Journal of Machine Learning Research 22 (2021) 1-45 Submitted 9/17; Revised 6/20; Published 1/21

On the Optimality of Kernel-Embedding Based Goodness-of-Fit Tests

Krishnakumar Balasubramanian KBALA@UCDAVIS.EDU
Department of Statistics
University of California, Davis
Davis, CA 95616, USA

Tong Li TONG.LI@COLUMBIA.EDU
Ming Yuan MING.YUAN@COLUMBIA.EDU
Department of Statistics
Columbia University
New York, NY 10027, USA

Editor: Arthur Gretton

Abstract

The reproducing kernel Hilbert space (RKHS) embedding of distributions offers a general and flexible framework for testing problems in arbitrary domains and has attracted considerable amount of attention in recent years. To gain insights into their operating characteristics, we study here the statistical performance of such approaches within a minimax framework. Focusing on the case of goodness-of-fit tests, our analyses show that a vanilla version of the kernel embedding based test could be minimax suboptimal, when considering χ^2 distance as the separation metric. Hence we suggest a simple remedy by moderating the embedding. We prove that the moderated approach provides optimal tests for a wide range of deviations from the null and can also be made adaptive over a large collection of interpolation spaces. Numerical experiments are presented to further demonstrate the merits of our approach.

Keywords: Adaptation, goodness of fit, maximum mean discrepancy, optimal rates of Convergence, reproducing kernel Hilbert space.

1. Introduction

In recent years, statistical tests based on the reproducing kernel Hilbert space (RKHS) embedding of distributions have attracted much attention because of their flexibility and broad applicability. Like other kernel methods, RKHS embedding based tests present a general and unifying framework for testing problems in arbitrary domains by using appropriate kernels defined on those domains. See Muandet et al. (2017) for a detailed review of kernel embedding and its applications. The idea of using kernel embedding for comparing probability distributions was initially introduced by Smola et al. (2007); Gretton et al. (2007, 2012a). Related extensions were also proposed by Harchaoui et al. (2007); Zaremba et al. (2013). Furthermore, Sejdinovic et al. (2013) established a close relationship between kernel-based hypothesis tests and energy distanced based test introduced by Székely et al. (2007). See also Lyons (2013). More recently, motivated by several applications based on quantifying the convergence of Monte Carlo simulations, Liu et al. (2016), Chwialkowski

et al. (2016) and Gorham and Mackey (2017) proposed goodness-of-fit tests which were based on combining the kernel based approach with Stein's identity. A linear-time method for goodness-of-fit was also proposed by Jitkrittum et al. (2017) recently. Finally, the idea of kernel embedding has also been used for constructing implicit generative models (*e.g.*, Dziugaite et al., 2015; Li et al., 2015).

Despite their popularity, fairly little is known about the statistical performance of these kernel embedding based tests. Our goal is to fill in this void. In particular, we focus on kernel embedding based goodness-of-fit tests and investigate their power under a general composite alternative. Our results not only provide new insights on the operating characteristics of these kernel embedding based tests but also suggest improved testing procedures that are minimax optimal and adaptive over a large collection of alternatives, when considering χ^2 distance as the separation metric.

More specifically, let $X_1, \cdots, X_n$ be n independent $\mathcal{X}$-valued observations from a certain probability measure P. We are interested in testing if the hypothesis $H_0 : \ P = P_0$ holds for a fixed P_0. Problems of this kind have a long and illustrious history in statistics and is often associated with household names such as *Kolmogrov-Smirnov tests, Pearson's Chi-square test* or *Neyman's smooth test*. A plethora of other techniques have also been proposed over the years in both parametric and nonparametric settings (*e.g.*, Ingster and Suslina, 2003; Lehmann and Romano, 2008). Most of the existing techniques are developed with the domain $\mathcal{X} = \mathbb{R}$ or $[0, 1]$ in mind and work the best in these cases. Modern applications, however, oftentimes involve domains different from these traditional ones. For example, when dealing with directional data, which arise naturally in applications such as diffusion tensor imaging, it is natural to consider $\mathcal{X}$ as the unit sphere in $\mathbb{R}^3$ (*e.g.*, Jupp, 2005). Another example occurs in the context of ranking or preference data (*e.g.*, Ailon et al., 2008). In these cases, $\mathcal{X}$ can be taken as the group of permutations. Furthermore, motivated by several applications, combinatorial testing problems have been investigated recently (*e.g.*, Addario-Berry et al., 2010), where the spaces under consideration are specific combinatorially structured spaces.

A particularly attractive approach to goodness-of-fit testing problems in general domains is through RKHS embedding of distributions. Specifically, let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric positive (semi-)definite kernel. The Moore-Aronszajn Theorem indicates that there is an RKHS, denoted by $(\mathcal{H}(K), \langle \cdot, \cdot \rangle_K)$, uniquely identified with the kernel K (*e.g.*, Aronszajn, 1950). The RKHS embedding of a probability measure P with respect to K is given by

$$\mu_P(\cdot) := \int_{\mathcal{X}} K(x, \cdot) dP(x).$$

It is well-known that, under mild regularity conditions, then $\mu_P \in \mathcal{H}(K)$ and furthermore for any $f \in \mathcal{H}(K)$,

$$\mathbb{E}_P f(X) = \langle \mu_P, f \rangle_K, \quad \forall \ f \in \mathcal{H}(K),$$

where $\mathbb{E}_P$ signifies that the expectation is taken over $X \sim P$. The so-called *maximum mean discrepancy* (MMD) between two probability measures P and Q is defined as

$$\gamma_K(P, Q) := \sup_{f \in \mathcal{H}(K): \|f\|_K \leq 1} \int_{\mathcal{X}} f(x) d(P - Q)(x),$$

where $\|\cdot\|_K$ is the norm associated with $(\mathcal{H}(K), \langle\cdot,\cdot\rangle_K)$. It is not hard to see

$$\gamma_K(P, Q) = \|\mu_P - \mu_Q\|_K.$$

See, *e.g.*, Sriperumbudur et al. (2010) or Gretton et al. (2012a) for details. The goodness-of-fit test can be carried out conveniently through RKHS embeddings of P and P_0 by first constructing an estimate of $\gamma_K(P, P_0)$:

$$\gamma_K(\widehat{P}_n, P_0) := \sup_{f\in\mathcal{H}(K):\|f\|_K\le 1} \int_{\mathcal{X}} f(x) d\left(\widehat{P}_n - P_0\right)(x),$$

where $\widehat{P}_n$ is the empirical distribution of $X_1, \cdots, X_n$, and then rejecting H_0 if the estimate exceeds a threshold calibrated to ensure a certain significance level, say α $(0 < \alpha < 1)$.

In this paper, we investigate the power of the above discussed testing strategy under a general composite alternative. Following the spirit of Ingster and Suslina (2003), we consider in particular a set of alternatives that are increasingly close to the null hypothesis. To fix ideas, we assume hereafter that P is dominated by P_0 under the alternative so that the Radon-Nikodym derivative dP/dP_0 is well defined. Recall that the χ^2 divergence between P and P_0 is defined as

$$\chi^2(P, P_0) := \int_{\mathcal{X}} \left(\frac{dP}{dP_0}\right)^2 dP_0 - 1.$$

We are particularly interested in the detection boundary, namely how close P and P_0 can be in terms of χ^2 distance, under the alternative, so that a test based on a sample of n observations can still consistently distinguish between the null hypothesis and the alternative. For example, in the parametric setting where P is known up to a finite dimensional parameters under the alternative, the detection boundary of the likelihood ratio test is n^{-1} under mild regularity conditions (*e.g.*, Theorem 13.5.4 in Lehmann and Romano, 2008, and the discussion leading to it). We are concerned here with alternatives that are nonparametric in nature. Our first result suggests that the detection boundary for aforementioned $\gamma_K(\widehat{P}_n, P_0)$ based test is of the order $n^{-1/2}$. However, our main results indicate, perhaps surprisingly at first, that this rate is far from optimal and the gap between it and the usual parametric rate can be largely bridged.

In particular, we argue that the distinguishability between P and P_0 depends on how close $u := dP/dP_0 - 1$ is to the RKHS $\mathcal{H}(K)$. The closeness of u to $\mathcal{H}(K)$ can be measured by the distance from u to an arbitrary ball in $\mathcal{H}(K)$. In particular, we shall consider the case where $\mathcal{H}(K)$ is dense in $L_2(P_0)$, and focus on functions that are polynomially approximable by $\mathcal{H}(K)$ for concreteness. More precisely, for some constants $M, \theta > 0$, denote by $\mathcal{F}(\theta; M)$ the collection of functions $f \in L_2(P_0)$ such that for any $R > 0$, there exists an $f_R \in \mathcal{H}(K)$ such that

$$\|f_R\|_K \le R, \qquad \text{and} \qquad \|f - f_R\|_{L_2(P_0)} \le MR^{-1/\theta}.$$

We also adopt the convention that

$$\mathcal{F}(0; M) = \{f \in \mathcal{H}(K) : \|f\|_K \le M\}.$$

See Section 5 for a more concrete example of the space $\mathcal{F}(\theta; M)$ when $\mathcal{H}(K)$ is the usual Sobolev space defined over $[0, 1]$. Interested readers are also referred to Cucker and Zhou

(2007) for further discussion on these so-called interpolation spaces and their use in statistical learning.

We investigate the optimal rate of detection for testing $H_0 : P = P_0$ against

$$H_1(\Delta_n, \theta, M) : P \in \mathcal{P}(\Delta_n, \theta, M), \tag{1}$$

where $\mathcal{P}(\Delta_n, \theta, M)$ is the collection of distributions P on $(\mathcal{X}, \mathcal{B})$ satisfying:

$$dP/dP_0 - 1 \in \mathcal{F}(\theta; M), \qquad \text{and} \qquad \chi^2(P, P_0) \geq \Delta_n.$$

We call r_n the optimal rate of detection if for any $c > 0$, there exists no consistent test whenever $\Delta_n \leq cr_n$; and on the other hand, a consistent test exists as long as $\Delta_n \gg r_n$.

Although one could consider a more general setup, for concreteness, we assume that the eigenvalues of K with respect to $L_2(P_0)$ decays polynomially in that $\lambda_k \asymp k^{-2s}$. We show that the optimal rate of detection for testing H_0 against $H_1(\Delta_n, \theta, M)$ for any $\theta \geq 0$ is $n^{-\frac{4s}{4s+\theta+1}}$. The rate of detection, although not achievable with a $\gamma_K(\widehat{P}_n, P_0)$ based test, can be attained via a moderated version of the MMD based approach. A practical challenge to the approach, however, is its reliance on the knowledge of θ. Unlike s which is determined by K and P_0 and therefore known a priori, θ depends on u and is not known in advance. This naturally brings about the issue of adaptation – is there an agnostic approach that can adaptively attain the optimal detection boundary without the knowledge of θ. We show that the answer is affirmative although a small price in the form of $\log \log n$ is required to achieve such adaptation.

The minimax framework we considered connects our work with the extensive statistics literature on minimax hypothesis testing. See, e.g., Ingster (1987); Ermakov (1991); Ingster (1993); Spokoiny (1996); Lepski and Spokoiny (1999); Ingster and Suslina (2000); Ingster (2000); Baraud (2002); Fromont and Laurent (2006); Fromont et al. (2012, 2013), among many others. As is customary in other areas of nonparametric statistics, these works usually start by characterizing function spaces for the alternatives such as Hölder, Sobolev or more generally Besov spaces, followed by devising an optimal testing procedure according to the specific class of alternatives. This creates a subtle yet important difference from the kernel based approaches where a method or algorithm is developed first with a particular kernel often specific to the applications in mind, and it is therefore of interest to investigate a posteriori the performance of such a method. The connection between the two paradigms is well understood in the context of supervised learning thanks to the one-to-one correspondence between a kernel and a RKHS: the two approaches are essentially equivalent in that kernel methods with appropriate regularization could achieve minimax optimality when considering functions from the RKHS with which the kernel identifies. See, e.g., Wahba (1990); Scholkopf and Smola (2001). In a sense, our work establishes a similar relationship in the context of kernel methods for hypothesis testing. Indeed, to achieve the aforementioned optimal rate of detection, we introduce a class of tests based on a modified MMD similar in spirit to that from Harchaoui et al. (2007). The modification we applied is akin to the RKHS norm regularization commonly used for supervised learning. The key idea is to allow the kernel in MMD to evolve with the number of observations so that MMD becomes increasingly similar to the χ^2 distance.

The rest of the paper is organized as follows. We first analyze the power of MMD based tests in Section 2. This analysis reveals a significant gap between the detection boundary

achieved by the MMD based test and the usual parametric $1/n$ rate. In turn, this prompts us to introduce, in Section 3, a class of tests based on a modified MMD that are rate optimal. To address the practical challenge of choosing an appropriate tuning parameter for these tests, we investigate the issue of optimal adaptation in Section 4, where we establish the optimal rates of detection for adaptively testing H_0 against a broader set of alternatives and propose a test based on the modified MMD that can attain these rates. To further illustrate the implications of our results, we consider in Section 5 the specific case of Sobolev kernels and compare our results with those known for nonparametric testing within Sobolev spaces. Numerical experiments are presented in Section 6. We conclude with some remarks in Section 7. All proofs are relegated to Section 8.

2. Operating characteristics of MMD based test

We begin by reviewing a simple MMD based test procedure and investigating its operating characteristics.

2.1 Background and notation

Throughout the paper, we shall assume

$$\int_{\mathcal{X}\times\mathcal{X}} K^2(x,x')dP_0(x)dP_0(x') < \infty.$$

Hence the Hilbert-Schmidt integral operator

$$L_K(f)(x) = \int_{\mathcal{X}} K(x,x')f(x')dP_0(x'), \quad \forall\ x \in \mathcal{X}$$

is well-defined. The spectral decomposition theorem ensures that L_K admits an eigenvalue decomposition. Let $\{\phi_k\}_{k\geq 1}$ denote the orthonormal eigenfunctions of L_K with eigenvalues λ_k's such that $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_k \geq \cdots > 0$. Then as proved in, *e.g.*, Dunford and Schwartz (1963),

$$K(x,x') = \sum_{k\geq 1} \lambda_k \varphi_k(x)\varphi_k(x') \tag{2}$$

in $L_2(P_0 \otimes P_0)$. We further assume that K is continuous and that P_0 is nondegenerate, meaning the support of P_0 is $\mathcal{X}$. Then Mercer's theorem ensures that (2) holds pointwisely. See, *e.g.*, Theorem 4.49 of Steinwart and Christmann (2008).

As mentioned in Section 1, the squared MMD between two probability distributions P and P_0 can be expressed as

$$\gamma_K^2(P,P_0) = \int K(x,x')d(P-P_0)(x)d(P-P_0)(x'). \tag{3}$$

Write

$$\bar{K}(x,x') = K(x,x') - \mathbb{E}_{P_0}K(x,X) - \mathbb{E}_{P_0}K(X,x') + \mathbb{E}_{P_0}K(X,X'), \tag{4}$$

where the subscript P_0 signifies the fact that the expectation is taken over $X, X' \sim P_0$ independently. By (4), $\gamma_K^2(P, P_0) = \gamma_{\bar{K}}^2(P, P_0)$. Therefore, without loss of generality, we can focus on kernels that are degenerate under P_0, *i.e.*,

$$\mathbb{E}_{P_0} K(X, \cdot) = 0. \tag{5}$$

For brevity, we shall omit the subscript K in γ in the rest of the paper, unless it is necessary to emphasize the dependence of MMD on the reproducing kernel. Passing from a nondegenerate kernel to a degenerate one however presents a subtlety regarding universality. Universality of a kernel is essential for MMD by ensuring that $dP/dP_0 - 1$ resides in the linear space spanned by its eigenfunctions. See, *e.g.*, Steinwart (2001) for the definition of universal kernel and Sriperumbudur et al. (2011) for a detailed discussion of different types of universality. Observe that $dP/dP_0 - 1$ necessarily lies in the orthogonal complement of constant functions in $L_2(P_0)$. A degenerate kernel K is universal if its eigenfunctions $\{\varphi_k\}_{k\geq 1}$ form an orthonormal basis of the orthogonal complement of linear space $\{c \cdot \varphi_0 : c \in \mathbb{R}\}$ where $\varphi_0(x) = 1$ in $L_2(P_0)$. In what follows, we shall assume that K is both degenerate and universal.

For the sake of concreteness, we shall also assume that K has infinitely many positive eigenvalues decaying polynomially, *i.e.*,

$$0 < \varliminf_{k\to\infty} k^{2s}\lambda_k \leq \varlimsup_{k\to\infty} k^{2s}\lambda_k < \infty \tag{6}$$

for some $s > 1/2$. In addition, we also assume that the eigenfunctions of K are uniformly bounded, *i.e.*,

$$\sup_{k\geq 1} \|\varphi_k\|_\infty < \infty, \tag{7}$$

Assumption (7) is satisfied for many commonly used kernels such as those associated with Sobolev spaces we shall describe in further details in Section 5. Together with Assumptions (6), (7) ensures that Mercer's decomposition (2) holds uniformly. In general, however, there are also situations under which (7) may not hold. For example, when considering discrete and countable domains, several standard kernels don't satisfy Assumption (7); see, *e.g.*, 2.24 in Scholkopf and Smola (2001), p. 58. Although it is plausible that most if not all of our results will continue to hold even without the uniform boundedness of φ_ks, a rigorous argument to do away with it has so far eluded us.

