that the above adaptive testing procedure is asymptotically minimax optimal. Specifically, Theorem 4.2 shows that τ_{n,m_n} achieves high power if the local alternative is separated from zero by an order $\delta(n, m_*)$ defined as

$$\delta(n, m_*) \equiv n^{-2m_*/(4m_*+1)} (\log\log n)^{m_*/(4m_*+1)}. \tag{23}$$

And, [44] showed that $\delta(n, m_*)$ is minimax optimal rate for adaptive testing.

Theorem 4.2. Suppose that $m_n \asymp (\log n)^{d_0}$ for a constant $d_0 \in (0, 1/2)$, and S_m satisfies Assumption A3 (b) with projection dimension s_m. Then, for any $\varepsilon > 0$, there exist positive constants $C_\varepsilon, N_\varepsilon$ for any $n \ge N_\varepsilon$, with probability approaching 1,

$$\inf_{\substack{f \in \mathcal{B}_{n,m_*} \\ \|f\|_n \ge C_\varepsilon \delta(n,m_*)}} \mathrm{P}_f(\tau_{n,m_n} \ge c_\alpha | \mathbf{x}, S) \ge 1 - \varepsilon,$$

where $\mathcal{B}_{n,m_*} = \{f \in \mathcal{H}_{m_*} : (\boldsymbol{f})^\top \mathbf{K}_{m^*}^{-1} \boldsymbol{f} \le 1\}$ and $\boldsymbol{f} = (f(x_1), \ldots, f(x_n))^\top$.

In the end, we point out that the lower bound for s_m given in (20) is slightly smaller than the sharp lower bound for s derived in the non-adaptive case; see Table 1. This is not surprising since the corresponding minimax rate $\delta(n, m_*)$, i.e., (23), is larger than the non-adaptive rate, i.e., $n^{-2m_*/(4m_*+1)}$.

5 Numerical Study

In this section, we examine the performance of the proposed testing procedure through simulation studies in Sections 5.1 and 5.2.

2. According to [45], B_n satisfying (22) has an approximation

$$B_n = \sqrt{2\log m_n} - \frac{1}{2}(\log\log m_n + \log 4\pi)/\sqrt{2\log m_n} + O(1/\log m_n) \asymp \sqrt{2\log m_n}.$$

5.1 Simulation Study I: PDK

Data were generated from the regression model (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 $i \ge 1$; see [46] 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.

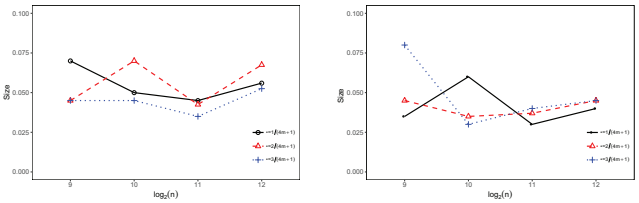
We examined the empirical performance of the distance-based test (DT) $T_{n,\lambda}$, and adaptive test (AT) τ_{n,m_n}. For DT, the projection dimension s was chosen as $2n^\gamma$ for $\gamma = 1/(4m+1), 2/(4m+1), 3/(4m+1)$, with $m = 2$ corresponding to cubic splines. For AT, the projection dimensions s_m was chosen as $2n^\gamma (\log\log n)^{-\frac{1}{4m+1}}$ for $m = 2, \cdots, \sqrt{\log n}$.

We propose a data-adaptive smoothing parameter slection rule to guarantee testing optimality. Specifically, we choose the optimal smoothing parameter λ^* satisfying

$$\lambda^* = \min\{\lambda \mid \lambda > \sigma_{n,\lambda}\},$$

where $\sigma_{n,\lambda}$ is defined in Theorem 3.3 as $\sigma_{n,\lambda} = 2trace(\Delta^4)/n^2$ with $\Delta = \mathbf{K}S^\top(S\mathbf{K}^2S^\top + \lambda S\mathbf{K}S^\top)^{-1}S\mathbf{K}$ as a projection version of the classical smoothing matrix.

