A Martingale Kernel Two-Sample Test

Anirban Chatterjee   ANIRBANC@UCHICAGO.EDU
Department of Statistics, University of Chicago

Aaditya Ramdas   ARAMDAS@STAT.CMU.EDU
Department of Statistics and Data Science, Machine Learning Department, Carnegie Mellon University

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

The Maximum Mean Discrepancy (MMD) is a widely used multivariate distance metric for two-sample testing. The standard MMD test statistic has an intractable null distribution typically requiring costly resampling or permutation approaches for calibration. In this work we leverage a martingale interpretation of the estimated squared MMD to propose martingale MMD (mMMD), a quadratic-time statistic which has a limiting standard Gaussian distribution under the null. Moreover we show that the test is consistent against any fixed alternative and for large sample sizes, mMMD offers substantial computational savings over the standard MMD test, with only a minor loss in power.

1. Introduction

Given two distributions P and Q over a metric space $\mathcal{X}$, the two-sample problem is to test

$$\boldsymbol{H}_0 : P = Q \text{ versus } \boldsymbol{H}_1 : P \neq Q \tag{1.1}$$

based on independent samples $\mathscr{X}_n = \{X_1, \ldots, X_n\} \sim P$ and $\mathscr{Y}_n = \{Y_1, \ldots, Y_n\} \sim Q$. This classical problem has been widely studied in the parametric setting, assuming low-dimensional structure. However, parametric methods often break down under model misspecification or non-Euclidean data, motivating flexible non-parametric approaches with minimal assumptions. For univariate data, classical non-parametric tests include the Kolmogorov–Smirnov test (Smirnov, 1948), Wald–Wolfowitz runs test (Wald and Wolfowitz, 1940), rank-sum test (Mann and Whitney, 1947; Wilcoxon, 1945), and Cramér–von Mises test (Anderson, 1962). Early multivariate extensions by Weiss (1960) and Bickel (1969) inspired modern methods based on geometric graphs (Friedman and Rafsky, 1979; Rosenbaum, 2005; Bhattacharya, 2019), energy distance (Székely, 2003; Székely and Rizzo, 2013), kernel MMD (Gretton et al., 2009, 2012a; Ramdas et al., 2015b; Shekhar et al., 2022; Chatterjee and Bhattacharya, 2025), ball divergence (Pan et al., 2018; Banerjee and Ghosh, 2025), classifier-based tests (Lopez-Paz and Oquab, 2017; Kim et al., 2021), and others.

Among the tests mentioned, kernel-based methods have recently emerged as powerful tools for detecting distributional differences in general domains. These methods use the Maximum Mean Discrepancy (MMD) (see (2.1)) to compare distributions P and Q by embedding them into a reproducing kernel Hilbert space (RKHS) $\mathcal{K}$ associated with a positive definite kernel K. Gretton et al. (2012a) proposed a quadratic-time estimator of the squared MMD, forming the basis of the kernel two-sample test (referred to as the quadratic-time MMD test) which rejects the null $\boldsymbol{H}_0$ when the statistic exceeds a threshold τ_α that controls the Type I error at level α.

Despite its popularity, a key practical challenge in the MMD test is determining the threshold τ_α. Under $\boldsymbol{H}_0$, the test statistic weakly converges to an infinite weighted sum of centered χ^2 variables, with weights depending on the unknown distribution $P = Q$ (Gretton et al., 2009, 2012a). As a result, estimating τ_α typically requires either computationally intensive resampling (Gretton et al., 2009; Chatterjee and Bhattacharya, 2025) or moment-based parametric approximations (for example using Gamma, Pearson curves) (Gretton et al., 2012a), which often lack consistency or accuracy guarantees. Calibration via concentration bounds is another option but tends to be overly conservative. Recent work has proposed more efficient alternatives (Gretton et al., 2012a, Section 6); (Zaremba et al., 2013; Jitkrittum et al., 2017), but these often overlook the pairwise kernel structure and can suffer from low power. We review these methods in detail in Section 1.2.

In this work, we propose a simple and novel variant of the kernel-based two-sample test, where the test statistic follows an asymptotic standard normal distribution under $\boldsymbol{H}_0$, making it easy to implement in practice. It remains valid under both fixed kernels (independent of sample size) and varying kernels (dependent on sample size) settings. We show that the test is consistent and establish asymptotic normality under fixed alternatives as well. Below, we provide a brief overview of our result, with full details deferred to Sections 2.2 and 3.

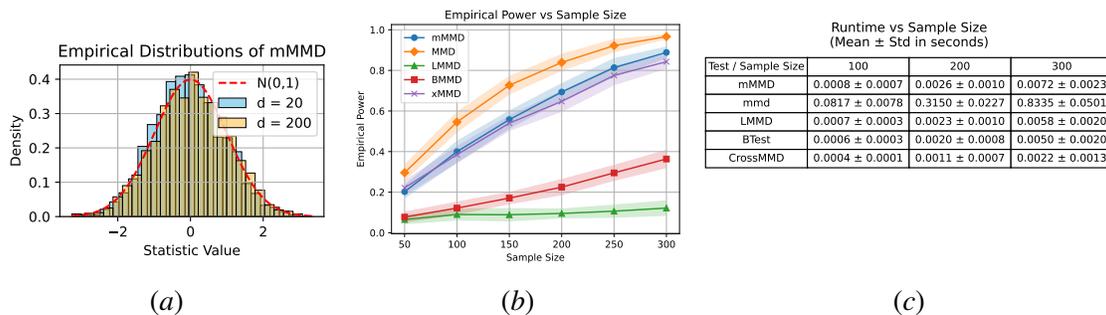## 1.1. Overview of the Proposed Method

Figure 1: The mMMD-test: (a) Empirical distribution of the mMMD-test statistic (see (2.4) for a formal definition) under $\boldsymbol{H}_0$ with $d = 20, 200$ and $P = Q$ are d-dimensional standard Gaussian distributions. (b) The second figure compares power of the mMMD test with the quadratic time MMD test (implemented with 200 permutations) from Gretton et al. (2012a) and computationally efficient variants from Gretton et al. (2012a), Zaremba et al. (2013) and Shekhar et al. (2022). The main takeaway is that the MMD is the most powerful but most computationally inefficient. The xMMD and mMMD are more powerful than the others, but the latter avoids sample splitting. (c) The third figure shows the computational efficiency of the proposed test against the permutation-based quadratic time MMD test. All tests use the Gaussian kernel with median bandwidth.

We propose a variant of the quadratic-time MMD-test (see Gretton et al. (2012a)) motivated by the key observation that a modified form of the estimator of the squared MMD admits a martingale structure. Specifically, we introduce the *martingale*-MMD (mMMD) statistic, defined as

$$T_n = \frac{1}{n} \sum_{i=2}^{n} \left\langle \hat{f}_i, \mathsf{K}(X_i, \cdot) - \mathsf{K}(Y_i, \cdot) \right\rangle_{\mathcal{K}}, \quad \text{where} \quad \hat{f}_i = \frac{1}{i} \sum_{j=1}^{i-1} \mathsf{K}(X_j, \cdot) - \mathsf{K}(Y_j, \cdot).$$

2

In other words, for each $2 \leq i \leq n$, our method estimates the *witness function* (see Section 2.2 and Kübler et al. (2022)) using the empirical mean embeddings computed from past data, and then averages the resulting estimates of the squared MMD. The name arises because nT_n is a (zero-mean) martingale under the null.

Our final test statistic (for (1.1)) is then defined as $\eta_n = T_n/\sigma_n$ where σ_n (see (2.3)) is an estimate of the asymptotic variance under $\boldsymbol{H}_0$. In Section 3, we establish general results showing that η_n converges in distribution to a standard Gaussian under the null. Theorem 2 shows this under mild moment conditions, when K, P and the data distribution remains fixed with respect to n. Theorem 4 then established conditions under which a Berry–Esseen-type convergence to $\mathrm{N}(0, 1)$ holds when the kernel and both data distribution and data dimension are allowed to change with n.

