On the Construction of Minimax Optimal Nonparametric Tests with Kernel Embedding Methods

Tong Li

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

Abstract

On the Construction of Minimax Optimal Nonparametric Tests with Kernel Embedding Methods

Tong Li

Kernel embedding methods have witnessed a great deal of practical success in the area of nonparametric hypothesis testing in recent years. But ever since its first proposal, there exists an inevitable problem that researchers in this area have been trying to answer–what kernel should be selected, because the performance of the associated nonparametric tests can vary dramatically with different kernels. While the way of kernel selection is usually ad hoc, we wonder if there exists a principled way of kernel selection so as to ensure that the associated nonparametric tests have good performance. As consistency results against fixed alternatives do not tell the full story about the power of the associated tests, we study their statistical performance within the minimax framework. First, focusing on the case of goodness-of-fit tests, our analyses show that a vanilla version of the kernel embedding based test could be suboptimal, and suggest a simple remedy by moderating the kernel. 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. Then, we study the asymptotic properties of goodness-of-fit, homogeneity and independence tests using Gaussian kernels, arguably the most popular and successful among such tests. Our results provide theoretical justifications for this common practice by showing that tests using a Gaussian kernel with an appropriately chosen scaling parameter are minimax optimal against smooth alternatives in all three settings. In addition, our analysis also pinpoints the importance of choosing a diverging scaling parameter when using Gaussian kernels and

suggests a data-driven choice of the scaling parameter that yields tests optimal, up to an iterated logarithmic factor, over a wide range of smooth alternatives. Numerical experiments are presented to further demonstrate the practical merits of our methodology.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First of all, I want to express my sincere gratitude to my Ph.D. advisor Prof. Ming Yuan. His guidance and support are invaluable to my research and my life. I have learned a lot from his deep insights and descent work ethics.

I am very grateful to Prof. Bodhisattva Sen, Prof. Victor de la Pena, Prof. Cynthia Rush and Prof. Bin Cheng for their serving on the committee. Their comments and feedback were very helpful for the modification of my thesis and left me inspirations for future research.

I also want to thank my close friends and my fellow students, Yi Li, Youran Qi, Chensheng Kuang, Shulei Wang, Cuize Han, Fan Gao, Luxi Cao, Yuan Li, Yuqing Xu, Chaoyu Yuan, Guanhua Fang, Yuanzhe Xu and Shun Xu. Their company has made this journey much more pleasurable. Their suggestions and help have been indispensable during the hard moments of my life.

Finally, I want to thank my parents Xiuyuan Li and Ping Zhou. They have always been supporting and encouraging me. I owe a lot to them.

Dedication

To my beloved parents who always give me unconditional support and encouragement.

Chapter 1: Introduction

1.1 Kernel Embedding

Tests for goodness-of-fit, homogeneity and independence are central to statistical inferences. Numerous techniques have been developed for these tasks and are routinely used in practice. In recent years, there is a renewed interest in them from both statistics and other related fields as they arise naturally in many modern applications where the performance of classical methods are less than satisfactory. In particular, nonparametric inferences via the embedding of distributions into a reproducing kernel Hilbert space (RKHS) have emerged as a popular and powerful technique to tackle these challenges. The approach immediately allows for easy access to the rich machinery for RKHS and has found great successes in a wide range of applications from causal discovery to deep learning. See, *e.g.*, Muandet et al. (2017) for a recent review.

More specifically, let $K(\cdot,\cdot)$ be a symmetric and positive definite function defined over $\mathcal{X}\times\mathcal{X}$, that is $K(x,y) = K(y,x)$ for all $x, y \in \mathcal{X}$, and the Gram matrix $[K(x_i,x_j)]_{1\le i,j\le n}$ is positive definite for any distinct $x_1,\ldots,x_n \in \mathcal{X}$. The Moore-Aronszajn Theorem indicates that such a function, referred to as a kernel, can always be uniquely identified with a RKHS $\mathcal{H}_K$ of functions over $\mathcal{X}$. The embedding

$$\mu_{\mathbb{P}}(\cdot) := \int_{\mathcal{X}} K(x,\cdot)\mathbb{P}(dx),$$

maps a probability distribution $\mathbb{P}$ into $\mathcal{H}_K$. The difference between two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ can then be conveniently measured by

$$\gamma_K(\mathbb{P},\mathbb{Q}) := \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_K}.$$

Under mild regularity conditions, it can be shown that $\gamma_K(\mathbb{P},\mathbb{Q})$ is an integral probability metric so

that it is zero if and only if $\mathbb{P} = \mathbb{Q}$, and

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

As such, $\gamma_K(\mathbb{P}, \mathbb{Q})$ is often referred to as the *maximum mean discrepancy* (MMD) between $\mathbb{P}$ and $\mathbb{Q}$. See, *e.g.*, Sriperumbudur et al. (2010) or Gretton et al. (2012a) for details. In what follows, we shall drop the subscript K whenever its choice is clear from the context. It was noted recently that MMD is also closely related to the so-called energy distance between random variables (Székely et al., 2007; Székely and Rizzo, 2009) commonly used to measure independence. See, *e.g.*, Sejdinovic et al. (2013) and Lyons (2013).

1.2 Nonparametric Hypothesis Testing

Given a sample from $\mathbb{P}$ and/or $\mathbb{Q}$, estimates of the $\gamma(\mathbb{P}, \mathbb{Q})$ can be derived by replacing $\mathbb{P}$ and $\mathbb{Q}$ with their respective empirical distributions. These estimates can subsequently be used for nonparametric hypothesis testing. Here are several notable examples that we shall focus on in this work.

Goodness-of-fit tests. The goal of goodness-of-fit tests is to check if a sample comes from a pre-specified distribution. Let $X_1, \cdots, X_n$ be n independent $\mathcal{X}$-valued samples from a certain distribution $\mathbb{P}$. We are interested in testing if the hypothesis $H_0^{\text{GOF}} : \mathbb{P} = \mathbb{P}_0$ holds for a fixed $\mathbb{P}_0$. Deviation from $\mathbb{P}_0$ can be conveniently measured by $\gamma(\mathbb{P}, \mathbb{P}_0)$ which can be readily estimated by:

$$\gamma(\widehat{\mathbb{P}}_n, \mathbb{P}_0) := \sup_{f \in \mathcal{H}(K) : \|f\|_K \leq 1} \int_{\mathcal{X}} f \, d\left(\widehat{\mathbb{P}}_n - \mathbb{P}_0\right),$$

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

Homogeneity tests. Homogeneity tests check if two independent samples come from a common population. Given two independent samples $X_1, \cdots, X_n \sim_{\text{iid}} \mathbb{P}$ and $Y_1, \cdots, Y_m \sim_{\text{iid}} \mathbb{Q}$, we are

interested in testing if the null hypothesis $H_0^{\text{HOM}} : \mathbb{P} = \mathbb{Q}$ holds. Discrepancy between $\mathbb{P}$ and $\mathbb{Q}$ can be measured by $\gamma(\mathbb{P}, \mathbb{Q})$, and similar to before, it can be estimated by the MMD between $\widehat{\mathbb{P}}_n$ and $\widehat{\mathbb{Q}}_m$:

$$\gamma(\widehat{\mathbb{P}}_n, \widehat{\mathbb{Q}}_m) := \sup_{f \in \mathcal{H}(K): \|f\|_K \leq 1} \int_{\mathcal{X}} f d \left(\widehat{\mathbb{P}}_n - \widehat{\mathbb{Q}}_m \right).$$

Again we reject H_0 if the estimate exceeds a threshold calibrated to ensure a certain significance level.

Independence tests. How to measure or test for independence among a set of random variables is another classical problem in statistics. Let $X = (X^1, \ldots, X^k)^\top \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$ be a random vector. If the random vectors $X^1, \ldots, X^k$ are jointly independent, then the distribution of X can be factorized:

$$H_0^{\text{IND}} : \qquad \mathbb{P}^X = \mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^k}.$$

Dependence among $X^1, \ldots, X^k$ can be naturally measured by the difference between the joint distribution and the product distribution evaluated under MMD:

$$\gamma(\mathbb{P}^X, \mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^k}) = \|\mu_{\mathbb{P}^X} - \mu_{\mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^k}}\|_{\mathcal{H}_K}.$$

When $d = 2$, $\gamma^2(\mathbb{P}^X, \mathbb{P}^{X^1} \otimes \mathbb{P}^{X^2})$ can be expressed as the squared Hilbert-Schmidt norm of the cross-covariance operator associated with X^1 and X^2 and is therefore referred to as Hilbert-Schmidt independence criterion (HSIC; Gretton et al., 2005). The more general case as given above is sometimes referred to as dHSIC (see, *e.g.*, Pfister et al., 2018). As before, we proceed to reject the independence assumption when $\gamma(\widehat{\mathbb{P}}_n^X, \widehat{\mathbb{P}}_n^{X^1} \otimes \cdots \otimes \widehat{\mathbb{P}}_n^{X^k})$ exceed a certain threshold where $\widehat{\mathbb{P}}_n^X$ and $\widehat{\mathbb{P}}_n^{X^j}$ are the empirical distribution of X and X^j respectively.

1.3 Minimax Framework

In all these cases the test statistic, namely $\gamma^2(\widehat{\mathbb{P}}_n, \mathbb{P}_0)$, $\gamma^2(\widehat{\mathbb{P}}_n, \widehat{\mathbb{Q}}_m)$ or $\gamma^2(\widehat{\mathbb{P}}_n, \widehat{\mathbb{P}}_n^{X^1} \otimes \cdots \otimes \widehat{\mathbb{P}}_n^{X^k})$, is a V-statistic. Following standard asymptotic theory for V-statistics (see, *e.g.*, Serfling, 2009), it

can be shown that under mild regularity conditions, when appropriately scaled by the sample size, they converge to a mixture of χ^2_1 distribution with weights determined jointly by the underlying probability distribution and the choice of kernel K. In contrast, it can also be derived that for a fixed alternative,

$$\gamma^2(\widehat{\mathbb{P}}_n, \mathbb{P}_0) \to_p \gamma^2(\mathbb{P}, \mathbb{P}_0), \qquad \gamma^2(\widehat{\mathbb{P}}_n, \widehat{\mathbb{Q}}_m) \to_p \gamma^2(\mathbb{P}, \mathbb{Q})$$
$$\text{and} \qquad \gamma^2(\widehat{\mathbb{P}}_n, \widehat{\mathbb{P}}_n^{X^1} \otimes \cdots \otimes \widehat{\mathbb{P}}_n^{X^k}) \to_p \gamma^2(\mathbb{P}, \mathbb{P}^{X^1} \otimes \cdots \times \mathbb{P}^{X^k}).$$

This immediately suggests that all aforementioned tests are consistent against fixed alternatives in that their power tends to one as sample sizes increase. Although useful, such consistency results do not tell the full story about the power of these tests, and if there are yet more powerful methods.

Specifically, although consistency against any fixed alternative is proved, the rate at which the power convergeces to 1 may vary for different alternatives and it remains a problem whether a given sample size is large enough to ensure a good power for the underlying alternative. In other words, can we detect the difference between the null and the alternative hypotheses with a high probability even in the worst scenario?

This concern naturally brings the notion of uniform consistency, meaning power converging to 1 uniformly over all alternaives to be considered, and leads us to adpot the minimax hypothesis testing framework pioneered by Burnashev (1979), Ingster (1987), and Ingster (1993). See also Ermakov (1991), Spokoiny (1996), Lepski and Spokoiny (1999), Ingster and Suslina (2000), Ingster (2000), Baraud (2002), Fromont and Laurent (2006), Fromont et al. (2012), and Fromont et al. (2013), and references therein. Within this framework, we consider testing against alternatives getting closer and closer to the null hypothesis as the sample size increases. The smallest departure from the null hypotheses that can be detected consistently, in a minimax sense, is referred to as the optimal detection boundary. And the test that maintains uniform consistency as the departure converges to 0 at the rate of optimal detection boundary is called minimax rate optimal.

1.4 Kernel Selection and Adaptation

The critical importance of kernel selection is widely recognized in practice, as the statistical performances of the associated tests can vary dramatically with different kernels. Yet, the way it is done is usually ad hoc and how to do so in a more principled way remains one of the chief practical challenges. See, *e.g.*, Gretton et al. (2008), Fukumizu et al. (2009), Gretton et al. (2012b), and Sutherland et al. (2017). In the following chapters, we address this problem by proposing two kernel selection methods in different settings such that the associated tests are shown to be minimax rate optimal.

However, such kernel selection methods depend on some regularity condition of the underlying space of probability distributions, and whether we can do it in an agnostic approach remains another challenge. This also naturally brings about the issue of adaptation. To address this challenge, we introduce a simple testing procedure by maximizing a normalized MMD over a pre-specified class of kernels. Similar idea of maximizing MMD over a class of kernels was first introduced by Sriperumbudur et al. (2009). Our analysis, however, suggests that it is more desirable to maximize *normalized* MMD instead. More specifically, we show that the proposed procedure can attain the optimal rate, up to an iterated logarithmic factor, over spaces of probability distributions with different regularity conditions.

The rest of the thesis is organized as follows. In Chapter 2, focusing on the case of goodness-of-fit tests, our analyses show that a vanilla version of the kernel embedding based test could be suboptimal, and suggest a simple remedy by moderating the kernel. We prove that the moderated approach provides optimal tests and can also be made adaptive over a wide range of deviations from the null. Then, in Chapter 3 we study the asymptotic properties of goodness-of-fit, homogeneity and independence tests using Gaussian kernels, arguably the most popular and successful among such tests. Our results provide theoretical justifications for this common practice by showing that tests using Gaussian kernel with an appropriately chosen scaling parameter are minimax optimal against smooth alternatives in all three settings. In addition, we suggests a data-driven choice of the

scaling parameter that yields tests optimal, up to an iterated logarithmic factor, over a wide range of smooth alternatives. Numerical experiments are presented in Chapter 4 to further demonstrate the practical merits of our methodology. We conclude with some summary discussion in Chapter 5. All the main proofs are relegated to Chapter 6. Other technical results and their proofs are put in the appendix.

Chapter 2: Moderated Kernel Embedding

2.1 Background and Problem Setting

In this chapter, we focus on goodness-of-fit test. Specifically, with n independent $\mathcal{X}$-valued samples $X_1, \cdots, X_n$ from a certain distribution $\mathbb{P}$, we are interested in testing if the hypothesis

$$H_0^{\text{GOF}} : \mathbb{P} = \mathbb{P}_0$$

holds for a fixed $\mathbb{P}_0$. 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.

We consider kernel embedding and maximum mean discrepancy (MMD) based goodness-of-fit test. Specifically, the goodness-of-fit test can be carried out conveniently by first constructing an estimate of $\gamma(\mathbb{P}, \mathbb{P}_0)$, $\gamma(\widehat{\mathbb{P}}, \mathbb{P}_0)$, and then rejecting H_0 if the estimate exceeds a threshold calibrated

to ensure a certain significance level, say α ($0 < \alpha < 1$).

We adpot the minimax framework to evaluate the above mentioned testing strategy. To fix ideas, we assume in this chapter that $\mathbb{P}$ is dominated by $\mathbb{P}_0$ under the alternative so that the Radon-Nikodym derivative $d\mathbb{P}/d\mathbb{P}_0$ is well defined and use the χ^2 divergence between $\mathbb{P}$ and $\mathbb{P}_0$,

$$\chi^2(\mathbb{P}, \mathbb{P}_0) := \int_{\mathcal{X}} \left(\frac{d\mathbb{P}}{d\mathbb{P}_0} \right)^2 d\mathbb{P}_0 - 1,$$

as the separation metric to quantify the departure from the null hypothesis. We are particularly interested in the detection boundary, namely how close $\mathbb{P}$ and $\mathbb{P}_0$ 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 $\mathbb{P}$ is known up to a finite dimensional parameters under the alternative, the detection boundary of the likelihood ratio test is $n^{-1/2}$ 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 $\gamma_K(\widehat{\mathbb{P}}_n, \mathbb{P}_0)$ based test is of the order $n^{-1/4}$. 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 $\mathbb{P}$ and $\mathbb{P}_0$ depends on how close $u := d\mathbb{P}/d\mathbb{P}_0 - 1$ 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 $L_2(\mathbb{P}_0)$, 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 $f \in L_2(\mathbb{P}_0)$ such that for any $R > 0$, there exists an $f_R \in \mathcal{H}_K$ such that

$$\|f_R\|_{\mathcal{H}_K} \leq R, \qquad \text{and} \qquad \|f - f_R\|_{L_2(\mathbb{P}_0)} \leq MR^{-1/\theta}.$$

We also adopt the convention that

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

We investigate the optimal rate of detection for testing $H_0^{\text{GOF}} : \mathbb{P} = \mathbb{P}_0$ against

$$H_1^{\text{GOF}}(\Delta_n, \theta, M) : \mathbb{P} \in \mathcal{P}(\Delta_n, \theta, M), \tag{2.1}$$

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

$$d\mathbb{P}/d\mathbb{P}_0 - 1 \in \mathcal{F}(\theta; M), \qquad \text{and} \qquad \chi(\mathbb{P}, \mathbb{P}_0) \geq \Delta_n.$$

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 $\Delta_n \gg r_n$.

Throughout this chapter, we shall assume

$$\int_{\mathcal{X}\times\mathcal{X}} K^2(x, x')d\mathbb{P}_0(x)d\mathbb{P}_0(x') < \infty.$$

Hence the Hilbert-Schmidt integral operator

$$L_K(f)(x) = \int_{\mathcal{X}} K(x, x')f(x')d\mathbb{P}_0(x'), \quad \forall\, x \in \mathcal{X}$$

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),

$$K(x, x') = \sum_{k\geq 1} \lambda_k \phi_k(x)\phi_k(x') \tag{2.2}$$

in $L_2(\mathbb{P}_0 \otimes \mathbb{P}_0)$. We further assume that K is continuous and that $\mathbb{P}_0$ is nondegenerate, meaning the

support of $\mathbb{P}_0$ is $\mathcal{X}$. Then Mercer's theorem ensures that (2.2) holds pointwisely. See, *e.g.*, Theorem 4.49 of Steinwart and Christmann (2008).

As shown in Gretton et al. (2012a), the squared MMD between two probability distributions $\mathbb{P}$ and $\mathbb{P}_0$ can be expressed as

$$\gamma_K^2(\mathbb{P}, \mathbb{P}_0) = \int K(x, x') d(\mathbb{P} - \mathbb{P}_0)(x) d(\mathbb{P} - \mathbb{P}_0)(x'). \tag{2.3}$$

Write

$$\bar{K}(x, x') = K(x, x') - \mathbb{E}_{\mathbb{P}_0} K(x, X) - \mathbb{E}_{\mathbb{P}_0} K(X, x') + \mathbb{E}_{\mathbb{P}_0} K(X, X'), \tag{2.4}$$

where the subscript $\mathbb{P}_0$ signifies the fact that the expectation is taken over $X, X' \sim \mathbb{P}_0$ independently. By (2.4), $\gamma_K^2(\mathbb{P}, \mathbb{P}_0) = \gamma_{\bar{K}}^2(\mathbb{P}, \mathbb{P}_0)$. Therefore, without loss of generality, we can focus on kernels that are degenerate under $\mathbb{P}_0$, *i.e.*,

$$\mathbb{E}_{\mathbb{P}_0} K(X, \cdot) = 0. \tag{2.5}$$

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 $d\mathbb{P}/d\mathbb{P}_0 - 1$ 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 $d\mathbb{P}/d\mathbb{P}_0 - 1$ necessarily lies in the orthogonal complement of constant functions in $L_2(\mathbb{P}_0)$. A degenerate kernel K is universal if its eigenfunctions $\{\phi_k\}_{k \geq 1}$ form an orthonormal basis of the orthogonal complement of linear space $\{c \cdot \phi_0 : c \in \mathbb{R}\}$ where $\phi_0(x) = 1$ in $L_2(\mathbb{P}_0)$. 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 eigen-

values decaying polynomially, *i.e.*,

$$0 < \liminf_{k\to\infty} k^{2s}\lambda_k \le \limsup_{k\to\infty} k^{2s}\lambda_k < \infty \tag{2.6}$$

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

$$\sup_{k\ge 1} \|\phi_k\|_\infty < \infty, \tag{2.7}$$

Together with Assumptions (2.6), (2.7) ensures that Mercer's decomposition (2.2) holds uniformly.

2.2 Operating Characteristics of MMD Based Test

2.2.1 Asymptotics under H_0^{GOF}

Note that (2.5) implies $\mathbb{E}_{P_0}\phi_k(X) = 0, \forall\ k \ge 1$. Hence

$$\gamma^2(\mathbb{P}, \mathbb{P}_0) = \sum_{k\ge 1} \lambda_k [\mathbb{E}_{\mathbb{P}}\phi_k(X)]^2$$

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

$$\gamma^2(\widehat{\mathbb{P}}_n, \mathbb{P}_0) = \sum_{k\ge 1} \lambda_k \left[\frac{1}{n}\sum_{i=1}^{n} \phi_k(X_i)\right]^2.$$

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

$$n\gamma^2(\widehat{\mathbb{P}}_n, \mathbb{P}_0) \xrightarrow{d} \sum_{k\ge 1} \lambda_k Z_k^2 := W$$

under H_0^{GOF}, where $Z_k \overset{i.i.d.}{\sim} N(0,1)$. Let Φ_{MMD} be an MMD based test, which rejects H_0^{GOF} if and only if $n\gamma^2(\widehat{\mathbb{P}}_n, \mathbb{P}_0)$ exceeds the $1-\alpha$ quantile $q_{w,1-\alpha}$ of W, *i.e.*,

$$\Phi_{\mathrm{MMD}} = \mathbb{1}_{\{n\gamma^2(\widehat{\mathbb{P}}_n,\mathbb{P}_0) > q_{w,1-\alpha}\}}.$$

The above limiting distribution of $n\gamma^2(\widehat{\mathbb{P}}_n, \mathbb{P}_0)$ immediately suggests that Φ_{MMD} is an asymptotic α-level test.

2.2.2 Power Analysis for MMD Based Tests

We now investigate the power of Φ_{MMD} in testing H_0^{GOF} against $H_1^{\mathrm{GOF}}(\Delta_n, \theta, M)$ given by (2.1). Recall that the type II error of a test $\Phi : \mathcal{X}^n \to [0,1]$ for testing H_0 against a composite alternative $H_1 : \mathbb{P} \in \mathcal{P}$ is given by

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

where $\mathbb{E}_{\mathbb{P}}$ means taking expectation over $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}$. For brevity, we shall write $\beta(\Phi; \Delta_n, \theta, M)$ instead of $\beta(\Phi; \mathcal{P}(\Delta_n, \theta, M))$ in what follows, with $\mathcal{P}(\Delta_n, \theta, M)$ defined right below (2.1). The performance of a test Φ 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 Φ_{MMD} in the case when $\theta = 0$. Hereafter, we abbreviate M in $\mathcal{P}(\Delta_n, \theta, M)$, $H_1^{\mathrm{GOF}}(\Delta_n, \theta, M)$ and $\beta(\Phi; \Delta_n, \theta, M)$, unless it is necessary to emphasize the dependence.

Theorem 1. *Consider testing H_0^{GOF} against $H_1^{\mathrm{GOF}}(\Delta_n, 0)$ by Φ_{MMD}.*

(i) If $n^{1/4}\Delta_n \to \infty$, then

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

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

$$\liminf_{n\to\infty} \beta(\Phi_{\mathrm{MMD}}; c_0 n^{-1/4}, 0) > 0.$$

Theorem 1 shows that when the alternative $H_1^{\mathrm{GOF}}(\Delta_n, 0)$ is considered, the detection boundary of Φ_{MMD} is of the order $n^{-1/4}$. It is of interest to compare the detection rate achieved by Φ_{MMD} with that in a parametric setting where consistent tests are available if $n^{1/2}\Delta_n \to \infty$. 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. We shall now argue that this gap actually is largely due to the sub-optimality of Φ_{MMD}, and the detection boundary of Φ_{MMD} could be significantly improved through a slight modification of the MMD.

2.3 Optimal Tests Based on Moderated MMD

2.3.1 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 $\mathbb{P}$ with respect to $\mathbb{P}_0$ is far away from $\mathcal{H}_K$, the distance between the two projections may not honestly reflect the distance between them. More specifically, $\gamma^2(\mathbb{P}, \mathbb{P}_0) = \sum_{k\geq 1} \lambda_k [\mathbb{E}_{\mathbb{P}} \phi_k(X)]^2$, while the χ^2 distance between $\mathbb{P}$ and $\mathbb{P}_0$ is $\chi^2(\mathbb{P}, \mathbb{P}_0) = \sum_{k\geq 1} [\mathbb{E}_{\mathbb{P}} \phi_k(X)]^2$. Considering that λ_k decreases with k, $\gamma^2(\mathbb{P}, \mathbb{P}_0)$ can be much smaller than $\chi^2(\mathbb{P}, \mathbb{P}_0)$. 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

$$\eta_{K,\varrho}(\mathbb{P}, \mathbb{Q}; \mathbb{P}_0) = \sup_{f \in \mathcal{H}_K : \|f\|^2_{L_2(\mathbb{P}_0)} + \varrho^2 \|f\|^2_{\mathcal{H}_K} \leq 1} \int_{\mathcal{X}} f d(\mathbb{P} - \mathbb{Q}) \tag{2.8}$$

for a given distribution $\mathbb{P}_0$ 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 $\eta_{K,\varrho}(\mathbb{P}, \mathbb{Q}; \mathbb{P}_0)$ and the distance from Harchaoui et al. (2007) is the set of f that we optimize over on the righthand side of (2.8). In the case of two sample test, there is no information about $\mathbb{P}_0$ and therefore one needs to replace the norm $\|\cdot\|_{L_2(\mathbb{P}_0)}$ with the

empirical L_2 norm.

It is worth noting that $\eta_{K,\varrho}(\mathbb{P}, \mathbb{Q}; \mathbb{P}_0)$ can also be identified with a particular type of MMD. Specifically, $\eta_{K,\varrho}(\mathbb{P}, \mathbb{Q}; \mathbb{P}_0) = \gamma_{\tilde{K}_\varrho}(\mathbb{P}, \mathbb{Q})$, where

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

We shall nonetheless still refer to $\eta_{K,\varrho}(\mathbb{P}, \mathbb{Q}; \mathbb{P}_0)$ 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 $\mathbb{P}_0$ unless necessary. The unit ball in (2.8) is defined in terms of both RKHS norm and $L^2(\mathbb{P}_0)$ norm. Recall that $u = d\mathbb{P}/d\mathbb{P}_0 - 1$ so that

$$\sup_{\|f\|_{L_2(\mathbb{P}_0)} \leq 1} \int_{\mathcal{X}} f d(\mathbb{P} - \mathbb{P}_0) = \sup_{\|f\|_{L_2(\mathbb{P}_0)} \leq 1} \int_{\mathcal{X}} f u dP_0 = \|u\|_{L_2(\mathbb{P}_0)} = \chi(\mathbb{P}, \mathbb{P}_0).$$

We can therefore expect that a smaller ϱ will make $\eta^2_\varrho(\mathbb{P}, \mathbb{P}_0)$ closer to $\chi^2(\mathbb{P}, \mathbb{P}_0)$, since the unit ball to be considered will become more similar to the unit ball in $L_2(\mathbb{P}_0)$. This can also be verified by noticing that

$$\lim_{\varrho \to 0} \eta^2_\varrho(\mathbb{P}, \mathbb{P}_0) = \lim_{\varrho \to 0} \sum_{k \geq 1} \frac{\lambda_k}{\lambda_k + \varrho^2} [\mathbb{E}_P \phi_k(X)]^2 = \sum_{k \geq 1} [\mathbb{E}_P \phi_k(X)]^2 = \chi^2(\mathbb{P}, \mathbb{P}_0).$$

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 Φ_{MMD}. The test statistic we propose is the empirical version of $\eta^2_{\varrho_n}(\mathbb{P}, \mathbb{P}_0)$:

$$\eta^2_{\varrho_n}(\widehat{\mathbb{P}}_n, P_0) = \frac{1}{n^2} \sum_{i,j=1} \tilde{K}_{\varrho_n}(X_i, X_j) = \sum_{k \geq 1} \frac{\lambda_k}{\lambda_k + \varrho_n^2} \left[\frac{1}{n} \sum_{i=1}^n \phi_k(X_i) \right]^2. \tag{2.9}$$

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 Φ_{MMD}.

2.3.2 Operating Characteristics of $\eta^2_{\varrho_n}(\widehat{\mathbb{P}}_n, \mathbb{P}_0)$ Based Tests

Although the expression for $\eta^2_{\varrho_n}(\widehat{\mathbb{P}}_n, \mathbb{P}_0)$ given by (2.9) looks similar to that of $\gamma^2(\widehat{\mathbb{P}}_n, \mathbb{P}_0)$, 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 $\gamma^2(\widehat{\mathbb{P}}_n, \mathbb{P}_0)$, holds for $\eta^2_{\varrho_n}(\widehat{\mathbb{P}}_n, \mathbb{P}_0)$ under $\mathbb{P}_0$, 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^{GOF},*

$$v_n^{-1/2}[n\eta^2_{\varrho_n}(\widehat{\mathbb{P}}_n, \mathbb{P}_0) - A_n] \xrightarrow{d} N(0, 2),$$

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

$$2^{-1/2} v_n^{-1/2}[n\eta^2_{\varrho_n}(\widehat{\mathbb{P}}_n, \mathbb{P}_0) - A_n]$$

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 $\Phi_{\mathrm{M^3d}}$ where the subscript $\mathrm{M^3d}$ stands for *M*oderated *MMD*. The performance of $\Phi_{\mathrm{M^3d}}$ under the alternative hypothesis is characterized by the follow-

ing theorem, showing that its detection boundary is much improved when compared with that of Φ_{MMD}.

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

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

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

2.3.3 Minimax Optimality

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

Theorem 4. *Consider testing* H_0^{GOF} *against* $H_1^{\mathrm{GOF}}(\Delta_n, \theta)$, *for some* $\theta < 2s-1$. *If* $\limsup_{n\to\infty} \Delta_n n^{\frac{2s}{4s+\theta+1}} < \infty$, *then there exists* $\alpha \in (0,1)$ *such that for any* Φ_n *of level* α *(asymptotically) based on* $X_1, \cdots, X_n$,

$$\limsup_{n\to\infty} \beta(\Phi_n; \Delta_n, \theta) > 0.$$

Together with Theorem 3, this suggests that $\Phi_{\mathrm{M^3d}}$ 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.

2.4 Adaptation

Despite the minimax optimality of $\Phi_{\mathrm{M^3d}}$, 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 $\mathbb{P}_0$ and K are known a priori, so is s. On the other hand, θ reflects the property of $d\mathbb{P}/d\mathbb{P}_0$ which is typically not known in advance. This naturally brings us to 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^{GOF} against $H_1^{\mathrm{GOF}}(\Delta_n(\theta), \theta)$ 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 statistic is taken to be the maximum of T_{n,ϱ_n} for $\rho_n = \rho_*, 2\rho_*, 2^2\rho_*, \ldots, 2^{m_*}\rho_*$:

$$T_n^{\mathrm{GOF(adapt)}} := \sup_{0 \leq k \leq m_*} T_{n,2^k\varrho_*}, \tag{2.10}$$

where

$$T_{n,\varrho_n} = (2v_n)^{-1/2}[n\eta_{\varrho_n}^2(\widehat{\mathbb{P}}_n, \mathbb{P}_0) - A_n].$$

It turns out if an appropriate rejection threshold is chosen, $T_n^{\mathrm{GOF(adapt)}}$ 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^{GOF},

$$\lim_{n\to\infty} P\left(T_n^{\mathrm{GOF(adapt)}} \geq \sqrt{3\log\log n}\right) = 0;$$

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

$$\lim_{n\to\infty} \inf_{\mathbb{P}\in\cup_{\theta\geq 0}\mathcal{P}(\Delta_n(\theta),\theta)} P\left(T_n^{\mathrm{GOF(adapt)}} \geq \sqrt{3\log\log n}\right) = 1,$$

provided that $\Delta_n(\theta) \geq c_1(n^{-1}\sqrt{\log\log n})^{\frac{2s}{4s+\theta+1}}$.