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 shows that the size of both DT and AT approach the nominal level 0.05 under various choices of (s, n), demonstrating the validity of the proposed testing procedure. Figure 4 displays the power of DT and AT. Under various choices of c and γ, it is not surprising to see from Figure 4 (a), (c), and (e) that the power of DT 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 $c \ge 0.02$. This is consistent with our theoretical result; see Theorem 3.10. Similar patterns have been observed for the power of AT in Figure 4 (b), (d), and (f). Of course, the power of AT is usually lower than that of DT under the same setup, especially when the signal strength is weak. This is the price paid for adaptivity.



(a)DT (b)AT

Fig. 3. Size for (a) DT and (b) AT with projection dimension varies. Signal strength $c = 0$.

5.2 Simulation Study II: EDK

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

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3063223, IEEE Transactions on Pattern Analysis and Machine Intelligence

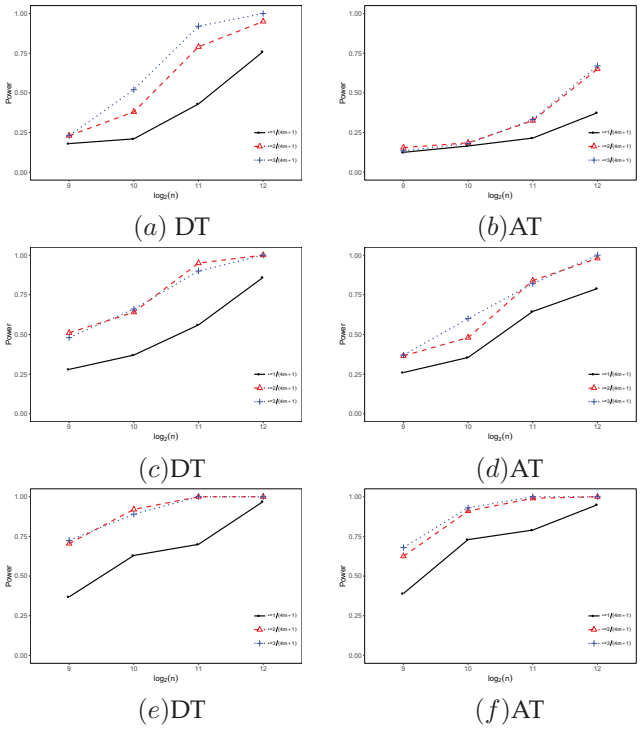


Fig. 4. Power for DT and AT with varying projection dimension. Signal strength $c = 0.01$ for (a) and (b); $c = 0.02$ for (c) and (d); $c = 0.03$ for (e) and (f).

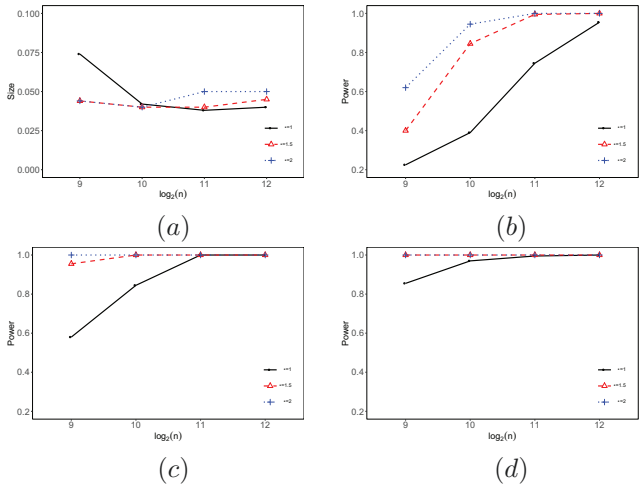


Fig. 5. Size and power for DT with varying projection dimensions. Signal strength $c = 0$ for (a); $c = 0.05$ for (b); $c = 0.1$ for (c); $c = 0.15$ for (d).

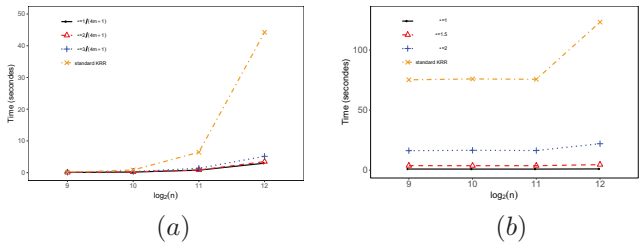


Fig. 6. Computing time for DT with varying projection dimensions: (a) is polynomially decay kernels; (b) is exponentially decay kernels.