Leveraging this asymptotic Gaussian behavior, our main methodological contribution is the *mMMD-test*, which rejects the null hypothesis if $\eta_n > z_{1-\alpha}$, where $z_{1-\alpha}$ is the $(1 - \alpha)$ quantile of the standard normal distribution. Formally, we define the test as

$$\Psi_n := \mathbf{1}\left\{\eta_n > z_{1-\alpha}\right\}. \tag{1.2}$$

Turning to performance under alternatives, in Theorem 8 we identify general conditions under which the test remains uniformly consistent over a class of alternatives, even in the challenging setting where K, P, Q and the data dimension may vary with the sample size n. Furthermore, Theorem 9 shows that under the alternative hypothesis $P_n \neq Q_n$, the test statistic (properly centered and scaled) converges in distribution to a standard Gaussian. This result directly implies (see Theorem 7) that the test defined in (1.2) is consistent in the simpler case when the parameters do not depend on n.

In Section 6, we empirically validate our method and compare it against the quadratic-time MMD test and other computationally efficient variants, demonstrating strong performance on both simulated and real data. Section 7 explores extensions to multiple-kernel settings and introduces a broader class of MMD-based statistics built on the same principles as the $m\mathrm{MMD}$. Theorem 13 shows that, under suitable conditions, this class also admits an asymptotic standard Gaussian distribution. Furthermore, in Section A.3, we show that certain statistics in this class remain consistent even when the L_2 distance between the densities decays at the minimax rate.

1.2. Related Works

The literature on kernel-based two-sample tests is quite extensive. In this section, we succinctly review the most relevant works. The first approach to obtaining a distribution-free kernel two-sample test relied on concentration-inequality-based large deviation bounds for the quadratic-time MMD statistic (Gretton et al., 2006; Kim, 2021). However, these bounds tend to be overly conservative and result in tests with low (Wolfer and Alquier, 2025). Balsubramani and Ramdas (2016) proposed a sequential two-sample test based on the Maximum Mean Discrepancy (MMD) with a linear kernel, using martingale concentration inequalities. More recently, Shekhar and Ramdas (2023) extended this approach by introducing a more general sequential MMD-based test that leverages Ville's inequality, resulting in a less conservative procedure. Gretton et al. (2012a) proposed moment-based parametric approximations such as Pearson curves and Gamma distributions, but these remain heuristic in nature and lack rigorous guarantees. Gretton et al. (2009) introduced a spectral method for approximating the asymptotic distribution, which required strong assumptions and needed additional resampling (with external randomness) for implementation.

To address computational inefficiency, later works proposed modifying the test statistic itself to yield a tractable null distribution. One such method partitions the data into disjoint blocks, computes

the kernel MMD statistic within each block, and averages the results. When block sizes are fixed, this leads to the linear-time MMD test (Gretton et al., 2012a). Zaremba et al. (2013); Ramdas et al. (2015a); Reddi et al. (2015) extended this idea to block sizes that grow with the sample size. The computational complexity of the block-MMD statistic then ranges from linear (for constant block size) to quadratic (when block size is $\Omega(n)$). If the block size $B = o(n)$, the test statistic admits a Gaussian limit under the null. These approaches fall under the umbrella of incomplete U-statistcs based approaches. We refer to Schrab et al. (2022) for a comprehensive overview of incomplete U-statistic-based kernel tests. We empirically compare our method with both linear-time and block-MMD tests in Section 6.2.

Recent works have also explored sample-splitting strategies. Notably, Kübler et al. (2022) used one half of the data to estimate the witness function and the other half to compute the test statistic. However, their analysis assumes the witness function is fixed and relies on permutation testing for calibration, which limits practical efficiency. In contrast, our test leverages asymptotic Gaussianity to avoid resampling, making it more straightforward to implement.

Shekhar et al. (2022) proposed a sample-splitting approach, where in each split they compute the difference between the mean embeddings and apply RKHS inner product between the two estimated embedding differences. Their dimension-agnostic Gaussian null distribution is something that we also achieve, but we also prove that our statistic has an alternative distribution that is Gaussian. We show empirically that our approach performs comparable to theirs, but ours avoids sample splitting and is based on different principles.

Beyond methods aimed at efficient threshold selection, several works have addressed the computational cost of the test statistic itself. The standard quadratic-time MMD statistic scales as $O(n^2)$. To reduce this, methods based on Random Fourier Features (Zhao and Meng, 2015; Zhao et al., 2022; Choi and Kim, 2024; Mukherjee and Sriperumbudur, 2025), coreset-based approximations (Domingo-Enrich et al., 2023), and Nyström approximations (Chatalic et al., 2025) have been proposed. However, most of these still rely on permutation or resampling techniques to calibrate the test statistic, which remains the main focus of this work.

2. The martingale Maximum Mean Discrepancy (mMMD) Test

2.1. Preliminaries

In this section, we review the fundamentals of kernel two-sample testing as introduced by Gretton et al. (2012a), and present the key concepts required for our proposed test. We begin by establishing the necessary formalism.

Let $\mathcal{X}$ be a Polish space, that is a complete seperable metric space, equipped with the corresponding Borel sigma algebra $\mathcal{B}_{\mathcal{X}}$. Let $\mathcal{P}(\mathcal{X})$ be the collection of all probability measures defined on $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$. Let K be a positive definite kernel with features maps $\phi_x, x \in \mathcal{X}$ such that $\mathsf{K}(x, \cdot) = \phi_x$ and $\mathsf{K}(x_1, x_2) = \langle \phi_{x_1}, \phi_{x_2} \rangle_{\mathcal{K}}$ where $\mathcal{K}$ is the associated RKHS. The notion of feature maps can be extended to embed any distribution $P \in \mathcal{P}(\mathcal{X})$ into $\mathcal{K}$. In particular, for any $P \in \mathcal{P}(\mathcal{X})$ define the kernel mean embedding ν_P as $\langle f, \nu_P \rangle = \mathbb{E}_{X \sim P}[f(X)]$ for all $f \in \mathcal{K}$. Moreover, by the cannonical form of the feature maps, for all $t \in \mathcal{X}, \nu_P(t) = \mathbb{E}_{X \sim P}[\mathsf{K}(X, t)]$. Throughout the rest of the article we make the following assumption on the kernel K.

Assumption 2.1 *Consider the positive definite kernel* $\mathsf{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *and let* $\mathcal{M}_{\mathsf{K}}^{\theta} = \{P \in \mathcal{P}(\mathcal{X}) : \mathbb{E}_{X \sim P}[\mathsf{K}(X, X)^{\theta}] < \infty\}$. *Then* K *is characteristic with respect to* $\mathcal{M}_{\mathsf{K}}^{1/2}$, *that is* $\nu : \mathcal{M}_{\mathsf{K}}^{1/2} \to \mathcal{K}$ *is a one-to-one (injective) mapping.*

Assumption 2.1 ensures that $\mu_P \in \mathcal{K}$ for all $P \in \mathcal{M}_{\mathsf{K}}^{1/2}$ (see Gretton et al. (2012a, Lemma 3) and Park and Muandet (2020, Lemma 2.1)). With the above definitions we can now define the Maximum Mean Discrepancy (MMD) between two distribution $P, Q \in \mathcal{M}_{\mathsf{K}}^{1/2}$ as,

$$\mathrm{MMD}[P, Q, \mathsf{K}] = \sup_{f \in \mathcal{F}_{\mathcal{K}}} \mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)], \tag{2.1}$$

where $\mathcal{F}_{\mathcal{K}}$ is the unit-ball in the RKHS $\mathcal{K}$. Assumption 2.1 also ensures that MMD forms a metric on $\mathcal{M}_{\mathsf{K}}^{1/2}$. Moreover, using the notion of mean embeddings the MMD can be equivalently expressed as $\mathrm{MMD}[P, Q, \mathsf{K}] = \|\nu_P - \nu_Q\|_{\mathcal{K}}$ where $\| \cdot \|_{\mathcal{K}}$ is the norm induced by the inner product $\langle \cdot, \cdot \rangle_{\mathcal{K}}$.