Theorem 5 immediately suggests that a test rejects H_0^{GOF} if and only if $T_n^{\mathrm{GOF(adapt)}} \geq \sqrt{3\log\log n}$ is consistent for testing it against $H_1^{\mathrm{GOF}}(\Delta_n(\theta),\theta)$ for all $\theta \geq 0$ provided that

$$\Delta_n(\theta) \geq c_1(n^{-1}\sqrt{\log\log n})^{\frac{2s}{4s+\theta+1}}.$$

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*

$$\limsup_{n\to\infty} \sup_{\theta\in[\theta_1,\theta_2]} \left\{\Delta_n(\theta)\left(\frac{n}{\sqrt{\log\log n}}\right)^{\frac{2s}{4s+\theta+1}}\right\} \leq c_2$$

then

$$\lim_{n\to\infty} \inf_{\Phi_n} \left[\mathbb{E}_{\mathbb{P}_0}\Phi_n + \sup_{\theta\in[\theta_1,\theta_2]} \beta(\Phi_n;\Delta_n(\theta),\theta)\right] = 1.$$

Similar to Theorem 4, Theorem 6 shows that there is no consistent test for H_0^{GOF} against $H_1^{\mathrm{GOF}}(\Delta_n,\theta)$ simultaneously over all $\theta \in [\theta_1,\theta_2]$, if $\Delta_n(\theta) \leq c_2\left(n^{-1}\sqrt{\log\log n}\right)^{\frac{2s}{4s+\theta+1}}$ $\forall\ \theta \in [\theta_1,\theta_2]$ for a sufficiently small c_2. Together with Theorem 5, this suggests that the above men-

tioned adaptive test is indeed rate optimal.

Chapter 3: Gaussian Kernel Embedding

3.1 Test for Goodness-of-fit

Throughout this chapter, we shall consider goodness-of-fit, homogeneity and independence tests. We focus on continuous data, *e.g.*, $\mathcal{X} = \mathbb{R}^d$, and Gaussian kernels, which are arguably the most popular and successful choice in practice.

Among the three testing problems that we consider, it is instructive to begin with the case of goodness-of-fit. Obviously, the choice of kernel K plays an essential role in kernel embedding of distributions. In particular, when data are continuous, Gaussian kernels are commonly used. More specifically, a Gaussian kernel with a scaling parameter $\nu > 0$ is given by

$$G_{d,\nu}(x, y) = \exp\left(-\nu\|x - y\|_d^2\right), \qquad \forall x, y \in \mathbb{R}^d.$$

Hereafter $\|\cdot\|_d$ stands for the usual Euclidean norm in $\mathbb{R}^d$. For brevity, we shall suppress the subscript d in both $\|\cdot\|$ and G when the dimensionality is clear from the context. When $\mathbb{P}$ and $\mathbb{Q}$ are probability distributions defined over $\mathcal{X} = \mathbb{R}^d$, we shall write the MMD between them with a Gaussian kernel and scaling parameter ν as $\gamma_\nu(\mathbb{P}, \mathbb{Q})$ where the subscript signifies the specific value of the scaling parameter.

We shall restrict our attention to distributions with smooth densities. Denote by $\mathcal{W}_d^{s,2}$ the sth order Sobolev space in $\mathbb{R}^d$, that is

$$\mathcal{W}_d^{s,2} = \left\{ f : \mathbb{R}^d \to \mathbb{R} \middle| f \text{ is almost surely continuous and } \int (1 + \|\omega\|^2)^s \|\mathcal{F}(f)(\omega)\|^2 d\omega < \infty \right\}$$

where $\mathcal{F}(f)$ is the Fourier transform of f:

$$\mathcal{F}(f)(\omega) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(x) e^{-ix^\top \omega} dx.$$

In what follows, we shall again abbreviate the subscript d in $\mathcal{W}_d^{s,2}$ when it is clear from the context. For any $f \in \mathcal{W}^{s,2}$, we shall write

$$\|f\|_{\mathcal{W}^{s,2}}^2 = \int_{\mathbb{R}^d} (1 + \|\omega\|^2)^s \|\mathcal{F}(f)(\omega)\|^2 d\omega.$$

Let p and p_0 be the density functions of $\mathbb{P}$ and $\mathbb{P}_0$ respectively. We are interested in the case when both p and p_0 are elements from $\mathcal{W}^{s,2}$.

Note that we can rewrite the null hypothesis H_0^{GOF} in terms of density functions: $H_0^{\text{GOF}} : p = p_0$ for some prespecified denstiy $p_0 \in \mathcal{W}^{s,2}$. To better quantify the power of a test, we shall consider testing against an alternative that is increasingly closer to the null as the sample size n increases:

$$H_1^{\text{GOF}}(\Delta_n; s) : p \in \mathcal{W}^{s,2}(M), \quad \|p - p_0\|_{L_2} \geq \Delta_n,$$

where

$$\mathcal{W}^{s,2}(M) = \left\{ f \in \mathcal{W}^{s,2} : \|f\|_{\mathcal{W}^{s,2}} \leq M \right\}.$$

and

$$\|f\|_{L_2}^2 = \int_{\mathbb{R}^d} f^2(x) dx.$$

The alternative hypothesis $H_1^{\text{GOF}}(\Delta_n; s)$ is composite and the power of a test Φ based on $X_1, \ldots, X_n \sim p$ is therefore defined as

$$\text{power}(\Phi; H_1^{\text{GOF}}(\Delta_n; s)) := \inf_{p \in \mathcal{W}^{s,2}(M), \|p - p_0\|_{L_2} \geq \Delta_n} \mathbb{P}\{\Phi \text{ rejects } H_0^{\text{GOF}}\}$$

Let

$$\bar{G}_\nu(x, y; \mathbb{P}_0) = G_\nu(x, y) - \mathbb{E}_{X\sim\mathbb{P}_0} G_\nu(X, y) - \mathbb{E}_{X\sim\mathbb{P}_0} G_\nu(x, X) + \mathbb{E}_{X,X'\sim_{\text{iid}}\mathbb{P}_0} G_\nu(X, X').$$

and recall that

$$\gamma_\nu^2(\widehat{\mathbb{P}}_n, \mathbb{P}_0) = \frac{1}{n^2} \sum_{i,j=1}^{n} \bar{G}_\nu(X_i, X_j).$$

Similarly with Chapter 2, we correct for bias and use instead the following U-statistic:

$$\widehat{\gamma_\nu^2}(\mathbb{P}, \mathbb{P}_0) := \frac{1}{n(n-1)} \sum_{1\le i\neq j\le n}^{n} \bar{G}_\nu(X_i, X_j),$$

which we shall focus on in what follows.

The choice of the scaling parameter ν is essential when using RKHS embedding for goodness-of-fit test. While the importance of data-driven choice of ν is widely recognized in practice, almost all existing theoretical studies assume that a fixed kernel, therefore a fixed scaling parameter, is used. Here we shall demonstrate the benefit of using a data-driven scaling parameter, and especially choosing a scaling parameter that diverges with the sample size.

More specifically, we argue that, with appropriate scaling, $\widehat{\gamma_\nu^2}(\mathbb{P}, \mathbb{P}_0)$ can be viewed as an estimate of $\|p - p_0\|_{L_2}^2$ when $\nu \to \infty$ as $n \to \infty$. Note that

$$\int (p - p_0)^2 = \int p^2 - 2\int p \cdot p_0 + \int p_0^2.$$

The first term can be estimated by

$$\int p^2 \approx \frac{1}{n} \sum_{i=1}^{n} p(X_i) \approx \frac{1}{n} \sum_{i=1}^{n} \widehat{p}_{h,-i}(X_i)$$

where $\widehat{p}_{h,-i}$ is a kernel density estimate of p with the ith observation removed and bandwidth h:

$$\widehat{p}_{h,-i}(x) = \frac{1}{n(2\pi h^2)^{d/2}} \sum_{j\neq i} G_{(2h^2)^{-1}}(x - X_j).$$

Thus, we can estimate $\int p^2$ by

$$\frac{1}{n(n-1)(2\pi h^2)^{d/2}} \sum_{1\le i\neq j\le n} G_{(2h^2)^{-1}}(X_i, X_j).$$

Similarly, the cross-product term can be estimated by

$$\int p \cdot p_0 \approx \int \widehat{p}_h(x) p_0(x) dx = \frac{1}{n(2\pi h^2)^{d/2}} \sum_{i=1}^{n} \int G_{(2h^2)^{-1}}(x, X_i) p_0(x) dx.$$

Together, we can view

$$\frac{1}{n(n-1)(2\pi h^2)^{d/2}} \sum_{1\le i\neq j\le n} \bar{G}_{(2h^2)^{-1}}(X_i, X_j)$$

as an estimate of $\int (p - p_0)^2$. Following standard asymptotic properties of the kernel density estimator (see, *e.g.*, Tsybakov, 2008), we know that

$$(\pi/\nu)^{-d/2} \widehat{\gamma^2_\nu}(\mathbb{P}, \mathbb{P}_0) \to_p \|p - p_0\|^2_{L_2}$$

if $\nu \to \infty$ in such a fashion that $\nu = o(n^{4/d})$. Motivated by this observation, we shall now consider testing H_0^{GOF} using $\widehat{\gamma^2_\nu}(\mathbb{P}, \mathbb{P}_0)$ with a diverging ν. To signify the dependence of ν on the sample size, we shall add a subscript n in what follows.

Under H_0^{GOF}, it is clear $\mathbb{E}\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{P}_0) = 0$. Note also that

$$\begin{aligned}
&\text{var}(\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{P}_0)) \\
=&\frac{2}{n(n-1)} \mathbb{E}\left[\bar{G}_{\nu_n}(X_1, X_2)\right]^2 \\
=&\frac{2}{n(n-1)} \left[\mathbb{E}\left[G_{\nu_n}(X_1, X_2)\right]^2 - 2\mathbb{E}[G_{\nu_n}(X_1, X_2) G_{\nu_n}(X_1, X_3)] + \left(\mathbb{E}\left[G_{\nu_n}(X_1, X_2)\right]\right)^2\right] \\
=&\frac{2}{n(n-1)} \left[\mathbb{E}G_{2\nu_n}(X_1, X_2) - 2\mathbb{E}[G_{\nu_n}(X_1, X_2) G_{\nu_n}(X_1, X_3)] + \left(\mathbb{E}\left[G_{\nu_n}(X_1, X_2)\right]\right)^2\right]. \qquad (3.1)
\end{aligned}$$

Simple calculations yield:

$$\mathrm{var}(\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{P}_0)) = \frac{2(\pi/(2\nu_n))^{d/2}}{n^2} \cdot \|p_0\|^2_{L_2} \cdot (1 + o(1)),$$

assuming that $\nu_n \to \infty$. We shall show that

$$\frac{n}{\sqrt{2}} \left(\frac{2\nu_n}{\pi}\right)^{d/4} \widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{P}_0) \to_d N\left(0, \|p_0\|^2_{L_2}\right).$$

To use this as a test statistic, however, we will need to estimate $\mathrm{var}(\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{P}_0))$. To this end, it is natural to consider estimating each of the three terms on the rightmost hand side of (3.1) by U-statistics:

$$\begin{aligned}
\tilde{s}^2_{n,\nu_n} =& \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} G_{2\nu_n}(X_i, X_j) \\
&- \frac{2(n-3)!}{n!} \sum_{\substack{1 \le i, j_1, j_2 \le n \\ |\{i, j_1, j_2\}| = 3}} G_{\nu_n}(X_i, X_{j_1}) G_{\nu_n}(X_i, X_{j_2}) \\
&+ \frac{(n-4)!}{n!} \sum_{\substack{1 \le i_1, i_2, j_1, j_2 \le n \\ |\{i_1, i_2, j_1, j_2\}| = 4}} G_{\nu_n}(X_{i_1}, X_{j_1}) G_{\nu_n}(X_{i_2}, X_{j_2}).
\end{aligned}$$

Note that $\tilde{s}^2_{n,\nu_n}$ is not always positive. To avoid a negative estimate of the variance, we can replace it with a sufficiently small value, say $1/n^2$, whenever it is negative or too small. Namely, let

$$\widehat{s}^2_{n,\nu_n} = \max\left\{\tilde{s}^2_{n,\nu_n}, 1/n^2\right\},$$

and consider a test statistic:

$$T^{\mathrm{GOF}}_{n,\nu_n} := \frac{n}{\sqrt{2}} \widehat{s}^{-1}_{n,\nu_n} \widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{P}_0).$$

We have

Theorem 7. *Let $\nu_n \to \infty$ as $n \to \infty$ in such a fashion that $\nu_n = o(n^{4/d})$. Then, under H_0^{GOF},*

$$\frac{n}{\sqrt{2}}\left(\frac{2\nu_n}{\pi}\right)^{d/4} \widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{P}_0) \to_d N(0, \|p_0\|^2_{L_2}). \tag{3.2}$$

Moreover,

$$T^{\mathrm{GOF}}_{n,\nu_n} \to_d N(0, 1). \tag{3.3}$$

Theorem 7 immediately implies a test, denoted by $\Phi^{\mathrm{GOF}}_{n,\nu_n,\alpha}$ ($\alpha \in (0, 1)$), that rejects H_0^{GOF} if and only if $T^{\mathrm{GOF}}_{n,\nu_n}$ exceeds z_α, the upper $1 - \alpha$ quantile of the standard normal distribution, is an asymptotic α-level test.

We now proceed to study its power against a smooth alternative. Following the same argument as before, it can be shown that

$$\frac{1}{n(n-1)(\pi/\nu_n)^{d/2}} \sum_{1 \le i \ne j \le n} \bar{G}_{\nu_n}(X_i, X_j) \to_p \|p - p_0\|^2_{L_2},$$

and

$$(2\nu_n/\pi)^{d/2} \widehat{s^2}_{n,\nu_n} \to_p \|p\|^2_{L_2},$$

so that

$$n^{-1}(\nu_n/(2\pi))^{d/4} T^{\mathrm{GOF}}_n \to_p \|p - p_0\|^2_{L_2}/\|p\|_{L_2}.$$

This immediately implies that, if $\nu_n \to \infty$ in such a manner that $\nu_n = o(n^{4/d})$, then $\Phi^{\mathrm{GOF}}_{n,\nu_n,\alpha}$ is consistent for a fixed $p \ne p_0$ in that its power converges to one. In fact, as n increases, more and more subtle deviation from p_0 can be detected by $\Phi^{\mathrm{GOF}}_{n,\nu_n,\alpha}$. A refined analysis of the asymptotic behavior of $T^{\mathrm{GOF}}_{n,\nu_n}$ yields that

Theorem 8. *Assume that $n^{2s/(d+4s)}\Delta_n \to \infty$. Then for any $\alpha \in (0, 1)$,*

$$\lim_{n\to\infty} \mathrm{power}\{\Phi^{\mathrm{GOF}}_{n,\nu_n,\alpha}; H_1^{\mathrm{GOF}}(\Delta_n; s)\} \to 1,$$

provided that $\nu_n \asymp n^{4/(d+4s)}$.

In other words, $\Phi^{\mathrm{GOF}}_{n,\nu_n,\alpha}$ has a detection boundary of the order $O(n^{-2s/(d+4s)})$ which turns out to be minimax optimal in that no other tests could attain a detection boundary with faster rate of convergence. More precisely, we have

Theorem 9. *Assume that* $\liminf_{n\to\infty} n^{2s/(d+4s)}\Delta_n < \infty$ *and* p_0 *is density such that* $\|p_0\|_{\mathcal{W}^{s,2}} < M$. *Then there exists some* $\alpha \in (0,1)$ *such that for any test* Φ_n *of level* α *(asymptotically) based on* $X_1, \ldots, X_n \sim p$,

$$\liminf_{n\to\infty} \mathrm{power}\{\Phi_n; H_1^{\mathrm{GOF}}(\Delta_n; s)\} < 1.$$

Together, Theorems 8 and 9 suggest that Gaussian kernel embedding of distributions is especially suitable for testing against smooth alternatives, and it yields a test that could consistently detect the smallest departures, in terms of rate of convergence, from the null distribution. The idea can also be readily applied to testing of homogeneity and independence which we shall examine next.

3.2 Test for Homogeneity

As in the case of goodness of fit test, we shall consider the case when the underlying distributions have smooth densities so that we can rewrite the null hypothesis as $H_0^{\mathrm{HOM}} : p = q \in \mathcal{W}^{s,2}(M)$, and the alternative hypothesis as

$$H_1^{\mathrm{HOM}}(\Delta_n; s) : p, q \in \mathcal{W}^{s,2}(M), \quad \|p - q\|_{L_2} \geq \Delta_n.$$

The power of a test Φ based on $X_1, \ldots, X_n \sim p$ and $Y_1, \ldots, Y_m \sim q$ is given by

$$\mathrm{power}(\Phi; H_1^{\mathrm{HOM}}(\Delta_n; s)) := \inf_{p,q\in\mathcal{W}^{s,2}(M), \|p-q\|_{L_2}\geq\Delta_n} \mathbb{P}\{\Phi \text{ rejects } H_0^{\mathrm{HOM}}\}$$

To fix ideas, we shall also assume that $c \leq m/n \leq C$ for some constants $0 < c \leq C < \infty$. In addition, we shall express explicitly only the dependence on n and not m, for brevity. Our treatment, however, can be straightforwardly extended to more general situations.

Recall that

$$\gamma^2_{\nu_n}(\widehat{\mathbb{P}}_n, \widehat{\mathbb{Q}}_m) = \frac{1}{n^2} \sum_{1 \leq i,j \leq n} G_{\nu_n}(X_i, X_j) + \frac{1}{m^2} \sum_{1 \leq i,j \leq m} G_{\nu_n}(Y_i, Y_j) - \frac{2}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} G_{\nu_n}(X_i, Y_j).$$

As before, to reduce bias, we shall focus instead on a closely related estimate of $\gamma_{\nu_n}(\mathbb{P}, \mathbb{Q})$:

$$\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{Q}) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} G_{\nu_n}(X_i, X_j) + \frac{1}{m(m-1)} \sum_{1 \leq i \neq j \leq m} G_{\nu_n}(Y_i, Y_j) - \frac{2}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} G_{\nu_n}(X_i, Y_j).$$

It is easy to see that under H_0^{HOM},

$$\mathbb{E}\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{Q}) = 0,$$

and

$$\mathrm{var}\left(\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{Q})\right) = 2\left(\frac{1}{n(n-1)} + \frac{2}{mn} + \frac{1}{m(m-1)}\right) \mathbb{E}_{(X,Y) \sim \mathbb{P} \otimes \mathbb{Q}} \bar{G}^2_{\nu_n}(X, Y),$$

where

$$\bar{G}_{\nu_n}(x, y) = G_{\nu}(x, y) - \mathbb{E}_{X \sim \mathbb{P}} G_{\nu_n}(X, y) - \mathbb{E}_{Y \sim \mathbb{Q}} G_{\nu_n}(x, Y) + \mathbb{E}_{(X,Y) \sim \mathbb{P} \otimes \mathbb{Q}} G_{\nu_n}(X, Y).$$

It is therefore natural to consider estimating the variance by $\widehat{s}^2_{n,m,\nu_n} = \max\left\{\tilde{s}^2_{n,m,\nu_n}, 1/n^2\right\}$ where

$$\begin{aligned}\tilde{s}^2_{n,m,\nu_n} =& \frac{1}{N(N-1)} \sum_{1\le i\neq j\le N} G_{2\nu_n}(Z_i, Z_j) \\ &- \frac{2(N-3)!}{N!} \sum_{\substack{1\le i,j_1,j_2\le N \\ |\{i,j_1,j_2\}|=3}} G_{\nu_n}(Z_i, Z_{j_1}) G_{\nu_n}(Z_i, Z_{j_2}) \\ &+ \frac{(N-4)!}{N!} \sum_{\substack{1\le i_1,i_2,j_1,j_2\le N \\ |\{i_1,i_2,j_1,j_2\}|=4}} G_{\nu_n}(Z_{i_1}, Z_{j_1}) G_{\nu_n}(Z_{i_2}, Z_{j_2}),\end{aligned}$$

$N = n + m$ and $Z_i = X_i$ if $i \le n$ and Y_{i-n} if $i > n$. This leads to the following test statistic

$$T^{\mathrm{HOM}}_{n,\nu_n} = \frac{nm}{\sqrt{2}(n+m)} \cdot \widehat{s}^{-1}_{n,m,\nu_n} \cdot \widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{Q}).$$

As before, we can show

Theorem 10. *Let $\nu_n \to \infty$ as $n \to \infty$ in such a fashion that $\nu_n = o(n^{4/d})$. Then under H_0^{HOM} : $p = q \in \mathcal{W}^{s,2}(M)$,*

$$T^{\mathrm{HOM}}_{n,\nu_n} \to_d N(0,1), \qquad \text{as } n \to \infty.$$

Motivated by Theorem 10, we can consider a test, denoted by $\Phi^{\mathrm{HOM}}_{n,\nu_n,\alpha}$, that rejects H_0^{HOM} if and only if $T^{\mathrm{HOM}}_{n,\nu_n}$ exceeds z_α. By construction, $\Phi^{\mathrm{HOM}}_{n,\nu_n,\alpha}$ is an asymptotic α level test. We now turn to study its power against H_1^{HOM}. As in the case of goodness of fit test, we can prove that $\Phi^{\mathrm{HOM}}_{n,\nu_n,\alpha}$ is minimax optimal in that it can detect the smallest difference between p and q in terms of rate of convergence. More precisely, we have

Theorem 11. *(i) Assume that $n^{2s/(d+4s)}\Delta_n \to \infty$. Then for any $\alpha \in (0,1)$,*

$$\lim_{n\to\infty} \mathrm{power}\{\Phi^{\mathrm{HOM}}_{n,\nu_n,\alpha}; H_1^{\mathrm{HOM}}(\Delta_n; s)\} \to 1,$$

provided that $\nu_n \asymp n^{4/(d+4s)}$.

(ii) Conversely, if $\liminf_{n\to\infty} n^{2s/(d+4s)}\Delta_n < \infty$, then there exists some $\alpha \in (0,1)$ such that for

any test Φ_n of level α (asymptotically) based on $X_1, \ldots, X_n \sim p$ and $Y_1, \ldots, Y_m \sim q$,

$$\liminf_{n\to\infty} \text{power}\{\Phi_n; H_1^{\text{HOM}}(\Delta_n; s)\} < 1.$$

3.3 Test for Independence

Similarly, we can also use Gaussian kernel embedding to construct minimax optimal tests of independence. Let $X = (X^1, \ldots, X^k)^\top \in \mathbb{R}^d$ be a random vector where the subvectors $X^j \in \mathbb{R}^{d_j}$ for $j = 1, \ldots, k$ so that $d_1 + \cdots + d_k = d$. Denote by p the joint density function of X, and p_j the marginal density of X^j. We assume that both the joint density and the marginal densities are smooth. Specifically, we shall consider testing

$$H_0^{\text{IND}} : p = p_1 \otimes \cdots \otimes p_k,\ p_j \in \mathcal{W}^{s,2}(M_j),\ 1 \le j \le k$$

against a smooth departure from independence:

$$H_1^{\text{IND}}(\Delta_n; s) : p \in \mathcal{W}^{s,2}(M),\ p_j \in \mathcal{W}^{s,2}(M_j),\ 1 \le j \le k \text{ and } \|p - p_1 \otimes \cdots \otimes p_k\|_{L_2} \ge \Delta_n,$$

where $M = \prod_{j=1}^{k} M_j$ so that $p_1 \otimes \cdots \otimes p_k \in \mathcal{W}^{s,2}(M)$ under both null and alternative hypotheses.

Given a sample $\{X_1, \ldots, X_n\}$ of independent copies of X, we can naturally estimate the so-called dHSIC $\gamma^2_{\nu_n}(\mathbb{P}, \mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^k})$ by

$$\begin{aligned}
\gamma^2_{\nu_n}(\widehat{\mathbb{P}}_n, \widehat{\mathbb{P}}_n^{X^1} \otimes \cdots \otimes \widehat{\mathbb{P}}_n^{X^k}) \;=\; & \frac{1}{n^2} \sum_{1\le i,j\le n} G_{\nu_n}(X_i, X_j) \\
& + \frac{1}{n^{2k}} \sum_{1\le i_1,\ldots,i_k,j_1\ldots,j_k\le n} G_{\nu_n}((X^1_{i_1}, \ldots, X^k_{i_k}), (X^1_{j_1}, \ldots, X^k_{j_k})) \\
& - \frac{2}{n^{k+1}} \sum_{1\le i,j_1,\ldots,j_k\le n} G_{\nu_n}(X_i, (X^1_{j_1}, \ldots, X^k_{j_k})).
\end{aligned}$$

To correct for the bias, we shall consider the following estimate of $\gamma^2_{\nu_n}(\mathbb{P}, \mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^k})$ instead.

$$\begin{aligned}
&\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^k}) \\
=&\frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} G_{\nu_n}(X_i, X_j) \\
&+ \frac{(n-2k)!}{n!} \sum_{\substack{1 \le i_1, \cdots, i_k, j_1, \cdots, j_k \le n \\ |\{i_1, \cdots, i_k, j_1, \cdots, j_k\}| = 2k}} G_{\nu_n}((X^1_{i_1}, \ldots, X^k_{i_k}), (X^1_{j_1}, \ldots, X^k_{j_k})) \\
&- \frac{2(n-k-1)!}{n!} \sum_{\substack{1 \le i, j_1, \cdots, j_k \le n \\ |\{i, j_1, \cdots, j_k\}| = k+1}} G_{\nu_n}(X_i, (X^1_{j_1}, \ldots, X^k_{j_k})).
\end{aligned}$$

Under H_0^{IND}, we have

$$\mathbb{E}\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^k}) = 0.$$

Deriving its variance, however, requires a bit more work. Write

$$h_j(x^j, y) = \mathbb{E}_{X \sim \mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^k}} G_{\nu_n}((X^1, \ldots, X^{j-1}, x^j, X^{j+1}, \ldots, X^k), y)$$

and

$$g_j(x^j, y) = h_j(x^j, y) - \mathbb{E}_{X^j \sim \mathbb{P}^{X^j}} h_j(X^j, y) - \mathbb{E}_{Y \sim \mathbb{P}} h_j(x^j, Y) + \mathbb{E}_{(X^j, Y) \sim \mathbb{P}^{X^j} \otimes \mathbb{P}} h_j(X^j, Y).$$

With slight abuse of notation, also denote by

$$\begin{aligned}
h_{j_1, j_2}(x^{j_1}, y^{j_2}) = \mathbb{E}_{X, Y \sim_{\mathrm{iid}} \mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^k}} G_{\nu_n}((X^1, \ldots, X^{j_1 - 1}, x^{j_1}, X^{j_1 + 1}, \ldots, X^k), \\
(Y^1, \ldots, Y^{j_2 - 1}, y^{j_2}, Y^{j_2 + 1}, \ldots, Y^k))
\end{aligned}$$

and

$$g_{j_1,j_2}(x^{j_1}, y^{j_2}) = h_{j_1,j_2}(x^{j_1}, y^{j_2}) - \mathbb{E}_{X^{j_1}\sim\mathbb{P}^{X^{j_1}}} h_{j_1,j_2}(X^{j_1}, y^{j_2}) - \mathbb{E}_{X^{j_2}\sim\mathbb{P}^{X^{j_2}}} h_{j_1,j_2}(x^{j_1}, X^{j_2}) + \mathbb{E}_{(X^{j_1},Y^{j_2})\sim\mathbb{P}^{X^{j_1}}\otimes\mathbb{P}^{X^{j_2}}} h_{j_1,j_2}(X^{j_1}, Y^{j_2}).$$

Then we have

Lemma 1. *Under* H_0^{IND},

$$\operatorname{var}\left(\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^k})\right) = \frac{2}{n(n-1)}\left(\mathbb{E}\bar{G}^2_{\nu_n}(X,Y) - 2\sum_{1\le j\le k}\mathbb{E}\left(g_j(X^j,Y)\right)^2 + \sum_{1\le j_1,j_2\le k}\mathbb{E}\left(g_{j_1,j_2}(X^{j_1},Y^{j_2})\right)^2\right) + O(\mathbb{E}G_{2\nu_n}(X,Y)/n^3). \quad (3.4)$$

In light of Lemma 1, a variance estimator can be derived by estimating the leading term on the righthand side of (3.4) term by term using U-statistics. Formulae for estimating the variance for general k are tedious and we defer them to the appendix for space consideration. In the special case when $k = 2$, the leading term on the righthand side of (3.4) takes a much simplified form:

$$\frac{2}{n(n-1)}\mathbb{E}\bar{G}_{\nu_n}(X^1,Y^1)\cdot\mathbb{E}\bar{G}_{\nu_n}(X^2,Y^2),$$

where $X^j, Y^j \sim_{\mathrm{iid}} \mathbb{P}^{X^j}$ for $j = 1, 2$. Thus, we can estimate $\mathbb{E}[\bar{G}_{\nu_n}(X^j,Y^j)]^2$ by

$$\tilde{s}^2_{n,j,\nu_n} = \frac{1}{n(n-1)}\sum_{1\le i_1\ne i_2\le n} G_{2\nu_n}(X^j_{i_1}, X^j_{i_2}) - \frac{2(n-3)!}{n!}\sum_{\substack{1\le i,l_1,l_2\le n\\ |\{i,l_1,l_2\}|=3}} G_{\nu_n}(X^j_i, X^j_{l_1})G_{\nu_n}(X^j_i, X^j_{l_2}) + \frac{(n-4)!}{n!}\sum_{\substack{1\le i_1,i_2,l_1,l_2\le n\\ |\{i_1,i_2,l_1,l_2\}|=4}} G_{\nu_n}(X^j_{i_1}, X^j_{l_1})G_{\nu_n}(X^j_{i_2}, X^j_{l_2})$$

and $\mathrm{var}(\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{P}^{X^1} \otimes \mathbb{P}^{X^2}))$ by $2/[n(n-1)]\widehat{s}^2_{n,\nu_n}$ where

$$\widehat{s}^2_{n,\nu_n} := \max\left\{\tilde{s}^2_{n,1,\nu_n}\tilde{s}^2_{n,2,\nu_n}, 1/n^2\right\}.$$

so that a test statistic for H_0^{IND} is

$$T^{\mathrm{IND}}_{n,\nu_n} := \frac{n}{\sqrt{2}}\widehat{s}^{-1}_{n,\nu_n}\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{P}^{X^1} \otimes \mathbb{P}^{X^2}).$$

Test statistics for general $k > 2$ can be defined accordingly. Again, we have

Theorem 12. *Let $\nu_n \to \infty$ as $n \to \infty$ in such a fashion that $\nu_n = o(n^{4/d})$. Then under H_0^{IND},*

$$T^{\mathrm{IND}}_{n,\nu_n} \to_d N(0,1), \qquad \text{as } n \to \infty.$$

Motivated by Theorem 12, we can consider a test, denoted by $\Phi^{\mathrm{IND}}_{n,\nu_n,\alpha}$, that rejects H_0^{IND} if and only if $T^{\mathrm{IND}}_{n,\nu_n}$ exceeds z_α. By construction, $\Phi^{\mathrm{IND}}_{n,\nu_n,\alpha}$ is an asymptotic α level test. We now turn to study its power against H_1^{IND}. As in the case of goodness of fit test, we can prove that $\Phi^{\mathrm{HOM}}_{n,\nu_n,\alpha}$ is minimax optimal in that it can detect the smallest departure from independence in terms of rate of convergence. More precisely, we have

Theorem 13. *(i) Assume that $n^{2s/(d+4s)}\Delta_n \to \infty$. Then for any $\alpha \in (0,1)$,*

$$\lim_{n\to\infty} \mathrm{power}\{\Phi^{\mathrm{IND}}_{n,\nu_n,\alpha}; H_1^{\mathrm{IND}}(\Delta_n; s)\} \to 1,$$

provided that $\nu_n \asymp n^{4/(d+4s)}$.

(ii) Conversely, if $\liminf_{n\to\infty} n^{2s/(d+4s)}\Delta_n < \infty$, then there exists some $\alpha \in (0,1)$ such that for any test Φ_n of level α (asymptotically) based on $X_1, \ldots, X_n \sim p$,

$$\liminf_{n\to\infty} \mathrm{power}\{\Phi_n; H_1^{\mathrm{IND}}(\Delta_n; s)\} < 1.$$

3.4 Adaptation

The results presented in the previous sections not only suggest that Gaussian kernel embedding of distributions is especially suitable for testing against smooth alternatives, but also indicate the importance of choosing an appropriate scaling parameter in order to detect small deviation from the null hypothesis. To achieve maximum power, the scaling parameter should be chosen according to the smoothness of underlying density functions. This, however, presents a practical challenge because the level of smoothness is rarely known a priori. This naturally brings about the questions of adaption: can we devise an agnostic testing procedure that does not require such knowledge but still attain similar performance? We shall show in this section that this is possible, at least for sufficiently smooth densities.