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 5 about the size and power are similar to those for Figures 3 and 4. 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 3.7. Figure 6 demonstrates the significant computational advantage of DT (corresponding to $\gamma < 1$) over the testing procedure based on standard KRR (corresponding to $\gamma = 1$).

In the supplementary, we conduct additional synthetic experiments under the same simulation setup as Sections 5.1 and 5.2 except for using the Bernoulli random matrix. As shown in Figures S.1-S.3, the interpretations remain the same.

6 Discussion

The main contribution of this paper is to apply random projection to nonparametric testing, and propose the first "sharp" result regarding projection dimension to guarantee minimax optimal testing respectively for a general class of random projections. In practice, many other random projections also satisfy K-satisfiability, for example, the randomized orthogonal system (ROS) sketches introduced in [47], [7], by randomly sampling and rescaling the rows of a fixed orthonormal matrix. For the ROS examples, Assumption A3 also holds with extra logarithm factor in projection dimension, see Lemma S.8 in supplementary.

Stochastic approximation is another computationally efficient method for nonparametric learning. A representative approach is the stochastic gradient descent (SGD) algorithm. In SGD, the total step size within n steps iteration plays the role of the regularization to avoid overfitting, and the SGD estimator ([16]) can achieve the optimality provided that the total step size has the same order of the effective dimension which is represented by s_λ in our work. [48] established an early stopping rule for gradient descent algorithm to achieve optimal nonparametric testing. The result shows that the total step size in gradient descent plays the same role as $1/\lambda$ in classic KRR, and the optimal testing rate can be achieved by the same "bias-standard deviation" tradeoff while the bias and the standard deviation are represented as functions of the total step sizes. In sketched KRR, we showed that the projection dimension is required to be greater than the effective dimension s_λ to guarantee high power performance. Therefore, although the sketched KRR and SGD are two different computationally efficient approaches, they are connected through the statistical effective dimension.

7 Proof of Main Results

In this section, we present the main proofs of Lemma 3.1, sharpness properties including Theorem 3.9, Theorem 3.10.

Proofs for the rest of Lemmas and Theorems can be found in the Supplementary.

7.1 Proof of Lemma 3.1

Before the proof of Lemma 3.1, we first state some definitions and preliminary lemmas. Define $\kappa_\lambda = \frac{s_\lambda}{n\lambda}$, for any $\lambda > 0$. In fact, κ_λ is the variance-to-bias ratio, where λ and s_λ/n correspond to (squared-)bias and variance of $\widehat{f}_R$, respectively; see Corollary 3.4 and Lemma S.3 for details. Consider a bundle of function classes indexed by κ_λ:

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

Define a generalized version of local Rademacher complexity function

$$\Psi_\lambda(r) = E\Big\{ \sup_{\substack{f\in\mathcal{F}_\lambda \\ Pf^2\le r}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)\Big\}, r \ge 0, \tag{24}$$

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

$$\widehat{\Psi}_\lambda(r) = E\Big\{ \sup_{\substack{f\in\mathcal{F}_\lambda \\ P_n f^2\le r}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)\Big| x_1, \cdots, x_n\Big\}, r \ge 0. \tag{25}$$

When $\kappa_\lambda \asymp 1$, $\Psi_\lambda(\cdot)$ and $\widehat{\Psi}_\lambda(\cdot)$ become the original local Rademacher complexity functions introduced in [32]. 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 [22], [24], which corresponds to a different choice of κ_λ in $\mathcal{F}_\lambda$ as demonstrated in Section 3.

In the following Lemma 5 we represent the generalized local Rademacher complexity function and its empirical version by a function of eigenvalues and κ_λ. In Lemma 6, we further show that both Ψ_λ and $\widehat{\Psi}_\lambda$ possess unique (positive) fixed points. This fixed point property is crucial in proving Lemma 3.1. We defer the proof of Lemma 5 and Lemma 6 to Section S.5.2 in the Supplementary.

Lemma 5.