2.2. Deriving the $m\mathrm{MMD}$ test

The standard quadratic-time MMD test statistic is obtained by substituting the empirical mean embeddings into the alternative expression for squared $\mathrm{MMD}[P, Q]$ derived earlier. In contrast, our approach returns to the original definition of $\mathrm{MMD}[P, Q]$ given in (2.1).
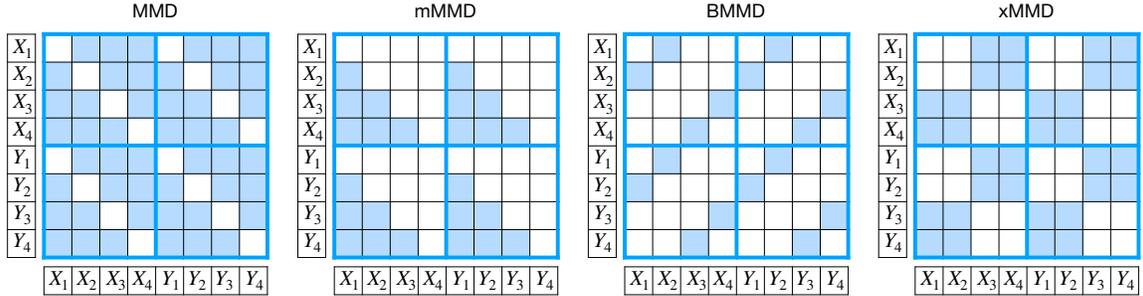

Figure 2: A visual illustration of the main difference in computing the quadratic-time MMD (MMD), our proposed mMMD, the block-MMD (BMMD) from Zaremba et al. (2013), and the cross-MMD (xMMD) from Shekhar et al. (2022). The highlighted regions indicate the sample pairs used in the computation. The figure shows that the quadratic-time MMD considers all off-diagonal pairwise kernel evaluations between the combined samples. BMMD partitions the data into blocks and averages the pairwise kernel evaluations within each block. xMMD splits the samples into two halves and evaluates kernels only across the splits. In contrast, our proposed mMMD computes pairwise kernel values by taking the lower triangular part of within-sample and between-sample blocks. However, we emphasize that, due to the symmetry of the kernel, the terms used by MMD and mMMD are equivalent. The principal distinction between the two lies in the normalization procedure: mMMD computes a row-wise mean followed by an average across rows (see (2.5)), whereas MMD applies a global mean over all elements.

Note that if the witness function f_0, that is, $f_0 = \arg\sup_{f \in \mathcal{F}} \{\mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)]\}$, were known, then one could estimate $\mathrm{MMD}[P, Q]$ using the statistic $\frac{1}{n} \sum_{i=1}^{n} \langle f_0, \mathsf{K}(X_i, \cdot) - \mathsf{K}(Y_i, \cdot) \rangle_{\mathcal{K}}$, based on samples $\mathscr{X}_n = \{X_1, \ldots, X_n\} \sim P$ and $\mathscr{Y}_n = \{Y_1, \ldots, Y_n\} \sim Q$. However, the witness

function f_0 is unknown. Nevertheless, as shown by Gretton et al. (2012a, Section 2.3) it can be expressed as $f_0 = (\nu_P - \nu_Q)/\|\nu_P - \nu_Q\|_{\mathcal{K}}$, where $\|\cdot\|_{\mathcal{K}}$ denotes the RKHS norm. Then for each $2 \leq i \leq n$ we can plug in the empirical mean embeddings to approximate the unnormalized version of f_0 by $\hat{f}_i = \frac{1}{i} \sum_{j=1}^{i-1} [\mathsf{K}(X_j, \cdot) - \mathsf{K}(Y_j, \cdot)]$, which leads to our proposed statistic:

$$T_n := \frac{1}{n} \sum_{i=2}^{n} \left\langle \hat{f}_i, \mathsf{K}(X_i, \cdot) - \mathsf{K}(Y_i, \cdot) \right\rangle_{\mathcal{K}}.$$

To further motivate this statistic, observe that by using only past data (notice that we only use indices less than i for the estimate $\hat{f}_i$) to estimate the unnormalized witness function f_0, we induce an underlying martingale structure. In particular, under $\boldsymbol{H}_0$, the collection

$$\left\{ \left\langle \hat{f}_i, \mathsf{K}(X_i, \cdot) - \mathsf{K}(Y_i, \cdot) \right\rangle_{\mathcal{K}} : i \geq 2 \right\} \tag{2.2}$$

forms a martingale difference sequence with respect to the natural filtration $\mathcal{F}_i = \sigma(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_i)$, where $\boldsymbol{Z}_i = (X_i, Y_i), 2 \leq i \leq n$. With this martingale formulation one can expect to apply the martingale CLT (Hall and Heyde, 2014) on T_n (with proper normalisation) to have asymptotic Gaussianity. However, directly applying a martingale central limit theorem (CLT) to T_n may yield an asymptotic Gaussian distribution with unknown variance. Therefore, to implement a test for (1.1) based on T_n, we must appropriately normalize the statistic. To that end, define

$$\sigma_n^2 := \frac{1}{n^2} \sum_{i=2}^{n} \left(\frac{1}{i} \sum_{j=1}^{i-1} \mathsf{K}(X_i, X_j) - \mathsf{K}(X_i, Y_j) - \mathsf{K}(X_j, Y_i) + \mathsf{K}(Y_i, Y_j) \right)^2. \tag{2.3}$$

With this notation, we define the normalized test statistic as

$$\eta_n = T_n/\sigma_n, \tag{2.4}$$

and use it to construct the $m\text{MMD}$ test as described in (1.2).

Remark 1 *Note that the computational complexity of computing η_n is $O(n^2)$. To see this, observe that using the kernel trick, the statistic T_n can be simplified as*

$$T_n = \frac{1}{n} \sum_{i=2}^{n} \frac{1}{i} \sum_{j=1}^{i-1} \mathsf{K}(X_i, X_j) - \mathsf{K}(X_i, Y_j) - \mathsf{K}(X_j, Y_i) + \mathsf{K}(Y_i, Y_j). \tag{2.5}$$

It is evident from this expression that computing T_n requires $O(n^2)$ operations. Moreover, from the definition of σ_n in (2.3), it follows that computing σ_n also requires $O(n^2)$ time. Hence, the total cost of computing the test statistic η_n is $O(n^2)$. This matches the computational complexity of the original quadratic-time MMD statistic proposed by Gretton et al. (2012a), as well as the more recent cross-MMD statistic introduced by Shekhar et al. (2022). In contrast, the alternative approaches proposed by Zaremba et al. (2013) and Jitkrittum et al. (2017) can achieve lower computational complexity, at the expense of highly suboptimal power.

3. Asymptotic Null Distribution of mMMD Test Statistic

In this section, we derive the limiting distribution of the statistic η_n (as defined in (2.4)) under the null hypothesis $\boldsymbol{H}_0$. Specifically, we establish the asymptotic normality of η_n, a phenomenon that is empirically suggested by Figure 1 (a).

For a kernel K, and define its centered version as $\bar{\mathsf{K}}(x, y) = \langle \mathsf{K}(x, \cdot) - \nu_P, \ \mathsf{K}(y, \cdot) - \nu_P \rangle_{\mathcal{K}}$, where ν_P denote the kernel mean embedding of P.

Theorem 2 *Take $\mathcal{X} = \mathbb{R}^d$ for some $d \geq 1$. Suppose the kernel K satisfies Assumption 2.1, and let $P \in \mathcal{M}_{\mathsf{K}}^{1/2}$. Moreover assume $\mathbb{E}\left[\bar{\mathsf{K}}\left(X_1, X_2\right)^4\right] < \infty$ for X_1, X_2 generated independently from P. Then under $\boldsymbol{H}_0$,*

$$\sup_{x \in \mathbb{R}} \left|\mathbb{P}\left(\eta_n \leq x\right) - \Phi(x)\right| \to 0.$$

The above theorem confirms that, under $\boldsymbol{H}_0$, the asymptotic distribution of the mMMD test statistic η_n is standard Gaussian. Consequently, the test Ψ_n defined in (1.2) asymptotically controls the type-I error at level α.