3.4.1 Test for Goodness-of-fit

We again begin with the test for goodness-of-fit. As we show in Section 3.1, under H_0^{GOF}, $T_{n,\nu_n}^{\mathrm{GOF}} \to_d N(0,1)$ if $1 \ll \nu_n \ll n^{4/d}$; whereas for any $p \in \mathcal{W}^{s,2}$ such that $\|p - p_0\|_{L_2} \gg n^{-2s/(d+4s)}$, $T_{n,\nu_n}^{\mathrm{GOF}} \to \infty$ provided that $\nu_n \asymp n^{4/(d+4s)}$. This motivates us to consider the following test statistic:

$$T_n^{\mathrm{GOF(adapt)}} = \max_{1 \le \nu_n \le n^{2/d}} T_{n,\nu_n}^{\mathrm{GOF}}.$$

In light of earlier discussion, it is plausible that such a statistic could be used to detect any smooth departure from the null provided that the level of smoothness $s \ge d/4$. We now argue that this is indeed the case. More specifically, we shall proceed to reject H_0^{GOF} if and only if $T_n^{\mathrm{GOF(adapt)}}$ exceeds the upper α quantile, denoted by $q_{n,\alpha}^{\mathrm{GOF}}$, of its null distribution. In what follows, we shall call this test $\Phi^{\mathrm{GOF(adapt)}}$. Note that, even though it is hard to derive the analytic form for $q_{n,\alpha}^{\mathrm{GOF}}$, it can be readily evaluated via Monte Carlo method.

To study the power of $\Phi^{\mathrm{GOF(adapt)}}$ against H_1^{GOF} with different levels of smoothness, we shall

consider the following alternative hypothesis

$$H_1^{\mathrm{GOF(adapt)}}(\Delta_{n,s} : s \geq d/4) : p \in \bigcup_{s \geq d/4} \{p \in \mathcal{W}^{s,2}(M) : \|p - p_0\|_{L_2} \geq \Delta_{n,s}\}.$$

The following theorem characterizes the power of $\Phi^{\mathrm{GOF(adapt)}}$ against $H_1^{\mathrm{GOF(adapt)}}(\Delta_{n,s} : s \geq d/4)$.

Theorem 14. *There exists a constant $c > 0$ such that if*

$$\liminf_{n \to \infty} \Delta_{n,s} (n/\log\log n)^{2s/(d+4s)} > c,$$

then

$$\mathrm{power}\{\Phi^{\mathrm{GOF(adapt)}}; H_1^{\mathrm{GOF(adapt)}}(\Delta_{n,s} : s \geq d/4)\} \to 1.$$

Theorem 14 shows that $\Phi^{\mathrm{GOF(adapt)}}$ has a detection boundary of the order $(\log\log n/n)^{\frac{2s}{d+4s}}$ when $p \in \mathcal{W}^{s,2}$ for any $s \geq d/4$. If s is known in advance, as we show in Section 3.1, the optimal test is based on $T_{n,\nu_n}^{\mathrm{GOF}}$ with $\nu_n \asymp n^{4/(d+4s)}$ and has a detection boundary of the order $O(n^{-2s/(d+4s)})$. The extra polynomial of iterated logarithmic factor $(\log\log n)^{2s/(d+4s)}$ is the price we pay to ensure that no knowledge of s is required and $\Phi^{\mathrm{GOF(adapt)}}$ is powerful against smooth alternatives for all $s \geq d/4$.

3.4.2 Test for Homogeneity

The treatment for homogeneity tests is similar. Instead of $T_{n,\nu_n}^{\mathrm{HOM}}$, we now consider a test based on

$$T_n^{\mathrm{HOM(adapt)}} = \max_{1 \leq \nu_n \leq n^{2/d}} T_{n,\nu_n}^{\mathrm{HOM}}.$$

If $T_n^{\mathrm{HOM(adapt)}}$ exceeds the upper α quantile, denoted by $q_{n,\alpha}^{\mathrm{HOM}}$, of its null distribution, then we reject H_0^{HOM}. In what follows, we shall refer to this test as $\Phi^{\mathrm{HOM(adapt)}}$. As before, we do not have a closed form expression for $q_{n,\alpha}^{\mathrm{HOM}}$, and it needs to be evaluated via Monte Carlo method. In particular, in the case of homogeneity test, we can approximate $q_{n,\alpha}^{\mathrm{HOM}}$ by permutation where we

randomly shuffle $\{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$ and compute the test statistic as if the first n shuffled observations are from the first population whereas the other m are from the second population. This is repeated multiple times in order to approximate the critical value $q_{n,\alpha}^{\mathrm{HOM}}$.

The following theorem characterize the power of $\Phi^{\mathrm{HOM(adapt)}}$ against an alternative with different levels of smoothness

$$H_1^{\mathrm{HOM(adapt)}}(\Delta_{n,s} : s \geq d/4) : (p,q) \in \bigcup_{s \geq d/4} \{(p,q) : p, q \in \mathcal{W}^{s,2}(M), \|p - q\|_{L_2} \geq \Delta_{n,s}\}.$$

Theorem 15. *There exists a constant $c > 0$ such that if*

$$\liminf_{n \to \infty} \Delta_{n,s} (n/\log\log n)^{2s/(d+4s)} > c,$$

then

$$\mathrm{power}\{\Phi^{\mathrm{HOM(adapt)}}; H_1^{\mathrm{HOM(adapt)}}(\Delta_{n,s} : s \geq d/4)\} \to 1.$$

Similar to the case of goodness-of-fit test, Theorem 15 shows that $\Phi^{\mathrm{HOM(adapt)}}$ has a detection boundary of the order $O((n/\log\log n)^{-2s/(d+4s)})$ when $p \neq q \in \mathcal{W}^{s,2}$ for any $s \geq d/4$. In light of the results from Section 3.2, this is optimal up to an extra polynomial of iterated logarithmic factor. The main advantage is that $\Phi^{\mathrm{HOM(adapt)}}$ is powerful against smooth alternatives simultaneously for all $s \geq d/4$.

3.4.3 Test for Independence

Similarly, for independence test, we shall adopt the following test statistic

$$T_n^{\mathrm{IND(adapt)}} = \max_{1 \leq \nu_n \leq n^{2/d}} T_{n,\nu_n}^{\mathrm{IND}}.$$

and reject H_0^{IND} if and only $T_n^{\mathrm{IND(adapt)}}$ exceeds the upper α quantile, denoted by $q_{n,\alpha}^{\mathrm{IND}}$, of its null distribution. In what follows, we shall refer to this test as $\Phi^{\mathrm{HOM(adapt)}}$. The critical value, $q_{n,\alpha}^{\mathrm{HOM}}$, can also be evaluated via permutation test. See, *e.g.*, Pfister et al. (2018) for detailed discussions.

We now show that $\Phi^{\mathrm{IND(adapt)}}$ is powerful in testing against the alternative with different levels of smoothness

$$H_1^{\mathrm{IND(adapt)}}(\Delta_{n,s} : s \geq d/4) : p \in \bigcup_{s \geq d/4} \Big\{ p \in \mathcal{W}^{s,2}(M), p_j \in \mathcal{W}^{s,2}(M_j), 1 \leq j \leq k, \\ \|p - p_1 \otimes \cdots \otimes p_k\|_{L_2} \geq \Delta_{n,s} \Big\}.$$

More specifically, we have

Theorem 16. *There exists a constant $c > 0$ such that if*

$$\liminf_{n \to \infty} \Delta_{n,s} (n/\log\log n)^{2s/(d+4s)} > c,$$

then

$$\mathrm{power}\{\Phi^{\mathrm{IND(adapt)}}; H_1^{\mathrm{IND(adapt)}}(\Delta_{n,s} : s \geq d/4)\} \to 1.$$

Similar to before, Theorem 16 shows that $\Phi^{\mathrm{IND(adapt)}}$ is optimal up to an extra polynomial of iterated logarithmic factor for detecting smooth departure from independence simultaneously for all $s \geq d/4$.

Chapter 4: Numerical Experiments

To further complement our theoretical development and demonstrate the practical merits of the proposed methodology, we conducted several sets of numerical experiments. We shall mainly consider Gaussian kernels in this chapter as they are the most popular choices in practice for continuous data.

4.1 Effect of Scaling Parameter

Our first set of experiments were designed to illustrate the importance of the scaling parameter and highlight the potential room for improvement over the "median" heuristic—one of the most common data-driven choice of the scaling parameter in practice (see, *e.g.*, Gretton et al., 2008; Pfister et al., 2018).

- *Experiment I*: the homogeneity test with underlying distributions being the normal distribution and the mixture of several normal distributions. Specifically,

$$p(x) = f(x; 0, 1), \quad q(x) = 0.5 \times f(x; 0, 1) + 0.1 \times \sum_{\mu \in \boldsymbol{\mu}} f(x; \mu, 0.05)$$

  where $f(x; \mu, \sigma)$ denotes the density of $N(\mu, \sigma^2)$ and $\boldsymbol{\mu} = \{-1, -0.5, 0, 0.5, 1\}$.

- *Experiment II:* the joint independence test of $X^1, \cdots, X^5$ where

$$X^1, \cdots, X^4, (X^5)' \sim_{\text{iid}} N(0, 1), \quad X^5 = \left|(X^5)'\right| \times \text{sign}\left(\prod_{l=1}^{4} X^l\right).$$

  Clearly $X^1, \cdots, X^5$ are jointly dependent since $\prod_{l=1}^{d} X^l \geq 0$.

In both experiments, our primary goal is to investigate how the power of Gaussian MMD based

test is influenced by a pre-fixed scaling parameter. These tests are also compared to the ones with scaling parameter selected via "median" heuristic. In order to evaluate tests with different scaling parameters under a unified framework, we determined the critical values for each test via permutation test.

For Experiment I we fixed the sample size at $n = m = 200$; and for Experiment II at $n = 400$. The number of permutations was set at 100, and significance level at $\alpha = 0.05$. We first repeated the experiments 100 times under the null to verify that permutation tests indeed yield the correct size, up to Monte Carlo error. Each experiment was then repeated for 100 times and the observed power ($\pm$ one standard error) for different choices of the scaling parameter. The results are summarized in Figure 4.1. It is perhaps not surprising that the scaling parameter selected via "median heuristic" has little variation across each simulation run, and we represent its performance by a single value.

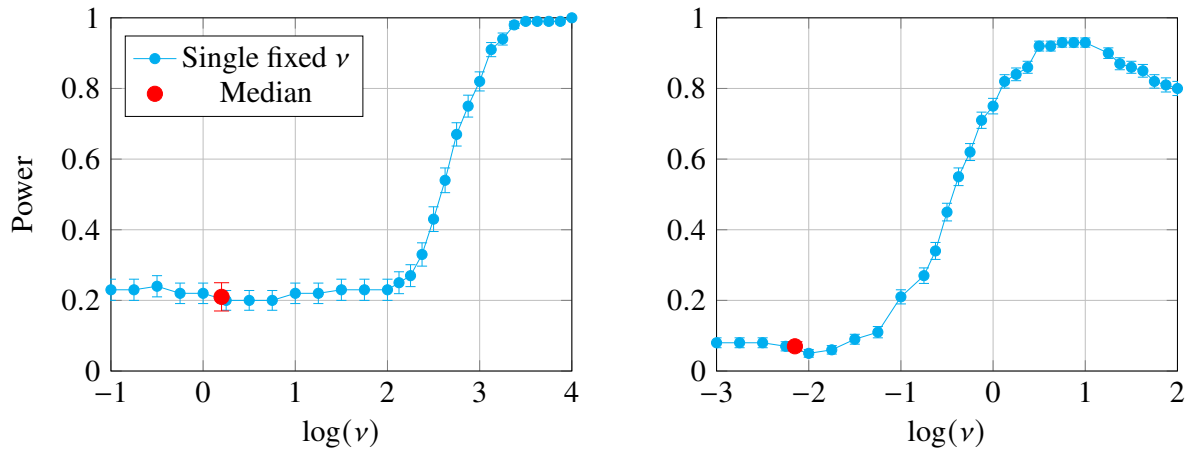

Figure 4.1: Observed power against $\log(\nu)$ in Experiment I (left) and Experiment II(right).

The importance of the scaling parameter is evident from Figure 4.1 with the observed power varies quite significantly for different choices. It is also of interest to note that in these settings the "median" heuristic typically does not yield a scaling parameter with great power. More specifically, in Experiment I, $\log(\nu_{\text{median}}) \approx 0.2$ and maximum power is attained at $\log(\nu) = 4$; in Experiment II, $\log(\nu_{\text{median}}) \approx -2.15$ and maximum power is attained at $\log(\nu) = 1$. This suggests that more appropriate choice of the scaling parameter may lead to much improved performance.

4.2 Efficacy of Adaptation

Our second experiment aims to illustrate that the adaptive procedures we proposed in Section 3.4 indeed yield more powerful tests when compared with other alternatives that are commonly used in practice. In particular, we compare the proposed self-normalized adaptive test (`S.A.`) with a couple of data-driven approaches, namely the "median" heuristic (`Median`) and the unnormalized adaptive test (`U.A.`) proposed in Sriperumbudur et al. (2009). When computing both self-normalized and unnormalized test statistics, we first rescaled the squared distance $\|X_i - X_j\|^2$ by the dimensionality d before taking maximum within a certain range of the scaling parameter. We considered two experiment setups:

- *Experiment III*: the homogeneity test with the underlying distributions being

$$P \sim N(\mathbf{0}, I_d), \quad Q \sim N\left(\mathbf{0}, \left(1 + 2d^{-1/2}\right) I_d\right).$$

  As the 'signal strength', the ratio between the variances of Q and P in each single direction is set to decrease to 1 at the order $1/\sqrt{d}$ with d, which is the decreasing order of variance ratio that can be detected by the classical F-test.

- *Experiment IV*: the independence test of $X^1, X^2 \in \mathbb{R}^{d/2}$, where $X = (X^1, X^2)$ follows a mixture of

$$N(\mathbf{0}, I_d) \quad \text{and} \quad N\left(\mathbf{0}, (1 + 6d^{-3/5}) I_d\right)$$

  with mixture probability being 0.5. Similarly, the ratio between the variances in each direction is set to decrease with d, but at a slightly higher rate.

To better compare different methods, we considered different combinations of sample size and dimensionality for each experiment. More specifically, for Experiment III, the sample sizes were set to be $m = n = 25, 50, 75, \cdots, 200$ and dimension $d = 1, 10, 100, 1000$; for Experiment IV, the sample size were $n = 100, 200, \cdots, 600$ and dimension $d = 2, 10, 100, 1000$. In both experiments,

we fixed the significance level at $\alpha = 0.05$, did 100 permutations to calibrate the critical values as before. Again we simulated under H_0 to verify that the resulting tests have the targeted size, up to Monte Carlo error. The power of each method, estimated from 100 such experiments, is reported in Figures 4.2 and 4.3.

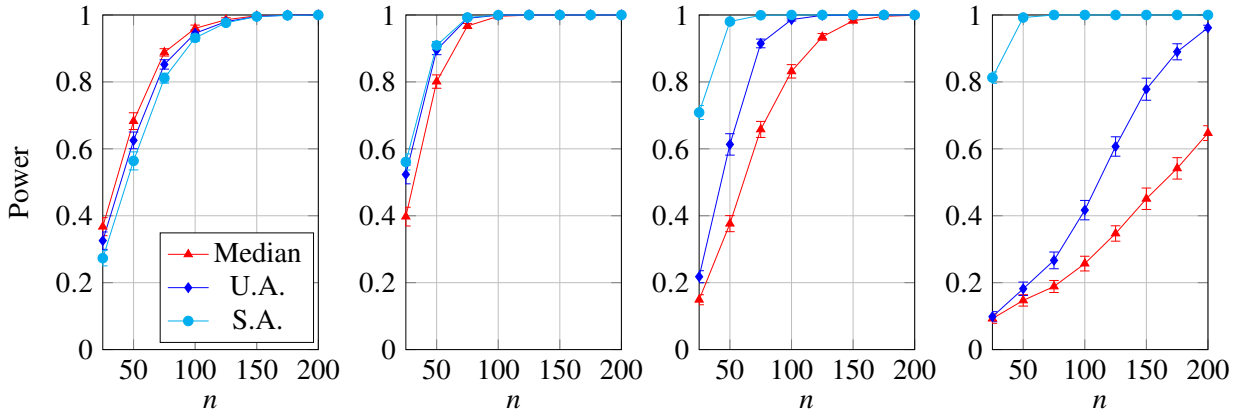

Figure 4.2: Observed power versus sample size in Experiment III for $d = 1, 10, 100, 1000$ from left to right.

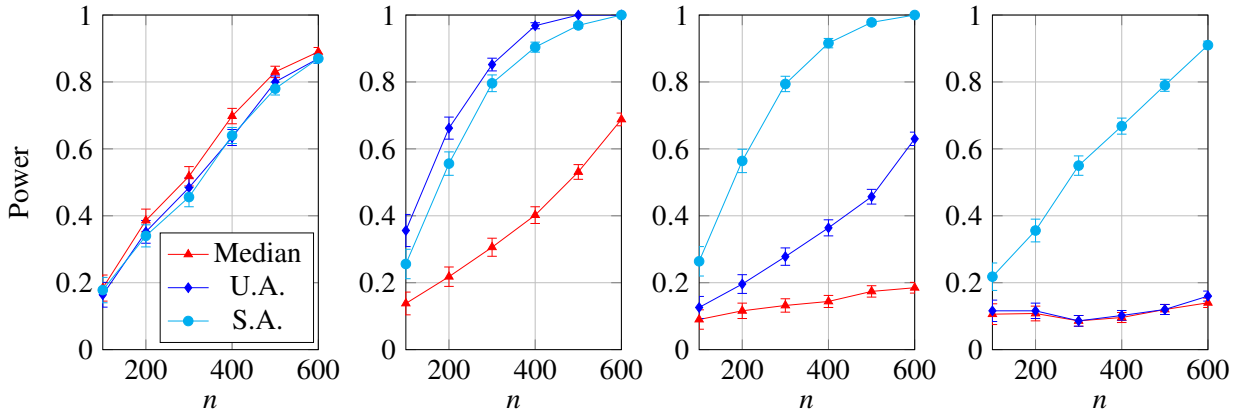

Figure 4.3: Observed power versus sample size in Experiment IV for $d = 2, 10, 100, 1000$ from left to right.

As Figures 4.2 and 4.3 show, for both experiments, these tests are comparable in low-dimensional settings. But as d increases, the proposed self-normalized adaptive test becomes more and more preferable to the two alternatives. For example, for Experiment IV, when $d = 1000$, the observed power of the proposed self-normalized adaptive test is about 90% when $n = 600$, while the other two tests have power around only 15%.

4.3 Data Example

Finally, we considered applying the proposed self-normalized adaptive test in a data example from Mooij et al. (2016). The data set consists of three variables, altitude (Alt), average temperature (Temp) and average duration of sunshine (Sun) from different weather stations. One goal of interest is to figure out the causal relationship among the three variables by figuring out a suitable directed acyclic graph (DAG) among them. Following Peters et al. (2014), if a set of random variables $X^1, \cdots, X^d$ follow a DAG $\mathcal{G}_0$, then we assume that they follow a sequence of additive models:

$$X^l = \sum_{r \in \mathrm{PA}^l} f_{l,r}(X^r) + N^l, \quad \forall\ 1 \le l \le d,$$

where N^l's are independent Gaussian noises and PA^l denotes the collection of parent nodes of node l specified by $\mathcal{G}_0$. As shown by (Peters et al., 2014), $\mathcal{G}_0$ is identifiable from the joint distribution of $X^1, \cdots, X^d$ under the assumption of $f_{l,r}$'s being non-linear. Therefore a natural method of deciding a specific DAG underlying a set of random variables is by testing the independence of the regression residuals after fitting the DAG induced additive models. In our case, there are totally 25 possible DAGs for the three variables. We can apply independence tests for the residuals for each of the 25 DAGs and choose the one with the largest p-value as the most plausible underlying DAG. See Peters et al. (2014) for more details.

As before, we considered three different ways for independence tests: the proposed self-normalized adaptive test (`S.A.`), Gaussian kernel embedding based independent test with the scaling parameter determined by the "median" heuristic (`Median`), and the unnormalized adaptive test from Sriperumbudur et al. (2009) (`U.A.`). Note that the three variables have different scales and we standardize them before applying the tests of independence.

The overall sample size of the data set is 349. Each time we randomly select 150 samples and compute the p-value associated with each DAG. The p-value is again computed based on 100 permutations. We repeated the experiment for 1000 times and recorded for each test the DAG with the largest p-value. All three tests agree on the top three most selected DAGs and they are shown

in Figure 4.4.

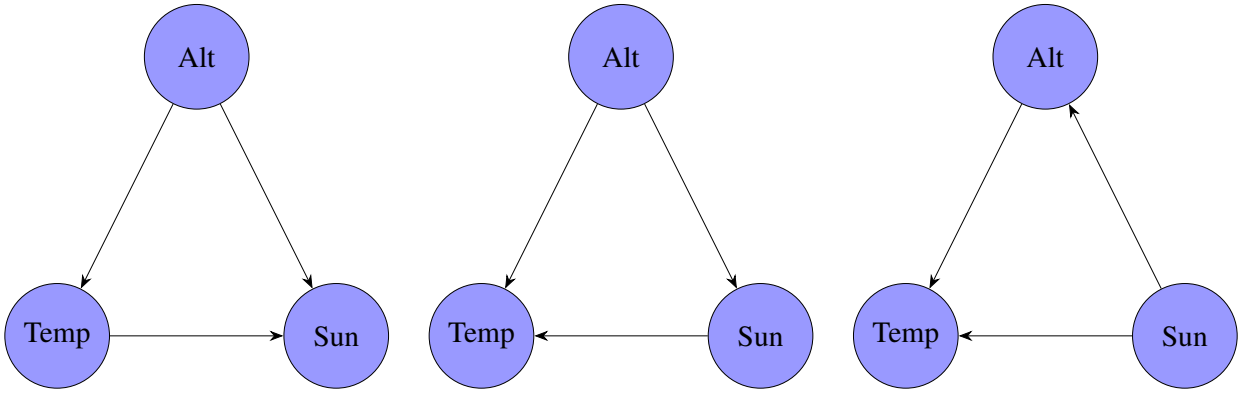

Figure 4.4: DAGs with the top 3 highest probabilities of being selected.

In addition, we report in Table 4.1 the frequencies that these three DAGs were selected by each of the tests. They are generally comparable with the proposed method more consistently selecting DAG I, the one heavily favored by all three methods.

Prob(%) DAG / Test	I	II	III
Median	78.5	4.7	14.5
U.A.	81.4	8.1	8.5
S.A.	83.4	9.8	4.7

Table 4.1: Frequency that each DAG in Figure 4.4 was selected by three tests.

Chapter 5: Conclusion and Discussion

In this thesis, we aim to address the problem of kernel selection when using kernel embedding for the purpose of nonparametric hypothesis testing, which is an inevitable problem that researchers and practitioners have been trying to answer ever since the first proposal of kernel embedding method while most of the existing solutions are ad-hoc. We propose principled ways of kernel selection in two different settings which are proved to ensure minimax rate optimality for the associated tests. We also propose adaptive test statistics to address the issue of aforementioned kernel selection methods depending on the regularity condition of the underlying space of probability distributions, whose sacrifice in terms of detection boundary is only some polynomial of iterated logarithmic factor of the sample size.

There are still many interesting problems in this area which remain to be explored further. For example, can we adopt fast computation techniques to compute the kernel based test statistic approximately so as to reduce the computation complexity to a large extent while maintaining the statistical optimality? Parallel results in the context of regression have been derived but there seems to be a lack of such results in hypothesis testing. The second one involves resampling methods such as permutation method and bootstrap method. In practice, with the concern that the sample size may not be large enough, resampling methods are usually used to decide the rejection boundary. Can we still ensure the statistical optimality of the proposed tests when incorporating resampling methods?

In addition to that, it is also wondered whether similar principled kernel selection methods can be proposed in a broader range of nonparametric testing problems such as conditional independent test, which can be very useful in Bayesian network learning and causal discovery.

Chapter 6: Proofs

Throughout this chapter, we shall write $a_n \lesssim b_n$ if there exists a universal constant $C > 0$ such that $a_n \leq Cb_n$. Similarly, we write $a_n \gtrsim b_n$ if $b_n \lesssim a_n$, and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. When the the constant depends on another quantity D, we shall write $a_n \lesssim_D b_n$. Relations $\gtrsim_D$ and $\asymp_D$ are defined accordingly.

Proof of Theorem 1. **Part (i).** The proof of the first part consists of two key steps. First, we show that the population counterpart $n\gamma^2(\mathbb{P}, \mathbb{P}_0)$ of the test statistic converges to ∞ uniformly, *i.e.*,

$$n \inf_{P \in \mathcal{P}(\Delta_n, 0)} \gamma^2(\mathbb{P}, \mathbb{P}_0) \to \infty.$$

Then, we argue that the deviation from $\gamma^2(\mathbb{P}, \mathbb{P}_0)$ to $\gamma^2(\widehat{\mathbb{P}}_n, \mathbb{P}_0)$ is uniformly negligible compared with $\gamma^2(\mathbb{P}, \mathbb{P}_0)$ itself.

It is not hard to see that

$$\begin{aligned}\gamma(\widehat{\mathbb{P}}_n, \mathbb{P}_0) =& \sqrt{\sum_{k \geq 1} \lambda_k \Big[\frac{1}{n} \sum_{i=1}^n \phi_k(X_i)\Big]^2} \\ \geq& \sqrt{\sum_{k \geq 1} \lambda_k [\mathbb{E}_{\mathbb{P}} \phi_k(X)]^2} - \sqrt{\sum_{k \geq 1} \lambda_k \Big[\frac{1}{n} \sum_{i=1}^n \phi_k(X_i) - \mathbb{E}_{\mathbb{P}} \phi_k(X)\Big]^2}.\end{aligned}$$

Thus,

$$
\begin{aligned}
&P\left\{n\gamma^2(\widehat{\mathbb{P}}_n, \mathbb{P}_0) < q_{w,1-\alpha}\right\} \\
\leq &P\left\{\sqrt{n\sum_{k\geq 1}\lambda_k[\mathbb{E}_{\mathbb{P}}\phi_k(X)]^2} - \sqrt{n\sum_{k\geq 1}\lambda_k\Big[\frac{1}{n}\sum_{i=1}^n \phi_k(X_i) - \mathbb{E}_{\mathbb{P}}\phi_k(X)\Big]^2} < \sqrt{q_{w,1-\alpha}}\right\} \\
= &P\left\{\sqrt{n\sum_{k\geq 1}\lambda_k\Big[\frac{1}{n}\sum_{i=1}^n \phi_k(X_i) - \mathbb{E}_{\mathbb{P}}\phi_k(X)\Big]^2} > \sqrt{n\sum_{k\geq 1}\lambda_k[\mathbb{E}_{\mathbb{P}}\phi_k(X)]^2} - \sqrt{q_{w,1-\alpha}}\right\}.
\end{aligned}
$$

Suppose that

$$
n\sum_{k\geq 1}\lambda_k[\mathbb{E}_{\mathbb{P}}\phi_k(X)]^2 > q_{w,1-\alpha}.
$$

Then

$$
P\left\{n\gamma^2(\widehat{\mathbb{P}}_n, \mathbb{P}_0) < q_{w,1-\alpha}\right\} \leq \frac{\mathbb{E}_{\mathbb{P}}\Big\{n\sum\limits_{k\geq 1}\lambda_k\Big[\frac{1}{n}\sum\limits_{i=1}^n \phi_k(X_i) - \mathbb{E}_{\mathbb{P}}\phi_k(X)\Big]^2\Big\}}{\left\{\sqrt{n\sum\limits_{k\geq 1}\lambda_k[\mathbb{E}_{\mathbb{P}}\phi_k(X)]^2} - \sqrt{q_{w,1-\alpha}}\right\}^2}.
$$

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

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

This implies that

$$\begin{aligned}\lim_{n\to\infty}\beta(\Phi_{\mathrm{MMD}};\Delta_n,0) &= \lim_{n\to\infty}\sup_{\mathbb{P}\in\mathcal{P}(\Delta_n,0)} P\left\{n\gamma^2(\widehat{\mathbb{P}}_n,\mathbb{P}_0) < q_{w,1-\alpha}\right\}\\ &\le \lim_{n\to\infty}\frac{\sup_{\mathbb{P}\in\mathcal{P}(\Delta_n,0)}\mathbb{E}_{\mathbb{P}}\left\{n\sum_{k\ge1}\lambda_k\left[\frac{1}{n}\sum_{i=1}^{n}\phi_k(X_i)-\mathbb{E}_{\mathbb{P}}\phi_k(X)\right]^2\right\}}{\inf_{\mathbb{P}\in\mathcal{P}(\Delta_n,0)}\left\{\sqrt{n\sum_{k\ge1}\lambda_k[\mathbb{E}_{\mathbb{P}}\phi_k(X)]^2}-\sqrt{q_{w,1-\alpha}}\right\}^2}\\ &=0,\end{aligned}$$

provided that

$$\inf_{\mathbb{P}\in\mathcal{P}(\Delta_n,0)} n\sum_{k\ge1}\lambda_k[\mathbb{E}_{\mathbb{P}}\phi_k(X)]^2\to\infty, \qquad \text{as } n\to\infty. \tag{6.1}$$

It now suffices to show that (6.1) holds if $n\Delta_n^4\to\infty$ as $n\to\infty$.

To this end, let $u = d\mathbb{P}/d\mathbb{P}_0 - 1$ and

$$a_k = \langle u,\phi_k\rangle_{L_2(\mathbb{P}_0)} = \mathbb{E}_{\mathbb{P}}\phi_k(X) - \mathbb{E}_{\mathbb{P}_0}\phi_k(X) = \mathbb{E}_{\mathbb{P}}(\phi_k(X)).$$

It is clear the that

$$\sum_{k\ge1}\lambda_k^{-1}a_k^2 = \|u\|_K^2, \qquad \text{and} \qquad \sum_{k\ge1}a_k^2 = \|u\|_{L_2(\mathbb{P}_0)}^2 = \chi^2(\mathbb{P},\mathbb{P}_0).$$

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

$$\sup_{\mathbb{P}\in\mathcal{P}(\Delta_n,0)}\sum_{k\ge1}\lambda_k^{-1}a_k^2 \le M^2, \qquad \text{and} \qquad \inf_{\mathbb{P}\in\mathcal{P}(\Delta_n,0)}\sum_{k\ge1}a_k^2 \ge \Delta_n.$$

Since $n\Delta_n^4 \to \infty$ as $n \to \infty$, we get

$$\begin{aligned}\inf_{\mathbb{P}\in\mathcal{P}(\Delta_n,0)} n\sum_{k\geq 1}\lambda_k[\mathbb{E}_{\mathbb{P}}\phi_k(X)]^2 &= \inf_{\mathbb{P}\in\mathcal{P}(\Delta_n,0)} n\sum_{k\geq 1}\lambda_k a_k^2 \\ &\geq \inf_{\mathbb{P}\in\mathcal{P}(\Delta_n,0)} n\frac{\Big(\sum_{k\geq 1} a_k^2\Big)^2}{\sum_{k\geq 1}\lambda_k^{-1}a_k^2} \\ &\geq \frac{n\Delta_n^4}{M^2} \to \infty\end{aligned}$$

as $n \to \infty$.

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 $\mathbb{P}_n$ such that $\mathbb{P}_n$ converges to $\mathbb{P}_0$ in the order $n^{-1/4}$.

Lemma 2. *Consider a sequence of probability measures $\{\mathbb{P}_n : n \geq 1\}$ contiguous to $\mathbb{P}_0$ satisfying $u_n = d\mathbb{P}_n/d\mathbb{P}_0 - 1 \to 0$ in $L_2(\mathbb{P}_0)$. Suppose that for any fixed k,*

$$\lim_{n\to\infty}\sqrt{n}\langle u_n,\phi_k\rangle_{L_2(\mathbb{P}_0)} = \tilde{a}_k, \qquad \text{and} \qquad \lim_{n\to\infty}\sum_{k\geq 1}\lambda_k(\sqrt{n}\langle u_n,\phi_k\rangle_{L_2(\mathbb{P}_0)})^2 = \sum_{k\geq 1}\lambda_k\tilde{a}_k^2 + \tilde{a}_0 < \infty,$$

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

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

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

Write $L(k) = \lambda_k k^{2s}$. By assumption (2.6),

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

Consider a sequence of $\{\mathbb{P}_n : n \geq 1\}$ such that

$$d\mathbb{P}_n/d\mathbb{P}_0 - 1 = C_1\sqrt{\lambda_{k_n}}[L(k_n)]^{-1}\phi_{k_n},$$

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 $\sup_{k\geq 1} \|\phi_k\|_\infty < \infty$ and $\lim_{k\to\infty} \lambda_k = 0$, there exists $N_0 > 0$ such that $\mathbb{P}_n$'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

$$\|u_n\|_{L_2(\mathbb{P}_0)}^2 = \frac{C_1^2 \lambda_{k_n}}{L^2(k_n)} = \frac{C_1^2}{L(k_n)} k_n^{-2s} \geq \overline{L}^{-1} C_1^2 k_n^{-2s} \sim \overline{L}^{-1} C_1^2 C_2^{-2s} n^{-1/2},$$

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 $\mathbb{P}_n \in \mathcal{P}(c_0 n^{-1/4}, 0)$ for sufficiently large n.