Note that (5) implies $\mathbb{E}_{P_0}\varphi_k(X) = 0$, $\forall\ k \geq 1$. The uniform convergence in (2) together with (3) give

$$\gamma^2(P, P_0) = \sum_{k\geq 1} \lambda_k [\mathbb{E}_P \varphi_k(X)]^2$$

for any P. Accordingly, when P is replaced by the empirical distribution $\widehat{P}_n$, the empirical squared MMD can be expressed as

$$\gamma^2(\widehat{P}_n, P_0) = \sum_{k\geq 1} \lambda_k \left[\frac{1}{n}\sum_{i=1}^n \varphi_k(X_i)\right]^2.$$

Classic results on the asymptotics of V-statistic (Serfling, 2009) imply that

$$n\gamma^2(\widehat{P}_n, P_0) \xrightarrow{d} \sum_{k\geq 1} \lambda_k Z_k^2 := W$$

under H_0, where $Z_k \overset{i.i.d.}{\sim} N(0,1)$. Let T_{MMD} be an MMD based test, which rejects H_0 if and only if $n\gamma^2(\widehat{P}_n, P_0)$ exceeds the $1-\alpha$ quantile $q_{w,1-\alpha}$ of W, *i.e.*,

$$T_{\mathrm{MMD}} = \mathbb{1}_{\{n\gamma^2(\widehat{P}_n, P_0) > q_{w,1-\alpha}\}}.$$

The above limiting distribution of $n\gamma^2(\widehat{P}_n, P_0)$ immediately suggests that T_{MMD} is an asymptotic α-level test.

2.2 Power analysis for MMD based tests

We now investigate the power of T_{MMD} in testing H_0 against $H_1(\Delta_n, \theta, M)$ given by (1). Recall that the type II error of a test $T : \mathcal{X}^n \to [0,1]$ for testing H_0 against a composite alternative $H_1 : P \in \mathcal{P}$ is given by

$$\beta(T; \mathcal{P}) = \sup_{P\in\mathcal{P}} \mathbb{E}_P[1 - T(X_1, \ldots, X_n)],$$

where $\mathbb{E}_P$ means taking expectation over $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$. For brevity, we shall write $\beta(T; \Delta_n, \theta, M)$ instead of $\beta(T; \mathcal{P}(\Delta_n, \theta, M))$ in what follows. The performance of a test T can then be evaluated by its detection boundary, that is, the smallest Δ_n under which the type II error converges to 0 as $n \to \infty$. Our first result establishes the convergence rate of the detection boundary for T_{MMD} in the case when $\theta = 0$. Hereafter, we abbreviate M in $\mathcal{P}(\Delta_n, \theta, M)$, $H_1(\Delta_n, \theta, M)$ and $\beta(T; \Delta_n, \theta, M)$, unless it is necessary to emphasize the dependence.

Theorem 1 *Consider testing $H_0 : P = P_0$ against $H_1(\Delta_n, 0)$ by T_{MMD}.*

(i) If $\sqrt{n}\Delta_n \to \infty$, then

$$\beta(T_{\mathrm{MMD}}; \Delta_n, 0) \to 0 \qquad \text{as} \quad n \to \infty;$$

(ii) conversely, there exists a constant $c_0 > 0$ such that

$$\varliminf_{n\to\infty} \beta(T_{\mathrm{MMD}}; c_0 n^{-1/2}, 0) > 0.$$

Theorem 1 shows that when the alternative $H_1(\Delta_n, 0)$ is considered, the detection boundary of T_{MMD} is of the order $n^{-1/2}$. It is of interest to compare the detection rate achieved by T_{MMD} with that in a parametric setting where consistent tests are available if $n\Delta_n \to \infty$. See, *e.g.*, Theorem 13.5.4 in Lehmann and Romano (2008) and the discussion leading to it. It is natural to raise the question to what extent such a gap can be entirely attributed to the fundamental difference between parametric and nonparametric testing problems. When considering χ^2 distance as the separation metric, we shall now argue that this gap actually is largely due to the sub-optimality of T_{MMD}, and the detection boundary of T_{MMD} could be significantly improved through a slight modification of the MMD.

3. Optimal tests based on moderated MMD

In this section, we introduce the moderated MMD (also referred to as M^3d) approach.

3.1 Moderated MMD test statistic

The basic idea behind MMD is to project two probability measures onto a unit ball in $\mathcal{H}(K)$ and use the distance between the two projections to measure the distance between the original probability measures. If the Radon-Nikodym derivative of P with respect to P_0 is far away from $\mathcal{H}(K)$, the distance between the two projections may not honestly reflect the distance between them. More specifically, $\gamma^2(P, P_0) = \sum_{k\geq 1} \lambda_k [\mathbb{E}_P \varphi_k(X)]^2$, while the χ^2 distance between P and P_0 is $\chi^2(P, P_0) = \sum_{k\geq 1} [\mathbb{E}_P \varphi_k(X)]^2$. Considering that λ_k decreases with k, $\gamma^2(P, P_0)$ can be much smaller than $\chi^2(P, P_0)$. To overcome this problem, we consider a moderated version of the MMD which allows us to project the probability measures onto a larger ball in $\mathcal{H}(K)$. In particular, write

$$\eta_{K,\varrho}(P, Q; P_0) = \sup_{f\in\mathcal{H}(K):\|f\|^2_{L_2(P_0)}+\varrho^2\|f\|^2_K\leq 1} \int_{\mathcal{X}} f d(P - Q) \tag{8}$$

for a given distribution P_0 and a constant $\varrho > 0$. Distance between probability measures of this type was first introduced by Harchaoui et al. (2007) when considering kernel methods for two sample test. A subtle difference between $\eta_{K,\varrho}(P, Q; P_0)$ and the distance from Harchaoui et al. (2007) is the set of f that we optimize over on the righthand side of (8). In the case of two sample test, there is no information about P_0 and therefore one needs to replace the norm $\|\cdot\|_{L_2(P_0)}$ with the empirical L_2 norm.

It is worth noting that $\eta_{K,\varrho}(P, Q; P_0)$ can also be identified with a particular type of MMD. Specifically, $\eta_{K,\varrho}(P, Q; P_0) = \gamma_{\tilde{K}_\varrho}(P, Q)$, where

$$\tilde{K}_\varrho(x, x') := \sum_{k\geq 1} \frac{\lambda_k}{\lambda_k + \varrho^2} \varphi_k(x)\varphi_k(x').$$

We shall nonetheless still refer to $\eta_{K,\varrho}(P, Q; P_0)$ as a moderated MMD in what follows to emphasize the critical importance of moderation. We shall also abbreviate the dependence of η on K and P_0 unless necessary. The unit ball in (8) is defined in terms of both RKHS norm and $L^2(P_0)$ norm. Recall that $u = dP/dP_0 - 1$ so that

$$\sup_{\|f\|_{L_2(P_0)}\leq 1} \int_{\mathcal{X}} f d(P - P_0) = \sup_{\|f\|_{L_2(P_0)}\leq 1} \int_{\mathcal{X}} f u dP_0 = \|u\|_{L_2(P_0)} = \chi(P, P_0).$$

We can therefore expect that a smaller ϱ will make $\eta^2_\varrho(P, P_0)$ closer to $\chi^2(P, P_0)$, since the unit ball to be considered will become more similar to the unit ball in $L_2(P_0)$. This can also be verified by noticing that

$$\lim_{\varrho\to\infty} \eta^2_\varrho(P, P_0) = \lim_{\varrho\to\infty} \sum_{k\geq 1} \frac{\lambda_k}{\lambda_k + \varrho^2} [\mathbb{E}_P \varphi_k(X)]^2 = \sum_{k\geq 1} [\mathbb{E}_P \varphi_k(X)]^2 = \chi^2(P, P_0).$$

Therefore, we choose ϱ converging to 0 when constructing our test statistic.

Hereafter we shall attach the subscript n to ϱ to signify its dependence on n. We shall argue that letting ρ_n converge to 0 at an appropriate rate as n increases indeed results in a test more powerful than T_{MMD}. The test statistic we propose is the empirical version of $\eta^2_{\varrho_n}(P, P_0)$:

$$\eta^2_{\varrho_n}(\widehat{P}_n, P_0) = \frac{1}{n^2}\sum_{i,j=1} \tilde{K}_{\varrho_n}(X_i, X_j) = \sum_{k\geq 1} \frac{\lambda_k}{\lambda_k + \varrho_n^2}\left[\frac{1}{n}\sum_{i=1}^{n}\varphi_k(X_i)\right]^2. \tag{9}$$

This test statistics is similar in spirit to the homogeneity test proposed previously by Harchaoui et al. (2007), albeit motivated from a different viewpoint. In either case, it is intuitive to expect improved performance over the vanilla version of the MMD when ϱ_n converges to zero at an appropriate rate. The main goal of the present work to precisely characterize the amount of moderation needed to ensure maximum power. We first argue that letting ϱ_n converge to 0 at an appropriate rate indeed results in a test more powerful than T_{MMD}.

3.2 Operating characteristics of $\eta^2_{\varrho_n}(\widehat{P}_n, P_0)$ based tests

Although the expression for $\eta^2_{\varrho_n}(\widehat{P}_n, P_0)$ given by (9) looks similar to that of $\gamma^2(\widehat{P}_n, P_0)$, their asymptotic behaviors are quite different. At a technical level, this is due to the fact that the eigenvalues of the underlying kernel

$$\tilde{\lambda}_{nk} := \frac{\lambda_k}{\lambda_k + \varrho_n^2}$$

depend on n and may not be uniformly summable over n. As presented in the following theorem, a certain type of asymptotic normality, instead of a sum of chi-squares as in the case of $\gamma^2(\widehat{P}_n, P_0)$, holds for $\eta^2_{\varrho_n}(\widehat{P}_n, P_0)$ under P_0, which helps determine the rejection region of the $\eta^2_{\varrho_n}$ based test.

Theorem 2 *Assume that $\varrho_n \to 0$ as $n \to \infty$ in such a fashion that $n\varrho_n^{1/(2s)} \to \infty$. Then under H_0 where $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_0$,*

$$v_n^{-1/2}[n\eta^2_{\varrho_n}(\widehat{P}_n, P_0) - A_n] \xrightarrow{d} N(0, 2),$$

where

$$v_n = \sum_{k\geq 1}\left(\frac{\lambda_k}{\lambda_k + \varrho_n^2}\right)^2, \qquad \text{and} \qquad A_n = \frac{1}{n}\sum_{i=1}^{n}\tilde{K}_{\varrho_n}(X_i, X_i).$$

In the light of Theorem 2, a test that rejects H_0 if and only if

$$2^{-1/2}v_n^{-1/2}[n\eta^2_{\varrho_n}(\widehat{P}_n, P_0) - A_n]$$

exceeds $z_{1-\alpha}$ is an asymptotic α-level test, where $z_{1-\alpha}$ stands for the $1-\alpha$ quantile of a standard normal distribution. We refer to this test as $T_{\mathrm{M^3d}}$where the subscript $\mathrm{M^3d}$ stands for *M*oderated *MMD*. The performance of $T_{\mathrm{M^3d}}$ under the alternative hypothesis is characterized by the following theorem, showing that its detection boundary is much improved when compared with that of T_{MMD}.

Theorem 3 *Consider testing H_0 against $H_1(\Delta_n, \theta)$ by $T_{\mathrm{M^3d}}$ with $\varrho_n = cn^{-\frac{2s(\theta+1)}{4s+\theta+1}}$ for an arbitrary constant $c > 0$. If $n^{\frac{4s}{4s+\theta+1}}\Delta_n \to \infty$, then $T_{\mathrm{M^3d}}$ is consistent in that*

$$\beta(T_{\mathrm{M^3d}}; \Delta_n, \theta) \to 0, \qquad \text{as } n \to \infty.$$

Theorem 3 indicates that the detection boundary for $T_{\mathrm{M^3d}}$ is $n^{-4s/(4s+\theta+1)}$. In particular, when testing H_0 against $H_1(\Delta_n, 0)$, *i.e.*, $\theta = 0$, it becomes $n^{-4s/(4s+1)}$. This is to be contrasted with the detection boundary for T_{MMD}, which, as suggested by Theorem 1, is of the order $n^{-1/2}$. It is also worth noting that the detection boundary for $T_{\mathrm{M^3d}}$ deteriorates as θ increases, implying that it is harder to test against a larger interpolation space.

3.3 Minimax optimality

It is of interest to investigate if the detection boundary of $T_{\mathrm{M^3d}}$ can be further improved. We now show that the answer is negative in a certain sense.

Indeed, the problem of identifying necessary condition on the decay rate of separation metric under which there exist consistent tests has been investigated under various settings. In general, both the constructive method of tackling such problem and the necessary condition depend not only on the regularity condition of the alternative space but also on the separation metric that measures the difference between null and alternative distributions. When distributions considered (in our case, Radon-Nikodym derivatives with respect to P_0) are assumed to lie in a space with an orthonormal basis, both regularity condition and separation metric can be expressed in terms of the corresponding coefficient sequences with respect to the basis. Specifically, the separation metric is usually the ℓ_r distance between coefficient sequences, and the regularity condition can be assuming that coefficient sequences lie in an ℓ_p ellipsoid. Indeed, results in the previous literatures suggest that what really contribute to the optimal rate of detection are r, p and the lengths of semi-axes of the ellipsoid. See Ingster (1987, 1993, 1995); Ermakov (1991); Lepski and Spokoiny (1999). See also Ingster and Suslina (2000) for the regularity condition being Besov bodies.

In our case, $\mathcal{F}(\theta, M)$ can be converted to certain conditions on the coefficient sequence with respect to $\{\varphi_k\}_{k\geq 1}$, and $r = p = 2$ since both separation metric and regularity condition can be expressed in terms of χ^2 distance. In particular, $\mathcal{F}(0, M)$ corresponds exactly to an ℓ_2 ellipsoid. Hence the whole procedure of identifying the necessary condition on the decay rate of Δ_n can be greatly simplified by borrowing classical arguments. Though not exactly the same, the procedure for $\theta > 0$ can be regarded as an adaption of that for $\theta = 0$.

Theorem 4 *Consider testing $H_0 : P = P_0$ against $H_1(\Delta_n, \theta)$, for some $\theta < 2s - 1$. If $\overline{\lim_{n\to\infty}} \Delta_n n^{\frac{4s}{4s+\theta+1}} < \infty$, then*

$$\underline{\lim_{n\to\infty}} \inf_{T\in\mathcal{T}_n} [\mathbb{E}_{P_0} T + \beta(T; \Delta_n, \theta)] > 0,$$

where $\mathcal{T}_n$ denotes the collection of all test functions based on $X_1, \ldots, X_n$.

Recall that for a test T, $\mathbb{E}_{P_0} T$ is its Type I error. Theorem 4 shows that, if $\Delta_n = O\left(n^{-4s/(4s+\theta+1)}\right)$, then the sum of Type I and Type II errors of any test does not vanish as n increases. In other words, there is no consistent test if $\Delta_n = O\left(n^{-4s/(4s+\theta+1)}\right)$.

Together with Theorem 3, this suggests that $T_{\mathrm{M^3d}}$ is rate optimal in the minimax sense, when considering χ^2 distance as the separation metric and $\mathcal{F}(\theta, M)$ as the regularity condition of alternative space.

3.4 Practical considerations

The moderation we applied is similar in spirit to regularization commonly seen in the context of supervised learning. And as in the case of regularization, moderation can also be done in multiple ways. Although in this work, we focus primarily on $T_{\mathrm{M^3d}}$, it is worth pointing out that similar optimality can also be expected for others forms of moderation. For example, a careful inspection of our analysis suggests an alternative form of moderation, $\breve{\eta}^2_{\varrho_n,N}(\widehat{P}_n, P_0) := \gamma_{\breve{K}_{\rho,N}}(P, Q)$ where

$$\breve{K}_{\rho,N}(x, x') := \sum_{k=1}^{N} \frac{\lambda_k}{\lambda_k + \rho^2} \varphi_k(x)\varphi_k(x').$$

Following the same argument as that of Theorem 2, it can be shown that, under H_0 where $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_0$,

$$\breve{v}_n^{-1/2} \left[n\breve{\eta}^2_{\varrho_n,N}(\widehat{P}_n, P_0) - \breve{A}_n \right] \xrightarrow{d} N(0, 2),$$

where

$$\breve{v}_n = \sum_{k=1}^{N} \left(\frac{\lambda_k}{\lambda_k + \rho_n^2} \right)^2 \quad \text{and} \quad \breve{A}_n = \frac{1}{n} \sum_{i=1}^{n} \breve{K}_{\rho_n,N}(X_i, X_i)$$

as $n \to \infty$ provided that the conditions of Theorem 2 hold and $\lambda_N \lesssim \rho_n^2$. A test that rejects H_0 when

$$2^{-1/2}\breve{v}_n^{-1/2} \left[n\breve{\eta}^2_{\varrho_n,N}(\widehat{P}_n, P_0) - \breve{A}_n \right] \tag{10}$$

exceeds $z_{1-\alpha}$ is therefore also an asymptotic α-level test. By the same argument as that of Theorem 3, this test can also be shown to be consistent whenever $n^{\frac{4s}{4s+\theta+1}} \Delta_n \to \infty$. In other words, $\breve{\eta}^2_{\varrho_n,N}(\widehat{P}_n, P_0)$ can achieve the same rate of detection as $\eta^2_{\varrho_n,N}(\widehat{P}_n, P_0)$. Based on the polynomial decay rate assumption (6), that $\lambda_N \lesssim \rho_n^2$ is equivalent to

$$N \gtrsim \varrho_n^{-1/s} \asymp n^{\frac{2(\theta+1)}{4s+\theta+1}}, \tag{11}$$

suggesting how large N needs to be in order for the truncated version of $\mathrm{M^3d}$ test to achieve minimax optimality. This test could be more appealing in practice when the infinite sum in defining $\tilde{K}_\rho$ does not have a convenient closed form expression and requires nontrivial numerical approximation. On the other hand, the presence of the extra tuning parameter N makes the exposition cumbersome at places. For brevity, we shall focus our attention on $T_{\mathrm{M^3d}}$ although all our discussion applies equally to both tests, and more generally other suitable forms of moderation.

