Journal of Machine Learning Research 25 (2024) 1-62 Submitted 10/20; Revised 8/24; Published 10/24

On the Optimality of Gaussian Kernel Based Nonparametric Tests against Smooth Alternatives

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

Editor: John Shawe-Taylor

Abstract

Nonparametric tests via kernel embedding of distributions have witnessed a great deal of practical successes in recent years. However, statistical properties of these tests are largely unknown beyond consistency against a fixed alternative. To fill in this void, we study here 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 also presented to further demonstrate the practical merits of the methodology.

Keywords: Gaussian kernel embedding, maximum mean discrepancy (MMD), nonparametric tests, diverging scaling parameter, minimax optimality, adaptation

1. Introduction

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 has been a renewed interest on them from both statistics and other related fields as they arise naturally in many modern applications where the performance of the 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.

1.1 Nonparametric Tests via Kernel Embedding

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 $x_1, \dots, x_n \in \mathcal{X}$. 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

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

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); Lyons (2013).

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 various statistical inferences. 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 $H_0^{\rm GOF} : \mathbb{P} = \mathbb{P}_0$ 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 $X_1, \cdots, X_n \sim_{\rm iid} \mathbb{P}$ and $Y_1, \cdots, Y_m \sim_{\rm iid} \mathbb{Q}$, we are interested in testing if the null hypothesis $H_0^{\rm HOM} : \mathbb{P} = \mathbb{Q}$ 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 of independence among a set of random variables is another classical problem in statistics. Let $X = (X^1, \dots, X^k)^\top \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$. If $X^1, \dots, X^k$ are jointly independent, then the distribution of X can be factorized:

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

Dependence among $X^1, \dots, X^k$ can be naturally measured by the discrepancy 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 $k = 2$, the squared discrepancy $\gamma^2(\mathbb{P}^X, \mathbb{P}^{X^1} \otimes \mathbb{P}^{X^2})$ 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 $\gamma(\widehat{\mathbb{P}}_n^X, \widehat{\mathbb{P}}_n^{X^1} \otimes \cdots \otimes \widehat{\mathbb{P}}_n^{X^k})$ exceeds 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.

In all these cases the squared 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_1^2 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,

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

where $\to_p$ stands for convergence in probability. 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.

For example, as recently shown by Balasubramanian et al. (2017), any goodness-of-fit test based on statistic $\gamma_K^2(\widehat{\mathbb{P}}_n, \mathbb{P}_0)$ with a *fixed* kernel K is necessarily suboptimal. Here, the subscript K signifies the choice of kernel. Balasubramanian et al. (2017) also argued that much more powerful tests can be constructed by *regularized embedding*. The appropriate regularization they employed, however, relies on the knowledge of $\mathbb{P}_0$, and therefore is specialized to goodness-of-fit tests. While it is plausible that MMD based tests for homogeneity or independence may suffer from similar deficiencies, it remains unclear how to construct tests that are more powerful in these settings. The goal of the current work is specifically to address this question. In particular, we show that embedding using a Gaussian kernel with an appropriately chosen scaling parameter provides a unified treatment to all three testing problems.

1.2 Our Contribution: Optimality and Adaptivity of Gaussian Kernel Embedding

When data are continuous, *e.g.*, $\mathcal{X} = \mathbb{R}^d$, Gaussian kernels are arguably the most popular and successful choice in practice. On the one hand, we show that this choice of kernel is justified because in all three scenarios, MMD based tests can be optimal for testing against smooth alternatives provided that an appropriate scaling parameter is elicited. On the other hand, we argue that existing ways of selecting the scaling parameter may not exploit the full potential of Gaussian kernel based approaches and yet more powerful tests can be constructed with appropriate choice of the scaling parameter.

In particular, we investigate how the power of these tests increases with the sample size by characterizing the asymptotic behavior of the smallest amount of departure from the null hypothesis that can be consistently detected. More specifically, we adopt the minimax hypothesis testing framework pioneered by Burnashev (1979); Ingster (1987, 1993). See also Ermakov (1991); Spokoiny (1996); Lepski and Spokoiny (1999); Ingster and Suslina (2000); Ingster (2000); Baraud (2002); Ingster and Suslina (2003); Fromont and Laurent (2006); Fromont et al. (2012, 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. In all three settings, goodness of fit, homogeneity and independence testing, we show that Gaussian kernels with an appropriately chosen scaling parameter yield tests that are rate optimal in detecting smooth departures from null hypotheses. It is worth pointing out that even though the goodness-of-fit and homogeneity tests have been considered within this framework before, it is always done under the assumption that the underlying distributions are compactly supported. The use of Gaussian kernel enables us to do away this restriction. Our results not only provide rigorous justifications to the practical successes of Gaussian kernels based testing procedures but also offer guidelines on how to choose the scaling parameter in a principled way.

The critical importance of selecting an appropriate scaling parameter is widely recognized in practice. 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); Sriperumbudur et al. (2009); Gretton et al. (2012b); Sutherland et al. (2017). Our result shows that it is essential that we take a diverging scaling parameter as the sample size increases, and the choice of the scaling parameter may determine against which types of deviation from the null hypothesis the resulting test is most powerful.

This also naturally brings about the issue of adaptation and whether or not there is an agnostic approach towards testing of the aforementioned null hypotheses without the need to specify a scaling parameter. To address this challenge, we introduce a simple testing procedure by maximizing a studentized MMD over a pre-specified range of scaling parameters. 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, simultaneously over the collection of parameter spaces corresponding to different levels of smoothness.

1.3 Relation to Earlier Work: A Tale of Two "Kernels"

A simple yet useful observation for our analysis is the close connection between MMD and another type of kernel method that is common in the literature on nonparametric statistics, namely, kernel density estimation (KDE). Through the lens of KDE, sample MMD can be viewed as an estimate of the L_2 distance between two smooth densities. A similar observation was first made by Gretton et al. (2012a) in the context of homogeneity tests. We argue that MMD based goodness-of-fit and independence test statistics can also be viewed as such, albeit slightly unconventional ones. This relationship allows us to blend insights and techniques from the two rich but largely separate strands of literature, which in turn leads to better understanding of the operating characteristics of Gaussian kernel based nonparametric tests.

We usually think of sample MMD as an estimate of MMD between two probability measures. Note that the Gaussian kernel is a characteristic kernel (Fukumizu et al., 2007). This means the sample MMD can be used to consistently differentiate between two fixed probability measures as sample size increases. However, with an increasing sample size, there is also the opportunity to differentiate between two probability measures that are closer to each other. While there are many benefits to quantify the "closeness" by MMD, it is nonetheless a rather weak distance metric. Consider for example two probability measures $\mathbb{P}$ and $\mathbb{Q}$ with densities p and q respectively. It is not hard to see that $\gamma(\mathbb{P}, \mathbb{Q})$ is always upper-bounded by the L_2 distance $\|p - q\|_{L_2}$ between p and q. This means that, when measured by $\gamma(\mathbb{P}, \mathbb{Q})$, $\mathbb{P}$ and $\mathbb{Q}$ may appear much closer to each other than they actually are. This also implies that a test based on the magnitude of an estimate of $\gamma(\mathbb{P}, \mathbb{Q})$, such as those described earlier, may not be as powerful as a test based on estimating a stronger distance measure between $\mathbb{P}$ and $\mathbb{Q}$. This insight was exploited earlier by Balasubramanian et al. (2017) in the context goodness-of-fit test where they construct a test based on the χ^2 distance between distributions. As mentioned earlier, their construction is specialized to goodness-of-fit test and requires evaluation of the eigenvalue decomposition of the kernel. Our development shares a similar spirit. But, by leveraging the property that sample MMD with an appropriately chosen kernel can estimate $\|p - q\|_{L_2}$ well, our approach is much simpler and more broadly applicable: using Gaussian kernel with a diverging scaling parameter.

The problem of estimating the L_2 distance between two densities is closely related to estimating $\|p\|_{L_2}^2$ given a sample from the density p, which has been well studied in the literature since the pioneering work of Bickel and Ritov (1988) who first showed that such functionals can be estimated at the parametric rate for smooth functions. However, almost none of the existing work employs KDE based methods. The lone exception and the work most related to our treatment is Giné and Nickl (2008) who showed that when $d = 1$, KDE based methods can also attain parametric rate. Similar to Giné and Nickl (2008), our development is based on a combination of U-statistic theory and Fourier analytical methods. However, there is also a crucial distinction: our goal is testing not estimation. This difference manifests prominently in our power analysis of MMD based tests. As first observed by Ingster (1987), optimal testing is often more subtle and requires more careful analysis of the behavior of higher order terms. Indeed this is also the case in our setting and as a result we show that it is possible to consistently differentiate between two probability measures even in situations where their difference cannot be consistently estimated.

The marriage between these two perspectives also leads to intriguing new findings. In particular, we show that, with Gaussian kernel, adaptativity can be attained by simply maximizing studentized sample MMD. This is to be contrasted with the more sophisticated procedure known as Lepski's method (Lepskii, 1991; Lepski and Spokoiny, 1997) that is typically used in nonparametric statistics. Our work here offers a partial explanation of the success of Gaussian kernel in practice.

1.4 Organization of the Paper

The rest of this paper is organized as follows. In the next three sections, we shall investigate the statistical properties of Gaussian kernel based tests for goodness-of-fit, homogeneity and independence respectively, and show that with appropriate choice of the scaling parameter, these tests are minimax optimal if the underlying densities are smooth. Since the optimal choice of scaling parameter requires the knowledge of smoothness which is rarely available, in Section 5 we introduce new tests that do not require such knowledge yet attain optimal power, up to an iterated logarithmic factor, for a wide range of smooth alternatives. Numerical experiments presented in Section 6 further illustrate the practical merits of our method and theoretical developments. We conclude with some summary discussion in Section 7 and all proofs are relegated to Section 8.

2. Test for Goodness-of-fit

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 io 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 suppress 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

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

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^{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:

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

where

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

and

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

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

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

Of particular interest here is the smallest Δ_n so that a test is consistent in that the above quantity converges to one.

Consider embedding with Gaussian kernel and a fixed scaling parameter $\nu > 0$. Following standard asymptotic theory for V-statistics (see, *e.g.*, Serfling, 2009), it can be shown that under H_0^{GOF} and certain regularity conditions,

$$n\gamma_\nu^2(\widehat{\mathbb{P}}, \mathbb{P}_0) \to_d \sum_{k \geq 1} \lambda_k^2 Z_k^2$$

where $\to_d$ stands for convergence in distribution and $\lambda_1 \geq \lambda_2 \geq \cdots$ are the singular values of the linear operator:

$$\mathcal{L}_\nu f = \int_{\mathbb{R}^d} \bar{G}_\nu(x, x'; \mathbb{P}_0) f(x') dx', \qquad \forall f \in L_2(\mathbb{R}^d)$$

and

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

and Z_ks are independent standard normal random variables. Hereafter, for brevity, we shall omit the last argument of $\bar{G}$ when it is clear from the context. As such, we may proceed to reject H_0^{GOF} if and only if $n\widehat{\gamma}_\nu^2(\widehat{\mathbb{P}}_n, \mathbb{P}_0)$ exceeds the upper α quantile of its asymptotic distribution, which yields an (asymptotic) α-level test. Following the same argument as that from Balasubramanian et al. (2017), we can show that under mild regularity conditions such a test has power tending to one if and only if $\Delta_n \gg n^{-1/4}$. In addition, as shown by Balasubramanian et al. (2017), much more powerful tests exist when assuming that the underlying densities are compactly supported and bounded away from 0 and 1. Here we show that the same is true for broader classes of distributions using Gaussian kernel embedding with a diverging scaling parameter.

