Minimax Nonparametric Two-Sample Test under Adversarial Losses

Rong Tang
University of Illinois Urbana-Champaign

Yun Yang
University of Illinois Urbana-Champaign

Abstract

In this paper, we consider the problem of two-sample hypothesis testing that aims at detecting the difference between two probability densities based on finite samples. The proposed test statistic is constructed by first truncating a sample version of a negative Besov norm and then normalizing it. Here, the negative Besov norm is the norm associated with a Besov space with negative exponent, and is shown to be closely related to a class of commonly used adversarial losses (or integral probability metrics) with smooth discriminators. Theoretically, we characterize the optimal detection boundary of two-sample testing in terms of the dimensionalities and smoothness levels of the underlying densities and the discriminator class defining the adversarial loss. We also show that the proposed approach can simultaneously attain the optimal detection boundary under many common adversarial losses, including those induced by the ℓ_1, ℓ_2 distances and Wasserstein distances. Our numerical experiments show that the proposed test procedure tends to exhibit higher power and robustness in difference detection than existing state-of-the-art competitors.

1 Introduction

The problem of two-sample hypothesis testing, which aims at determining whether two underlying probability densities are significantly different based on their samples, has been a central topic in statistics and machine learning. Many classic two-sample tests follow parametric approaches, which are designed based on prior information about the parametric form of the underlying distributions, like Gaussianity. Examples of classic parametric two-sample tests include Pearson's chi-squared test (Pearson, 1900), Student's t-test (Student, 1908) and Hotelling's two-sample test (Hotelling, 1931).

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

On the other hand, nonparametric two-sample test procedures avoid making any restrictive parametric assumptions on the distributions, and therefore tends to be more robust while less efficient when the parametric assumption indeed holds. There are rich literatures regarding the nonparametric two-sample testing problem. Nonparametric comparison for one-dimensional samples was done in the minimax sense in ℓ_2 distance by Ingster (1986) via a χ^2-type test statistic. Butucea and Tribouley (2006) proposed a minimax univariate two-sample testing procedure in ℓ_2 and ℓ_∞ distances based on the wavelet expansion, and the proposed procedure is adaptive to the smoothness of the underlying densities. Multivariate nonparametric two-sample testing problems have also been investigated in the literature. Friedman and Rafsky (1979) used the idea of minimal spanning tree (MST) to generalize the univariate test. Xing et al. (2019) addresses the problem of comparing probability density distributions by establishing a connection with interaction testing, and they propose a minimax optimal penalized likelihood ratio test for conducting interaction testing in this scenario. Gretton et al. (2012a); Li and Yuan (2019); Gretton et al. (2009a, 2012b) proposed two-sample tests based on Maximum Mean Discrepancy (MMD). In particular, Li and Yuan (2019) showed that two-sample tests via Gaussian kernel embedding with an appropriately chosen scaling parameter can attain the minimax optimal rate $n^{-\frac{2\alpha}{4\alpha+d}}$ in ℓ_2 loss for α-smooth d-dimensional densities. Other nonparametric approaches for two-sample testing include Schilling (1986); Henze (1988); Liu and Modarres (2011); Biswas and Ghosh (2014); Wang et al. (2021).

The test statistics for nonparametric two-sample tests are usually constructed based on finite-sample surrogates to some metrics quantifying the discrepancy between the two populations, including the ℓ_p distance (Györfi and Van Der Meulen, 1991), Wasserstein distance (Ramdas et al., 2017) and Maximum Mean Discrepancy (MMD, Gretton et al., 2012a, 2009b; Li and Yuan, 2019). These metrics can all be embraced into a general family of discrepancy measures on distributions, called *adversarial losses*, which are also called *integral probability metrics* (IPM) in the probability literature, defined as

$$d_{\mathcal{F}}(p,q) = \sup_{f\in\mathcal{F}} \Big| \int_{\mathcal{X}} f(x)\,\mathrm{d}p(x) - \int_{\mathcal{X}} f(x)\,\mathrm{d}q(x) \Big|, \quad (1)$$

where $\mathcal{X} \subset \mathbb{R}^d$ denotes the data space, and $\mathcal{F}$ is the discriminator class composed of a subset of all Borel-measurable functions. Note that if the discriminator class satisfies $\mathcal{F} = -\mathcal{F}$, then it is not necessary to take the absolute value inside (1) . Different choices of $\mathcal{F}$ leads to different adversarial losses. However, except for some special cases such as when $\mathcal{F}$ is the unit ball of a reproducing kernel Hilbert space (RKHS) (Li and Yuan, 2019; Sriperumbudur et al., 2010) or the dimensionality d equals to one (Del Barrio et al., 1999), the adversarial loss in (1) lacks a closed-form expression. For practical computations, we need to numerically solve the optimization problem of maximizing the difference over the discriminator class.

In this work, we consider a broad class of adversarial losses indexed by a smoothness (level) parameter $\gamma \in [0, \infty)$, which are shown to equivalent to the negative Besov norms. Here, the negative Besov norm is the norm associated with a Besov space (Triebel, 2006, 2010) with negative exponent. Since the smoothness parameter γ can be interpreted as the weights that penalize the high-order wavelet coefficients (high frequency components) in the wavelet expansion of the difference between the two distributions of concern, we propose to approximate the population level negative Besov norm by truncating its empirical version (c.f. Lemma 2). By further normalizing this truncated finite-sample surrogate to the negative Besov norm, we define a set of test statistics that are asymptotically standard normal under the null hypothesis and tends to infinity in the presence of any significant difference between the two distributions. In addition, when the populations of concern have d-dimensional densities that are at least α-smooth, for some suitably chosen penalizing weights (i.e., the exponent of the negative Besov norm), the constructed tests can detect the distributional difference at our derived optimal separation rate $O\big(n^{-\frac{2(\alpha+\gamma)}{4\alpha+d}} + n^{-\frac{1}{2}}\big)$ up to inessential logarithmic terms, simultaneously under all adversarial losses with γ-smooth ($\gamma \in [0, \infty]$) discriminators, which includes the commonly-used ℓ_1, ℓ_2 distances and the 1-Wasserstein distance (Santambrogio, 2015; Villani, 2009) as special cases. The result also rigorously verifies conventional wisdom that testing is usually easier than estimation, as our derived rate for testing is smaller than the minimax rate for estimation $O(n^{-\frac{\alpha+\gamma}{2\alpha+d}} + n^{-\frac{1}{2}})$ (Uppal et al., 2019) under the same loss functions. Empirically, we compare our approach with the state-of-the-art nonparametric two-sample test based on MMD with Gaussian kernels; it turns out that our approach outperforms the Gaussian-MMD test in terms of both detection power and robustness to hyperparameters.

The rest of the paper is organized as follows. In Section 2, we give a brief introduction to the adversarial loss. In Section 3, we show the equivalence between a class of representative adversarial losses and the negative Besov norm. We also provide an empirical surrogate to the negative Besov norm based on finite samples in Section 4. In Section 5, we first derive the minimax rate of nonparametric two-sample testing under adversarial losses, and then propose a minimax-optimal test procedure based on the empirical surrogate from Section 4. Simulations and a real data application are included in Section 6 and 7.

2 Adversarial losses

Many recent machine learning studies employ the adversarial loss as an alternative to the conventional ℓ_p distances for characterizing the closeness between probability measures (Arjovsky et al., 2017; Tolstikhin et al., 2017). The adversarial loss defined in (1) can realize a large family of probability metrics by suitably choosing the discriminator class $\mathcal{F}$. We focus on the following adversarial losses where $\mathcal{F}$ is the unit ball within the Sobolev-2 class or Hölder class, denoted as $\mathcal{W}_1^\gamma(\Omega)$ and $\mathcal{C}_1^\gamma(\Omega)$ respectively, with smoothness level $\gamma \geq 0$ (the formal definition of $\mathcal{W}_1^\gamma(\Omega)$ and $\mathcal{C}_1^\gamma(\Omega)$ can be found in the supplementary material), for which the corresponding metrics are respectively denoted as $d_\gamma^W(\cdot,\cdot)$ and $d_\gamma^H(\cdot,\cdot)$.

Adversarial losses are more suitable to characterize discrepancies between nearly singular distributions, such as those arising from high-dimensional data with low-dimensional structures, than many conventional metrics including the ℓ_p distances due to their robustness against distribution perturbations. In particular, the metric d_γ^W (or d_γ^H) becomes stronger as γ decreases. By taking $\gamma = 0$, d_γ^W and d_γ^H are equivalent to the ℓ_2 and ℓ_1 distance, respectively; by taking $\gamma = 1$, the metric d_γ^H corresponds to the 1-Wasserstein distance (Santambrogio, 2015). The smoothness parameter γ controls the sensitivity of the metric to oscillations: a smaller γ makes d_γ^W, d_γ^H more sensitive to high frequency components of the density. For example, consider a d-dimensional random variable X with support lying close to a low dimensional submanifold. More specifically, we consider the following probabilistic model

$$p(X|z) = \mathcal{N}(G(z), \sigma^2 I_d), \qquad z \sim \mathcal{N}(0, I_{\bar{d}}), \tag{2}$$

with $d > \bar{d}$. Model (2) is commonly employed in generative modelling literature for learning data generators for the images of objects (Kingma and Welling, 2013; Doersch, 2016), where the latent variable z can be interpreted as (low dimensional) global characteristics such as camera projection, lighting condition, texture, object position and orientation. Suppose we translate the mean parameter of the conditional distribution of X along a direction inside the normal space of the underlying submanifold $\mathcal{M} = \{G(z) : z \in \mathbb{R}^{\bar{d}}\}$ at point $G(z)$ by a tiny amount $u > 0$, that is, we consider the following conditional distribution: $p(X'|z) = \mathcal{N}(G(z) + \omega(z) \cdot u, \sigma^2 I_d)$, where $\omega(z)$ is a unit vector perpendicular to the tangent space of $\mathcal{M}$ at $G(z)$ (see Figure 1 for an illustration). In this example, the Wasserstein distance between marginal distributions of X

and X' is of order $O(|u|)$ regardless of the order of σ; while the corresponding ℓ_p distance can be of order $O(1)$ given that $|u/\sigma| = O(1)$. This suggests that when used as a discrepancy measure for distribution estimation, the ℓ_p distance is much more sensitive to oscillations and support mismatching. As a consequence, the detection boundary under the ℓ_p distance will be extremely large for nearly singular distributions. This is because some "physically close" distributions that are difficult to distinguish may have large ℓ_p distances due to their supports not perfectly aligning with the support of the original data. On the other hand, the adversarial loss with smoother discriminator class honestly quantifies the amount of support mismatch and therefore more suitable for quantifying the discrepancy between nearly singular distributions.

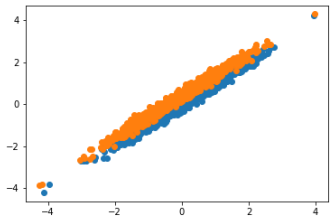


Figure 1: The figure shows the random samples from the distribution of X (Blue points) and X' (Orange points) with $d = 2$, $\bar{d} = 1$, $G(z) = (z, z)$, $u = 0.3$ and $\sigma = 0.1$. We can see that the shape and the location of the two scatter plots are quite similar, yet the ℓ_p distance is quite large due to the support mismatching.

3 Wavelet Transform and Besov Norm

The wavelet transform is a powerful exploratory data analysis tool that can efficiently represent signals with slowly varying trend and abrupt changes interrupting smooth regions. Roughly speaking, a wavelet is a rapidly decaying wave like oscillation that exists for a finite duration. Commonly-used wavelets includes Haar wavelet (Triebel, 2010), Meyer wavelet (Triebel, 2006; Meyer, 1992), Daubechies wavelet (Daubechies, 1988), etc. One of the key concepts in wavelet transform is the scaling, which refers to the process of stretching or shrinking the wavelet along the features. A stretched wavelet helps in capturing the slowly varying trends in a signal; while a shrinking wavelet helps in detecting the abrupt changes.

Concretely, let $\mathcal{L}^2(\mathbb{R}^d)$ denote the set of all square integrable functions on $\mathbb{R}^d$. It is possible to define a complete orthonormal basis $\{\overline{\Psi}_j\}_{j\geq 0}$ for $\mathcal{L}^2(\mathbb{R}^d)$ so that: the set of level zero basis $\overline{\Psi}_0$ is formed by shifting some compact scaling function; for any $j \in \mathbb{N}^+$, the set of level j basis $\overline{\Psi}_j$ is obtained by shifting some compact wavelet function and scaling it by a factor of $2^{-(j-1)}$; and any function $p \in \mathcal{L}^2(\mathbb{R}^d)$ can be uniquely expressed as

$$p(x) = \sum_{j=0}^{\infty} \sum_{\psi \in \overline{\Psi}_j} p_\psi \psi(x) \quad \text{with} \quad p_\psi = \int_{\mathbb{R}^d} \psi(x) p(x)\, \mathrm{d}x.$$

Further detail is included in Appendix B. As described above, for $\psi \in \Psi_j$, when the level j is small, the wavelet coefficient p_ψ can capture general trends of the function $p(\cdot)$; on the contrary, with a large level j, the wavelet coefficient p_ψ can capture abrupt changes/oscillations. Therefore, for a smooth function $p(\cdot)$ not containing large and abrupt oscillations, the wavelet coefficient p_ψ tends to be small for those ψ corresponding to a large level basis. To formally quantify such function oscillations, we can use the so-called Besov norm with exponent s, defined for a smooth level $s \in [0, \infty)$ and $l, m \in \mathbb{N}^+$ as follows:

$$\|p\|_{B^s_{l,m}} := \Big[\sum_{j=0}^{\infty} 2^{jm(s+\frac{d}{2}-\frac{d}{l})} \Big(\sum_{\psi \in \overline{\Psi}_j} |p_\psi|^l \Big)^{\frac{m}{l}} \Big]^{\frac{1}{m}}. \quad (3)$$

We can correspondingly define the Besov space $B^s_{l,m}(\mathbb{R}^d)$ as a subspace of $\mathcal{L}^2(\mathbb{R}^d)$ equipped with the norm $\|\cdot\|_{B^s_{l,m}}$. The Besov space is closely related to the Sobolev space: when $l = m = 2$, the Besov space $B^s_{2,2}(\mathbb{R}^d)$ is equivalent to the Sobolev-2 space $\mathcal{W}^s(\mathbb{R}^d)$; when $l = m = \infty$, the Besov space $B^s_{\infty,\infty}(\mathbb{R}^d)$ is equivalent to the Hölder space $\mathcal{C}^s(\mathbb{R}^d)$.

Apart from quantifying the smoothness level of a function, by allowing a negative exponent s inside (3), the Besov norm can be used to measure the difference of two functions. In particular, by choosing $l = m = 2$, $s = -\gamma$, we can obtain the following distance between two probability density functions $p, q \in \mathcal{L}^2(\mathbb{R}^d)$,

$$\|p - q\|_{B^{-\gamma}_{2,2}} = \Big[\sum_{j=0}^{\infty} 2^{-2j\gamma} \sum_{\psi \in \overline{\Psi}_j} (p_\psi - q_\psi)^2 \Big]^{\frac{1}{2}}. \quad (4)$$

We call the norm in (4) the negative Besov norm with exponent $\gamma \in [0, \infty)$. The decaying level γ controls the sensitivity of the metric to abrupt changes. By taking $\gamma = 0$ and $l = m = 2$, we attain the conventional ℓ_2 loss (i.e., $\big[\int_{\mathbb{R}^d} (p(x) - q(x))^2\, \mathrm{d}x\big]^{\frac{1}{2}}$). However, as described in Section 2, the ℓ_2 distance is sensitive to small wiggles/oscialltions, and may not be suitable for cases where we are also concerned about the slowly varying trends. On the other hand, by choosing a positive γ, the influence of high level wavelet coefficients (high frequency component) is controlled by the weight $2^{-2j\gamma}$. The following Lemma shows that the negative Besov norm is equivalent to the adversarial loss with the discriminator class being the Sobolev-2 space.

Lemma 1. *For probability density functions $p, q \in \mathcal{L}^2(\mathbb{R}^d)$ and $\gamma \geq 0$, we have*

$$c\, d^W_\gamma(p, q) \leq \|p - q\|_{B^{-\gamma}_{2,2}} \leq C\, d^W_\gamma(p, q),$$

where positive constants c and C only depend on γ, d.

4 Empirical Surrogate to Squared Negative Besov Norm

The adversarial loss, even though conceptual appealing, may suffer from lacking a closed-form expression for computations. According to Lemma 1, the adversarial loss d_γ^W is equivalent to the negative Besov norm $\|\cdot\|_{B_{2,2}^{-\gamma}}$ with exponent γ, up to some multiplicative constant. However, the negative Besov norm is a sum of an infinite series, and can not be computed in a finite number of operations. To this end, we restrict our attention to distributions supported on a bounded domain $\Omega \subset \mathbb{R}^d$ with smooth densities. Denote by $\mathcal{W}_L^{u,\alpha}(\Omega)$ the subset of αth order Sobolev-2 space $W_L^\alpha(\mathbb{R}^d)$ so that each function is uniformly bounded by L and supported on Ω, that is,

$$\mathcal{W}_L^{u,\alpha}(\Omega) = \big\{p \in \mathcal{W}_L^\alpha(\mathbb{R}^d) \,:\, \sup_{x\in\Omega}|p(x)| \le L, \\ \operatorname{supp}(p) \subset \Omega\big\}, \quad \alpha > 0.$$

Note that here the uniform boundness of the density function is only a technique artifact to simplify the proof, so that magnitudes of high-order (empirical) wavelet coefficients can be properly bounded; it also trivially holds for Hölder smooth density functions. The compactness of the support of the density is for ensuring that only finitely many wavelet coefficients are non-vanishing at a given scale. Consider densities $p, q \in \mathcal{W}_L^{u,\alpha}(\Omega)$, we further choose to truncate the wavelet expansion at a finite level J to attain a best accuracy versus efficiency trade-off, where J is an integer depending on the sample size and smoothness level α that will be chosen later. This leads to the following approximation to the squared negative Besov norm:

$$\|p-q\|^2_{\widehat{B}_{2,2}^{-\gamma}} = \sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi\in\Psi_j} (p_\psi - q_\psi)^2, \qquad (5)$$

where $\Psi_j = \{\psi \in \overline{\Psi}_j \,:\, \operatorname{supp}(\psi)\cap\Omega \neq \emptyset\}$ for $j \ge 0$.[1] Given the wavelet coefficients $\{p_\psi, q_\psi\}_{\psi\in\Psi_j, j\in\mathbb{N}}$, (5) can be computed in $O(2^{dJ})$ number of operations. In statistical applications, the wavelet coefficients are not directly computable, but instead two sets of i.i.d samples $X^{(n)} = \{X_1, \cdots, X_n\} \sim p$ and $Y^{(m)} = \{Y_1, \cdots, Y_m\} \sim q$ are available. Based on the definition $p_\psi = \mathbb{E}_p[\psi(X)]$ of the wavelet coefficient, we can estimate p_ψ by replacing the population level expectation with the empirical mean $\widehat{p}_\psi = n^{-1}\sum_{i=1}^n \psi(X_i)$. However, it is not hard to see that $\widehat{p}_\psi^2$ is a biased estimator of p_ψ^2. We then correct for the bias and use instead the U-statistic to approximate p_ψ^2, which leads to the following statistic that forms an unbiased estimator to (5),

$$T_{\gamma,J} = \sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi\in\Psi_j} \Big[\frac{1}{n(n-1)} \sum_{i_1\neq i_2} \psi(X_{i_1})\psi(X_{i_2}) + \\ \frac{1}{m(m-1)} \sum_{i_1\neq i_2} \psi(Y_{i_1})\psi(Y_{i_2}) - \frac{2}{nm}\sum_{i_1,i_2}\psi(X_{i_1})\psi(Y_{i_2})\Big]. \qquad (6)$$

For brevity, we consider the balanced case where $c \le n/m \le C$ for some constants $0 < c \le C < \infty$, and express explicitly only the dependence on n and not m in our theoretical results. For general situations, the rate will only depend on the minimum of n and m.

The next lemma shows that by choosing a suitable truncation level J, the statistic $T_{\gamma,J}$ is a valid estimator to the squared Besov norm in the sense that (1) under the case where the two distributions p, q are the same, $T_{\gamma,J}$ converges to zero in probability; (2) when the two distributions are sufficiently separated, the ratio of the statistic and squared negative Besov norm converges to one in probability.

Lemma 2. *For distributions $p, q \in \mathcal{W}_L^{u,\alpha}(\Omega)$. Suppose $c \le n/m \le C$ for some constants $0 < c \le C < \infty$. For any $\gamma \ge 0$, the statistic $T_{\gamma,J}$ with $J = \lceil \log_2(n^{\frac{2}{4\alpha+d}})\rceil$ satisfy the following properties:*

1. *If $p = q$, we have $T_{\gamma,J} \xrightarrow{P} 0$, where $\xrightarrow{P}$ means converging in probability;*
2. *If $d_\gamma^W(p,q)\cdot(n^{\frac{2(\alpha+\gamma)}{4\alpha+d}} \wedge n^{\frac{1}{2}})\cdot(\log n)^{-\frac{1}{2}} \to \infty$, then we have $\frac{T_{\gamma,J}}{\|p-q\|^2_{B_{2,2}^{-\gamma}}} \xrightarrow{P} 1$.*

The statistic $T_{\gamma,J}$ can then be deployed to construct a test statistic for the two-sample hypothesis testing, which we describe in detail in the following section.