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

$$\Psi_\lambda(r) \asymp \sqrt{\frac{1}{n}\sum_{i=1}^\infty \kappa_\lambda \min\{\frac{r}{\kappa_\lambda}, \mu_i\}}. \tag{26}$$

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

$$\widehat{\Psi}_\lambda(r) \asymp \sqrt{\frac{1}{n}\sum_{i=1}^n \kappa_\lambda \min\{\frac{r}{\kappa_\lambda}, \widehat{\mu}_i\}}. \tag{27}$$

Lemma 6. 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 ready to prove Lemma 3.1.

Proof Plugging these fixed points into (26) and (27) in Lemma 5, we have

$$r_\lambda \asymp \sqrt{\frac{1}{n}\sum_{i=1}^\infty \kappa_\lambda \min\{\frac{r_\lambda}{\kappa_\lambda}, \mu_i\}}, \tag{28}$$

$$\widehat{r}_\lambda \asymp \sqrt{\frac{1}{n}\sum_{i=1}^n \kappa_\lambda \min\{\frac{\widehat{r}_\lambda}{\kappa_\lambda}, \widehat{\mu}_i\}}. \tag{29}$$

By Lemma 6, we have $r_\lambda \asymp s_\lambda/n$, leading to $r_\lambda/\kappa_\lambda \asymp \lambda$; for the empirical version, with probability at least $1 - e^{-cs_\lambda}$, $\widehat{r}_\lambda \asymp s_\lambda/n$ leading to $\widehat{r}_\lambda/\kappa_\lambda \asymp \lambda$. Recall that $\widehat{s}_\lambda = \mathrm{argmin}\{i : \widehat{\mu}_i \le \lambda\} - 1$. Then by (29), with probability at least $1 - e^{-cs_\lambda}$,

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

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

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

based on the definition (10) that $\lambda < \mu_{s_\lambda}$. ■

7.2 Proof of Theorem 3.9

Proof We construct the true $f_0(\cdot) = \sum_{i=1}^n K(x_i, \cdot)w_i$ with $w = (w_1, \cdots, w_n)^\top$. Then

$$\|f_0\|_{\mathcal{H}}^2 = nw^\top \mathbf{K} w = \beta^{*\top} D \beta^*,$$

where $\beta^* = \sqrt{n}U^\top w$. Therefore, the constrain $\|f_0\|_{\mathcal{H}} \le C$ is equivalent to β^* ranges over all vectors satisfying $\|D^{1/2}\beta^*\|_2 \le C$. Then we have

$$\begin{aligned}\|f_0 - \widehat{f}_R\|_n^2 =& \|UD\beta^* - UDUS^\top \widehat{\alpha}\|_2^2 \\ =& \|D^{1/2}(D^{1/2}\beta^*) - D^{1/2}(SUD^{1/2})^\top \widehat{\alpha}\|_2^2,\end{aligned}$$

where $\widehat{\alpha} = (S\mathbf{K}^2 S^\top + \lambda S\mathbf{K}S^\top)^{-1} S\mathbf{K}\mathbf{y}$. Since the vector $(SUD^{1/2})^\top \widehat{\alpha}$ belongs to the column space of $D^{1/2}US^\top \in \mathbb{R}^{n\times s}$, and $dim(D^{1/2}US^\top) = s$. Then

$$\begin{aligned}\|f_0 - \widehat{f}_R\|_n^2 =& \|D^{1/2}(D^{1/2}\beta^*)\|_2^2 + \|D^{1/2}(SUD^{1/2})^\top \widehat{\alpha}\|_2^2 \\ &+ \langle D^{1/2}(D^{1/2}\beta^*), D^{1/2}(SUD^{1/2})^\top \widehat{\alpha}\rangle_2 \\ :=& A_1 + A_2 + A_3.\end{aligned}$$