Recall that

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

It is not hard to see that this is a biased estimate of $\gamma_\nu^2(\mathbb{P}, \mathbb{P}_0)$ due to the oversized influence of the summands when $i = j$. It is often common to correct for bias and use instead the following U-statistic:

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

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 a data-driven choice of ν is widely recognized in practice, almost all existing theoretical studies assume a fixed kernel, therefore those studies are valid for a fixed scaling parameter in the case of the Gaussian kernel. 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 bandwidth h and with the ith observation removed:

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

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

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

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

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

as an estimate of $\int(p-p_0)^2$. It is worth pointing out that, in spite the connection with kernel density estimation (KDE, for short), (1) differs from the usual and well studied KDE based estimate of $\int(p-p_0)^2$, namely, $\int(\widehat{p}_h-p_0)^2$.

Nonetheless, this close relationship between the two types of kernel methods, one popular in machine learning literature and the other common in statistics, allows us to draw insights from the well established theory for kernel density estimator (see, *e.g.*, Tsybakov, 2008) in studying goodness-of-fit tests based on kernel embedding. In particular, following standard asymptotic properties of the kernel density estimator, 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 $H_0^{\rm GOF}$ 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 $H_0^{\rm GOF}$, it is clear that $\mathbb{E}\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{P}_0)=0$. Note also that

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

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 in (2) by U-statistics:

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

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 the following result.

Theorem 1 *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},*

$$\frac{n}{\sqrt{2}}\left(\frac{2\nu_n}{\pi}\right)^{d/4}\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{P}_0) \to_d N(0, \|p_0\|^2_{L_2}). \tag{3}$$

Moreover,

$$T^{\mathrm{GOF}}_{n,\nu_n} \to_d N(0,1). \tag{4}$$

Theorem 1 immediately implies a test, denoted by $\Phi^{\mathrm{GOF}}_{n,\nu_n,\alpha}$ (with $\alpha \in (0,1)$), that rejects H_0^{GOF} if and only if $T^{\mathrm{GOF}}_{n,\nu_n}$ exceeds z_α, the upper α 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 the following result.

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

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

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 3 *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* $X_1, \dots, X_n \sim p$*,*

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

The lower bound given by Theorem 3 is similar in spirit to the classical result by Ingster (1987) who considers the case when both $\mathbb{P}$ and $\mathbb{P}_0$ are compactly supported. Together, Theorems 2 and 3 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 from the null distribution. The idea can also be readily applied to testing of homogeneity and independence which we shall examine next.

3. 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^{HOM} : $p = q \in \mathcal{W}^{s,2}(M)$, and the alternative hypothesis as

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

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

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

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 (for example, of Δ) on n and not m, for brevity. Our treatment, however, can be straightforwardly extended to more general situations.

We shall focus on an unbiased estimate of $\gamma^2_{\nu_n}(\mathbb{P}, \mathbb{Q})$, namely,

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

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

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

and

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

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

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

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

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

As before, we can show

Theorem 4 *Let $\nu_n \to \infty$ as $n \to \infty$ in such a fashion that $\nu_n = o(n^{4/d})$. Then under $H^{\text{HOM}}_0 : p = q \in \mathcal{W}^{s,2}(M)$,*

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

Note that the condition $c \le m/n \le C$ implies that $m \to \infty$ when $n \to \infty$. Motivated by Theorem 4, we can consider a test, denoted by $\Phi^{\text{HOM}}_{n,\nu_n,\alpha}$, that rejects H^{HOM}_0 if and only if T^{HOM}_{n,ν_n} exceeds z_α. By construction, $\Phi^{\text{HOM}}_{n,\nu_n,\alpha}$ is an asymptotic α level test. We now turn to study its power against H^{HOM}_1. As in the case of goodness of fit test, we can prove that $\Phi^{\text{HOM}}_{n,\nu_n,\alpha}$ 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 5 *(i) Assume that $n^{2s/(d+4s)}\Delta_n \to \infty$. Then for any $\alpha \in (0,1)$,*

$$\lim_{n\to\infty} \text{power}\{\Phi^{\text{HOM}}_{n,\nu_n,\alpha}; H^{\text{HOM}}_1(\Delta_n; s)\} = 1,$$

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 $Y_1, \ldots, Y_m \sim q$,

$$\liminf_{n\to\infty} \text{power}\{\Phi_n; H^{\text{HOM}}_1(\Delta_n; s)\} < 1.$$

Similar to the setting for goodness-of-fit test, Theorem 5 suggests that Gaussian kernel embedding of distributions with appropriate choice of the scaling parameter is also minimax rate optimal for testing of homogeneity. Our result, again, differs from previous studies that often require that the $\mathbb{P}$ and $\mathbb{Q}$ are compactly supported.

4. Test for Independence

Similarly, we can also use Gaussian kernel embedding to construct minimax optimal tests of independence. Let $X = (X^1, \dots, X^k)^\top \in \mathbb{R}^d$ be a random vector where the subvectors $X^j \in \mathbb{R}^{d_j}$ for $j = 1, \dots, k$ 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

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

against a smooth departure from independence:

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

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 the null and the alternative hypotheses.

Given a sample $\{X_1, \dots, X_n\}$ of independent copies of X, we shall consider the following unbiased estimate of $\gamma^2_{\nu_n}(\mathbb{P}, \mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^k})$,

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

Under $H_0^{\rm IND}$, 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

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

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, we write

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

and

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

Then we have

Lemma 6 *Under $H_0^{\rm IND}$,*

$$\begin{aligned}\operatorname{var}\left(\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^k})\right) =& \frac{2}{n(n-1)}\Bigg(\mathbb{E}\bar{G}^2_{\nu_n}(X,Y) - 2\sum_{1\le j\le k}\mathbb{E}\left(g_j(X^j,Y)\right)^2 \\ & + \sum_{1\le j_1,j_2\le k}\mathbb{E}\left(g_{j_1,j_2}(X^{j_1},Y^{j_2})\right)^2\Bigg) + O(\mathbb{E}G_{2\nu_n}(X,Y)/n^3).\end{aligned} \tag{5}$$

In light of Lemma 6, a variance estimator can be derived by estimating the leading term on the righthand side of (5) 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 (5) 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 $X^j, Y^j \sim_{\rm iid} \mathbb{P}^{X^j}$ for $j=1,2$. Thus, we can estimate $\mathbb{E}[\bar{G}_{\nu_n}(X^j,Y^j)]^2$ by

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

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

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

Then, a test statistic for $H_0^{\rm IND}$ is

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

Test statistics for general $k>2$ can be defined similarly.

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

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

As before, let $\Phi^{\rm IND}_{n,\nu_n,\alpha}$ be the test that rejects $H_0^{\rm IND}$ if and only if $T^{\rm IND}_{n,\nu_n}$ exceeds z_α. We have the following.

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

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

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 $X_1, \dots, X_n \sim p$,

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

As before, Theorem 8 shows that $\Phi^{\rm IND}_{n,\nu_n,\alpha}$ is also minimax optimal.

5. 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.

5.1 Test for Goodness-of-fit

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

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

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. Thus far, we do not know if adaptation can extend beyond $d/4$ and we shall leave this for future investigation.

More specifically, we shall proceed to reject $H_0^{\rm GOF}$ if and only if $T_n^{\rm GOF(adapt)}$ exceeds the upper α quantile, denoted by $q^{\rm GOF}_{n,\alpha}$, of its null distribution. In what follows, we shall call this test $\Phi^{\rm GOF(adapt)}$. Note that, even though it is hard to derive the analytic form for $q^{\rm GOF}_{n,\alpha}$, it can be readily evaluated via Monte Carlo method, *i.e.*, it can be approximated by the sample quantile of $T_n^{\rm GOF(adapt)}$ simulated under the null hypothesis. To study the power of $\Phi^{\rm GOF(adapt)}$ against $H_1^{\rm GOF}$ with different levels of smoothness, we shall consider the following alternative hypothesis

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

The following theorem characterizes the power of $\Phi^{\rm GOF(adapt)}$ against this alternative.

Theorem 9 *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

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

Theorem 9 shows that $\Phi^{\rm GOF(adapt)}$ 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 showed in Section 2, the optimal test is based on $T_{n,\nu_n}^{\rm GOF}$ 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 $\Phi^{\rm GOF(adapt)}$ is powerful against smooth alternatives for all $s \geq d/4$.

5.2 Test for Homogeneity

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

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

If $T_n^{\rm HOM(adapt)}$ exceeds the upper α quantile, denoted by $q_{n,\alpha}^{\rm HOM}$, of its null distribution, then we reject $H_0^{\rm HOM}$. In what follows, we shall refer to this test as $\Phi^{\rm HOM(adapt)}$. As before, we do not have a closed form expression for $q_{n,\alpha}^{\rm HOM}$, and it needs to be evaluated via Monte Carlo method. In particular, in the case of homogeneity test, we can approximate $q_{n,\alpha}^{\rm HOM}$ by permutation where we randomly shuffle $\{X_1, \dots, X_n, Y_1, \dots, Y_m\}$ 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 $q_{n,\alpha}^{\rm HOM}$.

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

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

Theorem 10 *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

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

Similar to the case of goodness-of-fit test, Theorem 10 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, 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 $s \geq d/4$.

5.3 Test for Independence

Similarly, for independence test, we 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 $\Phi^{\mathrm{IND(adapt)}}$. The critical value, $q_{n,\alpha}^{\mathrm{IND}}$, 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 the following result.

Theorem 11 *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 11 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 $s \geq d/4$.

6. Numerical Experiments

To complement our theoretical development and demonstrate the practical merits of the proposed methodology in choosing the scaling parameter, we conducted several sets of numerical experiments.

6.1 Effect of Scaling Parameter

Our first set of experiments was designed to illustrate the importance of the scaling parameter and highlight the potential room for improvement over the "median" heuristic—a 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 a 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 (± one standard error) for different choices of the scaling parameter. The results are summarized in Figure 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.

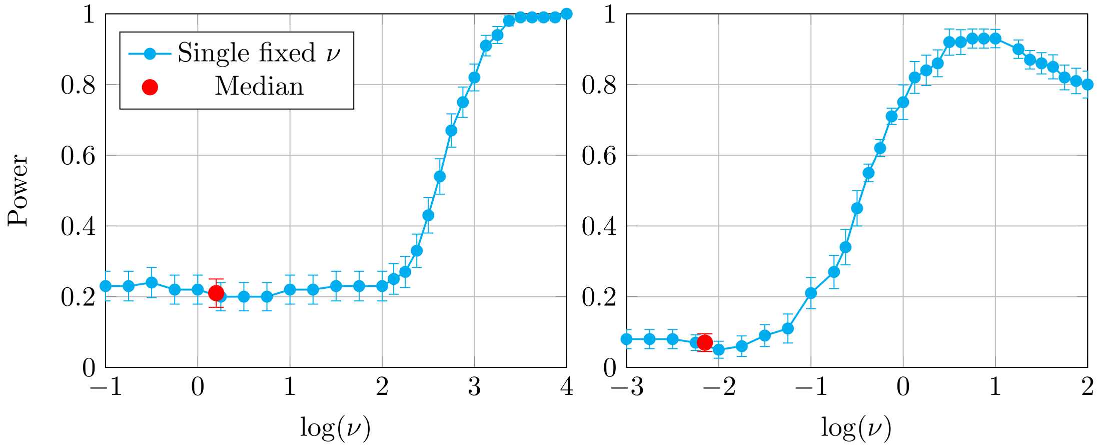


Figure 1: Observed power against log(ν) in Experiment I (left) and Experiment II (right).

The importance of the scaling parameter is evident from Figure 1 where 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.