A commonly-used metric in literature for measuring the discrepancy between two distributions p, q is the maximum mean discrepancy (MMD). With finite data, the squared MMD between p, q is commonly approximated by the U-statistic:

$$T_h^{\mathrm{MMD}} = \frac{1}{n(n-1)}\sum_{i_1\neq i_2} k_h(X_{i_1}, X_{i_2}) \\ + \frac{1}{m(m-1)}\sum_{i_1\neq i_2} k_h(Y_{i_1}, Y_{i_2}) - \frac{2}{nm}\sum_{i_1,i_2} k_h(X_{i_1}, Y_{i_2}), \qquad (7)$$

where $k_h(x,y) = k(\frac{x-y}{h})$ is a positive semi-definite kernel, typically chosen as a Gaussian kernel with bandwidth h. It has been shown that for properly chosen bandwidth, T_h^{MMD} can well approximate the squared ℓ_2 distance (Gretton et al., 2012a). Compared with T_h^{MMD}, the statistic $T_{\gamma,J}$ can represent a large class of metrics including the ℓ_2 distance by allowing different γ. Computationally, due to the compactness of the wavelet function, for any $j \in \mathbb{N}$ and $X \in \Omega$,

[1] We use $\operatorname{supp}(\psi)$ to denote the support of function ψ.

there are only a constant number of $\psi \in \Psi_j$ such that $\psi(X) \neq 0$. Therefore, the statistic $T_{\gamma,J}$ can be computed in $O(nJ + 2^{dJ})$ number of operations: we need $O(nJ)$ number of operations for obtaining estimators of wavelet coefficients, and $O(2^{dJ})$ number of operations to compute the truncated negative Besov norm given the wavelet coefficients. Plugging in the choice of $J = \lceil \log_2(n^{\frac{2}{4\alpha+d}}) \rceil$, the dependence of the computational complexity of $T_{\gamma,J}$ in n is $O(n \log n + n^{\frac{2d}{4\alpha+d}})$. While T_h^{MMD} requires $O(n^2)$ number of operations, which is larger than the number required by $T_{\gamma,J}$ for $\alpha > 0$. Moreover, by choosing a positive γ, the weights $2^{-2\gamma j}$ penalize coefficients at high levels and reduce the variance of the statistic. Therefore, $T_{\gamma,J}$ tends to be robust against the choice of sufficiently large J's. In comparison, T_h^{MMD} is known to be sensitive to the choice of bandwidth h: a bandwidth value which is too small leads to an estimation with small bias and large variance; while a large bandwidth leads to low variance at the expense of increased bias. It is also worth mentioning that when the populations of concern are α-smooth, the optimal choice of the bandwidth is given by $h = O(n^{-\frac{2}{4\alpha+d}})$ (Li and Yuan, 2019), which relates to the optimal choice of the truncation level J in our proposed statistic through $2^{-J} \asymp h$. Therefore, any rule for selecting the bandwidth in the MMD test with Gaussian kernel (e.g., the median heuristic, Arlot et al., 2019) can be deployed for selecting J. Additionally, by rearranging the order of summations in the statistic $T_{\gamma,J}$, it actually corresponds to a special MMD test statistic with a kernel constructed via wavelet truncation: $k_{\gamma,J}(x,y) = \sum_{j=0}^{J} \sum_{\psi \in \Psi_j} 2^{-2j\gamma} \psi(x)\psi(y)$. So we can adopt rules in literature of MMD tests for optimizing the choice of the kernel (Gretton et al., 2012b).

5 Minimax Nonparametric Two-Sample Test

The two-sample test is a statistical hypothesis test used to determine whether two independent samples $X^{(n)}$ and $Y^{(m)}$ come from a common population. Let p, q be two probability density functions in $\mathcal{W}_L^{u,\alpha}(\Omega)$. To better quantify the power of a two-sample test, We consider the null hypothesis $\mathbb{H}_0 : p = q$ and a local alternative hypothesis that is increasingly closer to the null as data accrue:

$$\mathbb{H}_1(\Delta_n; \mathcal{D}) : \mathcal{D}(p,q) \geq \Delta_n,$$

where $\mathcal{D}$ is some discrepancy measure. For a test Φ based on data $X^{(n)}$, we can define its power as

$$\mathrm{power}\big(\Phi; \mathbb{H}_1(\Delta_n; \mathcal{D})\big) := \inf_{\substack{p,q \in \mathcal{W}_L^{u,\alpha}(\Omega) \\ \mathcal{D}(p,q) \geq \Delta_n}} P(\Phi \text{ reject } \mathbb{H}_0). \tag{8}$$

Of particular interest here is the smallest separation Δ_n from the null hypothesis that can be detected consistently in a minimax sense. We care about metrics $d_{\gamma_1}^W$ and $d_{\gamma_1}^H$ with $\gamma_1 \in [0, \infty)$. Note that by Sobolev embedding theorem (Adams and Fournier, 2003), $C_L^{\gamma_1}(\Omega) \subset \mathcal{W}_{L_1}^{\gamma_1}(\Omega)$, and therefore $d_{\gamma_1}^H(p, p_0) \lesssim d_{\gamma_1}^W(p, p_0)$, which leads to

$$\mathrm{power}\big(\Phi; \mathbb{H}_1(\Delta_n; d_{\gamma_1}^W)\big) \leq \mathrm{power}\big(\Phi; \mathbb{H}_1(c\,\Delta_n; d_{\gamma_1}^H)\big), \tag{9}$$

for some constant c depend on Ω. We first provide a lower bound to the optimal detection boundary (or separation threshold) when $\mathcal{D}(\cdot,\cdot)$ is chosen to be the adversarial loss $d_{\gamma_1}^H$ with Hölder smooth discriminators. Here again for the sake of simplicity, we consider the balanced case where $c \leq n/m \leq C$ for some constants $0 < c \leq C < \infty$ and express explicitly only the dependence on n.

Theorem 1. *For any $\gamma_1 \geq 0$, if $\Delta_n = o\big(n^{-\frac{2(\alpha+\gamma_1)}{4\alpha+d}} \vee \frac{1}{\sqrt{n}}\big)$[2], then there exists some $\eta \in (0,1)$ so that for any test Φ_n based on data $X^{(n)}$ and $Y^{(m)}$ that has asymptotic significance level η, i.e., $\lim_{n\to\infty} P(\Phi_n \text{ reject } \mathbb{H}_0) = \eta$ for any $p = q \in \mathcal{W}_L^{u,\alpha}(\Omega)$, we have*

$$\liminf_{n\to\infty} \mathrm{power}\big(\Phi_n; \mathbb{H}_1(\Delta_n; d_{\gamma_1}^H)\big) < 1.$$

A similar result holds when the discrepancy measure is chosen to be $d_{\gamma_1}^W$ (recall inequality (9)). Now we demonstrate that, with a proper choice of γ and J, we can construct a test statistic based on $T_{\gamma,J}$ that simultaneously attains the optimal detection boundary (up to logarithmic term) for all the $d_{\gamma_1}^H$ and $d_{\gamma_1}^W$ metrics with γ_1 ranging over $[0, \infty)$.

Given a specified significance level η, to obtain an asymptotic η-level test, we may proceed to reject $\mathbb{H}_0$ if and only if $T_{\gamma,J}$ exceeds the η-upper quantile of its asymptotic distribution under $\mathbb{H}_0$. However, the asymptotic distribution of $T_{\gamma,J}$ remains unknown. To obtain a "normalized" test statistic that is asymptotically standard normal under $\mathbb{H}_0$, we should estimate the variance of $T_{\gamma,J}$. Denote $r_{n,m} = \frac{2}{n(n-1)} + \frac{2}{m(m-1)} + \frac{4}{mn}$, a simple calculation yields that, under $\mathbb{H}_0$,

$$\begin{aligned}\mathrm{Var}(T_{\gamma,J}) = r_{nm} \cdot \mathbb{E}\Big[&\Big(\sum_{j=0}^{J} \sum_{\psi \in \Psi_j} 2^{-2j\gamma} \psi(X)\psi(Y) \Big)^2 \\ &- \Big(\sum_{j=0}^{J} \sum_{\psi \in \Psi_j} 2^{-2j\gamma} \psi(X) \cdot q_\psi \Big)^2 - \\ &\Big(\sum_{j=0}^{J} \sum_{\psi \in \Psi_j} 2^{-2j\gamma} \psi(Y) \cdot p_\psi \Big)^2 + \Big(\sum_{j=0}^{J} \sum_{\psi \in \Psi_j} 2^{-2j\gamma} p_\psi q_\psi \Big)^2 \Big].\end{aligned} \tag{10}$$

Note that the last term in (10) is a higher-order term, so we only need to estimate the first three terms. To this end, we replace the population means with the empirical means and approximate the wavelet coefficients by their sample

[2] $a_n = o(b_n)$ if $a_n/b_n \to 0$ as n increases.

versions, which leads to

$$\widehat{\mathscr{S}}^2_{\gamma,J} = r_{nm}\cdot\Big\{\frac{1}{nm}\sum_{i_1,i_2}\Big(\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(X_{i_1})\psi(Y_{i_2})\Big)^2 - \frac{1}{n}\sum_{i_1=1}^{n}\Big[\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(X_{i_1})\cdot\big(\frac{1}{m}\sum_{i_2=1}^{m}\psi(Y_{i_2})\big)\Big]^2 - \frac{1}{m}\sum_{i_2=1}^{m}\Big[\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(Y_{i_2})\cdot\big(\frac{1}{n}\sum_{i_1=1}^{n}\psi(X_{i_1})\big)\Big]^2\Big\}.$$

However, we need to avoid a negative or zero estimate of the variance. To this end, we replace $\widehat{\mathscr{S}}^2_{\gamma,J}$ with a small value $1/n^3$ whenever it is too small or negative. Namely, let $\widetilde{\mathscr{S}}^2_{\gamma,J} = \max(\widehat{\mathscr{S}}^2_{\gamma,J}, \frac{1}{n^3})$. Now we can define the following "normalized" test statistic:

$$\widetilde{T}_{\gamma,J} = \widetilde{\mathscr{S}}^{-1}_{\gamma,J} T_{\gamma,J}.$$

The following theorem summarizes our main results on the validity and power of the test induced from the test statistic $\widetilde{T}_{\gamma,J}$.

Theorem 2. *For $p, q \in \mathcal{W}^{u,\alpha}_L(\Omega)$, consider test statistic $\widetilde{T}_{\gamma,J}$ with $2^J \asymp n^{\frac{2}{4\alpha+d}}$ and $0 \le \gamma \le \frac{d}{4}$,*

1. *under $\mathbb{H}_0$, we have $\widetilde{T}_{\gamma,J} \xrightarrow{d} N(0,1)$, where $\xrightarrow{d}$ means converging in distribution;*
2. *for any significance level $\eta \in (0,1)$, consider test $\Phi^J_{\gamma,\eta}$ where $\mathbb{H}_0$ is rejected if and only if $\widetilde{T}_{\gamma,J}$ exceeds z_η, the upper η-quantile of the standard normal distribution (i.e., $P(Z \ge z_\eta) = \eta$ with $Z \in N(0,1)$). Let*

$$\delta_n(\gamma_1) = \begin{cases} n^{-\frac{4\alpha+4(\gamma\wedge\gamma_1)}{4\alpha+d}}, & 0 < \gamma < \frac{d}{4} \\ \log n \cdot n^{-\frac{4\alpha}{4\alpha+d}}, & \gamma = 0 \\ \log n \cdot n^{-1+\frac{4(\gamma-\gamma\wedge\gamma_1)}{4\alpha+d}}, & \gamma = \frac{d}{4} \end{cases} \tag{11}$$

then for any $\gamma_1 \ge 0$ and Δ_n satisfies $\Delta_n^2 \cdot \delta_n(\gamma_1)^{-1} \to \infty$, we have (1) $\widetilde{T}_{\gamma,J} \xrightarrow{P} +\infty$ given that $d^W_{\gamma_1}(p,q) \ge \Delta_n$; (2) the power defined in (8) *satisfies*

$$\lim_{n\to\infty}\mathrm{power}\big(\Phi^J_{\gamma,\eta}; \mathbb{H}_1(\Delta_n; d^W_{\gamma_1})\big) = 1.$$

Theorem 2 suggests that (1) the test $\Phi^J_{\gamma,\eta}$ has asymptotic level η; (2) for any $0 < \gamma_1 < \frac{d}{4}$, by choosing $\gamma_1 \le \gamma < \frac{d}{4}$, the test $\Phi^J_{\gamma,\eta}$ can attain the optimal detection boundary $n^{-\frac{2(\alpha+\gamma_1)}{4\alpha+d}} \vee \frac{1}{\sqrt{n}}$ under both $d^W_{\gamma_1}$ and $d^H_{\gamma_1}$ metrics (recall inequality (9)); the case for $\gamma_1 \ge \frac{d}{4}$ or $\gamma_1 = 0$ only introduces an extra logarithmic term. In particular, by taking $\gamma = \frac{d}{4}$, the test $\Phi^J_{\gamma,\eta}$ can simultaneously attain the optimal detection boundary up to a logarithmic term under $d^W_{\gamma_1}$ and $d^H_{\gamma_1}$ metrics with γ_1 ranging over $[0,\infty)$.

Corollary 1. *Consider test statistics $\widetilde{T}_{\gamma,J}$ with $\gamma = \frac{d}{4}$ and $2^J \asymp n^{\frac{2}{4\alpha+d}}$. For any $\gamma_1 \ge 0$, denote $\Delta_n = (\log n)\cdot\big(n^{-\frac{2(\alpha+\gamma_1)}{4\alpha+d}} \vee \frac{1}{\sqrt{n}}\big)$, then for any significance level $\eta \in (0,1)$ the associated test $\Phi^J_{\frac{d}{4},\eta}$ satisfies that*

$$\lim_{n\to\infty}\mathrm{power}\big(\Phi^J_{\frac{d}{4},\eta}; \mathbb{H}_1(\Delta_n; d^W_{\gamma_1})\big) = 1.$$

In practice, rather than estimating the variance of $T_{\gamma,J}$, we can also estimate the testing threshold by bootstrap methods (Arcones and Gine, 1992; Efron, 1979) as the bootstrap threshold may be more accurate for small samples: we compute the statistics $T_{\gamma,J}$ based on datasets randomly sampled from the joint sample $\{X^{(n)}, Y^{(m)}\}$, and then we evaluate the upper η-quantile of the empirical distribution of $T_{\gamma,J}$ based on the bootstrapping datasets.

Remark 1. *The reason for considering Sobolev discriminators in constructing the test statistic is that, the norm $B^{-\gamma}_{2,2}$ associated with d^W_γ has a nice squared form, which enables us to utilize techniques from U-statistics for computation and theoretical analysis. Since we develop a matching lower bound for Hölder smooth discriminators, our result can be generalized to any adversarial loss with the discriminator class being an interpolation space between Sobolev-2 and Hölder space. Moreover, for $\gamma > d/2$, the Sobolev-2 space $B^\gamma_{2,2}(\mathbb{R}^d)$ coincides with the reproducing kernel Hilbert space (RKHS) generated by the Matérn kernel of order $\gamma - d/2$ (Kanagawa et al., 2018), so the proposed statistics $T_{\gamma,J}$ would be asymptotically equivalent to the Matérn kernel based MMD statistic with bandwidth $h \asymp 2^{-J}$. However, the most interesting case lies in $\gamma \le \frac{d}{4}$, as the optimal detection boundary is the parametric root-n rate when $\gamma > \frac{d}{4}$. So increasing γ above $\frac{d}{4}$ leads to a weaker loss but has no improvement in the optimal detection rate.*

Remark 2. *Another closely related problem is the goodness-of-fit test. The goodness-of-fit test is a statistical hypothesis test used to determine whether the sample data fits a specified distribution p_0 from an expected population (e.g. a population with a normal distribution). The two-sample test can also be used to do a goodness-of-fit test: a random sample $Z^{(m)}$ is first drawn from the known reference distribution p_0 and then a two–sample test is performed on data sets $X^{(n)}$ and $Z^{(m)}$. We can construct a minimax optimal nonparametric goodness-of-fit test in a similar way as the two-sample test, where the extra prior information of p_0 is incorporated. Further detail is available in Appendix C.*

Remark 3. *The assumption about the compactness of the support of p, q can be relaxed. For example, if p, q have exponential decay tails, then we can consider $\Omega = [-c\log n, c\log n]$ so that the probability mass outside of Ω is a negligible higher-order term, and it will only introduce extra logarithmic terms in the detection power.*

6 Numerical Illustration

In this section, we aim at: (1) verifying empirically that the proposed test $\Phi^J_{\gamma,\eta}$ has asymptotic significance level η; (2) evaluating the power of the proposed test and comparing it with the MMD test. We carry out our experiment using two hypothesis testing procedures: one is the test $\Phi^J_{\gamma,\eta}$ with $\gamma = \frac{d}{4}$ as suggested in Theorem 2 and Corollary 1, where the wavelet basis is chosen to be the Haar wavelet[3], the other one is the hypothesis test based on MMD with Gaussian kernel $k_h(x,y) = \exp\big(-(\|x-y\|/h)^2\big)$ (Li and Yuan, 2019): $\mathbb{H}_0$ is rejected if and only if $\widetilde{T}^{\mathrm{MMD}}_h = T^{\mathrm{MMD}}_h / \widehat{S}_{\mathrm{MMD}}$ exceeds z_η, where T^{MMD}_h is defined in (7), and $\widehat{S}_{\mathrm{MMD}}$ is an estimator to the standard variance of T^{MMD}_h for normalizing the statistic, we denote the corresponding test by $\Phi^{\mathrm{MMD}}_{h,\eta}$. We then apply the two procedures to synthetic datasets. Specifically, let μ_0 be the uniform distribution on $[0,1]^d$ and $\mu_1(x) = \prod_{j=1}^d \nu(x_j)$ be a d-dimensional distribution with ν being beta distribution with shape parameters $\alpha = 2.5$, $\beta = 2.5$. Then we scale and translate the random variable $X \sim \mu_1$ by the transform $Y = X/k + 0.5$ with $k \in \mathbb{R}^+$, the corresponding distribution of Y is denote by $\mu_1^{[k]}$. We set $p = \frac{1}{2}\mu_0 + \frac{1}{2}\mu_1^{[5]}$ and $q = \frac{1}{2}\mu_0 + \frac{1}{2}\mu_1^{[k]}$ with $k \in \{3, 3.5, 3.8, 4, 5\}$.

We first check the normality of the test statistic $\widetilde{T}_{\gamma,J}$ and $\widetilde{T}^{\mathrm{MMD}}_h$ under $\mathbb{H}_0$. With $n = 50$, we independently sample $2n$ data from p, and split it into two equally-sized data sets $X^{(n)}$ and $Y^{(n)}$. The bandwidth h and truncation level J are selected based on the median heuristic (Arlot et al., 2019; Garreau et al., 2017): define $H_n = \mathrm{Median}(\|Z_i - Z_j\|^2 \,|\, 1 \le i \le j \le 2n)$, where $Z = \{Z_1, Z_2, \cdots, Z_{2n}\} = \{X^{(n)}, Y^{(n)}\}$. Following (Garreau et al., 2017), we set $h = \sqrt{H_n/2}$, and similarly choose $J = \lceil \log_2(1/\sqrt{H_n}) + 1 \rceil$. The corresponding bandwidth is around $h = 0.25$, and truncation level is around $J = 3$. The density and normal quantile-quantile (Q-Q) plots of the test statistics of concern based on 1000 replicates are shown in figure 2(a) and 2(b). We can see the density for $\widetilde{T}_{\gamma,J}$ under $\mathbb{H}_0$ is closer to the standard normal. In addition, the 1-Wasserstein distance between the distribution of the test statistics (under $\mathbb{H}_0$) and standard normal is 0.2538 for $\widetilde{T}^{\mathrm{MMD}}_h$, and 0.2138 for $\widetilde{T}_{\gamma,J}$. Therefore, our method delivered better uncertainty quantification.

To assess the power of the proposed testing procedure. We sample $n = 50$ samples from p and $q = \frac{1}{2}\mu_0 + \frac{1}{2}\mu_1^{[k]}$ with $k \in \{3, 3.5, 3.8, 4\}$ respectively. The selected bandwidth based on the median heuristic is also around $h = 0.25$, and the truncation level is around $J = 3$. The densities of the two test statistics of concern are given in Figure 2(c). We can see the density of $\widetilde{T}_{\gamma,J}$ has a much heavier tail, that is, $\widetilde{T}_{\gamma,J}$ tends to return a larger value that leads to the rejection of the null hypothesis. Moreover, we consider the powers (i.e., the probability that success to reject the null hypothesis) of the tests $\Phi^J_{\gamma,\eta}$ and $\Phi^{\mathrm{MMD}}_{h,\eta}$ with level of significance $\eta \in \{0.05, 0.01\}$. The results are shown in Table 1, from which we can see that the test $\Phi^J_{\gamma,\eta}$ has much larger powers than $\Phi^{\mathrm{MMD}}_{h,\eta}$. Now we study the sensitivity of the hypothesis

	Powers: $\Phi^J_{\gamma,\eta} \uparrow$	
k	$\eta = 0.01$	$\eta = 0.05$
4	0.307 ± 0.016	0.416 ± 0.016
3.8	0.434 ± 0.020	0.548 ± 0.014
3.5	0.602 ± 0.013	0.711 ± 0.011
3	0.822 ± 0.012	0.891 ± 0.010
	Powers: $\Phi^{\mathrm{MMD}}_{h,\eta} \uparrow$	
k	$\eta = 0.01$	$\eta = 0.05$
4	0.085 ± 0.011	0.163 ± 0.012
3.8	0.134 ± 0.008	0.246 ± 0.013
3.5	0.283 ± 0.013	0.440 ± 0.018
3	0.716 ± 0.009	0.837 ± 0.008

Table 1: Under $p = \frac{1}{2}\mu_0 + \frac{1}{2}\mu_1^{[5]}$ and $q = \frac{1}{2}\mu_0 + \frac{1}{2}\mu_1^{[k]}$ with $k \in \{3, 3.5, 3.8, 4\}$, the table shows the powers for tests $\Phi^J_{\gamma,\eta}$ and $\Phi^{\mathrm{MMD}}_{h,\eta}$ with $\gamma = 1/2$, $J = 3$, $h = 0.25$ and level of significance $\eta \in \{0.05, 0.01\}$.