To apply Lemma 2, we note that

$$\lim_{n\to\infty} \|u_n\|_{L_2(\mathbb{P}_0)}^2 = \lim_{n\to\infty} \frac{C_1^2 \lambda_{k_n}}{L^2(k_n)} = 0.$$

In addition, for any fixed k,

$$\tilde{a}_{n,k} = \sqrt{n}\langle u_n, \phi_k \rangle_{L_2(\mathbb{P}_0)} = 0$$

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 2 implies that

$$n\gamma(\widehat{\mathbb{P}}_n, \mathbb{P}_0) \xrightarrow{d} \sum_{k\geq 1} \lambda_k Z_k^2 + C_1^2 C_2^{-4s}.$$

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

$$\begin{aligned}\liminf_{n\to\infty} \beta(\Phi_{\mathrm{MMD}}; c_0 n^{-1/2}, 0) \geq & \lim_{n\to\infty} P(n\gamma(\widehat{\mathbb{P}}_n, \mathbb{P}_0) < q_{w,1-\alpha}) \\ = & P\Big(\sum_{k\geq 1} \lambda_k Z_k^2 < \frac{1}{2} q_{w,1-\alpha} \Big) > 0,\end{aligned}$$

which concludes the proof. □

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

$$v_n^{-1/2}[n\eta_{\varrho_n}^2(\widehat{\mathbb{P}}_n, \mathbb{P}_0) - A_n] = 2(n^2 v_n)^{-1/2} \sum_{j=2}^{n} \sum_{i=1}^{j-1} \tilde{K}_n(X_i, X_j).$$

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 $\mathbb{E}\phi_k(X) = 0$ 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

$$v_n^{-1/2}[n\eta^2_{\varrho_n}(\widehat{\mathbb{P}}_n, \mathbb{P}_0) - A_n] = 2(n^2 v_n)^{-1/2} U_{nn}.$$

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

$$\left[\frac{1}{2}n^2\mathbb{E}\tilde{K}_n^2(X, X')\right]^{-1/2} U_{nn} \xrightarrow{d} N(0, 1), \tag{6.2}$$

provided that

$$[\mathbb{E}G_n^2(X, X') + n^{-1}\mathbb{E}\tilde{K}_n^2(X, X')\tilde{K}_n^2(X, X'') + n^{-2}\mathbb{E}\tilde{K}_n^4(X, X')]/[\mathbb{E}\tilde{K}_n^2(X, X')]^2 \to 0, \tag{6.3}$$

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

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

(6.2) implies that

$$v_n^{-1/2}[n\eta^2_{\varrho_n}(\widehat{\mathbb{P}}_n, \mathbb{P}_0) - A_n] = \sqrt{2} \cdot \left(\frac{1}{2}n^2\mathbb{E}\tilde{K}_n^2(X, X')\right)^{-1/2} U_{nn} \xrightarrow{d} N(0, 2).$$

It therefore suffices to verify (6.3).

Note that

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

where the last step holds by considering that $\lambda_k \asymp k^{-2s}$. Similarly,

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

and

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

Thus there exists a positive constant C_3 such that

$$\mathbb{E}G_n^2(X,X')/[\mathbb{E}\tilde{K}_n^2(X,X')]^2 \leq C_3\varrho_n^{1/s} \to 0, \tag{6.4}$$

and

$$n^{-1}\mathbb{E}\tilde{K}_n^2(X,X')\tilde{K}_n^2(X,X'')/[\mathbb{E}\tilde{K}_n^2(X,X')]^2 \leq C_3 n^{-1} \to 0, \tag{6.5}$$

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

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

where

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

This implies that for some positive constant C_4,

$$n^{-2}\mathbb{E}\tilde{K}_n^4(X,X')\}/[\mathbb{E}\tilde{K}_n^2(X,X')]^2 \leq n^{-2}\|\tilde{K}_n\|_\infty^2/\mathbb{E}\tilde{K}_n^2(X,X') \leq C_4(n^2\varrho_n^{1/s})^{-1} \to 0. \tag{6.6}$$

as $n \to \infty$. Together, (6.4), (6.5) and (6.6) ensure that condition (6.3) holds. □

Proof of Theorem 3. Note that

$$
\begin{aligned}
& n\eta^2_{\varrho_n}(\widehat{\mathbb{P}}_n, \mathbb{P}_0) - \frac{1}{n}\sum_{i=1}^{n} \tilde{K}_n(X_i, X_i) \\
=& \frac{1}{n}\sum_{k\geq 1} \frac{\lambda_k}{\lambda_k + \varrho_n^2} \sum_{\substack{1\leq i,j\leq n \\ i\neq j}} \phi_k(X_i)\phi_k(X_j) \\
=& \frac{1}{n}\sum_{k\geq 1} \frac{\lambda_k}{\lambda_k + \varrho_n^2} \sum_{\substack{1\leq i,j\leq n \\ i\neq j}} [\phi_k(X_i) - \mathbb{E}_{\mathbb{P}}\phi_k(X)][\phi_k(X_j) - \mathbb{E}_{\mathbb{P}}\phi_k(X)] \\
& + \frac{2(n-1)}{n}\sum_{k\geq 1} \frac{\lambda_k}{\lambda_k + \varrho_n^2} [\mathbb{E}_{\mathbb{P}}\phi_k(X)] \sum_{1\leq i\leq n} [\phi_k(X_i) - \mathbb{E}_{\mathbb{P}}\phi_k(X)] \\
& + \frac{n(n-1)}{n}\sum_{k\geq 1} \frac{\lambda_k}{\lambda_k + \varrho_n^2} [\mathbb{E}_{\mathbb{P}}\phi_k(X)]^2 \\
:=& V_1 + V_2 + V_3.
\end{aligned}
$$

Obviously, $\mathbb{E}_{\mathbb{P}} V_1 V_2 = 0$. We first argue that the following three statements together implies the desired result:

$$
\begin{aligned}
\lim_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}(\Delta_n,\theta)} v_n^{-1/2} V_3 &= \infty, && (6.7) \\
\sup_{\mathbb{P}\in\mathcal{P}(\Delta_n,\theta)} (\mathbb{E}_{\mathbb{P}} V_1^2 / V_3^2) &= o(1), && (6.8) \\
\sup_{\mathbb{P}\in\mathcal{P}(\Delta_n,\theta)} (\mathbb{E}_{\mathbb{P}} V_2^2 / V_3^2) &= o(1). && (6.9)
\end{aligned}
$$

To see this, note that (6.7) implies that

$$
\begin{aligned}
& \lim_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}(\Delta_n,\theta)} P(v_n^{-1/2}[n\eta^2_{\varrho_n}(\widehat{\mathbb{P}}_n, \mathbb{P}_0) - A_n] \geq \sqrt{2} z_{1-\alpha}) \\
\geq & \lim_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}(\Delta_n,\theta)} P\Big(v_n^{-1/2} V_3 \geq 2\sqrt{2} z_{1-\alpha}, V_1 + V_2 + V_3 \geq \frac{1}{2} V_3\Big) \\
= & \lim_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}(\Delta_n,\theta)} P\Big(V_1 + V_2 + V_3 \geq \frac{1}{2} V_3\Big).
\end{aligned}
$$

On the other hand, (6.8) and (6.9) imply that

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

This immediately suggests that $\Phi_{\mathrm{M^3d}}$ is consistent. We now show that (6.7)-(6.9) indeed hold.

Verifying (6.7). We begin with (6.7). Since $v_n \asymp \varrho_n^{-1/s}$ and $V_3 = (n-1)\eta^2_{\varrho_n}(\mathbb{P},\mathbb{P}_0)$, (6.7) is equivalent to

$$\lim_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}(\Delta_n,\theta)} n\varrho_n^{\frac{1}{2s}}\eta^2_{\varrho_n}(\mathbb{P},\mathbb{P}_0)=\infty.$$

For any $\mathbb{P}\in\mathcal{P}(\Delta_n,\theta)$, let $u = d\mathbb{P}/d\mathbb{P}_0 - 1$ and $a_k = \langle u,\phi_k\rangle_{L_2(\mathbb{P}_0)} = \mathbb{E}_{\mathbb{P}}\phi_k(X)$. Based on the assumption that K is universal, $u = \sum_{k\geq 1} a_k\phi_k$. We consider the case $\theta = 0$ and $\theta > 0$ separately.

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

$$\begin{aligned}\eta^2_{\varrho_n}(\mathbb{P},\mathbb{P}_0) =& \sum_{k\geq 1} a_k^2 - \sum_{k\geq 1}\frac{\varrho_n^2}{\lambda_k+\varrho_n^2}a_k^2\\ \geq& \|u\|^2_{L_2(\mathbb{P}_0)} - \varrho_n^2\sum_{k\geq 1}\frac{1}{\lambda_k}a_k^2\\ \geq& \|u\|^2_{L_2(\mathbb{P}_0)} - \varrho_n^2 M^2.\end{aligned}$$

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

$$\inf_{\mathbb{P}\in\mathcal{P}(\Delta_n,0)}\eta^2_{\varrho_n}(\mathbb{P},\mathbb{P}_0)\geq\frac{1}{2}\inf_{\mathbb{P}\in\mathcal{P}(\Delta_n,0)}\|u\|^2_{L_2(\mathbb{P}_0)}=\frac{1}{2}\Delta_n^2.$$

(2) Now consider the case when $\theta > 0$. For $\mathbb{P}\in\mathcal{P}(\Delta_n,\theta)$, $\forall\ R>0$, $\exists\ f_R\in\mathcal{H}(K)$ such that

$\|u - f_R\|_{L_2(\mathbb{P}_0)} \le MR^{-1/\theta}$ and $\|f_R\|_K \le R$. Let $b_k = \langle f_R, \phi_k \rangle_{L_2(\mathbb{P}_0)}$.

$$\begin{aligned}
\eta^2_{\varrho_n}(\mathbb{P}, \mathbb{P}_0) =& \sum_{k\ge 1} a_k^2 - \sum_{k\ge 1} \frac{\varrho_n^2}{\lambda_k + \varrho_n^2} a_k^2 \\
\ge& \|u\|^2_{L_2(\mathbb{P}_0)} - 2\sum_{k\ge 1} \frac{\varrho_n^2}{\lambda_k + \varrho_n^2}(a_k - b_k)^2 - 2\sum_{k\ge 1} \frac{\varrho_n^2}{\lambda_k + \varrho_n^2} b_k^2 \\
\ge& \|u\|^2_{L_2(\mathbb{P}_0)} - 2\sum_{k\ge 1}(a_k - b_k)^2 - 2\varrho_n^2 \sum_{k\ge 1} \frac{1}{\lambda_k} b_k^2 \\
=& \|u\|^2_{L_2(\mathbb{P}_0)} - 2\|u - f_R\|^2_{L_2(\mathbb{P}_0)} - 2\varrho_n^2 \|f_R\|^2_K.
\end{aligned}$$

Taking $R = (2M/\|u\|_{L_2(\mathbb{P}_0)})^\theta$ yields that

$$\eta^2_{\varrho_n}(\mathbb{P}, \mathbb{P}_0) \ge \|u\|^2_{L_2(\mathbb{P}_0)} - 2M^2 R^{-2/\theta} - 2\varrho_n^2 R^2 = \frac{1}{2}\|u\|^2_{L_2(\mathbb{P}_0)} - 2\varrho_n^2 R^2.$$

Now by choosing

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

we can ensure that

$$2\varrho_n^2 R^2 \le \frac{1}{4}\|u\|^2_{L_2(\mathbb{P}_0)}.$$

So that

$$\inf_{\mathbb{P}\in\mathcal{P}(\Delta_n,\theta)} \eta^2_{\varrho_n}(\mathbb{P}, \mathbb{P}_0) \ge \inf_{\mathbb{P}\in\mathcal{P}(\Delta_n,\theta)} \frac{1}{4}\|u\|^2_{L_2(\mathbb{P}_0)} \ge \frac{1}{4}\Delta_n^2.$$

In both cases, with $\varrho_n \le C\Delta_n^{\theta+1}$ for a sufficiently small $C = C(M) > 0$, $\lim_{n\to\infty} \varrho_n^{\frac{1}{2s}} n\Delta_n^2 = \infty$ suffices to ensure (6.7) holds. Under the condition that $\lim_{n\to\infty} \Delta_n n^{\frac{2s}{4s+\theta+1}} = \infty$,

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

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

Verifying (6.8)**.** Rewrite V_1 as

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

Then

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

where $X, X' \overset{i.i.d.}{\sim} \mathbb{P}$. 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

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

we have

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

Thus, to prove (6.8), it suffices to show that

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

For any $g \in L_2(\mathbb{P}_0)$ and positive definite kernel $G(\cdot,\cdot)$ such that $\mathbb{E}_{\mathbb{P}_0}G^2(X,X') < \infty$, let

$$\|g\|_G := \sqrt{\mathbb{E}_{\mathbb{P}_0}[g(X)g(X')G(X,X')]}.$$

By the positive definiteness of $G(\cdot,\cdot)$, triangular inequality holds for $\|\cdot\|_G$, *i.e.*, for any $g_1, g_2 \in L_2(\mathbb{P}_0)$,

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

Thus by taking $G = \tilde{K}_n^2$, $g_1 = d\mathbb{P}/d\mathbb{P}_0$ and $g_2 = 1$, we have

$$\left|\sqrt{\mathbb{E}_{\mathbb{P}}\tilde{K}_n^2(X,X')} - \sqrt{\mathbb{E}_{\mathbb{P}_0}\tilde{K}_n^2(X,X')}\right| \le \sqrt{\mathbb{E}_{\mathbb{P}_0}[u(X)u(X')\tilde{K}_n^2(X,X')]}. \tag{6.10}$$

We now appeal to the following lemma to bound the right hand side of (6.10):

Lemma 3. *Let G be a Mercer kernel defined over $\mathcal{X}\times\mathcal{X}$ with eigenvalue-eigenfunction pairs $\{(\mu_k,\phi_k) : k \ge 1\}$ with respect to $L_2(\mathbb{P})$ 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 $g \in L_2(\mathbb{P})$*

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

By Lemma 3, we get

$$\mathbb{E}_{\mathbb{P}_0}[u(X)u(X')\tilde{K}_n^2(X,X')] \le C_5\left(\sum_k \frac{\lambda_k}{\lambda_k+\varrho_n^2}\right)\|u\|_{L_2(\mathbb{P}_0)}^2 \asymp \varrho_n^{-1/s}\|u\|_{L_2(\mathbb{P}_0)}^2.$$

Recall that

$$\mathbb{E}_{\mathbb{P}_0}\tilde{K}_n^2(X,X') = \sum_k \left(\frac{\lambda_k}{\lambda_k+\varrho_n^2}\right)^2 \asymp \varrho_n^{-1/s}.$$

In the light of (6.10), they imply that

$$\mathbb{E}_{\mathbb{P}}\tilde{K}_n^2(X,X') \le 2\{\mathbb{E}_{\mathbb{P}_0}\tilde{K}_n^2(X,X') + \mathbb{E}_{\mathbb{P}_0}[u(X)u(X')\tilde{K}_n^2(X,X')]\} \le C_6\varrho_n^{-1/s}[1+\|u\|_{L_2(\mathbb{P}_0)}^2].$$

On the other hand, as already shown in the part of verifying (6.7), $\varrho_n \ll \Delta_n^{\theta+1}$ suffices to ensure that for sufficiently large n,

$$\frac{1}{4}\|u\|_{L_2(\mathbb{P}_0)}^2 \le \eta_{\varrho_n}^2(\mathbb{P},\mathbb{P}_0) \le \|u\|_{L_2(\mathbb{P}_0)}^2, \quad \forall\, \mathbb{P}\in\mathcal{P}(\Delta_n,\theta).$$

Thus

$$\begin{aligned}
&\lim_{n\to\infty}\sup_{\mathbb{P}\in\mathcal{P}(\Delta_n,\theta)}\mathbb{E}_{\mathbb{P}}\tilde{K}_n^2(X,X')/V_3^2\\
\le& 16C_6\left\{\left(\lim_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}(\Delta_n,\theta)}\varrho_n^{1/s}n^2\|u\|_{L_2(\mathbb{P}_0)}^4\right)^{-1} + \left(\lim_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}(\Delta_n,\theta)}\varrho_n^{1/s}n^2\|u\|_{L_2(\mathbb{P}_0)}^2\right)^{-1}\right\} = 0
\end{aligned}$$

provided that $\lim_{n\to\infty} n^{\frac{2s}{4s+\theta+1}}\Delta_n = \infty$. This immediately implies (6.8).

Verifying (6.9)**.** Observe that

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}V_2^2 \le& 4n\mathbb{E}_{\mathbb{P}}\left\{\sum_{k\ge1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_{\mathbb{P}}\phi_k(X)][\phi_k(X)-\mathbb{E}_{\mathbb{P}}\phi_k(X)]\right\}^2\\
\le& 4n\mathbb{E}_{\mathbb{P}}\left\{\sum_{k\ge1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_{\mathbb{P}}\phi_k(X)][\phi_k(X)]\right\}^2\\
=& 4n\mathbb{E}_{\mathbb{P}_0}\left([1+u(X)]\left\{\sum_{k\ge1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_{\mathbb{P}}\phi_k(X)][\phi_k(X)]\right\}^2\right).
\end{aligned}$$

It is clear that

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_0}\Big\{\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_{\mathbb{P}}\phi_k(X)][\phi_k(X)]\Big\}^2\\
=&\sum_{k,k'\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}\frac{\lambda_{k'}}{\lambda_{k'}+\varrho_n^2}\mathbb{E}_{\mathbb{P}}\phi_k(X)\mathbb{E}_{\mathbb{P}}\phi_{k'}(X)\mathbb{E}_{\mathbb{P}_0}[\phi_k(X)\phi_{k'}(X)]\\
=&\sum_{k\geq 1}\Big(\frac{\lambda_k}{\lambda_k+\varrho_n^2}\Big)^2[\mathbb{E}_{\mathbb{P}}\phi_k(X)]^2\leq \eta_{\varrho_n}^2(\mathbb{P},\mathbb{P}_0).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_0}\left(u(X)\Big\{\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_{\mathbb{P}}\phi_k(X)][\phi_k(X)]\Big\}^2\right)\\
\leq&\sqrt{\mathbb{E}_{\mathbb{P}_0}\left(u^2(X)\Big\{\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_{\mathbb{P}}\phi_k(X)][\phi_k(X)]\Big\}^2\right)}\times\\
&\times\sqrt{\mathbb{E}_{\mathbb{P}_0}\Big\{\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_{\mathbb{P}}\phi_k(X)][\phi_k(X)]\Big\}^2}\\
\leq&\|u\|_{L_2(\mathbb{P}_0)}\sup_x\Big|\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_{\mathbb{P}}\phi_k(X)][\phi_k(x)]\Big|\cdot\eta_{\varrho_n}(\mathbb{P},\mathbb{P}_0)\\
\leq&\left(\sup_k\|\phi_k\|_\infty\right)\|u\|_{L_2(\mathbb{P}_0)}\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}|\mathbb{E}_{\mathbb{P}}\phi_k(X)|\cdot\eta_{\varrho_n}(\mathbb{P},\mathbb{P}_0)\\
\leq&\left(\sup_k\|\phi_k\|_\infty\right)\|u\|_{L_2(\mathbb{P}_0)}\sqrt{\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}}\sqrt{\sum_{k\geq 1}\frac{\lambda_k}{\lambda_k+\varrho_n^2}[\mathbb{E}_{\mathbb{P}}\phi_k(X)]^2}\cdot\eta_{\varrho_n}(\mathbb{P},\mathbb{P}_0)\\
\leq&C_7\|u\|_{L_2(\mathbb{P}_0)}\varrho_n^{-\frac{1}{2s}}\eta_{\varrho_n}^2(\mathbb{P},\mathbb{P}_0).
\end{aligned}$$

Together, they imply that

$$\begin{aligned}
&\lim_{n\to\infty}\sup_{\mathbb{P}\in\mathcal{P}(\Delta_n,\theta)}\mathbb{E}_{\mathbb{P}}V_1^2/V_3^2\\
\leq&4\max\{1,C_7\}\left\{\Big(\lim_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}(\Delta_n,\theta)}n\eta_{\varrho_n}^2(\mathbb{P},\mathbb{P}_0)\Big)^{-1}+\lim_{n\to\infty}\sup_{\mathbb{P}\in\mathcal{P}(\Delta_n,\theta)}\left(\frac{\|u\|_{L_2(\mathbb{P}_0)}}{\varrho_n^{\frac{1}{2s}}n\eta_{\varrho_n}^2(\mathbb{P},\mathbb{P}_0)}\right)\right\}=0,
\end{aligned}$$

under the assumption that $\lim_{n\to\infty}n^{\frac{2s}{4s+\theta+1}}\Delta_n=\infty$. □

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 $\Delta_n = cn^{-\frac{2s}{4s+\theta+1}}$ 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 $B_n = \lfloor C_8 \Delta_n^{-\frac{1}{s}} \rfloor$ for a sufficiently small constant $C_8 > 0$ and $a_n = \sqrt{\Delta_n^2 / B_n}$. For any $\xi_n := (\xi_{n1}, \xi_{n2}, \cdots, \xi_{nB_n})^\top \in \{\pm 1\}^{B_n}$, write

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

It is clear that

$$\|u_{n,\xi_n}\|^2_{L_2(\mathbb{P}_0)} = B_n a_n^2 = \Delta_n^2$$

and

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

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 $\mathbb{P}_{n,\xi_n} \in \mathcal{P}(\Delta_n, 0)$ such that $d\mathbb{P}_{n,\xi_n} / d\mathbb{P}_0 = 1 + u_{n,\xi_n}$. Following a standard argument for minimax lower bound, it suffices to show that

$$\limsup_{n\to\infty} \mathbb{E}_{\mathbb{P}_0} \left(\frac{1}{2^{B_n}} \sum_{\xi_n \in \{\pm 1\}^{B_n}} \left\{ \prod_{i=1}^n [1 + u_{n,\xi_n}(X_i)] \right\} \right)^2 < \infty. \tag{6.11}$$

Note that

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{P}_0}\left(\frac{1}{2^{B_n}}\sum_{\xi_n\in\{\pm1\}^{B_n}}\left\{\prod_{i=1}^{n}[1+u_{n,\xi_n}(X_i)]\right\}\right)^2\\
=&\mathbb{E}_{\mathbb{P}_0}\left(\frac{1}{2^{2B_n}}\sum_{\xi_n,\xi_n'\in\{\pm1\}^{B_n}}\left\{\prod_{i=1}^{n}[1+u_{n,\xi_n}(X_i)]\right\}\left\{\prod_{i=1}^{n}[1+u_{n,\xi_n'}(X_i)]\right\}\right)\\
=&\frac{1}{2^{2B_n}}\sum_{\xi_n,\xi_n'\in\{\pm1\}^{B_n}}\prod_{i=1}^{n}\mathbb{E}_{\mathbb{P}_0}\left\{[1+u_{n,\xi_n}(X_i)][1+u_{n,\xi_n'}(X_i)]\right\}\\
=&\frac{1}{2^{2B_n}}\sum_{\xi_n,\xi_n'\in\{\pm1\}^{B_n}}\left(1+a_n^2\sum_{k=1}^{B_n}\xi_{nk}\xi_{nk}'\right)^n\\
\le&\frac{1}{2^{2B_n}}\sum_{\xi_n,\xi_n'\in\{\pm1\}^{B_n}}\exp\left(na_n^2\sum_{k=1}^{B_n}\xi_{n,k}\xi_{n,k}'\right)\\
=&\left\{\frac{\exp(na_n^2)+\exp(-na_n^2)}{2}\right\}^{B_n}\\
\le&\exp\left(\frac{1}{2}B_nn^2a_n^4\right),
\end{aligned}
$$

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 (6.11).

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 4. *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 $B_n = \lfloor C_{10}\Delta_n^{-\frac{\theta+1}{s}} \rfloor$ and $a_n = \sqrt{\Delta_n^2/B_n}$. By Lemma 4, we can find $\mathbb{P}_{n,\xi_n} \in \mathcal{P}(\Delta_n, \theta)$ such that $d\mathbb{P}_{n,\xi_n}/d\mathbb{P}_0 = 1 + u_{n,\xi_n}$, for appropriately chosen C_{10}. Following the same argument as in the previous case, we can again verify (6.11). □

Proof of Theorem 5. Without loss of generality, assume that $\Delta_n(\theta) = c_1(n^{-1}\sqrt{\log\log n})^{\frac{2s}{4s+\theta+1}}$ 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

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

and

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

As shown in the proof of Theorem 2,

$$\frac{\mathbb{E}G_n^2(X,X')}{v_n^2} \leq C_3\varrho_n^{1/s}, \quad \frac{\mathbb{E}\tilde{K}_n^4(X,X')}{n^2v_n^2} \leq C_4 n^{-2}\varrho_n^{-1/s}, \quad \text{and} \quad \frac{\mathbb{E}\tilde{K}_n^2(X,X')\tilde{K}_n^2(X,X'')}{nv_n^2} \leq C_3 n^{-1}.$$

Therefore,

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

which implies that

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

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

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

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 $\theta \ge 0$,

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

By Markov inequality, this can accomplished by verifying

$$\inf_{\theta\in[0,\infty)}\inf_{\mathbb{P}\in\mathcal{P}(\Delta_n(\theta),\theta)} \mathbb{E}_{\mathbb{P}}T_{n,\tilde{\varrho}_n(\theta)} \ge \tilde{M}\sqrt{\log\log n} \tag{6.12}$$

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

$$\lim_{n\to\infty}\sup_{\theta\ge 0}\sup_{\mathbb{P}\in\mathcal{P}(\Delta_n(\theta),\theta)} \frac{\mathrm{Var}\left(T_{n,\tilde{\varrho}_n(\theta)}\right)}{\left(\mathbb{E}_{\mathbb{P}}T_{n,\tilde{\varrho}_n(\theta)}\right)^2} = 0. \tag{6.13}$$

We now show that both (6.12) and (6.13) hold with

$$\Delta_n(\theta) = c_1 \left(\frac{\sqrt{\log \log n}}{n} \right)^{\frac{2s}{4s+\theta+1}}$$

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

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

$$\frac{1}{2} \varrho_n(\theta) \leq \tilde{\varrho}_n(\theta) \leq \varrho_n(\theta), \tag{6.14}$$

which immediately suggests

$$\eta^2_{\tilde{\varrho}_n(\theta)}(\mathbb{P}, \mathbb{P}_0) \geq \eta^2_{\varrho_n(\theta)}(\mathbb{P}, \mathbb{P}_0). \tag{6.15}$$

Following the arguments in the proof of Theorem 3,

$$\mathbb{E}_{\mathbb{P}} T_{n, \tilde{\varrho}_n(\theta)} \geq C_{17} n [\tilde{\varrho}_n(\theta)]^{1/(2s)} \eta^2_{\tilde{\varrho}_n(\theta)}(\mathbb{P}, \mathbb{P}_0) \geq 2^{-1/(2s)} C_{17} n [\varrho_n(\theta)]^{1/2s} \eta^2_{\varrho_n(\theta)}(\mathbb{P}, \mathbb{P}_0),$$

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

$$\eta^2_{\varrho_n(\theta)}(\mathbb{P}, \mathbb{P}_0) \geq \frac{1}{4} \|u\|^2_{L_2(P_0)} \tag{6.16}$$

provided that $\Delta_n(\theta) \geq C'(M) \left(\frac{\sqrt{\log \log n}}{n} \right)^{\frac{2s}{4s+\theta+1}}$.

Therefore,

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

if $c_1 \geq C_{18}^{-1} \tilde{M}$. Hence to ensure (6.12) holds, it suffices to take

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

With (6.14), (6.15) and (6.16), the results in the proof of Theorem 3 imply that for sufficiently large n

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

which shows (6.13). □

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

Nevertheless, in order to show

$$\inf_{\Phi_n}\left[\mathbb{E}_{\mathbb{P}_0}\Phi_n + \sup_{\theta\in[\theta_1,\theta_2]}\beta(\Phi_n;\Delta_n(\theta),\theta)\right]$$

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

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

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 $\mathbb{P}_\pi$ and $\mathbb{P}_0^{\otimes n}$ converges to 0. See Ingster (2000).

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

$$\Delta_n(\theta) = c_2\left(\frac{n}{\sqrt{\log\log n}}\right)^{-\frac{2s}{4s+\theta+1}}, \qquad \forall\theta\in[\theta_1,\theta_2],$$

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

Let $r_n = \lfloor C_{20}\log n\rfloor$ and $B_{n,1} = \lfloor C_{21}\Delta_n^{-\frac{\theta_1+1}{s}}(\theta_1)\rfloor$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.

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

Note that by choosing C_{20} sufficiently small,

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

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 $\xi_{n,r} = (\xi_{n,r,1}, \cdots, \xi_{n,r,B_{n,r}}) \in \{\pm 1\}^{B_{n,r}}$, 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 $a_{n,r} = \sqrt{\Delta_n^2(\theta_{n,r})/B_{n,r}}$. Following the same argument as that in the proof of Theorem 4, we can verify that with a sufficiently small C_{21}, each $\mathbb{P}_{n,r,\xi_{n,r}} \in \mathcal{P}(\Delta_n(\theta_{n,r}), \theta_{n,r})$, where $f_{n,r,\xi_{n,r}}$ is the Radon-Nikodym derivative $d\mathbb{P}_{n,r,\xi_{n,r}}/d\mathbb{P}_0$. 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

$$\|f_n - 1\|^2_{L_2(\mathbb{P}_0)} = \|f_n\|^2_{L_2(\mathbb{P}_0)} - 1 \to 0, \qquad \text{as } n \to \infty,$$

where $\|f_n\|^2_{L_2(\mathbb{P}_0)} = \mathbb{E}_{\mathbb{P}_0} f_n^2(X_1, X_2, \cdots, X_n)$.

Note that

$$\begin{aligned}\|f_n\|^2_{L_2(\mathbb{P}_0)} =& \frac{1}{r_n^2} \sum_{1 \le r, r' \le r_n} \langle f_{n,r}, f_{n,r'} \rangle_{L_2(\mathbb{P}_0)} \\ =& \frac{1}{r_n^2} \sum_{1 \le r \le r_n} \|f_{n,r}\|^2_{L_2(\mathbb{P}_0)} + \frac{1}{r_n^2} \sum_{\substack{1 \le r, r' \le r_n \\ r \ne r'}} \langle f_{n,r}, f_{n,r'} \rangle_{L_2(\mathbb{P}_0)}.\end{aligned}$$

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

$$\langle f_{n,r}, f_{n,r'} \rangle_{L_2(\mathbb{P}_0)} = 1.$$

It therefore suffices to show that

$$\sum_{1 \le r \le r_n} \|f_{n,r}\|^2_{L_2(\mathbb{P}_0)} = o(r_n^2).$$

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

$$\|f_{n,r}\|^2_{L_2(\mathbb{P}_0)} \le \left(\frac{\exp(n a_{n,r}^2) + \exp(-n a_{n,r}^2)}{2} \right)^{B_{n,r}} \le \exp\left(\frac{1}{2} B_{n,r} n^2 a_{n,r}^4 \right)$$

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

$$\sum_{1 \le r \le r_n} \|f_{n,r}\|_{L_2(\mathbb{P}_0)}^2 \le r_n^{3/2} = o(r_n^2).$$

□

Proof of Theorem 7. We begin with (3.2). Note that $\widehat{\gamma_{\nu_n}^2}(\mathbb{P}, \mathbb{P}_0)$ is a U-statistic. We can apply the general techniques for U-statistics to establish its asymptotic normality. In particular, as shown in Hall (1984), it suffices to verify the following four conditions:

$$\left(\frac{2\nu_n}{\pi}\right)^{d/2} \mathbb{E}\bar{G}_{\nu_n}^2(X_1, X_2) \to \|p_0\|_{L_2}^2, \tag{6.17}$$

$$\frac{\mathbb{E}\bar{G}_{\nu_n}^4(X_1, X_2)}{n^2[\mathbb{E}\bar{G}_{\nu_n}^2(X_1, X_2)]^2} \to 0, \tag{6.18}$$

$$\frac{\mathbb{E}[\bar{G}_{\nu_n}^2(X_1, X_2)\bar{G}_{\nu_n}^2(X_1, X_3)]}{n[\mathbb{E}\bar{G}_{\nu_n}^2(X_1, X_2)]^2} \to 0, \tag{6.19}$$

$$\frac{\mathbb{E}H_{\nu_n}^2(X_1, X_2)}{[\mathbb{E}\bar{G}_{\nu_n}^2(X_1, X_2)]^2} \to 0, \tag{6.20}$$

as $n \to \infty$, where

$$H_{\nu_n}(x, y) = \mathbb{E}\bar{G}_{\nu_n}(x, X_3)\bar{G}_{\nu_n}(y, X_3), \quad \forall\, x, y \in \mathbb{R}^d.$$

Verifying Condition (6.17). Note that

$$\mathbb{E}\bar{G}_{\nu_n}^2(X_1, X_2) = \mathbb{E}G_{\nu_n}^2(X_1, X_2) - 2\mathbb{E}\{\mathbb{E}[G_{\nu_n}(X_1, X_2)|X_1]\}^2 + [\mathbb{E}G_{\nu_n}(X_1, X_2)]^2.$$

By Lemma 7,

$$\mathbb{E}G_{\nu_n}(X_1, X_2) = \left(\frac{\pi}{\nu_n}\right)^{\frac{d}{2}} \int \exp\left(-\frac{\|\omega\|^2}{4\nu_n}\right) \|\mathcal{F}p_0(\omega)\|^2 \, d\omega,$$

which immediately yields

$$\left(\frac{\nu_n}{\pi}\right)^{\frac{d}{2}} \mathbb{E}G_{\nu_n}(X_1, X_2) \to \|p_0\|_{L_2}^2$$

and

$$\left(\frac{2\nu_n}{\pi}\right)^{\frac{d}{2}} \mathbb{E}G_{\nu_n}^2(X_1, X_2) = \left(\frac{2\nu_n}{\pi}\right)^{\frac{d}{2}} \mathbb{E}G_{2\nu_n}(X_1, X_2) \to \|p_0\|_{L_2}^2,$$

as $\nu_n \to \infty$.