6.2 Efficacy of Adaptation

Our second set of experiments aims to illustrate that the adaptive procedures we proposed in Section 5 indeed yield more powerful tests when compared with other procedures that are commonly used in practice. In particular, we compare the proposed self-normalized adaptive test (`S.A.`) with several data-driven approaches, namely the “median” heuristic (`Median`), the training-testing approach (`T.T.`) in Sutherland et al. (2017) and the unnormalized adaptive test (`U.A.`) proposed in Sriperumbudur et al. (2009). For `T.T.`, `U.A.` and `S.A.`, 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

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

  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

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

  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 $m = n \in \{25, 50, 75, \cdots, 200\}$ and dimension $d \in \{1, 10, 100, 1000\}$; for Experiment IV, the sample size were $n \in \{100, 200, \cdots, 600\}$ and dimension $d \in \{2, 10, 100, 1000\}$. 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 2 and 3.

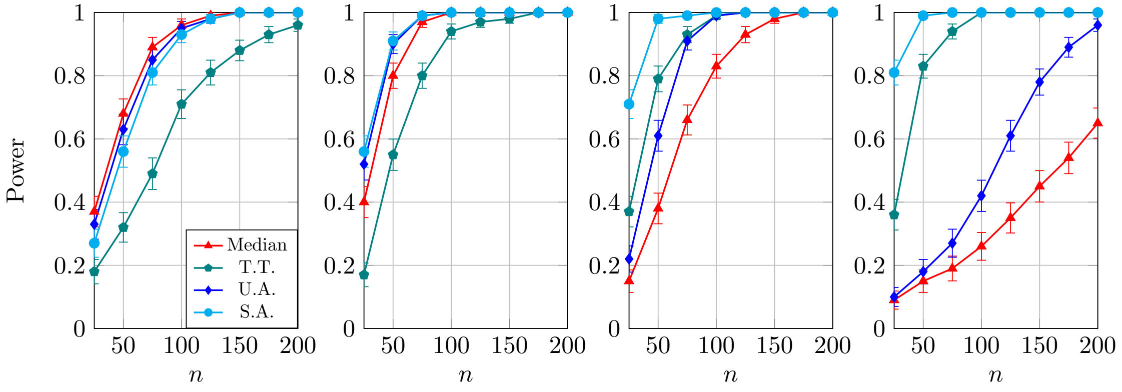


Figure 2: Observed power versus sample size in Experiment III for $d = 1, 10, 100, 1000$ from left to right.

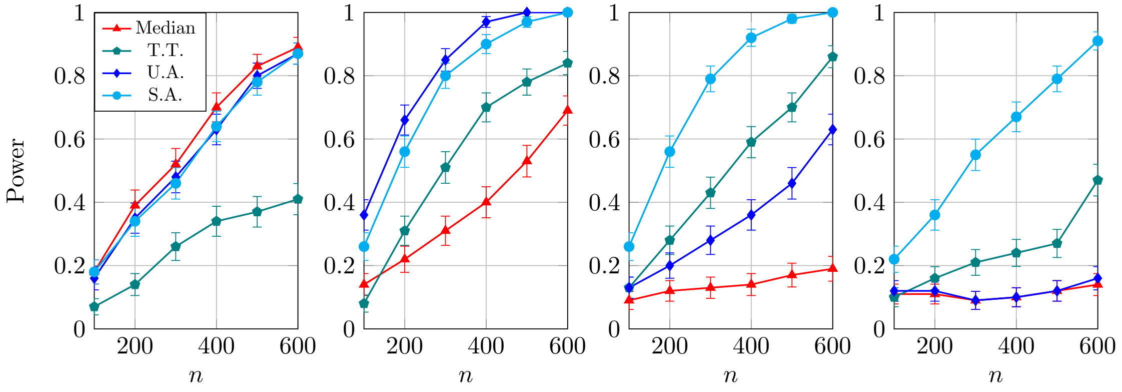


Figure 3: Observed power versus sample size in Experiment IV for $d = 2, 10, 100, 1000$ from left to right.

As Figures 2 and 3 show, for both experiments, these tests are comparable in low-dimensional settings. But as d increases, the proposed self-normalized adaptive test (`S.A.`) and the training-testing approach (`T.T.`) become more and more preferable to `U.A.` and `Median`, and the self-normalized adaptive test even performs better than the training-testing approach. For example, for Experiment IV, when $d = 1000$, the observed power of the self-normalized adaptive test is about 90% when $n = 600$, while the power of the training-testing approach is about 50% and the other two tests have power around only 15%.

Another interesting phenomenon to observe is that in both experiments with sample size fixed, the power of our proposed adaptive test `S.A.` maintains or even increases as the dimensionality increases, while that of `Median` and `U.A.` exhibits a quite clear downtrend. Experiment III taken as an example, although the difference between $\mathbb{P}$ and $\mathbb{Q}$ on a single dimension decreases at the rate $1/\sqrt{d}$, the aggregated difference can still be identifiable. For example, if we knew *a priori* that all dimensions of $\mathbb{P}(\mathbb{Q})$ are independent and they follow the identical normal distribution, we could conduct the classical F-test based on $n \times d$ samples from each univariate normal distribution. Without relying on such *a priori* information, it is still achievable to identify the difference between $\mathbb{P}$ and $\mathbb{Q}$ using kernel embedding related test, whereas a key step is to ensure that an appropriate scaling parameter is selected. In particular, after all squared distance $\|X_i - X_j\|^2$ are rescaled by d, some basic calculations suggest that with $\nu \asymp \sqrt{d}$, the 'signal-to-noise ratio', defined as $\gamma_\nu^2(\mathbb{P},\mathbb{Q})/\text{s.d.}\left(\widehat{\gamma_\nu^2}(\mathbb{P},\mathbb{Q})\middle| H_0^{\rm HOM}\right)$, can be bounded away from 0 as d varies, which essentially guarantees the power of the test. Our adaptive test is exactly designed for the purpose of selecting the appropriate scaling parameter automatically.

6.3 Real 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:

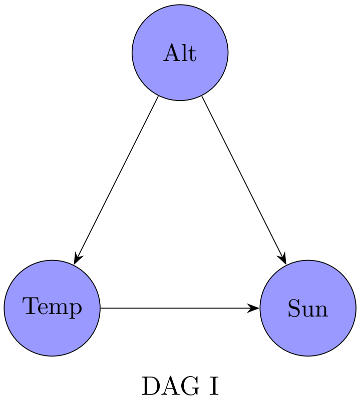
$$X^l = \sum_{r\in \mathrm{PA}^l} f_{l,r}(X^r) + N^l, \quad \forall\, 1 \le l \le d,$$

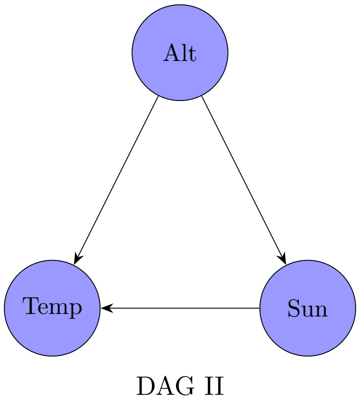
where N^l's are independent Gaussian noises and PA^l denotes the collection of parent nodes of node l specified by $\mathcal{G}_0$. As shown by (Peters et al., 2014), $\mathcal{G}_0$ is identifiable from the joint distribution of $X^1, \cdots, X^d$ under the assumption of $f_{l,r}$'s being non-linear. 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 four 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`), the training-testing approach (`T.T.`) 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 four tests agree on the top three most selected DAGs and they are shown in Figure 4.

In addition, we report in Table 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 four methods.





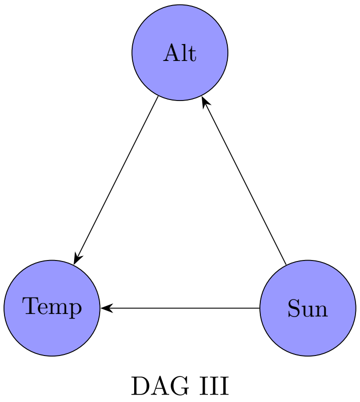


Figure 4: DAGs with the top 3 highest probabilities of being selected.

Prob(%) DAG Test	I	II	III
Median	78.5	4.7	14.5
T.T.	62.4	16.0	8.3
U.A.	81.4	8.1	8.5
S.A.	83.4	9.8	4.7

Table 1: Frequency that each DAG in Figure 4 was selected by four tests.

7. Concluding Remarks

In this paper, we provide a systematic investigation of the statistical properties of Gaussian kernel embedding based nonparametric tests. Our contribution is twofold.

First of all, we provide theoretical justifications for this popular class of methods by showing that they are capable of detecting the smallest possible deviation from the null hypotheses in the context of goodness-of-fit, homogeneity, and independence test. Our analyses also suggest that the existing theoretical studies do not fully explain the practical success of these methods because they assume a fixed kernel or scaling parameter for Gaussian kernel and these methods, as we argue, are most powerful with a scaling parameter that increases appropriately with the sample size.

From a more practical viewpoint, we offer general guidelines on choosing the scaling parameter for Gaussian kernels: our results highlight the importance of using larger scaling parameter for larger sample size and establish the relationship between the smoothness of the underlying densities and the appropriate scaling parameter. Furthermore, we introduce new adaptive testing procedures for goodness-of-fit, homogeneity, and independence, respectively, that are optimal up to a polynomial of iterated logarithmic factor, for a wide range of smooth densities while not needing to know the level of smoothness.

RKHS embedding has emerged as a powerful tool for nonparametric inferences and has found success in numerous applications. Our work here provides insights into their operating characteristics and leads to improved testing procedures within the framework. Our work also pointed to a number of interesting directions that are worth investigating further.

Efficient computation is essential when applying to large datasets. While naive computation of $\tilde{s}^2_{n,\nu_n}$ requires $O(n^4)$ operations, the computational complexity of an equivalent form is only $O(n^2)$. See Appendix B for more details. Similar techniques have been developed in Sutherland et al. (2017). Therefore, the computation of the proposed test $\Phi^{\rm GOF}_{n,\nu_n,\alpha}$ can be completed with $O(n^2)$ operations. The same is true for the homogeneity test $\Phi^{\rm HOM}_{n,\nu_n,\alpha}$ and independence test $\Phi^{\rm IND}_{n,\nu_n,\alpha}$. Particularly, the independence test statistic itself has very similar expression with the aforementioned variance estimator and an equivalent form guarantees $O(n^2)$ computational complexity. See also Song et al. (2007). It is of great interest to see if the computation cost can be further reduced.

Another potential direction is to explore to what extent our findings can be applied to kernels other than Gaussian. A possible class of kernels that may benefit from the technical tools we developed is the translation invariant kernels. This is broad class of kernels that have been widely used, and of which the Gaussian kernel is a specific example. As indicated by Bochner's theorem, every continuous translation invariant kernel is the Fourier transform of some positive finite measure. For example, the Laplacian kernel can be expressed as

$$\exp(-\gamma\|x-x'\|)=\pi^{-(d+1)/2}\gamma^{-d}\Gamma\left(\frac{d+1}{2}\right)\int\left(1+\frac{\|\omega\|^2}{\gamma^2}\right)^{-(d+1)/2}\exp(-i(x-x')^\top\omega)d\omega.$$

Namely, the Laplacian kernel is the Fourier transform of $\gamma^{-d}\left(1+\|\omega/\gamma\|^2\right)^{-(d+1)/2}$ up to some constant, where γ is the scaling parameter. By tuning the positive finite measure at every given sample size, we may be able to make the nonparametric test based on the associated kernel minimax optimal, which warrants further exploration in future studies.