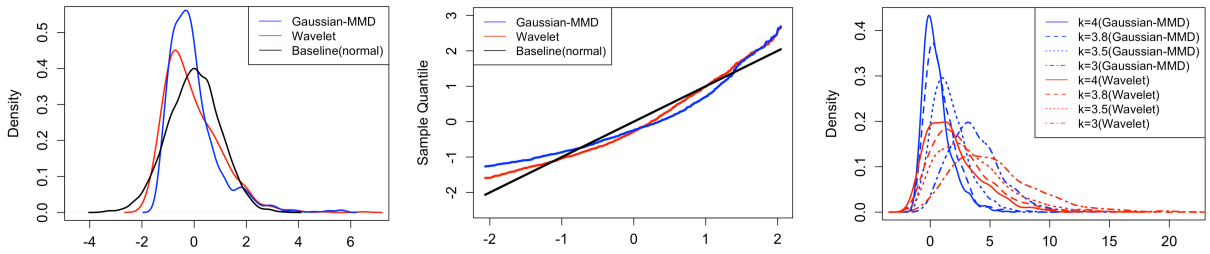
testing procedures to the hyperparameter. We also include in comparison the hypothesis testing procedure $\Phi^J_\eta = \Phi^J_{0,\eta}$, which means that we do not include the decaying factor $2^{-2\gamma j}$ for the level j of the wavelet in the proposed test statistic. Figure 3 shows the trends of the powers as the level of significance varies for different methods and hyperparameters. We can see that the tests Φ^J_η and $\Phi^{\mathrm{MMD}}_{h,\eta}$ exhibit similar patterns: both increasing and decreasing J (or h) from the optimal level would lead to an obvious deteriorate in the performance. On the other hand, the test $\Phi^J_{\gamma,\eta}$ with $\gamma = \frac{d}{4}$ is much more robust to large truncation levels: choosing an arbitrary J from $\{3, 4 \cdots, 7\}$ outperforms the MMD test $\Phi^{\mathrm{MMD}}_{h,\eta}$ with the bandwidth h being selected through the median heuristic (i.e., $h = 0.25$).

7 Application

As shown in Lemmas 1 and 2, the statistic $T_{\gamma,J}$ provides a reasonable metric for quantifying the distance between the underlying populations based on finite samples. Therefore, $T_{\gamma,J}$ can be deployed in practical problems as an evaluation criterion for checking the goodness of model fit.

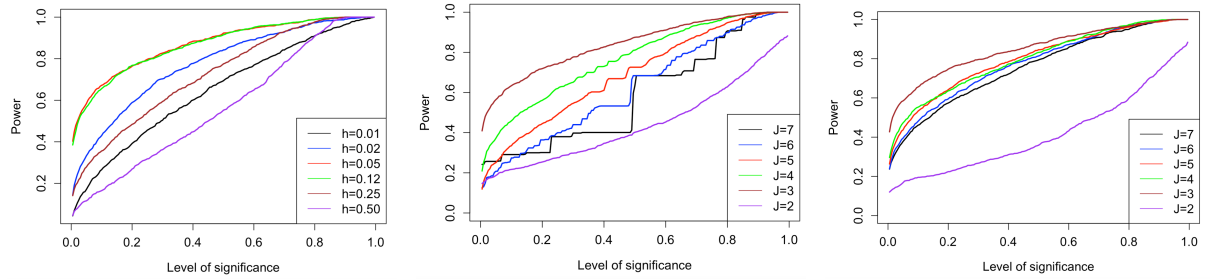
We consider the MNIST handwritten digit (LeCun et al., 1995) dataset, which is composed of 60k grey-scale images of handwritten digits $(0-9)$, along with a test set of 10k images. A popular method for modelling and efficient sampling from the complex distribution $\mathcal{P}^*$ over the handwritten digit is the variational autoencoder (VAE,

[3]The Haar wavelet's (mother) wavelet function is described as $\phi_{\mathfrak{M}}(x) = \mathbf{1}(0 \le x < 1/2) - \mathbf{1}(1/2 \le x < 1)$, and its scaling function can be described as $\phi_{\mathfrak{F}}(x) = \mathbf{1}(0 \le x < 1)$.



(a) Density plot: $q = p = \frac{1}{2}\mu_0 + \frac{1}{2}\mu_1^{[5]}$ (b) Q-Q plot: $q = p = \frac{1}{2}\mu_0 + \frac{1}{2}\mu_1^{[5]}$ (c) Density plot: $p = \frac{1}{2}\mu_0 + \frac{1}{2}\mu_1^{[5]}$, $q = \frac{1}{2}\mu_0 + \frac{1}{2}\mu_1^{[k]}$

Figure 2: Densities and Normal quantile-quantile (Q-Q) plots of test statistics: blue curves correspond to $\widetilde{T}_h^{\mathrm{MMD}}$ with $h = 0.25$, red curves correspond to $\widetilde{T}_{\gamma,J}$ with $J = 3$ and $\gamma = 1/2$, black curve corresponds to the baseline of standard normal.



(a) Test $\Phi_{h,\eta}^{\mathrm{MMD}}$ (b) Test Φ_η^J (c) Test $\Phi_{\gamma,\eta}^J$

Figure 3: For $p = \frac{1}{2}\mu_0 + \frac{1}{2}\mu_1^{[5]}$ and $q = \frac{1}{2}\mu_0 + \frac{1}{2}\mu_1^{[k]}$ with $k = 3.8$. The figure illustrates the trends of powers as the level η of significance varies for (1) test $\Phi_{h,\eta}^{\mathrm{MMD}}$ under different choices of bandwidth h; (2) test $\Phi_\eta^J = \Phi_{0,\eta}^J$ with different choices of truncation level J; (3) test $\Phi_{\gamma,\eta}^J$ with $\gamma = \frac{d}{4}$ and different choices of truncation level J. The cases for $k = 3, 3.5, 4$ exhibit similar trends.

Kingma and Welling, 2013; Rezende et al., 2014). In plain language, VAE is a latent variable generative modelling approach that defines a joint density $p(x, z)$ over the data space $\mathcal{X} \subset \mathbb{R}^D$ and the latent space $\mathcal{Z} \subset \mathbb{R}^d$ by specifying a prior $\pi(z)$ over latent variables and a conditional density (decoder) $p(x|z)$ of data given latent variables. To avoid marginalizing out latent variables, VAE introduces a family of encoders $q(z|x)$ for approximating the posterior of latent variables and jointly optimizing the so-called evidence lower bound (ELBO, Ormerod and Wand, 2010). The commonly-used choice of $\pi(z)$ is the isotropic Gaussian distribution. In this experiment, we consider jointly optimizing the prior inside a mixture of Gaussian family $\{\pi(z) = \frac{1}{K}\sum_{j=1}^K N(\mu_j, \sigma_j^2 I_d) \,|\, \mu_j \in \mathbb{R}, \sigma_j \in \mathbb{R}^+\}$ (Jiang et al., 2016; Tomczak and Welling, 2018). Of particular interest here is the choice of the hyperparameter $K \in \mathbb{N}^+$: whether uses a mixture of Gaussian can outperform the standard Gaussian and what is the optimal choice of K. To this end, let the latent dimension $d = 2$ and denote the fitted encoder and prior based on the training set as $\hat{q}_{[K]}(z|x)$ and $\widehat{\pi}_{[K]}(z)$, respectively. Here the encoder and prior are modelled by neural networks, details are available in Appendix A. A good choice of prior family will result in a small distance between the marginal of the learned encoder $\mathbb{E}_{\mathcal{P}^*}[\widehat{q}_{[K]}(z|x)]$ and the prior $\widehat{\pi}_{[K]}(z)$. Therefore, we sample 10k i.i.d data from $\mathbb{E}_{\mathcal{P}^*}[\widehat{q}_{[K]}(z|x)]$ by first randomly pick a data point x from the test set and sample z from $\widehat{q}_{[K]}(z|x)$, the obtaining data set is denote by $Z_1^{(n)}$ with $n = 10$k. Similarly, we sample data set $Z_2^{(n)}$ from $\widehat{\pi}_{[K]}(z)$. We record the value of $T_{\gamma,J}$ with $\gamma = \frac{d}{4}$ and $J = 8$ for different choices of K. As a comparison, we record the values of the negative test marginal log-likelihood (LL) (Burda et al., 2015; Tomczak and Welling, 2018), which is a commonly-used metric for quantitatively evaluating the VAE model. The results are shown in Figure 4.

According to the plot, the statistic $T_{\gamma,J}$ decreases rapidly when K increases from 1 to 10. As we can see in Figure 5, when $K = 1$, where the fitted prior is a single mode normal distribution, the marginal of the fitted encoder has an obvious clustering structure. In contrast, when $K = 10$, the fitted prior and marginal of the fitted encoder has a similar clustering structure. In addition, the trend of $T_{\gamma,J}$ approaches a horizontal line when $K \geq 10$. This is consis-

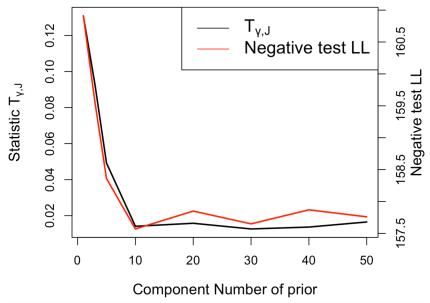
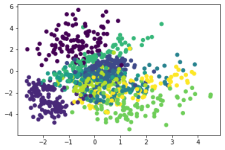
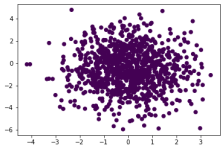


Figure 4: The statistic $T_{\gamma,J}$ (black curve) and negative test marginal log-likelihood (red curve) as K varies. For both metrics, a smaller value implies better performance.

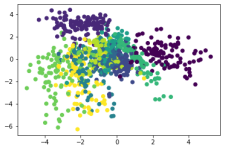
tent with the fact that the dataset consists of 10 digits '0' to '9'. Interestingly, for this dataset our method correctly identified/learned the number of clusters, a key clustering tuning parameter usually set from domain knowledge. The trend of the negative test LL exhibits similar pattern as the statistic $T_{\gamma,J}$, while it is more computational demanding: we need 170s for computing the test LL using an NVIDIA-A100 GPU, while the statistic $T_{\gamma,J}$ can be computed in 9s.



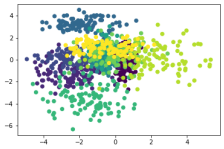
(a) Marginal of the fitted encoder ($K = 1$)



(b) Fitted prior ($K = 1$)



(c) Marginal of the fitted encoder ($K = 10$)



(d) Fitted prior ($K = 10$)

Figure 5: Random samples from the marginal of the fitted encoder and priors for different K.

8 Conclusion

In this paper, we propose a minimax nonparametric two-sample test that can simultaneously attain the optimal detection boundary under many common adversarial losses. We conducted experiments to show that in comparison to the conventional MMD test with Gaussian kernel, the proposed testing procedure tends to exhibit higher power and robustness against tuning parameters. In our theoretical analysis, the optimal choice of the truncation level J depends on the smoothness α of the underlying population, which may be unknown in practical problems. The development of a data-driven adaptive test to the distribution smoothness level may be left to future research.

References

Adams, R. A. and Fournier, J. J. (2003) *Sobolev spaces*. Elsevier.

Arcones, M. A. and Gine, E. (1992) On the bootstrap of u and v statistics. *The Annals of Statistics*, 655–674.

Arjovsky, M., Chintala, S. and Bottou, L. (2017) Wasserstein gan. URL: https://arxiv.org/abs/1701.07875.

Arlot, S., Celisse, A. and Harchaoui, Z. (2019) A kernel multiple change-point algorithm via model selection. *Journal of Machine Learning Research*, **20**, 1–56. URL: http://jmlr.org/papers/v20/16-155.html.

Biswas, M. and Ghosh, A. K. (2014) A nonparametric two-sample test applicable to high dimensional data. *Journal of Multivariate Analysis*, **123**, 160–171. URL: https://www.sciencedirect.com/science/article/pii/S0047259X13001966.

Bouzebda, S. and Didi, S. (2017) Multivariate wavelet density and regression estimators for stationary and ergodic discrete time processes: Asymptotic results. *Communications in Statistics - Theory and Methods*, **46**, 1367–1406. URL: https://doi.org/10.1080/03610926.2015.1019144.

Burda, Y., Grosse, R. and Salakhutdinov, R. (2015) Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.

Butucea, C. and Tribouley, K. (2006) Nonparametric homogeneity tests. *Journal of Statistical Planning and Inference*, **136**, 597–639. URL: https://www.sciencedirect.com/science/article/pii/S0378375804003374.

Cohen, A. (2003) *Numerical analysis of wavelet methods*. Elsevier.

Daubechies, I. (1988) Orthonormal bases of compactly supported wavelets. *Communications on pure and applied mathematics*, **41**, 909–996.

Del Barrio, E., Cuesta-Albertos, J. A., Matrán, C. and Rodríguez-Rodríguez, J. M. (1999) Tests of goodness of fit based on the l2-wasserstein distance. *Annals of Statistics*, 1230–1239.

Doersch, C. (2016) Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*.

Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, **7**, 1 – 26. URL: https://doi.org/10.1214/aos/1176344552.

Evans, L. C. (2010) *Partial differential equations*. Providence, R.I.: American Mathematical Society.

Friedman, J. H. and Rafsky, L. C. (1979) Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests. *The Annals of Statistics*, **7**, 697 – 717. URL: https://doi.org/10.1214/aos/1176344722.

Garreau, D., Jitkrittum, W. and Kanagawa, M. (2017) Large sample analysis of the median heuristic. URL: https://arxiv.org/abs/1707.07269.

Giné, E. and Nickl, R. (2015) *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

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

Gretton, A., Fukumizu, K., Harchaoui, Z. and Sriperumbudur, B. K. (2009a) A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems* (eds. Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams and A. Culotta), vol. 22. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2009/file/9246444d94f081e3549803b928260f56-Paper.pdf.

— (2009b) A fast, consistent kernel two-sample test. *Advances in neural information processing systems*, **22**.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K. and Sriperumbudur, B. K. (2012b) Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems* (eds. F. Pereira, C. Burges, L. Bottou and K. Weinberger), vol. 25. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2012/file/dbe272bab69f8e13f14b405e038deb64-Paper.pdf.

Györfi, L. and Van Der Meulen, E. C. (1991) *A Consistent Goodness of Fit Test Based on the Total Variation Distance*, 631–645. Dordrecht: Springer Netherlands. URL: https://doi.org/10.1007/978-94-011-3222-0_47.

Hall, P. (1984) Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of Multivariate Analysis*, **14**, 1–16. URL: https://www.sciencedirect.com/science/article/pii/0047259X84900447.

Hall, P. and Heyde, C. (1980) *Martingale Limit Theory and its Application*. Academic Press.

Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (2012) *Wavelets, approximation, and statistical applications*, vol. 129. Springer Science & Business Media.

Henze, N. (1988) A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, **16**, 772–783.

Hotelling, H. (1931) The Generalization of Student's Ratio. *The Annals of Mathematical Statistics*, **2**, 360 – 378. URL: https://doi.org/10.1214/aoms/1177732979.

Hütter, J.-C. and Rigollet, P. (2021) Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, **49**, 1166–1194.

Ingster, Y. I. (1986) An asymptotic minimax testing of nonparametric hypotheses on the density of the distribution of an independent sample. *Journal of Soviet Mathematics*.

— (1987) Minimax testing of nonparametric hypotheses on a distribution density in the l_p metrics. *Theory of Probability & Its Applications*, **31**, 333–337. URL: https://doi.org/10.1137/1131042.

Jiang, Z., Zheng, Y., Tan, H., Tang, B. and Zhou, H. (2016) Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148*.

Kanagawa, M., Hennig, P., Sejdinovic, D. and Sriperumbudur, B. K. (2018) Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*.

Kingma, D. P. and Welling, M. (2013) Auto-encoding variational bayes. URL: https://arxiv.org/abs/1312.6114.

LeCun, Y., Jackel, L. D., Bottou, L., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Muller, U. A., Sackinger, E., Simard, P. et al. (1995) Learning algorithms for classification: A comparison on handwritten digit recognition. *Neural networks: the statistical mechanics perspective*, **261**, 2.

Li, T. and Yuan, M. (2019) On the optimality of gaussian kernel based nonparametric tests against smooth alternatives. *arXiv preprint arXiv:1909.03302*.

Liu, Z. and Modarres, R. (2011) A triangle test for equality of distribution functions in high dimensions. *Journal of Nonparametric Statistics*, **23**, 605–615. URL: https://doi.org/10.1080/10485252.2010.485644.

Meyer, Y. (1992) *Wavelets and Operators: Volume 1*. No. 37. Cambridge university press.

Ormerod, J. T. and Wand, M. P. (2010) Explaining variational approximations. *The American Statistician*, **64**, 140–153.

Pearson, K. (1900) X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably

supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **50**, 157–175.

Ramdas, A., García Trillos, N. and Cuturi, M. (2017) On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, **19**, 47.

Rezende, D. J., Mohamed, S. and Wierstra, D. (2014) Stochastic backpropagation and approximate inference in deep generative models.

Santambrogio, F. (2015) *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Cham: Springer International Publishing.

Schilling, M. F. (1986) Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, **81**, 799–806.

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

Student (1908) The probable error of a mean. *Biometrika*, **6**, 1–25. URL: http://www.jstor.org/stable/2331554.

Tolstikhin, I., Bousquet, O., Gelly, S. and Schoelkopf, B. (2017) Wasserstein auto-encoders. URL: https://arxiv.org/abs/1711.01558.

Tomczak, J. and Welling, M. (2018) Vae with a vampprior. In *International Conference on Artificial Intelligence and Statistics*, 1214–1223. PMLR.

Triebel, H. (2006) *Theory of Function Spaces III*. Basel: Birkhäuser Basel.

— (2010) *Bases in function spaces, sampling, discrepancy, numerical integration*, vol. 11. European Mathematical Society.

Uppal, A., Singh, S. and Póczos, B. (2019) Nonparametric density estimation & convergence rates for gans under besov ipm losses. *Advances in neural information processing systems*, **32**.

Villani, C. (2009) *Optimal Transport: Old and New*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Wang, J., Gao, R. and Xie, Y. (2021) Two-sample test using projected wasserstein distance. In *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE. URL: https://doi.org/10.1109%2Fisit45174.2021.9518186.

Xing, X., Shang, Z., Du, P., Ma, P., Zhong, W. and Liu, J. S. (2019) Minimax nonparametric two-sample test under smoothing. URL: https://arxiv.org/abs/1911.02171.

Appendix

Notations: We adopt the notations in the manuscript, and further introduce the following additional notations for the technical proofs. For $\alpha \in \mathbb{R}$, the floor and ceiling functions are denoted by $\lfloor \alpha \rfloor$ and $\lceil \alpha \rceil$, indicating rounding α to the next smaller and larger integer. For two sequences $\{a_n\}$ and $\{b_n\}$, we use the notation $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ to mean $a_n \leq C b_n$ and $a_n \geq C b_n$, respectively, for some constant $C > 0$ independent of n. In addition, $a_n \asymp b_n$ means that both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. We use $\mathcal{L}^2(\mathbb{R}^d)$ to denote the set of square integrable functions on $\mathbb{R}^d$. When no ambiguity arises, for an absolutely continuous probability measure ν, we may also use ν to refer its density function. For a multi-index $a = (a_1, \cdots, a_d) \in \mathbb{N}_0^d = \{(a_1, \cdots, a_d) \,|\, \forall j \in [d],\, a_j \in \mathbb{N}_0\}$, we define $|a| = \sum_{k=1}^d a_j$ and use $f^{(a)}$ to denote the mixed partial derivative of function f with order a. For $\alpha \in [0, \infty)$, we use the notation $\mathcal{C}^\alpha(\Omega)$ to denote the α-smooth Hölder (function) class (see e.g., Evans (2010)) equipped with the Hölder norm $\|\cdot\|_{C^\alpha(\Omega)}$:

$$\|f\|_{C^\alpha(\Omega)} = \sum_{|a|=\lfloor \alpha \rfloor} \max_{\substack{x,y \in \Omega, \\ x \neq y}} \frac{|f^{(a)}(x) - f^{(a)}(y)|}{\|x - y\|^{\alpha - \lfloor \alpha \rfloor}} + \sum_{|a| \leq \lfloor \alpha \rfloor} \max_{x \in \Omega} |f^{(a)}(x)|,$$

and let $\mathcal{C}_r^\alpha(\Omega) := \big\{f : \Omega \to \mathbb{R} : \|f\|_{\mathcal{C}^\alpha(\Omega)} \leq r\big\}$. Similarly, we use the notation $\mathcal{W}^\alpha(\Omega)$ to denote the α-smooth Sobolev(-2) class equipped with the Sobolev norm $\|\cdot\|_{\mathcal{W}^\alpha(\Omega)}$:

$$\|f\|_{\mathcal{W}^\alpha(\Omega)} = \sum_{|a|=\lfloor \alpha \rfloor} \sqrt{\int_\Omega \int_\Omega \frac{|f^{(a)}(x) - f^{(a)}(y)|^2}{\|x - y\|^{2(\alpha - \lfloor \alpha \rfloor) + d}} \,\mathrm{d}x \,\mathrm{d}y} + \sum_{|a| \leq \lfloor \alpha \rfloor} \sqrt{\int_\Omega |f^{(a)}(x)|^2 \,\mathrm{d}x},$$

and let $\mathcal{W}_r^\alpha(\Omega) := \big\{f : \Omega \to \mathbb{R} : \|f\|_{\mathcal{W}^\alpha(\Omega)} \leq r\big\}$. We use $\|\cdot\|_p$ to denote the usual vector ℓ_p norm, and reserve $\|\cdot\|$ for the ℓ_2 norm (that is, suppress the subscript when $p = 2$). For two sequences $\{a_n\}$ and $\{b_n\}$, we use the notation $a_n = o(b_n)$ if $a_n / b_n \to 0$ ad n increases. For two random sequences $\{a_n\}$ and $\{b_n\}$, we use the notation $a_n = o_p(b_n)$ if $a_n / b_n \xrightarrow{P} 0$. For any positive integer m, we use the shorthand $[m] := \{1, \cdots, m\}$. Throughout, C, L c, C_0, L_0, c_0, C_1, L_1, c_1,... are generically used to denote positive constants whose values might change from one line to another, but are independent from everything else.