By choosing β^* orthogonal to the span of $US^\top$, so that $D^{1/2}\beta^*$ is orthogonal to $D^{1/2}US^\top$, and then we have $A_3 = 0$. Applying the minimax characterization of eigenvalues of the $n \times n$ matrix $D^{1/2}$, we have with probability greater than $1 - e^{-cn\delta_n}$, when $s \ll s^\dagger$,

$$\begin{aligned}\sup_{\|f_0\|_{\mathcal{H}}\le 1} \|f_0 - \widehat{f}_R\|_n^2 \ge& \min_{V: dim(V) = s} \max_{v\in V^\perp : \|v\|_2 \le 1} \|D^{1/2}v\|_2^2 \\ =& \widehat{\mu}_{s+1} \ge c'\mu_{s+1} \gg c'\lambda^\dagger,\end{aligned}$$

where c' is a constant. ■

7.3 Proof of Theorem 3.10

Proof Without loss of generality, here we consider $H_0 : f = f_0$ with $f_0 = 0$. We need to prove that when projection dimension s is too small, for any random matrix S, there exists true function f, such that $\|f\|_{\mathcal{H}} \leq C$, $\|f\|_n \geq d^*$, our testing rule still cannot detect it.

We show the existence of such true function f as follows. We construct the true $f(\cdot) = \sum_{i=1}^n K(x_i, \cdot) w_i$ with $w = (w_1, \cdots, w_n)^\top$. Then $\|f\|_{\mathcal{H}}^2 = n w^\top \mathbf{K} w = \beta^{*\top} D \beta^*$, where $\beta^* = \sqrt{n} U^\top w$. Therefore, the constrain $\|f\|_{\mathcal{H}} \leq C$ is equivalent to β^* ranges over all vectors satisfying $\|D^{1/2}\beta^*\|_2 \leq C$.

Notice that $n T_{n,\lambda} = n\|\widehat{f}_R\|_n^2 = T_1 + T_2 + 2T_3$, where

$$\begin{aligned}
T_1 =& n\|E_\epsilon \widehat{f}_R\|_n^2 = \|\mathbf{K} S^\top (S\mathbf{K}^2 S^\top + \lambda S \mathbf{K} S^\top)^{-1} S \mathbf{K} \boldsymbol{f}\|_2^2, \\
T_2 =& n\|\mathbf{K} S^\top (S\mathbf{K}^2 S^\top + \lambda S \mathbf{K} S^\top)^{-1} S \mathbf{K} \epsilon\|_n^2, \\
T_3 =& (\mathbf{K} S^\top (S\mathbf{K}^2 S^\top + \lambda S \mathbf{K} S^\top)^{-1} S \mathbf{K} \boldsymbol{f})^\top \\
& \cdot \mathbf{K} S^\top (S\mathbf{K}^2 S^\top + \lambda S \mathbf{K} S^\top)^{-1} S \mathbf{K} \epsilon.
\end{aligned}$$

For T_1, plugging in the expression $\boldsymbol{f} = n\mathbf{K}w = \sqrt{n} U D^{1/2} D^{1/2} \beta^*$, we have

$$T_1 = n\|D\widetilde{S}^\top (\widetilde{S} D^2 \widetilde{S}^\top + \lambda \widetilde{S} D \widetilde{S}^\top)^{-1} \widetilde{S} D^{3/2} D^{1/2} \beta^*\|_2^2,$$

where $\widetilde{S} = SU$. Denote $\boldsymbol{\Delta} = (D\widetilde{S}(\widetilde{S} D^2 \widetilde{S}^\top + \lambda \widetilde{S} D \widetilde{S})^{-1} \widetilde{S} D^{3/2})^\top$, then $T_1 = n\|\boldsymbol{\Delta}^\top D^{1/2} \beta^*\|_2^2$. Notice that the $span\{\boldsymbol{\Delta}\} \subset span\{D^{3/2}\widetilde{S}^\top\}$, and $dim(D^{3/2}\widetilde{S}^\top) = s$. There always exists $\beta^* \in \mathbb{R}^{n\times 1}$ such that $D^{1/2}\beta^*$ is orthogonal to the column space of $D^{3/2}\widetilde{S}^\top$, so that $\boldsymbol{\Delta}^\top D^{1/2}\beta^* = 0$ leading to $T_1 = 0$. Furthermore, based on the above argument, we have $T_3 = 0$. Then we have

$$\frac{T_{n,\lambda} - \mu_{n,\lambda}}{\sigma_{n,\lambda}} = \frac{T_2/n - \mu_{n,\lambda}}{\sigma_{n,\lambda}} \xrightarrow{d} N(0,1).$$