8. Proofs

Throughout this section, we shall write $a_n\lesssim b_n$ if there exists a universal constant $C>0$ such that $a_n\le Cb_n$. 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 We begin with (3). Note that $\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{P}_0)$ 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:

$$\left(\frac{2\nu_n}{\pi}\right)^{d/2}\mathbb{E}\bar{G}^2_{\nu_n}(X_1,X_2)\to\|p_0\|^2_{L_2}, \tag{6}$$

$$\frac{\mathbb{E}\bar{G}^4_{\nu_n}(X_1,X_2)}{n^2[\mathbb{E}\bar{G}^2_{\nu_n}(X_1,X_2)]^2}\to 0, \tag{7}$$

$$\frac{\mathbb{E}[\bar{G}^2_{\nu_n}(X_1,X_2)\bar{G}^2_{\nu_n}(X_1,X_3)]}{n[\mathbb{E}\bar{G}^2_{\nu_n}(X_1,X_2)]^2}\to 0, \tag{8}$$

$$\frac{\mathbb{E}H^2_{\nu_n}(X_1,X_2)}{[\mathbb{E}\bar{G}^2_{\nu_n}(X_1,X_2)]^2}\to 0, \tag{9}$$

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)**.** Note that

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

By Lemma 14,

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

which immediately yields

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

and

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

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 $Z\sim N(0,4\nu_nI_d)$. 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

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

Condition (6) then follows.

Verifying Conditions (7) **and** (8)**.** 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

$$n^{-2}\mathbb{E}\bar{G}^4_{\nu_n}(X_1,X_2)/(\mathbb{E}\bar{G}^2_{\nu_n}(X_1,X_2))^2\lesssim_{d,p_0}\nu_n^{d/2}/n^2\to0.$$

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 (8).

Verifying Condition (9)**.** We now prove (9). 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

$$\mathbb{E}(G_{\nu_n}(X_1+X_4,X_2+X_3)|X_1-X_4,X_2-X_3)\to 0 \tag{10}$$

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

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

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 (10), it suffices to show

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

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

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

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

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

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

which ensures the uniform convergence of $h_n(y)$ to $h(y)$ over $\{y: \ \|y\| \le \delta\}$, and hence (9).

Indeed, we have shown that

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

By Slutsky Theorem, in order to prove (4), it sufficies to show

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

which is equivalent to

$$\tilde{s}^2_{n,\nu_n}/\mathbb{E}\left[\bar{G}^2_{\nu_n}(X_1, X_2)\right] \to_p 1 \tag{11}$$

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

It follows from

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

and

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

that (11) holds. ■

Proof of Theorem 2 Recall that

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

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 $\mathbb{E}V^{(1)}_{\nu_n} = \mathbb{E}V^{(2)}_{\nu_n} = 0$.

Note that

$$\begin{aligned}
&\mathbb{P}\left(T^{\mathrm{GOF}}_{n,\nu_n} \geq z_\alpha\right)\\
=&\mathbb{P}\left(\frac{n}{\sqrt{2}}\widehat{s}^{-1}_{n,\nu_n}\left(\gamma^2_{\nu_n}(\mathbb{P},\mathbb{P}_0)+V^{(1)}_{\nu_n}+V^{(2)}_{\nu_n}\right)\geq z_\alpha\right)\\
\geq&\mathbb{P}\left(\frac{n}{\sqrt{2}}\widehat{s}^{-1}_{n,\nu_n}\left(\gamma^2_{\nu_n}(\mathbb{P},\mathbb{P}_0)+V^{(1)}_{\nu_n}+V^{(2)}_{\nu_n}\right)\geq z_\alpha,\ V^{(1)}_{\nu_n}+V^{(2)}_{\nu_n}\geq-\frac{1}{2}\gamma^2_{\nu_n}(\mathbb{P},\mathbb{P}_0)\right)\\
\geq&\mathbb{P}\left(\frac{n}{2\sqrt{2}}\widehat{s}^{-1}_{n,\nu_n}\gamma^2_{\nu_n}(\mathbb{P},\mathbb{P}_0)\geq z_\alpha,\ V^{(1)}_{\nu_n}+V^{(2)}_{\nu_n}\geq-\frac{1}{2}\gamma^2_{\nu_n}(\mathbb{P},\mathbb{P}_0)\right)\\
\geq&1-\mathbb{P}\left(\frac{n}{2\sqrt{2}}\widehat{s}^{-1}_{n,\nu_n}\gamma^2_{\nu_n}(\mathbb{P},\mathbb{P}_0)< z_\alpha\right)-\mathbb{P}\left(V^{(1)}_{\nu_n}+V^{(2)}_{\nu_n}<-\frac{1}{2}\gamma^2_{\nu_n}(\mathbb{P},\mathbb{P}_0)\right)\\
\geq&1-\frac{2\sqrt{2}z_\alpha\sqrt{\mathbb{E}\left(\widehat{s}^2_{n,\nu_n}\right)}}{n\gamma^2_{\nu_n}(\mathbb{P},\mathbb{P}_0)}-\frac{\mathbb{E}\left(V^{(1)}_{\nu_n}+V^{(2)}_{\nu_n}\right)^2}{\gamma^4_{\nu_n}(\mathbb{P},\mathbb{P}_0)/4}.
\end{aligned}$$

Then once we prove

$$\sup_{\substack{p\in\mathcal{W}^{s,2}(M)\\ \|p-p_0\|\geq\Delta_n}}\frac{\mathbb{E}\left(V^{(1)}_{\nu_n}\right)^2+\mathbb{E}\left(V^{(2)}_{\nu_n}\right)^2}{\gamma^4_{\nu_n}(\mathbb{P},\mathbb{P}_0)}\to 0 \tag{12}$$

and

$$\inf_{\substack{p\in\mathcal{W}^{s,2}(M)\\ \|p-p_0\|\geq\Delta_n}}\frac{n\gamma^2_{\nu_n}(\mathbb{P},\mathbb{P}_0)}{\sqrt{\mathbb{E}\left(\widehat{s}^2_{n,\nu_n}\right)}}\to\infty \tag{13}$$

as $n\to\infty$, it immediately follows that

$$\begin{aligned}
&\mathrm{power}\{\Phi^{\mathrm{GOF}}_{n,\nu_n,\alpha};H^{\mathrm{GOF}}_1(\Delta_n;s)\}\\
=&\inf_{\substack{p\in\mathcal{W}^{s,2}(M)\\ \|p-p_0\|\geq\Delta_n}}\mathbb{P}\left(T^{\mathrm{GOF}}_{n,\nu_n}\geq z_\alpha\right)\\
\geq&1-2\sqrt{2}z_\alpha\cdot\sup_{\substack{p\in\mathcal{W}^{s,2}(M)\\ \|p-p_0\|\geq\Delta_n}}\frac{\sqrt{\mathbb{E}\left(\widehat{s}^2_{n,\nu_n}\right)}}{n\gamma^2_{\nu_n}(\mathbb{P},\mathbb{P}_0)}-8\cdot\sup_{\substack{p\in\mathcal{W}^{s,2}(M)\\ \|p-p_0\|\geq\Delta_n}}\frac{\mathbb{E}\left(V^{(1)}_{\nu_n}\right)^2+\mathbb{E}\left(V^{(2)}_{\nu_n}\right)^2}{\gamma^4_{\nu_n}(\mathbb{P},\mathbb{P}_0)}\to 1.
\end{aligned}$$

We first prove (12). Note that $\|p\|_{L_2}\leq\|p\|_{\mathcal{W}^{s,2}(M)}\leq M$. Following arguments similar to those in the proof of Theorem 1, we get

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

and

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

By Lemma 15, there exists a constant $C>0$ depending on s and M only such that for $f\in\mathcal{W}^{s,2}(M)$,

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

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

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

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

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

as $n\to\infty$.

Next we prove (13). It follows from

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

that (13) holds. ■

Proof of Theorem 3 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

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

It is clear that $H_1^{\text{GOF}}(\Delta_n; s) \supset H_1^{\text{GOF}}(\Delta_n; s, M - \|p_0\|_{\mathcal{W}^{s,2}})^*$. Hence it suffices to prove Theorem 3 with $H_1^{\text{GOF}}(\Delta_n; s)$ replaced by $H_1^{\text{GOF}}(\Delta_n; s, M)^*$ for an arbitrary $M > 0$. We shall abbreviate $H_1^{\text{GOF}}(\Delta_n; s, M)^*$ as $H_1^{\text{GOF}}(\Delta_n; s)^*$ 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 $p_{n,\boldsymbol{a}_n} \in H_1^{\text{GOF}}(\Delta_n; s)^*$ for all possible values of $\boldsymbol{a}_n$ such that

$$\mathbb{E}_{p_0}\left(\frac{p_n(X_1, \cdots, X_n)}{\prod_{i=1}^n p_0(X_i)}\right)^2 = O(1), \tag{14}$$

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

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

In order to ensure (14), it suffices to have

$$B_n^{1/2} n r_n^2 = O(1). \tag{15}$$

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 (15) 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

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

Since for $k \neq k'$, 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

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

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

$$r_n b_n^{d/2}\frac{\|\phi\|_\infty}{\|\phi\|_{L_2}} \to 0 \text{ as } n \to \infty, \tag{16}$$

$$r_n^2 b_n^d = \Delta_n^2, \tag{17}$$

$$r_n^2 b_n^{d+2s}\frac{\|\phi\|_{\mathcal{W}^{s,2}}^2}{\|\phi\|_{L_2}^2} \le M^2. \tag{18}$$

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 (17) and (18) 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 (15) and (16).

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 4 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

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

As we assume $r_n \to r$ as $n \to \infty$, Theorem 1 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

$$\frac{\mathbb{E}\bar{G}^4_{\nu_n}(X_1,X_2)}{N^2\mathbb{E}\bar{G}^2_{\nu_n}(X_1,X_2)}+\frac{\mathbb{E}\bar{G}^2_{\nu_n}(X_1,X_2)\bar{G}^2_{\nu_n}(X_1,X_3)}{N\mathbb{E}\bar{G}^2_{\nu_n}(X_1,X_2)}+\frac{\mathbb{E}H^2_{\nu_n}(X_1,X_2)}{\mathbb{E}\bar{G}^2_{\nu_n}(X_1,X_2)}\to 0 \tag{19}$$

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 1, given $\nu_n \to \infty$ and $\nu_n/n^{4/d} \to 0$, (19) 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 1,

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

The proof is therefore concluded. ■

Proof of Theorem 5 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

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

Similarly to the proof of Theorem 2, in order to prove the consistency of $\Phi^{\mathrm{HOM}}_{n,\nu_n,\alpha}$, it suffices to show

$$\sup_{\substack{p,q\in\mathcal{W}^{s,2}(M)\\ \|p-q\|_{L_2}\geq\Delta_n}} \frac{\mathbb{E}\left(L^{(1)}_{n,\nu_n}\right)^2+\mathbb{E}\left(L^{(2)}_{n,\nu_n}\right)^2}{\gamma^4_{\nu_n}(\mathbb{P},\mathbb{Q})}\to 0, \tag{20}$$

$$\inf_{\substack{p,q\in\mathcal{W}^{s,2}(M)\\ \|p-q\|_{L_2}\geq\Delta_n}} \frac{\gamma^2_{\nu_n}(\mathbb{P},\mathbb{Q})}{(1/n+1/m)\sqrt{\mathbb{E}\left(\widehat{s}^2_{n,m,\nu_n}\right)}}\to\infty, \tag{21}$$

as $n\to\infty$. Following arguments similar to those in the proof of Theorem 2, we can ensure that (20) and (21) hold.