A Implementation Details for the Real Data Application

Follow Kingma and Welling (2013), for our encoder, we use the multivariate Gaussian distribution, with the mean and covariance matrices parameterized by the outputs of a convolutional neural network with Probabilistic (tfp) Layers; the decoder is a multivariate Bernoulli whose probabilities are computed with a deconvolutional neural network. The specification of our models are described in Table 2. Note that the regularizer "KLDivergenceRegularizer" in the probabilistic layer of the encoder should contribute a "regularization" term to the final loss. Specifically, we are adding the KL divergence between the encoder and the prior to the loss, which is the KL term in the ELBO. Moreover, in the computation of the statistic $T_{\gamma,J}$, for consistency, we transform the data sets by setting $Z_1^{(n)} = Z_1^{(n)} - Z_{\min}/(Z_{\max} - Z_{\min})$ and $Z_2^{(n)} = Z_2^{(n)} - Z_{\min}/(Z_{\max} - Z_{\min})$, where $Z_{\max}$ and $Z_{\min}$ are the maximum and minimum of the joint dataset $\{Z_1^{(n)}, Z_2^{(n)}\}$ through all the data points (by dimension). After the transformation, the data sets $Z_1^{(n)}$ and $Z_2^{(n)}$ are all included in $[0, 1]^d$. The code for reproducing the experiment is available in `https://github.com/rtang1997/Two_sample_test_adversarial`.

Operation	Kernel	Strides	Feature maps	Activation
Decoder $p(x\|z) : z \in \mathbb{R}^d$			2	
Fully connected			$6 \times 6 \times 32$	Leaky ReLU
Transposed convolution	3×3	2×2	$13 \times 13 \times 64$	Leaky ReLU
Transposed convolution	3×3	2×2	$27 \times 27 \times 32$	Leaky ReLU
Transposed convolution	2×2	1×1	$28 \times 28 \times 1$	Leaky ReLU
Probabilistic Layers: IndependentBernoulli			$28 \times 28 \times 1$	
Encoder $q(z\|x)$			$28 \times 28 \times 1$	
Minus x by 0.5			$28 \times 28 \times 1$	
Convolution	3×3	2×2	$14 \times 14 \times 32$	LeakyReLU
Convolution	3×3	2×2	$7 \times 7 \times 64$	LeakyReLU
Fully connected			5	
Probabilistic Layers:MultivariateNomalTriL			2	KLDivergenceRegularizer
Batch size	128			
Number of epochs	50			
Number of training samples and test samples	60k and 10k respectively.			

Table 2: Network architecture and hyperparameters the encoder and decoder.

B Wavelet and Besov Function Space

In this section, we give a brief introduction to the wavelet and Besov function Space. Further details are available in Cohen (2003); Triebel (2006); Härdle et al. (2012). Let $\phi_{\mathfrak{M}}$ and $\phi_{\mathfrak{F}}$ be a compactly supported wavelet and scaling function, respectively, for example Daubechies wavelets (Bouzebda and Didi, 2017; Hütter and Rigollet, 2021). This implies that

$$\begin{cases} \psi_{\mathfrak{F}}(x-k) & j=0, k \in \mathbb{Z}, \\ 2^{(j-1)/2}\psi_{\mathfrak{M}}(2^{j-1}x-k), & j \in \mathbb{N}^+, k \in \mathbb{Z}, \end{cases}$$

is an orthonormal basis of $\mathcal{L}^2(\mathbb{R})$. To obtain a basis of $\mathcal{L}^2(\mathbb{R}^d)$ for an integer $d > 1$, set $\mathfrak{G} = \{\mathfrak{F}, \mathfrak{M}\}^d \setminus \{(\mathfrak{F}, \ldots, \mathfrak{F})\}$. Then for any multi-index $k \in \mathbb{Z}^d$, the level zero basis $\phi_k^{[d]}$ is obtained by translating the d-fold tensor product $\phi_{\mathfrak{F}}^{\otimes d}$ by k as $\phi_k^{[d]}(x) = \prod_{i=1}^d \phi_{\mathfrak{F}}(x_i - k_i)$ for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, and for any $j \geq 1$, the level j basis $\{\psi_{ljk}^{[d]} : l \in [2^d - 1]\}$ with translation k is any ordering of the following $2^d - 1$ functions,

$$\psi_{gjk}^{[d]}(x) = 2^{\frac{d(j-1)}{2}} \prod_{i=1}^d \phi_{g_i}^{[d]}\big(2^{j-1}x_i - k_i\big), \quad \forall g \in \mathfrak{G}.$$

This gives the orthornormal basis

$$\begin{cases} \phi_k^{[d]}(x), & j=0, l=0, k \in \mathbb{Z}^d, \\ \psi_{ljk}^{[d]}(x), & j \in \mathbb{N}^+, l \in [2^d-1], k \in \mathbb{Z}^d, \end{cases}$$

for $\mathcal{L}^2(\mathbb{R}^d)$. Denote $\overline{\Psi}_0 = \{\phi_k^{[d]}(\cdot) : k \in \mathbb{Z}^d\}$ as the set of level zero basis and $\overline{\Psi}_j = \{\psi_{ljk}^{[d]}(\cdot) : l \in [2^d-1], k \in \mathbb{Z}^d\}$ as the set of level j basis for $j \in \mathbb{N}^+$. We are then ready to define the Besov space $B_{l,m}^s(\mathbb{R}^d)$ consists of functions f that admits the wavelet expansion

$$f(x) = \sum_{j \geq 0} \sum_{\psi \in \overline{\Psi}_j} f_\psi \psi(x),$$

where $f_\psi := \int f(x)\psi(x)\,\mathrm{d}x$, and is equipped with the norm

$$\|f\|_{B_{p,q}^s} := \Big\| 2^{js} 2^{dj(\frac{1}{2}-\frac{1}{p})} \|f_j\|_l \Big\|_m,$$

with $f_j = \{f_\psi\}_{\psi \in \overline{\Psi}_j}$.

The following Theorem collects the relationship between the Besov space, Hölder space and Sobolev-2 space.

Theorem 3. *(Triebel, 2006; Giné and Nickl, 2015) Let $\alpha > 0$, $\mathcal{W}^\alpha(\mathbb{R}^d) = B_{2,2}^\alpha(\mathbb{R}^d)$. If α is not integer, then $C^\alpha(\mathbb{R}^d) = B_{\infty,\infty}^\alpha(\mathbb{R}^d)$; if α is integer, then $B_{1,\infty}^\alpha(\mathbb{R}^d) \subset C^\alpha(\mathbb{R}^d) \subset B_{\infty,\infty}^\alpha(\mathbb{R}^d)$.*

C Goodness-of-fit Test

The goodness of fit test is a statistical hypothesis test used to determine whether the sample data fits a distribution from an expected population (e.g. a population with a normal distribution). Given data sets $X^{(n)} = \{X_1, X_2, \cdots, X_n\}$ i.i.d sampled from an unknown distribution p. The goal is to check if $X^{(n)}$ come from a distribution p_0, in other words, we are interested in testing the null hypothesis $\mathbb{H}_0^{\mathrm{GoF}} : p = p_0$.

As in the case of two-sample test, we restrict our attention to smooth densities $p, p_0 \in \mathcal{W}_L^{u,\alpha}(\Omega)$, and consider an alternative test

$$\mathbb{H}_1^{\mathrm{GoF}}(\Delta_n; \mathcal{D}) : \mathcal{D}(p, p_0) \geq \Delta_n,$$

where $\mathcal{D}$ is some discrepancy measure. Then for a test Φ based on data $X^{(n)}$, the power of Φ is defined as

$$\mathrm{power}\big(\Phi; \mathbb{H}_1^{\mathrm{GoF}}(\Delta_n; \mathcal{D})\big) := \inf_{\substack{p \in \mathcal{W}_L^{u,\alpha}(\Omega) \\ \mathcal{D}(p,p_0) \geq \Delta_n}} P(\Phi \,\text{reject}\, \mathbb{H}_0^{\mathrm{GoF}}).$$

Similar to the statistic $T_{\gamma,J}$, we can define the following statistic for approximating the squared negative Besov norm $\|p - p_0\|^2_{B^{-\gamma}_{2,2}(\Omega)}$:

$$T_{\gamma,J}^{\mathrm{GoF}} = \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} \Big\{ \sum_{j=0}^{J} 2^{-2j\gamma} \cdot \sum_{\psi \in \Psi_j} \Big[\big(\psi(X_{i_1}) - p_{0\psi}\big)\big(\psi(X_{i_2}) - p_{0\psi}\big) \Big] \Big\}. \tag{12}$$

As before, we normalize the statistic $T_{\gamma,J}^{\mathrm{GoF}}$ to construct an optimal test statistic. Note that under $\mathbb{H}_0^{\mathrm{GoF}}$,

$$\mathrm{Var}(T_{\gamma,J}^{\mathrm{GoF}}) = \frac{2}{n(n-1)} \mathbb{E}_{X_1, X_2 \sim p_0} \Big[\Big(\sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi \in \Psi_j} \big(\psi(X_1) - p_{0\psi}\big) \cdot \big(\psi(X_2) - p_{0\psi}\big) \Big)^2 \Big].$$

It is then natural to consider estimating $\mathrm{Var}(T_{\gamma,J}^{\mathrm{GoF}})$ by U-statistics:

$$\begin{aligned} \widehat{S}_{\gamma,J}^2 = \frac{2}{n(n-1)} \Big\{ & \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} \Big[\sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi \in \Psi_j} \psi(X_{i_1}) \psi(X_{i_2}) \Big]^2 \\ & - \frac{2}{n} \sum_{i=1}^{n} \Big[\sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi \in \Psi_j} \psi(X_i) p_{0\psi} \Big]^2 + \Big(\sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi \in \Psi_j} p_{0\psi}^2 \Big)^2 \Big\}. \end{aligned}$$

Similarly to the two-sample test, we slightly modify the variance estimator by considering

$$\widetilde{S}_{\gamma,J}^2 = \max(\widehat{S}_{\gamma,J}^2, \frac{1}{n^3})$$

to ensure the positiveness. In the end, we can define the test statistic

$$\widetilde{T}_{\gamma,J}^{\mathrm{GoF}} = \widetilde{S}_{\gamma,J}^{-1} T_{\gamma,J}^{\mathrm{GoF}}.$$

The following theorem show the validity and power of the test induced from $\widetilde{T}_{\gamma,J}^{\mathrm{GoF}}$.

Theorem 4. *Consider Test statistic $\widetilde{T}_{\gamma,J}^{\mathrm{GoF}}$ with $2^J \asymp n^{\frac{2}{4\alpha+d}}$ and $0 \leq \gamma \leq \frac{d}{4}$, and level $\eta \in (0,1)$,*

1. *under $\mathbb{H}_0^{\mathrm{GoF}}$, we have $\widetilde{T}_{\gamma,J}^{\mathrm{GoF}} \xrightarrow{d} N(0,1)$;*
2. *consider test $\Phi_{\gamma,\eta,J}^{\mathrm{GoF}}$ where $\mathbb{H}_0^{\mathrm{GoF}}$ is rejected if and only if $\widetilde{T}_{\gamma,J}^{\mathrm{GoF}}$ exceeds the η-upper quantile of the standard normal, then for any $\gamma_1 \geq 0$ and Δ_n satisfies $\Delta_n^2 \cdot \delta_n(\gamma_1)^{-1} \to \infty$ with*

$$\delta_n(\gamma_1) = \begin{cases} n^{-\frac{4\alpha+4(\gamma\wedge\gamma_1)}{4\alpha+d}}, & 0 < \gamma < \frac{d}{4} \\ \log n \cdot n^{-\frac{4\alpha}{4\alpha+d}} & \gamma = 0 \\ \log n \cdot n^{-1+\frac{4(\gamma-\gamma\wedge\gamma_1)}{4\alpha+d}}, & \gamma = \frac{d}{4}, \end{cases}$$

we have

$$\lim_{n\to\infty} \mathrm{power}\big(\Phi_{\gamma,\eta,J}^{\mathrm{GoF}}; \mathbb{H}_1^{\mathrm{GoF}}(\Delta_n; d_{\gamma_1}^W)\big) = 1$$

3. choose $\gamma = \frac{d}{4}$, let $\Delta_n = (\log n) \cdot \big(n^{-\frac{2(\alpha+\gamma_1)}{4\alpha+d}} \vee \frac{1}{\sqrt{n}}\big)$, then the associated test $\Phi^{\mathrm{GoF}}_{\frac{d}{4},\eta,J}$ satisfies that for any $\gamma_1 \geq 0$,

$$\lim_{n\to\infty} \mathrm{power}\big(\Phi^{\mathrm{GoF}}_{\frac{d}{4},\eta,J}; \mathbb{H}_1^{\mathrm{GoF}}(\Delta_n; d^W_{\gamma_1})\big) = 1.$$

Theorem 5. *For any $\gamma_1 > 0$, if $\Delta_n = o(n^{-\frac{2(\alpha+\gamma_1)}{4\alpha+d}} \vee \frac{1}{\sqrt{n}})$, then there exists some $\eta \in (0,1)$ so that for any test Φ_n based on data $X^{(n)}$ that has asymptotically significance level η, we have*

$$\liminf_{n\to\infty} \mathrm{power}\big(\Phi_n; \mathbb{H}_1^{\mathrm{GoF}}(\Delta_n; d^H_{\gamma_1})\big) < 1.$$

D Proof for Goodness-of-fit Test

D.1 Proof of Theorem 5

Since p_0 is almost surely continuous, there exists $x_0 \in \Omega$ and $\delta, c > 0$ so that $p_0(x) \geq c > 0$ for any $\|x - x_0\| \leq \delta$. So without loss of generality, we assume $[0,1]^d \subset \Omega$ and p_0 is bounded away from zero in $[0,1]^d$. We first consider the case when $d \geq 4\gamma_1$. Then $\Delta_n = n^{-\frac{2(\alpha+\gamma_1)}{4\alpha+d}}$. Similar as the proof of Theorem 3 of Li and Yuan (2019), as proved in Ingster (1987), we only need to construct a set of density function $\{p_\omega\}_{\omega\in\mathscr{W}}$ belong to $\mathcal{C}^\alpha_L(\mathbb{R}^d)$ with compact support and indexed by a multi-index ω so that

$$\mathbb{E}_{p_0}\Big(\frac{\frac{1}{|\mathscr{W}|}\sum_{\omega\in\mathscr{W}}\prod_{i=1}^n p_\omega(X_i)}{\prod_{i=1}^n p_0(X_i)}\Big)^2 = O(1),$$

and for any $\omega \in \mathscr{W}$

$$d^H_{\gamma_1}(p_0, p_\omega) \gtrsim n^{-\frac{2(\alpha+\gamma_1)}{4\alpha+d}}.$$

To construct p_ω satisfies above conditions, we set $m = \lceil n^{\frac{2}{4\alpha+d}} \rceil$,

$$\mathscr{W} = \{-1, 1\}^{m^d},$$

$$\omega = \{\omega_\xi\}_{\xi\in[m]^d},$$

and

$$p_\omega(x) = p_0(x) + \big(\frac{1}{m}\big)^{\alpha+\frac{d}{2}} \sum_{\xi\in[m-1]^d} \omega_\xi \cdot \phi_\xi(x)$$

$$\text{with} \quad \phi_\xi(x) = m^{\frac{d}{2}} \cdot \prod_{j=1}^d k(mx_j - \xi_j)$$

$$\text{where} \quad k(t) = \begin{cases} \exp(-\frac{1}{1-(4t-1)^2}) & 0 < t < \frac{1}{2} \\ -\exp(-\frac{1}{1-(4t-3)^2}) & \frac{1}{2} < t < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then we can check that $\{p_\omega\}_{\omega\in\mathscr{W}} \subset C^\alpha_L(\mathbb{R}^d)$ and $\bigcup_{\omega\in\mathscr{W}} \mathrm{supp}(p_\omega) \subset \Omega$. Moreover, by equation (14) of Li and Yuan (2019), we have

$$\begin{aligned}\mathbb{E}_{p_0}\Big(\frac{\frac{1}{|\mathscr{W}|}\sum_{\omega\in\mathscr{W}}\prod_{i=1}^n p_\omega(X_i)}{\prod_{i=1}^n p_0(X_i)}\Big)^2 &\leq \exp\Big(\frac{1}{2}m^d n^2 m^{-4\alpha+2d} \max_{\xi\in[m-1]^d}\Big(\int \phi^2_\xi(x)/p_0(x)\mathrm{d}x\Big)^2\Big) \\ &= O(1).\end{aligned}$$

Furthermore, for any $\omega \in \mathscr{W}$, we have

$$f_\omega(x) = \big(\frac{1}{m}\big)^{\gamma_1+\frac{d}{2}} \sum_{\xi\in[m-1]^d} \omega_\xi \cdot \phi_\xi(x) \in \mathcal{C}^{\gamma_1}_{L_1}(\mathbb{R}^d),$$

and thus

$$\begin{aligned} d_{\gamma_1}^H(p_0, p_\omega) &= \sup_{f \in \mathcal{C}_1^{\gamma_1}(\mathbb{R}^d)} \int f(x) \cdot \big(p_\omega(x) - p_0(x)\big)\, \mathrm{d}x \\ &\gtrsim \int f_\omega(x) \cdot \big(p_\omega(x) - p_0(x)\big)\, \mathrm{d}x \\ &= \big(\frac{1}{m}\big)^{\gamma_1+d+\alpha} \sum_{\xi \in [m-1]^d} \sum_{\xi_1 \in [m-1]^d} \omega_\xi \cdot \omega_{\xi_1} \cdot \int \phi_\xi(x) \cdot \phi_{\xi_1}(x)\, \mathrm{d}x \\ &= \big(\frac{1}{m}\big)^{\gamma_1+d+\alpha} \sum_{\xi \in [m-1]^d} \omega_\xi^2 \int \phi_\xi^2(x) \mathrm{d}x \\ &\gtrsim m^{-(\gamma_1+\alpha)} \asymp n^{-\frac{2(\alpha+\gamma_1)}{4\alpha+d}}. \end{aligned}$$

For the case $d < 4\gamma_1$, we have $\Delta_n = \frac{1}{\sqrt{n}}$. Consider

$$p(x) = p_0(x) + \frac{1}{\sqrt{n}} \prod_{j=1}^d k(x_j) \in \mathcal{C}_L^\alpha(\mathbb{R}^d).$$

Then we have $\mathrm{supp}(p) \subset \Omega$, and

$$\begin{aligned} d_{\chi^2}(p(x), p_0(x)) &= \int (\frac{p(x)}{p_0(x)} - 1)^2 p_0(x)\, \mathrm{d}x \\ &= \int \frac{1}{n} \frac{\prod_{j=1}^d k^2(x_j)}{p_0(x)}\, \mathrm{d}x \\ &\lesssim \frac{1}{n}, \end{aligned}$$

and thus

$$d_{\chi^2}(p^{\otimes n}(x), p_0^{\otimes n}(x)) = O(1).$$

Moreover, since $f(x) = \prod_{j=1}^d k(x_j) \in \mathcal{C}_{L_1}^\gamma(\mathbb{R}^d)$, we have

$$\begin{aligned} d_\gamma^H(p_0, p) &\geq \int \prod_{j=1}^d k(x_j) \cdot (p(x) - p_0(x))\, \mathrm{d}x \\ &= \int \frac{1}{\sqrt{n}} \cdot \prod_{j=1}^d k(x_j)^2\, \mathrm{d}x \\ &\gtrsim \frac{1}{\sqrt{n}}. \end{aligned}$$

We can then get the desired conclusion by combining all pieces.

D.2 Proof of Theorem 4

Throughout the proof, we use X to denote the random variable sampled from p, and $X_1, X_2, \cdots$ to denote independent random variables from p. Without loss of generality, we assume $\Omega = [0,1]^d$. The proof contains two part: one part is about the normality under $\mathbb{H}_0^{\mathrm{GoF}}$, another part is the power analysis. We first show the normality.

D.2.1 Proof of the normality under $\mathbb{H}_0^{\mathrm{GoF}}$

To begin with, we show $\widetilde{S}_{\gamma,J}^2$ is a valid approximate for the variance of $T_{\gamma,J}^{\mathrm{GoF}}$ though the following lemma.