4. Adaptation

Despite the minimax optimality of $T_{\mathrm{M^3d}}$, a practical challenge in using it is the choice of an appropriate tuning parameter ϱ_n. In particular, Theorem 3 suggests that ϱ_n needs to be taken at the order of $n^{-2s(\theta+1)/(4s+\theta+1)}$ which depends on the value of s and θ. On the one hand, since P_0 and K are known a priori, so is s. On the other hand, θ reflects the property of dP/dP_0 which is typically not known in advance. This naturally brings about the issue of adaptation (see, *e.g.*, Spokoiny, 1996; Ingster, 2000). In other words, we are interested in a single testing procedure that can achieve the detection boundary for testing H_0 against $H_1(\Delta_n(\theta), \theta)$ simultaneously over all $\theta \geq 0$. We emphasize the dependence of Δ_n on θ since the detection boundary may depend on θ, as suggested by the results from the previous section. In fact, we should build upon the test statistic introduced before.

More specifically, write

$$\rho_* = \left(\frac{\sqrt{\log\log n}}{n}\right)^{2s},$$

and

$$m_* = \left\lceil \log_2 \left[\rho_*^{-1} \left(\frac{\sqrt{\log\log n}}{n}\right)^{\frac{2s}{4s+1}} \right] \right\rceil.$$

Then our test statistics is taken to be the maximum of T_{n,ϱ_n} for $\rho_n = \rho_*, 2\rho_*, 2^2\rho_*, \ldots, 2^{m_*}\rho_*$:

$$\tilde{T}_n := \sup_{0 \leq k \leq m_*} T_{n, 2^k \varrho_*}, \tag{12}$$

where, with slight abuse of notation,

$$T_{n,\varrho_n} = (2v_n)^{-1/2}[n\eta^2_{\varrho_n}(\widehat{P}_n, P_0) - A_n].$$

It turns out if an appropriate rejection threshold is chosen, $\tilde{T}_n$ can achieve a detection boundary very similar to the one we have before, but now simultaneously over all $\theta > 0$.

Theorem 5 (i) *Under H_0,*

$$\lim_{n\to\infty} P\left(\tilde{T}_n \geq \sqrt{3\log\log n}\right) = 0;$$

(ii) *on the other hand, there exists a constant $c_1 > 0$ such that,*

$$\lim_{n\to\infty} \inf_{P \in \cup_{\theta\geq 0} \mathcal{P}(\Delta_n(\theta),\theta)} P\left(\tilde{T}_n \geq \sqrt{3\log\log n}\right) = 1,$$

provided that $\Delta_n(\theta) \geq c_1(n^{-1}\sqrt{\log\log n})^{\frac{4s}{4s+\theta+1}}$.

Theorem 5 immediately suggests that a test rejects H_0 if and only if $\tilde{T}_n \geq \sqrt{3\log\log n}$ is consistent for testing it against $H_1(\Delta_n(\theta), \theta)$ for all $\theta \geq 0$ provided that $\Delta_n(\theta) \geq c_1(n^{-1}\sqrt{\log\log n})^{\frac{4s}{4s+\theta+1}}$. It is worth noting that same detection boundary can be attained by replacing $T_{n,2^k\varrho_*}$ with the test statistic defined by (10) with $\rho_n = 2^k\varrho_*$ and $N \gtrsim 2^{-k/s}\varrho_*^{-1/s}$.

We can further calibrate the rejection region to yield a test at a given significance level. More precisely, let $\tilde{q}_{1-\alpha}$ be the $1-\alpha$ quantile of $\tilde{T}_n$ under H_0, we can proceed to reject H_0

whenever the observed test statistic exceeds $\tilde{q}_{1-\alpha}$. Denote such a test by $\tilde{T}_{\mathrm{M^3d}}$. By definition, $\tilde{T}_{\mathrm{M^3d}}$ is an α-level test. Theorem 5 implies that the type II error of $\tilde{T}_{\mathrm{M^3d}}$ vanishes as $n \to \infty$ uniformly over all $\theta \geq 0$. In practice, the quantile $\tilde{q}_{1-\alpha}$ can be evaluated by Monte Carlo methods as we shall discuss in further details in Section 6. We note that the detection boundary given in Theorem 5 is similar, but inferior by a factor of $(\log\log n)^{\frac{2s}{4s+\theta+1}}$, to that from Theorem 4. As our next result indicates such an extra factor is indeed unavoidable and is the price one needs to pay for adaptation.

Theorem 6 *Let $0 < \theta_1 < \theta_2 < 2s - 1$. Then there exists a positive constant c_2, such that if*

$$\overline{\lim_{n\to\infty}} \sup_{\theta\in[\theta_1,\theta_2]} \left\{ \Delta_n(\theta) \left(\frac{n}{\sqrt{\log\log n}} \right)^{\frac{4s}{4s+\theta+1}} \right\} \leq c_2$$

then

$$\lim_{n\to\infty} \inf_{T\in\mathcal{T}_n} \left[\mathbb{E}_{P_0} T + \sup_{\theta\in[\theta_1,\theta_2]} \beta(T; \Delta_n(\theta), \theta) \right] = 1.$$

Similar to Theorem 4, Theorem 6 shows that there is no consistent test for H_0 against $H_1(\Delta_n, \theta)$ simultaneously over all $\theta \in [\theta_1, \theta_2]$, if $\Delta_n(\theta) \leq c_2 \left(n^{-1}\sqrt{\log\log n}\right)^{\frac{4s}{4s+\theta+1}}$ $\forall\ \theta \in [\theta_1, \theta_2]$ for a sufficiently small c_2. Together with Theorem 5, this suggests that the test $\tilde{T}_{\mathrm{M^3d}}$ is indeed rate optimal.

5. A Specific Example

To better illustrate the implications of our general results, it is instructive to consider a more specific example where we are interested in testing uniformity on the unit interval $\mathcal{X} = [0, 1]$ using a periodic spline kernel. Recall that the periodic Sobolev space of order s is given by

$$W_0^s([0,1]) := \left\{ u \in L^2([0,1]) : \sum_{m=0}^{s} \int_0^1 \left(u^{(m)}(x)\right)^2 dx < \infty,\ \int_0^1 u^{(m)}(x) = 0,\ \forall\ 0 \leq m < s \right\}.$$

When endowed with the inner product

$$\langle u_1, u_2 \rangle_{W_0^s([0,1])} = \int_0^1 u_1^{(s)}(x) u_2^{(s)}(x) dx, \qquad \forall u_1, u_2 \in W_0^s([0,1])$$

$W_0^s([0,1])$ forms a RKHS with reproducing kernel

$$K(x, x') = \frac{(-1)^{s-1}}{(2s)!} B_{2s}([x - x'])$$

where B_r is the Bernoulli polynomial and $[t]$ is the fractional part of t. See, *e.g.*, Wahba (1990), for further details.

In this case, the interpolation spaces $\mathcal{F}(\theta, M)$ is closely related to Sobolev spaces of lower order. Recall that the eigenvalues and eigenfunctions of K with respect to $L_2(P_0)$ are also known explicitly:

$$\varphi_k(x) = \begin{cases} \sqrt{2}\cos(2j\pi x), & k = 2j-1,\ j \in \mathbb{N} \\ \sqrt{2}\sin(2j\pi x), & k = 2j,\ j \in \mathbb{N} \end{cases},$$

and $\lambda_k = (2\pi j)^{-2s}$ where $k = 2j-1$ or $2j$. It is clear that $u \in \mathcal{F}(0, M)$ is equivalent to $\|u\|_{W_0^s([0,1])} \leq M$. More generally, there exists a $M'(M, s, \theta) > 0$ such that $\|u\|_{W_0^{s'}([0,1])} \leq M'$ implies that $u \in \mathcal{F}(\theta, M)$ where $s' = s/(1+\theta)$. In other words, for any $s' \leq s$, the ball in $W_0^{s'}([0,1])$ with an appropriate radius is contained in $\mathcal{F}(s/s'-1, M)$.

The problem of minimax hypothesis testing of uniformity against a deviation from Sobolev spaces is well studied. See, *e.g.*, Ingster and Suslina (2003). Specifically, consider the goodness-of-fit test with P_0 being uniform distribution on $[0,1]$ and alternative hypothesis

$$\{P : \|dP/dP_0 - 1\|_{W_0^s([0,1])} \leq M,\ \chi^2(P, P_0) \geq \Delta_n\}.$$

It is well-known that the optimal rate of detection is $n^{-4s/(4s+1)}$ when χ^2 distance is considered. See, *e.g.*, Ingster (1993). On the other hand, since the alternative hypothesis is exactly $\mathcal{P}(\Delta_n, 0)$, such detection boundary can also be derived from Theorem 3 and Theorem 4, and the proposed M^{3}d test can achieve minimax optimality in this specific example.

6. Numerical Experiments

To complement the earlier theoretical development, we also performed several sets of simulation experiments to demonstrate the merits of the proposed adaptive test. As mentioned before, we shall consider the test based on $\breve{\eta}^2_{\varrho_n,N}(\widehat{P}_n, P_0)$ instead of $\eta^2_{\varrho_n}(\widehat{P}_n, P_0)$ for the ease of practical implementation. With slight abuse of notation, we shall still refer to the adaptive test as a test based on moderated MMD and denote it by $\tilde{T}_n$ for brevity.

Though the form of $\breve{\eta}^2_{\varrho_n,N}(\widehat{P}_n, P_0)$ looks similar to that of $\gamma^2(\widehat{P}_n, P_0)$, from the point of view of computing it numerically, there is a subtle issue. The kernel $\breve{K}_{\varrho_n,N}(x, x')$ is defined only in its Mercer decomposed form, which is based on the Mercer decomposition of $K(x, x')$. Hence, in order to compute the kernel $\breve{K}_{\varrho_n,N}(x, x')$, we need to first choose a kernel $K(x, x')$ and compute its Mercer decomposition numerically. Specifically, we use *chebfun* framework in Matlab (with slight modifications) to compute Mercer decompositions associated with kernels based on their integral operator representations (Driscoll et al., 2014; Trefethen and Battles, 2004).

6.1 Simulation Studies

We first conducted two simulation studies. One was with Euclidean data and the other with directional data. In both cases, the number of eigenvalues used for kernel approximation (denoted by N) is given in Table 1. We assessed the null distribution of $\tilde{T}_n$ by simulating it under the null hypothesis H_0. In particular, we repeated for each case 200 runs and estimated the 95% quantile of $\tilde{T}_n$ under H_0 by the corresponding sample quantile. We then proceeded to reject H_0 when an observed test statistic exceeds the estimate 95% quantile.

Sample size	200	400	600	800	1000
N	15	22	25	28	36

Table 1: Number of eigenvalues (N) used for the simulation experiments.

By construction, the procedure gave a 5%-level test, up to Monte Carlo error. For all other tests that we compared with, we used the same approach to determine the rejection threshold.

Similar idea of using resampling method to decide the rejection threshold has been considered in Fromont et al. (2012), Fromont et al. (2013), where they used bootstrap approaches to resample in a two sample problem. To the best of our knowledge, it is proved there the resulted single kernel tests are exactly of α level. Whether we can give theoretical justification of the resampling method applied to the adaptive test is of stong interest in our future work.

Euclidean data: Consider the one sample test where P_0 is the uniform distribution on $[0,1]^d$, with $d = 100$ and 200. We followed the examples for densities put forward in Marron and Wand (1992) in the context of nonparametric density estimation, for the alternatives. Specifically we set the alternative hypothesis to be (1) mixture of five Gaussians, (2) skewed unimodal, (3) asymmetric claw density and (4) smooth comb density. The value of α was set to 0.05. The sample size n varied from 200 to 1000 (in steps of 200) and for each value of sample size 100 simulations were conducted to estimate the probability of rejecting H_0.

We chose Gaussian kernel as the original kernel to construct the adaptive test. And for practical purpose, the bandwidth of Gaussian kernel was selected via median heuristic. See, *e.g.*, Gretton et al. (2012a). We now describe the specific tests that we experimented with, among which all other MMD based tests considered Gaussian kernel as well.

1. *M^3D*: Adaptive M^{3}d test based on $\breve{\eta}^2_{\varrho_n,N}(\widehat{P}_n, P_0)$.
2. *MMD1*: Vanilla MMD with the median heuristic for selecting the bandwidth.
3. *MMD2*: Sup-MMD as outlined in Sriperumbudur et al. (2009).
4. *MMD3*: The method proposed in Sutherland et al. (2017) to select the kernel that maximizes the power.
5. *KS*: Kolmogorov-Smirnov test, a classical test for goodness of fit testing.

We remark that, for MMD2 and MMD3, we split the whole dataset into two parts. We use the training dataset to first select the bandwidth and then use the testing samples to perform the actual hypothesis testing. See Sutherland et al. (2017) for more details.

We first conducted a type I error analysis. Results in Figure 1 suggest that all tests control the probability of type I error at 0.05 approximately. Then we investigated the performances of these tests under the alternative hypothesis. Figure 2 illustrates a plot of the estimated probability of accepting the null hypothesis under alternatives mentioned above for different values of sample size n . We note from Figure 2 that the estimated probability of type II error converges to zero at a faster rate for M^3D compared to other tests

on all the different simulation settings that are considered. Note that it has been previously observed that MMD test performs better than Kolmogorov-Smirnov test in various setting in Gretton et al. (2012a), which we observe in our setting as well.

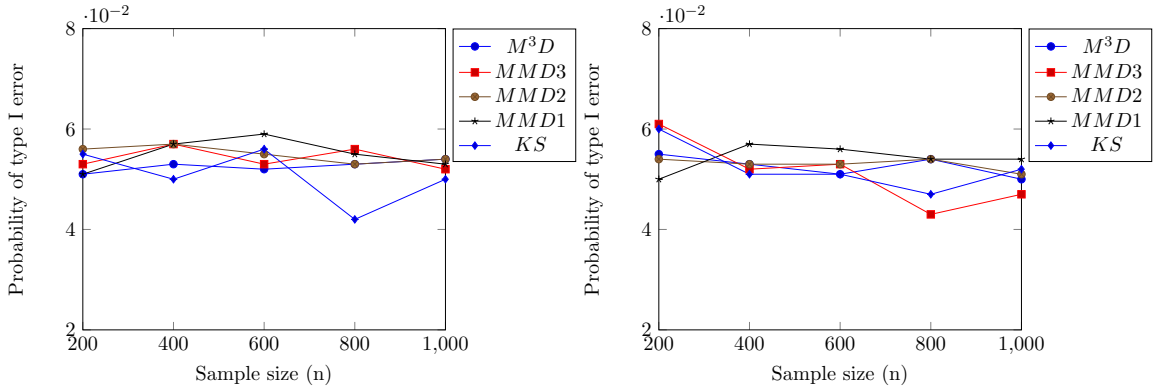

Figure 1: Estimated probability of type I error versus sample size for different tests with dimensionality 100 (left) and 200 (right), in the case of Euclidean data.

Directional data: One of the advantages of the proposed RKHS embedding based approach is that it could be used on domains other than the d-dimensional Euclidean space. For example when $\mathcal{X} = \mathbb{S}^{d-1}$ where $\mathbb{S}^{d-1}$ corresponds to the d-dimensional unit sphere, one can perform hypothesis testing using the above framework, as long as we can compute the Mercer decomposition of a kernel K defined on the domain. In several applications, like protein folding, often times data are modeled as coming from the unit-sphere and testing goodness-of-fit for such data needs specialized methods different from the standard non-parametric testing methods (Mardia and Jupp, 2009; Jupp, 2005).

In order to highlight the advantage of the proposed approach, we assumed P_0 to be the uniform distribution on the unit sphere of dimension $d = 100$ and 150 respectively. For each d, we tested P_0 against the alternative that data were from:

(1) multivariate von Mises-Fisher distribution (which is the Gaussian analogue on the unit-sphere) given by $f_{vM\text{-}F}(x, \mu, \kappa) = C_{vM\text{-}F}(\kappa) \exp(\kappa \mu^\top x)$ for data $x \in \mathbb{S}^{d-1}$, where $\kappa \geq 0$ is concentration parameter and μ is the mean parameter. The term $C_{vM\text{-}F}$ is the normalization constant given by $\frac{\kappa^{d/2-1}}{2\pi^{d/2} I_{d/2-1}(\kappa)}$ where I is modified Bessel function;

(2) multivariate Watson distribution (used to model axially symmetric data on sphere) given by $f_W(x, \mu, \kappa) = C_W(\kappa) \exp(\kappa(\mu^\top x)^2)$ for data $x \in \mathbb{S}^{d-1}$, where $\kappa \geq 0$ is concentration parameter and μ is the mean parameter as before. The term $C_W(\kappa)$ is the normalization constant given by $\frac{\Gamma(d/2)}{2\pi^{p/2} M(1/2, d/2, \kappa)}$ where M is Kummer's confluent hypergeometric function;

(3) mixture of five von Mises-Fisher distribution which are used in modeling and clustering spherical data (Banerjee et al., 2005);

(4) mixture of five Watson distribution which are used in modeling and clustering spherical data (Sra and Karp, 2013).