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

$$H^{\mathrm{HOM}}_1(\Delta_n;s)\supset\{(p,p_0):\ p\in H^{\mathrm{GOF}}_1(\Delta_n;s)\}.$$

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,

$$\mathrm{power}\{\Phi_n;H^{\mathrm{HOM}}_1(\Delta_n;s)\}\leq\mathrm{power}\{\tilde{\Phi}_n;H^{\mathrm{GOF}}_1(\Delta_n;s)\}$$

Since $0<c\leq m/n\leq C<\infty$, Theorem 3 ensures that there exists some $\alpha\in(0,1)$ such that for any sequence of asymptotic α-level tests $\{\Phi_n\}_{n\geq1}$,

$$\liminf_{n\to\infty}\mathrm{power}\{\Phi_n;H^{\mathrm{HOM}}_1(\Delta_n;s)\}\leq\liminf_{n\to\infty}\mathrm{power}\{\tilde{\Phi}_n;H^{\mathrm{GOF}}_1(\Delta_n;s)\}<1$$

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

Proof of Theorem 7 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^{IND}_0:

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

where

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

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 E for more details.

Moreover, borrowing arguments in the proof of Lemma 6, we obtain

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

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 1 and 4, we have

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

Since

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

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

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

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

Proof of Theorem 8 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 5. It sufficies to show

$$\sup_{p\in H_1^{\mathrm{IND}}(\Delta_n,s)} \frac{\mathrm{var}(\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2}))}{\gamma^4_{\nu_n}(\mathbb{P},\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2})} \to 0, \tag{22}$$

$$\inf_{p\in H_1^{\mathrm{IND}}(\Delta_n,s)} \frac{n\gamma^2_{\nu_n}(\mathbb{P},\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2})}{\mathbb{E}\left(\widehat{s}_{n,\nu_n}\right)} \to \infty, \tag{23}$$

as $n\to\infty$.

We begin with (22). Let $f = p - p_1\otimes p_2$. Lemma 15 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 $\nu \geq C\|f\|_{L_2}^{-2/s}$, which is satisfied by all $p\in H_1^{\mathrm{IND}}(\Delta_n,s)$ given $\nu=\nu_n$ and $\lim\limits_{n\to\infty}\Delta_n n^{\frac{2s}{4s+d}} = \infty$. 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 E. We follow the same notations here.

Under the alternative hypothesis, the "first order" term

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

no longer vanish, but based on arguments similar to those in the proof of Theorem 2,

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

Moreover, the “second order” term $D_2(\nu_n)$ is not solely $\sum\limits_{1\leq i\neq j\leq n} G^*_{\nu_n}(X_i,X_j)/(n(n-1))$, 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

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

which proves (22).

Now consider (23). Since

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

we have

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

where

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

Therefore (23) 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 5, 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

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

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 3 because of the additional constraints

$$p_1 = p_{1,0}, \quad p_2 = p_{2,0} \tag{24}$$

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

However, by modifying the proof of Theorem 3, we only need to further require each $p_{n,\boldsymbol{a}_n}$ in the proof of Theorem 3 satisfying (24), 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 $p_{n,\boldsymbol{a}_n} = p_0 + r_n \sum\limits_{k=1}^{B_n} a_{n,k}\phi_{n,k}$, 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 9 The proof of Theorem 9 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

$$\lim_{n\to\infty} P(T_n^{\mathrm{GOF(adapt)}} \geq C(d)\log\log n) = 0 \tag{25}$$

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

$$\inf_{p\in H_1^{\mathrm{GOF(adapt)}}(\Delta_{n,s}:s\geq d/4)} P(T_n^{\mathrm{GOF(adapt)}} \geq C(d)\log\log n) \to 1 \tag{26}$$

as $n \to \infty$.

Verifying (25). In order to prove (25), 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 1 that for ν_n increasing at a proper rate,

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

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

Lemma 12 *We have that $\widehat{s}^2_{n,\nu_n}/\mathbb{E}\left[\bar{G}^2_{\nu_n}(X_1, X_2)\right]$ converges to 1 uniformly over $\nu_n \in [1, n^{2/d}]$, i.e.,*

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

We defer the proof of Lemma 12 to the appendix. Note that

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

Lemma 12 first ensures that

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

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

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

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 13 *There exists $C = C(d) > 0$ such that*

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

under $H_0^{\rm GOF}$.

Verifying (26). 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

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

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

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

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

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

Let

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

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

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

as guaranteed by Lemma 15. Further considering that

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

we immediately have

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

■

Proof of Theorem 10 and Theorem 11 The proof of Theorem 10 and Theorem 11 is very similar to that of Theorem 9. Hence we only emphasize the main differences here.

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

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

under $H_0^{\rm HOM}$, observe that

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

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

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

Apply Lemma 13 to bound each term of the right hand side of the above inequality. Then we conclude that for some $C=C(d)>0$,

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

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

$$\lim_{n\to\infty} P(T_n^{\mathrm{IND(adapt)}} \geq C \log\log n) = 0 \tag{27}$$

under H_0^{IND}, recall the decomposition

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

where we express R_n as $R_n = D_3(\nu_n) + D_4(\nu_n)$ in the proof of Theorem 8.

Following arguments similar to those in the proof of Lemma 13, we obtain that there exists $C(d) > 0$ such that for sufficiently large n,

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

Similarly,

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

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

for sufficiently large n.

On the other hand, note that

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

and based on results in the proof of Lemma 12, $\sup_{1\leq\nu_n\leq n^{2/d}} \left|\tilde{s}^2_{n,j,\nu_n}/\mathbb{E}\left[\bar{G}^2_{\nu_n}(X_1^j,X_2^j)\right] - 1\right| = o_p(1)$ for $j = 1, 2$. Further considering that

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

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

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

They combined together ensure that (27) 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

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

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 9. Therefore,

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

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

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

■

Acknowledgments

We acknowledge support for this project from the National Science Foundation (NSF grants DMS-2015285 and DMS-2052955).

Appendix A. 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 14 *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 14 is a close connection between Gaussian kernel MMD and L_2 norm.

Lemma 15 *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 14,

$$\left(\frac{\nu}{\pi}\right)^{d/2} \int G_\nu(x,y)f(x)f(y)dxdy = \int \exp\left(-\frac{\|\omega\|^2}{4\nu}\right) \|\mathcal{F}f(\omega)\|^2 \, d\omega.$$

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

$$\int_{\|\omega\|\le T} \|\mathcal{F}f(\omega)\|^2 \, d\omega \ge \frac{3}{4}\|f\|_{L^2}^2.$$

Hence

$$\begin{aligned}\int \exp\left(-\frac{\|\omega\|^2}{4\nu}\right) \|\mathcal{F}f(\omega)\|^2 \, d\omega &\ge \exp\left(-\frac{T^2}{4\nu}\right) \int_{\|\omega\|\le T} \|\mathcal{F}f(\omega)\|^2 \, d\omega \\ &\ge \frac{3}{4}\exp\left(-\frac{T^2}{4\nu}\right) \|f\|_{L^2}^2.\end{aligned}$$

In particular, if

$$\nu \ge \frac{(2M)^{2/s}}{4\log 3} \cdot \|f\|_{L^2}^{-2/s},$$

then

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

which concludes the proof. ■

Appendix B. Computation Complexity of the Variance Estimator

We show that

$$\sum_{\substack{1\le i_1,i_2,j_1,j_2\le n \\ |\{i_1,i_2,j_1,j_2\}|=4}} G_{\nu_n}(X_{i_1}, X_{j_1}) G_{\nu_n}(X_{i_2}, X_{j_2})$$

can be computed with $O(n^2)$ operations. The result for

$$\sum_{\substack{1\le i,j_1,j_2\le n \\ |\{i,j_1,j_2\}|=3}} G_{\nu_n}(X_i, X_{j_1}) G_{\nu_n}(X_i, X_{j_2})$$

follows similarly.

Proof In this proof, we shall skip writing the constraint that $1 \le i_1, i_2, j_1, j_2 \le n$ but that should be assumed by default.

But checking straightforwardly, we can observe that

$$\begin{aligned}
&\sum_{|\{i_1,i_2,j_1,j_2\}|=4} f(i_1,i_2,j_1,j_2)\\
=&\sum f(i_1,i_2,j_1,j_2)\\
&-\left(\sum_{i_1=i_2} f(i_1,i_2,j_1,j_2)+\sum_{i_1=j_1} f(i_1,i_2,j_1,j_2)+\sum_{i_1=j_2} f(i_1,i_2,j_1,j_2)+\cdots\right)\\
&+\left(\sum_{i_1=i_2=j_1} f(i_1,i_2,j_1,j_2)+\sum_{i_1=i_2=j_2} f(i_1,i_2,j_1,j_2)+\cdots\right)\times 2\\
&+\left(\sum_{i_1=i_2,j_1=j_2} f(i_1,i_2,j_1,j_2)+\sum_{i_1=j_1,i_2=j_2} f(i_1,i_2,j_1,j_2)+\sum_{i_1=j_2,i_2=j_1} f(i_1,i_2,j_1,j_2)\right)\\
&-\left(\sum_{i_1=i_2=j_1=j_2} f(i_1,i_2,j_1,j_2)\right)\times 6
\end{aligned}$$

for a general f.

Let $f(i_1,i_2,j_1,j_2)=G_{\nu_n}(X_{i_1},X_{j_1})G_{\nu_n}(X_{i_2},X_{j_2})$. Then it follows that

$$\begin{aligned}
&\sum_{|\{i_1,i_2,j_1,j_2\}|=4} G_{\nu_n}(X_{i_1},X_{j_1})G_{\nu_n}(X_{i_2},X_{j_2})\\
=&\left(\sum G_{\nu_n}(X_i,X_j)\right)^2-\left(\sum G_{\nu_n}(X_i,X_i)\right)\times\left(\sum G_{\nu_n}(X_i,X_j)\right)\times 2\\
&-\left(\sum_i\left(\sum_j G_{\nu_n}(X_i,X_j)\right)^2\right)\times 4+\left(\sum G_{\nu_n}(X_i,X_i)G_{\nu_n}(X_i,X_j)\right)\times 8\\
&+\left(\sum G_{\nu_n}(X_i,X_i)\right)^2+\left(\sum G^2_{\nu_n}(X_i,X_j)\right)\times 2-\left(\sum G^2_{\nu_n}(X_i,X_i)\right)\times 6\\
=&\left(\sum G_{\nu_n}(X_i,X_j)\right)\times\left(\sum G_{\nu_n}(X_i,X_j)-2n+8\right)-\left(\sum_i\left(\sum_j G_{\nu_n}(X_i,X_j)\right)^2\right)\times 4\\
&+\left(\sum G_{2\nu_n}(X_i,X_j)\right)\times 2+n^2-6n,
\end{aligned}$$

which can be computed with $O(n^2)$ operations. $\blacksquare$

Appendix C. Proof of Lemma 12

We first prove that $\sup_{1\le\nu_n\le n^{2/d}}\left|\tilde{s}^2_{n,\nu_n}/\mathbb{E}\left[\bar{G}^2_{\nu_n}(X_1,X_2)\right]-1\right|=o_p(1)$ and then show the difference caused by the modification from $\tilde{s}^2_{n,\nu_n}$ to $\widehat{s}^2_{n,\nu_n}$ is asymptotically negligible.