On the other hand, the signal of such f satisfies

$$\begin{aligned}
\sup_{\|f\|_{\mathcal{H}} \leq 1} \|f\|_n^2 =& \sup_{\substack{\beta^* \in \mathbb{R}^{n\times 1} \\ \|D^{1/2}\beta^*\|_2 \leq 1}} \|D^{1/2} D^{1/2} \beta^*\|_2^2 \\
\geq& \min_{V: dim(V) = s} \max_{v \in V^\perp : \|v\|_2 \leq 1} \|D^{1/2} v\|_2^2 = \widehat{\mu}_{s+1}.
\end{aligned}$$

Let $\beta_{n,\lambda} = \widehat{\mu}_{s+1}/d_n^{*2}$, with probability greater than $1 - e^{-n\delta_n}$, $\beta_{n,\lambda} \to \infty$ when $s \ll s^*$ and $n \to \infty$.

Therefore, there exists $f \in \mathcal{B}$ satisfying $\|f\|_n^2 \geq \widehat{\mu}_{s+1} = \beta_{n,\lambda} d^{*2} \geq d^{*2}$, but

$$P_f\big(\phi_{n,\lambda} = 1 | \mathbf{x}, S\big) \leq P_f\big(\phi_{n,\lambda} = 1 | \mathbf{x}, S\big) \to \alpha.$$

■

REFERENCES

[1] Y. Zhang, J. Duchi, and M. Wainwright, "Divide and conquer kernel ridge regression," *Conference on Learning Theory*, pp. 592–617, 2013.

[2] M. I. Jordan and R. Jacobs, "Supervised learning and divide-and-conquer: A statistical approach," *Proceedings of the Tenth International Conference on Machine Learning*, pp. 159–166, 2014.

[3] X. Chang, S. Lin, and D. Zhou, "Distributed semi-supervised learning with kernel ridge regression," *Journal of Machine Learning Research*, vol. 18, no. 46, pp. 1–22, 2017.

[4] Z. Shang and G. Cheng, "Computational limits of a distributed algorithm for smoothing spline," *Journal of Machine Learning Research*, vol. 18, no. 108, pp. 1–37, 2017.

[5] C. Musco and C. Musco, "Recursive sampling for the nyström method," *Advances in Neural Information Processing Systems*, pp. 3836–3848, 2017.

[6] M. Lopes, L. Jacob, and M. J. Wainwright, "A more powerful two-sample test in high dimensions using random projection," *Advances in Neural Information Processing Systems*, pp. 1206–1214, 2011.

[7] Y. Yang, M. Pilanci, and M. J. Wainwright, "Randomized sketches for kernels: Fast and optimal non-parametric regression," *The Annals of Statistics*, vol. 456, pp. 991 – 1023, 2017.

[8] C. Huang and X. Huo, "A statistically and numerically efficient independence test based on random projections and distance covariance," *arXiv preprint arXiv:1701.06054*, 2017.

[9] Y. Kim and C. Gu, "Smoothing spline gaussian regression: more scalable computation via efficient approximation," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 66, no. 2, pp. 337–356, 2004.

[10] P. Ma, N. Zhang, J. Z. Huang, and W. Zhong, "Adaptive basis selection for exponential family smoothing splines with application in joint modeling of multiple sequencing samples," *Statistica Sinica*, vol. 27, pp. 1757–1777, 2017.

[11] A. Alaoui and M. W. Mahoney, "Fast randomized kernel ridge regression with statistical guarantees," *Advances in Neural Information Processing Systems*, pp. 775–783, 2015.

[12] A. Gittens and M. W. Mahoney, "Revisiting the nyström method for improved large-scale machine learning," *Journal of Machine Learning Research*, vol. 28, no. 3, pp. 567–575, 2013.

[13] A. Rudi, R. Camoriano, and L. Rosasco, "Less is more: Nyström computational regularization," *Advances in Neural Information Processing Systems*, pp. 1657–1665, 2015.

[14] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "Moa: Massive online analysis," *Journal of Machine Learning Research*, vol. 11, no. May, pp. 1601–1604, 2010.