On the other hand,

$$\begin{aligned}
&\mathbb{E}\{\mathbb{E}[G_{\nu_n}(X_1, X_2)|X_1]\}^2 \\
&= \int \left(\int G_{\nu_n}(x, x') G_{\nu_n}(x, x'') p_0(x) dx \right) p_0(x') p_0(x'') dx' dx'' \\
&= \int \left(\int G_{2\nu_n}(x, (x' + x'')/2) p_0(x) dx \right) G_{\nu_n/2}(x', x'') p_0(x') p_0(x'') dx' dx''.
\end{aligned}$$

Let $Z \sim N(0, 4\nu_n I_d)$. Then

$$\begin{aligned}
\int G_{2\nu_n}(x, (x' + x'')/2) p_0(x) dx &= (2\pi)^{d/2} \mathbb{E}\left[\mathcal{F}p_0(Z) \exp\left(\frac{x' + x''}{2} iZ\right)\right] \\
&\le (2\pi)^{d/2} \sqrt{\mathbb{E}\|\mathcal{F}p_0(Z)\|^2} \\
&\lesssim_d \|p_0\|_{L_2} / \nu_n^{d/4}.
\end{aligned}$$

Thus

$$\mathbb{E}\{\mathbb{E}[G_{\nu_n}(X_1, X_2)|X_1]\}^2 \lesssim_d \|p_0\|_{L_2}^3 / \nu_n^{3d/4}.$$

Condition (6.17) then follows.

Verifying Conditions (6.18) and (6.19). Since

$$\mathbb{E}\bar{G}^2_{\nu_n}(X_1, X_2) \asymp_{d,p_0} \nu_n^{-d/2}.$$

and

$$\mathbb{E}\bar{G}^4_{\nu_n}(X_1, X_2) \lesssim \mathbb{E}G^4_{\nu_n}(X_1, X_2) \lesssim_d \nu_n^{-d/2},$$

we obtain

$$n^{-2}\mathbb{E}\bar{G}^4_{\nu_n}(X_1, X_2)/(\mathbb{E}\bar{G}^2_{\nu_n}(X_1, X_2))^2 \lesssim_{d,p_0} \nu_n^{d/2}/n^2 \to 0.$$

Similarly,

$$\begin{aligned}\mathbb{E}\bar{G}^2_{\nu_n}(X_1, X_2)\bar{G}^2_{\nu_n}(X_1, X_3) &\lesssim \mathbb{E}G^2_{\nu_n}(X_1, X_2)G^2_{\nu_n}(X_1, X_3) \\ &= \mathbb{E}G_{2\nu_n}(X_1, X_2)G_{2\nu_n}(X_1, X_3) \\ &\lesssim_{d,p_0} \nu_n^{-3d/4}.\end{aligned}$$

This implies

$$n^{-1}\mathbb{E}\bar{G}^2_{\nu_n}(X_1, X_2)\bar{G}^2_{\nu_n}(X_1, X_3)/(\mathbb{E}\bar{G}^2_{\nu_n}(X_1, X_2))^2 \lesssim_{d,p_0} \nu_n^{d/4}/n \to 0,$$

which verifies (6.19).

Verifying Condition (6.20). We now prove (6.20). It suffices to show

$$\nu_n^d \mathbb{E}(\mathbb{E}(\bar{G}_{\nu_n}(X_1, X_2)\bar{G}_{\nu_n}(X_1, X_3)|X_2, X_3))^2 \to 0$$

as $n \to \infty$. Note that

$$\begin{aligned}
&\mathbb{E}(\mathbb{E}(\bar{G}_{\nu_n}(X_1, X_2)\bar{G}_{\nu_n}(X_1, X_3)|X_2, X_3))^2 \\
\lesssim&\mathbb{E}(\mathbb{E}(G_{\nu_n}(X_1, X_2)G_{\nu_n}(X_1, X_3)|X_2, X_3))^2 \\
=&\mathbb{E}G_{\nu_n}(X_1, X_2)G_{\nu_n}(X_1, X_3)G_{\nu_n}(X_4, X_2)G_{\nu_n}(X_4, X_3) \\
=&\mathbb{E}(G_{\nu_n}(X_1, X_4)G_{\nu_n}(X_2, X_3)\mathbb{E}(G_{\nu_n}(X_1 + X_4, X_2 + X_3)|X_1 - X_4, X_2 - X_3)).
\end{aligned}$$

Since for any $\delta > 0$,

$$\begin{aligned}
\nu_n^d\mathbb{E}(G_{\nu_n}(X_1, X_4)G_{\nu_n}(X_2, X_3)\mathbb{E}(G_{\nu_n}(X_1 + X_4, X_2 + X_3)|X_1 - X_4, X_2 - X_3) \\
(\mathbb{1}_{\{\|X_1 - X_4\| > \delta\}} + \mathbb{1}_{\|X_2 - X_3\| > \delta\}})) \to 0,
\end{aligned}$$

it remains to show that

$$\begin{aligned}
\nu_n^d\mathbb{E}(G_{\nu_n}(X_1, X_4)G_{\nu_n}(X_2, X_3)\mathbb{E}(G_{\nu_n}(X_1 + X_4, X_2 + X_3)|X_1 - X_4, X_2 - X_3) \\
\mathbb{1}_{\{\|X_1 - X_4\| \le \delta, \|X_2 - X_3\| \le \delta\}})) \to 0
\end{aligned}$$

for some $\delta > 0$, which holds as long as

$$\mathbb{E}(G_{\nu_n}(X_1 + X_4, X_2 + X_3)|X_1 - X_4, X_2 - X_3) \to 0 \tag{6.21}$$

uniformly on $\{\|X_1 - X_4\| \le \delta, \|X_2 - X_3\| \le \delta\}$.

Let

$$Y_1 = X_1 - X_4, \quad Y_2 = X_2 - X_3, \quad Y_3 = X_1 + X_4, \quad Y_4 = X_2 + X_3.$$

Then

$$\begin{aligned}
&\mathbb{E}(G_{\nu_n}(X_1+X_4, X_2+X_3)|X_1-X_4, X_2-X_3)\\
=&\left(\frac{\pi}{\nu_n}\right)^{\frac{d}{2}}\int \exp\left(-\frac{\|\omega\|^2}{4\nu_n}\right)\mathcal{F}p_{Y_1}(\omega)\overline{\mathcal{F}p_{Y_2}}(\omega)d\omega\\
\leq&\sqrt{\left(\frac{\pi}{\nu_n}\right)^{\frac{d}{2}}\int \exp\left(-\frac{\|\omega\|^2}{4\nu_n}\right)\left\|\mathcal{F}p_{Y_1}(\omega)\right\|^2 d\omega}\sqrt{\left(\frac{\pi}{\nu_n}\right)^{\frac{d}{2}}\int \exp\left(-\frac{\|\omega\|^2}{4\nu_n}\right)\left\|\mathcal{F}p_{Y_2}(\omega)\right\|^2 d\omega}
\end{aligned}$$

where

$$p_y(y') = \frac{p(Y_1=y, Y_3=y')}{p(Y_1=y)} = \frac{p_0\left(\frac{y+y'}{2}\right)p_0\left(\frac{y'-y}{2}\right)}{\int p_0\left(\frac{y+y'}{2}\right)p_0\left(\frac{y'-y}{2}\right)dy'}$$

is the conditional density of Y_3 given $Y_1 = y$. Thus to prove (6.21), it suffices to show

$$\begin{aligned}
h_n(y) :=& \left(\frac{\pi}{\nu_n}\right)^{\frac{d}{2}}\int \exp\left(-\frac{\|\omega\|^2}{4\nu_n}\right)\left\|\mathcal{F}p_y(\omega)\right\|^2 d\omega\\
=&\ \pi^{\frac{d}{2}}\int \exp\left(-\frac{\|\omega\|^2}{4}\right)\left\|\mathcal{F}p_y(\sqrt{\nu_n}\omega)\right\|^2 d\omega\\
\to&\ 0
\end{aligned}$$

uniformly over $\{y:\ \|y\| \leq \delta\}$.

Note that

$$h_n(y) = \mathbb{E}G_{\nu_n}(X, X')$$

where $X, X' \sim_{\text{iid}} p_y$, which suggests $h_n(y) \to 0$ pointwisely. To prove the uniform convergence of $h_n(y)$, we only need to show

$$\lim_{y_1 \to y}\sup_n |h_n(y_1) - h_n(y)| = 0$$

for any y.

Since $p_0 \in L_2$, $P(Y_1 = y)$ is continuous. Therefore, the almost surely continuity of p_0 immediately suggests that for every y, $p_{y_1}(\cdot) \to p_y(\cdot)$ almost surely as $y_1 \to y$. Considering that p_{y_1}

and p_y are both densities, it follows that

$$|\mathcal{F}p_{y_1}(\omega) - \mathcal{F}p_y(\omega)| \leq (2\pi)^{-d/2} \int |p_{y_1}(y') - p_y(y')|dy' \to 0,$$

i.e., $\mathcal{F}p_{y_1} \to \mathcal{F}p_y$ uniformly as $y_1 \to y$. Therefore we have

$$\sup_{n\to\infty} |h_n(y_1) - h_n(y)| \lesssim \left\|\mathcal{F}p_{y_1} - \mathcal{F}p_y\right\|_{L_\infty} \to 0,$$

which ensures the uniform convergence of $h_n(y)$ to $h(y)$ over $\{y : \|y\| \leq \delta\}$, and hence (6.20).

Indeed, we have shown that

$$\frac{n\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{P}_0)}{\sqrt{2\mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2}} \to_d N(0, 1).$$

By Slutsky Theorem, in order to prove (3.3), it sufficies to show

$$\widehat{s}^2_{n,\nu_n}/\mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2 \to_p 1,$$

which is equivalent to

$$\tilde{s}^2_{n,\nu_n}/\mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2 \to_p 1 \tag{6.22}$$

since $1/n^2 = o(\mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2)$.

It follows from

$$\mathbb{E}\left(\tilde{s}^2_{n,\nu_n}\right) = \mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2$$

and

$$
\begin{aligned}
&\operatorname{var}\left(\tilde{s}^2_{n,\nu_n}\right)\\
\lesssim& n^{-4}\operatorname{var}\left(\sum_{1\le i\neq j\le n} G_{2\nu_n}(X_i,X_j)\right)+n^{-6}\operatorname{var}\left(\sum_{\substack{1\le i,j_1,j_2\le n\\ |\{i,j_1,j_2\}|=3}} G_{\nu_n}(X_i,X_{j_1})G_{\nu_n}(X_i,X_{j_2})\right)\\
&+n^{-8}\operatorname{var}\left(\sum_{\substack{1\le i_1,i_2,j_1,j_2\le n\\ |\{i_1,i_2,j_1,j_2\}|=4}} G_{\nu_n}(X_{i_1},X_{j_1})G_{\nu_n}(X_{i_2},X_{j_2})\right)\\
\lesssim& n^{-2}\mathbb{E}G_{4\nu_n}(X_1,X_2)+n^{-1}\mathbb{E}G_{2\nu_n}(X_1,X_2)G_{2\nu_n}(X_1,X_3)+n^{-1}(\mathbb{E}G_{2\nu_n}(X_1,X_2))^2\\
=& o((\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2)^2).
\end{aligned}
$$

that (6.22) holds. □

Proof of Theorem 8. Recall that

$$
\begin{aligned}
\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{P}_0)=&\frac{1}{n(n-1)}\sum_{i\neq j}\bar{G}_{\nu_n}(X_i,X_j;\mathbb{P}_0)\\
=&\gamma^2_{\nu_n}(\mathbb{P},\mathbb{P}_0)+\frac{1}{n(n-1)}\sum_{i\neq j}\bar{G}_{\nu_n}(X_i,X_j;\mathbb{P})\\
&+\frac{2}{n}\sum_{i=1}^n\Big(\mathbb{E}_{X\sim\mathbb{P}}[G_{\nu_n}(X_i,X)|X_i]-\mathbb{E}_{X\sim\mathbb{P}_0}[G_{\nu_n}(X_i,X)|X_i]\\
&-\mathbb{E}_{X,X'\sim_{\text{iid}}\mathbb{P}}G_{\nu_n}(X,X')+\mathbb{E}_{(X,Y)\sim\mathbb{P}\otimes\mathbb{P}_0}G_{\nu_n}(X,Y)\Big).
\end{aligned}
$$

Denote by the last two terms on the rightmost hand side by $V^{(1)}_{\nu_n}$ and $V^{(2)}_{\nu_n}$ respectively. It is clear that $\mathbb{E}V^{(1)}_{\nu_n}=\mathbb{E}V^{(2)}_{\nu_n}=0$. Then it suffices to show that

$$
\sup_{\substack{p\in\mathcal{W}^{s,2}(M)\\ \|p-p_0\|\ge\Delta_n}}\frac{\mathbb{E}\left(V^{(1)}_{\nu_n}\right)^2+\mathbb{E}\left(V^{(2)}_{\nu_n}\right)^2}{\gamma^4_{\nu_n}(\mathbb{P},\mathbb{P}_0)}\to 0 \tag{6.23}
$$

and

$$\inf_{\substack{p\in\mathcal{W}^{s,2}(M)\\ \|p-p_0\|\geq\Delta_n}} \frac{n\gamma^2_{G_{\nu_n}}(\mathbb{P},\mathbb{P}_0)}{\sqrt{\mathbb{E}\left(\widehat{s}^2_{n,\nu_n}\right)}} \to \infty \tag{6.24}$$

as $n \to \infty$.

We first prove (6.23). Note that $\|p\|_{L_2} \leq \|p\|_{\mathcal{W}^{s,2}(M)} \leq M$. Following arguments similar to those in the proof of Theorem 7, we get

$$\mathbb{E}\left(V^{(1)}_{\nu_n}\right)^2 \lesssim n^{-2}\mathbb{E}G^2_{\nu_n}(X_1,X_2) \lesssim_d M^2 n^{-2}\nu_n^{-d/2},$$

and

$$\begin{aligned}
\mathbb{E}\left(V^{(2)}_{\nu_n}\right)^2 &\leq \frac{4}{n}\mathbb{E}\left[\mathbb{E}_{X\sim\mathbb{P}}[G_{\nu_n}(X_i,X)|X_i] - \mathbb{E}_{X\sim\mathbb{P}_0}[G_{\nu_n}(X_i,X)|X_i]\right]^2\\
&= \frac{4}{n}\int\left(\int G_{2\nu_n}(x,(x'+x'')/2)p(x)dx\right)G_{\nu_n/2}(x',x'')f(x')f(x'')dx'dx''\\
&\lesssim_d \frac{4M}{n\nu^{d/4}}\int G_{\nu_n/2}(x',x'')|f(x')||f(x'')|dx'dx''\\
&\lesssim_d \frac{4M}{n\nu^{3d/4}}\|f\|^2_{L_2}.
\end{aligned}$$

By Lemma 8, there exists a constant $C > 0$ depending on s and M only such that for $f \in \mathcal{W}^{s,2}(M)$,

$$\int \exp\left(-\frac{\|\omega\|^2}{4\nu_n}\right)\|\mathcal{F}f(\omega)\|^2\,d\omega \geq \frac{1}{4}\|f\|^2_{L_2}$$

given that $\nu_n \geq C\|f\|^{-2/s}_{L_2}$. Because $\nu_n\Delta_n^{s/2} \to \infty$, we obtain

$$\gamma^2_{\nu_n}(\mathbb{P},\mathbb{P}_0) \gtrsim_d \nu_n^{-d/2}\|f\|^2_{L_2},$$

for sufficiently large n. Thus

$$\sup_{\substack{p\in\mathcal{W}^{s,2}(M)\\ \|p-p_0\|\geq\Delta_n}} \frac{\mathbb{E}\left(V_{\nu_n}^{(1)}\right)^2}{\gamma_{\nu_n}^4(\mathbb{P},\mathbb{P}_0)} \lesssim_d M^2(n^2\nu_n^{-d/2}\Delta_n^4)^{-1} \to 0$$

and

$$\sup_{\substack{p\in\mathcal{W}^{s,2}(M)\\ \|p-p_0\|\geq\Delta_n}} \frac{\mathbb{E}\left(V_{\nu_n}^{(2)}\right)^2}{\gamma_{G_{\nu_n}}^4(\mathbb{P},\mathbb{P}_0)} \lesssim_d M(n\nu_n^{-d/4}\Delta_n^2)^{-1} \to 0,$$

as $n \to \infty$.

Next we prove (6.24). It follows from

$$\mathbb{E}\left(\widehat{s}_{n,\nu_n}^2\right) \leq \mathbb{E}\max\left\{\left|\tilde{s}_{n,\nu_n}^2\right|, 1/n^2\right\} \lesssim \mathbb{E}G_{2\nu_n}(X_1,X_2) + 1/n^2 \lesssim_d M^2\nu_n^{-d/2} + 1/n^2$$

that (6.24) holds. □

Proof of Theorem 9. This, in a certain sense, can be viewed as an extension of results from Ingster (1987), and the proof proceeds in a similar fashion. While Ingster (1987) considered the case when p_0 is the uniform distribution on $[0,1]$, we shall show that similar bounds hold for a wider class of p_0.

For any $M > 0$ and p_0 such that $\|p_0\|_{\mathcal{W}^{s,2}} < M$, let

$$H_1^{\mathrm{GOF}}(\Delta_n; s, M - \|p_0\|_{\mathcal{W}^{s,2}})^*$$
$$:= \{p \in \mathcal{W}^{s,2} : \|p - p_0\|_{\mathcal{W}^{s,2}} \leq M - \|p_0\|_{\mathcal{W}^{s,2}},\ \|p - p_0\|_{L_2} \geq \Delta_n\}.$$

It is clear that $H_1^{\mathrm{GOF}}(\Delta_n; s) \supset H_1^{\mathrm{GOF}}(\Delta_n; s, M - \|p_0\|_{\mathcal{W}^{s,2}})^*$. Hence it suffices to prove Theorem 9 with $H_1^{\mathrm{GOF}}(\Delta_n; s)$ replaced by $H_1^{\mathrm{GOF}}(\Delta_n; s, M)^*$ for an arbitrary $M > 0$. We shall abbreviate $H_1^{\mathrm{GOF}}(\Delta_n; s, M)^*$ as $H_1^{\mathrm{GOF}}(\Delta_n; s)^*$ in the rest of the proof.

Since p_0 is almost surely continuous, there exists $x_0 \in \mathbb{R}^d$ and $\delta, c > 0$ such that

$$p_0(x) \geq c > 0, \quad \forall \, \|x - x_0\| \leq \delta.$$

In light of this, we shall assume $p_0(x) \geq c > 0$, for all $x \in [0,1]^d$ without loss of generality.

Let $\boldsymbol{a}_n$ be a multivariate random index. As proved in Ingster (1987), in order to prove the existence of $\alpha \in (0,1)$ such that no asymptotic α-level test can be consistent, it suffices to identify $p_{n,\boldsymbol{a}_n} \in H_1^{\mathrm{GOF}}(\Delta_n; s)^*$ for all possible values of $\boldsymbol{a}_n$ such that

$$\mathbb{E}_{p_0} \left(\frac{p_n(X_1, \cdots, X_n)}{\prod_{i=1}^n p_0(X_i)} \right)^2 = O(1), \tag{6.25}$$

where

$$p_n(x_1, \cdots, x_n) = \mathbb{E}_{\boldsymbol{a}_n} \left(\prod_{i=1}^n p_{n,\boldsymbol{a}_n}(x_i) \right), \ \forall \, x_1, \cdots, x_n,$$

i.e., p is the mixture of all $p_{n,\boldsymbol{a}_n}$'s.

Let $\mathbb{1}_{\{x \in [0,1]^d\}}, \phi_{n,1}, \cdots, \phi_{n,B_n}$ be an orthonormal sets of functions in $L_2(\mathbb{R}^d)$ such that the supports of $\phi_{n,1}, \cdots, \phi_{n,B_n}$ are disjoint and all included in $[0,1]^d$. Let $\boldsymbol{a}_n = (a_{n,1}, \cdots, a_{n,B_n})$ satisfy that $a_{n,1}, \cdots, a_{n,B_n}$ are independent and that

$$p(a_{n,k} = 1) = p(a_{n,k} = -1) = \frac{1}{2}, \quad \forall \, 1 \leq k \leq B_n.$$

Define

$$p_{n,\boldsymbol{a}_n} = p_0 + r_n \sum_{k=1}^{B_n} a_{n,k} \phi_{n,k}.$$

Then

$$\frac{p_{n,\boldsymbol{a}_n}}{p_0} = 1 + r_n \sum_{k=1}^{B_n} a_{n,k} \frac{\phi_{n,k}}{p_0},$$

where $1, \frac{\phi_{n,1}}{p_0}, \cdots, \frac{\phi_{n,B_n}}{p_0}$ are orthogonal in $L_2(P_0)$.

By arguments similar to those in Ingster (1987), we find

$$\mathbb{E}_{p_0}\left(\frac{p_n(X_1,\cdots,X_n)}{\prod_{i=1}^n p_0(X_i)}\right)^2 \le \exp\left(\frac{1}{2}B_n n^2 r_n^4 \max_{1\le k\le B_n}\left(\int \phi_{n,k}^2/p_0 dx\right)^2\right)$$
$$\le \exp\left(\frac{1}{2c^2}B_n n^2 r_n^4\right).$$

In order to ensure (6.25), it suffices to have

$$B_n^{1/2} n r_n^2 = O(1). \tag{6.26}$$

Therefore, given $\Delta_n = O\left(n^{-\frac{2s}{4s+d}}\right)$, once we can find proper r_n, B_n and $\phi_{n,1},\cdots,\phi_{n,B_n}$ such that $p_{n,\boldsymbol{a}_n} \in H_1^{\mathrm{GOF}}(\Delta_n; s)^*$ for all $\boldsymbol{a}_n$ and (6.26) holds, the proof is finished.

Let $b_n = B_n^{1/d}$, ϕ be an infinitely differentiable function supported on $[0,1]^d$ that is orthogonal to $\mathbb{1}_{\{x\in[0,1]^d\}}$ in L_2, and for each $x_{n,k} \in \{0,1,\cdots,b_n-1\}^{\otimes d}$, let

$$\phi_{n,k}(x) = \frac{b_n^{d/2}}{\|\phi\|_{L_2}}\phi(b_n x - x_{n,k}), \quad \forall\, x \in \mathbb{R}^d.$$

Then all $\phi_{n,k}$'s are supported on $[0,1]^d$ and

$$\langle \phi_{n,k}, 1\rangle_{L_2} = \frac{b_n^{d/2}}{\|\phi\|_{L_2}}\int_{\mathbb{R}^d}\phi(b_n x - x_{n,k})dx = \frac{1}{b_n^{d/2}\|\phi\|_{L_2}}\int_{\mathbb{R}^d}\phi(x)dx = 0,$$
$$\|\phi_{n,k}\|_{L_2}^2 = \frac{b_n^d}{\|\phi\|_{L_2}^2}\int_{[0,1/b_n]^d}\phi^2(b_n x)dx = 1,$$
$$\|\phi_{n,k}\|_{\mathcal{W}^{s,2}}^2 \le b_n^{2s}\frac{\|\phi\|_{\mathcal{W}^{s,2}}^2}{\|\phi\|_{L_2}^2}.$$

Since for $k \ne k'$, the supports of $\phi_{n,k}$ and $\phi_{n,k'}$ are disjoint,

$$\|p_{n,\boldsymbol{a}_n} - p_0\|_\infty = r_n b_n^{d/2}\frac{\|\phi\|_\infty}{\|\phi\|_{L_2}},$$

and

$$\langle \phi_{n,k}, \phi_{n,k'} \rangle_{L_2} = 0, \qquad \langle \phi_{n,k}, \phi_{n,k'} \rangle_{\mathcal{W}^{s,2}} = 0,$$

from which we immediately obtain

$$\begin{aligned} \|p_{n,\boldsymbol{a}_n} - p_0\|_{L_2}^2 &= r_n^2 b_n^d \\ \|p_{n,\boldsymbol{a}_n} - p_0\|_{\mathcal{W}^{s,2}}^2 &\leq r_n^2 b_n^{d+2s} \frac{\|\phi\|_{\mathcal{W}^{s,2}}^2}{\|\phi\|_{L_2}^2}. \end{aligned}$$

To ensure $p_{n,\boldsymbol{a}_n} \in H_1^{\mathrm{GOF}}(\Delta_n; s)^*$, it suffices to make

$$r_n b_n^{d/2} \frac{\|\phi\|_\infty}{\|\phi\|_{L_2}} \to 0 \text{ as } n \to \infty, \tag{6.27}$$

$$r_n^2 b_n^d = \Delta_n^2, \tag{6.28}$$

$$r_n^2 b_n^{d+2s} \frac{\|\phi\|_{\mathcal{W}^{s,2}}^2}{\|\phi\|_{L_2}^2} \leq M^2. \tag{6.29}$$

Let

$$b_n = \left\lfloor \left(\frac{M\|\phi\|_{L_2}^2}{\|\phi\|_{\mathcal{W}^{s,2}}} \right)^{1/s} \Delta_n^{-1/s} \right\rfloor, \quad r_n = \frac{\Delta_n}{b_n^{d/2}}.$$

Then (6.28) and (6.29) are satisfied. Moreover, given $\Delta_n = O\left(n^{-\frac{2s}{4s+d}}\right)$,

$$B_n^{1/2} n r_n^2 = b_n^{-d/2} n \Delta_n^2 \lesssim_{d,\phi,M} n\Delta_n^{\frac{4s+d}{2s}} = O(1),$$

and

$$r_n b_n^{d/2} \frac{\|\phi\|_\infty}{\|\phi\|_{L_2}} \lesssim_\phi \Delta_n = o(1)$$

ensuring both (6.26) and (6.27).

Finally, we show the existence of such ϕ. Let

$$\phi_0(x_1) = \begin{cases} \exp\left(-\frac{1}{1-(4x_1-1)^2}\right) & 0 < x_1 < \frac{1}{2} \\ -\exp\left(-\frac{1}{1-(4x_1-3)^2}\right) & \frac{1}{2} < x_1 < 1 \\ 0 & \text{otherwise} \end{cases}.$$

Then ϕ_0 is supported on $[0,1]$, infinitely differentiable and orthogonal to the indicator function of $[0,1]$.

Let

$$\phi(x) = \prod_{l=1}^{d} \phi_0(x_l), \quad \forall\, x = (x_1, \cdots, x_d) \in \mathbb{R}^d.$$

Then ϕ is supported on $[0,1]^d$, infinitely differentiable and $\langle \phi, 1\rangle_{L_2} = \langle \phi_0, 1\rangle^d_{L_2[0,1]} = 0$. □

Proof of Theorem 10. Let $N = m + n$ denote the total sample size. It suffices to prove the result under the assumption that $n/N \to r \in (0,1)$.

Note that under H_0,

$$\begin{aligned}\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{Q}) =& \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} \bar{G}_{\nu_n}(X_i, X_j) + \frac{1}{m(m-1)} \sum_{1\le i\ne j\le m} \bar{G}_{\nu_n}(Y_i, Y_j) \\ & - \frac{2}{nm} \sum_{1\le i\le n} \sum_{1\le j\le m} \bar{G}_{\nu_n}(X_i, Y_j).\end{aligned}$$

Let $n/N = r_n$. Then we have

$$\begin{aligned}&\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{Q}) \\ =& N^{-2}\left(\frac{1}{r_n(r_n - N^{-1})} \sum_{1\le i\ne j\le n} \bar{G}_{\nu_n}(X_i, X_j) + \right. \\ & \left. \frac{1}{(1-r_n)(1-r_n-N^{-1})} \sum_{1\le i\ne j\le m} \bar{G}_{\nu_n}(Y_i, Y_j) - \frac{2}{r_n(1-r_n)} \sum_{1\le i\le n} \sum_{1\le j\le m} \bar{G}_{\nu_n}(X_i, Y_j)\right).\end{aligned}$$

Let

$$\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{Q})' = N^{-2}\left(\frac{1}{r^2}\sum_{1\le i\ne j\le n}\bar{G}_{\nu_n}(X_i,X_j)+\frac{1}{(1-r)^2}\sum_{1\le i\ne j\le m}\bar{G}_{\nu_n}(Y_i,Y_j)\right.$$
$$\left.-\frac{2}{r(1-r)}\sum_{1\le i\le n}\sum_{1\le j\le m}\bar{G}_{\nu_n}(X_i,Y_j)\right).$$

As we assume $r_n \to r$ as $n \to \infty$, Theorem 7 ensures that

$$\frac{nm}{\sqrt{2}(n+m)}\left[\mathbb{E}\bar{G}^2_{\nu_n}(X_1,X_2)\right]^{-\frac{1}{2}}\left(\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{Q})-\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{Q})'\right)=o_p(1)$$

A slight adaption of arguments in Hall (1984) suggests that

$$\frac{\mathbb{E}\bar{G}^4_{\nu_n}(X_1,X_2)}{N^2\mathbb{E}\bar{G}^2_{\nu_n}(X_1,X_2)}+\frac{\mathbb{E}\bar{G}^2_{\nu_n}(X_1,X_2)\bar{G}^2_{\nu_n}(X_1,X_3)}{N\mathbb{E}\bar{G}^2_{\nu_n}(X_1,X_2)}+\frac{\mathbb{E}H^2_{\nu_n}(X_1,X_2)}{\mathbb{E}\bar{G}^2_{\nu_n}(X_1,X_2)}\to 0 \tag{6.30}$$

ensures that

$$\frac{nm}{\sqrt{2}(n+m)}\left[\mathbb{E}\bar{G}^2_{\nu_n}(X_1,X_2)\right]^{-\frac{1}{2}}\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{Q})'\to_d N(0,1).$$

Following arguments similar to those in the proof of Theorem 7, given $\nu_n \to \infty$ and $\nu_n/n^{4/d} \to 0$, (6.30) holds and therefore

$$\frac{nm}{\sqrt{2}(n+m)}\left[\mathbb{E}\bar{G}^2_{\nu_n}(X_1,X_2)\right]^{-\frac{1}{2}}\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{Q})\to_d N(0,1).$$

Additionally, based on the same arguments as in the proof of Theorem 7,

$$\widehat{s}^2_{n,m,\nu_n}/\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2\to_p 1.$$

The proof is therefore concluded. □

Proof of Theorem 11. With slight abuse of notation, we shall write

$$\bar{G}_{\nu_n}(x,y;\mathbb{P},\mathbb{Q}) = G_{\nu_n}(x,y) - \mathbb{E}_{Y\sim\mathbb{Q}}G_{\nu_n}(x,Y) - \mathbb{E}_{X\sim\mathbb{P}}G_{\nu_n}(X,y) + \mathbb{E}_{(X,Y)\sim\mathbb{P}\otimes\mathbb{Q}}G_{\nu_n}(X,Y),$$

We consider the two parts separately.