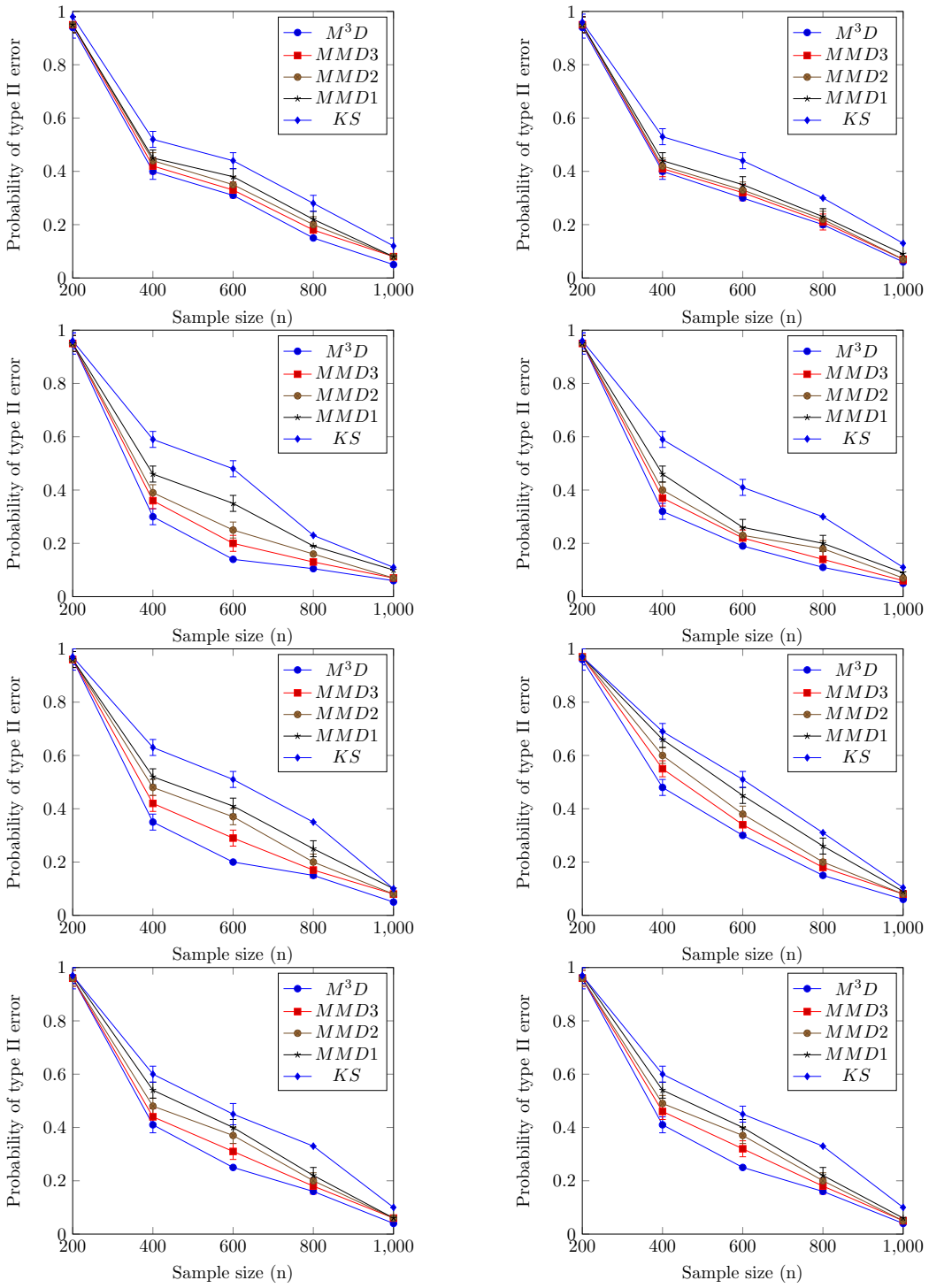

Figure 2: Estimated probability of type II error versus sample size: mixture of Gaussian (row 1), skewed unimodal (row 2), asymmetric claw (row 3) and smooth comb (row 4) with dimensionality 100 (left) and 200 (right), in the case of Euclidean data.

Besides the adaptive M^{3}d test and all other MMD based tests that involved in the first simulation study, we also considered the Sobolev test approach (denoted as ST hereafter) proposed in Jupp (2005). The original kernel of M^3D and the kernel of $MMD1$ were again chosen as Gaussian kernel with bandwidth selected via median heuristic. Other MMD based tests also used Gaussian kernel. Note that in this setup one can analytically compute the Mercer decomposition of the Gaussian kernel on the unit sphere with respect to the uniform distribution. Specifically, the eigenvalues are given by Theorem 2 in Minh et al. (2006) and the eigenfunctions are the standard spherical harmonics of order k (see section 2.1 in Minh et al. (2006) for details). Rest of the simulation setup is similar to the previous setting (of Euclidean data).

The results of type I error analysis are reported in Figure 3, indicating all tests are approximately α-level tests. Figure 4 illustrates a plot of the estimated probability of type II error for different values of sample size, from which we see the adaptive M^{3}d test performs better.

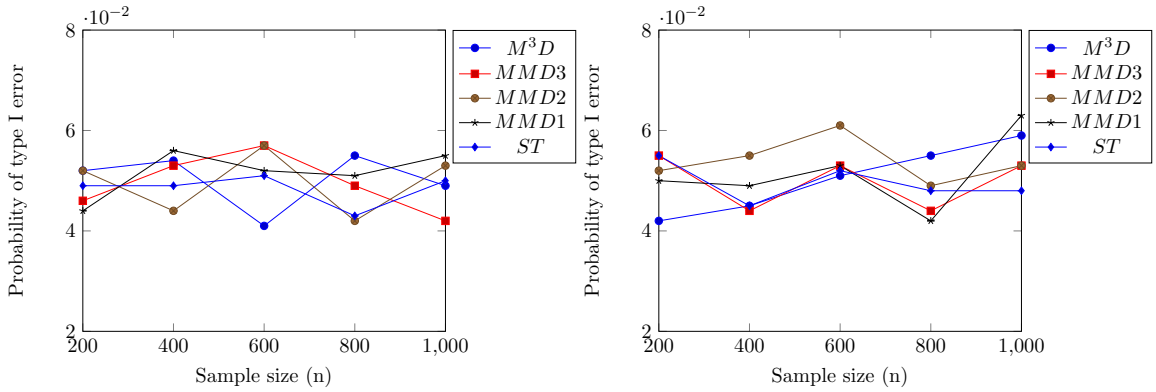

Figure 3: Estimated probability of type I error versus sample size for different tests with dimensionality 100 (left) and 150 (right), in the case of directional data.

6.2 Real Data Experiments

In addition to the simulation examples, we also performed experiments on several real-world data examples. All tests considered were the same as those in the simulation studies in both the cases of Euclidean data and directional data. Furthermore, similar to the simulation experiments, we used the Monte Carlo quantiles for all tests. Gaussian kernels were considered and the median heuristic was used as before. The number of eigenvalues used in kernel approximation is reported along with the main results. Different from the simulation examples, in the real data setting, the null hypothesis P_0 is not known. Indeed, we are testing goodness of fit to a simple null hypothesis P_0, rather than a composite family of densities. Hence, ideally one must construct tests that take into account the estimated P_0. Deriving the asymptotics of kernel-embedding based test and other classical tests, in particular establishing optimality and adaptivity properties, in this setting is beyond the scope of the current draft. As a compromise, we first use 50% of the total dataset first to get an estimate of the parameter of P_0. We then use the rest of the data to do the goodness of fit test to the simple hypothesis P_0 defined with the estimated parameter. In doing so, we

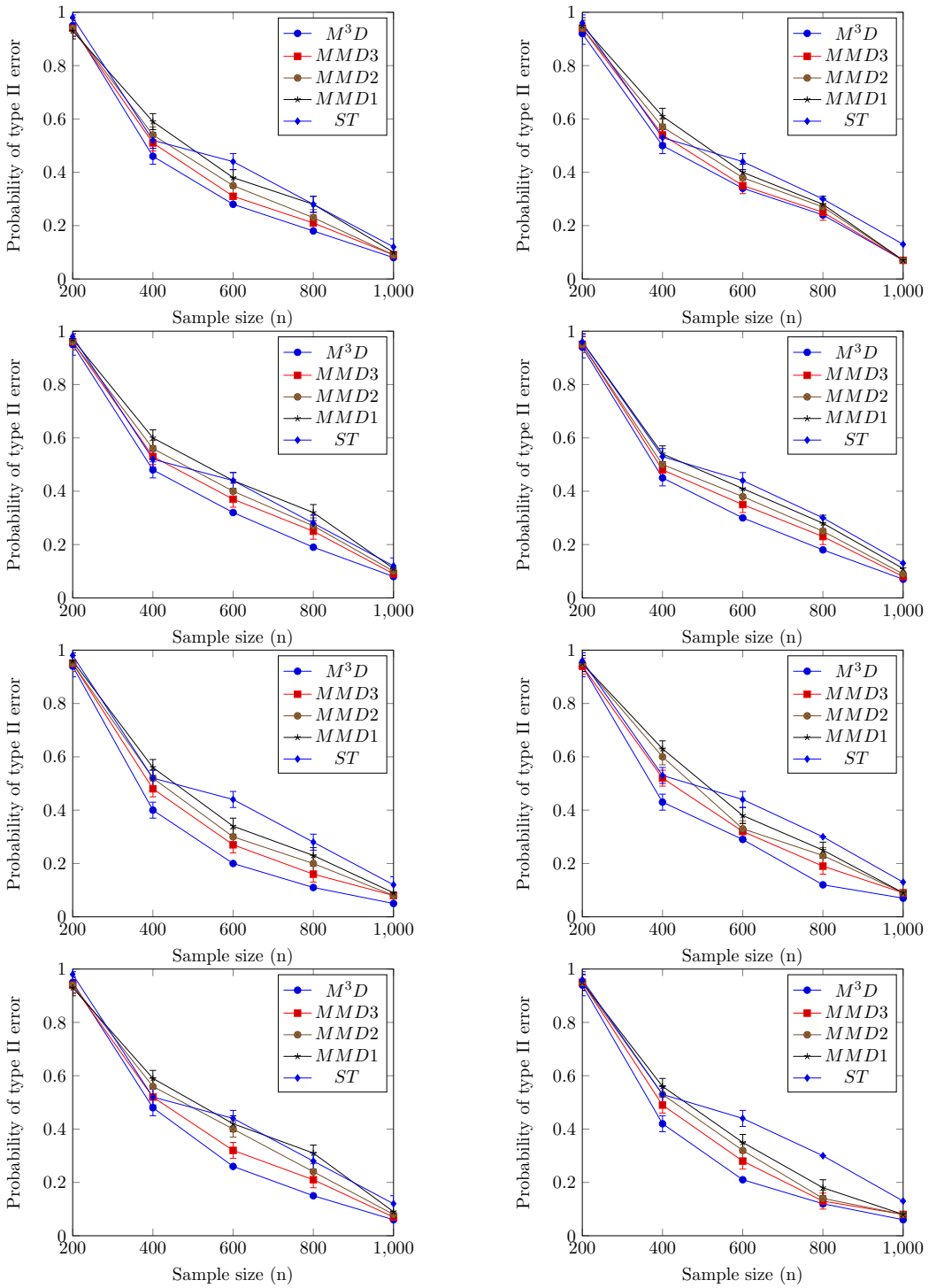

Figure 4: Estimated probability of type II error versus sample size: von Mises-Fisher distribution (row 1), Watson distribution (row 2), mixture of von Mises-Fisher distribution (row 3) and mixture of Watson distribution (row 4) with dimensionality 100 (left) and 150 (right), in the case of directional data.

are essentially treating P_0 with the estimated parameters as being given to us by an oracle and carry out goodness of fit testing. We take the above approach, as our purpose in this section is mainly to compare the adaptive M^{3}d test against the other tests for a simple P_0 and not to address the issue of testing with composite null hypothesis. Note that a similar approach was also performed in Kellner and Celisse (2015) for testing for Gaussianity with kernel-based testing.

For the case of Euclidean data, we used the MNIST digits data set from the following webpage: `http://yann.lecun.com/exdb/mnist/`. Model-based clustering (Fraley and Raftery, 2002) is a widely-used and practical successful clustering technique in the literature. Furthermore, the MNIST data set is a standard data set for testing clustering algorithms and consists of image of digits. Several works have implicitly assumed that the data come from a mixture of Gaussian distributions, because of the observed superior empirical performance under such an assumption. But the validity of such a mixture model assumption is invariably not tested statistically. In this experiment we selected three digits (which correspond to a cluster) randomly and conditioned on the selected digit (cluster), we tested if the data come from a Gaussian distribution (that is, P_0 is Gaussian). For our experiments, we down sampled the images and use pixels as feature vectors with dimensionality 64 as is commonly done in the literature. Table 2 reports the probability with which the null hypothesis is accepted. Such probability is estimated through 100 trials. The observed result reiterates in a statistically significant way that it is reasonable to make a mixture of Gaussian assumption in this case.

For the case of directional data, we used the Human Fibroblasts dataset from Iyer et al. (1999); Dhillon et al. (2003) and the Yeast Cell Cycle dataset from Spellman et al. (1998). The Fibroblast data set contains 12 expression data corresponding to 517 samples (genes) report in the response of human fibroblasts following addition of serum to the growth media. We refer to Iyer et al. (1999) for more details about the scientific procedure with which these data were obtained. The Yeast Cell Cycle dataset consists of 82-dimensional data corresponding to 696 subjects. Previous data analysis studies (Sra and Karp, 2013; Dhillon et al., 2003) have used mixtures of spherical distributions for clustering the above data set. Specifically, it has been observed in Sra and Karp (2013) that clustering using a mixture of Watson distribution has superior performance. While that has proved to be useful scientifically, it was not statistically tested if such an assumption is valid. Here, we conducted goodness of fit test of Watson distribution (that is, P_0 is a Watson distribution) for the largest cluster from the above data sets. Table 3 shows the estimated probability of acceptance of the null hypothesis, which is computed through 100 random trials. The observed results provide a statistical justification for the use of Watson distribution in modeling the above datasets.

One thing to comment is that wo do not aim to argue that our proposed test outperforms the other considered tests in the above real-world experiments. The information conveyed by the results rather lies in the following two aspects. First, if we assume that H_0 is true, then our proposed test is valid in the sense that the estimated probability of Type I error is controlled around 5%, although two approximations are involved in the whole testing procedure. One is using one half of data to estimate P_0 and the other is using Monte Carlo simulations to determine the rejection region. Second, that H_0 is accepted with high probability by our proposed test and other involved tests does provide certain evidence that

Sample size =	300	400	500	300	400	500	300	400	500
KS	0.91	0.94	0.96	0.91	0.91	0.95	0.91	0.92	0.95
$MMD1$	0.92	0.95	0.96	0.91	0.92	0.95	0.90	0.92	0.94
$MMD2$	0.93	0.95	0.96	0.92	0.93	0.95	0.93	0.94	0.96
$MMD3$	0.92	0.95	0.97	0.90	0.94	0.97	0.91	0.92	0.96
M^3D	0.94	0.96	0.98	0.93	0.95	0.98	0.93	0.95	0.98
N	32	36	40	32	36	40	32	36	40

Table 2: Estimated probability with which the corresponding test accepts the null hypothesis. The level of the test $\alpha = 0.05$. Digit 4 on the left, Digit 6 in the middle and Digit 7 on the right,for various values of sample size. N refers to the number of eigenvalues/eigenfunctions used in kernel approximation.

Sample size=	75	150	200	150	200	250
ST	0.90	0.93	0.98	0.84	0.89	0.91
$MMD1$	0.91	0.94	0.98	0.85	0.91	0.93
$MMD2$	0.91	0.94	0.98	0.86	0.90	0.94
$MMD3$	0.92	0.96	0.98	0.87	0.91	0.96
M^3D	0.92	0.96	0.99	0.88	0.93	0.96
N	16	19	22	20	23	27

Table 3: Estimated probability with which the corresponding test accepts the null hypothesis. The level of the test $\alpha = 0.05$. Human Fibroblasts dataset on the left and Yeast Cell Cycle dataset on the right. N refers to the number of eigenvalues/eigenfunctions used in kernel approximation.

H_0 is true, considering that all these tests tend to have small probabilities of accepting H_0 under the alternative hypothesis.

In this sense, these real-word experiments are more like a beginning example to show the validity of the proposed test. And we may seek some other examples to demonstrate the optimality of the proposed test in our following work.

7. Concluding Remarks

In this paper, we investigated the performance of kernel embedding based approaches to goodness-of-fit testing from a minimax perspective. When considering χ^2 distance as the separation metric, we showed that while the vanilla MMD tests could be suboptimal when testing against an alternative coming from the RKHS identified with the kernel, a simple moderation leads to tests that are not only optimal when the alternative resides in the RKHS but also adaptive with respect to its deviation from the RKHS under suitable distances. Our analysis provides new insights into the operating characteristics of as well as a benchmark for evaluating the popular kernel based approach to hypothesis testing. Our work also points to a number interesting directions that warrant further investigation.

Our work highlighted the importance of moderation in kernel based testing, akin to regularization in kernel based supervised learning. The specific type of moderation we considered here is defined in terms of eigenfunctions of the kernel. In some settings this is convenient to do as we illustrated in Section 5. In other settings, however, eigenfunctions of a kernel may not be trivial to compute. It is therefore of interest to investigate alternative ways of moderating the MMD. For example, one intriguing possibility is to apply appropriate moderation to the so-called random Fourier features (see, *e.g.*, Chwialkowski et al., 2015; Jitkrittum et al., 2016; Bach, 2017) instead of eigenfunctions.

Another practical issue is how to devise computationally more efficient goodness-of-tests for dealing with large-scale datasets. For example, it would be of interest to investigate if the ideas of linear-time approximation (Gretton et al., 2012a) or block test (Zaremba et al., 2013) can be applied to yield minimax optimal and adaptive goodness of fit tests.