[15] E. D. Schifano, J. Wu, C. Wang, J. Yan, and M. Chen, "Online updating of statistical inference in the big data setting," *Technometrics*, vol. 58, no. 3, pp. 393–403, 2016.

[16] A. Dieuleveut and F. Bach, "Nonparametric stochastic approximation with large step-sizes," *The Annals of Statistics*, vol. 44, no. 4, pp. 1363–1399, 2016.

[17] S. Volgushev, S.-K. Chao, and G. Cheng, "Distributed inference for quantile regression processes," *arXiv preprint arXiv:1701.06088*, 2017.

[18] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[19] D. Cox, E. Koh, G. Wahba, and B. S. Yandell, "Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models," *The Annals of Statistics*, vol. 16, no. 1, pp. 113–119, 1988.

[20] A. Liu and Y. Wang, "Hypothesis testing in smoothing spline models," *Journal of Statistical Computation and Simulation*, vol. 74, no. 8, pp. 581–597, 2004.

[21] J. Fan, C. Zhang, and J. Zhang, "Generalized likelihood ratio statistics and wilks phenomenon," *The Annals of statistics*, vol. 29, no. 1, pp. 153–193, 2001.

[22] Z. Shang and G. Cheng, "Local and global asymptotic inference in smoothing spline models," *The Annals of Statistics*, vol. 41, no. 5, pp. 2608–2638, 2013.

[23] T. W. Anderson and D. A. Darling, "Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes," *The Annals of Mathematical Statistics*, vol. 23, no. 2, pp. 193–212, 1952.

[24] Y. I. Ingster, "Asymptotically minimax hypothesis testing for nonparametric alternatives. i, ii, iii," *Mathematical Methods of Statistics*, vol. 2, no. 2, pp. 85–114, 1993.

[25] O. V. Lepski, V. G. Spokoiny *et al.*, "Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative," *Bernoulli*, vol. 5, no. 2, pp. 333–358, 1999.

[26] Y. Wei and M. J. Wainwright, "The local geometry of testing in ellipses: Tight control via localized kolomogorov widths," *arXiv preprint arXiv:1712.00711*, 2017.

[27] R. L. Eubank and C. H. Spiegelman, "Testing the goodness of fit of a linear model via nonparametric regression techniques," *Journal of the American Statistical Association*, vol. 85, no. 410, pp. 387–392, 1990.

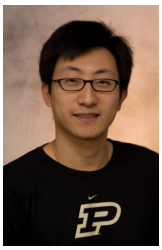
[28] A. Azzalini and A. Bowman, "On the use of nonparametric regression for checking linear relationships," *Journal of the Royal*