Part (i). We first verify the consistency of $\Phi^{\mathrm{HOM}}_{n,\nu_n,\alpha}$ with $\nu_n \asymp n^{4/(d+4s)}$ given $\Delta_n \gg n^{-2s/(d+4s)}$.

Observe the following decomposition of $\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{Q})$,

$$\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{Q}) = \gamma^2_{\nu_n}(\mathbb{P},\mathbb{Q}) + L^{(1)}_{n,\nu_n} + L^{(2)}_{n,\nu_n},$$

where

$$\begin{aligned} L^{(1)}_{n,\nu_n} =& \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} \bar{G}_{\nu_n}(X_i,X_j;\mathbb{P}) - \frac{2}{mn}\sum_{1\le i\le n}\sum_{1\le j\le m}\bar{G}_{\nu_n}(X_i,Y_j;\mathbb{P},\mathbb{Q}) \\ &+ \frac{1}{m(m-1)}\sum_{1\le i\ne j\le m}\bar{G}_{\nu_n}(Y_i,Y_j;\mathbb{Q}) \end{aligned}$$

and

$$\begin{aligned} L^{(2)}_{n,\nu_n} =& \frac{2}{n}\sum_{i=1}^{n}\left(\mathbb{E}[G_{\nu_n}(X_i,X)|X_i] - \mathbb{E}G_{\nu_n}(X,X') - \mathbb{E}[G_{\nu_n}(X_i,Y)|X_i] + \mathbb{E}G_{\nu_n}(X,Y)\right) \\ &+ \frac{2}{m}\sum_{j=1}^{m}\left(\mathbb{E}[G_{\nu_n}(Y_j,Y)|Y_j] - \mathbb{E}G_{\nu_n}(Y,Y') - \mathbb{E}[G_{\nu_n}(X,Y_j)|Y_j] + \mathbb{E}G_{\nu_n}(X,Y)\right). \end{aligned}$$

In order to prove the consistency of $\Phi^{\mathrm{HOM}}_{n,\nu_n,\alpha}$, it suffices to show

$$\sup_{\substack{p,q\in\mathcal{W}^{s,2}(M)\\ \|p-q\|_{L_2}\ge\Delta_n}} \frac{\mathbb{E}\left(L^{(1)}_{n,\nu_n}\right)^2 + \mathbb{E}\left(L^{(2)}_{n,\nu_n}\right)^2}{\gamma^4_{G_{\nu_n}}(\mathbb{P},\mathbb{Q})} \to 0, \tag{6.31}$$

$$\inf_{\substack{p,q\in\mathcal{W}^{s,2}(M)\\ \|p-q\|_{L_2}\ge\Delta_n}} \frac{\gamma^2_{G_{\nu_n}}(\mathbb{P},\mathbb{Q})}{(1/n+1/m)\sqrt{\mathbb{E}\left(\widehat{s}^2_{n,m,\nu_n}\right)}} \to \infty, \tag{6.32}$$

as $n \to \infty$. We now prove (6.31) and (6.32) with arguments similar to those obtained in the proof of Theorem 8.

Note that

$$\begin{aligned}
&\mathbb{E}(L^{(1)}_{n,\nu_n})^2 \\
\lesssim &\mathbb{E}\left(\frac{1}{n(n-1)}\sum_{1\le i\ne j\le n}\bar{G}_{\nu_n}(X_i,X_j;\mathbb{P})\right)^2 + \mathbb{E}\left(\frac{2}{mn}\sum_{1\le i\le n}\sum_{1\le j\le m}\bar{G}_{\nu_n}(X_i,Y_j;\mathbb{P},\mathbb{Q})\right)^2 \\
&+ \mathbb{E}\left(\frac{1}{m(m-1)}\sum_{1\le i\ne j\le m}\bar{G}_{\nu_n}(Y_i,Y_j;\mathbb{Q})\right)^2 \\
\lesssim &\frac{1}{n^2}\mathbb{E}G^2_{\nu_n}(X_1,X_2) + \frac{1}{m^2}\mathbb{E}G^2_{\nu_n}(Y_1,Y_2).
\end{aligned}$$

Given $p, q \in \mathcal{W}^{s,2}(M)$,

$$\mathbb{E}G^2_{\nu_n}(X_1,X_2) \lesssim_d M^2\nu_n^{-d/2}, \quad \mathbb{E}G^2_{\nu_n}(Y_1,Y_2) \lesssim_d M^2\nu_n^{-d/2}.$$

Hence

$$\mathbb{E}(L^{(1)}_{n,\nu_n})^2 \lesssim_d M^2\nu_n^{-d/2}\left(\frac{1}{n^2}+\frac{1}{m^2}\right). \tag{6.33}$$

Now consider bounding $L^{(2)}_{n,\nu_n}$. Let $f = p - q$. Then we have

$$\mathbb{E}(L^{(2)}_{n,\nu_n})^2 \lesssim_d \nu_n^{-\frac{3d}{4}} M\|f\|^2_{L_2}\left(\frac{1}{n}+\frac{1}{m}\right). \tag{6.34}$$

Since $\nu_n \asymp n^{4/(4s+d)} \gg \Delta_n^{-2/s}$, Lemma 8 ensures that for sufficiently large n,

$$\gamma^2_{G_{\nu_n}}(\mathbb{P},\mathbb{Q}) \gtrsim_d \nu_n^{-d/2}\|f\|^2_{L_2}, \quad \forall\, p, q \in \mathcal{W}^{s,2}(M).$$

This together with (6.33) and (6.34) gives

$$\sup_{\substack{p,q\in\mathcal{W}^{s,2}(M)\\ \|p-q\|_{L_2}\geq\Delta_n}} \frac{\mathbb{E}\left(L^{(1)}_{n,\nu_n}\right)^2+\mathbb{E}\left(L^{(2)}_{n,\nu_n}\right)^2}{\gamma^4_{G_{\nu_n}}(\mathbb{P},\mathbb{Q})} \lesssim_d \frac{M^2\nu_n^{d/2}}{n^2\Delta_n^4}+\frac{M\nu_n^{d/4}}{n\Delta_n^2}\to 0$$

as $n\to\infty$, which proves (6.31).

Finally, consider (6.32). It follows from

$$\begin{aligned}\mathbb{E}\left(\widetilde{s}^2_{n,m,\nu_n}\right) &\leq \mathbb{E}\max\left\{\left|\tilde{s}^2_{n,m,\nu_n}\right|, 1/n^2\right\}\\ &\lesssim \max\{\mathbb{E}G^2_{\nu_n}(X_1,X_2),\mathbb{E}G^2_{\nu_n}(Y_1,Y_2)\}+1/n^2\\ &\lesssim_d M^2\nu_n^{-d/2}+1/n^2\end{aligned}$$

that (6.32) holds.

Part (ii). Next, we prove that if $\liminf_{n\to\infty}\Delta_n n^{2s/(d+4s)}<\infty$, then there exists some $\alpha\in(0,1)$ such that no asymptotic α-level test can be consistent. To prove this, we shall verify that consistency of homogeneity test is harder to achieve than that of goodness-of-fit test.

Consider an arbitrary $p_0\in\mathcal{W}^{s,2}(M/2)$. It immediately follows

$$H_1^{\mathrm{HOM}}(\Delta_n;s)\supset\{(p,p_0):\ p\in H_1^{\mathrm{GOF}}(\Delta_n;s)\}.$$

Let $\{\Phi_n\}_{n\geq1}$ be any sequence of asymptotic α-level homogeneity tests, where

$$\Phi_n=\Phi_n(X_1,\cdots,X_n,Y_1,\cdots,Y_m).$$

Then if $Y_1,\cdots,Y_m\sim_{\mathrm{iid}}P_0$, $\{\Phi_n\}_{n\geq1}$ can also be treated as a sequence of (random) goodness-of-fit tests

$$\Phi_n(X_1,\cdots,X_n,Y_1,\cdots,Y_m)=\tilde{\Phi}_n(X_1,\cdots,X_n)$$

whose probabilities of type I error with respect to P_0 are controlled at α asymptotically. Moreover,

$$\text{power}\{\Phi_n; H_1^{\text{HOM}}(\Delta_n; s)\} \leq \text{power}\{\tilde{\Phi}_n; H_1^{\text{GOF}}(\Delta_n; s)\}$$

Since $0 < c \leq m/n \leq C < \infty$, Theorem 9 ensures that there exists some $\alpha \in (0, 1)$ such that for any sequence of asymptotic α-level tests $\{\Phi_n\}_{n\geq 1}$,

$$\liminf_{n\to\infty} \text{power}\{\Phi_n; H_1^{\text{HOM}}(\Delta_n; s)\} \leq \liminf_{n\to\infty} \text{power}\{\tilde{\Phi}_n; H_1^{\text{GOF}}(\Delta_n; s)\} < 1$$

given $\liminf_{n\to\infty} \Delta_n n^{2s/(d+4s)} < \infty$. □

Proof of Theorem 12. For brevity, we shall focus on the case when $k = 2$ in the rest of the proof. Our argument, however, can be straightforwardly extended to the more general cases. The proof relies on the following decomposition of $\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{P}^{X^1} \otimes \mathbb{P}^{X^2})$ under H_0^{IND}:

$$\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{P}^{X^1} \otimes \mathbb{P}^{X^2}) = \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} G^*_{\nu_n}(X_i, X_j) + R_n,$$

where

$$G^*_{\nu_n}(x, y) = \bar{G}_{\nu_n}(x, y) - \sum_{1\leq j\leq 2} g_j(x^j, y) - \sum_{1\leq j\leq 2} g_j(y^j, x) + \sum_{1\leq j_1, j_2\leq 2} g_{j_1,j_2}(x^{j_1}, y^{j_2})$$

and the remainder R_n satisfies

$$\mathbb{E}(R_n)^2 \lesssim \mathbb{E}G_{2\nu}(X_1, X_2)/n^3 \lesssim_d \|p\|^2_{L_2} \nu_n^{-d/2}/n^3.$$

See Appendix B.4 for more details.

Moreover, borrowing arguments in the proof of Lemma 1, we obtain

$$
\begin{aligned}
&\mathbb{E}(G^*_{\nu_n}(X_1, X_2) - \bar{G}_{\nu_n}(X_1, X_2))^2 \\
\lesssim &\sum_{1\le j\le 2} \mathbb{E}\Big(g_j(X_1^j, X_2)\Big)^2 + \sum_{1\le j_1, j_2\le 2} \mathbb{E}\Big(g_{j_1,j_2}(X_1^{j_1}, X_2^{j_2})\Big)^2 \\
\le &\sum_{1\le j_1\ne j_2\le 2} \mathbb{E}G_{2\nu_n}(X_1^{j_1}, X_2^{j_1}) \cdot \mathbb{E}\left\{\mathbb{E}\left[G_{\nu_n}(X_1^{j_2}, X_2^{j_2})\Big|X_1^{j_2}\right]\right\}^2 + \\
&\sum_{1\le j_1\ne j_2\le 2} \mathbb{E}G_{2\nu_n}(X_1^{j_1}, X_2^{j_1})[\mathbb{E}G_{\nu_n}(X_1^{j_2}, X_2^{j_2})]^2 + \\
&2\mathbb{E}\left\{\mathbb{E}\left[G_{\nu_n}(X_1^1, X_2^1)\Big|X_1^1\right]\right\}^2 \mathbb{E}\left\{\mathbb{E}\left[G_{\nu_n}(X_1^2, X_2^2)\Big|X_1^2\right]\right\}^2 \\
\lesssim_d &\ \nu_n^{-d_1/2-3d_2/4}\|p_1\|_{L_2}^2\|p_2\|_{L_2}^3 + \nu_n^{-3d_1/4-d_2/2}\|p_1\|_{L_2}^3\|p_2\|_{L_2}^2
\end{aligned}
$$

Together with the fact that

$$(2\nu_n/\pi)^{d/2}\mathbb{E}\bar{G}^2_{\nu_n}(X_1, X_2) \to \|p\|^2_{L_2}$$

as $\nu_n \to \infty$, we conclude that

$$\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{P}^{X^1} \otimes \mathbb{P}^{X^2}) = D(\nu_n) + o_p\left(\sqrt{\mathbb{E}D^2(\nu_n)}\right),$$

where

$$D(\nu_n) = \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} \bar{G}_{\nu_n}(X_i, X_j).$$

Applying arguments similar to those in the proofs of Theorem 7 and 10, we have

$$\frac{D(\nu_n)}{\sqrt{\mathbb{E}D^2(\nu_n)}} \to_d N(0, 1).$$

Since

$$\mathbb{E}D^2(\nu_n) = \frac{2}{n(n-1)}\mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2 \quad \text{and} \quad \mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2/\mathbb{E}[G^*_{\nu_n}(X_1, X_2)]^2 \to 1,$$

it remains to prove

$$\widehat{s}^2_{n,\nu_n}/\mathbb{E}[G^*_{\nu_n}(X_1,X_2)]^2 \to_p 1,$$

which immediately follows by observing

$$\tilde{s}^2_{n,\nu_n}/\mathbb{E}[G^*_{\nu_n}(X_1,X_2)]^2 = \prod_{j=1}^{2} \tilde{s}^2_{n,j,\nu_n}/\mathbb{E}[\bar{G}_{\nu_n}(X_1^j,X_2^j)]^2 \to_p 1$$

and $1/n^2 = o(\mathbb{E}[G^*_{\nu_n}(X_1,X_2)]^2)$. The proof is therefore concluded. □

Proof of Theorem 13. We prove the two parts separately. **Part (i).** The proof of consistency of $\Phi^{\mathrm{IND}}_{n,\nu_n,\alpha}$ is very similar to its counterpart in the proof of Theorem 11. It sufficies to show

$$\sup_{p\in H_1^{\mathrm{IND}}(\Delta_n,s)} \frac{\mathrm{var}(\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2}))}{\gamma^4_{\nu_n}(\mathbb{P},\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2})} \to 0, \tag{6.35}$$

$$\inf_{p\in H_1^{\mathrm{IND}}(\Delta_n,s)} \frac{n\gamma^2_{\nu_n}(\mathbb{P},\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2})}{\mathbb{E}\left(\widehat{s}_{n,\nu_n}\right)} \to \infty, \tag{6.36}$$

as $n\to\infty$.

We begin with (6.35). Let $f = p - p_1\otimes p_2$. Lemma 8 then implies that there exists $C = C(s,M) > 0$ such that

$$\gamma^2_\nu(\mathbb{P},\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2}) \asymp_d \nu^{-d/2}\|f\|^2_{L_2}$$

for $\nu \geq C\|f\|^{-2/s}_{L_2}$, which is satisfied by all $p\in H_1^{\mathrm{IND}}(\Delta_n,s)$ given $\nu=\nu_n$ and $\lim_{n\to\infty}\Delta_n n^{\frac{2s}{4s+d}} = \infty$. On the other hand, we can still do the decomposition of $\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2})$ as in Appendix B.4. We follow the same notations here.

Under the alternative hypothesis, the "first order" term

$$
\begin{aligned}
&D_1(\nu_n)\\
=&\frac{2}{n}\sum_{1\le i\le n}\left(\mathbb{E}_{X_i,X\sim_{\text{iid}}\mathbb{P}}[G_{\nu_n}(X_i,X)|X_i]-\mathbb{E}_{X,X'\sim_{\text{iid}}\mathbb{P}}G_{\nu_n}(X,X')\right)\\
&-\frac{2}{n}\sum_{1\le i\le n}\left(\mathbb{E}_{X_i\sim\mathbb{P},Y\sim\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2}}[G_{\nu_n}(X_i,Y)|X_i]-\mathbb{E}_{X\sim\mathbb{P},Y\sim\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2}}G_{\nu_n}(X,Y)\right)\\
&-\sum_{1\le j\le 2}\left(\frac{2}{n}\sum_{1\le i\le n}\left(\mathbb{E}_{X_i\sim\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2},X\sim\mathbb{P}}[G_{\nu_n}(X_i,X)|X_i^j]-\mathbb{E}_{X\sim\mathbb{P},Y\sim\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2}}G_{\nu_n}(X,Y)\right)\right)\\
&+\sum_{1\le j\le 2}\left(\frac{2}{n}\sum_{1\le i\le n}\left(\mathbb{E}_{X_i,Y\sim_{\text{iid}}\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2}}[G_{\nu_n}(X_i,Y)|X_i^j]-\mathbb{E}_{Y,Y'\sim_{\text{iid}}\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2}}G_{\nu_n}(Y,Y')\right)\right)
\end{aligned}
$$

no longer vanish, but based on arguments similar to those in the proof of Theorem 8,

$$\mathbb{E}D_1^2(\nu_n)\lesssim_d Mn^{-1}\nu_n^{-3d/4}\|f\|_{L_2}^2.$$

Moreover, the "second order" term $D_2(\nu_n)$ is not solely $\sum_{1\le i\ne j\le n}G^*_{\nu_n}(X_i,X_j)/(n(n-1))$, but we still have

$$\mathbb{E}D_2^2(\nu_n)\lesssim n^{-2}\max\{\mathbb{E}G_{2\nu_n}(X_1,X_2),\mathbb{E}G_{2\nu_n}(X_1^1,X_2^1)\mathbb{E}G_{2\nu_n}(X_1^2,X_2^2)\}\lesssim_d M^2n^{-2}\nu_n^{-d/2}.$$

Similarly, define the third order term $D_3(\nu_n)$ and the fourth order term $D_4(\nu_n)$ as the aggregation of all 3-variate centered components and the aggregation of all 4-variate centered components in $\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2})$ respectively, which together constitue R_n. Then we have

$$\mathbb{E}D_3^2(\nu_n)\lesssim_d M^2n^{-3}\nu_n^{-d/2},\quad \mathbb{E}D_4^2(\nu_n)\lesssim_d M^2n^{-4}\nu_n^{-d/2}.$$

Hence we finally obtain

$$\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2})=\gamma^2_{\nu_n}(\mathbb{P},\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2})+\sum_{l=1}^4 D_l(\nu_n)$$

and

$$\mathrm{var}\left(\widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{P}^{X^1} \otimes \mathbb{P}^{X^2})\right) = \sum_{l=1}^{4} \mathbb{E}D_l^2(\nu_n) \lesssim_d Mn^{-1}\nu_n^{-3d/4}\|f\|_{L_2}^2 + M^2 n^{-2}\nu_n^{-d/2}$$

which proves (6.35).

Now consider (6.36). Since

$$\widehat{s}_{n,\nu_n} \le \max\left\{\prod_{j=1}^{2}\sqrt{\left|\tilde{s}^2_{n,j,\nu_n}\right|}, 1/n\right\},$$

we have

$$\mathbb{E}\left(\widehat{s}_{n,\nu_n}\right) \le \prod_{j=1}^{2}\sqrt{\mathbb{E}\left|\tilde{s}^2_{n,j,\nu_n}\right|} + 1/n,$$

where

$$\prod_{j=1}^{2}\mathbb{E}\left|\tilde{s}^2_{n,j,\nu_n}\right| \lesssim \prod_{j=1}^{2}\mathbb{E}G_{2\nu_n}(X_1^j, X_2^j) = \mathbb{E}_{Y_1,Y_2\sim_{\mathrm{iid}}\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2}}G_{2\nu_n}(Y_1, Y_2) \lesssim_d M^2\nu_n^{-d/2}.$$

Therefore (6.36) holds.

Part (ii). Then we verify that $n^{2s/(d+4s)}\Delta_n \to \infty$ is also the necessary condition for the existence of consistent asymptotic α-level tests for any $\alpha \in (0,1)$. Similarly to the proof of Theorem 11, the idea is to relate the existence of consistent independence test to the existence of consistent goodness-of-fit test.

Let $p_{j,0} \in \mathcal{W}^{s,2}\left(M_j/\sqrt{2}\right)$ be density on $\mathbb{R}^{d_j}$ for $j = 1, 2$ and p_0 be the product of $p_{1,0}$ and $p_{2,0}$, *i.e.*,

$$p_0(x^1, x^2) = p_{1,0}(x^1)p_{2,0}(x^2), \quad \forall\, x^1 \in \mathbb{R}^{d_1}, x^2 \in \mathbb{R}^{d_2}.$$

Hence $p_0 \in \mathcal{W}^{s,2}(M/2)$.

Let

$$H_1^{\mathrm{GOF}}(\Delta_n; s)' := \{p : \ p \in \mathcal{W}^{s,2}(M),\ p_1 = p_{1,0},\ p_2 = p_{2,0}, \|p - p_0\|_{L_2} \ge \Delta_n\}.$$

We immediately have

$$H_1^{\mathrm{IND}}(\Delta_n; s) \supset H_1^{\mathrm{GOF}}(\Delta_n; s)'$$

Let $\{\Phi_n\}_{n\geq 1}$ be any sequence of asymptotic α-level independence tests, where

$$\Phi_n = \Phi_n(X_1, \cdots, X_n).$$

Then $\{\Phi_n\}_{n\geq 1}$ can also be treated as a sequence of asymptotic α-level goodness-of-fit tests with the null density being p_0. Moreover,

$$\mathrm{power}\{\Phi_n; H_1^{\mathrm{IND}}(\Delta_n; s)\} \leq \mathrm{power}\{\Phi_n; H_1^{\mathrm{GOF}}(\Delta_n; s)'\}.$$

It remains to show that given $\liminf_{n\to\infty} n^{2s/(d+4s)}\Delta_n < \infty$, there exists some $\alpha \in (0, 1)$ such that

$$\liminf_{n\to\infty} \mathrm{power}\{\Phi_n; H_1^{\mathrm{GOF}}(\Delta_n; s)'\} < 1,$$

which cannot be directly obtained from Theorem 9 because of the additional constraints

$$p_1 = p_{1,0}, \quad p_2 = p_{2,0} \tag{6.37}$$

in $H_1^{\mathrm{GOF}}(\Delta_n; s)'$.

However, by modifying the proof of Theorem 9, we only need to further require each $p_{n,\boldsymbol{a}_n}$ in the proof of Theorem 9 satisfying (6.37), or equivalently,

$$\int_{\mathbb{R}^{d_2}} (p - p_0)(x^1, x^2)dx^2 = 0, \quad \int_{\mathbb{R}^{d_1}} (p - p_0)(x^1, x^2)dx^1 = 0.$$

Recall that each $p_{n,\boldsymbol{a}_n} = p_0 + r_n \sum_{k=1}^{B_n} a_{n,k}\phi_{n,k}$, where

$$\phi_{n,k}(x) = \frac{b_n^{d/2}}{\|\phi\|_{L_2}}\phi(b_n x - x_{n,k}).$$

Write $x_{n,k} = (x_{n,k}^1, x_{n,k}^2) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$. Since ϕ can be decomposed as

$$\phi(x^1, x^2) = \phi_1(x^1)\phi_2(x^2),$$

we have

$$\phi_{n,k}(x) = \frac{b_n^{d/2}}{\|\phi\|_{L_2}}\phi_1(b_n x^1 - x_{n,k}^1)\phi_2(b_n x^2 - x_{n,k}^2)$$

Hence

$$\begin{aligned}\int_{\mathbb{R}^{d_2}} (p_{n,\boldsymbol{a}_n} - p_0)(x^1, x^2)dx^2 =& r_n \sum_{k=1}^{B_n} a_{n,k} \int_{\mathbb{R}^{d_2}} \phi_{n,k}(x^1, x^2)dx^2 \\ =& r_n \sum_{k=1}^{B_n} a_{n,k} \frac{b_n^{d/2}}{\|\phi\|_{L_2}} \cdot \phi_1(b_n x^1 - x_{n,k}^1) \cdot \frac{1}{b_n^{d_2}} \int_{\mathbb{R}^{d_2}} \phi_2(x^2)dx^2 \\ =& 0\end{aligned}$$

since $\int_{\mathbb{R}^{d_2}} \phi_2(x^2)dx^2 = 0$. Similarly, $\int_{\mathbb{R}^{d_1}} (p_{n,\boldsymbol{a}_n} - p_0)(x^1, x^2)dx^1 = 0$. The proof is therefore finished. □

Proof of Theorem 14. The proof of Theorem 14 consists of two steps. First, we bound $q_{n,\alpha}^{\mathrm{GOF}}$. To be more specific, we show that there exists $C = C(d) > 0$ such that

$$q_{n,\alpha}^{\mathrm{GOF}} \leq C(d) \log\log n$$

for sufficiently large n, which holds if

$$\lim_{n\to\infty} P(T_n^{\mathrm{GOF(adapt)}} \geq C(d)\log\log n) = 0 \tag{6.38}$$

under H_0^{GOF}. Second, we show that there exists $c > 0$ such that

$$\liminf_{n\to\infty} \Delta_{n,s}(n/\log\log n)^{2s/(d+4s)} > c$$

ensures

$$\inf_{p\in H_1^{\mathrm{GOF(adapt)}}(\Delta_{n,s}:s\geq d/4)} P(T_n^{\mathrm{GOF(adapt)}} \geq C(d)\log\log n) \to 1 \tag{6.39}$$

as $n \to \infty$.

Verifying (6.38). In order to prove (6.38), we first show the following two lemmas. The first lemma suggests that $\widehat{s}^2_{n,\nu_n}$ is a consistent estimator of $\mathbb{E}\bar{G}^2_{\nu_n}(X_1, X_2)$ uniformly over all $\nu_n \in [1, n^{2/d}]$. Recall we have shown in the proof of Theorem 7 that for ν_n increasing at a proper rate,

$$\widehat{s}^2_{n,\nu_n}/\mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2 \to_p 1.$$

Hence the first lemma is a uniform version of such result.

Lemma 5. *We have that* $\widehat{s}^2_{n,\nu_n}/\mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2$ *converges to* 1 *uniformly over* $\nu_n \in [1, n^{2/d}]$, *i.e.,*

$$\sup_{1\leq\nu_n\leq n^{2/d}} \left|\widehat{s}^2_{n,\nu_n}/\mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2 - 1\right| = o_p(1).$$

We defer the proof of Lemma 5 to the appendix. Note that

$$T_n^{\mathrm{GOF(adapt)}} = \sup_{1\le \nu_n \le n^{2/d}} \frac{n\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{P}_0)}{\sqrt{2\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2}} \cdot \sqrt{\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2/\widehat{s}^2_{n,\nu_n}}$$
$$\le \sup_{1\le \nu_n \le n^{2/d}} \left| \frac{n\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{P}_0)}{\sqrt{2\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2}} \right| \cdot \sup_{1\le \nu_n \le n^{2/d}} \sqrt{\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2/\widehat{s}^2_{n,\nu_n}}.$$

Lemma 5 first ensures that

$$\sup_{1\le \nu_n \le n^{2/d}} \sqrt{\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2/\widehat{s}^2_{n,\nu_n}} = 1 + o_p(1).$$

It therefore suffices to show that under H_0^{GOF},

$$\widetilde{T}_n^{\mathrm{GOF(adapt)}} := \sup_{1\le \nu_n \le n^{2/d}} \left| \frac{n\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{P}_0)}{\sqrt{2\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2}} \right|$$

is also of order $\log\log n$. This is the crux of our argument yet its proof is lengthy. For brevity, we shall state it as a lemma here and defer its proof to the appendix.

Lemma 6. *There exists $C = C(d) > 0$ such that*

$$\lim_{n\to\infty} P\left(\widetilde{T}_n^{\mathrm{GOF(adapt)}} \ge C\log\log n\right) = 0$$

under H_0^{GOF}.

Verifying (6.39). Let

$$\nu_n(s)' = \left(\frac{\log\log n}{n}\right)^{-4/(4s+d)},$$

which is smaller than $n^{2/d}$ for $s \ge d/4$. Hence it suffices to show

$$\inf_{s\ge d/4} \inf_{p\in H_1^{\mathrm{GOF}}(\Delta_{n,s};s)} P(T^{\mathrm{GOF}}_{n,\nu_n(s)'} \ge C(d)\log\log n) \to 1$$

as $n \to \infty$.

First of all, observe

$$0 \le \mathbb{E}\left(\tilde{s}^2_{n,\nu_n(s)'}\right) \le \mathbb{E}G_{2\nu_n(s)'}(X_1, X_2) \le M^2(2\nu_n(s)'/\pi)^{-d/2}$$

and

$$\mathrm{var}\left(\tilde{s}^2_{n,\nu_n(s)'}\right) \lesssim_d M^3 n^{-1}(\nu_n(s)')^{-3d/4} + M^2 n^{-2}(\nu_n(s)')^{-d/2}$$

for any s and $p \in H_1^{\mathrm{GOF}}(\Delta_{n,s}, s)$. Further considering $1/n^2 = o(M^2(2\nu_n(s)'/\pi)^{-d/2})$ uniformly over all s, we obtain that

$$\inf_{s \ge d/4} \inf_{p \in H_1^{\mathrm{GOF}}(\Delta_{n,s};s)} P\left(\widehat{s}^2_{n,\nu_n(s)'} \le 2M^2(2\nu_n(s)'/\pi)^{-d/2}\right) \to 1.$$

Let

$$\Delta_{n,s} \ge c(\sqrt{M} + M)(\log\log n/n)^{2s/(d+4s)}$$

for some sufficiently large $c = c(d)$. Then

$$\widehat{\mathbb{E}\gamma^2_{\nu_n(s)'}}(\mathbb{P}, \mathbb{P}_0) = \gamma^2_{\nu_n(s)'}(\mathbb{P}, \mathbb{P}_0) \ge \left(\frac{\pi}{\nu_n(s)'}\right)^{d/2} \cdot \frac{\|p - p_0\|^2_{L_2}}{4},$$

as guaranteed by Lemma 8. Further considering that

$$\mathrm{var}\left(\widehat{\gamma^2_{\nu_n(s)'}}(\mathbb{P}, \mathbb{P}_0)\right) \lesssim_d M^2 n^{-2}(\nu_n(s)')^{-d/2} + Mn^{-1}(\nu_n(s)')^{-3d/4}\|p - p_0\|^2_{L_2},$$

we immediately have

$$\begin{aligned}
&\lim_{n\to\infty} \inf_{s \ge d/4} \inf_{p \in H_1^{\mathrm{GOF}}(\Delta_{n,s};s)} P(T^{\mathrm{GOF}}_{n,\nu_n(s)'} \ge C(d)\log\log n) \\
&\ge \lim_{n\to\infty} \inf_{s \ge d/4} \inf_{p \in H_1^{\mathrm{GOF}}(\Delta_{n,s};s)} P\left(\frac{n\gamma^2_{\nu_n(s)'}(\mathbb{P}, \mathbb{P}_0)/2}{\sqrt{2\widehat{s}^2_{n,\nu_n(s)'}}} \ge C(d)\log\log n\right) = 1.
\end{aligned}$$

□

Proof of Theorem 15 and Theorem 16. The proof of Theorem 15 and Theorem 16 is very similar to that of Theorem 14. Hence we only emphasize the main differences here.

For adaptive homogeneity test: to verify that there exists $C = C(d) > 0$ such that

$$\lim_{n\to\infty} P(T_n^{\mathrm{HOM(adapt)}} \geq C \log\log n) = 0$$

under H_0^{HOM}, observe that

$$T_n^{\mathrm{HOM(adapt)}} \leq \sup_{1\leq\nu_n\leq n^{2/d}} \sqrt{\frac{\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2}{\widehat{s}^2_{n,m,\nu_n}}} \cdot \left(\frac{1}{n}+\frac{1}{m}\right)^{-1} \sup_{1\leq\nu_n\leq n^{2/d}} \frac{|\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{Q})|}{\sqrt{2\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2}}.$$

Denote $X_1,\cdots,X_n,Y_1,\cdots,Y_m$ as $Z_1,\cdots,Z_N$. Hence

$$2\sum_{i=1}^{n}\sum_{j=1}^{m} G_{\nu_n}(X_i,Y_j) = \sum_{1\leq i\neq j\leq N} G_{\nu_n}(Z_i,Z_j) - \sum_{1\leq i\neq j\leq n} G_{\nu_n}(X_i,X_j) - \sum_{1\leq i\neq j\leq m} G_{\nu_n}(Y_i,Y_j)$$

and

$$\begin{aligned}
&\sup_{1\leq\nu_n\leq n^{2/d}} \frac{|\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{Q})|}{\sqrt{2\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2}} \\
&\leq \left(\frac{1}{n(n-1)}+\frac{1}{mn}\right) \sup_{1\leq\nu_n\leq n^{2/d}} \left|\sum_{1\leq i\neq j\leq n} \frac{\bar{G}_{\nu_n}(X_i,X_j)}{\sqrt{2\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2}}\right| \\
&\quad + \left(\frac{1}{m(m-1)}+\frac{1}{mn}\right) \sup_{1\leq\nu_n\leq n^{2/d}} \left|\sum_{1\leq i\neq j\leq m} \frac{\bar{G}_{\nu_n}(Y_i,Y_j)}{\sqrt{2\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2}}\right| \\
&\quad + \frac{1}{mn} \sup_{1\leq\nu_n\leq n^{2/d}} \left|\sum_{1\leq i\neq j\leq N} \frac{\bar{G}_{\nu_n}(Z_i,Z_j)}{\sqrt{2\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2}}\right|
\end{aligned}$$

Apply Lemma 6 to bound each term of the right hand side of the above inequality. Then we

conclude that for some $C = C(d) > 0$,

$$\lim_{n\to\infty} P\left(\left(\frac{1}{n}+\frac{1}{m}\right)^{-1} \sup_{1\le \nu_n \le n^{2/d}} \frac{|\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{Q})|}{\sqrt{2\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2}} \ge C\log\log n\right) = 0.$$

For adaptive independence test: to verify that there exists $C = C(d) > 0$ such that

$$\lim_{n\to\infty} P(T_n^{\mathrm{IND(adapt)}} \ge C\log\log n) = 0 \tag{6.40}$$

under H_0^{IND}, recall the decomposition

$$\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2}) = D_2(\nu_n) + R_n = \frac{1}{n(n-1)}\sum_{1\le i\ne j\le n} G^*_{\nu_n}(X_i,X_j) + R_n,$$

where we express R_n as $R_n = D_3(\nu_n) + D_4(\nu_n)$ in the proof of Theorem 13.