Note that

$$\begin{aligned}&\sup_{1\le\nu_n\le n^{2/d}}\left|\tilde{s}^2_{n,\nu_n}/\mathbb{E}\left[\bar{G}^2_{\nu_n}(X_1,X_2)\right]-1\right|\\ \le&\left(\inf_{1\le\nu_n\le n^{2/d}}\nu_n^{d/2}\mathbb{E}\left[\bar{G}^2_{\nu_n}(X_1,X_2)\right]\right)^{-1}\cdot\sup_{1\le\nu_n\le n^{2/d}}\nu_n^{d/2}\left|\tilde{s}^2_{n,\nu_n}-\mathbb{E}\left[\bar{G}^2_{\nu_n}(X_1,X_2)\right]\right|.\end{aligned}$$

For $X\sim\mathbb{P}_0$, denote the distribution of (X,X) as $\mathbb{P}_1$. Then we have

$$\mathbb{E}\left[\bar{G}^2_{\nu_n}(X_1,X_2)\right]=\gamma^2_{\nu_n}(\mathbb{P}_1,\mathbb{P}_0\otimes\mathbb{P}_0).$$

Hence $\mathbb{E}\left[\bar{G}^2_{\nu_n}(X_1,X_2)\right]>0$ for any $\nu_n>0$ since G_{ν_n} is characteristic.

In addition, $\nu_n^{d/2}\mathbb{E}\left[\bar{G}^2_{\nu_n}(X_1,X_2)\right]$ is continuous with respect to ν_n and

$$\lim_{\nu_n\to\infty}\nu_n^{d/2}\mathbb{E}\left[\bar{G}^2_{\nu_n}(X_1,X_2)\right]=\left(\frac{\pi}{2}\right)^{d/2}\|p_0\|^2_{L_2}.$$

Therefore,

$$\inf_{1\le\nu_n\le n^{2/d}}\nu_n^{d/2}\mathbb{E}\left[\bar{G}^2_{\nu_n}(X_1,X_2)\right]\ge\inf_{\nu_n\in[0,\infty)}\nu_n^{d/2}\mathbb{E}\left[\bar{G}^2_{\nu_n}(X_1,X_2)\right]>0,$$

and it remains to prove

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

Recall the expression of $\tilde{s}^2_{n,\nu_n}$. It suffcies to show that

$$\sup_{1\le\nu_n\le n^{2/d}}\nu_n^{d/2}\left|\frac{1}{n(n-1)}\sum_{1\le i\ne j\le n}G_{2\nu_n}(X_i,X_j)-\mathbb{E}G_{2\nu_n}(X_1,X_2)\right|\tag{28}$$

$$\sup_{1\le\nu_n\le n^{2/d}}\nu_n^{d/2}\left|\frac{2(n-3)!}{n!}\sum_{\substack{1\le i,j_1,j_2\le n\\|\{i,j_1,j_2\}|=3}}G_{\nu_n}(X_i,X_{j_1})G_{\nu_n}(X_i,X_{j_2})-\mathbb{E}G_{\nu_n}(X_1,X_2)G_{\nu_n}(X_1,X_3)\right|\tag{29}$$

$$\sup_{1\le\nu_n\le n^{2/d}}\nu_n^{d/2}\left|\frac{(n-4)!}{n!}\sum_{\substack{1\le i_1,i_2,j_1,j_2\le n\\|\{i_1,i_2,j_1,j_2\}|=4}}G_{\nu_n}(X_{i_1},X_{j_1})G_{\nu_n}(X_{i_2},X_{j_2})-[\mathbb{E}G_{\nu_n}(X_1,X_2)]^2\right|\tag{30}$$

are all $o_p(1)$. We shall first control (28) and then bound (29) and (30) in the same way.

Let

$$\widehat{\mathbb{E}}_nG_{2\nu_n}(X,X')=\frac{1}{n(n-1)}\sum_{1\le i\ne j\le n}G_{2\nu_n}(X_i,X_j).$$

In the rest of this proof, abbreviate $\widehat{\mathbb{E}}_nG_{2\nu_n}(X,X')$ and $\mathbb{E}G_{2\nu_n}(X_1,X_2)$ as $\widehat{\mathbb{E}}_nG_{2\nu_n}$ 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 $\nu_n \in [u_{a-1}, u_a]$,

$$\begin{aligned} \nu_n^{d/2}\widehat{\mathbb{E}}_n G_{2\nu_n} - \nu_n^{d/2}\mathbb{E}G_{2\nu_n} \geq & - \nu_n^{d/2}\left|\widehat{\mathbb{E}}_n G_{2u_a} - \mathbb{E}G_{2u_a}\right| - \nu_n^{d/2}\left|\mathbb{E}G_{2u_a} - \mathbb{E}G_{2u_{a-1}}\right| \\ \geq & - u_a^{d/2}\left|\widehat{\mathbb{E}}_n G_{2u_a} - \mathbb{E}G_{2u_a}\right| - u_a^{d/2}\left|\mathbb{E}G_{2u_a} - \mathbb{E}G_{2u_{a-1}}\right| \end{aligned}$$

and

$$\nu_n^{d/2}\widehat{\mathbb{E}}_n G_{2\nu_n} - \nu_n^{d/2}\mathbb{E}G_{2\nu_n} \leq u_a^{d/2}\left|\widehat{\mathbb{E}}_n G_{2u_{a-1}} - \mathbb{E}G_{2u_{a-1}}\right| + u_a^{d/2}\left|\mathbb{E}G_{2u_a} - \mathbb{E}G_{2u_{a-1}}\right|,$$

which together ensure that

$$\begin{aligned} & \sup_{1\leq\nu_n\leq n^{2/d}}\left|\nu_n^{d/2}\widehat{\mathbb{E}}_n G_{2\nu_n} - \nu_n^{d/2}\mathbb{E}G_{2\nu_n}\right| \\ \leq & \sup_{1\leq a\leq A}\left(\frac{u_a}{u_{a-1}}\right)^{d/2}\cdot\sup_{0\leq a\leq A} u_a^{d/2}\left|\widehat{\mathbb{E}}_n G_{2u_a} - \mathbb{E}G_{2u_a}\right| + \sup_{1\leq a\leq A} u_a^{d/2}\left|\mathbb{E}G_{2u_a} - \mathbb{E}G_{2u_{a-1}}\right| \\ \leq & \sup_{1\leq a\leq A}\left(\frac{u_a}{u_{a-1}}\right)^{d/2}\cdot\sup_{0\leq a\leq A} u_a^{d/2}\left|\widehat{\mathbb{E}}_n G_{2u_a} - \mathbb{E}G_{2u_a}\right| + \sup_{1\leq a\leq A}\left|u_a^{d/2}\mathbb{E}G_{2u_a} - u_{a-1}^{d/2}\mathbb{E}G_{2u_{a-1}}\right| \\ & + \sup_{1\leq a\leq A}\left(\left(u_a^{d/2} - u_{a-1}^{d/2}\right)\mathbb{E}G_{2u_{a-1}}\right). \end{aligned}$$

Bound the three terms in the right hand side of the last inequality separately.

Let $\{u_a\}_{a\geq 0}$ be a geometric sequence, namely,

$$A := \inf\{a \in \mathbb{N} : r^a \geq n^{2/d}\},$$

and

$$u_a = \begin{cases} r^a, & \forall\ 0 \leq a \leq A-1 \\ n^{2/d}, & a = A \end{cases},$$

with $r > 1$ to be determined later.

Since $\lim\limits_{\nu\to\infty}\nu^{d/2}\mathbb{E}G_{2\nu_n} = (\pi/2)^{d/2}\|p_0\|^2$ 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

$$\sup_{1\leq a\leq A}\left|u_a^{d/2}\mathbb{E}G_{2u_a} - u_{a-1}^{d/2}\mathbb{E}G_{2u_{a-1}}\right| \leq \varepsilon.$$

At the same time, we can also ensure

$$\sup_{1\leq a\leq A}\left(\left(u_a^{d/2} - u_{a-1}^{d/2}\right)\mathbb{E}G_{2u_{a-1}}\right) \leq (r^{d/2} - 1)\left(\frac{\pi}{2}\right)^{d/2}\|p_0\|^2 \leq \varepsilon$$

by choosing r sufficiently small.

Finally consider

$$\sup_{1\leq a\leq A}\left(\frac{u_a}{u_{a-1}}\right)^{d/2}\cdot\sup_{0\leq a\leq A} u_a^{d/2}\left|\widehat{\mathbb{E}}_n G_{2u_a} - \mathbb{E}G_{2u_a}\right|.$$

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

$$\begin{aligned}\mathrm{var}\left(\widehat{\mathbb{E}}_n G_{2\nu_n}\right)&\lesssim \frac{1}{n}\mathbb{E}G_{2\nu_n}(X,X')G_{2\nu_n}(X,X'')+\frac{1}{n^2}\mathbb{E}G_{4\nu_n}(X,X')\\&\lesssim_d \frac{\nu_n^{-3d/4}\|p_0\|^3}{n}+\frac{\nu_n^{-d/2}\|p_0\|^2}{n^2}\end{aligned}$$

for any $\nu_n\in(0,\infty)$, we have

$$\begin{aligned}&P\left(\sup_{0\le a\le A}u_a^{d/2}\left|\widehat{\mathbb{E}}_n G_{2u_a}-\mathbb{E}G_{2u_a}\right|\ge\varepsilon\right)\\&\le\frac{\sum_{a=0}^{A}u_a^d\mathrm{var}\left(\widehat{\mathbb{E}}_n G_{2u_a}\right)}{\varepsilon^2}\lesssim_{d,r}\frac{1}{\varepsilon^2}\left(\frac{u_A^{d/4}\|p_0\|^3}{n}+\frac{u_A^{d/2}\|p_0\|^2}{n^2}\right)\to 0\end{aligned}$$

as $n\to\infty$. Hence we conclude $\sup_{1\le\nu_n\le n^{2/d}}\left|\nu_n^{d/2}\widehat{\mathbb{E}}_n G_{2\nu_n}-\nu_n^{d/2}\mathbb{E}G_{2\nu_n}\right|=o_p(1)$.

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 (29) and (30) are also $o_p(1)$, based on almost the same arguments. Hence

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

On the other hand, since $\mathbb{E}\left[\bar{G}^2_{\nu_n}(X_1,X_2)\right]\gtrsim_{p_0,d}\nu_n^{-d/2}$ for $\nu_n\in[1,n^{2/d}]$,

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

Hence we finally conclude that

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

■

Appendix D. Proof of Lemma 13

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

$$\tilde{T}_n^{\mathrm{GOF(adapt)}} = \sup_{1\le\nu_n\le n^{2/d}} \left| \frac{1}{n-1}\sum_{i\neq j} \bar{K}_{\nu_n}(X_i,X_j)\cdot\sqrt{\frac{\mathbb{E}G_{2\nu_n}(X_1,X_2)}{\mathbb{E}\left[\bar{G}^2_{\nu_n}(X_1,X_2)\right]}}\right|.$$

To finish this proof, we first bound

$$\sup_{1\le\nu_n\le n^{2/d}} \left| \frac{1}{n-1}\sum_{i\neq j} \bar{K}_{\nu_n}(X_i,X_j)\right| \tag{31}$$

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 (31).