At a more technical level, our analysis has focused on detection boundaries under χ^2 distance. There are other commonly used distances suitable for goodness-of-fit tests such as KL-divergence, or total variation. It would be interesting to examine to what extent the phenomena we observed here occur under theses alternative distance measures.

As in any kernel based approaches, the choice of the kernel plays a crucial role. In the current work we have assumed implicitly that a suitable kernel is selected. How to choose an appropriate kernel is a critical problem in practice. It would be of great interest to investigate to what extent the ideas from Sriperumbudur et al. (2009); Gretton et al. (2012b); Sutherland et al. (2017) can be adapted in the current setting.

Throughtout numerical experiments in our paper, we have considered truncated version of the moderated kernel. Both ϱ_n^2 and truncation level N play the role of regularization and we can show that as long as N increases at a proper rate with n, even if ϱ_n in $\breve{K}_{\rho_n,N}$ is set to 0, the corresponding test can still achieve optimal rate. Note that in this scenario, $\breve{K}_{0,N}$ becomes the so-called projection kernel and MMD^2 associated with $\breve{K}_{0,N}$ is exactly the projection of $\chi^2(P,P_0)$ onto the subspace spanned by $\{\varphi_k\}_{1\le k\le N}$.

8. Proofs

We now present the proofs of the main results.

Proof [Proof of Theorem 1]

Part (i). The proof of the first part consists of two key steps. First, we show that the population counterpart $n\gamma^2(P,P_0)$ of the test statistic converges to ∞ uniformly, *i.e.*,

$$n \inf_{P\in\mathcal{P}(\Delta_n,0)} \gamma^2(P,P_0) \to \infty \tag{13}$$

as $n\to\infty$. Then, we argue that the deviation from $\gamma^2(P,P_0)$ to $\gamma^2(\widehat{P}_n,P_0)$ is uniformly negligible compared with $\gamma^2(P,P_0)$ itself.

Let $u = dP/dP_0 - 1$ and

$$a_k = \langle u,\varphi_k\rangle_{L_2(P_0)} = \mathbb{E}_P\varphi_k(X) - \mathbb{E}_{P_0}\varphi_k(X) = \mathbb{E}_P\varphi_k(X).$$

Then

$$\sum_{k\ge1}\lambda_k^{-1}a_k^2 = \|u\|_K^2, \qquad \text{and} \qquad \sum_{k\ge1}a_k^2 = \|u\|_{L_2(P_0)}^2 = \chi^2(P,P_0).$$

By the definition of $\mathcal{P}(\Delta_n, 0)$,

$$\sup_{P\in\mathcal{P}(\Delta_n,0)} \sum_{k\geq 1} \lambda_k^{-1} a_k^2 \leq M^2, \qquad \text{and} \qquad \inf_{P\in\mathcal{P}(\Delta_n,0)} \sum_{k\geq 1} a_k^2 \geq \Delta_n.$$

Since $n\Delta_n^2 \to \infty$ as $n \to \infty$, we get

$$\begin{aligned}
\inf_{P\in\mathcal{P}(\Delta_n,0)} n \sum_{k\geq 1} \lambda_k [\mathbb{E}_P \varphi_k(X)]^2 &= \inf_{P\in\mathcal{P}(\Delta_n,0)} n \sum_{k\geq 1} \lambda_k a_k^2 \\
&\geq \inf_{P\in\mathcal{P}(\Delta_n,0)} n \frac{\Big(\sum\limits_{k\geq 1} a_k^2\Big)^2}{\sum\limits_{k\geq 1} \lambda_k^{-1} a_k^2} \\
&\geq \frac{n\Delta_n^2}{M^2} \to \infty
\end{aligned}$$

as $n \to \infty$, verifying (13).

On the other hand,

$$\begin{aligned}
\gamma(\widehat{P}_n, P_0) =& \sqrt{\sum_{k\geq 1} \lambda_k \Big[\frac{1}{n}\sum_{i=1}^n \varphi_k(X_i)\Big]^2} \\
\geq& \sqrt{\sum_{k\geq 1} \lambda_k [\mathbb{E}_P \varphi_k(X)]^2} - \sqrt{\sum_{k\geq 1} \lambda_k \Big[\frac{1}{n}\sum_{i=1}^n \varphi_k(X_i) - \mathbb{E}_P \varphi_k(X)\Big]^2}.
\end{aligned}$$

Thus,

$$\begin{aligned}
& P\left\{n\gamma^2(\widehat{P}_n, P_0) < q_{w,1-\alpha}\right\} \\
\leq & P\left\{\sqrt{n\sum_{k\geq 1} \lambda_k [\mathbb{E}_P \varphi_k(X)]^2} - \sqrt{n\sum_{k\geq 1} \lambda_k \Big[\frac{1}{n}\sum_{i=1}^n \varphi_k(X_i) - \mathbb{E}_P \varphi_k(X)\Big]^2} < \sqrt{q_{w,1-\alpha}}\right\} \\
= & P\left\{\sqrt{n\sum_{k\geq 1} \lambda_k \Big[\frac{1}{n}\sum_{i=1}^n \varphi_k(X_i) - \mathbb{E}_P \varphi_k(X)\Big]^2} > \sqrt{n\sum_{k\geq 1} \lambda_k [\mathbb{E}_P \varphi_k(X)]^2} - \sqrt{q_{w,1-\alpha}}\right\}.
\end{aligned}$$

Note that (13) ensures for sufficiently lagre n,

$$n\sum_{k\geq 1} \lambda_k [\mathbb{E}_P \varphi_k(X)]^2 > q_{w,1-\alpha}, \quad \forall\ P \in \mathcal{P}(\Delta_n, 0),$$

which together with Markov inequality gives

$$P\left\{n\gamma^2(\widehat{P}_n, P_0) < q_{w,1-\alpha}\right\} \leq \frac{\mathbb{E}_P\Big\{n\sum\limits_{k\geq 1} \lambda_k \Big[\frac{1}{n}\sum\limits_{i=1}^n \varphi_k(X_i) - \mathbb{E}_P \varphi_k(X)\Big]^2\Big\}}{\left\{\sqrt{n\sum\limits_{k\geq 1} \lambda_k [\mathbb{E}_P \varphi_k(X)]^2} - \sqrt{q_{w,1-\alpha}}\right\}^2}, \quad \forall\ P \in \mathcal{P}(\Delta_n, 0).$$

Observe that for any $P \in \mathcal{P}(\Delta_n, 0)$,

$$\begin{aligned}\mathbb{E}_P\Big\{n\sum_{k\geq 1}\lambda_k\Big[\frac{1}{n}\sum_{i=1}^n \varphi_k(X_i) - \mathbb{E}_P\varphi_k(X)\Big]^2\Big\} &= \sum_{k\geq 1}\lambda_k \mathrm{Var}[\varphi_k(X)] \\ &\leq \sum_{k\geq 1}\lambda_k \mathbb{E}_P\varphi_k^2(X) \\ &\leq \Big(\sup_{k\geq 1}\|\varphi_k\|_\infty\Big)^2 \sum_{k\geq 1}\lambda_k < \infty.\end{aligned}$$

Hence we obtain

$$\begin{aligned}\lim_{n\to\infty}\beta(T_{\mathrm{MMD}};\Delta_n,0) &= \lim_{n\to\infty}\sup_{P\in\mathcal{P}(\Delta_n,0)} P\Big\{n\gamma^2(\widehat{P}_n,P_0) < q_{w,1-\alpha}\Big\} \\ &\leq \lim_{n\to\infty}\frac{\sup_{P\in\mathcal{P}(\Delta_n,0)}\mathbb{E}_P\Big\{n\sum\limits_{k\geq 1}\lambda_k\Big[\frac{1}{n}\sum\limits_{i=1}^n\varphi_k(X_i)-\mathbb{E}_P\varphi_k(X)\Big]^2\Big\}}{\inf_{P\in\mathcal{P}(\Delta_n,0)}\Big\{\sqrt{n\sum\limits_{k\geq 1}\lambda_k[\mathbb{E}_P\varphi_k(X)]^2}-\sqrt{q_{w,1-\alpha}}\Big\}^2} \\ &= 0.\end{aligned}$$

Part (ii). In proving the second part, we will make use of the following lemma that can be obtained by adapting the argument in Gregory (1977). It gives the limit distribution of V-statistic under P_n such that P_n converges to P_0 in the order $n^{-1/2}$.

Lemma 7 *Consider a sequence of probability measures $\{P_n : n \geq 1\}$ contiguous to P_0 satisfying $u_n = dP_n/dP_0 - 1 \to 0$ in $L^2(P_0)$. Suppose that for any fixed k,*

$$\lim_{n\to\infty}\sqrt{n}\langle u_n,\varphi_k\rangle_{L^2(P_0)} = \tilde{a}_k, \qquad \text{and} \qquad \lim_{n\to\infty}\sum_{k\geq 1}\lambda_k(\sqrt{n}\langle u_n,\varphi_k\rangle_{L^2(P_0)})^2 = \sum_{k\geq 1}\lambda_k\tilde{a}_k^2+\tilde{a}_0 < \infty,$$

for some sequence $\{\tilde{a}_k : k \geq 0\}$, then

$$\frac{1}{n}\sum_{k\geq 1}\lambda_k\Big[\sum_{i=1}^n\varphi_k(X_i)\Big]^2 \xrightarrow{d} \sum_{k\geq 1}\lambda_k(Z_k+\tilde{a}_k)^2+\tilde{a}_0,$$

where $X_1,\ldots,X_n \overset{i.i.d}{\sim} P_n$, and Z_ks are independent standard normal random variables.

Write $L(k) = \lambda_k k^{2s}$. By assumption (6),

$$0 < \underline{L} := \inf_{k\geq 1} L(k) \leq \sup_{k\geq 1} L(k) := \overline{L} < \infty.$$

Consider a sequence of $\{P_n : n \geq 1\}$ such that

$$dP_n/dP_0 - 1 = C_1\sqrt{\lambda_{k_n}}[L(k_n)]^{-1}\varphi_{k_n},$$

where C_1 is a positive constant and $k_n = \lfloor C_2 n^{\frac{1}{4s}} \rfloor$ for some positive constant C_2. Both C_1 and C_2 will be determined later. Since $\sup_{k\geq 1} \|\varphi_k\|_\infty < \infty$ and $\lim_{k\to\infty} \lambda_k = 0$, there exists $N_0 > 0$ such that P_n's are well-defined probability measures for any $n \geq N_0$.

Note that

$$\|u_n\|_K^2 = \frac{C_1^2}{L^2(k_n)} \leq \underline{L}^{-2} C_1^2$$

and

$$\|u_n\|_{L^2(P_0)}^2 = \frac{C_1^2 \lambda_{k_n}}{L^2(k_n)} = \frac{C_1^2}{L(k_n)} k_n^{-2s} \geq \overline{L}^{-1} C_1^2 k_n^{-2s} \sim \overline{L}^{-1} C_1^2 C_2^{-2s} n^{-1/2},$$

where $A_n \sim B_n$ means that $\lim_{n\to\infty} A_n/B_n = 1$. Thus, by choosing C_1 sufficiently small and $c_0 = \frac{1}{2}\overline{L}^{-1} C_1^2 C_2^{-2s}$, we ensure that $P_n \in \mathcal{P}(c_0 n^{-1/2}, 0)$ for sufficiently large n.

To apply Lemma 7, we note that

$$\lim_{n\to\infty} \|u_n\|_{L^2(P_0)}^2 = \lim_{n\to\infty} \frac{C_1^2 \lambda_{k_n}}{L^2(k_n)} = 0.$$

In addition, for any fixed k,

$$\tilde{a}_{n,k} = \sqrt{n} \langle u_n, \varphi_k \rangle_{L^2(P_0)} = 0$$

for sufficiently large n, and

$$\sum_{k\geq 1} \lambda_k \tilde{a}_{n,k}^2 = \frac{n C_1^2 \lambda_{k_n}^2}{L^2(k_n)} = n C_1^2 k_n^{-4s} \to C_1^2 C_2^{-4s}$$

as $n \to \infty$. Thus, Lemma 7 implies that

$$n\gamma(\widehat{P}_n, P_0) \xrightarrow{d} \sum_{k\geq 1} \lambda_k Z_k^2 + C_1^2 C_2^{-4s}.$$

Now take $C_2 = \left(2C_1^2/q_{w,1-\alpha}\right)^{1/4s}$ so that $C_1^2 C_2^{-4s} = \frac{1}{2} q_{w,1-\alpha}$. Then

$$\begin{aligned}\underline{\lim_{n\to\infty}} \beta(T_{\mathrm{MMD}}; c_0 n^{-1/2}, 0) &\geq \lim_{n\to\infty} P_n(n\gamma(\widehat{P}_n, P_0) < q_{w,1-\alpha}) \\ &= P\Big(\sum_{k\geq 1} \lambda_k Z_k^2 < \frac{1}{2} q_{w,1-\alpha}\Big) > 0,\end{aligned}$$

which concludes the proof. ■

Proof [Proof of Theorem 2] Let $\tilde{K}_n(\cdot,\cdot) := \tilde{K}_{\varrho_n}(\cdot,\cdot)$. Note that

$$v_n^{-1/2}[n\eta_{\varrho_n}^2(\widehat{P}_n, P_0) - A_n] = 2(n^2 v_n)^{-1/2} \sum_{j=2}^{n} \sum_{i=1}^{j-1} \tilde{K}_n(X_i, X_j).$$

Let $\zeta_{nj} = \sum_{i=1}^{j-1} \tilde{K}_n(X_i, X_j)$. Consider a filtration $\{\mathcal{F}_j : j \geq 1\}$ where $\mathcal{F}_j = \sigma\{X_i : 1 \leq i \leq j\}$. Due to the assumption that K is degenerate, we have $\mathbb{E}\varphi_k(X) = 0$ for any $k \geq 1$, which implies that

$$\mathbb{E}(\zeta_{nj}|\mathcal{F}_{j-1}) = \sum_{i=1}^{j-1} \mathbb{E}[\tilde{K}_n(X_i, X_j)|\mathcal{F}_{j-1}] = \sum_{i=1}^{j-1} \mathbb{E}[\tilde{K}_n(X_i, X_j)|X_i] = 0,$$

for any $j \geq 2$.

Write

$$U_{nm} = \begin{cases} 0 & m = 1 \\ \sum_{j=2}^{m} \zeta_{nj} & m \geq 2 \end{cases}.$$

Then for any fixed n, $\{U_{nm}\}_{m\geq 1}$ is a martingale with respect to $\{\mathcal{F}_m : m \geq 1\}$ and

$$v_n^{-1/2}[n\eta_{\varrho_n}^2(\widehat{P}_n, P_0) - A_n] = 2(n^2 v_n)^{-1/2} U_{nn}.$$

We now apply martingale central limit theorem to U_{nn}. Following the argument from Hall (1984), it can be shown that

$$\Big[\frac{1}{2}n^2\mathbb{E}\tilde{K}_n^2(X, X')\Big]^{-1/2} U_{nn} \xrightarrow{d} N(0, 1), \tag{14}$$

provided that

$$[\mathbb{E}G_n^2(X, X') + n^{-1}\mathbb{E}\tilde{K}_n^2(X, X')\tilde{K}_n^2(X, X'') + n^{-2}\mathbb{E}\tilde{K}_n^4(X, X')]/[\mathbb{E}\tilde{K}_n^2(X, X')]^2 \to 0, \tag{15}$$

as $n \to \infty$, where $G_n(x, x') = \mathbb{E}\tilde{K}_n(X, x)\tilde{K}_n(X, x')$. Since

$$\mathbb{E}\tilde{K}_n^2(X, X') = \sum_{k\geq 1} \Big(\frac{\lambda_k}{\lambda_k + \varrho_n^2}\Big)^2 = v_n,$$

(14) implies that

$$v_n^{-1/2}[n\eta_{\varrho_n}^2(\widehat{P}_n, P_0) - A_n] = \sqrt{2} \cdot \Big(\frac{1}{2}n^2\mathbb{E}\tilde{K}_n^2(X, X')\Big)^{-1/2} U_{nn} \xrightarrow{d} N(0, 2).$$

It therefore suffices to verify (15).