Remark 3 (Martingale Energy Two-Sample Test) *Take $\mathcal{X} = \mathbb{R}^d$ and consider the distance induced kernel $\mathsf{K}(x, y) = \frac{1}{2}\left(\|x\|_2 + \|y\|_2 - \|x - y\|_2\right)$, so that the squared maximum mean discrepancy satisfies $\mathrm{MMD}^2[P, Q; \mathsf{K}] = \frac{1}{2}\rho(P, Q)$, where $\rho(P, Q)$ denotes the classical energy distance between P and Q (Székely, 2003; Székely and Rizzo, 2013; Sejdinovic et al., 2013). This kernel is characteristic with respect to $\mathcal{M}_{\mathsf{K}}^1$, providing an RKHS representation of the energy distance. With this choice of kernel, we can define a* martingale energy two-sample test *using the normalised martingale energy statistic as,*

$$\eta_{n,\rho} = \frac{\frac{1}{n}\sum_{i=2}^{n}\frac{1}{i}\sum_{j=1}^{i-1}\|X_i - Y_j\|_2 + \|X_j - Y_i\|_2 - \|X_i - X_j\|_2 - \|Y_i - Y_j\|_2}{\sqrt{\frac{1}{n^2}\sum_{i=2}^{n}\left(\frac{1}{i}\sum_{j=1}^{i-1}\|X_i - Y_j\|_2 + \|X_j - Y_i\|_2 - \|X_i - X_j\|_2 - \|Y_i - Y_j\|_2\right)^2}}.$$

Note that assumptions of Theorem 2 are satisfied whenever $\mathbb{E}\left[\|X\|_2^4\right] < \infty$, and under the null hypothesis it is asymptotically standard normal. This allows us to define the martingale energy test $\Psi_{n,\rho} = \mathbf{1}\{\eta_{n,\rho} > z_{1-\alpha}\}$, where (as before) $z_{1-\alpha}$ denotes the $(1 - \alpha)$-quantile of the standard normal distribution.

The above result can be significantly extended to accommodate more general dependencies of the kernel, the distribution, and even the sample space on the sample size n. Specifically, we now allow the kernel K_n, the null distribution $P_n(= Q_n)$, and the dimension d_n to vary with n[1]. In this more general setting, the next result shows that, under suitable conditions, the test statistic η_n continues to converge in distribution to a standard Gaussian under the null.

Theorem 4 *Take $n \geq 1$ and $\mathcal{X}_n = \mathbb{R}^{d_n}$. Suppose the kernel K_n satisfies Assumption 2.1 on $\mathcal{X}_n$ and $P_n \in \mathcal{M}_{\mathsf{K}_n}^{1/2}$. With X_1, X_2, X_3 generated independently from P_n assume that $\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^4\right] <$*

---

1. To denote dependence of a parameter θ on n we use the notation θ_n.

∞. *Furthermore assume the following holds,*

$$\lim_{n \to \infty} \frac{\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^4\right]}{(\log n)^2 \mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2\right]^2} = 0 \text{ and } \lim_{n \to \infty} \frac{\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2 \bar{\mathsf{K}}_n(X_1, X_3)^2\right]}{(\log n)^2 \mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2\right]^2} = 0. \quad (3.1)$$

Then under $\boldsymbol{H}_0$,

$$\sup_{x \in \mathbb{R}} |\mathbb{P}\left(\eta_n \leq x\right) - \Phi(x)| \to 0.$$

Note that the condition in (3.1) is well-defined by Lemma 21. Moreover, Theorem 2 follows as an immediate corollary of Theorem 4, which in turn can be viewed as a special case of Corollary 15 by setting $\gamma = 1$. The proof of Theorem 4 (and more generally Theorem 13) proceeds by recalling the martingale difference sequence in the construction of η_n from (2.2), and applying Berry–Esseen bounds for self-normalized martingales from Fan and Shao (2018).

Remark 5 *In the case where $d_n = d$ is fixed, the condition (3.1) is satisfied for uniformly bounded kernels, including commonly used examples such as Gaussian and Laplace kernels, when appropriate n-dependent bandwidths are used (see Section C.3 and Li and Yuan (2024)). Similar conditions have been considered in the context of kernel two-sample tests by Shekhar et al. (2022) and Li and Yuan (2024). Furthermore, a condition akin to (3.1) has appeared in the broader literature on studentized central limit theorems, notably in Bentkus and Götze (1996), in the setting of triangular arrays of i.i.d. random variables. While their assumption is phrased in terms of third moments, our condition is slightly stronger, due to the fact that the summands in our case form a martingale difference sequence and thus exhibit dependencies not present in the i.i.d. setting.*

Remark 6 *Note that Theorem 4 permits the kernel to depend on the sample size n, but not on the observed samples $\mathscr{X}_n$ and $\mathscr{Y}_n$. This is particularly relevant as it excludes the common practice of selecting kernel bandwidths, such as for Gaussian or Laplace kernels, based on the median of pairwise distances from the pooled samples (the median heuristic). Nevertheless, we find in our experiments that using the median heuristic retains the same null distribution, so it is possible that the above theoretical guarantees can be extended to certain data-dependent kernels as well.*

4. Consistency of the mMMD Test

In this section, we show that the mMMD test defined in (1.2) is consistent, that is, it achieves asymptotic power one, against fixed alternatives. Furthermore, we establish conditions under which the test achieves *uniform consistency* against a class of alternatives in the more challenging setting where the kernel, distributions, and sample space may all vary with the sample size n.

We first show that the mMMD test is consistent against fixed alternatives.

Theorem 7 *Take $\mathcal{X} = \mathbb{R}^d$ for some $d \geq 1$ and for a kernel K adopt Assumption 2.1. Let $P, Q \in \mathcal{M}_{\mathsf{K}}^1$ such that $P \neq Q$. Then $\mathbb{E}[1 - \Psi_n] \to 1$, that is the test Ψ_n from (1.2) is consistent.*

The above theorem requires only mild assumptions, which are trivially satisfied by commonly used kernels such as the Gaussian and Laplace kernels. Analogous to the setting discussed in Theorem 13, the following result establishes conditions under which consistency continues to hold even when the kernel, distributions, and sample space are allowed to vary with the sample size n.

Theorem 8 *Take $n \geq 1$ and $\mathcal{X}_n = \mathbb{R}^{d_n}$. Consider a positive definite kernel K_n satisfying Assumption 2.1 on $\mathcal{X}_n$ and take $\mathcal{P}_n \subseteq \mathcal{M}_{\mathsf{K}_n}^{1/2} \times \mathcal{M}_{\mathsf{K}_n}^{1/2}$. For $(P_n, Q_n) \in \mathcal{P}_n$ assume that $\delta_n := \mathrm{MMD}[P_n, Q_n, \mathsf{K}_n] > 0$. If,*

$$\lim_{n \to \infty} \sup_{(P_n, Q_n) \in \mathcal{P}_n} \frac{\mathbb{E}[\sigma_n^2]}{\delta_n^4} + \frac{\mathrm{Var}\,(T_n)}{\delta_n^4} = 0,$$

then $\lim_{n \to \infty} \sup_{(P_n, Q_n) \in \mathcal{P}_n} \mathbb{E}\left[1 - \Psi_n\right] = 0$, that is Ψ_n is consistent.

Theorem 8 is an immediate consequence of Theorem 16. Theorem 7 indeed follows by checking the conditions of Theorem 8, which we detail in Section D.1. We note that the conditions required for the consistency of the test Ψ_n are identical to those needed for the consistency of the cross-MMD test proposed by Shekhar et al. (2022).

5. Distribution under the alternative

In this section, we show that the m-MMD statistic T_n converges to a Gaussian distribution even under the alternative hypothesis, that is, when $P \neq Q$. To state the result, we consider the setting, allowing the distributions P_n, Q_n, and the kernel K_n to vary with the sample size n. However note that in contrast to the previous general settings here $\mathcal{X} = \mathbb{R}^d$ remains fixed with respect to n.

We begin by introducing the following notation:

$$\mathsf{H}_n(\boldsymbol{z}_1, \boldsymbol{z}_2) := \mathsf{K}_n(x_1, x_2) - \mathsf{K}_n(x_1, y_2) - \mathsf{K}_n(x_2, y_1) + \mathsf{K}_n(y_1, y_2), \tag{5.1}$$

where $\boldsymbol{z}_i = (x_i, y_i)$ for $i = 1, 2$. Furthermore, for $\boldsymbol{Z} = (X, Y) \sim P_n \times Q_n$, define

$$h_n(\boldsymbol{z}) := \mathbb{E}\left[\mathsf{H}_n(\boldsymbol{z}, \boldsymbol{Z})\right] - \mathrm{MMD}[P_n, Q_n, \mathsf{K}_n]^2. \tag{5.2}$$

With this setup in place, we now state the result characterizing the asymptotic distribution of T_n under the alternative $P_n \neq Q_n$.