Lemma 3. *Under $\mathbb{H}_0^{\mathrm{GoF}}$, the quantity $\frac{\widetilde{S}_{\gamma,J}^2}{\mathrm{Var}(T_{\gamma,J}^{\mathrm{GoF}})}$ converges in probability to 1 as n goes to infinity.*

The proof of Lemma 3 is given in Section D.2.3. Write

$$\widetilde{S}_{\gamma,J}^{-1} T_{\gamma,J}^{\mathrm{GoF}} = \frac{T_{\gamma,J}^{\mathrm{GoF}}}{\sqrt{\mathrm{Var}(T_{\gamma,J}^{\mathrm{GoF}})}} + \Big(\frac{\sqrt{\mathrm{Var}(T_{\gamma,J}^{\mathrm{GoF}})}}{\widetilde{S}_{\gamma,J}} - 1\Big) \cdot \frac{T_{\gamma,J}^{\mathrm{GoF}}}{\sqrt{\mathrm{Var}(T_{\gamma,J}^{\mathrm{GoF}})}}.$$

By Lemma 3, we only need to prove that

$$\frac{T^{\mathrm{GoF}}_{\gamma,J}}{\sqrt{\mathrm{Var}(T^{\mathrm{GoF}}_{\gamma,J})}} \xrightarrow{d} N(0,1). \tag{13}$$

Let

$$H(X_1,X_2)=\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}\big(\psi(X_1)-p_{0\psi}\big)\cdot\big(\psi(X_2)-p_{0\psi}\big).$$

Then by a slight adaptation of the proof of Theorem 1 of Hall (1984), we have the following lemma.

Lemma 4. *Suppose*

$$\frac{\mathbb{E}[H^4(X_1,X_2)]}{n^2(\mathbb{E}[H^2(X_1,X_2)])^2}\to 0; \tag{14}$$

$$\frac{\mathbb{E}[H^2(X_1,X_2)H^2(X_1,X_3)]}{n(\mathbb{E}[H^2(X_1,X_2)])^2}\to 0; \tag{15}$$

$$\frac{\mathbb{E}[G^2(X_1,X_2)]}{(\mathbb{E}[H^2(X_1,X_2)])^2}\to 0, \text{ where } G(x,y)=\mathbb{E}[H(X,x)H(X,y)]. \tag{16}$$

Then we have under $\mathbb{H}_0$, $\frac{T^{\mathrm{GoF}}_{\gamma,J}}{\sqrt{\mathrm{Var}(T^{\mathrm{GoF}}_{\gamma,J})}} \xrightarrow{d} N(0,1)$.

We first show statement (14). By equation (21) in the proof Lemma 3, we have

$$\mathbb{E}[H^2(X_1,X_2)]\gtrsim 2^{J(d-4\gamma)}; \tag{17}$$

moreover, we can obtain

$$\begin{aligned}
\mathbb{E}[H^4(X_1,X_2)] &= \mathbb{E}\Big[\Big(\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}\big(\psi(X_1)-p_{0\psi}\big)\cdot\big(\psi(X_2)-p_{0\psi}\big)\Big)^4\Big]\\
&\lesssim \mathbb{E}\Big[\Big(\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(X_1)\cdot\psi(X_2)\big)\Big)^4\Big]\\
&\quad+\mathbb{E}\Big[\Big(\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(X)\cdot p_{0\psi}\big)\Big)^4\Big]+\Big(\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}p_{0\psi}^2\Big)^4\\
&\lesssim 2^{J(3d-8\gamma)}\cdot J,
\end{aligned}$$

where the last inequality uses the bounds for terms (A), (C) in the proof of Lemma 3. So statement (14) holds by plugging in $2^J\asymp n^{\frac{2}{4\alpha+d}}$ with $\alpha>0$.

Now we show statement (15). Let

$$\widetilde{\Psi}_j(\psi)=\big\{\psi'\in\Psi_j\,:\,\mathrm{supp}(\psi)\cap\mathrm{supp}(\psi')\neq\emptyset\big\}.$$

Then for any $j\geq j'$ and $\psi\in\Psi_{j'}$, we have

$$\big|\widetilde{\Psi}_j(\psi)\big|\lesssim 2^{d(j-j')}.$$

We have

$$
\begin{aligned}
&\mathbb{E}[H^2(X_1,X_2)H^2(X_1,X_3)] \\
&= \mathbb{E}\Big[\Big(\sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi\in\Psi_j} \big(\psi(X_1)-p_{0\psi}\big)\cdot\big(\psi(X_2)-p_{0\psi}\big)\Big)^2 \\
&\qquad \cdot \Big(\sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi\in\Psi_j} \big(\psi(X_1)-p_{0\psi}\big)\cdot\big(\psi(X_3)-p_{0\psi}\big)\Big)^2\Big] \\
&\le \mathbb{E}\Big[\Big(\sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi\in\Psi_j} \big(\psi(X_1)-p_{0\psi}\big)\cdot\psi(X_2)\Big)^2 \cdot \Big(\sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi\in\Psi_j} \big(\psi(X_1)-p_{0\psi}\big)\cdot\psi(X_3)\Big)^2\Big] \\
&= \sum_{j_1=0}^{J}\sum_{j_2=0}^{J}\sum_{j_3=0}^{J}\sum_{j_4=0}^{J} 2^{-2(j_1+j_2+j_3+j_4)\gamma} \sum_{\psi_1\in\Psi_{j_1}}\sum_{\psi_2\in\Psi_{j_2}}\sum_{\psi_3\in\Psi_{j_3}}\sum_{\psi_4\in\Psi_{j_4}} \mathbb{E}\Big[\psi_1(X_2)\psi_2(X_2)\psi_3(X_3)\psi_4(X_3) \\
&\qquad \cdot\big(\psi_1(X_1)-p_{0\psi_1}\big)\cdot\big(\psi_2(X_1)-p_{0\psi_2}\big)\cdot\big(\psi_3(X_1)-p_{0\psi_3}\big)\cdot\big(\psi_4(X_1)-p_{0\psi_4}\big)\Big] \\
&\overset{(i)}{\lesssim} \sum_{j_1=0}^{J}\sum_{\psi_1\in\Psi_{j_1}}\sum_{j_2=j_1}^{J}\sum_{\psi_2\in\widetilde{\Psi}_{j_2}(\psi_1)}\sum_{j_3=j_2}^{J}\sum_{\psi_3\in\widetilde{\Psi}_{j_3}(\psi_2)}\sum_{j_4=j_3}^{J}\sum_{\psi_4\in\widetilde{\Psi}_{j_4}(\psi_3)} 2^{-2(j_1+j_2+j_3+j_4)\gamma}\mathbb{E}\Big[\psi_1(X_2)\psi_2(X_2)\psi_3(X_3)\psi_4(X_3) \\
&\qquad \cdot\big(\psi_1(X_1)-p_{0\psi_1}\big)\cdot\big(\psi_2(X_1)-p_{0\psi_2}\big)\cdot\big(\psi_3(X_1)-p_{0\psi_3}\big)\cdot\big(\psi_4(X_1)-p_{0\psi_4}\big)\Big] \\
&\lesssim \sum_{j_1=0}^{J}\sum_{j_2=j_1}^{J}\sum_{j_3=j_2}^{J}\sum_{j_4=j_3}^{J} 2^{dj_4}\cdot 2^{-2(j_1+j_2+j_3+j_4)\gamma}\cdot 2^{-dj_4}\cdot 2^{\frac{d}{2}(j_1+j_2+j_3+j_4)}\cdot 2^{-dj_2}\cdot 2^{\frac{d}{2}(j_1+j_2)}\cdot 2^{-dj_4}\cdot 2^{\frac{d}{2}(j_3+j_4)} \\
&\lesssim 2^{J(2d-8\gamma)}\cdot J^2.
\end{aligned}
\tag{18}
$$

Combined with (17), we can get the desired statement. The it remains to show statement (16). We have

$$
\begin{aligned}
&\mathbb{E}[G^2(X_1,X_2)] \\
&= \mathbb{E}_{X_2,X_3}\Big[\Big(\mathbb{E}_{X_1}\big[H(X_1,X_2)H(X_1,X_3)\big]\Big)^2\Big] \\
&= \mathbb{E}_{X_2,X_3}\Big[\Big(\sum_{j_1=0}^{J}\sum_{j_2=0}^{J}\sum_{\psi_1\in\Psi_{j_1}}\sum_{\psi_2\in\Psi_{j_2}} 2^{-2(j_1+j_2)\gamma}\cdot\mathbb{E}_{X_1}\big[(\psi_1(X_1)-p_{0\psi_1})(\psi_2(X_1)-p_{0\psi_2})\big] \\
&\qquad\qquad \cdot(\psi_1(X_2)-p_{0\psi_1})(\psi_2(X_3)-p_{0\psi_2})\Big)^2\Big] \\
&\le \mathbb{E}_{X_2,X_3}\Big[\Big(\sum_{j_1=0}^{J}\sum_{j_2=0}^{J}\sum_{\psi_1\in\Psi_{j_1}}\sum_{\psi_2\in\Psi_{j_2}} 2^{-2(j_1+j_2)\gamma}\mathbb{E}_{X_1}\big[(\psi_1(X_1)-p_{0\psi_1})(\psi_2(X_1)-p_{0\psi_2})\big]\psi_1(X_2)\psi_2(X_3)\Big)^2\Big] \\
&= \sum_{j_1=0}^{J}\sum_{j_2=0}^{J}\sum_{j_3=0}^{J}\sum_{j_4=0}^{j}\sum_{\psi_1\in\Psi_{j_1}}\sum_{\psi_2\in\Psi_{j_2}}\sum_{\psi_3\in\Psi_{j_3}}\sum_{\psi_4\in\Psi_{j_4}} 2^{-2(j_1+j_2+j_3+j_4)\gamma}\cdot\mathbb{E}\Big[\psi_1(X_2)\psi_2(X_3)\psi_3(X_2)\psi_4(X_3) \\
&\qquad \cdot\big(\psi_1(X_1)-p_{0\psi_1}\big)\cdot\big(\psi_2(X_1)-p_{0\psi_2}\big)\cdot\big(\psi_3(X_4)-p_{0\psi_3}\big)\cdot\big(\psi_4(X_4)-p_{0\psi_4}\big)\Big] \\
&\overset{(ii)}{\lesssim} \sum_{j_1=0}^{J}\sum_{j_2=j_1}^{J}\sum_{j_3=j_2}^{J}\sum_{j_4=j_3}^{J} 2^{dj_4}\cdot 2^{-2(j_1+j_2+j_3+j_4)\gamma}\cdot 2^{d(j_1+j_2+j_3+j_4)}\cdot 2^{-d(j_3+j_4+j_2+j_4)} \\
&\lesssim J^3\cdot\sum_{j_1=0}^{J} 2^{(d-8\gamma)j_1} = o(2^{J(2d-8\gamma)}),
\end{aligned}
$$

where (ii) uses the same strategy as in (i) of inequality (18). We can then get the desired result by combining all pieces.

D.2.2 Proof for the Power analysis

Since $p \in \mathcal{W}_L^{u,\alpha}(\Omega)$, we can write

$$p = \sum_{j=0}^{J} \sum_{\psi \in \Psi_j} p_\psi \psi(x).$$

For any $f \in \mathcal{W}_1^{\gamma_1}(\mathbb{R}^d)$ and $\psi \in \Psi_j$,denote $f_\psi = \int f(x)\psi(x)\,\mathrm{d}x$. We have

$$\begin{aligned}
&\sup_{f \in \mathcal{W}_1^{\gamma_1}(\mathbb{R}^d)} \int f \,\mathrm{d}p_0 - \int f \,\mathrm{d}p \\
&= \sup_{f \in \mathcal{W}_1^{\gamma_1}(\mathbb{R}^d)} \sum_{j=0}^{\infty} \sum_{\psi \in \Psi_j} (p_\psi - p_{0\psi}) \cdot f_\psi \\
&\lesssim \sup_{f \in \mathcal{W}_1^{\gamma_1}(\mathbb{R}^d)} \sum_{j=0}^{J} \sum_{\psi \in \Psi_j} (p_\psi - p_{0\psi}) \cdot f_\psi \\
&\qquad + 2^{-J(\gamma_1+\alpha)} \sqrt{\sum_{j=J}^{\infty} \sum_{\psi \in \Psi_j} (p_\psi - p_{0\psi})^2 \cdot 2^{2j\alpha}} \sqrt{\sup_{f \in \mathcal{W}_1^{\gamma_1}(\mathbb{R}^d)} \sum_{j=J}^{\infty} \sum_{\psi \in \Psi_j} 2^{2j\gamma_1} \cdot f_\psi^2} \\
&\lesssim \sqrt{\sum_{j=0}^{J} \sum_{\psi \in \Psi_j} 2^{-2\gamma_1 j} \cdot (p_\psi - p_{0\psi})^2} \sqrt{\sup_{f \in \mathcal{W}_1^{\gamma_1}(\mathbb{R}^d)} \sum_{j=0}^{J} \sum_{\psi \in \Psi_j} 2^{2\gamma_1 j} \cdot f_\psi^2} + O\big(2^{-J(\gamma_1+\alpha)}\big) \\
&\lesssim \sqrt{\sum_{j=0}^{J} \sum_{\psi \in \Psi_j} 2^{-2\gamma_1 j} \cdot (p_\psi - p_{0\psi})^2} + O\big(2^{-J(\gamma_1+\alpha)}\big)
\end{aligned} \tag{19}$$

Then when $d_{\gamma_1}(p_0, p) \geq \Delta_n$ with $\Delta_n^2 \cdot \delta_n(\gamma_1)^{-1} \to \infty$. We can obtain

$$\sum_{j=0}^{J} \sum_{\psi \in \Psi_j} 2^{-2\gamma_1 j} \cdot (p_\psi - p_{0\psi})^2 \gtrsim \Delta_n^2.$$

So if $\gamma_1 \geq \gamma$, we have

$$\sum_{j=0}^{J} \sum_{\psi \in \Psi_j} 2^{-2\gamma j} \cdot (p_\psi - p_{0\psi})^2 \gtrsim \Delta_n^2;$$

and if $\gamma_1 < \gamma$, we have

$$\sum_{j=0}^{J} \sum_{\psi \in \Psi_j} 2^{-2\gamma j} \cdot (p_\psi - p_{0\psi})^2 \gtrsim \Delta_n^2 \cdot 2^{-2J(\gamma-\gamma_1)};$$

Denote $\widetilde{\Delta}_n = \Delta_n^2 \cdot 2^{-2J(\gamma-\gamma_1\wedge\gamma)}$. To show the desired result, we only need to prove that when $\sum_{j=0}^{J} \sum_{\psi \in \Psi_j} 2^{-2\gamma j} \cdot (p_\psi - p_{0\psi})^2 \gtrsim \widetilde{\Delta}_n$, $\widetilde{S}_{\gamma,J}^{-1} T_{\gamma,J}^{\mathrm{GoF}} \xrightarrow{P} \infty$.

Note that we can rewrite the statistic $T_{\gamma,J}^{\mathrm{GoF}}$ as

$$\begin{aligned}
T_{\gamma,J}^{\mathrm{GoF}} = &\underbrace{\frac{1}{n(n-1)} \sum_{i_1 \neq i_2} \Big\{ \sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi \in \Psi_j} \big(\psi(X_{i_1}) - p_\psi\big) \cdot \big(\psi(X_{i_2}) - p_\psi\big) \Big\}}_{(A')} \\
&+ \underbrace{\sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi \in \Psi_j} (p_\psi - p_{0\psi})^2}_{(B')} + \underbrace{\frac{2}{n} \sum_{i=1}^{n} \sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi \in \Psi_j} (p_\psi - p_{0\psi}) \cdot (\psi(X_i) - p_\psi)}_{(C')}.
\end{aligned}$$

We first consider term (A'), we have

$$\mathbb{E}[(A')] = 0$$

and

$$\begin{aligned}
\mathrm{Var}(A') &= \frac{2}{n(n-1)}\mathbb{E}\Big[\Big(\sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi\in\Psi_j} \big(\psi(X_1) - p_\psi\big)\cdot\big(\psi(X_2) - p_\psi\big)\Big)^2\Big] \\
&\le \frac{2}{n(n-1)}\mathbb{E}\Big[\Big(\sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi\in\Psi_j} \psi(X_1)\cdot\psi(X_2)\Big)^2\Big] \\
&\lesssim n^{-2}\cdot \sum_{j_1=0}^{J}\sum_{\psi_1\in\Psi_{j_1}}\sum_{j_2=j_1}^{J}\sum_{\psi_2\in\widetilde{\Psi}_{j_2}(\psi_1)} 2^{-2(j_1+j_2)\gamma}\cdot\Big(\mathbb{E}_{p_0}\big[\psi_1(X)\psi_2(X)\big]\Big)^2 \\
&\lesssim n^{-2}\cdot \sum_{j_1=0}^{J}\sum_{j_2=j_1}^{J} 2^{dj_2}\cdot 2^{-2(j_1+j_2)\gamma}\Big(2^{-dj_2}\cdot 2^{\frac{d}{2}(j_1+j_2)}\Big)^2 \\
&\lesssim \begin{cases} n^{-2}\cdot\frac{2^{J(d-4\gamma)}-1}{2^{d-4\gamma}-1} & \gamma>0 \\ n^{-2}\cdot 2^{dJ}\cdot J & \gamma = 0 \end{cases} \\
&= o(\widetilde{\Delta}_n^2),
\end{aligned}$$

where $\widetilde{\Psi}_j(\psi)$ is defined in (23). The above inequality leads to

$$\frac{(A')}{\widetilde{\Delta}_n} \xrightarrow{P} 0.$$

For term (B'), we have

$$(B') \gtrsim \widetilde{\Delta}_n.$$

For term (C'), we have we have

$$\mathbb{E}[(C')] = 0$$

and

$$\begin{aligned}
\mathrm{Var}(C') &\le \frac{4}{n}\,\mathbb{E}\Big[\Big(\sum_{j=0}^{J} 2^{-2j\gamma}\sum_{\psi\in\Psi_j}(p_\psi - p_{0\psi})\cdot\psi(X)\Big)^2\Big] \\
&= \frac{4}{n}\sum_{j_1=0}^{J}\sum_{j_2=0}^{J}\sum_{\psi_1\in\Psi_{j_1}}\sum_{\psi_2\in\Psi_{j_2}} 2^{-2\gamma(j_1+j_2)}\cdot(p_{\psi_1}-p_{0\psi_1})\cdot(p_{\psi_2}-p_{0\psi_2})\cdot\mathbb{E}[\psi_1(X)\psi_2(X)] \\
&\lesssim n^{-1}\sqrt{\sum_{j_1=0}^{J}\sum_{j_2=j_1}^{J}\sum_{\psi_1\in\Psi_{j_1}}\sum_{\psi_2\in\widetilde{\Psi}_{j_2}(\psi_1)} 2^{-2\gamma(j_1+j_2)}\cdot\mathbb{E}\big[|\psi_1(X)|\cdot|\psi_2(X)|\big]} \\
&\quad\cdot\sqrt{\sum_{j_1=0}^{J}\sum_{j_2=0}^{J}\sum_{\psi_1\in\Psi_{j_1}}\sum_{\psi_2\in\Psi_{j_2}} 2^{-2\gamma(j_1+j_2)}\cdot(p_{\psi_1}-p_{0\psi_1})^2\cdot(p_{\psi_2}-p_{0\psi_2})^2\cdot\mathbb{E}\big[|\psi_1(X)|\cdot|\psi_2(X)|\big]} \\
&\lesssim n^{-1}\cdot\sqrt{\sum_{j_1=0}^{J}\sum_{j_2=j_1}^{J} 2^{(\frac{d}{2}-2\gamma)(j_1+j_2)}}\cdot\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j} 2^{-2\gamma j}\cdot(p_\psi - p_{0\psi})^2\Big) \\
&\lesssim n^{-1}\cdot\frac{2^{J(\frac{d}{2}-2\gamma)}-1}{2^{\frac{d}{2}-2\gamma}-1}\cdot\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j} 2^{-2\gamma j}\cdot(p_\psi - p_{0\psi})^2\Big).
\end{aligned}$$

Combining with the fact that $\sum_{j=0}^{J}\sum_{\psi\in\Psi_j} 2^{-2\gamma j}\cdot(p_\psi - p_{0\psi})^2 \gtrsim \widetilde{\Delta}_n$, we have

$$\frac{\mathrm{Var(C')}}{\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j} 2^{-2\gamma j}\cdot(p_\psi - p_{0\psi})^2\Big)^2} = o(1),$$

which can lead to

$$\frac{(C')}{\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2\gamma j}\cdot(p_\psi-p_{0\psi})^2}\xrightarrow{P}0.$$

Combined with the bounds for terms (A'), (B') and (C'), we can finally obtain

$$\frac{T_{\gamma,J}^{\mathrm{GoF}}}{\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2\gamma j}\cdot(p_\psi-p_{0\psi})^2}\xrightarrow{P}1 \tag{20}$$

Now we provide bound to the term $\widetilde{S}_{\gamma,J}$. Recall that $|\widetilde{S}_{\gamma,J}-\widehat{S}_{\gamma,J}|\lesssim n^{-3}$ and equation (22), we consider

$$\begin{aligned}
\mathbb{E}[\widehat{S}_{\gamma,J}^2] &= \frac{2}{n(n-1)}\sum_{j_1=0}^{J}\sum_{j_2=0}^{J}\sum_{\psi_1\in\Psi_{j_1}}\sum_{\psi_2\in\Psi_{j_2}}2^{-2(j_1+j_2)\gamma}\cdot\Big(\mathbb{E}_{p_0}\big[\psi_1(X)\psi_2(X)\big]-p_{0\psi_1}p_{0\psi_2}\Big)^2\\
&\lesssim n^{-2}\cdot\sum_{j_1=0}^{J}\sum_{\psi_1\in\Psi_{j_1}}\sum_{j_2=j_1}^{J}\sum_{\psi_2\in\widetilde{\Psi}_{j_2}(\psi_1)}2^{-2(j_1+j_2)\gamma}\cdot\big(\mathbb{E}\big[\psi_1(X_1)\psi_1(X_2)\psi_2(X_1)\psi_2(X_2)\big]+p_{0\psi_1}^2p_{0\psi_2}^2\big)\\
&\lesssim n^{-2}\cdot\sum_{j_1=0}^{J}\sum_{j_2=j_1}^{J}2^{dj_2}\cdot2^{-2(j_1+j_2)\gamma}\cdot2^{-2dj_2}\cdot2^{d(j_1+j_2)}\\
&\lesssim\begin{cases}n^{-2}\cdot\frac{2^{J(d-4\gamma)}-1}{2^{d-4\gamma}-1} & \gamma>0\\ n^{-2}\cdot2^{dJ}\cdot J & \gamma=0\end{cases}\\
&=o(\widetilde{\Delta}_n^2).
\end{aligned}$$

So we have

$$\frac{\widehat{S}_{\gamma,J}}{\widetilde{\Delta}_n}\xrightarrow{P}0,$$

which leads to

$$\frac{\widehat{S}_{\gamma,J}}{\widetilde{\Delta}_n}\xrightarrow{P}0.$$

Combined with statement (20) and the fact that $\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2\gamma j}\cdot(p_\psi-p_{0\psi})^2\gtrsim\widetilde{\Delta}_n$, we can obtain the desired result.