Note that

$$\begin{aligned}\mathbb{E}\tilde{K}_n^2(X, X') = \sum_{k\geq 1} \Big(\frac{\lambda_k}{\lambda_k + \varrho_n^2}\Big)^2 \geq & \sum_{\lambda_k \geq \varrho_n^2} \frac{1}{4} + \frac{1}{4\varrho_n^4} \sum_{\lambda_k < \varrho_n^2} \lambda_k^2 \\ = & \frac{1}{4}|\{k : \lambda_k \geq \varrho_n^2\}| + \frac{1}{4\varrho_n^4} \sum_{\lambda_k < \varrho_n^2} \lambda_k^2 \asymp \varrho_n^{-1/s},\end{aligned}$$

where the last step holds by considering that $\lambda_k \asymp k^{-2s}$. Hereafter, we shall write $a_n \asymp b_n$ if $0 < \underline{\lim}_{n\to\infty} a_n/b_n \le \overline{\lim}_{n\to\infty} a_n/b_n < \infty$, for two positive sequences $\{a_n\}$ and $\{b_n\}$. Similarly,

$$\mathbb{E}G_n^2(X, X') = \sum_{k\ge 1} \Big(\frac{\lambda_k}{\lambda_k + \varrho_n^2}\Big)^4 \le |\{k : \lambda_k \ge \varrho_n^2\}| + \varrho_n^{-8} \sum_{\lambda_k < \varrho_n^2} \lambda_k^4 \asymp \varrho_n^{-1/s},$$

and

$$\begin{aligned}\mathbb{E}\tilde{K}_n^2(X, X')\tilde{K}_n^2(X, X'') =& \mathbb{E}\Big\{\sum_{k\ge 1} \Big(\frac{\lambda_k}{\lambda_k + \varrho_n^2}\Big)^2 \varphi_k^2(X)\Big\}^2 \\ &\le \Big(\sup_{k\ge 1} \|\varphi_k\|_\infty\Big)^4 \Big\{\sum_{k\ge 1} \Big(\frac{\lambda_k}{\lambda_k + \varrho_n^2}\Big)^2\Big\}^2 \asymp \varrho_n^{-2/s}.\end{aligned}$$

Thus there exists a positive constant C_3 such that

$$\mathbb{E}G_n^2(X, X')/[\mathbb{E}\tilde{K}_n^2(X, X')]^2 \le C_3 \varrho_n^{1/s} \to 0, \tag{16}$$

and

$$n^{-1}\mathbb{E}\tilde{K}_n^2(X, X')\tilde{K}_n^2(X, X'')/[\mathbb{E}\tilde{K}_n^2(X, X')]^2 \le C_3 n^{-1} \to 0, \tag{17}$$

as $n \to \infty$. On the other hand,

$$\mathbb{E}\tilde{K}_n^4(X, X') \le \|\tilde{K}_n\|_\infty^2 \mathbb{E}\tilde{K}_n^2(X, X'),$$

where

$$\|\tilde{K}_n\|_\infty = \sup_x \Big\{\sum_{k\ge 1} \frac{\lambda_k}{\lambda_k + \varrho_n^2} \varphi_k^2(x)\Big\} \le \Big(\sup_{k\ge 1} \|\varphi_k\|_\infty\Big)^2 \sum_{k\ge 1} \frac{\lambda_k}{\lambda_k + \varrho_n^2} \asymp \varrho_n^{-1/s}.$$

This implies that for some positive constant C_4,

$$n^{-2}\mathbb{E}\tilde{K}_n^4(X, X')\}/[\mathbb{E}\tilde{K}_n^2(X, X')]^2 \le n^{-2}\|\tilde{K}_n\|_\infty^2/\mathbb{E}\tilde{K}_n^2(X, X') \le C_4(n^2\varrho_n^{1/s})^{-1} \to 0. \tag{18}$$

as $n \to \infty$. Together, (16), (17) and (18) ensure that condition (15) holds. ■

Proof [Proof of Theorem 3] Note that

$$\begin{aligned}
&n\eta^2_{\varrho_n}(\widehat{P}_n, P_0) - \frac{1}{n}\sum_{i=1}^{n}\tilde{K}_n(X_i, X_i)\\
=&\frac{1}{n}\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}\sum_{\substack{1\leq i,j\leq n\\ i\neq j}}\varphi_k(X_i)\varphi_k(X_j)\\
=&\frac{1}{n}\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}\sum_{\substack{1\leq i,j\leq n\\ i\neq j}}[\varphi_k(X_i)-\mathbb{E}_P\varphi_k(X)][\varphi_k(X_j)-\mathbb{E}_P\varphi_k(X)]\\
&+\frac{2(n-1)}{n}\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_P\varphi_k(X)]\sum_{1\leq i\leq n}[\varphi_k(X_i)-\mathbb{E}_P\varphi_k(X)]\\
&+\frac{n(n-1)}{n}\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_P\varphi_k(X)]^2\\
:=&V_1+V_2+V_3.
\end{aligned}$$

Obviously, $\mathbb{E}_P V_1V_2 = 0$. We first argue that the following three statements together implies the desired result:

$$\lim_{n\to\infty}\inf_{P\in\mathcal{P}(\Delta_n,\theta)} v_n^{-1/2}V_3 = \infty, \tag{19}$$

$$\sup_{P\in\mathcal{P}(\Delta_n,\theta)}(\mathbb{E}_P V_1^2/V_3^2) = o(1), \tag{20}$$

$$\sup_{P\in\mathcal{P}(\Delta_n,\theta)}(\mathbb{E}_P V_2^2/V_3^2) = o(1). \tag{21}$$

To see this, note that (19) implies that

$$\begin{aligned}
&\lim_{n\to\infty}\inf_{P\in\mathcal{P}(\Delta_n,\theta)} P(v_n^{-1/2}[n\eta^2_{\varrho_n}(\widehat{P}_n,P_0)-A_n]\geq\sqrt{2}z_{1-\alpha})\\
\geq&\lim_{n\to\infty}\inf_{P\in\mathcal{P}(\Delta_n,\theta)} P\Big(v_n^{-1/2}V_3\geq 2\sqrt{2}z_{1-\alpha}, V_1+V_2+V_3\geq\frac{1}{2}V_3\Big)\\
=&\lim_{n\to\infty}\inf_{P\in\mathcal{P}(\Delta_n,\theta)} P\Big(V_1+V_2+V_3\geq\frac{1}{2}V_3\Big).
\end{aligned}$$

On the other hand, (20) and (21) imply that

$$\begin{aligned}
\lim_{n\to\infty}\inf_{P\in\mathcal{P}(\Delta_n,\theta)} P\Big(V_1+V_2+V_3\geq\frac{1}{2}V_3\Big) =&1-\lim_{n\to\infty}\sup_{P\in\mathcal{P}(\Delta_n,\theta)} P\Big(V_1+V_2+V_3<\frac{1}{2}V_3\Big)\\
\geq&1-\lim_{n\to\infty}\sup_{P\in\mathcal{P}(\Delta_n,\theta)}\frac{\mathbb{E}_P(V_1+V_2)^2}{(V_3/2)^2}=1.
\end{aligned}$$

This immediately suggests that $T_{\mathrm{M^3d}}$ is consistent. We now show that (19)-(21) indeed hold.

Verifying (19)**.** We begin with (19). Since $v_n \asymp \varrho_n^{-1/s}$ and $V_3 = (n-1)\eta^2_{\varrho_n}(P, P_0)$, (19) is equivalent to

$$\lim_{n\to\infty} \inf_{P\in\mathcal{P}(\Delta_n,\theta)} n\varrho_n^{\frac{1}{2s}} \eta^2_{\varrho_n}(P, P_0) = \infty.$$

For any $P \in \mathcal{P}(\Delta_n, \theta)$, let $u = dP/dP_0 - 1$ and $a_k = \langle u, \varphi_k\rangle_{L_2(P_0)} = \mathbb{E}_P\varphi_k(X)$. Based on the assumption that K is universal, $u = \sum_{k\geq 1} a_k\varphi_k$. We consider the case $\theta = 0$ and $\theta > 0$ separately.

(1) First consider $\theta = 0$. It is clear that

$$\begin{aligned}\eta^2_{\varrho_n}(P, P_0) =& \sum_{k\geq 1} a_k^2 - \sum_{k\geq 1} \frac{\varrho_n^2}{\lambda_k + \varrho_n^2} a_k^2\\ \geq& \|u\|^2_{L_2(P_0)} - \varrho_n^2 \sum_{k\geq 1} \frac{1}{\lambda_k} a_k^2\\ \geq& \|u\|^2_{L_2(P_0)} - \varrho_n^2 M^2.\end{aligned}$$

Take $\varrho_n \leq \sqrt{\Delta_n/(2M^2)}$ so that $\rho_n^2 M^2 \leq \frac{1}{2}\Delta_n$. Then we have

$$\inf_{P\in\mathcal{P}(\Delta_n,0)} \eta^2_{\varrho_n}(P, P_0) \geq \frac{1}{2} \inf_{P\in\mathcal{P}(\Delta_n,0)} \|u\|^2_{L_2(P_0)} = \frac{1}{2}\Delta_n.$$

(2) Now consider the case when $\theta > 0$. For $P \in \mathcal{P}(\Delta_n, \theta)$, $\forall\ R > 0$, $\exists\ f_R \in \mathcal{H}(K)$ such that $\|u - f_R\|_{L_2(P_0)} \leq MR^{-1/\theta}$ and $\|f_R\|_K \leq R$. Let $b_k = \langle f_R, \varphi_k\rangle_{L_2(P_0)}$.

$$\begin{aligned}\eta^2_{\varrho_n}(P, P_0) =& \sum_{k\geq 1} a_k^2 - \sum_{k\geq 1} \frac{\varrho_n^2}{\lambda_k + \varrho_n^2} a_k^2\\ \geq& \|u\|^2_{L_2(P_0)} - 2\sum_{k\geq 1} \frac{\varrho_n^2}{\lambda_k + \varrho_n^2}(a_k - b_k)^2 - 2\sum_{k\geq 1} \frac{\varrho_n^2}{\lambda_k + \varrho_n^2} b_k^2\\ \geq& \|u\|^2_{L_2(P_0)} - 2\sum_{k\geq 1}(a_k - b_k)^2 - 2\varrho_n^2 \sum_{k\geq 1} \frac{1}{\lambda_k} b_k^2\\ =& \|u\|^2_{L_2(P_0)} - 2\|u - f_R\|^2_{L_2(P_0)} - 2\varrho_n^2\|f_R\|^2_K.\end{aligned}$$

Taking $R = (2M/\|u\|_{L_2(P_0)})^\theta$ yields that

$$\eta^2_{\varrho_n}(P, P_0) \geq \|u\|^2_{L_2(P_0)} - 2M^2R^{-2/\theta} - 2\varrho_n^2R^2 = \frac{1}{2}\|u\|^2_{L_2(P_0)} - 2\varrho_n^2R^2.$$

Now by choosing

$$\varrho_n \leq \frac{1}{2\sqrt{2}}(2M)^{-\theta}\Delta_n^{\frac{1+\theta}{2}},$$

we can ensure that

$$2\varrho_n^2R^2 \leq \frac{1}{4}\|u\|^2_{L_2(P_0)}.$$

So that

$$\inf_{P\in\mathcal{P}(\Delta_n,\theta)} \eta^2_{\varrho_n}(P, P_0) \geq \inf_{P\in\mathcal{P}(\Delta_n,\theta)} \frac{1}{4}\|u\|^2_{L_2(P_0)} \geq \frac{1}{4}\Delta_n.$$

In both cases, with $\varrho_n \le C\Delta_n^{\frac{\theta+1}{2}}$ for a sufficiently small $C = C(M) > 0$, $\lim_{n\to\infty} \varrho_n^{\frac{1}{2s}} n\Delta_n = \infty$ suffices to ensure (19) holds. Under the condition that $\lim_{n\to\infty} \Delta_n n^{\frac{4s}{4s+\theta+1}} = \infty$,

$$\varrho_n = cn^{-\frac{2s(\theta+1)}{4s+\theta+1}} \le C\Delta_n^{\frac{\theta+1}{2}}$$

for sufficiently large n and $\lim_{n\to\infty} \varrho_n^{\frac{1}{2s}} n\Delta_n = \infty$ holds as well.

Verifying (20)**.** Rewrite V_1 as

$$\begin{aligned} V_1 =& \frac{1}{n} \sum_{\substack{1\le i,j\le n \\ i\neq j}} \sum_{k\ge 1} \frac{\lambda_k}{\lambda_k + \varrho_n^2} [\varphi_k(X_i) - \mathbb{E}_P\varphi_k(X)][\varphi_k(X_j) - \mathbb{E}_P\varphi_k(X)] \\ :=& \frac{1}{n} \sum_{\substack{1\le i,j\le n \\ i\neq j}} F_n(X_i, X_j). \end{aligned}$$

Then

$$\begin{aligned} \mathbb{E}_P V_1^2 =& \frac{1}{n^2} \sum_{\substack{i\neq j \\ i'\neq j'}} \mathbb{E}_P F_n(X_i, X_j) F_n(X_{i'}, X_{j'}) \\ =& \frac{2n(n-1)}{n^2} \mathbb{E}_P F_n^2(X, X') \\ \le& 2\mathbb{E}_P F_n^2(X, X'), \end{aligned}$$

where $X, X' \overset{i.i.d.}{\sim} P$. Recall that, for any two random variables Y_1, Y_2 such that $\mathbb{E}Y_1^2 < \infty$,

$$\mathbb{E}[Y_1 - \mathbb{E}(Y_1|Y_2)]^2 = \mathbb{E}Y_1^2 - \mathbb{E}[\mathbb{E}(Y_1|Y_2)^2] \le \mathbb{E}Y_1^2.$$

Together with the fact that

$$\begin{aligned} F_n(X, X') =& \tilde{K}_n(X, X') - \mathbb{E}_P[\tilde{K}_n(X, X')|X] - \mathbb{E}_P[\tilde{K}_n(X, X')|X'] + \mathbb{E}_P\tilde{K}_n(X, X') \\ =& \tilde{K}_n(X, X') - \mathbb{E}_P[\tilde{K}_n(X, X')|X] - \mathbb{E}\left[\tilde{K}_n(X, X') - \mathbb{E}_P[\tilde{K}_n(X, X')|X]\middle| X'\right], \end{aligned}$$

we have

$$\mathbb{E}_P F_n^2(X, X') \le \mathbb{E}_P\{\tilde{K}_n(X, X') - \mathbb{E}_P[\tilde{K}_n(X, X')|X]\}^2 \le \mathbb{E}_P\tilde{K}_n^2(X, X').$$

Thus, to prove (20), it suffices to show that

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}(\Delta_n,\theta)} \mathbb{E}_P\tilde{K}_n^2(X, X')/V_3^2 = 0.$$

For any $g \in L_2(P_0)$ and positive definite kernel $G(\cdot,\cdot)$ such that $\mathbb{E}_{P_0}G^2(X, X') < \infty$, let

$$\|g\|_G := \sqrt{\mathbb{E}_{P_0}[g(X)g(X')G(X, X')]}.$$

By the positive definiteness of $G(\cdot,\cdot)$, triangular inequality holds for $\|\cdot\|_G$, *i.e.*, for any g_1, $g_2 \in L_2(P_0)$,

$$|\|g_1\|_G - \|g_2\|_G| \le \|g_1 - g_2\|_G.$$

Thus by taking $G = \tilde{K}_n^2$, $g_1 = dP/dP_0$ and $g_2 = 1$, we have

$$\left|\sqrt{\mathbb{E}_P \tilde{K}_n^2(X,X')} - \sqrt{\mathbb{E}_{P_0} \tilde{K}_n^2(X,X')}\right| \le \sqrt{\mathbb{E}_{P_0}[u(X)u(X')\tilde{K}_n^2(X,X')]}. \tag{22}$$

We now appeal to the following lemma to bound the right hand side of (22):

Lemma 8 *Let G be a Mercer kernel defined over $\mathcal{X}\times\mathcal{X}$ with eigenvalue-eigenfunction pairs $\{(\mu_k,\varphi_k) : k \ge 1\}$ with respect to $L_2(P)$ such that $\mu_1 \ge \mu_2 \ge \cdots$. If G is a trace kernel in that $\mathbb{E}G(X,X) < \infty$, then for any $g \in L_2(P)$*

$$\mathbb{E}_P[g(X)g(X')G^2(X,X')] \le \mu_1 \left(\sum_{k\ge1}\mu_k\right)\left(\sup_{k\ge1}\|\varphi_k\|_\infty\right)^2 \|g\|^2_{L_2(P)}.$$

By Lemma 8, we get

$$\mathbb{E}_{P_0}[u(X)u(X')\tilde{K}_n^2(X,X')] \le C_5\left(\sum_k \frac{\lambda_k}{\lambda_k+\varrho_n^2}\right)\|u\|^2_{L_2(P_0)} \asymp \varrho_n^{-1/s}\|u\|^2_{L_2(P_0)}.$$

Recall that

$$\mathbb{E}_{P_0}\tilde{K}_n^2(X,X') = \sum_k \left(\frac{\lambda_k}{\lambda_k+\varrho_n^2}\right)^2 \asymp \varrho_n^{-1/s}.$$

In the light of (22), they imply that

$$\mathbb{E}_P\tilde{K}_n^2(X,X') \le 2\{\mathbb{E}_{P_0}\tilde{K}_n^2(X,X') + \mathbb{E}_{P_0}[u(X)u(X')\tilde{K}_n^2(X,X')]\} \le C_6\varrho_n^{-1/s}[1+\|u\|^2_{L_2(P_0)}].$$

On the other hand, as already shown in the part of verifying (19), $\varrho_n \ll \Delta_n^{\frac{\theta+1}{2}}$ suffices to ensure that for sufficiently large n,

$$\frac{1}{4}\|u\|^2_{L_2(P_0)} \le \eta^2_{\varrho_n}(P,P_0) \le \|u\|^2_{L_2(P_0)}, \quad \forall\ P \in \mathcal{P}(\Delta_n,\theta).$$

Thus

$$\begin{aligned}
&\lim_{n\to\infty}\sup_{P\in\mathcal{P}(\Delta_n,\theta)} \mathbb{E}_P\tilde{K}_n^2(X,X')/V_3^2 \\
\le& 16C_6\Big\{\Big(\lim_{n\to\infty}\inf_{P\in\mathcal{P}(\Delta_n,\theta)}\varrho_n^{1/s}n^2\|u\|^4_{L_2(P_0)}\Big)^{-1} + \Big(\lim_{n\to\infty}\inf_{P\in\mathcal{P}(\Delta_n,\theta)}\varrho_n^{1/s}n^2\|u\|^2_{L_2(P_0)}\Big)^{-1}\Big\} = 0
\end{aligned}$$

provided that $\lim_{n\to\infty} n^{\frac{4s}{4s+\theta+1}}\Delta_n = \infty$. This immediately implies (20).