Theorem 9 *Adopt Assumption 2.1 with K_n. Let $P_n, Q_n \in \mathcal{M}_{\mathsf{K}_n}^{3/2}$ such that $P_n \neq Q_n$ and assume that for $\boldsymbol{Z}, \boldsymbol{Z}_1, \boldsymbol{Z}_2 \sim P_n \times Q_n$,*

$$\frac{\mathbb{E}\left[|h_n(\boldsymbol{Z})|^3\right]}{\mathrm{Var}\,(h_n(\boldsymbol{Z}))^{3/2}} \ll \frac{n^{3/2}}{(\log n)^3}, \quad \frac{\mathbb{E}\left[g_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2\right]}{\mathrm{Var}\,(h_n(\boldsymbol{Z}))} \ll \frac{n}{\log n}, \quad \frac{\mathrm{MMD}[P_n, Q_n, \mathsf{K}_n]^2}{\sqrt{\mathrm{Var}\,(h_n(\boldsymbol{Z})^2)}} \ll \frac{\sqrt{n}}{\log n},$$

where $g_n(\boldsymbol{z}_1, \boldsymbol{z}_2) = \mathsf{H}_n(\boldsymbol{z}_1, \boldsymbol{z}_2) - h_n(\boldsymbol{z}_1) - h_n(\boldsymbol{z}_2) + \mathrm{MMD}[P_n, Q_n, \mathsf{K}_n]^2$. Then,

$$\sqrt{\frac{n}{\mathrm{Var}\,(h_n(\boldsymbol{Z}))}} \left(T_n - \mathrm{MMD}[P_n, Q_n, \mathsf{K}_n]^2\right) \xrightarrow{D} \mathrm{N}(0, 5).$$

Using the above theorem, when P, Q and K does not depend on n we can easily conclude that,

$$\sqrt{n}\left(T_n - \mathrm{MMD}[P, Q, \mathsf{K}]^2\right) \xrightarrow{D} \mathrm{N}\left(0, 5\mathrm{Var}\,(h(\boldsymbol{Z}))\right) \text{ under } \boldsymbol{H}_1, \tag{5.3}$$

where $h(\boldsymbol{Z})$ is defined in (5.2) (without the dependence on n). We note that the asymptotic distribution in (5.3) closely resembles that of the original quadratic-time MMD estimator from Gretton et al. (2012a), having a constant factor 5 in the asymptotic variance in place of 4. While the results in Section 3 are similar to those for the cross-MMD test by Shekhar et al. (2022), the latter do not establish distributional convergence under the alternative.

9

6. Experiments

We empirically evaluate the proposed mMMD-test against the linear-time MMD test (Gretton et al., 2012a), Block MMD with $B = \lfloor\sqrt{n}\rfloor$ (Zaremba et al., 2013), and cross-MMD (Shekhar et al., 2022), collectively referred to as computationally efficient MMD tests. Section 6.1 validates the null distribution from Section 3, while Sections 6.2 and 6.3 compare performance on simulated and real data. In all experiments, we use the median heuristic for bandwidth selection. Although this choice is not supported by our theory (see Remark 6), nor by the theoretical guarantees in most prior works, it remains the standard in empirical studies due to its strong practical performance. Code for all experiments in this work is available at https://github.com/anirbanc96/mMMD.

6.1. Limiting Null Distribution of η_n

In Theorem 2, we established that the asymptotic null distribution of the m-MMD test statistic η_n is standard Gaussian under mild conditions. In this section, we empirically validate this result across a range of settings, examining the impact of dimensionality, data distribution, and kernel choice. Specifically, we consider dimensions $d \in \{10, 100, 250, 500\}$; data distributions $\mathrm{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ and $t_d(10)$; and both Gaussian and Laplace kernels. All experiments in this section use a sample size of $n = 200$ and 2000 resamples for the empirical distributions.
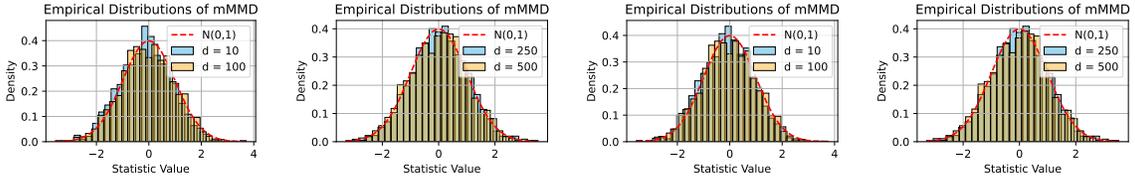
Figure 3: Empirical null distribution of the m-MMD test statistic η_n from equation (2.4). The left two figures use a Gaussian kernel with the median heuristic, while the right two figures use a Laplace kernel, also with the median heuristic. The underlying data distribution is $\mathrm{N}(\mathbf{0}_d, \mathbf{I}_d)$.
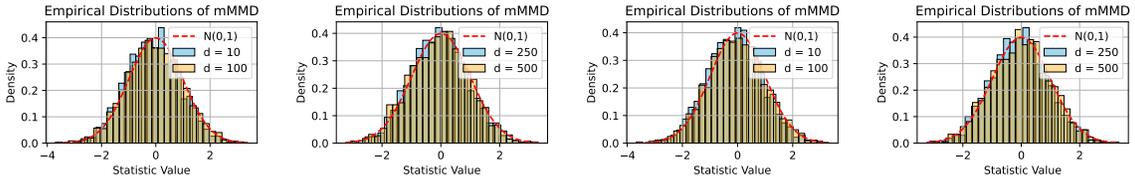
Figure 4: Empirical null distribution of the m-MMD test statistic η_n from equation (2.4). The left two figures use a Gaussian kernel with the median heuristic, while the right two use a Laplace kernel, also with the median heuristic. The underlying data distribution is $t_d(10)$.

In Figure 3, we plot the empirical null distribution of η_n from equation (2.4), where the underlying data distribution is $\mathrm{N}_d(\mathbf{0}_d, \mathbf{I}_d)$. In Figure 4, we consider the same setup but with multivariate t-distribution having degrees of freedom 10 as the underlying data.

Across all these settings, the empirical distribution of η_n closely matches the standard Gaussian, demonstrating robustness to the data distribution, kernel choice, and dimensionality. This stands

in sharp contrast to the null distribution of the quadratic-time MMD statistic from Gretton et al. (2012a), which varies significantly, particularly with increasing dimension (Shekhar et al., 2022).

6.2. Power Experiments

In this section, we compare the performance of our proposed test with the quadratic-time MMD test with 200 permutations, the sequential MMD test from Shekhar and Ramdas (2023) (referred as BetMMD) and the computationally efficient MMD tests.

For the experimental setup, let $d \geq 1$, and define $\boldsymbol{\mu}_{d,j,\varepsilon} \in \mathbb{R}^d$ such that the first j coordinates are equal to ε and the remaining coordinates are zero. We consider a two-sample testing problem with distributions $P = \mathrm{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ and $Q = \mathrm{N}_d(\boldsymbol{\mu}_{d,j,\varepsilon}, \mathbf{I}_d)$. The specific configurations we evaluate are $(d, j, \varepsilon) = (10, 5, 0.3)$, $(50, 5, 0.3)$, and $(100, 5, 0.5)$. This experimental setting has also been considered in Shekhar et al. (2022). All experiments in this section use the Gaussian kernel.
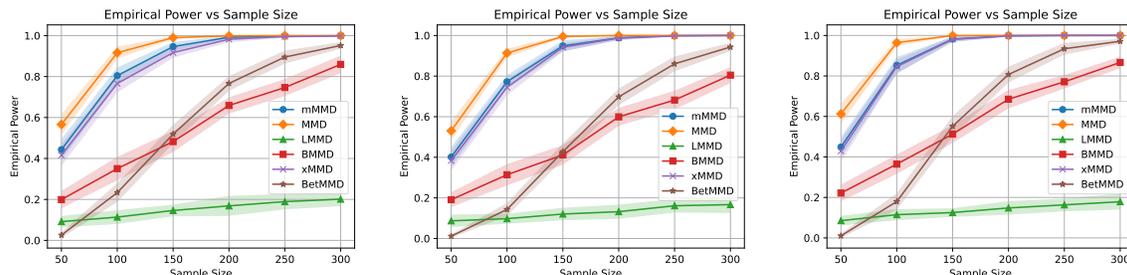

Figure 5: Empirical power of the m-MMD test compared to the original quadratic-time MMD, the linear-time LMMD, the block-based BMMD, and the cross-MMD (xMMD) tests. From left to right, the figures correspond to $(d, j, \varepsilon) = (10, 5, 0.3)$, $(50, 5, 0.4)$, and $(100, 5, 0.5)$.