D.2.3 Proof of Lemma 3

Under $\mathbb{H}_0^{\mathrm{GoF}}$, we have

$$\mathbb{E}(T_{\gamma,J}^{\mathrm{GoF}})=\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}\big(\mathbb{E}_p[\psi(X)-p_{0\psi}]\big)^2=0,$$

and

$$
\begin{aligned}
\mathrm{Var}(T^{\mathrm{GoF}}_{\gamma,J}) &= \frac{2}{n(n-1)}\mathbb{E}_{X_1,X_2\sim p_0}\Big[\Big(\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}\big(\psi(X_1)-p_{0\psi}\big)\cdot\big(\psi(X_2)-p_{0\psi}\big)\Big)^2\Big] \\
&= \frac{2}{n(n-1)}\sum_{j_1=0}^{J}\sum_{j_2=0}^{J}2^{-2(j_1+j_2)\gamma}\sum_{\psi_1\in\Psi_{j_1}}\sum_{\psi_2\in\Psi_{j_2}}\Big(\mathbb{E}_{p_0}\Big[\big(\psi_1(X)-p_{0\psi_1}\big)\cdot\big(\psi_2(X)-p_{0\psi_2}\big)\Big]\Big)^2 \\
&\overset{(i)}{\geq} \frac{2}{n(n-1)}\sum_{j_1=0}^{J}2^{-4j_1\gamma}\sum_{\psi_1\in\Psi_{j_1}}\Big(\mathbb{E}_{p_0}\big[\psi_1^2(X)\big]-p_{0\psi_1}^2\Big)^2 \\
&\geq \frac{2}{n(n-1)}\sum_{j=\lfloor\frac{J}{2}\rfloor}^{J}2^{-4j\gamma}\sum_{\psi\in\Psi_j}\Big(\mathbb{E}_{p_0}\big[\psi^2(X)\big]-p_{0\psi}^2\Big)^2 \\
&\geq \frac{1}{n(n-1)}\sum_{j=\lfloor\frac{J}{2}\rfloor}^{J}2^{-4j\gamma}\sum_{\psi\in\Psi_j}\big(\mathbb{E}_{p_0}\big[\psi^2(X)\big]\big)^2-\frac{2}{n(n-1)}\sum_{j=\lfloor\frac{J}{2}\rfloor}^{J}2^{-4j\gamma}\sum_{\psi\in\Psi_j}p_{0\psi}^4 \\
&\overset{(ii)}{\geq} \frac{1}{n(n-1)}\sum_{j=\lfloor\frac{J}{2}\rfloor}^{J}2^{-4j\gamma}\frac{\big(\mathbb{E}_{p_0}\big[\sum_{\psi\in\Psi_j}\psi^2(X)\big]\big)^2}{C2^{jd}}-C_1n^{-2} \\
&\gtrsim n^{-2}\cdot\frac{2^{J(d-4\gamma)}-2^{\frac{J}{2}(d-4\gamma)}}{2^{d-4\gamma}-1}.
\end{aligned}
\tag{21}
$$

where (i) is obtained by taking $j_1=j_2$, (ii) uses the uniform boundedness of p_0 that leads to $p_{0\psi}=\mathbb{E}_{p_0}[\psi(X)]\lesssim 2^{-\frac{dj}{2}}$ and $\sum_{j=0}^{\infty}p_{0\psi}^2\cdot 2^{2j\alpha}=O(1)$. Therefore, we can obtain

$$
\frac{\widetilde{S}^2_{\gamma,J}}{\mathrm{Var}(T^{\mathrm{GoF}}_{\gamma,J})}=\frac{\widehat{S}^2_{\gamma,J}}{\mathrm{Var}(T^{\mathrm{GoF}}_{\gamma,J})}+\frac{\widetilde{S}^2_{\gamma,J}-\widehat{S}^2_{\gamma,J}}{\mathrm{Var}(T^{\mathrm{GoF}}_{\gamma,J})}=\frac{\widehat{S}^2_{\gamma,J}}{\mathrm{Var}(T^{\mathrm{GoF}}_{\gamma,J})}+o(1),
$$

and we only need to prove

$$
\frac{\widehat{S}^2_{\gamma,J}}{\mathrm{Var}(T^{\mathrm{GoF}}_{\gamma,J})}\xrightarrow{P}1.
$$

Note that

$$
\begin{aligned}
\mathbb{E}[\widehat{S}^2_{\gamma,J}] &= \frac{2}{n(n-1)}\mathbb{E}_{X_1,X_2\sim p_0}\Big[\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2j\gamma}\psi(X_1)\psi(X_2)\Big)^2 \\
&\qquad -2\cdot\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2j\gamma}\psi(X_1)\cdot p_{0\psi}\Big)^2+\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2j\gamma}p_{0\psi}^2\Big)^2\Big] \\
&= \frac{2}{n(n-1)}\sum_{j_1=0}^{J}\sum_{j_2=0}^{J}\sum_{\psi_1\in\Psi_{j_1}}\sum_{\psi_2\in\Psi_{j_2}}2^{-2(j_1+j_2)\gamma}\cdot\Big\{\Big(\mathbb{E}_{p_0}\big[\psi_1(X)\psi_2(X)\big]\Big)^2 \\
&\qquad -2p_{0\psi_1}p_{0\psi_2}\mathbb{E}_{p_0}[\psi_1(X)\psi_2(X)]+p_{0\psi_1}^2p_{0\psi_2}^2\Big\} \\
&= \frac{2}{n(n-1)}\sum_{j_1=0}^{J}\sum_{j_2=0}^{J}\sum_{\psi_1\in\Psi_{j_1}}\sum_{\psi_2\in\Psi_{j_2}}2^{-2(j_1+j_2)\gamma}\cdot\Big(\mathbb{E}_{p_0}\big[\psi_1(X)\psi_2(X)\big]-p_{0\psi_1}p_{0\psi_2}\Big)^2 \\
&\overset{\text{under}H_0^{\mathrm{GoF}}}{=} \frac{2}{n(n-1)}\sum_{j_1=0}^{J}\sum_{j_2=0}^{J}\sum_{\psi_1\in\Psi_{j_1}}\sum_{\psi_2\in\Psi_{j_2}}2^{-2(j_1+j_2)\gamma}\cdot\Big(\mathbb{E}_{p_0}\Big[\big(\psi_1(X)-p_{0\psi_1}\big)\cdot\big(\psi_2(X)-p_{0\psi_2}\big)\Big]\Big)^2 \\
&= \mathrm{Var}(T^{\mathrm{GoF}}_{\gamma,J}).
\end{aligned}
\tag{22}
$$

So the estimator $\widehat{S}^2_{\gamma,J}$ is unbiased. Now we bound the variance of $\widehat{S}^2_{\gamma,J}$.

$$\begin{aligned}\mathrm{Var}(\widehat{S}^2_{\gamma,J}) \lesssim \frac{1}{n^2(n-2)^2}\Bigg\{ & \underbrace{n^{-2}\cdot\mathbb{E}_{X_1,X_2\sim p_0}\Big[\Big(\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(X_1)\psi(X_2)\Big)^4\Big]}_{(A)} \\ & + \underbrace{n^{-1}\cdot\mathbb{E}_{X_2\sim p_0}\Big[\Big(\sum_{j_1=0}^{J}\sum_{j_2=0}^{J}2^{-2(j_1+j_2)\gamma}\sum_{\psi_1\in\Psi_{j_1}}\sum_{\psi_2\in\Psi_{j_2}}\mathbb{E}_{p_0}\big[\psi_1(X_1)\psi_2(X_1)\big]\psi_1(X_2)\psi_2(X_2)\Big)^2\Big]}_{(B)} \\ & + \underbrace{n^{-1}\cdot\mathbb{E}_{p_0}\Big[\Big(\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(X)\cdot p_{0\psi}\Big)^4\Big]}_{(C)}\Bigg\}.\end{aligned}$$

Let

$$\widetilde{\Psi}_j(\psi) = \big\{\psi'\in\Psi_j \,:\, \mathrm{supp}(\psi)\cap\mathrm{supp}(\psi')\neq\emptyset\big\}. \tag{23}$$

Then for any $j\geq j'$ and $\psi\in\Psi_{j'}$, we have

$$\big|\widetilde{\Psi}_j(\psi)\big| \lesssim 2^{d(j-j')}.$$

Therefore, we can bound term (C) as

$$\begin{aligned}(C) &\leq \frac{24}{n}\sum_{j_1=0}^{J}\sum_{\psi_1\in\Psi_{j_1}}\sum_{j_2=j_1}^{J}\sum_{\psi_2\in\widetilde{\Psi}_{j_2}(\psi_1)}\sum_{j_3=j_2}^{J}\sum_{\psi_3\in\widetilde{\Psi}_{j_3}(\psi_2)}\sum_{j_4=j_3}^{J}\sum_{\psi_4\in\widetilde{\Psi}_{j_4}(\psi_3)}2^{-2(j_1+j_2+j_3+j_4)\gamma} \\ &\qquad\cdot\mathbb{E}_{p_0}\big[\psi_1(X)\psi_2(X)\psi_3(X)\psi_4(X)\big]\cdot p_{0\psi_1}p_{0\psi_2}p_{0\psi_3}p_{0\psi_4} \\ &\leq \frac{24}{n}\sum_{j_1=0}^{J}2^{dj_1}\sum_{j_2=j_1}^{J}2^{d(j_2-j_1)}\sum_{j_3=j_2}^{J}2^{d(j_3-j_2)}\sum_{j_4=j_3}^{J}2^{d(j_4-j_3)}\cdot 2^{-2(j_1+j_2+j_3+j_4)\gamma} \\ &\qquad\cdot 2^{-dj_4+\frac{d}{2}(j_1+j_2+j_3+j_4)}\cdot 2^{-\frac{d}{2}(j_1+j_2+j_3+j_4)} \\ &\lesssim \frac{2^{(2d-8\gamma)J}}{n}.\end{aligned}$$

Similarly, we can bound term (B) as

$$\begin{aligned}(B) &\lesssim n^{-1}\sum_{j_1=0}^{J}\sum_{\psi_1\in\Psi_{j_1}}\sum_{j_2=0}^{J}\sum_{\psi_2\in\Psi_{j_2}}\sum_{j_3=0}^{J}\sum_{\psi_3\in\Psi_{j_3}}\sum_{j_4=0}^{J}\sum_{\psi_4\in\Psi_{j_4}}2^{-2(j_1+j_2+j_3+j_4)\gamma} \\ &\qquad\cdot\mathbb{E}_{p_0}[\psi_1(X_1)\psi_2(X_1)]\cdot\mathbb{E}_{p_0}[\psi_3(X_1)\psi_4(X_1)]\cdot\mathbb{E}_{p_0}[\psi_1(X_2)\psi_2(X_2)\psi_3(X_2)\psi_4(X_2)] \\ &\lesssim n^{-1}\cdot\sum_{j_1=0}^{J}\sum_{\psi_1\in\Psi_{j_1}}\sum_{j_2=j_1}^{J}\sum_{\psi_2\in\widetilde{\Psi}_{j_2}(\psi_1)}\sum_{j_3=j_2}^{J}\sum_{\psi_3\in\widetilde{\Psi}_{j_3}(\psi_2)}\sum_{j_4=j_3}^{J}\sum_{\psi_4\in\widetilde{\Psi}_{j_4}(\psi_3)}2^{-2(j_1+j_2+j_3+j_4)\gamma} \\ &\qquad\cdot\mathbb{E}_{p_0}[\psi_1(X_1)\psi_2(X_1)]\cdot\mathbb{E}_{p_0}[\psi_3(X_1)\psi_4(X_1)]\cdot\mathbb{E}_{p_0}[\psi_1(X_2)\psi_2(X_2)\psi_3(X_2)\psi_4(X_2)] \\ &\quad + n^{-1}\cdot\sum_{j_1=0}^{J}\sum_{\psi_1\in\Psi_{j_1}}\sum_{j_3=j_1}^{J}\sum_{\psi_3\in\widetilde{\Psi}_{j_3}(\psi_1)}\sum_{j_2=j_3}^{J}\sum_{\psi_2\in\widetilde{\Psi}_{j_2}(\psi_3)}\sum_{j_4=j_2}^{J}\sum_{\psi_4\in\widetilde{\Psi}_{j_4}(\psi_2)}2^{-2(j_1+j_2+j_3+j_4)\gamma} \\ &\qquad\cdot\mathbb{E}_{p_0}[\psi_1(X_1)\psi_2(X_1)]\cdot\mathbb{E}_{p_0}[\psi_3(X_1)\psi_4(X_1)]\cdot\mathbb{E}_{p_0}[\psi_1(X_2)\psi_2(X_2)\psi_3(X_2)\psi_4(X_2)] \\ &\lesssim n^{-1}\cdot\sum_{j_1=0}^{J}\sum_{j_2=j_1}^{J}\sum_{j_3=j_2}^{J}\sum_{j_4=j_3}^{J}2^{dj_4}\cdot 2^{-2(j_1+j_2+j_3+j_4)\gamma}\cdot 2^{-d(j_4+j_2)}\cdot 2^{-dj_4}\cdot 2^{d(j_1+j_2+j_3+j_4)} \\ &\quad + n^{-1}\cdot\sum_{j_1=0}^{J}\sum_{j_3=j_1}^{J}\sum_{j_2=j_3}^{J}\sum_{j_4=j_2}^{J}2^{dj_4}\cdot 2^{-2(j_1+j_2+j_3+j_4)\gamma}\cdot 2^{-d(j_4+j_2)}\cdot 2^{-dj_4}\cdot 2^{d(j_1+j_2+j_3+j_4)} \\ &\lesssim \frac{J^2}{n}2^{(2d-8\gamma)J},\end{aligned}$$

where the last inequality uses $d \geq 4\gamma$, and the $\log n$ term occurs at the boundary $d = 4\gamma$. For the term (A),

$$\begin{aligned}(A) &\lesssim n^{-2} \cdot \sum_{j_1=0}^{J} \sum_{\psi_1 \in \Psi_{j_1}} \sum_{j_2=j_1}^{J} \sum_{\psi_2 \in \widetilde{\Psi}_{j_2}(\psi_1)} \sum_{j_3=j_2}^{J} \sum_{\psi_3 \in \widetilde{\Psi}_{j_3}(\psi_2)} \sum_{j_4=j_3}^{J} \sum_{\psi_4 \in \widetilde{\Psi}_{j_4}(\psi_3)} 2^{-2(j_1+j_2+j_3+j_4)\gamma} \\ &\quad \cdot \Big(\mathbb{E}_{p_0}\big[\psi_1(X)\psi_2(X)\psi_3(X)\psi_4(X)\big]\Big)^2 \\ &\lesssim n^{-2} \cdot \sum_{j_1=0}^{J} \sum_{j_2=j_1}^{J} \sum_{j_3=j_2}^{J} \sum_{j_4=j_3}^{J} 2^{dj_4} \cdot 2^{-2(j_1+j_2+j_3+j_4)\gamma} \Big(2^{-dj_4} \cdot 2^{\frac{d}{2}(j_1+j_2+j_3+j_4)}\Big)^2 \\ &\lesssim \frac{J}{n^2} \cdot 2^{J(3d-8\gamma)}.\end{aligned}$$

So, combine with the bound to terms (A), (B), and (C) and plug in $2^J \asymp n^{\frac{2}{4\alpha+d}}$, we have

$$\mathrm{Var}(\widehat{S}^2_{\gamma,J}) = o(n^{-4+\frac{4d-16\gamma}{4\alpha+d}}).$$

Combined with the upper bound (21) to $\mathrm{Var}(T^{\mathrm{GoF}}_{\gamma,J})$, we have

$$\mathrm{Var}(\widehat{S}^2_{\gamma,J}) = o\Big((\mathrm{Var}(T^{\mathrm{GoF}}_{\gamma,J}))^2\Big).$$

Then combined with the unbiasedness of $\widehat{S}^2_{\gamma}$, we can obtain the desired conclusion.

E Proof for Two-sample Test

E.1 Proof of Theorem 1

The proof of Theorem 1 directly follows from the result of Theorem 5 and the argument of the proof of Theorem 5 in Li and Yuan (2019).

E.2 Proof of Theorem 2

Throughout the proof, we denote

$$\widetilde{\Psi}_j(\psi) = \big\{\psi' \in \Psi_j \,:\, \mathrm{supp}(\psi) \cap \mathrm{supp}(\psi') \neq \emptyset\big\}. \tag{24}$$

We use X, Y to denote random variables from p and q respectively. We use $X_1, X_2, \cdots$ to denote independent random variables from p; and we use $Y_1, Y_2, \cdots$ to denote independent random variables from q. Without loss of generality, we assume $\Omega = [0,1]^d$. The proof contains two part: one part is about the normality under $\mathbb{H}_0$, another part is the power analysis. We first show the normality.

E.2.1 Proof of the normality under $\mathbb{H}_0$

Since $p, q \in \mathcal{W}^{u,\alpha}_L(\Omega)$, we can write

$$p = \sum_{j=0}^{J} \sum_{\psi \in \Psi_j} p_\psi \psi(x),$$
$$q = \sum_{j=0}^{J} \sum_{\psi \in \Psi_j} q_\psi \psi(x).$$

Similar to the proof of Theorem 4, we first show $\widetilde{\mathscr{S}}^2_{\gamma,J}$ is a valid approximate for the variance of $T_{\gamma,J}$ though the following lemma.

Lemma 5. *Under $\mathbb{H}_0$, the quantity $\frac{\widetilde{\mathscr{S}}^2_{\gamma,J}}{\mathrm{Var}(T_{\gamma,J})}$ converges in probability to 1 as n goes to infinity.*

The proof of Lemma 5 is given in Section E.2.3. Write

$$\widetilde{\mathscr{S}}_{\gamma,J}^{-1}T_{\gamma,J} = \frac{T_{\gamma,J}}{\sqrt{\mathrm{Var}(T_{\gamma,J})}} + \Big(\frac{\sqrt{\mathrm{Var}(T_{\gamma,J})}}{\widetilde{\mathscr{S}}_{\gamma,J}} - 1\Big)\cdot\frac{T_{\gamma,J}}{\sqrt{\mathrm{Var}(T_{\gamma,J})}}.$$

By Lemma 5, we only need to prove that

$$\frac{T_{\gamma,J}}{\sqrt{\mathrm{Var}(T_{\gamma,J})}} \xrightarrow{d} N(0,1). \tag{25}$$

Let

$$H(X_1,X_2) = \sum_{j=0}^{J} 2^{-2j\gamma}\sum_{\psi\in\Psi_j}\big(\psi(X_1)-q_\psi\big)\cdot\big(\psi(X_2)-q_\psi\big).$$

Without loss of generality, we can assume $n\geq m$. Then under $\mathbb{H}_0$, we can rewrite

$$\begin{aligned}
T_{\gamma,J} &= \frac{2}{n(n-1)}\sum_{1\leq i_1<i_2\leq n}H(X_{i_1},X_{i_2}) + \frac{2}{m(m-1)}\sum_{1\leq w_1<w_2\leq m}H(Y_{w_1},Y_{w_2}) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{w=1}^{m}H(X_i,Y_w)\\
&= \sum_{i=2}^{m}\sum_{j=1}^{i-1}\Big[\frac{2}{n(n-1)}H(X_i,X_j) + \frac{2}{m(m-1)}H(Y_i,Y_j) - \frac{2}{nm}\big(H(X_i,Y_j)+H(X_j,Y_i)\big)\Big]\\
&+ \sum_{i=m+1}^{n}\Big[\sum_{j=1}^{i-1}\frac{2}{n(n-1)}H(X_i,X_j) - \frac{2}{nm}\sum_{j=1}^{m}H(X_i,Y_j)\big)\Big] - \frac{2}{nm}\sum_{i=1}^{m}H(X_i,Y_i).
\end{aligned}$$

Then by a adaptation of the proof of Theorem 1 of Hall (1984), we have the following lemma.

Lemma 6. *Suppose $0<c\leq\frac{n}{m}\leq C<\infty$ for constants c, C, and*

$$\frac{\frac{2}{nm}\sum_{i=1}^{m}H(X_i,Y_i)}{\sqrt{\mathrm{Var}(T_{\gamma,J})}} = o_p(1);$$

$$\frac{\mathbb{E}[H^4(X_1,X_2)]}{n^2(\mathbb{E}[H^2(X_1,X_2)])^2}\to 0;$$

$$\frac{\mathbb{E}[H^2(X_1,X_2)H^2(X_1,X_3)]}{n(\mathbb{E}[H^2(X_1,X_2)])^2}\to 0;$$

$$\frac{\mathbb{E}[G^2(X_1,X_2)]}{(\mathbb{E}[H^2(X_1,X_2)])^2}\to 0, \text{ where } G(x,y)=\mathbb{E}[H(X,x)H(X,y)].$$

Then we have under $\mathbb{H}_0$, $\frac{T_{\gamma,J}}{\sqrt{\mathrm{Var}(T_{\gamma,J})}}\xrightarrow{d}N(0,1)$.