Following arguments similar to those in the proof of Lemma 6, we obtain that there exists $C(d) > 0$ such that for sufficiently large n,

$$P\left(\sup_{1\le\nu_n\le n^{2/d}}\left|\frac{nD_2(\nu_n)}{\sqrt{2\mathbb{E}[G^*_{\nu_n}(X_1,X_2)]^2}}\right| \ge C(d)(\log\log n + t\log\log\log n)\right) \lesssim \exp(-t^{2/3}),$$

Similarly,

$$P\left(\sup_{1\le\nu_n\le n^{2/d}}\left|\frac{n^{3/2}D_3(\nu_n)}{\sqrt{2\mathbb{E}[G^*_{\nu_n}(X_1,X_2)]^2}}\right| \ge C(d)(\log\log n + t\log\log\log n)\right) \lesssim \exp(-t^{1/2})$$

$$P\left(\sup_{1\le\nu_n\le n^{2/d}}\left|\frac{n^{2}D_4(\nu_n)}{\sqrt{2\mathbb{E}[G^*_{\nu_n}(X_1,X_2)]^2}}\right| \ge C(d)(\log\log n + t\log\log\log n)\right) \lesssim \exp(-t^{2/5})$$

for sufficiently large n.

On the other hand, note that

$$\mathbb{E}[G^*_{\nu_n}(X_1, X_2)]^2 = \prod_{j=1}^{2} \mathbb{E}[\bar{G}_{\nu_n}(X_1^j, X_2^j)]^2,$$

and based on results in the proof of Lemma 5, $\sup_{1\le\nu_n\le n^{2/d}} \left|\tilde{s}^2_{n,j,\nu_n}/\mathbb{E}[\bar{G}_{\nu_n}(X_1^j, X_2^j)]^2 - 1\right| = o_p(1)$ for $j = 1, 2$. Further considering that

$$1/n^2 = o(\mathbb{E}[G^*_{\nu_n}(X_1, X_2)]^2)$$

uniformly over all $\nu_n \in [1, n^{2/d}]$, we obtain

$$\sup_{1\le\nu_n\le n^{2/d}} \left|\widehat{s}^2_{n,\nu_n}/\mathbb{E}[G^*_{\nu_n}(X_1, X_2)]^2 - 1\right| = o_p(1).$$

They combined together ensure that (6.40) holds.

To show that the detection boundary of $\Phi^{\mathrm{IND(adapt)}}$ is of order $O((n/\log\log n)^{-2s/(d+4s)})$, observe that

$$0 \le \mathbb{E}\left(\tilde{s}^2_{n,j,\nu_n(s)'}\right) \le \mathbb{E}G_{2\nu_n(s)'}(X_1^j, X_2^j) \le M_j^2(2\nu_n(s)'/\pi)^{-d_j/2}$$

and

$$\mathrm{var}\left(\tilde{s}^2_{n,j,\nu_n(s)'}\right) \lesssim_{d_j} M_j^3 n^{-1}(\nu_n(s)')^{-3d_j/4} + M_j^2 n^{-2}(\nu_n(s)')^{-d_j/2}$$

for $j = 1, 2$, where $\nu_n(s)' = (\log\log n/n)^{-4/(4s+d)}$ as in the proof of Theorem 14. Therefore,

$$\inf_{s\ge d/4} \inf_{p\in H_1^{\mathrm{IND}}(\Delta_{n,s};s)} P\left(\left|\tilde{s}^2_{n,j,\nu_n(s)'}\right| \le \sqrt{3/2}M_j^2(2\nu_n(s)'/\pi)^{-d_j/2}\right) \to 1, \quad j = 1, 2.$$

Further considering $1/n^2 = o(M^2(2\nu_n(s)'/\pi)^{-d/2})$ uniformly over all s, we obtain that

$$\inf_{s\ge d/4} \inf_{p\in H_1^{\mathrm{IND}}(\Delta_{n,s};s)} P\left(\widehat{s}^2_{n,\nu_n(s)'} \le 2M^2(2\nu_n(s)'/\pi)^{-d/2}\right) \to 1.$$

□

References

L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi (2010). "On combinatorial testing problems". In: *The Annals of Statistics* 38.5, pp. 3063–3092.

N. Ailon, M. Charikar, and A. Newman (2008). "Aggregating inconsistent information: ranking and clustering". In: *Journal of ACM* 55.5, 23:1–23:27.

M. A. Arcones and E. Gine (1993). "Limit Theorems for U-Processes". In: *The Annals of Probability* 21.3, pp. 1494–1542.

Y. Baraud (2002). "Non-asymptotic minimax rates of testing in signal detection". In: *Bernoulli* 8.5, pp. 577–606.

M. V. Burnashev (1979). "On the minimax detection of an inaccurately known signal in a white Gaussian noise background". In: *Theory of Probability & Its Applications* 24.1, pp. 107–119.

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

M. S. Ermakov (1991). "Minimax detection of a signal in a Gaussian white noise". In: *Theory of Probability & Its Applications* 35.4, pp. 667–679.

M. Fromont and B. Laurent (2006). "Adaptive goodness-of-fit tests in a density model". In: *The Annals of Statistics* 34.2, pp. 680–720.

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

M. Fromont, B. Laurent, and P. Reynaud-Bouret (2013). "The two-sample problem for poisson processes: Adaptive tests with a nonasymptotic wild bootstrap approach". In: *The Annals of Statistics* 41.3, pp. 1431–1461.

K. Fukumizu, A. Gretton, G. R. Lanckriet, B. Schölkopf, and B. K. Sriperumbudur (2009). "Kernel choice and classifiability for RKHS embeddings of probability distributions". In: *Advances in Neural Information Processing Systems*, pp. 1750–1758.

G. G. Gregory (1977). "Large sample theory for U-statistics and tests of fit". In: *The Annals of Statistics* 5.1, pp. 110–123.

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

A. Gretton, O. Bousquet, A. J. Smola, and B. Schölkopf (2005). "Measuring statistical dependence with Hilbert-Schmidt norms". In: *International Conference on Algorithmic Learning Theory*. Springer, pp. 63–77.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola (2008). "A kernel statistical test of independence". In: *Advances in Neural Information Processing Systems*, pp. 585–592.

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

E. Haeusler (1988). "On the rate of convergence in the central limit theorem for martingales with discrete and continuous time". In: *The Annals of Probability* 16.1, pp. 275–299.

P. Hall (1984). "Central limit theorem for integrated square error of multivariate nonparametric density estimators". In: *Journal of Multivariate Analysis* 14.1, pp. 1–16.

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

Y. I. Ingster (1987). "Minimax testing of nonparametric hypotheses on a distribution density in the L_p metrics". In: *Theory of Probability & Its Applications* 31.2, pp. 333–337.

— (1993). "Asymptotically minimax hypothesis testing for nonparametric alternatives. I, II, III". In: *Mathematical Methods of Statistics* 2.2, pp. 85–114.

— (2000). "Adaptive chi-square tests". In: *Journal of Mathematical Sciences* 99.2, pp. 1110–1119.

Y. I. Ingster and I. A. Suslina (2000). "Minimax nonparametric hypothesis testing for ellipsoids and Besov bodies". In: *ESAIM: Probability and Statistics* 4, pp. 53–135.

— (2003). *Nonparametric Goodness-of-Fit Testing under Gaussian Models*. New York, NY: Springer.

P. E. Jupp (2005). "Sobolev tests of goodness of fit of distributions on compact Riemannian manifolds". In: *The Annals of Statistics* 33.6, pp. 2957–2966.

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

O. V. Lepski and V. G. Spokoiny (1999). “Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative”. In: *Bernoulli* 5.2, pp. 333–358.

R. Lyons (2013). “Distance covariance in metric spaces”. In: *The Annals of Probability* 41.5, pp. 3284–3305.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf (2016). “Distinguishing cause from effect using observational data: methods and benchmarks”. In: *The Journal of Machine Learning Research* 17.1, pp. 1103–1204.

K. Muandet, K. Fukumizu, B. K. Sriperumbudur, and B. Schölkopf (2017). “Kernel mean embedding of distributions: a review and beyond”. In: *Foundations and Trends® in Machine Learning* 10.1-2, pp. 1–141.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf (2014). “Causal discovery with continuous additive noise models”. In: *The Journal of Machine Learning Research* 15.1, pp. 2009–2053.

N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters (2018). “Kernel-based tests for joint independence”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.1, pp. 5–31.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu (2013). “Equivalence of distance-based and RKHS-based statistics in hypothesis testing”. In: *The Annals of Statistics* 41.5, pp. 2263–2291.

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

V. G. Spokoiny (1996). “Adaptive hypothesis testing using wavelets”. In: *The Annals of Statistics* 24.6, pp. 2477–2498.

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

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

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

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

I. Steinwart (2001). “On the influence of the kernel on the consistency of support vector machines”. In: *Journal of machine learning research* 2.Nov, pp. 67–93.

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

G. J Székely and M. L. Rizzo (2009). “Brownian distance covariance”. In: *The Annals of Applied Statistics* 3.4, pp. 1236–1265.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov (2007). “Measuring and testing dependence by correlation of distances”. In: *The Annals of Statistics* 35.6, pp. 2769–2794.

M. Talagrand (2014). *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems*. Springer Science & Business Media.

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

Appendix A: Some Technical Results and Proofs Related to Chapter 2

Proof of Lemma 3. 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 of Lemma 4. For brevity, write

$$l_K = \sum_{k=1}^{K} \frac{a_k^2}{\lambda_k}.$$

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 K be such that $l_K^2 \le R^2 \le l_{K+1}^2$, and denote by

$$f_R = \sum_{k=1}^{K} a_k \varphi_k + a_{K+1}^*(R)\varphi_{K+1},$$

where

$$a_{K+1}^*(R) = \operatorname{sgn}(a_{K+1})\sqrt{\lambda_{K+1}(R^2 - l_K^2)}.$$

Clearly,

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

and

$$\|u - f_R\|_{L_2(P_0)}^2 = \sum_{k>K+1} a_k^2 + \left(|a_{K+1}| - \sqrt{\lambda_{K+1}(R^2 - l_K^2)}\right)^2 \le \sum_{k\ge K+1} a_k^2.$$

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

$$\sup_{l_K^2 \le R^2 \le l_{K+1}^2} \|u - f_R\|_{L_2(P_0)}^2 R^{2/\theta} \le M^2, \quad \forall\, K \ge 0,$$

which concludes the proof. □

Appendix B: Some Technical Results and Proofs Related to Chapter 3

B.1 Properties of Gaussian Kernel

We collect here a couple of useful properties of Gaussian kernel that we used repeated in the proof to the main results.

Lemma 7. *For any* $f \in L_2(\mathbb{R}^d)$,

$$\int G_\nu(x,y)f(x)f(y)dxdy = \left(\frac{\pi}{\nu}\right)^{\frac{d}{2}} \int \exp\left(-\frac{\|\omega\|^2}{4\nu}\right) \|\mathcal{F}f(\omega)\|^2 \, d\omega.$$

Proof. Denote by Z a Gaussian random vector with mean 0 and covariance matrix $2\nu I_d$. Then

$$\begin{aligned}
\int G_\nu(x,y)f(x)f(y)dxdy &= \int \exp\left(-\nu\|x-y\|^2\right) f(x)f(y)dxdy \\
&= \int \mathbb{E}\exp[iZ^\top(x-y)]f(x)f(y)dxdy \\
&= \mathbb{E}\left\|\int \exp(-iZ^\top x)f(x)dx\right\|^2 \\
&= \int \frac{1}{(4\pi\nu)^{d/2}} \exp\left(-\frac{\|\omega\|^2}{4\nu}\right) \left\|\int \exp(-i\omega^\top x)f(x)dx\right\|^2 \\
&= \left(\frac{\pi}{\nu}\right)^{\frac{d}{2}} \int \exp\left(-\frac{\|\omega\|^2}{4\nu}\right) \|\mathcal{F}f(\omega)\|^2 \, d\omega,
\end{aligned}$$

which concludes the proof. □

A useful consequence of Lemma 7 is a close connection between Gaussian kernel MMD and L_2 norm.

Lemma 8. *For any* $f \in \mathcal{W}^{s,2}(M)$

$$\left(\frac{\nu}{\pi}\right)^{d/2} \int G_\nu(x,y) f(x) f(y) dxdy \geq \frac{1}{4}\|f\|_{L_2}^2,$$

provided that

$$\nu^s \geq \frac{4^{1-s} M^2}{(\log 3)^s} \cdot \|f\|_{L_2}^{-2}.$$

Proof. In light of Lemma 7,

$$\left(\frac{\nu}{\pi}\right)^{d/2} \int G_\nu(x,y) f(x) f(y) dxdy = \int \exp\left(-\frac{\|\omega\|^2}{4\nu}\right) \|\mathcal{F} f(\omega)\|^2 \, d\omega.$$

By Plancherel Theorem, for any $T > 0$,

$$\int_{\|\omega\| \leq T} \|\mathcal{F} f(\omega)\|^2 \, d\omega = \|f\|_{L^2}^2 - \int_{\|\omega\| > T} \|\mathcal{F} f(\omega)\|^2 \, d\omega \geq \|f\|_{L^2}^2 - \frac{M^2}{T^{2s}},$$

Choosing

$$T = \left(\frac{2M}{\|f\|_{L^2}}\right)^{1/s},$$

yields

$$\int_{\|\omega\| \leq T} \|\mathcal{F} f(\omega)\|^2 \, d\omega \geq \frac{3}{4}\|f\|_{L^2}^2.$$

Hence

$$\begin{aligned}\int \exp\left(-\frac{\|\omega\|^2}{4\nu}\right) \|\mathcal{F} f(\omega)\|^2 \, d\omega &\geq \exp\left(-\frac{T^2}{4\nu}\right) \int_{\|\omega\| \leq T} \|\mathcal{F} f(\omega)\|^2 \, d\omega \\ &\geq \frac{3}{4} \exp\left(-\frac{T^2}{4\nu}\right) \|f\|_{L^2}^2.\end{aligned}$$

In particular, if

$$\nu \geq \frac{(2M)^{2/s}}{4 \log 3} \cdot \|f\|_{L^2}^{-2/s},$$

then

$$\int \exp\left(-\frac{\|\omega\|^2}{4\nu}\right) \|\mathcal{F}f(\omega)\|^2 \, d\omega \geq \frac{1}{4}\|f\|_{L^2}^2,$$

which concludes the proof. □

B.2 Proof of Lemma 5

We first prove that $\sup_{1\leq \nu_n \leq n^{2/d}} \left|\tilde{s}_{n,\nu_n}^2 / \mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2 - 1\right| = o_p(1)$ and then show the difference caused by the modification from $\tilde{s}_{n,\nu_n}^2$ to $\widehat{s}_{n,\nu_n}^2$ is asymptotically negligible.

Note that

$$\begin{aligned}
&\sup_{1\leq \nu_n \leq n^{2/d}} \left|\tilde{s}_{n,\nu_n}^2 / \mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2 - 1\right| \\
&\leq \left(\inf_{1\leq \nu_n \leq n^{2/d}} \nu_n^{d/2} \mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2\right)^{-1} \cdot \sup_{1\leq \nu_n \leq n^{2/d}} \nu_n^{d/2} \left|\tilde{s}_{n,\nu_n}^2 - \mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2\right|.
\end{aligned}$$

For $X \sim \mathbb{P}_0$, denote the distribution of (X, X) as $\mathbb{P}_1$. Then we have

$$\mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2 = \gamma_{\nu_n}^2(\mathbb{P}_1, \mathbb{P}_0 \otimes \mathbb{P}_0).$$

Hence $\mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2 > 0$ for any $\nu_n > 0$ since G_{ν_n} is characteristic.

In addition, $\nu_n^{d/2}\mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2$ is continuous with respect to ν_n and

$$\lim_{\nu_n \to \infty} \nu_n^{d/2} \mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2 = \left(\frac{\pi}{2}\right)^{d/2} \|p_0\|_{L_2}^2.$$

Therefore,

$$\inf_{1\leq \nu_n \leq n^{2/d}} \nu_n^{d/2} \mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2 \geq \inf_{\nu_n \in [0,\infty)} \nu_n^{d/2} \mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2 > 0,$$

and it remains to prove

$$\sup_{1\le \nu_n \le n^{2/d}} \nu_n^{d/2} \left| \tilde{s}^2_{n,\nu_n} - \mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2 \right| = o_p(1).$$

Recall the expression of $\tilde{s}^2_{n,\nu_n}$. It suffcies to show that

$$\sup_{1\le \nu_n \le n^{2/d}} \nu_n^{d/2} \left| \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} G_{2\nu_n}(X_i, X_j) - \mathbb{E}G_{2\nu_n}(X_1, X_2) \right| \tag{B.1}$$

$$\sup_{1\le \nu_n \le n^{2/d}} \nu_n^{d/2} \left| \frac{2(n-3)!}{n!} \sum_{\substack{1\le i,j_1,j_2\le n \\ |\{i,j_1,j_2\}|=3}} G_{\nu_n}(X_i, X_{j_1}) G_{\nu_n}(X_i, X_{j_2}) - \mathbb{E}G_{\nu_n}(X_1, X_2) G_{\nu_n}(X_1, X_3) \right| \tag{B.2}$$

$$\sup_{1\le \nu_n \le n^{2/d}} \nu_n^{d/2} \left| \frac{(n-4)!}{n!} \sum_{\substack{1\le i_1,i_2,j_1,j_2\le n \\ |\{i_1,i_2,j_1,j_2\}|=4}} G_{\nu_n}(X_{i_1}, X_{j_1}) G_{\nu_n}(X_{i_2}, X_{j_2}) - [\mathbb{E}G_{\nu_n}(X_1, X_2)]^2 \right| \tag{B.3}$$

are all $o_p(1)$. We shall first control (B.1) and then bound (B.2) and (B.3) in the same way.

Let

$$\widehat{\mathbb{E}}_n G_{2\nu_n}(X, X') = \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} G_{2\nu_n}(X_i, X_j).$$

In the rest of this proof, abbreviate $\widehat{\mathbb{E}}_n G_{2\nu_n}(X, X')$ and $\mathbb{E}G_{2\nu_n}(X_1, X_2)$ as $\widehat{\mathbb{E}}_n G_{2\nu_n}$ and $\mathbb{E}G_{2\nu_n}$ respectively when no confusion occurs.

Divide the whole interval $[1, n^{2/d}]$ into A sub-intervals, $[u_0, u_1], [u_1, u_2], \cdots, [u_{A-1}, u_A]$ with $u_0 = 1, u_A = n^{2/d}$. For any $\nu_n \in [u_{a-1}, u_a]$,

$$\begin{aligned}
\nu_n^{d/2} \widehat{\mathbb{E}}_n G_{2\nu_n} - \nu_n^{d/2} \mathbb{E}G_{2\nu_n} &\ge -\nu_n^{d/2} \left| \widehat{\mathbb{E}}_n G_{2u_a} - \mathbb{E}G_{2u_a} \right| - \nu_n^{d/2} \left| \mathbb{E}G_{2u_a} - \mathbb{E}G_{2u_{a-1}} \right| \\
&\ge -u_a^{d/2} \left| \widehat{\mathbb{E}}_n G_{2u_a} - \mathbb{E}G_{2u_a} \right| - u_a^{d/2} \left| \mathbb{E}G_{2u_a} - \mathbb{E}G_{2u_{a-1}} \right|
\end{aligned}$$

and

$$\nu_n^{d/2} \widehat{\mathbb{E}}_n G_{2\nu_n} - \nu_n^{d/2} \mathbb{E}G_{2\nu_n} \le u_a^{d/2} \left| \widehat{\mathbb{E}}_n G_{2u_{a-1}} - \mathbb{E}G_{2u_{a-1}} \right| + u_a^{d/2} \left| \mathbb{E}G_{2u_a} - \mathbb{E}G_{2u_{a-1}} \right|,$$

which together ensure that

$$
\begin{aligned}
&\sup_{1\le \nu_n \le n^{2/d}} \left| \nu_n^{d/2} \widehat{\mathbb{E}}_n G_{2\nu_n} - \nu_n^{d/2} \mathbb{E} G_{2\nu_n} \right| \\
&\le \sup_{1\le a\le A} \left(\frac{u_a}{u_{a-1}} \right)^{d/2} \cdot \sup_{0\le a\le A} u_a^{d/2} \left| \widehat{\mathbb{E}}_n G_{2u_a} - \mathbb{E} G_{2u_a} \right| + \sup_{1\le a\le A} u_a^{d/2} \left| \mathbb{E} G_{2u_a} - \mathbb{E} G_{2u_{a-1}} \right| \\
&\le \sup_{1\le a\le A} \left(\frac{u_a}{u_{a-1}} \right)^{d/2} \cdot \sup_{0\le a\le A} u_a^{d/2} \left| \widehat{\mathbb{E}}_n G_{2u_a} - \mathbb{E} G_{2u_a} \right| + \sup_{1\le a\le A} \left| u_a^{d/2} \mathbb{E} G_{2u_a} - u_{a-1}^{d/2} \mathbb{E} G_{2u_{a-1}} \right| \\
&\quad + \sup_{1\le a\le A} \left(\left(u_a^{d/2} - u_{a-1}^{d/2} \right) \mathbb{E} G_{2u_{a-1}} \right).
\end{aligned}
$$

Bound the three terms in the right hand side of the last inequality separately.

Let $\{u_a\}_{a\ge 0}$ be a geometric sequence, namely,

$$
A := \inf\{a \in \mathbb{N} : r^a \ge n^{2/d}\},
$$

and

$$
u_a = \begin{cases} r^a, & \forall\, 0 \le a \le A-1 \\ n^{2/d}, & a = A \end{cases},
$$

with $r > 1$ to be determined later.

Since $\lim_{\nu\to\infty} \nu^{d/2} \mathbb{E} G_{2\nu_n} = (\pi/2)^{d/2} \|p_0\|^2$ and $\nu^{d/2} \mathbb{E} G_{2\nu}$ is continuous, we obtain that for any $\varepsilon > 0$, there exsits sufficiently small $r > 1$ such that

$$
\sup_{1\le a\le A} \left| u_a^{d/2} \mathbb{E} G_{2u_a} - u_{a-1}^{d/2} \mathbb{E} G_{2u_{a-1}} \right| \le \varepsilon.
$$

At the same time, we can also ensure

$$
\sup_{1\le a\le A} \left(\left(u_a^{d/2} - u_{a-1}^{d/2} \right) \mathbb{E} G_{2u_{a-1}} \right) \le (r^{d/2} - 1) \left(\frac{\pi}{2} \right)^{d/2} \|p_0\|^2 \le \varepsilon
$$

by choosing r sufficiently small.

Finally consider

$$\sup_{1\le a\le A}\left(\frac{u_a}{u_{a-1}}\right)^{d/2}\cdot\sup_{0\le a\le A}u_a^{d/2}\left|\widehat{\mathbb{E}}_nG_{2u_a}-\mathbb{E}G_{2u_a}\right|.$$

On the one hand,

$$\sup_{1\le a\le A}\left(\frac{u_a}{u_{a-1}}\right)^{d/2}\le r^{d/2}.$$

On the other hand, since

$$\begin{aligned}\mathrm{var}\left(\widehat{\mathbb{E}}_nG_{2\nu_n}\right)&\lesssim\frac{1}{n}\mathbb{E}G_{2\nu_n}(X,X')G_{2\nu_n}(X,X'')+\frac{1}{n^2}\mathbb{E}G_{4\nu_n}(X,X')\\&\lesssim_d\frac{\nu_n^{-3d/4}\|p_0\|^3}{n}+\frac{\nu_n^{-d/2}\|p_0\|^2}{n^2}\end{aligned}$$

for any $\nu_n\in(0,\infty)$, we have

$$\begin{aligned}&P\left(\sup_{0\le a\le A}u_a^{d/2}\left|\widehat{\mathbb{E}}_nG_{2u_a}-\mathbb{E}G_{2u_a}\right|\ge\varepsilon\right)\\&\le\frac{\sum_{a=0}^{A}u_a^d\mathrm{var}\left(\widehat{\mathbb{E}}_nG_{2u_a}\right)}{\varepsilon^2}\lesssim_{d,r}\frac{1}{\varepsilon^2}\left(\frac{u_A^{d/4}\|p_0\|^3}{n}+\frac{u_A^{d/2}\|p_0\|^2}{n^2}\right)\to0\end{aligned}$$

as $n\to\infty$. Hence we conclude $\sup_{1\le\nu_n\le n^{2/d}}\left|\nu_n^{d/2}\widehat{\mathbb{E}}_nG_{2\nu_n}-\nu_n^{d/2}\mathbb{E}G_{2\nu_n}\right|=o_p(1)$.

Considering that

$$\lim_{\nu_n\to\infty}\nu_n^{d/2}\mathbb{E}G_{\nu_n}(X_1,X_2)G_{\nu_n}(X_1,X_3)=0,\quad\lim_{\nu_n\to\infty}\nu_n^{d/2}[\mathbb{E}G_{\nu_n}(X_1,X_2)]^2=0,$$

we obtain that (B.2) and (B.3) are also $o_p(1)$, based on almost the same arguments. Hence

$$\sup_{1\le\nu_n\le n^{2/d}}\left|\tilde{s}^2_{n,\nu_n}/\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2-1\right|=o_p(1).$$

On the other hand, since $\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2\gtrsim_{p_0,d}\nu_n^{-d/2}$ for $\nu_n\in[1,n^{2/d}]$,

$$\sup_{1\le\nu_n\le n^{2/d}}\frac{1}{n^2\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2}=o_p(1).$$

Hence we finally conclude that

$$\sup_{1\le \nu_n \le n^{2/d}} \left| \widehat{s}^2_{n,\nu_n} / \mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2 - 1 \right| = o_p(1).$$

□

B.3 Proof of Lemma 6

Let

$$K_{\nu_n}(x, x') = \frac{G_{\nu_n}(x, x')}{\sqrt{2\mathbb{E}G_{2\nu_n}(X_1, X_2)}}, \quad \forall\, x, x' \in \mathbb{R}^d,$$

and accordingly,

$$\bar{K}_{\nu_n}(x, x') = \frac{\bar{G}_{\nu_n}(x, x')}{\sqrt{2\mathbb{E}G_{2\nu_n}(X_1, X_2)}}.$$

Hence

$$\tilde{T}_n^{\mathrm{GOF(adapt)}} = \sup_{1\le \nu_n \le n^{2/d}} \left| \frac{1}{n-1} \sum_{i\neq j} \bar{K}_{\nu_n}(X_i, X_j) \cdot \sqrt{\frac{\mathbb{E}G_{2\nu_n}(X_1, X_2)}{\mathbb{E}[\bar{G}_{\nu_n}(X_1, X_2)]^2}} \right|.$$

To finish this proof, we first bound

$$\sup_{1\le \nu_n \le n^{2/d}} \left| \frac{1}{n-1} \sum_{i\neq j} \bar{K}_{\nu_n}(X_i, X_j) \right| \tag{B.4}$$

and then control $\tilde{T}_n^{\mathrm{GOF(adapt)}}$.

Step (i). There are two main tools that we borrow in this step. First, we apply results in Arcones and Gine (1993) to obtain a Bernstein-type inequality for

$$\left| \frac{1}{n-1} \sum_{i\neq j} \bar{K}_{\nu_0}(X_i, X_j) \right| \text{ and } \left| \frac{1}{n-1} \sum_{i\neq j} \left(\bar{K}_{\nu_n}(X_i, X_j) - \bar{K}_{\nu_n'}(X_i, X_j) \right) \right|$$

for some ν_0 and arbitrary $\nu_n, \nu_n' \in [1, \infty)$. And based on that, we borrow Talagrand's techniques on handling Bernstein-type inequality (*e.g.*, see Talagrand, 2014) to give a generic chaining bound of (B.4).

To be more specific, for any $\nu_0, \nu_n, \nu_n' \in [1, n^{2/d}]$, define

$$d_1(\nu_n, \nu_n') = \|\bar{K}_{\nu_n'} - \bar{K}_{\nu_n}\|_{L_\infty}, \quad d_2(\nu_n, \nu_n') = \|\bar{K}_{\nu_n'} - \bar{K}_{\nu_n}\|_{L_2}.$$

Then Proposition 2.3 (c) of Arcones and Gine (1993) ensures that for any $t > 0$,

$$P\left(\left|\frac{1}{n-1}\sum_{i\neq j}\bar{K}_{\nu_0}(X_i, X_j)\right| \geq t\right) \leq C\exp\left(-C\min\left\{\frac{t}{\|\bar{K}_{\nu_0}\|_{L_2}}, \left(\frac{\sqrt{n}t}{\|\bar{K}_{\nu_0}\|_{L_\infty}}\right)^{\frac{2}{3}}\right\}\right) \tag{B.5}$$

and

$$\begin{aligned} &P\left(\left|\frac{1}{n-1}\sum_{i\neq j}\left(\bar{K}_{\nu_n}(X_i, X_j) - \bar{K}_{\nu_n'}(X_i, X_j)\right)\right| \geq t\right) \\ \leq &C\exp\left(-C\min\left\{\frac{t}{d_2(\nu_n, \nu_n')}, \left(\frac{\sqrt{n}t}{d_1(\nu_n, \nu_n')}\right)^{\frac{2}{3}}\right\}\right) \end{aligned}$$

for some $C > 0$, and based on a chaining type argument see, *e.g.*, Theorem 2.2.28 in Talagrand, 2014 the latter inequality suggests there exists $C > 0$ such that

$$\begin{aligned} P\Bigg(\sup_{1\leq\nu_n\leq n^{2/d}}\left|\frac{1}{n-1}\sum_{i\neq j}\left(\bar{K}_{\nu_n}(X_i, X_j) - \bar{K}_{\nu_0}(X_i, X_j)\right)\right| \geq \\ C\left(\frac{\gamma_{2/3}([1, n^{2/d}], d_1)}{\sqrt{n}}t + \gamma_1([1, n^{2/d}], d_2) + D_2 t\right)\Bigg) \lesssim \exp(-t^{2/3}), \end{aligned} \tag{B.6}$$

where $\gamma_{2/3}([1, n^{2/d}], d_1)$, $\gamma_1([1, n^{2/d}], d_2)$ are the so-called γ-functionals and

$$D_2 = \sum_{l\geq 0} e_l([1, n^{2/d}], d_2)$$

with e_l being the so-called entropy numbers.

A straightforward combination of (B.5) and (B.6) then gives

$$P\Bigg(\sup_{1\le \nu_n \le n^{2/d}} \left| \frac{1}{n-1} \sum_{i\neq j} \bar{K}_{\nu_n}(X_i, X_j) \right| \geq C\left(\frac{\gamma_{2/3}([1, n^{2/d}], d_1)}{\sqrt{n}} t + \gamma_1([1, n^{2/d}], d_2) + D_2 t + \frac{\|\bar{K}_{\nu_0}\|_{L_\infty}}{\sqrt{n}} + \|\bar{K}_{\nu_0}\|_{L_2} t \right) \Bigg) \lesssim \exp(-t^{2/3}).$$

Therefore, given that the bounds on $\|\bar{K}_{\nu_0}\|_{L_2}$ and $\|\bar{K}_{\nu_0}\|_{L_\infty}$ can be obtained quite directly, *e.g.*, with $\nu_0 = 1$,

$$\|\bar{K}_{\nu_0}\|_{L_\infty} \le 4\|K_{\nu_0}\|_{L_\infty} = \frac{4}{\sqrt{2\mathbb{E}G_2}}, \qquad \|\bar{K}_{\nu_0}\|_{L_2} \le \|K_{\nu_0}\|_{L_2} = \frac{\sqrt{2}}{2},$$

the main focus is to bound $\gamma_{2/3}([1, n^{2/d}], d_1)$, $\gamma_1([1, n^{2/d}], d_2)$ and D_2 properly.