We present the performance comparison in Figure 5. Across all settings, the m-MMD test consistently outperforms both the linear-time MMD and Block MMD tests, and achieves performance comparable to the cross-MMD test. While its performance is slightly below that of the quadratic-time MMD test, this is expected due to the trade-off between statistical and computational efficiency.

6.3. Real Data Experiment

In this section we compare the performance of our proposed test with computationally efficient MMD tests on the MNIST (LeCun, 1998) data set. We evaluate the performance of tests for the test from (1.1) with the following 5 pairs of sets of digits: Group 1: $P = \{2, 4, 8, 9\}$ and $Q = \{3, 4, 7, 9\}$, Group 2: $P = \{1, 2, 4, 8, 9\}$ and $Q = \{1, 3, 4, 7, 9\}$, Group 3: $P = \{0, 1, 2, 4, 8, 9\}$ and $Q = \{0, 1, 3, 4, 7, 9\}$, Group 4: $P = \{0, 1, 2, 4, 5, 8, 9\}$ and $Q = \{0, 1, 3, 4, 5, 7, 9\}$, and Group 5: $P = \{0, 1, 2, 4, 5, 6, 8, 9\}$ and $Q = \{0, 1, 3, 4, 5, 6, 7, 9\}$. Similar experimental settings have been considered in Schrab et al. (2023) and Chatterjee and Bhattacharya (2025).

The entire procedure is repeated 100 times, and the resulting empirical power is reported in Figure 6.

For each of the five methods, we draw 100 samples with replacement from the sets P and Q and use the asymptotic Gaussian distribution to calibrate the test under $\boldsymbol{H}_0$ at a significance level of 0.05. Notice that by construction, with increasing group number it becomes harder to detect difference between the distributions. Indeed in Figure 6 we note that the power of all methods decreases with increasing group index. However for the proposed mMMD-test and the cross-MMD tests the power decrement is much slower and overall they perform significantly better than other computationally efficient variants.
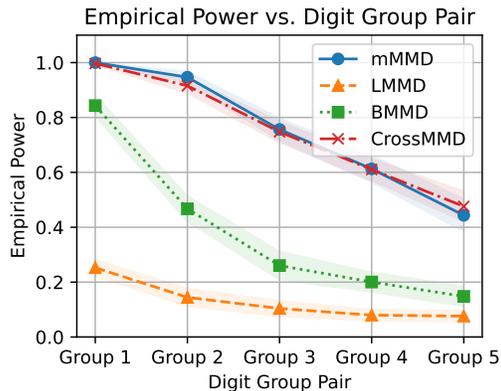

Figure 6: Empirical power comparison on MNIST.

7. Extensions

In this section, we explore two broad applications of the *mMMD*-test. In Section 7.1, we propose combining multiple kernels with the mMMD-test to potentially enhance power, and establish the null distribution of a corresponding Mahalanobis-distance-type test statistic. In Section 7.2, we introduce a general family of MMD-based test statistics that interpolate between the quadratic-time MMD and the mMMD statistic. We study this family in more detail in Section A.

7.1. mmMMD : Multi-kernel test using mMMD

We explore the broader applicability of the *mMMD* test within two-sample testing, particularly as a complement to recent multi-kernel approaches for boosting power. As noted in Section 6, we used the widely adopted *median heuristic* to select bandwidths for Gaussian and Laplace kernels. Though common, this heuristic lacks theoretical support and can underperform when the distributional difference occurs at scales far from the median pairwise distance. An alternative is to split the data and tune kernel parameters to maximize test power on a held-out set (Gretton et al., 2012b; Liu et al., 2020). More recent works improve robustness and power by aggregating over multiple kernels or bandwidths (Schrab et al., 2023, 2022; Biggs et al., 2023; Chatterjee and Bhattacharya, 2025), enabling detection of both local and global differences. However, these methods still rely on the quadratic-time MMD and face significant computational costs due to the need for resampling or permutation to calibrate the test statistic.

This limitation presents a promising opportunity to combine multiple kernels with the mMMD test introduced in (1.2). To that end in the following we establish the asymptotic distribution of mMMD test with multiple kernels. Consider $\mathcal{K}_r = \{\mathsf{K}_1, \ldots, \mathsf{K}_r\}$ to be a collection of bounded positive definite kernels on $\mathcal{X} = \mathbb{R}^d$ for some $d \geq 1$, such that each $\mathsf{K} \in \mathcal{K}_r$ is characteristic for $\mathcal{P}(\mathcal{X})$. Let $T_{n,\ell}$ (recall (2.5)) be the (un-normalised) mMMD statistic computed using the kernel K_ℓ. With the notation in place we now establish the asymptotic null distribution of $\boldsymbol{T}_{n,r} = (T_{n,1}, \ldots, T_{n,r})$.
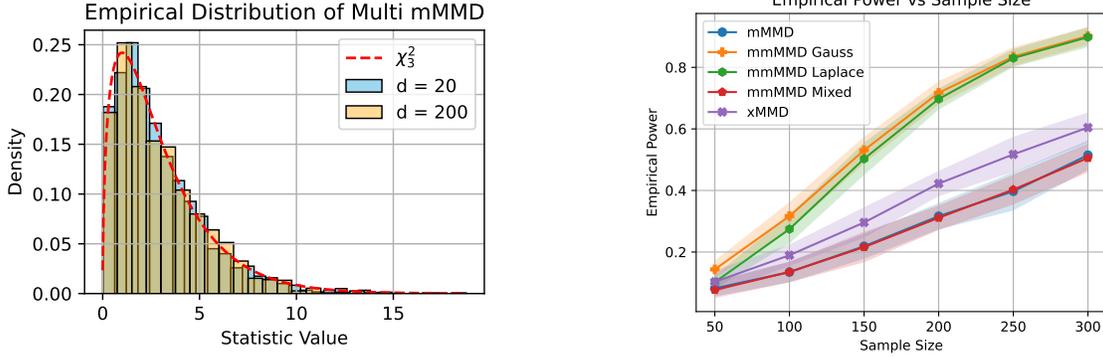
Figure 7: (Left) Empirical null distribution of `mmMMD` and (Right) Empirical power of `mmMMD`.

Theorem 10 *Let $\Sigma_{n,r} = (\sigma_n(a,b))_{a,b=1}^r$ be the non-negative definite matrix with entries $\sigma_n(a,b) := \frac{1}{n^2} \sum_{i=2}^n \left(\frac{1}{i} \sum_{j=1}^{i-1} \mathsf{H}_a(\boldsymbol{Z}_i, \boldsymbol{Z}_j) \right) \left(\frac{1}{i} \sum_{j=1}^{i-1} \mathsf{H}_b(\boldsymbol{Z}_i, \boldsymbol{Z}_j) \right)$ where H_ℓ is defined as in (5.1) using the kernel K_ℓ and $\boldsymbol{Z}_i = (X_i, Y_i), 1 \le i \le n$. Then under $\boldsymbol{H}_0$ from (1.1), $T_{n,r}^\top \Sigma_{n,r}^{-1} T_{n,r} \xrightarrow{D} \chi_r^2$.*

The proof of Theorem 10 follows from the Cramér–Wold device and the Martingale Central Limit Theorem (Hall and Heyde, 2014, Theorem 3.2), along with the arguments in Section C.1.1 and Section C.1.2, and is therefore omitted. Theorem 10 enables a test of $\boldsymbol{H}_0$ by comparing the statistic to the appropriate quantile of the χ_r^2 distribution. Importantly, this avoids the need for resampling or permutation, making the test simple to implement. For brevity, Theorem 10 is stated under specific assumptions. A more general version, analogous to Theorem 4, is possible but involves significantly more notation which are omitted.

We empirically assess the performance of the Multi-kernel mMMD (hereafter referred to as `mmMMD`) in Figure 7. In our experimental setup, we consider two distributions: $P = \mathcal{N}_{10}(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $Q = \mathcal{N}_{10}(\boldsymbol{0}, 1.3\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ denotes a covariance matrix with entries defined by $\Sigma_{ij} = 0.5^{|i-j|}$. This setting is adapted from Chatterjee and Bhattacharya (2025).