Note that

$$\mathbb{E}[\frac{2}{nm}\sum_{i=1}^{m}H(X_i,Y_i)] = 0$$

and

$$\begin{aligned}
\mathrm{Var}\Big(\frac{2}{nm}\sum_{i=1}^{m}H(X_i,Y_i)\Big) &\lesssim n^{-3}\mathrm{Var}[H(X_1,X_2)]\\
&\lesssim n^{-3}\cdot\sum_{j_1=0}^{J}\sum_{j_2=j_1}^{J}2^{dj_1}\cdot 2^{-2(j_1+j_2)\gamma}\\
&\lesssim \begin{cases} n^{-3}\cdot\frac{2^{J(d-4\gamma)}-2^{\frac{J}{2}(d-4\gamma)}}{2^{d-4\gamma}-1} & \gamma>0\\ n^{-3}\cdot 2^{dJ}\cdot J & \gamma=0\end{cases}\\
&= o\big(\mathrm{Var}(T_{\gamma,J})\big),
\end{aligned}$$

where the last inequality uses (28). So combined with equations (14), (15) and (16) in the proof of Theorem 4. We can obtain the desired result.

E.2.2 Proof for the Power analysis

Follow the proof of Theorem 4. When $d_{\gamma_1}(p,q) \geq \Delta_n$ with $\Delta_n^2 \cdot \delta_n(\gamma_1)^{-1} \to \infty$, we can obtain

$$\sum_{j=0}^{J} \sum_{\psi \in \Psi_j} 2^{-2\gamma_1 j} \cdot (p_\psi - q_\psi)^2 \gtrsim \Delta_n^2.$$

So if $\gamma_1 \geq \gamma$, we have

$$\sum_{j=0}^{J} \sum_{\psi \in \Psi_j} 2^{-2\gamma j} \cdot (p_\psi - q_\psi)^2 \gtrsim \Delta_n^2;$$

and if $\gamma_1 < \gamma$, we have

$$\sum_{j=0}^{J} \sum_{\psi \in \Psi_j} 2^{-2\gamma j} \cdot (p_\psi - q_\psi)^2 \gtrsim \Delta_n^2 \cdot 2^{-2J(\gamma - \gamma_1)};$$

Denote $\widetilde{\Delta}_n = \Delta_n^2 \cdot 2^{-2J(\gamma - \gamma_1 \wedge \gamma)}$. To show the desired result, we only need to prove that when $\sum_{j=0}^{J} \sum_{\psi \in \Psi_j} 2^{-2\gamma j} \cdot (p_\psi - q_\psi)^2 \gtrsim \widetilde{\Delta}_n$, $\widetilde{\mathscr{S}}_{\gamma,J}^{-1} T_{\gamma,J} \xrightarrow{P} \infty$.

Note that we can rewrite the statistic $T_{\gamma,J}$ as

$$\begin{aligned}
T_{\gamma,J} = & \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} \Big\{ \sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi \in \Psi_j} \big(\psi(X_{i_1}) - p_\psi\big) \cdot \big(\psi(X_{i_2}) - p_\psi\big) \Big\} \\
& + \frac{1}{m(m-1)} \sum_{w_1 \neq w_2} \Big\{ \sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi \in \Psi_j} \big(\psi(Y_{w_1}) - q_\psi\big) \cdot \big(\psi(Y_{w_2}) - q_\psi\big) \Big\} \\
& + \sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi \in \Psi_j} (p_\psi - q_\psi)^2 + \frac{2}{n} \sum_{i=1}^{n} \sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi \in \Psi_j} (p_\psi - q_\psi) \cdot (\psi(X_i) - p_\psi) \\
& + \frac{2}{m} \sum_{w=1}^{m} \sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi \in \Psi_j} (q_\psi - p_\psi) \cdot (\psi(Y_w) - q_\psi) \\
& \underbrace{- \frac{2}{nm} \sum_{i=1}^{n} \sum_{w=1}^{m} \Big\{ \sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi \in \Psi_j} \big(\psi(X_i) - p_\psi\big) \cdot \big(\psi(Y_w) - q_\psi\big) \Big\}}_{(F)}
\end{aligned}$$

By tracking the proof of Theorem 4 in Section D.2.2, it remains to show

$$\frac{(F)}{\widetilde{\Delta}_n} \xrightarrow{P} 0, \tag{26}$$

and

$$\frac{\widehat{\mathscr{S}_{\gamma,J}}}{\widetilde{\Delta}_n} \xrightarrow{P} 0. \tag{27}$$

Since

$$E[(F)] = 0,$$

and

$$\begin{aligned}
\mathrm{Var}(F) &= \frac{4}{nm}\cdot\mathbb{E}\Big[\Big(\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}\big(\psi(X)-p_\psi\big)\cdot\big(\psi(Y)-q_\psi\big)\Big)^2\Big]\\
&\lesssim n^{-2}\sum_{j_1=0}^{J}\sum_{j_2=j_1}^{J}2^{dj_2}\cdot 2^{-2(j_1+j_2)\gamma}\cdot 2^{-2dj_2}\cdot 2^{d(j_1+j_2)}\\
&\lesssim \begin{cases} n^{-2}\cdot\frac{2^{J(d-4\gamma)}-1}{2^{d-4\gamma}-1} & \gamma>0\\ n^{-2}\cdot 2^{dJ}\cdot J & \gamma=0\end{cases}\\
&= o(\widetilde{\Delta}_n^2),
\end{aligned}$$

which proves statement (26). Now we provide bound to the term $\widehat{\mathscr{S}}_{\gamma,J}$. Recall $\overline{\mathscr{S}}^2_{\gamma,J}$ defined in equation (29), we have

$$\begin{aligned}
&\Big|\mathbb{E}[\widehat{\mathscr{S}}^2_\gamma]-\mathbb{E}[\overline{\mathscr{S}}^2_\gamma]\Big|\\
&\lesssim 2\big(\frac{1}{n(n-1)}+\frac{1}{m(m-1)}+\frac{2}{mn}\big)\\
&\cdot\Big\{\sum_{j_1=0}^{J}\sum_{j_2=0}^{J}2^{-2(j_1+j_2)\gamma}\sum_{\psi_1\in\Psi_{j_1}}\sum_{\psi_2\in\Psi_{j_2}}\Big[\frac{1}{m}\cdot\mathbb{E}[\psi_1(X)\psi_2(X)]\cdot\big(q_{\psi_1}q_{\psi_2}+\mathbb{E}[\psi_1(Y)\psi_2(Y)]\big)\\
&\qquad+\frac{1}{n}\cdot\mathbb{E}[\psi_1(Y)\psi_2(Y)]\cdot\big(p_{\psi_1}p_{\psi_2}+\mathbb{E}[\psi_1(X)\psi_2(X)]\big)\Big]+\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2j\gamma}p_\psi q_\psi\Big)^2\Big\}\\
&\lesssim n^{-3}\sum_{j_1=0}^{J}\sum_{j_2=j_1}^{J}\big[2^{dj_2}\cdot 2^{-2(j_1+j_2)\gamma}\cdot 2^{-2dj_2}\cdot 2^{d(j_1+j_2)}\big]+n^{-2}\\
&= o(\widetilde{\Delta}_n^2)
\end{aligned}$$

Therefore, we only need to consider

$$\begin{aligned}
\mathbb{E}[\overline{\mathscr{S}}^2_\gamma] &= 2\big(\frac{1}{n(n-1)}+\frac{1}{m(m-1)}+\frac{2}{mn}\big)\cdot\mathbb{E}\Big[\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2j\gamma}\psi(X)\psi(Y)\Big)^2\\
&\qquad-\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2j\gamma}\psi(X)\cdot q_\psi\Big)^2-\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2j\gamma}\psi(Y)\cdot p_\psi\Big)^2+\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2j\gamma}p_\psi q_\psi\Big)^2\Big]\\
&= 2\big(\frac{1}{n(n-1)}+\frac{1}{m(m-1)}+\frac{2}{mn}\big)\cdot\mathbb{E}\Big[\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2j\gamma}(\psi(X)-p_\psi)(\psi(Y)-q_\psi)\Big)^2\Big]\\
&\lesssim n^{-2}\cdot\mathbb{E}\Big[\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2j\gamma}\cdot\psi(X)\cdot\psi(Y)\Big)^2\Big]\\
&\lesssim n^{-2}\cdot\sum_{j_1=0}^{J}\sum_{\psi_1\in\Psi_{j_1}}\sum_{j_2=j_1}^{J}\sum_{\psi_2\in\widetilde{\Psi}_{j_2}(\psi_1)}2^{-2(j_1+j_2)\gamma}\cdot\mathbb{E}\big[\psi_1(X)\psi_1(Y)\psi_2(X)\psi_2(Y)\big]\\
&\lesssim n^{-2}\cdot\sum_{j_1=0}^{J}\sum_{j_2=j_1}^{J}2^{dj_2}\cdot 2^{-2(j_1+j_2)\gamma}\cdot 2^{-2dj_2}\cdot 2^{d(j_1+j_2)}\\
&\lesssim \begin{cases} n^{-2}\cdot\frac{2^{J(d-4\gamma)}-1}{2^{d-4\gamma}-1} & \gamma>0\\ n^{-2}\cdot 2^{dJ}\cdot J & \gamma=0\end{cases}\\
&= o(\widetilde{\Delta}_n^2),
\end{aligned}$$

which leads to statement (27). We can then obtain the desired result.

E.2.3 Proof of Lemma 5

Under $\mathbb{H}_0$, we have

$$\mathbb{E}(T_{\gamma,J}) = \sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi\in\Psi_j} \big(p_\psi - q_\psi\big)^2 = 0,$$

and

$$\begin{aligned}
\mathrm{Var}(T_{\gamma,J}) = \mathbb{E}\Bigg[\Big(&\frac{1}{n(n-1)} \sum_{i_1\neq i_2} \sum_{j=0}^{J} \sum_{\psi\in\Psi_j} 2^{-2j\gamma} \cdot (\psi(X_{i_1}) - p_\psi) \cdot (\psi(X_{i_2}) - p_\psi) \\
&+ \frac{1}{m(m-1)} \sum_{w_1\neq w_2} \sum_{j=0}^{J} \sum_{\psi\in\Psi_j} 2^{-2j\gamma} \cdot (\psi(Y_{w_1}) - q_\psi) \cdot (\psi(Y_{w_2}) - q_\psi) \\
&- \frac{2}{nm} \sum_{i=1}^{n} \sum_{w=1}^{m} \sum_{j=0}^{J} \sum_{\psi\in\Psi_j} 2^{-2j\gamma} \cdot (\psi(X_i) - p_\psi) \cdot (\psi(Y_w) - q_\psi)\Big)^2\Bigg] \\
= &\frac{2}{n(n-1)} \cdot \mathbb{E}_{X_1,X_2\sim p}\Bigg[\Big(\sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi\in\Psi_j} \big(\psi(X_1) - p_\psi\big) \cdot \big(\psi(X_2) - p_\psi\big)\Big)^2\Bigg] \\
&+ \frac{2}{m(m-1)} \cdot \mathbb{E}_{Y_1,Y_2\sim q}\Bigg[\Big(\sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi\in\Psi_j} \big(\psi(Y_1) - q_\psi\big) \cdot \big(\psi(Y_2) - q_\psi\big)\Big)^2\Bigg] \\
&+ \frac{4}{nm} \cdot \mathbb{E}_{X\sim p,Y\sim q}\Bigg[\Big(\sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi\in\Psi_j} \big(\psi(X) - p_\psi\big) \cdot \big(\psi(Y) - q_\psi\big)\Big)^2\Bigg] \\
= &\Big(\frac{2}{n(n-1)} + \frac{2}{m(m-1)} + \frac{4}{nm}\Big) \\
&\cdot \sum_{j_1=0}^{J} \sum_{j_2=0}^{J} 2^{-2(j_1+j_2)\gamma} \sum_{\psi_1\in\Psi_{j_1}} \sum_{\psi_2\in\Psi_{j_2}} \Big(\mathbb{E}_{p_0}\Big[\big(\psi_1(X) - p_{\psi_1}\big) \cdot \big(\psi_2(X) - p_{\psi_2}\big)\Big]\Big)^2 \\
\geq &\Big(\frac{2}{n(n-1)} + \frac{2}{m(m-1)} + \frac{4}{nm}\Big) \sum_{j_1=0}^{J} 2^{-4j_1\gamma} \sum_{\psi_1\in\Psi_{j_1}} \Big(\mathbb{E}_{p_0}\big[\psi_1^2(X)\big] - p_{\psi_1}^2\Big)^2 \\
\gtrsim &\Big(\frac{2}{n(n-1)} + \frac{2}{m(m-1)} + \frac{4}{nm}\Big) \cdot \frac{2^{J(d-4\gamma)} - 2^{\frac{J}{2}(d-4\gamma)}}{2^{d-4\gamma} - 1},
\end{aligned} \tag{28}$$

where the last inequality is obtained by using the same strategy as in (21). Therefore, we can obtain

$$\frac{\widetilde{\mathscr{S}}^2_{\gamma,J}}{\mathrm{Var}(T_{\gamma,J})} = \frac{\widehat{\mathscr{S}}^2_{\gamma,J}}{\mathrm{Var}(T_{\gamma,J})} + \frac{\widetilde{\mathscr{S}}^2_{\gamma,J} - \widehat{\mathscr{S}}^2_{\gamma,J}}{\mathrm{Var}(T_{\gamma,J})} = \frac{\widehat{\mathscr{S}}^2_{\gamma,J}}{\mathrm{Var}(T_{\gamma,J})} + o(1),$$

and we only need to prove

$$\frac{\widehat{\mathscr{S}}^2_{\gamma,J}}{\mathrm{Var}(T_{\gamma,J})} \xrightarrow{P} 1.$$

Denote

$$\begin{aligned}
\overline{\mathscr{S}}^2_{\gamma,J} = 2\Big(&\frac{1}{n(n-1)} + \frac{1}{m(m-1)} + \frac{2}{mn}\Big) \cdot \Bigg\{\frac{1}{nm} \sum_{i_1=1}^{n} \sum_{i_2=1}^{m} \Big(\sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi\in\Psi_j} \psi(X_{i_1})\psi(Y_{i_2})\Big)^2 \\
&- \frac{1}{n} \sum_{i_1=1}^{n} \Big[\sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi\in\Psi_j} \psi(X_{i_1}) \cdot q_\psi\Big]^2 - \frac{1}{m} \sum_{i_2=1}^{m} \Big[\sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi\in\Psi_j} \psi(Y_{i_2}) \cdot p_\psi\Big]^2 \\
&+ \Big(\sum_{j=0}^{J} \sum_{\psi\in\Psi_j} 2^{-2j\gamma} \cdot q_\psi \cdot p_\psi\Big)^2\Bigg\}.
\end{aligned} \tag{29}$$

We now show that $\widehat{\mathscr{S}}^2_{\gamma,J}$ is close to $\overline{\mathscr{S}}^2_{\gamma,J}$. Note that

$$\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^n\Big[\sum_{j=0}^J 2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(X_i)\cdot\Big(\frac{1}{m}\sum_{w=1}^m\psi(Y_w)\Big)\Big]^2-\frac{1}{n}\sum_{i=1}^n\Big[\sum_{j=0}^J 2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(X_i)\cdot q_\psi\Big]^2\Big|\\
&=\Big|\frac{1}{n}\sum_{i=1}^n\Big[\sum_{j=0}^J 2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(X_i)\cdot\Big(\frac{1}{m}\sum_{w=1}^m\psi(Y_w)-q_\psi\Big)\Big]\\
&\qquad\cdot\Big[\sum_{j=0}^J 2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(X_i)\cdot\Big(\frac{1}{m}\sum_{w=1}^m\psi(Y_w)+q_\psi\Big)\Big]\Big|
\end{aligned}$$

Since for any $j\in\{0,1,\cdots,J\}$, $\psi(Y)\lesssim 2^{\frac{dj}{2}}$ and by the uniform boundedness of q, we have $\mathbb{E}_q[\psi(Y)^2]=O(1)$ and $\mathbb{E}_q[\psi(Y)]=O(2^{-\frac{dj}{2}})$. Then by Bernstein's inequality and a union bound, we have it holds with probability at least $1-n^{-1}$ that for any $j\in\{0,1,\cdots,J\}$ and $\psi\in\Psi_j$,

$$\Big|\frac{1}{m}\sum_{w=1}^m\psi(Y_w)-q_\psi\Big|\lesssim\sqrt{\frac{\log n}{n}}+\frac{\log n}{n}\cdot 2^{\frac{dj}{2}}.$$

Therefore it holds with probability at least $1-n^{-1}$ that

$$\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^n\Big[\sum_{j=0}^J 2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(X_i)\cdot\Big(\frac{1}{m}\sum_{w=1}^m\psi(Y_w)\Big)\Big]^2-\frac{1}{n}\sum_{i=1}^n\Big[\sum_{j=0}^J 2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(X_i)\cdot q_\psi\Big]^2\Big|\\
&\lesssim\frac{1}{n}\sum_{i=1}^n\Big[\sum_{j=0}^J 2^{-2j\gamma}\sum_{\psi\in\Psi_j}|\psi(X_i)|\cdot\Big(\sqrt{\frac{\log n}{n}}+\frac{\log n}{n}\cdot 2^{\frac{dj}{2}}\Big)\Big]\\
&\qquad\cdot\Big[\sum_{j=0}^J 2^{-2j\gamma}\sum_{\psi\in\Psi_j}|\psi(X_i)|\cdot\Big(\sqrt{\frac{\log n}{n}}+\frac{\log n}{n}\cdot 2^{\frac{dj}{2}}+2^{-\frac{dj}{2}}\Big)\Big]\\
&\lesssim\underbrace{\frac{1}{n}\sum_{i=1}^n\sum_{j_1=0}^J\sum_{\psi_1\in\Psi_{j_1}}\sum_{j_2=0}^J\sum_{\psi_2\in\widetilde{\Psi}_{j_2}(\psi_1)}2^{-2(j_1+j_2)\gamma}\Big(\frac{\log n}{n}+(\frac{\log n}{n})^2 2^{\frac{d(j_1+j_2)}{2}}+\sqrt{\frac{\log n}{n}}2^{-\frac{d}{2}j_2}\Big)|\psi_1(X_i)||\psi_2(X_i)|}_{(D)}.
\end{aligned}$$

Then we have

$$\begin{aligned}
E[(D)]&\lesssim\sum_{j_2=0}^J\sum_{j_1\geq j_2}^J 2^{dj_1}\cdot 2^{-2(j_1+j_2)\gamma}\Big(\frac{\log n}{n}+(\frac{\log n}{n})^2 2^{\frac{d(j_1+j_2)}{2}}+\sqrt{\frac{\log n}{n}}2^{-\frac{d}{2}j_2}\Big)\cdot 2^{-dj_1}\cdot 2^{\frac{d(j_1+j_2)}{2}}\\
&=o\Big(\frac{2^{J(d-4\gamma)}-2^{\frac{J}{2}(d-4\gamma)}}{2^{d-4\gamma}-1}\Big)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}(D)&\lesssim n^{-1}\sum_{j_1=0}^J\sum_{\psi_1\in\Psi_{j_1}}\sum_{j_2=j_1}^J\sum_{\psi_2\in\widetilde{\Psi}_{j_2}(\psi_1)}\sum_{j_3=j_2}^J\sum_{\psi_3\in\widetilde{\Psi}_{j_3}(\psi_2)}\sum_{j_4=j_3}^J\sum_{\psi_4\in\widetilde{\Psi}_{j_4}(\psi_3)}2^{-2(j_1+j_2+j_3+j_4)\gamma}\\
&\qquad\cdot\Big(\frac{\log n}{n}+(\frac{\log n}{n})^2 2^{\frac{d(j_1+j_2)}{2}}+\sqrt{\frac{\log n}{n}}2^{-\frac{d}{2}j_2}\Big)^2\cdot\mathbb{E}\big[|\psi_1(X)\psi_2(X)\psi_3(X)\psi_4(X)|\big]\\
&\lesssim n^{-1}\sum_{j_1=0}^J\sum_{j_2=j_1}^J\sum_{j_3=j_2}^J\sum_{j_4=j_3}^J 2^{dj_4}\cdot 2^{-2(j_1+j_2+j_3+j_4)\gamma}\cdot 2^{-dj_4}\cdot 2^{\frac{d(j_1+j_2+j_3+j_4)}{2}}\\
&=o\Big(\Big(\frac{2^{J(d-4\gamma)}-2^{\frac{J}{2}(d-4\gamma)}}{2^{d-4\gamma}-1}\Big)^2\Big).
\end{aligned}\tag{30}$$

Therefore, we can get

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\Big[\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(X_i)\cdot\Big(\frac{1}{m}\sum_{w=1}^{m}\psi(Y_w)\Big)\Big]^2-\frac{1}{n}\sum_{i=1}^{n}\Big[\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(X_i)\cdot q_\psi\Big]^2\Big|$$
$$=o_p\Big(\frac{2^{J(d-4\gamma)}-2^{\frac{J}{2}(d-4\gamma)}}{2^{d-4\gamma}-1}\Big).$$