Verifying (21)**.** Observe that

$$\begin{aligned}\mathbb{E}_P V_2^2 \leq & 4n\mathbb{E}_P\Big\{\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_P\varphi_k(X)][\varphi_k(X)-\mathbb{E}_P\varphi_k(X)]\Big\}^2\\ \leq & 4n\mathbb{E}_P\Big\{\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_P\varphi_k(X)][\varphi_k(X)]\Big\}^2\\ = & 4n\mathbb{E}_{P_0}\left([1+u(X)]\Big\{\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_P\varphi_k(X)][\varphi_k(X)]\Big\}^2\right).\end{aligned}$$

It is clear that

$$\begin{aligned}&\mathbb{E}_{P_0}\Big\{\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_P\varphi_k(X)][\varphi_k(X)]\Big\}^2\\ =&\sum_{k,k'\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}\frac{\lambda_{k'}}{\lambda_{k'}+\varrho_n^2}\mathbb{E}_P\varphi_k(X)\mathbb{E}_P\varphi_{k'}(X)\mathbb{E}_{P_0}[\varphi_k(X)\varphi_{k'}(X)]\\ =&\sum_{k\geq 1}\Big(\frac{\lambda_k}{\lambda_k+\varrho_n^2}\Big)^2[\mathbb{E}_P\varphi_k(X)]^2\leq \eta_{\varrho_n}^2(P,P_0).\end{aligned}$$

On the other hand,

$$\begin{aligned}&\mathbb{E}_{P_0}\left(u(X)\Big\{\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_P\varphi_k(X)][\varphi_k(X)]\Big\}^2\right)\\ \leq&\sqrt{\mathbb{E}_{P_0}\left(u^2(X)\Big\{\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_P\varphi_k(X)][\varphi_k(X)]\Big\}^2\right)}\times\\ &\times\sqrt{\mathbb{E}_{P_0}\Big\{\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_P\varphi_k(X)][\varphi_k(X)]\Big\}^2}\\ \leq&\|u\|_{L_2(P_0)}\sup_x\Big|\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_P\varphi_k(X)][\varphi_k(x)]\Big|\cdot\eta_{\varrho_n}(P,P_0)\\ \leq&\left(\sup_k\|\varphi_k\|_\infty\right)\|u\|_{L_2(P_0)}\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}|\mathbb{E}_P\varphi_k(X)|\cdot\eta_{\varrho_n}(P,P_0)\\ \leq&\left(\sup_k\|\varphi_k\|_\infty\right)\|u\|_{L_2(P_0)}\sqrt{\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}}\sqrt{\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_P\varphi_k(X)]^2}\cdot\eta_{\varrho_n}(P,P_0)\\ \leq&C_7\|u\|_{L_2(P_0)}\varrho_n^{-\frac{1}{2s}}\eta_{\varrho_n}^2(P,P_0).\end{aligned}$$

Together, they imply that

$$\begin{aligned}&\lim_{n\to\infty}\sup_{P\in\mathcal{P}(\Delta_n,\theta)}\mathbb{E}_P V_1^2/V_3^2\\ \le& 4\max\{1,C_7\}\left\{\left(\lim_{n\to\infty}\inf_{P\in\mathcal{P}(\Delta_n,\theta)} n\eta^2_{\varrho_n}(P,P_0)\right)^{-1}+\lim_{n\to\infty}\sup_{P\in\mathcal{P}(\Delta_n,\theta)}\left(\frac{\|u\|_{L_2(P_0)}}{\varrho_n^{\frac{1}{2s}} n\eta^2_{\varrho_n}(P,P_0)}\right)\right\}=0,\end{aligned}$$

under the assumption that $\lim_{n\to\infty} n^{\frac{4s}{4s+\theta+1}}\Delta_n=\infty$. ■

Proof [Proof of Theorem 4] The main architect is now standard in establishing minimax lower bounds for nonparametric hypothesis testing. The main idea is to carefully construct a set of points under the alternative hypothesis and argue that a mixture of these alternatives cannot be reliably distinguished from the null. See, *e.g.*, Ingster (1993); Ingster and Suslina (2003); Tsybakov (2008). Without loss of generality, assume $M=1$ and $\Delta_n = cn^{-\frac{4s}{4s+\theta+1}}$ for some $c>0$.

Let us consider the cases of $\theta=0$ and $\theta>0$ separately.

The case of $\theta=0$. We first treat the case when $\theta=0$. Let $B_n=\lfloor C_8\Delta_n^{-\frac{1}{2s}}\rfloor$ for a sufficiently small constant $C_8>0$ and $a_n=\sqrt{\Delta_n/B_n}$. For any $\xi_n:=(\xi_{n1},\xi_{n2},\cdots,\xi_{nB_n})^\top\in\{\pm1\}^{B_n}$, write

$$u_{n,\xi_n}=a_n\sum_{k=1}^{B_n}\xi_{nk}\varphi_k.$$

It is clear that

$$\|u_{n,\xi_n}\|^2_{L_2(P_0)}=B_n a_n^2=\Delta_n$$

and

$$\|u_{n,\xi_n}\|_\infty\le a_nB_n\left(\sup_k\|\varphi_k\|_\infty\right)\asymp\Delta_n^{\frac{2s-1}{4s}}\to 0.$$

By taking C_8 small enough, we can also ensure

$$\|u_{n,\xi_n}\|_K^2=a_n^2\sum_{k=1}^{B_n}\lambda_k^{-1}\le 1,$$

Therefore, there exists a probability measure $P_{n,\xi_n}\in\mathcal{P}(\Delta_n,0)$ such that $dP_{n,\xi_n}/dP_0=1+u_{n,\xi_n}$. Following a standard argument for minimax lower bound, it suffices to show that

$$\overline{\lim_{n\to\infty}}\,\mathbb{E}_{P_0}\left(\frac{1}{2^{B_n}}\sum_{\xi_n\in\{\pm1\}^{B_n}}\left\{\prod_{i=1}^n[1+u_{n,\xi_n}(X_i)]\right\}\right)^2<\infty. \tag{23}$$

Note that

$$\begin{aligned}
&\mathbb{E}_{P_0}\left(\frac{1}{2^{B_n}}\sum_{\xi_n\in\{\pm1\}^{B_n}}\left\{\prod_{i=1}^{n}[1+u_{n,\xi_n}(X_i)]\right\}\right)^2\\
=&\mathbb{E}_{P_0}\left(\frac{1}{2^{2B_n}}\sum_{\xi_n,\xi_n'\in\{\pm1\}^{B_n}}\left\{\prod_{i=1}^{n}[1+u_{n,\xi_n}(X_i)]\right\}\left\{\prod_{i=1}^{n}[1+u_{n,\xi_n'}(X_i)]\right\}\right)\\
=&\frac{1}{2^{2B_n}}\sum_{\xi_n,\xi_n'\in\{\pm1\}^{B_n}}\prod_{i=1}^{n}\mathbb{E}_{P_0}\left\{[1+u_{n,\xi_n}(X_i)][1+u_{n,\xi_n'}(X_i)]\right\}\\
=&\frac{1}{2^{2B_n}}\sum_{\xi_n,\xi_n'\in\{\pm1\}^{B_n}}\left(1+a_n^2\sum_{k=1}^{B_n}\xi_{nk}\xi_{nk}'\right)^n\\
\le&\frac{1}{2^{2B_n}}\sum_{\xi_n,\xi_n'\in\{\pm1\}^{B_n}}\exp\left(na_n^2\sum_{k=1}^{B_n}\xi_{n,k}\xi_{n,k}'\right)\\
=&\left\{\frac{\exp(na_n^2)+\exp(-na_n^2)}{2}\right\}^{B_n}\\
\le&\exp\left(\frac{1}{2}B_nn^2a_n^4\right),
\end{aligned}$$

where the last inequality is ensured by that

$$\cosh(t)\le\exp\left(\frac{t^2}{2}\right),\quad\forall\ t\in\mathbb{R}.$$

See, *e.g.*, Baraud (2002). With the particular choice of B_n, a_n, and the conditions on Δ_n, this immediately implies (23).

The case of $\theta > 0$. The main idea is similar to before. To find a set of probability measures in $\mathcal{P}(\Delta_n,\theta)$, we appeal to the following lemma.

Lemma 9 *Let $u=\sum_k a_k\varphi_k$. If*

$$\sup_{B\ge1}\left\{\left(\sum_{k=1}^{B}\frac{a_k^2}{\lambda_k}\right)^{2/\theta}\left(\sum_{k\ge B}a_k^2\right)\right\}\le M^2,$$

then $u\in\mathcal{F}(\theta,M)$.

Similar to before, we shall now take $B_n=\lfloor C_{10}\Delta_n^{-\frac{\theta+1}{2s}}\rfloor$ and $a_n=\sqrt{\Delta_n/B_n}$. By Lemma 9, we can find $P_{n,\xi_n}\in\mathcal{P}(\Delta_n,\theta)$ such that $dP_{n,\xi_n}/dP_0=1+u_{n,\xi_n}$, for appropriately chosen C_{10}. Following the same argument as in the previous case, we can again verify (23). ■

Proof [Proof of Theorem 5] Without loss of generality, assume that $\Delta_n(\theta) = c_1(n^{-1}\sqrt{\log\log n})^{\frac{4s}{4s+\theta+1}}$ for some constant $c_1 > 0$ to be determined later.

Type I Error. We first prove the first statement which shows that the Type I error converges to 0. Following the same notations as defined in the proof of Theorem 2, let

$$N_{n,2} = \mathbb{E}\Big\{\sum_{j=2}^{n}\mathbb{E}\Big(\tilde{\zeta}_{nj}^2|\mathcal{F}_{j-1}\Big) - 1\Big\}^2, \quad L_{n,2} = \sum_{j=2}^{n}\mathbb{E}\tilde{\zeta}_{nj}^4$$

where $\tilde{\zeta}_{nj} = \sqrt{2}\zeta_{nj}/(n\sqrt{v_n})$. As shown by Haeusler (1988),

$$\sup_t |P(T_{n,\varrho_n} > t) - \bar{\Phi}(t)| \le C_{11}(L_{n,2} + N_{n,2})^{1/5},$$

where $\bar{\Phi}(t)$ is the survival function of the standard normal, *i.e.*, $\bar{\Phi}(t) = P(Z > t)$ where $Z \sim N(0,1)$. Again by the argument from Hall (1984),

$$\mathbb{E}\Big\{\sum_{j=2}^{n}\mathbb{E}(\zeta_{nj}^2|\mathcal{F}_{j-1}) - \frac{1}{2}n(n-1)v_n\Big\}^2 \le C_{12}[n^4\mathbb{E}G_n^2(X,X') + n^3\mathbb{E}\tilde{K}_n^2(X,X')\tilde{K}_n^2(X,X'')],$$

where $G_n(\cdot,\cdot)$ is defined in the proof of Theorem 2, and

$$\sum_{j=2}^{n}\mathbb{E}\zeta_{nj}^4 \le C_{13}[n^2\mathbb{E}\tilde{K}_n^4(X,X') + n^3\mathbb{E}\tilde{K}_n^2(X,X')\tilde{K}_n^2(X,X'')],$$

which ensures

$$\begin{aligned} N_{n,2} =& \frac{4\mathbb{E}\Big\{\sum_{j=2}^{n}\mathbb{E}(\zeta_{nj}^2|\mathcal{F}_{j-1}) - \frac{1}{2}n(n-1)v_n - \frac{1}{2}nv_n\Big\}^2}{n^4v_n^2} \\ \le& 8\max\left\{C_{12}, \frac{1}{4}\right\}\left\{\frac{\mathbb{E}G_n^2(X,X')}{v_n^2} + \frac{\mathbb{E}\tilde{K}_n^2(X,X')\tilde{K}_n^2(X,X'')}{nv_n^2} + \frac{1}{n^2}\right\}, \end{aligned}$$

and

$$L_{n,2} = \frac{4\sum_{j=2}^{n}\mathbb{E}\tilde{\zeta}_{nj}^4}{n^4v_n^2} \le 4C_{13}\left\{\frac{\mathbb{E}\tilde{K}_n^4(X,X')}{n^2v_n^2} + \frac{\mathbb{E}\tilde{K}_n^2(X,X')\tilde{K}_n^2(X,X'')}{nv_n^2}\right\}.$$

As shown in the proof of Theorem 2,

$$\frac{\mathbb{E}G_n^2(X,X')}{v_n^2} \le C_3\varrho_n^{1/s}, \quad \frac{\mathbb{E}\tilde{K}_n^4(X,X')}{n^2v_n^2} \le C_4n^{-2}\varrho_n^{-1/s}, \quad \text{and} \quad \frac{\mathbb{E}\tilde{K}_n^2(X,X')\tilde{K}_n^2(X,X'')}{nv_n^2} \le C_3n^{-1}.$$

Therefore,

$$\sup_t |P(T_{n,\varrho_n} > t) - \bar{\Phi}(t)| \le C_{14}(\varrho_n^{\frac{1}{5s}} + n^{-\frac{1}{5}} + n^{-\frac{2}{5}}\varrho_n^{-\frac{1}{5s}}),$$

which implies that

$$P\left(\sup_{0\le k\le m_*} T_{n,2^k\varrho_*} > t\right) \le m_*\bar{\Phi}(t) + C_{15}(2^{\frac{m_*}{5s}}\varrho_*^{\frac{1}{5s}} + m_* n^{-\frac{1}{5}} + n^{-\frac{2}{5}}\varrho_*^{-\frac{1}{5s}}), \qquad \forall t.$$

It is not hard to see, by the definitions of m_* and ϱ_*,

$$2^{m_*}\varrho_* \le 2\left(\frac{\sqrt{\log\log n}}{n}\right)^{\frac{2s}{4s+1}}$$

and

$$\begin{aligned} m_* =& (\log 2)^{-1}\{2s\log n - \frac{2s}{4s+1}\log n + o(\log n)\} \\ =& (\log 2)^{-1}\frac{8s^2}{4s+1}\log n + o(\log n) \asymp \log n. \end{aligned}$$

Together with the fact that $\bar{\Phi}(t) \le \frac{1}{2}e^{-t^2/2}$ for $t \ge 0$, we get

$$\begin{aligned} &P\left(\sup_{0\le k\le m_*} T_{n,2^k\varrho_*} > \sqrt{3\log\log n}\right) \\ \le& C_{16}\left[e^{-\frac{3}{2}\log\log n}\log n + \left(\frac{\sqrt{\log n}}{n}\right)^{\frac{2}{5(4s+1)}} + n^{-\frac{1}{5}}\log\log n + n^{-\frac{2}{5}}\left(\frac{\sqrt{\log\log n}}{n}\right)^{-\frac{2}{5}}\right] \to 0, \end{aligned}$$

as $n \to \infty$.

Type II Error. Next consider Type II error. To this end, write $\varrho_n(\theta) = (\frac{\sqrt{\log\log n}}{n})^{\frac{2s(\theta+1)}{4s+\theta+1}}$. Let

$$\tilde{\varrho}_n(\theta) = \sup_{0\le k\le m_*}\{2^k\varrho_* : \varrho_n \le \varrho_n(\theta)\}.$$

It is clear that $\tilde{T}_n \ge T_{n,\tilde{\varrho}_n(\theta)}$ for any $\theta \ge 0$. It therefore suffices to show that for any $\theta \ge 0$,

$$\lim_{n\to\infty}\inf_{\theta\ge 0}\inf_{P\in\mathcal{P}(\Delta_n,\theta)} P\left\{T_{n,\tilde{\varrho}_n(\theta)} \ge \sqrt{3\log\log n}\right\} = 1.$$

By Markov inequality, this can accomplished by verifying

$$\inf_{\theta\in[0,\infty)}\inf_{P\in\mathcal{P}(\Delta_n(\theta),\theta)} \mathbb{E}_P T_{n,\tilde{\varrho}_n(\theta)} \ge \tilde{M}\sqrt{\log\log n} \tag{24}$$

for some $\tilde{M} > \sqrt{3}$; and

$$\lim_{n\to\infty}\sup_{\theta\ge 0}\sup_{P\in\mathcal{P}(\Delta_n(\theta),\theta)} \frac{\mathrm{Var}\left(T_{n,\tilde{\varrho}_n(\theta)}\right)}{\left(\mathbb{E}_P T_{n,\tilde{\varrho}_n(\theta)}\right)^2} = 0. \tag{25}$$

We now show that both (24) and (25) hold with

$$\Delta_n(\theta) = c_1\left(\frac{\sqrt{\log\log n}}{n}\right)^{\frac{4s}{4s+\theta+1}}$$

for a sufficiently large $c_1 = c_1(M, \tilde{M})$.

Note that $\forall\ \theta \in [0, \infty)$,

$$\frac{1}{2}\varrho_n(\theta) \leq \tilde{\varrho}_n(\theta) \leq \varrho_n(\theta), \tag{26}$$

which immediately suggests

$$\eta^2_{\tilde{\varrho}_n(\theta)}(P, P_0) \geq \eta^2_{\varrho_n(\theta)}(P, P_0). \tag{27}$$

Following the arguments in the proof of Theorem 3,

$$\mathbb{E}_P T_{n,\tilde{\varrho}_n(\theta)} \geq C_{17} n[\tilde{\varrho}_n(\theta)]^{1/(2s)} \eta^2_{\tilde{\varrho}_n(\theta)}(P, P_0) \geq 2^{-1/(2s)} C_{17} n[\varrho_n(\theta)]^{1/2s} \eta^2_{\varrho_n(\theta)}(P, P_0),$$

and $\forall\ P \in \mathcal{P}(\Delta_n(\theta), \theta)$,

$$\eta^2_{\varrho_n(\theta)}(P, P_0) \geq \frac{1}{4}\|u\|^2_{L_2(P_0)} \tag{28}$$

provided that $\Delta_n(\theta) \geq C'(M)\left(\frac{\sqrt{\log\log n}}{n}\right)^{\frac{4s}{4s+\theta+1}}$.