First consider $\gamma_{2/3}([1, n^{2/d}], d_1)$. Note that for any $1 \le \nu_n < \nu_n' < \infty$,

$$d_1(\nu_n, \nu_n') \le 4\|K_{\nu_n} - K_{\nu_n'}\|_{L_\infty} \le 4 \int_{\nu_n}^{\nu_n'} \left\| \frac{dK_u}{du} \right\|_{L_\infty} du$$

Since for any ν_n,

$$\begin{aligned} \frac{dK_{\nu_n}}{d\nu_n} =& (-\|x - x'\|^2) G_{\nu_n}(X_1, X_2) \left(\mathbb{E}G_{2\nu_n}(X_1, X_2) \right)^{-1/2} \\ & -\frac{1}{2} G_{\nu_n}(X_1, X_2) \left(\mathbb{E}G_{2\nu_n}(X_1, X_2) \right)^{-3/2} \frac{d}{d\nu_n} \mathbb{E}G_{2\nu_n}(X_1, X_2) \end{aligned}$$

where

$$\begin{aligned} \left(\mathbb{E}G_{2\nu_n}(X_1, X_2) \right)^{-1/2} &= \left(\frac{\pi}{2} \right)^{-d/4} \nu_n^{d/4} \left(\int \exp\left(-\frac{\|\omega\|^2}{8\nu_n} \right) \|\mathcal{F}p_0(\omega)\|^2 d\omega \right)^{-1/2} \\ &\lesssim_d \nu_n^{d/4} \left(\int \exp\left(-\frac{\|\omega\|^2}{8} \right) \|\mathcal{F}p_0(\omega)\|^2 d\omega \right)^{-1/2}, \end{aligned}$$

$$\left(\mathbb{E}G_{2\nu_n}(X_1, X_2)\right)^{-3/2} \lesssim_d \nu_n^{3d/4} \left(\int \exp\left(-\frac{\|\omega\|^2}{8}\right) \|\mathcal{F}p_0(\omega)\|^2 d\omega\right)^{-3/2},$$

and

$$\begin{aligned}
&\frac{d}{d\nu_n}\mathbb{E}_{2\nu_n}(X_1, X_2) \\
=&\left(\frac{\pi}{2}\right)^{d/2} \nu_n^{-d/2-1} \left(-\frac{d}{2} \cdot \int \exp\left(-\frac{\|\omega\|^2}{8\nu_n}\right) \|\mathcal{F}p_0(\omega)\|^2 d\omega \right. \\
&\left. + \int \exp\left(-\frac{\|\omega\|^2}{8\nu_n}\right) \left(\frac{\|\omega\|^2}{8\nu_n}\right) \|\mathcal{F}p_0(\omega)\|^2 d\omega\right),
\end{aligned}$$

which together ensure

$$\left\|\frac{dK_{\nu_n}}{d\nu_n}\right\|_{L_\infty} \lesssim_{d,p_0} \nu_n^{d/4-1}.$$

Hence

$$d_1(\nu_n, \nu_n') \lesssim_{d,p_0} |\nu_n^{d/4} - (\nu_n')^{d/4}|,$$

and $\gamma_{2/3}([1, n^{2/d}], d_1) \lesssim_{d,p_0} |(n^{2/d})^{d/4} - 1^{d/4}| \le \sqrt{n}$.

Then consider $\gamma_1([1, n^{2/d}], d_2)$. We have

$$d_2^2(\nu_n, \nu_n') \le \|K_{\nu_n'} - K_{\nu_n}\|_{L_2}^2 = 1 - \frac{\mathbb{E}G_{\nu_n}G_{\nu_n'}}{\sqrt{\mathbb{E}G_{2\nu_n}\mathbb{E}G_{2\nu_n'}}} \le -\log\left(\frac{\mathbb{E}G_{\nu_n}G_{\nu_n'}}{\sqrt{\mathbb{E}G_{2\nu_n}\mathbb{E}G_{2\nu_n'}}}\right)$$

Let $f_1(\nu_n) = \int \exp\left(-\frac{\|\omega\|^2}{8\nu_n}\right) \|\mathcal{F}p_0(\omega)\|^2 d\omega$. Then

$$\log\left(\mathbb{E}G_{2\nu_n}\right) = \frac{d}{2}\log\left(\frac{\pi}{2\nu_n}\right) + \log f_1(\nu_n)$$

and hence

$$\begin{aligned}
&-\log\left(\frac{\mathbb{E}G_{\nu_n}G_{\nu_n'}}{\sqrt{\mathbb{E}G_{2\nu_n}\mathbb{E}G_{2\nu_n'}}}\right) \\
=&\frac{d}{2}\left(-\frac{\log\nu_n + \log\nu_n'}{2} + \log\left(\frac{\nu_n + \nu_n'}{2}\right)\right) + \left(\frac{\log f_1(\nu_n) + \log f_1(\nu_n')}{2} - \log f_1\left(\frac{\nu_n + \nu_n'}{2}\right)\right).
\end{aligned}$$

Note that

$$\frac{\log f_1(\nu_n) + \log f_1(\nu_n')}{2} - \log f_1\left(\frac{\nu_n + \nu_n'}{2}\right) = \frac{1}{2}\int_0^{\frac{\nu_n' - \nu_n}{2}} \int_{-u}^{u} \left(\log f_1\left(\frac{\nu_n' + \nu_n}{2} + v\right)\right)'' dv du.$$

For any $\nu_n \geq 1$,

$$(\log f_1(\nu_n))'' = \frac{f_1(\nu_n) f_1''(\nu_n) - (f_1'(\nu_n))^2}{f_1^2(\nu_n)} \leq \frac{f_1''(\nu_n)}{f_1(\nu_n)},$$

and

$$f_1''(\nu_n) = \int \exp\left(-\frac{\|\omega\|^2}{8\nu_n}\right)\left(\frac{\|\omega\|^4}{64\nu_n^4} - \frac{\|\omega\|^2}{4\nu_n^3}\right)\|\mathcal{F}p_0(\omega)\|^2 d\omega \lesssim \nu_n^{-2}\|p_0\|_{L_2}^2.$$

Moreover, there exists $\nu_n^* = \nu_n^*(p_0) > 1$ such that $f_1(\nu_n^*) \geq \|p_0\|_{L_2}^2/2$, from which we obtain

$$(\log f_1(\nu_n))'' \lesssim \begin{cases} \nu_n^{-2}\|p_0\|_{L_2}^2/f_1(1), & 1 \leq \nu_n \leq \nu_n^* \\ \nu_n^{-2}, & \nu_n^* < \nu_n \leq n^{2/d} \end{cases},$$

which suggests that for any $\nu_n, \nu_n' \in [1, \nu_n^*]$

$$\begin{aligned} d_2^2(\nu_n, \nu_n') &\lesssim \left(\frac{d}{2} + \frac{\|p_0\|_{L_2}^2}{f_1(1)}\right)\left(-\frac{\log \nu_n + \log \nu_n'}{2} + \log\left(\frac{\nu_n + \nu_n'}{2}\right)\right) \\ &\lesssim \left(\frac{d}{2} + \frac{\|p_0\|_{L_2}^2}{f_1(1)}\right)|\log \nu_n - \log \nu_n'|, \end{aligned}$$

and for any $\nu_n, \nu_n' \in [\nu_n^*, n^{2/d}]$

$$d_2^2(\nu_n, \nu_n') \lesssim \left(\frac{d}{2} + 1\right)|\log \nu_n - \log \nu_n'|.$$

Note that in addition to the bound on d_2 obtained above, we also have

$$d_2(\nu_n, \nu_n') \leq \|\bar{K}_{\nu_n}\|_{L_2} + \|\bar{K}_{\nu_n'}\|_{L_2} \leq \|K_{\nu_n}\|_{L_2} + \|K_{\nu_n'}\|_{L_2} \leq \sqrt{2}.$$

Therefore,

$$\begin{aligned}
\gamma_1([1,n^{2/d}],d_2) &\le \sum_{l\ge 0} 2^l e_l([1,n^{2/d}],d_2)\\
&\lesssim e_0([1,n^{2/d}],d_2) + \sum_{l\ge 0} 2^l e_l([1,\nu_n^*],d_2) + \sum_{l\ge 0} 2^l e_l([\nu_n^*,n^{2/d}],d_2)\\
&\lesssim 1 + \sqrt{\frac{d}{2}+\frac{\|p_0\|_{L_2}^2}{f_1(1)}} \sum_{l\ge 0} 2^l \sqrt{\frac{\log \nu_n^* - \log 1}{2^{2^l}}}\\
&\quad + \sqrt{\frac{d}{2}+1}\left(\sum_{l\ge 0} 2^l \min\left\{1, \sqrt{\frac{\log n^{2/d} - \log \nu_n^*}{2^{2^l}}}\right\}\right)\\
&\lesssim 1 + \sqrt{\frac{d}{2}+\frac{\|p_0\|_{L_2}^2}{f_1(1)}}\sqrt{\log \nu_n^*} + \sqrt{\frac{d}{2}+1}\left(\sum_{l\ge 0} 2^l \min\left\{1, \sqrt{\frac{\log n^{2/d}}{2^{2^l}}}\right\}\right)\\
&\lesssim 1 + \sqrt{\frac{d}{2}+\frac{\|p_0\|_{L_2}^2}{f_1(1)}}\sqrt{\log \nu_n^*} + \sqrt{\frac{d}{2}+1}\left(\sum_{0\le l<l^*} 2^l + \sum_{l\ge l^*} 2^l \sqrt{\frac{\log n^{2/d}}{2^{2^l}}}\right)\\
&\lesssim 1 + \sqrt{\frac{d}{2}+\frac{\|p_0\|_{L_2}^2}{f_1(1)}}\sqrt{\log \nu_n^*} + \sqrt{\frac{d}{2}+1}\cdot 2^{l^*}
\end{aligned}$$

where l^* is the smallest l such that

$$\sqrt{\frac{\log n^{2/d}}{2^{2^l}}} \le 1.$$

Hence $2^{l^*} \asymp \log\log n$ and there exists $C = C(d) > 0$ such that

$$\gamma_1([1,n^{2/d}],d_2) \le C(d)\log\log n$$

for sufficiently large n.

By the similar approach, we get that

$$D_2 \lesssim 1 + \sqrt{\frac{d}{2}+\frac{\|p_0\|_{L_2}^2}{f_1(1)}}\sqrt{\log \nu_n^*} + \sqrt{\frac{d}{2}+1}\cdot l^*$$

which is upper-bounded by $C(d)\log\log\log n$ for sufficiently large n.

Therefore, we finally obtain that there exists $C(d) > 0$ such that for sufficiently large n,

$$P\left(\sup_{1\le\nu_n\le n^{2/d}}\left|\frac{1}{n-1}\sum_{i\neq j}\bar{K}_{\nu_n}(X_i,X_j)\right| \ge C(d)(\log\log n + t\log\log\log n)\right) \lesssim \exp(-t^{2/3}). \quad \text{(B.7)}$$

Step (ii). By slight abuse of notation, there exists $\nu_n^* = \nu_n^*(p_0) > 1$ such that

$$\frac{\mathbb{E}G_{2\nu_n}(X_1,X_2)}{\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2} \le 2$$

for $\nu_n \ge \nu_n^*$. Therefore,

$$\begin{aligned}
\tilde{T}_n^{\text{GOF(adapt)}} \le & \sup_{1\le\nu_n\le\nu_n^*}\sqrt{\frac{\mathbb{E}G_{2\nu_n}(X_1,X_2)}{\mathbb{E}[\bar{G}_{\nu_n}(X_1,X_2)]^2}}\cdot\sup_{1\le\nu_n\le\nu_n^*}\left|\frac{1}{n-1}\sum_{i\neq j}\bar{K}_{\nu_n}(X_i,X_j)\right| + \\
& \sqrt{2}\sup_{\nu_n^*\le\nu_n\le n^{2/d}}\left|\frac{1}{n-1}\sum_{i\neq j}\bar{K}_{\nu_n}(X_i,X_j)\right| \\
\le & C(p_0)\sup_{1\le\nu_n\le\nu_n^*}\left|\frac{1}{n-1}\sum_{i\neq j}\bar{K}_{\nu_n}(X_i,X_j)\right| + \\
& \sqrt{2}\sup_{\nu_n^*\le\nu_n\le n^{2/d}}\left|\frac{1}{n-1}\sum_{i\neq j}\bar{K}_{\nu_n}(X_i,X_j)\right|
\end{aligned}$$

for some $C(p_0) > 0$.

Based on arguments similar to those in the first step,

$$P\left(\sup_{1\le\nu_n\le\nu_n^*}\left|\frac{1}{n-1}\sum_{i\neq j}\bar{K}_{\nu_n}(X_i,X_j)\right| \ge C(d,p_0)t\right) \lesssim \exp(-t^{2/3})$$

for some $C(d,p_0) > 0$ and (B.7) still holds when ν_n is restricted to $[\nu_n^*, n^{2/d}]$. They together prove Lemma 6. □

B.4 Decomposition of dHSIC and Its Variance Estimation

In this section, we first derive an approximation of $\widehat{\gamma_\nu^2}(\mathbb{P}, \mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^k})$ under H_0 for general k, and then the approximation of $\text{var}\left(\widehat{\gamma_\nu^2}(\mathbb{P}, \mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^k})\right)$ can be obtained subsequently.

Note that

$$
\begin{aligned}
&G_\nu(x,y)\\
=&\int G_\nu(u,v)d(\delta_x-\mathbb{P}+\mathbb{P})(u)d(\delta_y-\mathbb{P}+\mathbb{P})(v)\\
=&\bar{G}_\nu(x,y)+(\mathbb{E}G_\nu(x,X)-\mathbb{E}G_\nu(X,X'))+(\mathbb{E}G_\nu(y,X)-\mathbb{E}G_\nu(X,X'))+\mathbb{E}G_\nu(X,X').
\end{aligned}
$$

Similarly write

$$
\begin{aligned}
&G_\nu(x,(y^1,\cdots,y^k))\\
=&\int G_\nu(u,(v^1,\cdots,v^k))d(\delta_x-\mathbb{P}+\mathbb{P})d(\delta_{y^1}-\mathbb{P}^{X^1}+\mathbb{P}^{X^1})\cdots d(\delta_{y^k}-\mathbb{P}^{X^k}+\mathbb{P}^{X^k})
\end{aligned}
$$

and expand it as the summation of all l-variate centered components where $l \le k+1$. Do the same expansion to $G_\nu((x^1,\cdots,x^k),(y^1,\cdots,y^k))$ and write it as the summation of all l-variate centered components where $l \le 2k$. Plug these expansions in $\widehat{\gamma_\nu^2}(\mathbb{P},\mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^k})$ and denote the summation of all l-variate centered components in such expression of $\widehat{\gamma_\nu^2}(\mathbb{P},\mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^k})$ by $D_l(\nu)$ for $l \le 2k$. Let the remainder $R_n = \sum_{l=3}^{2k} D_l(\nu)$ so that

$$
\widehat{\gamma_\nu^2}(\mathbb{P},\mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^k}) = \gamma_\nu^2(\mathbb{P},\mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^k}) + D_1(\nu) + D_2(\nu) + R_n.
$$

Straightforward calculation yields the following facts:

- $\mathbb{E}(R_n)^2 \lesssim_k n^{-3}\left(\mathbb{E}G_{2\nu}(X_1,X_2)+\prod_{l=1}^{k}\mathbb{E}G_{2\nu}(X_1^l,X_2^l)\right)$;
- under the null hypothesis, $D_1(\nu)=0$ and

$$
D_2(\nu) = \frac{1}{n(n-1)}\sum_{1\le i\ne j\le n} G_\nu^*(X_i,X_j)
$$

where

$$G_\nu^*(x,y) = \bar{G}_\nu(x,y) - \sum_{1\le j\le k} g_j(x^j, y) - \sum_{1\le j\le k} g_j(y^j, x) + \sum_{1\le j_1,j_2\le k} g_{j_1,j_2}(x^{j_1}, y^{j_2}).$$

Proof of Lemma 1. Observe that under H_0,

$$\mathrm{var}\left(\widehat{\gamma_\nu^2}(\mathbb{P}, \mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^k})\right) = \mathbb{E}(D_2(\nu))^2 + \mathbb{E}\,(R_n)^2 = \frac{2}{n(n-1)}\mathbb{E}[G_\nu^*(X_1,X_2)]^2 + \mathbb{E}\,(R_n)^2,$$

$$\mathbb{E}\,(R_n)^2 \lesssim_k n^{-3}\mathbb{E}G_{2\nu}(X_1,X_2),$$

and

$$\begin{aligned}
&\mathbb{E}[G_\nu^*(X_1,X_2)]^2\\
=&\mathbb{E}\left(\bar{G}_\nu(X_1,X_2) - \sum_{1\le j\le k} g_j(X_1^j, X_2)\right)^2\\
&-\mathbb{E}\left(\sum_{1\le j\le k} g_j(X_2^j, X_1) + \sum_{1\le j_1,j_2\le k} g_{j_1,j_2}(X_1^{j_1}, X_2^{j_2})\right)^2\\
=&\mathbb{E}\bar{G}_\nu^2(X_1,X_2) - 2\sum_{1\le j\le k}\mathbb{E}\Big(g_j(X_1^j,X_2)\Big)^2 + \sum_{1\le j_1,j_2\le k}\mathbb{E}\Big(g_{j_1,j_2}(X_1^{j_1},X_2^{j_2})\Big)^2.
\end{aligned}$$

They together conclude the proof. □

Below we shall further expand $\mathbb{E}\bar{G}_\nu^2(X_1,X_2)$, $\mathbb{E}\Big(g_j(X_1^j,X_2)\Big)^2$ and $\mathbb{E}\Big(g_{j_1,j_2}(X_1^{j_1},X_2^{j_2})\Big)^2$ in Lemma 1, based on which consistent estimator of $\mathrm{var}\left(\widehat{\gamma_\nu^2}(\mathbb{P}, \mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^k})\right)$ can be derived naturally.

First,

$$\begin{aligned}
&\mathbb{E}\bar{G}_\nu^2(X_1,X_2)\\
=&\mathbb{E}G_{2\nu}(X_1,X_2) - 2\mathbb{E}G_\nu(X_1,X_2)G_\nu(X_1,X_3) + (\mathbb{E}G_\nu(X_1,X_2))^2\\
=&\prod_{1\le l\le k}\mathbb{E}G_{2\nu}(X_1^l,X_2^l) - 2\prod_{1\le l\le k}\mathbb{E}G_\nu(X_1^l,X_2^l)G_\nu(X_1^l,X_3^l) + \prod_{1\le l\le k}\Big(\mathbb{E}G_\nu(X_1^l,X_2^l)\Big)^2.
\end{aligned}$$

Second,

$$\begin{aligned}
&\mathbb{E}\Big(g_j(X_1^j, X_2)\Big)^2 \\
=&\mathbb{E}G_{2\nu}(X_1^j, X_2^j) \cdot \prod_{l\neq j} \mathbb{E}G_\nu(X_1^l, X_2^l)G_\nu(X_1^l, X_3^l) - \prod_{1\leq l\leq k} \mathbb{E}G_\nu(X_1^l, X_2^l)G_\nu(X_1^l, X_3^l) \\
&- \mathbb{E}G_\nu(X_1^j, X_2^j)G_\nu(X_1^j, X_3^j) \cdot \prod_{l\neq j} (\mathbb{E}G_\nu(X_1^l, X_2^l))^2 + \prod_{1\leq l\leq k} \Big(\mathbb{E}G_\nu(X_1^l, X_2^l)\Big)^2.
\end{aligned}$$

Hence

$$\begin{aligned}
&\sum_{1\leq j\leq k} \mathbb{E}\Big(g_j(X_1^j, X_2)\Big)^2 \\
=&\left(\prod_{1\leq l\leq k} \mathbb{E}G_\nu(X_1^l, X_2^l)G_\nu(X_1^l, X_3^l)\right)\left(\sum_{1\leq j\leq k} \frac{\mathbb{E}G_{2\nu}(X_1^j, X_2^j)}{\mathbb{E}G_\nu(X_1^j, X_2^j)G_\nu(X_1^j, X_3^j)} - k\right) \\
&- \left(\prod_{1\leq l\leq k} \Big(\mathbb{E}G_\nu(X_1^l, X_2^l)\Big)^2\right)\left(\sum_{1\leq j\leq k} \frac{\mathbb{E}G_\nu(X_1^j, X_2^j)G_\nu(X_1^j, X_3^j)}{(\mathbb{E}G_\nu(X_1^j, X_2^j))^2} - k\right).
\end{aligned}$$

Finally,

$$\begin{aligned}
&\mathbb{E}\Big(g_{j_1,j_2}(X_1^{j_1}, X_2^{j_2})\Big)^2 \\
=&\begin{cases}
\mathbb{E}(\bar{G}_\nu(X_1^{j_1}, X_2^{j_1}))^2 \cdot \prod\limits_{l\neq j_1} \Big(\mathbb{E}G_\nu(X_1^l, X_2^l)\Big)^2, & j_1 = j_2 \\
\prod\limits_{l\in\{j_1,j_2\}} \Big(\mathbb{E}G_\nu(X_1^l, X_2^l)G_\nu(X_1^l, X_3^l) - (\mathbb{E}G_\nu(X_1^l, X_2^l))^2\Big) \prod\limits_{l\neq j_1,j_2} \Big(\mathbb{E}G_\nu(X_1^l, X_2^l)\Big)^2, & j_1 \neq j_2.
\end{cases}
\end{aligned}$$

Hence

$$\begin{aligned}
&\sum_{1\le j_1,j_2\le k}\mathbb{E}\Big(g_{j_1,j_2}(X_1^{j_1},X_2^{j_2})\Big)^2\\
&=\left(\prod_{1\le l\le k}\Big(\mathbb{E}G_\nu(X_1^l,X_2^l)\Big)^2\right)\left(\sum_{1\le j_1\le k}\frac{\mathbb{E}(\bar G_\nu(X_1^{j_1},X_2^{j_1}))^2}{(\mathbb{E}G_\nu(X_1^{j_1},X_2^{j_1}))^2}\right.\\
&\quad\left.+\sum_{1\le j_1\ne j_2\le k}\prod_{l\in\{j_1,j_2\}}\left(\frac{\mathbb{E}G_\nu(X_1^l,X_2^l)G_\nu(X_1^l,X_2^l)}{\Big(\mathbb{E}G_\nu(X_1^l,X_2^l)\Big)^2}-1\right)\right).
\end{aligned}$$

Then the consistent estimator $\tilde s^2_{n,\nu}$ of $\mathbb{E}\left(G^*_\nu(X_1,X_2)\right)^2$ is constructed by replacing

$$\mathbb{E}G_{2\nu}(X_1^l,X_2^l),\quad \mathbb{E}G_\nu(X_1^l,X_2^l)G_\nu(X_1^l,X_3^l),\quad (\mathbb{E}G_\nu(X_1^l,X_2^l))^2$$

in the above expansions of

$$\mathbb{E}\bar G^2_\nu(X_1,X_2),\quad \sum_{1\le j\le k}\mathbb{E}\Big(g_j(X_1^j,X_2)\Big)^2,\quad \sum_{1\le j_1,j_2\le k}\mathbb{E}\Big(g_{j_1,j_2}(X_1^{j_1},X_2^{j_2})\Big)^2$$

with the corresponding unbiased estimators

$$\begin{aligned}
&\frac{1}{n(n-1)}\sum_{1\le i\ne j\le n}G_{2\nu_n}(X_i^l,X_j^l),\quad \frac{(n-3)!}{n!}\sum_{\substack{1\le i,j_1,j_2\le n\\ |\{i,j_1,j_2\}|=3}}G_{\nu_n}(X_i^l,X_{j_1}^l)G_{\nu_n}(X_i^l,X_{j_2}^l)\\
&\qquad\frac{(n-4)!}{n!}\sum_{\substack{1\le i_1,i_2,j_1,j_2\le n\\ |\{i_1,i_2,j_1,j_2\}|=4}}G_{\nu_n}(X_{i_1}^l,X_{j_1}^l)G_{\nu_n}(X_{i_2}^l,X_{j_2}^l)
\end{aligned}$$

for $1\le l\le k$. Again, to avoid a negative estimate of the variance, we can replace $\tilde s^2_{n,\nu_n}$ with $1/n^2$ whenever it is negative or too small. Namely, let

$$\widehat{s}^2_{n,\nu_n}=\max\left\{\tilde s^2_{n,\nu_n},1/n^2\right\},$$

and estimate $\mathrm{var}\left(\widehat{\gamma^2_\nu}(\mathbb{P},\mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^k})\right)$ by $2\widehat{s}^2_{n,\nu}/(n(n-1))$.

Therefore for general k, the single kernel test statistic and the adaptive test statistic are constructed as

$$T^{\mathrm{IND}}_{n,\nu_n} = \frac{n}{\sqrt{2}} \widehat{s^{-1}_{n,\nu_n}} \widehat{\gamma^2_{\nu_n}}(\mathbb{P}, \mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^k}) \quad \text{and} \quad T^{\mathrm{IND(adapt)}}_{n} = \max_{1 \le \nu_n \le n^{2/d}} T^{\mathrm{IND}}_{n,\nu_n}$$

respectively. Accordingly, $\Phi^{\mathrm{IND}}_{n,\nu_n,\alpha}$ and $\Phi^{\mathrm{IND(adapt)}}$ can be constructed as in the case of $k = 2$.

B.5 Theoretical Properties of Independence Tests for General k

In this section, with $\Phi^{\mathrm{IND}}_{n,\nu_n,\alpha}$ and $\Phi^{\mathrm{IND(adapt)}}$ constructed in Appendix B.4 for general k, we confirm that Theorem 12, Theorem 13 and Theorem 16 still hold. We shall only emphasize the main differences between the new proofs and the original proofs in the case of $k = 2$.

Under the null hypothesis: we only need to re-ensure that $\tilde{s}^2_{n,\nu_n}$ is a consistent estimator of $\mathbb{E}[G^*_{\nu_n}(X_1, X_2)]^2$. Specifically, we show that

$$\tilde{s}^2_{n,\nu_n} / \mathbb{E}[G^*_{\nu_n}(X_1, X_2)]^2 \to_p 1$$

given $1 \ll \nu_n \ll n^{4/d}$ for Theorem 12 and

$$\sup_{1 \le \nu_n \le n^{2/d}} \left| \tilde{s}^2_{n,\nu_n} / \mathbb{E}[G^*_{\nu_n}(X_1, X_2)]^2 - 1 \right| = o_p(1)$$

for Theorem 16.

To prove the former one, since

$$\frac{\mathbb{E}[G^*_{\nu_n}(X_1, X_2)]^2}{(\pi/(2\nu_n))^{d/2} \|p\|^2_{L_2}} \to 1$$

as $\nu_n \to \infty$, it suffices to show

$$\nu_n^{d/2} \left| \tilde{s}^2_{n,\nu_n} - \mathbb{E}[G^*_{\nu_n}(X_1, X_2)]^2 \right| = o_p(1),$$

which follows considering that

$$\nu_n^{d_l/2}\mathbb{E}G_{2\nu_n}(X_1^l, X_2^l), \quad \nu_n^{d/2}\mathbb{E}G_{\nu_n}(X_1^l, X_2^l)G_{\nu_n}(X_1^l, X_3^l), \quad \nu_n^{d_l/2}(\mathbb{E}G_{\nu_n}(X_1^l, X_2^l))^2 \tag{B.8}$$

are all bounded and they are estimated consistently by their corresponding estimators. For example,

$$\nu_n^{d_l/2}\mathbb{E}G_{2\nu_n}(X_1^l, X_2^l) \to (\pi/2)^{d_l/2}\,\|p_l\|_{L_2}^2$$

and

$$\begin{aligned}
&\nu_n^{d_l}\mathbb{E}\left(\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n} G_{2\nu_n}(X_i^l, X_j^l) - \mathbb{E}G_{2\nu_n}(X_1^l, X_2^l)\right)^2 \\
=\ &\nu_n^{d_l}\mathrm{var}\left(\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n} G_{2\nu_n}(X_i^l, X_j^l)\right) \\
\lesssim\ &\nu_n^{d_l}\left(n^{-1}\mathbb{E}G_{2\nu_n}(X_1^l, X_2^l)G_{2\nu_n}(X_1^l, X_3^l) + n^{-2}\mathbb{E}G_{4\nu_n}(X_1^l, X_2^l)\right) \\
\lesssim_{d_l}\ &n^{-1}\nu_n^{d_l/4}\|p_l\|_{L_2}^3 + n^{-2}\nu_n^{d_l/2}\|p_l\|_{L_2}^2 \to 0.
\end{aligned}$$

The proof of the latter one is similar. It sufficies to have

- each term in (B.8) is bounded for $\nu_n \in [1, \infty)$, which immediately follows since each term is continuous and converges at ∞;
- the difference between each term in (B.8) and its corresponding estimator converges to 0 uniformly over $\nu_n \in [1, n^{2/d}]$, the proof of which is the same with that of Lemma 5.

Under the alternative hypothesis: we only need to re-ensure that $\widehat{s}_{n,\nu_n}$ is bounded. Specifically, we show

$$\inf_{p\in H_1^{\mathrm{IND}}(\Delta_n, s)} \frac{n\gamma_{\nu_n}^2(\mathbb{P}, \mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^k})}{\left[\mathbb{E}\left(\widehat{s}_{n,\nu_n}^2\right)^{1/k}\right]^{k/2}} \to \infty$$

for Theorem 13 and

$$\inf_{s\geq d/4}\inf_{p\in H_1^{\mathrm{IND}}(\Delta_{n,s};s)} P\left(\widehat{s}^2_{n,\nu_n(s)'} \leq 2M^2(2\nu_n(s)'/\pi)^{-d/2}\right) \to 1 \tag{B.9}$$

for Theorem 16, where $\nu_n(s)' = (\log\log n/n)^{-4/(4s+d)}$.

The former one holds because

$$\begin{aligned}
\mathbb{E}\left(\widehat{s}^2_{n,\nu_n}\right)^{1/k} &\leq \mathbb{E}\left(\max\left\{\left|\tilde{s}^2_{n,\nu}\right|, 1/n^2\right\}\right)^{1/k} \\
&\leq \mathbb{E}\left|\tilde{s}^2_{n,\nu}\right|^{1/k} + n^{-2/k} \\
&\lesssim_k \left(\prod_{l=1}^{k}\mathbb{E}G_{2\nu_n}(X_1^l, X_2^l)\right)^{1/k} + n^{-2/k} \\
&\leq \left(M^2(\pi/(2\nu_n))^{d/2}\right)^{1/k} + n^{-2/k}.
\end{aligned}$$

where the second to last inequality follows from generalized Hölder's inequality. For example,

$$\mathbb{E}\left(\prod_{l=1}^{k}\frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n}G_{2\nu_n}(X_i^l, X_j^l)\right)^{1/k} \leq \left(\prod_{l=1}^{k}\mathbb{E}G_{2\nu_n}(X_1^l, X_2^l)\right)^{1/k}.$$

To prove the latter one, note that for $\nu_n = \nu_n(s)'$, all three terms in (B.8) are bounded by $M_l^2(\pi/2)^{d_l/2}$ and the variances of their corresponding estimators are bounded by

$$C(d_l)\left(n^{-1}\left(\nu_n(s)'\right)^{d_l/4}M_l^3 + n^{-2}\left(\nu_n(s)'\right)^{d_l/2}M_l^2\right) = o(1)$$

uniformly over all s. Therefore,

$$\inf_{s\geq d/4}\inf_{p\in H_1^{\mathrm{IND}}(\Delta_{n,s};s)} P\left(\left(\nu_n(s)'\right)^{d/2}\left|\tilde{s}^2_{n,\nu_n(s)'} - \mathbb{E}[G^*_{\nu_n(s)'}(Y_1,Y_2)]^2\right| \leq M^2(\pi/2)^{d/2}\right) \to 1$$

where $Y_1, Y_2 \sim_{\text{iid}} \mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^k}$. Further considering that

$$\mathbb{E}[G^*_{\nu_n(s)'}(Y_1, Y_2)]^2 \leq \mathbb{E}[\bar{G}_{\nu_n(s)'}(Y_1, Y_2)]^2 \leq M^2(\pi/(2\nu_n(s)'))^{d/2}$$

and that

$$1/n^2 = o((\nu_n(s)')^{-d/2})$$

uniformly over all s, we prove (B.9).