Similarly, we can show

$$\Big|\frac{1}{n}\sum_{w=1}^{m}\Big[\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(Y_w)\cdot\Big(\frac{1}{n}\sum_{i=1}^{n}\psi(X_i)\Big)\Big]^2-\frac{1}{m}\sum_{w=1}^{m}\Big[\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(Y_w)\cdot p_\psi\Big]^2\Big|$$
$$=o_p\Big(\frac{2^{J(d-4\gamma)}-2^{\frac{J}{2}(d-4\gamma)}}{2^{d-4\gamma}-1}\Big).$$

Moreover, we have

$$\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2j\gamma}\cdot q_\psi\cdot p_\psi\Big)^2=O(1)=o\Big(\frac{2^{J(d-4\gamma)}-2^{\frac{J}{2}(d-4\gamma)}}{2^{d-4\gamma}-1}\Big). \tag{31}$$

Then combined with the upper bound (28) to $\mathrm{Var}(T_{\gamma,J})$, we have obtain

$$\frac{\widehat{\mathscr{S}}^2_{\gamma,J}-\overline{\mathscr{S}}^2_{\gamma,J}}{\mathrm{Var}(T_{\gamma,J})}\xrightarrow{P}0.$$

Therefore, it remains to show

$$\frac{\overline{\mathscr{S}}^2_{\gamma,J}}{\mathrm{Var}(T_{\gamma,J})}\xrightarrow{P}1.$$

Note that under $\mathbb{H}_0$,

$$\begin{aligned}\mathbb{E}[\overline{\mathscr{S}}^2_{\gamma,J}]&=2\,\big(\frac{1}{n(n-1)}+\frac{1}{m(m-1)}+\frac{2}{mn}\big)\cdot\mathbb{E}\Big[\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2j\gamma}\psi(X)\psi(Y)\Big)^2\\&\quad-\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2j\gamma}\psi(X)\cdot q_\psi\Big)^2-\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2j\gamma}\psi(Y)\cdot p_\psi\Big)^2+\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2j\gamma}p_\psi q_\psi\Big)^2\Big]\\&\overset{\text{under }\mathbb{H}_0}{=}2\,\big(\frac{1}{n(n-1)}+\frac{1}{m(m-1)}+\frac{2}{mn}\big)\cdot\mathbb{E}\Big[\Big(\sum_{j=0}^{J}\sum_{\psi\in\Psi_j}2^{-2j\gamma}(\psi(X_1)-p_\psi)(\psi(X_2)-p_\psi)\Big)^2\Big]\\&=\mathrm{Var}(T_{\gamma,J}).\end{aligned} \tag{32}$$

So the estimator $\overline{\mathscr{S}}^2_{\gamma,J}$ is unbiased. Now we bound the variance of $\overline{\mathscr{S}}^2_{\gamma,J}$.

$$\begin{aligned}\mathrm{Var}(\overline{\mathscr{S}}^2_{\gamma,J})&\lesssim n^{-4}\Big\{\frac{1}{nm}\cdot\mathbb{E}\Big[\Big(\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(X_1)\psi(X_2)\Big)^4\Big]\\&\quad+(n^{-1}+m^{-1})\cdot\mathbb{E}\Big[\Big(\sum_{j=0}^{J}2^{-2j\gamma}\sum_{\psi\in\Psi_j}\psi(X)\cdot p_\psi\Big)^4\Big]\Big\}\lesssim\frac{J}{n^6}2^{J(3d-8\gamma)}+\frac{J^2}{n^5}2^{J(2d-8\gamma)},\end{aligned}$$

where the last inequality uses the bounds of terms (A) and (C) in the proof of Lemma 3. So plug in $2^J\asymp n^{\frac{2}{4\alpha+d}}$, we have

$$\mathrm{Var}(\widehat{\mathscr{S}}^2_{\gamma,J})=o(n^{-4+\frac{4d-16\gamma}{4\alpha+d}}).$$

Combined with the upper bound (28) to $\mathrm{Var}(T_{\gamma,J})$, we have

$$\mathrm{Var}(\widehat{\mathscr{S}}^2_{\gamma,J})=o\Big((\mathrm{Var}(T_{\gamma,J}))^2\Big).$$

Then combined with the unbiasedness of $\widehat{\mathscr{S}}^2_{\gamma}$, we can obtain the desired conclusion.

F Proof of Technical Results

F.1 Proof of Lemma 1

For the left hand side, for any $f \in \mathcal{W}_1^{\gamma_1}(\mathbb{R}^d)$ and $\psi \in \overline{\Psi}_j$,denote $f_\psi = \int f(x)\psi(x)\,\mathrm{d}x$. We have

$$\begin{aligned}
d_\gamma^W(p,q) &= \sup_{f\in\mathcal{W}_1^{\gamma_1}(\mathbb{R}^d)} \int f\,\mathrm{d}p - \int f\,\mathrm{d}q \\
&= \sup_{f\in\mathcal{W}_1^{\gamma_1}(\mathbb{R}^d)} \sum_{j=0}^{\infty}\sum_{\psi\in\overline{\Psi}_j} (p_\psi - q_\psi)\cdot f_\psi \\
&\lesssim \sqrt{\sum_{j=0}^{\infty}\sum_{\psi\in\overline{\Psi}_j}(p_\psi-q_\psi)^2\cdot 2^{-2j\gamma}}\sqrt{\sup_{f\in\mathcal{W}_1^{\gamma_1}(\mathbb{R}^d)}\sum_{j=0}^{\infty}\sum_{\psi\in\overline{\Psi}_j} 2^{2j\gamma}\cdot f_\psi^2} \\
&\lesssim \sqrt{\sum_{j=0}^{\infty}\sum_{\psi\in\overline{\Psi}_j} 2^{-2\gamma j}\cdot(p_\psi-q_\psi)^2} = \|p-q\|_{\mathcal{B}_{2,2}^{-\gamma}}.
\end{aligned}$$

For the right hand side, consider

$$f_\psi = \frac{2^{-2j\gamma}(p_\psi - q_\psi)}{\sqrt{\sum_{j=0}^{\infty}\sum_{\psi\in\overline{\Psi}_j} 2^{-2\gamma j}(p_\psi-q_\psi)^2}}$$

and $f = \sum_{j=0}^{\infty} f_\psi\psi(x)$. We have

$$\|f\|_{\mathcal{B}_{2,2}^{\gamma}}^2 = \sum_{j=0}^{\infty}\sum_{\psi\in\overline{\Psi}_j} 2^{2j\gamma}\cdot f_\psi^2 = \sum_{j=0}^{\infty}\sum_{\psi\in\overline{\Psi}_j}\frac{2^{-2\gamma j}(p_\psi-q_\psi)^2}{\sum_{j=0}^{\infty}\sum_{\psi\in\overline{\Psi}_j}2^{-2\gamma j}(p_\psi-q_\psi)^2} = 1.$$

So we have

$$\begin{aligned}
d_\gamma^W(p,q) &\gtrsim \int f\,\mathrm{d}p - \int f\,\mathrm{d}q \\
&= \sum_{j=0}^{\infty}\sum_{\psi\in\overline{\Psi}_j}(p_\psi-q_\psi)\cdot f_\psi \\
&= \frac{\sum_{j=0}^{\infty}\sum_{\psi\in\overline{\Psi}_j}2^{-2\gamma j}(p_\psi-q_\psi)^2}{\sqrt{\sum_{j=0}^{\infty}\sum_{\psi\in\Psi_j}2^{-2\gamma j}(p_\psi-q_\psi)^2}} \\
&= \sqrt{\sum_{j=0}^{\infty}\sum_{\psi\in\Psi_j}2^{-2\gamma j}(p_\psi-q_\psi)^2} = \|p-q\|_{\mathcal{B}_{2,2}^{-\gamma}}.
\end{aligned}$$

F.2 Proof of Lemma 2

For the first statement, under $\mathbb{H}_0$, we have

$$\mathbb{E}(T_{\gamma,J}) = \sum_{j=0}^{J} 2^{-2j\gamma}\sum_{\psi\in\Psi_j}\big(\mathbb{E}_p[\psi(X)-q_\psi]\big)^2 = 0,$$

and

$$\begin{aligned}
\mathrm{Var}(T_{\gamma,J}) &= \Big(\frac{2}{n(n-1)} + \frac{2}{m(m-1)} + \frac{4}{nm}\Big) \\
&\qquad \cdot \mathbb{E}_{X_1,X_2\sim p_0}\Big[\Big(\sum_{j=0}^{J} 2^{-2j\gamma} \sum_{\psi\in\Psi_j} \big(\psi(X_1)-q_\psi\big)\cdot\big(\psi(X_2)-q_\psi\big)\Big)^2\Big] \\
&\lesssim n^{-2}\cdot \sum_{j_1=0}^{J}\sum_{j_2=0}^{J} 2^{-2(j_1+j_2)\gamma} \sum_{\psi_1\in\Psi_{j_1}}\sum_{\psi_2\in\Psi_{j_2}} \Big(\mathbb{E}_{p_0}\Big[\big(\psi_1(X)-q_{\psi_1}\big)\cdot\big(\psi_2(X)-q_{\psi_2}\big)\Big]\Big)^2 \\
&= n^{-2}\cdot \sum_{j_1=0}^{J}\sum_{j_2=0}^{J} 2^{-2(j_1+j_2)\gamma} \sum_{\psi_1\in\Psi_{j_1}}\sum_{\psi_2\in\Psi_{j_2}} \Big(\mathbb{E}_{p_0}\big[\psi_1(X)\cdot\psi_2(X)\big] - q_{\psi_1}q_{\psi_2}\Big)^2 \\
&\lesssim n^{-2}\cdot \sum_{j_1=0}^{J}\sum_{j_2=0}^{J} 2^{-2(j_1+j_2)\gamma} \sum_{\psi_1\in\Psi_{j_1}}\sum_{\psi_2\in\Psi_{j_2}} \Big(\mathbb{E}_{p_0}\big[\psi_1(X)\cdot\psi_2(X)\big]\Big)^2 + C\,n^{-2} \\
&\overset{(i)}{\lesssim} n^{-2}\sum_{j_1=0}^{J}\sum_{j_2\geq j_1}^{J} 2^{-2(j_1+j_2)\gamma}\cdot 2^{dj_2}\cdot 2^{d(j_1+j_2)}\cdot 2^{-2dj_2} \\
&= o(1)
\end{aligned}$$

where (i) uses the same strategy as that for bounding term (B) in the proof of Lemma 3. We can then get the first statement. For the second statement, by equation (19), we have

$$\sum_{j=0}^{J}\sum_{\psi\in\Psi_j} 2^{-2\gamma j}\cdot(p_\psi - q_\psi)^2 \gtrsim d_\gamma^W(p,p_0),$$

and

$$\sum_{j=J}^{\infty}\sum_{\psi\in\Psi_j} 2^{-2\gamma j}\cdot(p_\psi - q_\psi)^2 \lesssim n^{-\frac{4(\alpha+\gamma)}{4\alpha+d}}.$$

Then by $d_\gamma^W(p,q)\cdot n^{-\frac{2(\alpha+\gamma)}{4\alpha+d}} \to \infty$, we only need to prove

$$\frac{T_{\gamma,J}}{\sum_{j=0}^{J}\sum_{\psi\in\Psi_j} 2^{-2j\gamma}(p_\psi-q_\psi)^2} \xrightarrow{P} 1,$$

which directly follows from equation (20) and equation (26).

F.3 Proof of Lemma 4

The proof is a slight adaptation of the proof of Theorem 1 of Hall (1984). We include it here for completeness. Set $Y_{ni} = \sum_{j=1}^{i-1} H(X_i, X_j)$. By applying Brown's Martingale central limit theory. We only need to check two conditions

$$s_n^{-2}\sum_{i=2}^{n}\mathbb{E}\Big\{Y_{ni}^2\mathbf{1}\big(|Y_{ni}|>\varepsilon s_n\big)\Big\}\to 0 \tag{33}$$

as $n\to\infty$ for each $\varepsilon>0$, where $s_n^2 = \mathbb{E}[(\sum_{i\neq j}H(X_i,X_j))^2]$, and

$$s_n^{-2}V_n^2 \xrightarrow{P} 1 \tag{34}$$

as $n\to\infty$, where

$$V_n^2 = \sum_{i=2}^{n}\mathbb{E}[Y_{ni}^2\,|\,X_1,\cdots,X_{i-1}].$$

Since H is symmetric and $\mathbb{E}[H(X_1,X_2)|X_1]=0$, it has been shown in the proof of Theorem 1 of Hall (1984) that

$$s_n^2 = \frac{1}{2}n(n-1)\mathbb{E}[H^2(X_1,X_2)],$$

and

$$\sum_{i=2}^{n}\mathbb{E}[Y_{ni}^4] \lesssim n^2\cdot\mathbb{E}[H^4(X_1,X_2)] + n^3\cdot\mathbb{E}[H^2(X_1,X_2)H^2(X_1,X_3)],$$

which combines with conditions (14) and (15) can lead to $s_n^{-4}\sum_{i=2}^{n}\mathbb{E}[Y_{ni}^2]\to 0$ that implies (33). Moreover, it's shown in the proof of Theorem 1 of Hall (1984) that

$$\mathbb{E}[(V_n^2 - s_n^2)^2] \lesssim n^4\cdot\mathbb{E}[G^2(X_1,X_2)] + n^3\cdot\mathbb{E}[H^2(X_1,X_2)H^2(X_1,X_3)],$$

which combines with conditions (15) and (16) leads to $s_n^{-4}\mathbb{E}[(V_n^2 - s_n^2)^2]\to 0$ that implies (34). Proof is completed.

F.3.1 Proof of Lemma 6

The proof follows the proof of Theorem 1 of Hall (1984). Set

$$Y_{ni} = \begin{cases} \sum_{j=1}^{i-1}\Big(\frac{2}{n(n-1)}H(X_i,X_j) + \frac{2}{m(m-1)}H(Y_i,Y_j) - \frac{2}{nm}\big(H(X_i,Y_j)+H(X_j,Y_i)\big)\Big) & 2\le i\le m,\\ \sum_{j=1}^{i-1}\frac{2}{n(n-1)}H(X_i,X_j) - \frac{2}{nm}\sum_{j=1}^{m}H(X_i,Y_j) & m<i\le n.\end{cases}$$

Set $\overline{T}_{\gamma,J} = \sum_{i=2}^{n} Y_{ni}$. Then by the condition

$$\frac{\frac{2}{nm}\sum_{i=1}^{m}H(X_i,Y_i)}{\sqrt{\mathrm{Var}(T_{\gamma,J})}} = o_p(1),$$

it remains to prove $\frac{\overline{T}_{\gamma,J}}{\sqrt{\mathrm{Var}(\overline{T}_{\gamma,J})}}\xrightarrow{d} N(0,1)$. By applying Brown's Martingale central limit theory (see for example Corollary 3.1 of Hall and Heyde (1980)). We only need to check the following two conditions:

$$s_n^{-2}\sum_{i=2}^{n}\mathbb{E}\Big\{Y_{ni}^2\mathbf{1}\big(|Y_{ni}|>\varepsilon s_n\big)\Big\}\to 0 \tag{35}$$

as $n\to\infty$ for each $\varepsilon>0$, where $s_n^2 = \mathbb{E}[(\overline{T}_{\gamma,J})^2]$, and

$$s_n^{-2}V_n^2 \xrightarrow{P} 1 \tag{36}$$

as $n\to\infty$, where

$$V_n^2 = \sum_{i=2}^{m}\mathbb{E}[Y_{ni}^2\,|\,X_1,\cdots,X_{i-1},Y_1,\cdots,Y_{i-1}] + \sum_{i=m+1}^{n}\mathbb{E}[Y_{ni}^2\,|\,X_1,\cdots,X_{i-1},Y_1,\cdots,Y_m].$$

Since

$$\begin{aligned} s_n^2 &= \sum_{i=2}^{n}\mathbb{E}[Y_{ni}^2]\\ &= \sum_{i=2}^{m}\sum_{j=1}^{i-1}\big(\frac{4}{n^2(n-1)^2} + \frac{4}{m^2(m-1)^2} + \frac{8}{n^2m^2}\big)\cdot\mathbb{E}[H^2(X_1,X_2)]\\ &\quad + \sum_{i=m+1}^{n}\big(\sum_{j=1}^{i-1}\frac{4}{n^2(n-1)^2} + \sum_{j=1}^{m}\frac{4}{n^2m^2}\big)\cdot\mathbb{E}[H^2(X_1,X_2)]\\ &= \big(\frac{2}{n(n-1)} + \frac{2}{m(m-1)} + \frac{4(m-1)}{nm^2}\big)\cdot\mathbb{E}[H^2(X_1,X_2)].\end{aligned}$$

Furthermore, since

$$\mathbb{E}\big[H(X_1,X_2)H(X_1,X_3)H(X_1,X_4)H(X_1,X_5)\big] = \mathbb{E}\big[H(X_1,X_2)H^3(X_1,X_3)\big] = 0,$$

follow the proof of Theorem 1 of Hall (1984), we can obtain

$$\mathbb{E}[Y_{ni}^4] \lesssim n^{-8}\cdot i\cdot\mathbb{E}[H^4(X_1,X_2)] + n^{-8}\cdot i^2\cdot\mathbb{E}[H^2(X_1,X_2)H^2(X_1,X_3)],$$

whence

$$s_n^{-4}\sum_{i=2}^{n}\mathbb{E}[Y_{ni}^4] \lesssim \frac{\mathbb{E}[H^4(X_1,X_2)]}{n^2\cdot\big(\mathbb{E}[H^2(X_1,X_2)]\big)^2} + \frac{\mathbb{E}[H^2(X_1,X_2)H^2(X_1,X_3)]}{n\cdot\big(\mathbb{E}[H^2(X_1,X_2)]\big)^2} \to 0,$$

which implies condition (35). Write

$$v_{ni} = \begin{cases} \mathbb{E}\big[Y_{ni}^2 \,|\, X_1,\cdots,X_{i-1},Y_1,\cdots,Y_{i-1}\big], & 2\le i\le m \\ \mathbb{E}\big[Y_{ni}^2 \,|\, X_1,\cdots,X_{i-1},Y_1,\cdots,Y_m\big], & m+1\le i\le n. \end{cases}$$

Observe that when $i \le m$

$$\begin{aligned} v_{ni} = \sum_{j=1}^{i-1}\sum_{k=1}^{i-1}\Big[&\big(\frac{4}{n^2(n-1)^2}+\frac{4}{n^2m^2}\big)\cdot G(X_j,X_k) + \big(\frac{4}{m^2(m-1)^2}+\frac{4}{n^2m^2}\big)\cdot G(Y_j,Y_k) \\ &- \big(\frac{4}{n^2m(n-1)}+\frac{4}{m^2n(m-1)}\big)\cdot\big(G(Y_j,X_k)+G(Y_k,X_j)\big)\Big], \end{aligned}$$

and when $i > m$

$$\begin{aligned} v_{ni} = &\sum_{j=1}^{i-1}\sum_{k=1}^{i-1}\frac{4}{n^2(n-1)^2}\cdot G(X_j,X_k) + \sum_{j=1}^{m}\sum_{k=1}^{m}\frac{4}{m^2(m-1)^2}\cdot G(Y_j,Y_k) \\ &- \sum_{j=1}^{i-1}\sum_{k=1}^{m}\frac{8}{n^2m(n-1)}+\frac{8}{m^2n(m-1)}\cdot G(Y_k,X_j). \end{aligned}$$

Note that for $j_1\le k_1$ and $j_2\le k_2$,

$$\mathbb{E}[G(X_{j_1},X_{k_1})G(X_{j_2},X_{k_2})] = \begin{cases} \mathbb{E}[G^2(X_1,X_1)] & j_1=k_1=j_2=k_2 \\ (\mathbb{E}[G(X_1,X_1)])^2=(\mathbb{E}[H^2(X_1,X_2)])^2 & j_1=k_1\neq j_2=k_2 \\ \mathbb{E}[G^2(X_1,X_2)] & j_1=j_2,k_1=k_2,j_1<k_1 \\ 0 & \text{otherwise.} \end{cases}$$

We can write

$$\begin{aligned} \mathbb{E}[V_n^4] &= \mathbb{E}\big[\big(\sum_{i=1}^{n}v_{ni}\big)^2\big] \\ &= C_{nm1}\cdot\mathbb{E}[G^2(X_1,X_1)] + C_{nm2}\cdot(\mathbb{E}[H^2(X_1,X_2)])^2 + C_{nm3}\cdot\mathbb{E}[G^2(X_1,X_2)]. \end{aligned}$$

After some algebra, we can check that

$$\begin{aligned} C_{nm1} &\lesssim n^{-5} \\ C_{nm3} &\lesssim n^{-4} \\ \Big|\frac{C_{nm2}}{\big(\frac{2}{n(n-1)}+\frac{2}{m(m-1)}+\frac{4(m-1)}{nm^2}\big)^2}-1\Big| &= o(1) \end{aligned}$$

Thus we have

$$\begin{aligned} &\Big|s_n^{-4}\mathbb{E}(V_n^2-s_n^2)^2\Big| \\ &= \Big|\frac{\mathbb{E}[V_n^4]}{s_n^4}-1\Big| \\ &\lesssim \frac{\mathbb{E}[G^2(X_1,X_2)]}{(\mathbb{E}[H^2(X_1,X_2)])^2} + \frac{\mathbb{E}[G^2(X_1,X_1)]}{n(\mathbb{E}[H^2(X_1,X_2)])^2} + o(1) \\ &\lesssim \frac{\mathbb{E}[G^2(X_1,X_2)]}{(\mathbb{E}[H^2(X_1,X_2)])^2} + \frac{\mathbb{E}[H^2(X_1,X_2)H^2(X_1,X_3)]}{n(\mathbb{E}[H^2(X_1,X_2)])^2} + o(1) \\ &= o(1), \end{aligned}$$

which implies condition (36). Proof is completed.