Statistical Society: Series B (Methodological), vol. 55, no. 2, pp. 549–557, 1993.
[29] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge university press, 2004.
[30] N. Ailon and B. Chazelle, "The fast johnson–lindenstrauss transform and approximate nearest neighbors," *SIAM Journal on computing*, vol. 39, no. 1, pp. 302–322, 2009.
[31] Y. Koike, "Gaussian approximation of maxima of wiener functionals and its application to high-frequency data," *Annals of Statistics*, 2018.
[32] P. L. Bartlett, O. Bousquet, and S. Mendelson, "Local rademacher complexities," *The Annals of Statistics*, vol. 33, no. 4, pp. 1497–1537, 2005.
[33] G. Xu, Z. Shang, and G. Cheng, "Optimal tuning for divide-and-conquer kernel ridge regression with massive data," *arXiv preprint arXiv:1612.05907*, 2016.
[34] H. Q. Minh, P. Niyogi, and Y. Yao, "Mercer's theorem, feature maps, and smoothing," in *Learning theory*. Springer, 2006, pp. 154–168.
[35] G. Wahba, *Spline models for observational data*. SIAM, 1990.
[36] W. Guo, "Inference in smoothing spline analysis of variance," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 4, pp. 887–898, 2002.
[37] B. Schölkopf, C. J. Burges, and A. J. Smola, *Advances in kernel methods: support vector learning*. MIT press, 1999.
[38] C. Boutsidis and A. Gittens, "Improved matrix algorithms via the subsampled randomized hadamard transform," *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 3, pp. 1301–1340, 2013.
[39] Y. Wei, F. Yang, and M. J. Wainwright, "Early stopping for kernel boosting algorithms: A general analysis with localized complexities," in *Advances in Neural Information Processing Systems*, 2017, pp. 6067–6077.
[40] J. Shawe-Taylor, C. K. Williams, N. Cristianini, and J. Kandola, "On the eigenspectrum of the gram matrix and the generalization error of kernel-pca," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2510–2522, 2005.
[41] M. L. Braun, "Accurate error bounds for the eigenvalues of the kernel matrix," *Journal of Machine Learning Research*, vol. 7, no. Nov, pp. 2303–2328, 2006.
[42] A. Saumard and J. A. Wellner, "Log-concavity and strong log-concavity: a review," *Statistics surveys*, vol. 8, p. 45, 2014.
[43] W. Hardle, E. Mammen *et al.*, "Comparing nonparametric versus parametric regression fits," *The Annals of Statistics*, vol. 21, no. 4, pp. 1926–1947, 1993.
[44] V. G. Spokoiny, "Adaptive hypothesis testing using wavelets," *The Annals of Statistics*, vol. 24, no. 6, pp. 2477–2498, 1996.
[45] H. Cramér, *Mathematical Methods of Statistics*. Princeton university press, 2016, vol. 9.
[46] C. Gu, *Smoothing spline ANOVA models*. Springer Science & Business Media, 2013, vol. 297.
[47] M. Pilanci and M. J. Wainwright, "Randomized sketches of convex programs with sharp guarantees," *IEEE Transactions on Information Theory*, vol. 61, no. 9, pp. 5096–5115, 2015.
[48] M. Liu and G. Cheng, "Early stopping for nonparametric testing," in *Advances in Neural Information Processing Systems*, 2018, pp. 3985–3994.

Meimei Liu received B.S. degree in mathematics in Anhui University, Hefei, China, 2010, M.S. degree in statistics in University of Science and Technology of China, Hefei, China, 2013, and PhD degree from Purdue University in 2018. From 2018 to 2020, she was a postdoctoral researcher in Statistics at Duke University, working on problems in network embedding with application in neuroscience, variational inference and domain adaptation. She is currently an Assistant Professor in Statistics at Virginia Tech. Her current research interests include scalable statistical inference, statistical learning theory, and developing computationally efficient probabilistic models for complex data.

Zuofeng Shang received his BS and MS degrees in mathematics in Nankai University, China, in 2003 and 2006 respectively, and PhD degree in statistics from University of Wisconsin–Madison in 2011. He was a postdoctoral researcher in University of Notre Dame and Cornell University, and was a Visiting Assistant Professor in Purdue University. He is currently an Associate Professor in mathematics at New Jersey Institute of Technology, prior to that he was an Assistant Professor in mathematics at SUNY-Binghamton and IUPUI. He is interested in learning theory, primarily in statistical aspect. He acknowledges NSF DMS 1764280 and 1821157 for supporting this work.

Yun Yang received the B.S. degree in Mathematics from Tsinghua University, Beijing, China, in 2011, and the Ph.D. degree in Statistics from Duke University in 2014. From 2014 to 2016, he was a postdoctoral researcher at University of California, Berkeley, working on problems in machine learning, optimization and high-dimensional statistics. From 2016 to 2018, he was an Assistant Professor in Statistics at Florida State University. He is currently an Assistant Professor in Statistics at University of Illinois Urbana-Champaign. His current research interests include scalable statistical computation, statistical learning theory and machine learning. He acknowledges NSF DMS 1907316 for supporting this work.



Guang Cheng received BA degree in Economics from Tsinghua University, China, in 2002, and PhD degree from University of Wisconsin–Madison in 2006. He then joined Dept of Statistics at Duke University as Visiting Assisitant Professor and Postdoc Fellow in SAMSI. He is currently Professor in Statisics at Purdue University, directing Big Data Theory research group, whose main goal is to develop computationally efficient inferential tools for big data with statistical guarantees. He acknowledges support by NSF DMS-1712907, DMS-1811812, DMS-1821183, Office of Naval Research (ONR N00014-18-2759), and Adobe Data Science Award.