Therefore,

$$\inf_{P \in \mathcal{P}(\Delta_n(\theta), \theta)} \mathbb{E}_P T_{n,\tilde{\varrho}_n(\theta)} \geq C_{18} n[\varrho_n(\theta)]^{1/(2s)} \Delta_n(\theta) \geq C_{18} c_1 \sqrt{\log\log n} \geq \tilde{M}\sqrt{\log\log n}$$

if $c_1 \geq C_{18}^{-1}\tilde{M}$. Hence to ensure (24) holds, it suffices to take

$$c_1 = \max\{C'(M), C_{18}^{-1}\tilde{M}\}.$$

With (26), (27) and (28), the results in the proof of Theorem 3 imply that for sufficiently large n

$$\begin{aligned}\sup_{P \in \mathcal{P}(\Delta_n^*(\theta), \theta)} \frac{\mathrm{Var}\left(T_{n,\tilde{\varrho}_n(\theta)}\right)}{\left(\mathbb{E}_P T_{n,\tilde{\varrho}_n(\theta)}\right)^2} \leq & C_{19}\Big\{ \left([\varrho_n(\theta)]^{\frac{1}{2s}} n\Delta_n^*(\theta)\right)^{-2} + \left([\varrho_n(\theta)]^{\frac{1}{s}} n^2 \Delta_n^*(\theta)\right)^{-1} \\ & + (n\Delta_n^*(\theta))^{-1} + \left([\varrho_n(\theta)]^{\frac{1}{2s}} n\sqrt{\Delta_n^*(\theta)}\right)^{-1} \Big\} \\ \leq & 2C_{19}\left([\varrho_n(\theta)]^{\frac{1}{2s}} n\Delta_n^*(\theta)\right)^{-1} = 2C_{19}(c_1 \log\log n)^{-\frac{1}{2}} \to 0,\end{aligned}$$

which shows (25). ■

Proof [Proof of Theorem 6] The main idea of the proof is similar to that for Theorem 4.

Nevertheless, in order to show

$$\mathbb{E}_{P_0} T + \sup_{\theta \in [\theta_1, \theta_2]} \beta(T; \Delta_n(\theta), \theta)$$

converges to 1 rather than bounded below from 0, we need to find P_π, which is the marginal distribution on $\mathcal{X}^n$ with conditional distribution selected from

$$\{P^{\otimes n}:\ P \in \cup_{\theta\in[\theta_1,\theta_2]}\mathcal{P}(\Delta_n(\theta),\theta)\}$$

and prior distribution π on $\cup_{\theta\in[\theta_1,\theta_2]}\mathcal{P}(\Delta_n(\theta),\theta)$ such that the χ^2 distance between P_π and $P_0^{\otimes n}$ converges to 0. See Ingster (2000).

To this end, assume, without loss of generality, that

$$\Delta_n(\theta) = c_2\left(\frac{n}{\sqrt{\log\log n}}\right)^{-\frac{4s}{4s+\theta+1}}, \qquad \forall\theta\in[\theta_1,\theta_2],$$

where $c_2 > 0$ is a sufficiently small constant to be determined later.

Let $r_n = \lfloor C_{20}\log n\rfloor$ and $B_{n,1} = \lfloor C_{21}\Delta_n^{-\frac{\theta_1+1}{2s}}(\theta_1)\rfloor$for sufficiently small $C_{20}, C_{21} > 0$. Set $\theta_{n,1} = \theta_1$. For $2 \le r \le r_n$, let

$$B_{n,r} = 2^{r-2}B_{n,1}$$

and $\theta_{n,r}$ is selected such that the following equation holds.

$$B_{n,r} = \left\lfloor C_{21}[\Delta_n(\theta_{n,r})]^{-\frac{\theta_{n,r}+1}{2s}}\right\rfloor.$$

Note that by choosing C_{20} sufficiently small,

$$\begin{aligned}
B_{n,r_n} = 2^{r_n-2}B_{n,1} &\le \left\lfloor c_2^{\frac{2(\theta_1+1)}{4s+\theta_1+1}}\left(\frac{n}{\sqrt{\log\log n}}\right)^{\frac{2(\theta_1+1)}{4s+\theta_1+1}}\cdot 2^{r_n-2}\right\rfloor \\
&= \left\lfloor c_2^{\frac{2(\theta_1+1)}{4s+\theta_1+1}}\exp\left(\log\left(\frac{n}{\sqrt{\log\log n}}\right)\cdot\frac{2(\theta_1+1)}{4s+\theta_1+1} + (r_n-2)\log 2\right)\right\rfloor \\
&\le \left\lfloor C_{21}\exp\left(\log\left(\frac{n}{\sqrt{\log\log n}}\right)\cdot\frac{2(\theta_2+1)}{4s+\theta_2+1}\right)\right\rfloor = \lfloor C_{21}[\Delta_n(\theta_2)]^{-\frac{\theta_2+1}{2s}}\rfloor
\end{aligned}$$

for sufficiently large n. Thus, we can guarante that $\forall\ 1 \le r \le r_n$, $\theta_{n,r_n} \in [\theta_1,\theta_2]$.

We now construct a finite subset of $\cup_{\theta\in[\theta_1,\theta_2]}\mathcal{P}(\Delta_n(\theta),\theta)$ as follows. Let $B^*_{n,0} = 0$ and $B^*_{n,r} = B_{n,1} + \cdots + B_{n,r}$ for $r \ge 1$. For each $\xi_{n,r} = (\xi_{n,r,1},\cdots,\xi_{n,r,B_{n,r}}) \in \{\pm1\}^{B_{n,r}}$, let

$$f_{n,r,\xi_{n,r}} = 1 + \sum_{k=B^*_{n,r-1}+1}^{B^*_{n,r}} a_{n,r}\xi_{n,r,k-B^*_{n,r-1}}\varphi_k,$$

and $a_{n,r} = \sqrt{\Delta_n(\theta_{n,r})/B_{n,r}}$. Following the same argument as that in the proof of Theorem 4, we can verify that with a sufficiently small C_{21}, each $P_{n,r,\xi_{n,r}} \in \mathcal{P}(\Delta_n(\theta_{n,r}),\theta_{n,r})$, where $f_{n,r,\xi_{n,r}}$ is the Radon-Nikodym derivative $dP_{n,r,\xi_{n,r}}/dP_0$. With slight abuse of notation, write

$$f_n(X_1,X_2,\cdots,X_n) = \frac{1}{r_n}\sum_{r=1}^{r_n} f_{n,r}(X_1,X_2,\cdots,X_n),$$

where

$$f_{n,r}(X_1, X_2, \cdots, X_n) = \frac{1}{2^{B_{n,r}}} \sum_{\xi_{n,r} \in \{\pm 1\}^{B_{n,r}}} \prod_{i=1}^{n} f_{n,r,\xi_{n,r}}(X_i).$$

It now suffices to show that

$$\|f_n - 1\|^2_{L_2(P_0)} = \|f_n\|^2_{L_2(P_0)} - 1 \to 0, \qquad \text{as } n \to \infty,$$

where $\|f_n\|^2_{L_2(P_0)} = \mathbb{E}_{P_0} f_n^2(X_1, X_2, \cdots, X_n)$.

Note that

$$\begin{aligned}\|f_n\|^2_{L_2(P_0)} =& \frac{1}{r_n^2} \sum_{1 \le r, r' \le r_n} \langle f_{n,r}, f_{n,r'} \rangle_{L_2(P_0)} \\ =& \frac{1}{r_n^2} \sum_{1 \le r \le r_n} \|f_{n,r}\|^2_{L_2(P_0)} + \frac{1}{r_n^2} \sum_{\substack{1 \le r, r' \le r_n \\ r \ne r'}} \langle f_{n,r}, f_{n,r'} \rangle_{L_2(P_0)}.\end{aligned}$$

It is easy to verify that, for any $r \neq r'$,

$$\langle f_{n,r}, f_{n,r'} \rangle_{L_2(P_0)} = 1.$$

It therefore suffices to show that

$$\sum_{1 \le r \le r_n} \|f_{n,r}\|^2_{L_2(P_0)} = o(r_n^2).$$

Following the same derivation as that in the proof of Theorem 4, we can show that

$$\|f_{n,r}\|^2_{L_2(P_0)} \le \left(\frac{\exp(n a_{n,r}^2) + \exp(-n a_{n,r}^2)}{2} \right)^{B_{n,r}} \le \exp\left(\frac{1}{2} B_{n,r} n^2 a_{n,r}^4 \right)$$

for sufficiently large n. By setting c_2 in the expression of $\Delta_n(\theta)$ sufficiently small, we have

$$B_{n,r} n^2 a_{n,r}^4 \le \log r_n,$$

which ensures that

$$\sum_{1 \le r \le r_n} \|f_{n,r}\|^2_{L_2(P_0)} \le r_n^{3/2} = o(r_n^2).$$

■

Acknowledgments

We would like to thanks the editor, Arthur Gretton, and the anonymous reviewers for their insightful comments that helped greatly improved the paper. We also acknowledge support for this project from the National Science Foundation (NSF grants DMS-1803450 and DMS-2015285).

Appendix A.

Proof [Proof of Lemma 8] We have

$$G^2(x,x') = \sum_{k,l} \mu_k \mu_l \varphi_k(x)\varphi_l(x)\varphi_k(x')\varphi_l(x').$$

Thus

$$\begin{aligned}
&\int g(x)g(x')G^2(x,x')dP(x)dP(x') \\
=&\sum_{k,l} \mu_k\mu_l \Big(\int g(x)\varphi_k(x)\varphi_l(x)dP(x)\Big)^2 \\
\leq&\mu_1 \sum_k \mu_k \sum_l \Big(\int g(x)\varphi_k(x)\varphi_l(x)dP(x)\Big)^2 \\
\leq&\mu_1 \left(\sum_k \mu_k \int g^2(x)\varphi_k^2(x)dP(x)\right) \\
\leq&\mu_1 \left(\sum_k \mu_k \right) \left(\sup_k \|\varphi_k\|_\infty \right)^2 \|g\|^2_{L_2(P)}.
\end{aligned}$$

■

Proof [Proof of Lemma 9] For brevity, write

$$l_B = \sum_{k=1}^{B} \frac{a_k^2}{\lambda_k}.$$

By definition, it suffices to show that $\forall\ R > 0$, $\exists\ f_R \in \mathcal{H}(K)$ such that $\|f_R\|_K^2 \leq R^2$ and $\|u - f_R\|^2_{L_2(P_0)} \leq M^2 R^{-2/\theta}$.

To this end, let B be such that $l_B^2 \leq R^2 \leq l_{B+1}^2$, and denote by

$$f_R = \sum_{k=1}^{B} a_k\varphi_k + a^*_{B+1}(R)\varphi_{B+1},$$

where

$$a^*_{B+1}(R) = \mathrm{sgn}(a_{B+1})\sqrt{\lambda_{B+1}(R^2 - l_B^2)}.$$

Clearly,

$$\|f_R\|_K^2 = \sum_{k=1}^{B} \frac{a_k^2}{\lambda_k} + \frac{(a^*_{k+1}(R))^2}{\lambda_{B+1}} = R^2,$$

and

$$\|u - f_R\|^2_{L_2(P_0)} = \sum_{k>B+1} a_k^2 + \left(|a_{B+1}| - \sqrt{\lambda_{B+1}(R^2 - l_B^2)}\right)^2 \leq \sum_{k\geq B+1} a_k^2.$$

To ensure $u \in \mathcal{F}(\theta, M)$, it suffices to have

$$\sup_{l_B^2 \leq R^2 \leq l_{B+1}^2} \|u - f_R\|^2_{L_2(P_0)} R^{2/\theta} \leq M^2, \quad \forall\ B \geq 0,$$

which concludes the proof. ■

References

L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi. On combinatorial testing problems. *The Annals of Statistics*, 38(5):3063–3092, 2010.

N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: ranking and clustering. *Journal of ACM*, 55(5):23:1–23:27, 2008.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

F. Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017.

A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the unit hypersphere using von mises-fisher distributions. In *Journal of Machine Learning Research*, pages 1345–1382, 2005.

Y. Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5): 577–606, 2002.

K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pages 1981–1989, 2015.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning*, pages 2606–2615, 2016.

F. Cucker and D. Zhou. *Learning Theory: an Approximation Theory Viewpoint*. Cambridge University Press, New York, NY, 2007.

I. S. Dhillon, E. M. Marcotte, and U. Roshan. Diametrical clustering for identifying anti-correlated gene clusters. *Bioinformatics*, 19(13):1612–1619, 2003.

T. A. Driscoll, N. Hale, and L. N. Trefethen. *Chebfun Guide*. Pafnuty Publications, Oxford, 2014. URL `http://www.chebfun.org/docs/guide/`.

N. Dunford and J. T. Schwartz. *Linear Operators: Part II: Spectral Theory: Self Adjoint Operators in Hilbert Space*. Interscience Publishers, 1963.

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

M. S. Ermakov. Minimax detection of a signal in a gaussian white noise. *Theory of Probability & Its Applications*, 35(4):667–679, 1991.

C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

M. Fromont and B. Laurent. Adaptive goodness-of-fit tests in a density model. *The Annals of Statistics*, 34(2):680–720, 2006.

M. Fromont, B. Laurent, M. Lerasle, and P. Reynaud-Bouret. Kernels based tests with non-asymptotic bootstrap approaches for two-sample problem. In *JMLR: Workshop and Conference Proceedings*, volume 23, pages 23–1, 2012.

M. Fromont, B. Laurent, and P. Reynaud-Bouret. The two-sample problem for poisson processes: Adaptive tests with a nonasymptotic wild bootstrap approach. *The Annals of Statistics*, 41(3):1431–1461, 2013.

J. Gorham and L. Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, 2017.

G. G. Gregory. Large sample theory for U-statistics and tests of fit. *The Annals of Statistics*, 5(1):110–123, 1977.

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pages 513–520, 2007.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012a.

A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, pages 1205–1213, 2012b.

E. Haeusler. On the rate of convergence in the central limit theorem for martingales with discrete and continuous time. *The Annals of Probability*, 16(1):275–299, 1988.

P. Hall. Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of Multivariate Analysis*, 14(1):1–16, 1984.

Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel fisher discriminant analysis. In *Advances in Neural Information Processing Systems*, pages 609–616, 2007.

Yu. I. Ingster. Minimax testing of nonparametric hypotheses on a distribution density in the l_p metrics. *Theory of Probability & Its Applications*, 31(2):333–337, 1987.

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

Yu. I. Ingster. Minimax testing of hypotheses on the distribution density for ellipsoids in l_p. *Theory of Probability & Its Applications*, 39(3):417–436, 1995.

Yu. I. Ingster. Adaptive chi-square tests. *Journal of Mathematical Sciences*, 99(2):1110–1119, 2000.

Yu. I. Ingster and I. A. Suslina. Minimax nonparametric hypothesis testing for ellipsoids and besov bodies. *ESAIM: Probability and Statistics*, 4:53–135, 2000.

Yu. I. Ingster and I. A. Suslina. *Nonparametric Goodness-of-Fit Testing under Gaussian Models*. Springer, New York, NY, 2003.

V. R. Iyer, M. B. Eisen, D. T. Ross, G. Schuler, T. Moore, J. Lee, J. M. Trent, L. M. Staudt, J. Hudson, and M. S. Boguski. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283(5398):83–87, 1999.

W. Jitkrittum, Z. Szabó, K. P. Chwialkowski, and A. Gretton. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems*, pages 181–189, 2016.

W. Jitkrittum, W. Xu, Z. Szabo, K. Fukumizu, and A. Gretton. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems (to appear)*, 2017.

P. E. Jupp. Sobolev tests of goodness of fit of distributions on compact riemannian manifolds. *The Annals of Statistics*, 33(6):2957–2966, 2005.

J. Kellner and A. Celisse. A one-sample test for normality with kernel methods. *arXiv preprint arXiv:1507.02904*, 2015.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer Science & Business Media, New York, NY, 2008.

O. V. Lepski and V. G. Spokoiny. Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. *Bernoulli*, 5(2):333–358, 1999.

Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727, 2015.

Q. Liu, J. Lee, and M. Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pages 276–284, 2016.

R. Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3284–3305, 2013.

K. V. Mardia and P. E. Jupp. *Directional Statistics*. John Wiley & Sons, New York, NY, 2009.

J. S. Marron and M. P. Wand. Exact mean integrated squared error. *The Annals of Statistics*, 20(2):712–736, 1992.

H. Q. Minh, P. Niyogi, and Y. Yao. Mercer's theorem, feature maps, and smoothing. In *Conference on Learning Theory*, pages 154–168, 2006.

K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: a review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2001.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013.

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York, NY, 2009.

A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions". In *Proc. 18th International Conference on Algorithmic Learning Theory*, pages 13–31. Springer-Verlag, Berlin, Germany, 2007.

P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297, 1998.

V. G. Spokoiny. Adaptive hypothesis testing using wavelets. *The Annals of Statistics*, 24 (6):2477–2498, 1996.

S. Sra and D. Karp. The multivariate Watson distribution: maximum-likelihood estimation and other aspects. *Journal of Multivariate Analysis*, 114(C):256–269, 2013.

B. Sriperumbudur, K. Fukumizu, A. Gretton, G. Lanckriet, and B. Schoelkopf. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in Neural Information Processing Systems 22*, pages 1750–1758, 2009.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2(Nov):67–93, 2001.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*, 2017.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

L. N. Trefethen and Z. Battles. An extension of MATLAB to continuous functions and operators. *SIAM Journal on Scientific Computing*, 25:1743–1770, 2004. doi: 10.1137/S1064827503430126. URL `http://link.aip.org/link/?SCE/25/1743/1`.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, New York, NY, 2008.

G. Wahba. *Spline Models for Observational Data*, volume 59. Siam, 1990.

W. Zaremba, A. Gretton, and M. Blaschko. B-test: a non-parametric, low variance kernel two-sample test. In *Advances in Neural Information Processing Systems*, pages 755–763, 2013.