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 $t>0$,

$$P\left(\left| \frac{1}{n-1}\sum_{i\neq j} \bar{K}_{\nu_0}(X_i,X_j)\right|\ge t\right) \le C\exp\left(-C\min\left\{\frac{t}{\|\bar{K}_{\nu_0}\|_{L_2}}, \left(\frac{\sqrt{n}t}{\|\bar{K}_{\nu_0}\|_{L_\infty}}\right)^{\frac{2}{3}}\right\}\right) \tag{32}$$

and

$$\begin{aligned}&P\left(\left| \frac{1}{n-1}\sum_{i\neq j} \left(\bar{K}_{\nu_n}(X_i,X_j)-\bar{K}_{\nu_n'}(X_i,X_j)\right)\right|\ge t\right)\\ \le& C\exp\left(-C\min\left\{\frac{t}{d_2(\nu_n,\nu_n')}, \left(\frac{\sqrt{n}t}{d_1(\nu_n,\nu_n')}\right)^{\frac{2}{3}}\right\}\right)\end{aligned}$$

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

$$\begin{aligned}P\Bigg(\sup_{1\le\nu_n\le n^{2/d}} &\left| \frac{1}{n-1}\sum_{i\neq j} \left(\bar{K}_{\nu_n}(X_i,X_j)-\bar{K}_{\nu_0}(X_i,X_j)\right)\right| \ge \\ &C\left(\frac{\gamma_{2/3}([1,n^{2/d}],d_1)}{\sqrt{n}}t+\gamma_1([1,n^{2/d}],d_2)+D_2t\right)\Bigg) \lesssim \exp(-t^{2/3}),\end{aligned} \tag{33}$$

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 (32) and (33) then gives

$$P\Bigg(\sup_{1\leq\nu_n\leq n^{2/d}}\left|\frac{1}{n-1}\sum_{i\neq j}\bar{K}_{\nu_n}(X_i,X_j)\right|\geq$$
$$C\left(\frac{\gamma_{2/3}([1,n^{2/d}],d_1)}{\sqrt{n}}t+\gamma_1([1,n^{2/d}],d_2)+D_2t+\frac{\|\bar{K}_{\nu_0}\|_{L_\infty}}{\sqrt{n}}+\|\bar{K}_{\nu_0}\|_{L_2}t\right)\Bigg)\lesssim\exp(-t^{2/3}).$$

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 $\nu_0=1$,

$$\|\bar{K}_{\nu_0}\|_{L_\infty}\leq 4\|K_{\nu_0}\|_{L_\infty}=\frac{4}{\sqrt{2\mathbb{E}G_2}},\qquad\|\bar{K}_{\nu_0}\|_{L_2}\leq\|K_{\nu_0}\|_{L_2}=\frac{\sqrt{2}}{2},$$

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 $1\leq\nu_n<\nu_n'<\infty$,

$$d_1(\nu_n,\nu_n')\leq 4\|K_{\nu_n}-K_{\nu_n'}\|_{L_\infty}\leq 4\int_{\nu_n}^{\nu_n'}\left\|\frac{dK_u}{du}\right\|_{L_\infty}du$$

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

$$\begin{aligned}\left(\mathbb{E}G_{2\nu_n}(X_1,X_2)\right)^{-1/2}=&\left(\frac{\pi}{2}\right)^{-d/4}\nu_n^{d/4}\left(\int\exp\left(-\frac{\|\omega\|^2}{8\nu_n}\right)\|\mathcal{F}p_0(\omega)\|^2d\omega\right)^{-1/2}\\\lesssim_d&\ \nu_n^{d/4}\left(\int\exp\left(-\frac{\|\omega\|^2}{8}\right)\|\mathcal{F}p_0(\omega)\|^2d\omega\right)^{-1/2},\end{aligned}$$

$$\left(\mathbb{E}G_{2\nu_n}(X_1,X_2)\right)^{-3/2}\lesssim_d\nu_n^{3d/4}\left(\int\exp\left(-\frac{\|\omega\|^2}{8}\right)\|\mathcal{F}p_0(\omega)\|^2d\omega\right)^{-3/2},$$

and

$$\begin{aligned}&\frac{d}{d\nu_n}\mathbb{E}_{2\nu_n}(X_1,X_2)\\=&\left(\frac{\pi}{2}\right)^{d/2}\nu_n^{-d/2-1}\left(-\frac{d}{2}\cdot\int\exp\left(-\frac{\|\omega\|^2}{8\nu_n}\right)\|\mathcal{F}p_0(\omega)\|^2d\omega\right.\\&\left.+\int\exp\left(-\frac{\|\omega\|^2}{8\nu_n}\right)\left(\frac{\|\omega\|^2}{8\nu_n}\right)\|\mathcal{F}p_0(\omega)\|^2d\omega\right),\end{aligned}$$

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 $f_1(\nu_n)=\int\exp\left(-\frac{\|\omega\|^2}{8\nu_n}\right)\|\mathcal{F}p_0(\omega)\|^2d\omega$. 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

$$\frac{\log f_1(\nu_n)+\log f_1(\nu_n')}{2}-\log f_1\left(\frac{\nu_n+\nu_n'}{2}\right)=\frac{1}{2}\int_0^{\frac{\nu_n'-\nu_n}{2}}\int_{-u}^{u}\left(\log f_1\left(\frac{\nu_n'+\nu_n}{2}+v\right)\right)''dvdu.$$

For any $\nu_n\ge 1$,

$$(\log f_1(\nu_n))''=\frac{f_1(\nu_n)f_1''(\nu_n)-(f_1'(\nu_n))^2}{f_1^2(\nu_n)}\le\frac{f_1''(\nu_n)}{f_1(\nu_n)},$$

and

$$f_1''(\nu_n)=\int\exp\left(-\frac{\|\omega\|^2}{8\nu_n}\right)\left(\frac{\|\omega\|^4}{64\nu_n^4}-\frac{\|\omega\|^2}{4\nu_n^3}\right)\|\mathcal{F}p_0(\omega)\|^2d\omega\lesssim\nu_n^{-2}\|p_0\|_{L_2}^2.$$

Moreover, there exists $\nu_n^*=\nu_n^*(p_0)>1$ such that $f_1(\nu_n^*)\ge\|p_0\|_{L_2}^2/2$, from which we obtain

$$(\log f_1(\nu_n))''\lesssim\begin{cases}\nu_n^{-2}\|p_0\|_{L_2}^2/f_1(1), & 1\le\nu_n\le\nu_n^*\\ \nu_n^{-2}, & \nu_n^*<\nu_n\le n^{2/d}\end{cases},$$

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

$$d_2(\nu_n, \nu_n') \le \|\bar{K}_{\nu_n}\|_{L_2} + \|\bar{K}_{\nu_n'}\|_{L_2} \le \|K_{\nu_n}\|_{L_2} + \|K_{\nu_n'}\|_{L_2} \le \sqrt{2}.$$

Therefore,

$$\begin{aligned}
\gamma_1([1, n^{2/d}], d_2) \le& \sum_{l\ge 0} 2^l e_l([1, n^{2/d}], d_2) \\
\lesssim& e_0([1, n^{2/d}], d_2) + \sum_{l\ge 0} 2^l e_l([1, \nu_n^*], d_2) + \sum_{l\ge 0} 2^l e_l([\nu_n^*, n^{2/d}], d_2) \\
\lesssim& 1 + \sqrt{\frac{d}{2} + \frac{\|p_0\|_{L_2}^2}{f_1(1)}} \sum_{l\ge 0} 2^l \sqrt{\frac{\log \nu_n^* - \log 1}{2^{2^l}}} \\
&+ \sqrt{\frac{d}{2}+1}\left(\sum_{l\ge 0} 2^l \min\left\{1, \sqrt{\frac{\log n^{2/d} - \log \nu_n^*}{2^{2^l}}}\right\}\right) \\
\lesssim& 1 + \sqrt{\frac{d}{2} + \frac{\|p_0\|_{L_2}^2}{f_1(1)}} \sqrt{\log \nu_n^*} + \sqrt{\frac{d}{2}+1}\left(\sum_{l\ge 0} 2^l \min\left\{1, \sqrt{\frac{\log n^{2/d}}{2^{2^l}}}\right\}\right) \\
\lesssim& 1 + \sqrt{\frac{d}{2} + \frac{\|p_0\|_{L_2}^2}{f_1(1)}} \sqrt{\log \nu_n^*} + \sqrt{\frac{d}{2}+1}\left(\sum_{0\le l<l^*} 2^l + \sum_{l\ge l^*} 2^l \sqrt{\frac{\log n^{2/d}}{2^{2^l}}}\right) \\
\lesssim& 1 + \sqrt{\frac{d}{2} + \frac{\|p_0\|_{L_2}^2}{f_1(1)}} \sqrt{\log \nu_n^*} + \sqrt{\frac{d}{2}+1} \cdot 2^{l^*}
\end{aligned}$$

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,

$$P\left(\sup_{1\le\nu_n\le n^{2/d}}\left|\frac{1}{n-1}\sum_{i\neq j}\bar{K}_{\nu_n}(X_i,X_j)\right|\ge C(d)(\log\log n+t\log\log\log n)\right)\lesssim\exp(-t^{2/3}). \tag{34}$$

Step (ii). By slight abuse of notation, there exists $\nu_n^*=\nu_n^*(p_0)>1$ such that

$$\frac{\mathbb{E}G_{2\nu_n}(X_1,X_2)}{\mathbb{E}\left[\bar{G}_{\nu_n}^2(X_1,X_2)\right]}\le 2$$

for $\nu_n\ge\nu_n^*$. Therefore,

$$\begin{aligned}\tilde{T}_n^{\mathrm{GOF(adapt)}}\le&\sup_{1\le\nu_n\le\nu_n^*}\sqrt{\frac{\mathbb{E}G_{2\nu_n}(X_1,X_2)}{\mathbb{E}\left[\bar{G}_{\nu_n}^2(X_1,X_2)\right]}}\cdot\sup_{1\le\nu_n\le\nu_n^*}\left|\frac{1}{n-1}\sum_{i\neq j}\bar{K}_{\nu_n}(X_i,X_j)\right|+\\&\sqrt{2}\sup_{\nu_n^*\le\nu_n\le n^{2/d}}\left|\frac{1}{n-1}\sum_{i\neq j}\bar{K}_{\nu_n}(X_i,X_j)\right|\\\le&C(p_0)\sup_{1\le\nu_n\le\nu_n^*}\left|\frac{1}{n-1}\sum_{i\neq j}\bar{K}_{\nu_n}(X_i,X_j)\right|+\\&\sqrt{2}\sup_{\nu_n^*\le\nu_n\le n^{2/d}}\left|\frac{1}{n-1}\sum_{i\neq j}\bar{K}_{\nu_n}(X_i,X_j)\right|\end{aligned}$$

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 (34) still holds when ν_n is restricted to $[\nu_n^*,n^{2/d}]$. They together prove Lemma 13. ■

Appendix E. 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 $\mathrm{var}\left(\widehat{\gamma_\nu^2}(\mathbb{P},\mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^k})\right)$ 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 6

Observe that under H_0,

$$
\mathrm{var}\left(\widehat{\gamma_\nu^2}(\mathbb{P},\mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^k})\right)=\mathbb{E}(D_2(\nu))^2+\mathbb{E}\left(R_n\right)^2=\frac{2}{n(n-1)}\mathbb{E}\left[G_\nu^*(X_1,X_2)^2\right]+\mathbb{E}\left(R_n\right)^2,
$$

$$
\mathbb{E}\left(R_n\right)^2\lesssim_k n^{-3}\mathbb{E}G_{2\nu}(X_1,X_2),
$$

and

$$
\begin{aligned}
&\mathbb{E}\left[G_\nu^*(X_1,X_2)^2\right]\\
=&\mathbb{E}\left(\left(\bar{G}_\nu(X_1,X_2)-\sum_{1\le j\le k}g_j(X_1^j,X_2)\right)^2\right)\\
&-\mathbb{E}\left(\left(\sum_{1\le j\le k}g_j(X_2^j,X_1)+\sum_{1\le j_1,j_2\le k}g_{j_1,j_2}(X_1^{j_1},X_2^{j_2})\right)^2\right)\\
=&\mathbb{E}\bar{G}_\nu^2(X_1,X_2)-2\sum_{1\le j\le k}\mathbb{E}\Big(g_j^2(X_1^j,X_2)\Big)+\sum_{1\le j_1,j_2\le k}\mathbb{E}\Big(g_{j_1,j_2}^2(X_1^{j_1},X_2^{j_2})\Big).
\end{aligned}
$$