To evaluate statistical power, we consider three kernel configurations. For `mmMMD Gauss`, we use a combination of three Gaussian kernels with bandwidths $(1, 2, 4)\lambda_{\text{med}}$, where λ_{med} is the median heuristic bandwidth. Similarly, for `mmMMD Laplace`, we employ three Laplace kernels with the same set of bandwidths. The third configuration, `mmMMD Mixed`, combines both Gaussian and Laplace kernels, each with the median bandwidth λ_{med}. For estimation of the empirical null distribution, we adopt the same Gaussian kernel configuration as used in `mmMMD Gauss`.

Figure 7 validates the empirical null distribution derived in Theorem 10, and further suggests that a dimension-agnostic result similar to Theorem 4 may also hold for `mmMMD`. Moreover, the power comparison indicates that combining multiple kernels can lead to significant improvements in test power. However, it also suggests that arbitrary kernel combinations do not necessarily yield such improvements, as evidenced by the performance of `mmMMD Mixed` in the power plot. This observation highlights the need for a more principled strategy for selecting and combining kernels, pointing to an interesting future research direction in the study of optimal kernel combinations.

7.2. Generalizing the mMMD Statistic

In this section, we briefly introduce a generalization of the mMMD statistic, defining a family of statistics that interpolate between the original quadratic-time MMD statistic from Gretton et al. (2012a) and our proposed linear-time mMMD. Recalling the alternate form of T_n from (2.5), we define the generalized statistic as

$$T_{n,\gamma} = \frac{1}{n^{2-\gamma}} \sum_{i=2}^{n} \frac{1}{i^\gamma} \sum_{j=1}^{i-1} \mathsf{H}_n(\boldsymbol{Z}_i, \boldsymbol{Z}_j), \quad \text{for } 0 \le \gamma \le 1,$$

where H_n is defined in (5.1). This family recovers the quadratic-time statistic from Gretton et al. (2012a) when $\gamma = 0$ and the mMMD statistic when $\gamma = 1$, thus forming a continuous interpolation between the two.

We study this family in detail in Section A, where we establish the asymptotic null distribution under $\boldsymbol{H}_0$ as standard Gaussian (Theorem 13) and derive conditions for consistency under $\boldsymbol{H}_1$ (Theorem 16). These results generalize the corresponding theorems for the special case $\gamma = 1$ (the mMMD-test). Moreover, we show that a broad subclass of this family is minimax optimal for detecting distributional differences when the L_2 distance between the densities vanishes at the minimax rate (Li and Yuan, 2024).

Remark 11 *Although $T_{n,\gamma}$ defines a general family of statistics encompassing the mMMD test, a key limitation for $\gamma < 1$ is that our proof technique does not establish asymptotic Gaussianity under the null when the kernel is fixed (i.e., does not vary with n), unlike the $\gamma = 1$ case. In particular, for $\gamma = 0$, corresponding to the quadratic-time MMD test, the null distribution is known to be non-Gaussian in this setting. Similar behavior might hold for $\gamma < 1$ cases. On the other hand, in Section A.3, we show that $T_{n,\gamma}$ yields a minimax rate-optimal test (similar to Li and Yuan (2024)) whenever $\gamma \le 1/2$. However, this result does not apply to the mMMD test. In fact, our current analysis does not establish that the mMMD test is consistent when the L_2 distance between the underlying densities decays at the minimax rate. The current proof technique appears to be inadequate for verifying the validity of the mMMD test with respect to Type-I error control. Nonetheless, it can be shown that, up to logarithmic factors, the consistency part of the proof still holds. Whether or not $\gamma > 1/2$ is minimax optimal remains an open problem. Nonetheless, given the close connection to the cross-MMD test from Shekhar et al. (2022), which is provably minimax optimal, and the empirically similar performance of the mMMD test, a promising future direction is to characterize the class of alternatives for which the mMMD test is rate-optimal.*

8. Conclusion and Future Work

In this work, we proposed a novel variant of the kernel MMD test, called the mMMD test, by identifying a modified form of the squared MMD estimator as a martingale. We showed that this formulation admits a standard Gaussian distribution under the null, allowing for a resampling-free, computationally efficient, and consistent kernel test for the two-sample problem. Experiments demonstrate that our test achieves power close to the quadratic-time MMD test and often matches or surpasses the performance of other computationally efficient alternatives.

The established equivalence between kernel-based and distance-based tests (Sejdinovic et al., 2013) suggests that our method extends naturally to distance-based settings. Furthermore, building on the connection between kernel-based two-sample and independence testing (Gretton et al., 2007), an interesting future direction is to adapt our approach for computationally efficient independence testing. Additional computational gains might also be achieved by integrating our method with recent advances in kernel approximation techniques (Choi and Kim, 2024; Chatalic et al., 2025; Domingo-Enrich et al., 2023).

Acknowledgments

References

Theodore W Anderson. On the distribution of the two-sample Cramer-von Mises criterion. *The Annals of Mathematical Statistics*, pages 1148–1159, 1962.

Tom M Apostol. An elementary view of Euler's summation formula. *The American Mathematical Monthly*, 106(5):409–418, 1999.

Akshay Balsubramani and Aaditya Ramdas. Sequential nonparametric testing with the law of the iterated logarithm. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 42–51, 2016.

Bilol Banerjee and Anil K Ghosh. High dimensional behaviour of some two-sample tests based on ball divergence. *Statistica Sinica*, 35:1–22, 2025.

Vidmantas Bentkus and Friedrich Götze. The Berry-Esseen bound for Student's statistic. *The Annals of Probability*, 24(1):491–503, 1996.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Bhaswar B Bhattacharya. A general asymptotic framework for distribution-free graph-based two-sample tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(3): 575–602, 2019.

Peter J Bickel. A distribution free version of the Smirnov two sample test in the p-variate case. *The Annals of Mathematical Statistics*, 40(1):1–23, 1969.

Felix Biggs, Antonin Schrab, and Arthur Gretton. MMD-FUSE: Learning and combining kernels for two-sample testing without data splitting. *Advances in Neural Information Processing Systems*, 36:75151–75188, 2023.

Antoine Chatalic, Marco Letizia, Nicolas Schreuder, and Lorenzo Rosasco. An efficient permutation-based kernel two-sample test. *arXiv preprint arXiv:2502.13570*, 2025.

Anirban Chatterjee and Bhaswar B Bhattacharya. Boosting the power of kernel two-sample tests. *Biometrika*, 112(1):asae048, 2025.

Ikjun Choi and Ilmun Kim. Computational-statistical trade-off in kernel two-sample testing with random fourier features. *arXiv preprint arXiv:2407.08976*, 2024.

Carles Domingo-Enrich, Raaz Dwivedi, and Lester Mackey. Compress then test: Powerful kernel testing in near-linear time. In *International Conference on Artificial Intelligence and Statistics*, pages 1174–1218. PMLR, 2023.

Xiequan Fan and Qi-Man Shao. Berry–Esseen bounds for self-normalized martingales. *Communications in Mathematics and Statistics*, 6(1):13–27, 2018.

Jerome H Friedman and Lawrence C Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of statistics*, pages 697–717, 1979.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperumbudur. A fast, consistent kernel two-sample test. *Advances in neural information processing systems*, 22, 2009.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012a.

Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. *Advances in neural information processing systems*, 25, 2012b.

Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press, 2014.

Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of-fit test. *Advances in neural information processing systems*, 30, 2017.

Ilmun Kim. Comparing a large number of multivariate distributions. *Bernoulli*, 27(1):419–441, 2021.

Ilmun Kim, Aaditya Ramdas, Aarti Singh, and Larry Wasserman. Classification accuracy as a proxy for two-sample testing. *Annals of Statistics*, 49(1):411–434, 2021.

Jonas M Kübler, Wittawat Jitkrittum, Bernhard Schölkopf, and Krikamol Muandet. A witness two-sample test. In *International Conference on Artificial Intelligence and Statistics*, pages 1403–1419. PMLR, 2022.

Yann LeCun. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Tong Li and Ming Yuan. On the optimality of Gaussian kernel based nonparametric tests against smooth alternatives. *Journal of Machine Learning Research*, 25(334):1–62, 2024.

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning*, pages 6316–6326. PMLR, 2020.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

Soumya Mukherjee and Bharath K Sriperumbudur. Minimax optimal kernel two-sample tests with random features. *arXiv preprint arXiv:2502.20755*, 2025.

Wenliang Pan, Yuan Tian, Xueqin Wang, and Heping Zhang. Ball divergence: nonparametric two sample test. *Annals of statistics*, 46(3):1109, 2018.