They together conclude the proof. ■

Below we shall further expand $\mathbb{E}\bar{G}_\nu^2(X_1, X_2)$, $\mathbb{E}\Big(g_j^2(X_1^j, X_2)\Big)$ and $\mathbb{E}\Big(g_{j_1,j_2}^2(X_1^{j_1}, X_2^{j_2})\Big)$ in Lemma 6, based on which consistent estimator of 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,

$$\begin{aligned}
&\mathbb{E}\Big(g_j^2(X_1^j, X_2)\Big)\\
=&\mathbb{E}G_{2\nu}(X_1^j, X_2^j) \cdot \prod_{l\ne j}\mathbb{E}G_\nu(X_1^l, X_2^l)G_\nu(X_1^l, X_3^l) - \prod_{1\le l\le k} \mathbb{E}G_\nu(X_1^l, X_2^l)G_\nu(X_1^l, X_3^l)\\
&- \mathbb{E}G_\nu(X_1^j, X_2^j)G_\nu(X_1^j, X_3^j) \cdot \prod_{l\ne j}(\mathbb{E}G_\nu(X_1^l, X_2^l))^2 + \prod_{1\le l\le k}\Big(\mathbb{E}G_\nu(X_1^l, X_2^l)\Big)^2.
\end{aligned}$$

Hence

$$\begin{aligned}
&\sum_{1\le j\le k}\mathbb{E}\Big(g_j^2(X_1^j, X_2)\Big)\\
=&\left(\prod_{1\le l\le k}\mathbb{E}G_\nu(X_1^l, X_2^l)G_\nu(X_1^l, X_3^l)\right)\left(\sum_{1\le j\le k}\frac{\mathbb{E}G_{2\nu}(X_1^j, X_2^j)}{\mathbb{E}G_\nu(X_1^j, X_2^j)G_\nu(X_1^j, X_3^j)} - k\right)\\
&-\left(\prod_{1\le l\le k}\Big(\mathbb{E}G_\nu(X_1^l, X_2^l)\Big)^2\right)\left(\sum_{1\le j\le k}\frac{\mathbb{E}G_\nu(X_1^j, X_2^j)G_\nu(X_1^j, X_3^j)}{(\mathbb{E}G_\nu(X_1^j, X_2^j))^2} - k\right).
\end{aligned}$$

Finally,

$$\begin{aligned}
&\mathbb{E}\Big(g_{j_1,j_2}^2(X_1^{j_1}, X_2^{j_2})\Big)\\
=&\begin{cases}\mathbb{E}(\bar{G}_\nu^2(X_1^{j_1}, X_2^{j_1})) \cdot \prod\limits_{l\ne j_1}\Big(\mathbb{E}G_\nu(X_1^l, X_2^l)\Big)^2, & j_1 = j_2\\ \prod\limits_{l\in\{j_1,j_2\}}(\mathbb{E}G_\nu(X_1^l, X_2^l)G_\nu(X_1^l, X_3^l) - (\mathbb{E}G_\nu(X_1^l, X_2^l))^2)\prod\limits_{l\ne j_1,j_2}\Big(\mathbb{E}G_\nu(X_1^l, X_2^l)\Big)^2, & j_1 \ne j_2.\end{cases}
\end{aligned}$$

Hence

$$\begin{aligned}
&\sum_{1\le j_1,j_2\le k}\mathbb{E}\Big(g^2_{j_1,j_2}(X_1^{j_1},X_2^{j_2})\Big)\\
&=\left(\prod_{1\le l\le k}\Big(\mathbb{E}G_\nu(X_1^l,X_2^l)\Big)^2\right)\left(\sum_{1\le j_1\le k}\frac{\mathbb{E}(\bar{G}^2_\nu(X_1^{j_1},X_2^{j_1}))}{(\mathbb{E}G_\nu(X_1^{j_1},X_2^{j_1}))^2}\right.\\
&\quad\left.+\sum_{1\le j_1\ne j_2\le k}\prod_{l\in\{j_1,j_2\}}\left(\frac{\mathbb{E}G_\nu(X_1^l,X_2^l)G_\nu(X_1^l,X_2^l)}{\big(\mathbb{E}G_\nu(X_1^l,X_2^l)\big)^2}-1\right)\right).
\end{aligned}$$

Then the consistent estimator $\tilde{s}^2_{n,\nu}$ of $\mathbb{E}\left(G^*_\nu(X_1,X_2)^2\right)$ 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

$$\mathbb{E}\bar{G}^2_\nu(X_1,X_2),\quad \sum_{1\le j\le k}\mathbb{E}\Big(g_j^2(X_1^j,X_2)\Big),\quad \sum_{1\le j_1,j_2\le k}\mathbb{E}\Big(g^2_{j_1,j_2}(X_1^{j_1},X_2^{j_2})\Big)$$

with the corresponding unbiased estimators

$$\frac{1}{n(n-1)}\sum_{1\le i\ne j\le n}G_{2\nu_n}(X_i^l,X_j^l),\quad \frac{(n-3)!}{n!}\sum_{\substack{1\le i,j_1,j_2\le n\\ |\{i,j_1,j_2\}|=3}}G_{\nu_n}(X_i^l,X_{j_1}^l)G_{\nu_n}(X_i^l,X_{j_2}^l)$$

$$\frac{(n-4)!}{n!}\sum_{\substack{1\le i_1,i_2,j_1,j_2\le n\\ |\{i_1,i_2,j_1,j_2\}|=4}}G_{\nu_n}(X_{i_1}^l,X_{j_1}^l)G_{\nu_n}(X_{i_2}^l,X_{j_2}^l)$$

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 $\operatorname{var}\left(\widehat{\gamma^2_\nu}(\mathbb{P},\mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^k})\right)$ 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

$$T^{\rm IND}_{n,\nu_n}=\frac{n}{\sqrt{2}}\widehat{s}^{-1}_{n,\nu_n}\widehat{\gamma^2_{\nu_n}}(\mathbb{P},\mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^k})\quad\text{and}\quad T_n^{\rm IND(adapt)}=\max_{1\le\nu_n\le n^{2/d}}T^{\rm IND}_{n,\nu_n}$$

respectively. Accordingly, $\Phi^{\rm IND}_{n,\nu_n,\alpha}$ and $\Phi^{\rm IND(adapt)}$ can be constructed as in the case of $k=2$.

Appendix F. Theoretical Properties of Independence Tests for General k

In this section, with $\Phi^{\rm IND}_{n,\nu_n,\alpha}$ and $\Phi^{\rm IND(adapt)}$ constructed in Appendix E for general k, we confirm that Theorem 7, Theorem 8 and Theorem 11 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 7 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 11.

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

$$\nu_n^{d_l/2}\mathbb{E}G_{2\nu_n}(X_1^l,X_2^l), \quad \nu_n^{d/2}\mathbb{E}G_{\nu_n}(X_1^l,X_2^l)G_{\nu_n}(X_1^l,X_3^l), \quad \nu_n^{d_l/2}(\mathbb{E}G_{\nu_n}(X_1^l,X_2^l))^2 \qquad (35)$$

are all bounded and they are estimated consistently by their corresponding estimators. For example,

$$\nu_n^{d_l/2}\mathbb{E}G_{2\nu_n}(X_1^l,X_2^l) \to (\pi/2)^{d_l/2}\,\|p_l\|^2_{L_2}$$

and

$$\begin{aligned}
&\nu_n^{d_l}\mathbb{E}\left(\frac{1}{n(n-1)}\sum_{1\le i\ne j\le n} G_{2\nu_n}(X_i^l,X_j^l) - \mathbb{E}G_{2\nu_n}(X_1^l,X_2^l)\right)^2 \\
=\ & \nu_n^{d_l}\mathrm{var}\left(\frac{1}{n(n-1)}\sum_{1\le i\ne j\le n} G_{2\nu_n}(X_i^l,X_j^l)\right) \\
\lesssim\ & \nu_n^{d_l}\left(n^{-1}\mathbb{E}G_{2\nu_n}(X_1^l,X_2^l)G_{2\nu_n}(X_1^l,X_3^l) + n^{-2}\mathbb{E}G_{4\nu_n}(X_1^l,X_2^l)\right) \\
\lesssim_{d_l}& n^{-1}\nu_n^{d_l/4}\|p_l\|^3_{L_2} + n^{-2}\nu_n^{d_l/2}\|p_l\|^2_{L_2} \to 0.
\end{aligned}$$

The proof of the latter one is similar. It sufficies to have

- each term in (35) 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 (35) 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 12.

Under the alternative hypothesis: we only need to re-ensure that $\widehat{s}_{n,\nu_n}$ is bounded. Specifically, we show

$$\inf_{p\in H_1^{\rm IND}(\Delta_n,s)} \frac{n\gamma^2_{\nu_n}(\mathbb{P},\mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^k})}{\left[\mathbb{E}\left(\widehat{s}^2_{n,\nu_n}\right)^{1/k}\right]^{k/2}} \to \infty$$

for Theorem 8 and

$$\inf_{s\geq d/4}\inf_{p\in H_1^{\rm IND}(\Delta_{n,s};s)} P\left(\widehat{s}^2_{n,\nu_n(s)'} \leq 2M^2(2\nu_n(s)'/\pi)^{-d/2}\right) \to 1 \tag{36}$$

for Theorem 11, 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 (35) 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,

$$\inf_{s\geq d/4}\inf_{p\in H_1^{\rm IND}(\Delta_{n,s};s)} P\left(\left(\nu_n(s)'\right)^{d/2}\left|\tilde{s}^2_{n,\nu_n(s)'} - \mathbb{E}[G^*_{\nu_n(s)'}(Y_1,Y_2)^2]\right| \leq M^2(\pi/2)^{d/2}\right) \to 1$$

where $Y_1, Y_2 \sim_{\rm iid} \mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^k}$. Further considering that

$$\mathbb{E}[G^*_{\nu_n(s)'}(Y_1,Y_2)^2] \leq \mathbb{E}[\bar{G}^2_{\nu_n(s)'}(Y_1,Y_2)] \leq M^2(\pi/(2\nu_n(s)'))^{d/2}$$

and that

$$1/n^2 = o((\nu_n(s)')^{-d/2})$$

uniformly over all s, we prove (36).

References

M. A. Arcones and E. Gine. Limit theorems for U-processes. *The Annals of Probability*, 21 (3):1494–1542, 1993.

K. Balasubramanian, T. Li, and M. Yuan. On the optimality of kernel-embedding based goodness-of-fit tests. *arXiv preprint arXiv:1709.08148*, 2017.

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

P. J. Bickel and Y. Ritov. Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 381–393, 1988.

M. V. Burnashev. On the minimax detection of an inaccurately known signal in a white gaussian noise background. *Theory of Probability & Its Applications*, 24(1):107–119, 1979.

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

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

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

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

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. *Advances in Neural Information Processing Systems*, 20, 2007.

E. Giné and R. Nickl. A simple adaptive estimator of the integrated square of a density. *Bernoulli*, 14(1):47–61, 2008.

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

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

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

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

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

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

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

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

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

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

O. V. Lepski and V. G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, 25(6):2512–2546, 1997.

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

O. V. Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991.

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

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

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

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1): 5–31, 2018.

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

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

L. Song, A. Smola, A. Gretton, K. M. Borgwardt, and J. Bedo. Supervised feature selection via dependence estimation. In *Proceedings of the 24th International Conference on Machine Learning*, pages 823–830. ACM, 2007.

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

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

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

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

G. J Székely and M. L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265, 2009.

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

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

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