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. *Advances in neural information processing systems*, 33:21247–21259, 2020.

Aaditya Ramdas, Sashank J Reddi, Barnabas Poczos, Aarti Singh, and Larry Wasserman. Adaptivity and computation-statistics tradeoffs for kernel and distance based high dimensional two sample testing. *arXiv preprint arXiv:1508.00655*, 2015a.

Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015b.

Sashank Reddi, Aaditya Ramdas, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the high dimensional power of a linear-time two sample test under mean-shift alternatives. In *Artificial Intelligence and Statistics*, pages 772–780. PMLR, 2015.

Paul R Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(4): 515–530, 2005.

Antonin Schrab, Ilmun Kim, Benjamin Guedj, and Arthur Gretton. Efficient aggregated kernel tests using incomplete U-statistics. *Advances in Neural Information Processing Systems*, 35:18793–18807, 2022.

Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, and Arthur Gretton. MMD aggregated two-sample test. *Journal of Machine Learning Research*, 24(194): 1–81, 2023.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.

Shubhanshu Shekhar and Aaditya Ramdas. Nonparametric two-sample testing by betting. *IEEE Transactions on Information Theory*, 70(2):1178–1203, 2023.

Shubhanshu Shekhar, Ilmun Kim, and Aaditya Ramdas. A permutation-free kernel two-sample test. *Advances in Neural Information Processing Systems*, 35:18168–18180, 2022.

Nickolay Smirnov. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281, 1948.

Gábor J Székely. E-statistics: The energy of statistical samples. *Bowling Green State University, Department of Mathematics and Statistics Technical Report*, 3(05):1–18, 2003.

Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

Abraham Wald and Jacob Wolfowitz. On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2):147–162, 1940.

Lionel Weiss. Two-sample tests for multivariate distributions. *The Annals of Mathematical Statistics*, 31(1):159–164, 1960.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

Geoffrey Wolfer and Pierre Alquier. Variance-aware estimation of kernel mean embedding. *Journal of Machine Learning Research*, 26(57):1–48, 2025.

Wojciech Zaremba, Arthur Gretton, and Matthew Blaschko. B-test: A non-parametric, low variance kernel two-sample test. *Advances in Neural Information Processing Systems*, 26, 2013.

Ji Zhao and Deyu Meng. FastMMD: Ensemble of circular discrepancy for efficient two-sample test. *Neural computation*, 27(6):1345–1372, 2015.

Shengjia Zhao, Abhishek Sinha, Yutong He, Aidan Perreault, Jiaming Song, and Stefano Ermon. Comparing distributions by measuring differences that affect decision making. In *International Conference on Learning Representations*, 2022.

Supplementary Materials

Contents

Appendix A. A general class of Martingale based MMD two-sample test

In this section, we introduce a general class of statistics using the squared Maximum Mean Discrepancy (MMD) (recall (2.1)). The *mMMD* estimator defined in (2.5) is a special case within this broader family. Moreover, we show that the original quadratic-time MMD estimator proposed by Gretton et al. (2012a) also belongs to this class.

In Section A.1, we show that in the most general setting, where the kernel K_n, the null distribution $P_n(= Q_n)$, and the sample space $\mathcal{X}_n = \mathbb{R}^{d_n}$ may all vary with the sample size n, the entire family of test statistics converges in distribution to a standard Gaussian under the null. Then, in Section A.2, we provide general conditions under which this family of tests achieves uniform consistency for the two-sample testing problem (see Eq. (1.1)). Finally, in Section A.3, we focus on a specific subclass of tests based on the Gaussian kernel, and show that it can consistently detect differences between P_n and Q_n when they admit densities and the L_2 distance between those densities converges to zero at the minimax optimal rate.

We begin by recalling and setting up the notation. Let $n \geq 1$ and take the sample space $\mathcal{X}_n = \mathbb{R}^{d_n}$ (we allow the dimension to vary with n). Suppose we observe independent samples $\mathscr{X}_n := \{X_1, \ldots, X_n\} \sim P_n$ and $\mathscr{Y}_n := \{Y_1, \ldots, Y_n\} \sim Q_n$. Take K_n to be a positive definite kernel with associated RKHS $\mathcal{K}_n$ satisfying Assumption 2.1 and moreover suppose $P_n, Q_n \in \mathcal{M}_{\mathsf{K}_n}^{1/2}$. For $\boldsymbol{Z}_i = (X_i, Y_i), 1 \leq i \leq n$ define,

$$\mathsf{H}_n\left(\boldsymbol{Z}_i, \boldsymbol{Z}_j\right) := \mathsf{K}_n(X_i, X_j) - \mathsf{K}_n(X_i, Y_j) - \mathsf{K}_n(X_j, Y_i) + \mathsf{K}_n(Y_i, Y_j). \tag{A.1}$$

With the notation in place we now define the family of estimators,

$$T_{n,\gamma} := \frac{1}{n^{2-\gamma}} \sum_{i=2}^{n} \frac{1}{i^\gamma} \sum_{j=1}^{i-1} \mathsf{H}_n\left(\boldsymbol{Z}_i, \boldsymbol{Z}_j\right), \text{ for all } 0 \leq \gamma \leq 1. \tag{A.2}$$

Remark 12 *Recall that the standard quadratic-time estimator of* $\mathrm{MMD}^2[\mathsf{K}_n, P_n, Q_n]$, *introduced by Gretton et al. (2012a), is given by* $\frac{1}{n^2} \sum_{i \neq j} \mathsf{H}_n(\mathbf{Z}_i, \mathbf{Z}_j)$. *By definition, and using the symmetry of the kernel function* H_n, *this expression is equivalent to* $2T_{n,0}$. *Hence, the classical quadratic-time MMD estimator is a special case of our proposed family. Moreover, the* $m\mathrm{MMD}$ *statistic introduced in (2.4) (recall the equivalent form from (2.5)) can be recovered by setting* $\gamma = 1$ *in our framework. Specifically, observe that* $T_{n,1}$ *corresponds exactly to the* $m\mathrm{MMD}$ *statistic.*

A.1. Asymptotic Null Distribution of $T_{n,\gamma}$

In this section, we establish the asymptotic null distribution of $T_{n,\gamma}$, that is, its distribution under the null hypothesis $\boldsymbol{H}_0$ where $P_n = Q_n$. To this end, we first define the centered kernel $\bar{\mathsf{K}}_n$ as,

$$\bar{\mathsf{K}}_n(x, y) := \langle \mathsf{K}_n(x, \cdot) - \nu_{P_n}, \mathsf{K}_n(y, \cdot) - \nu_{P_n} \rangle_{\mathcal{K}_n}. \tag{A.3}$$

Next, we define the terms,

$$
t_{n,\gamma,1} = \begin{cases} n^{-2} & \text{if } 0 \le \gamma < \frac{1}{2} \\ \frac{\log n}{n^2} & \text{if } \gamma = \frac{1}{2} \\ n^{-4(1-\gamma)} & \text{if } \frac{1}{2} < \gamma < 1 \\ \frac{1}{(\log n)^2} & \text{if } \gamma = 1, \end{cases} \quad \text{and} \quad t_{n,\gamma,2} = \begin{cases} n^{-1} & \text{if } 0 \le \gamma < \frac{3}{4} \\ \frac{\log n}{n} & \text{if } \gamma = \frac{3}{4} \\ n^{-4(1-\gamma)} & \text{if } \frac{3}{4} < \gamma < 1 \\ \frac{1}{(\log n)^2} & \text{if } \gamma = 1. \end{cases} \quad (A.4)
$$

With the notation in place we now establish a Berry-Essen type distributional convergence results for the family $T_{n,\gamma}$ when $P_n = Q_n$.

Theorem 13 *Generate X_1, X_2, X_3 independently from P_n and suppose $\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^4\right] < \infty$. Then define,*

$$
R_{n,\gamma,1} = t_{n,\gamma,1} \frac{\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^4\right]}{\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2\right]^2} + t_{n,\gamma,2} \left(\frac{\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2 \bar{\mathsf{K}}_n(X_1, X_3)^2\right]}{\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2\right]^2} + 1 \right), \quad (A.5)
$$

where $t_{n,\gamma,1}$ and $t_{n,\gamma,2}$ are defined in (A.4) and also let,

$$
R_{n,\gamma,2} = \begin{cases} \frac{\mathbb{E}\left[\mathbb{E}\left[\bar{\mathsf{K}}(X_1, X_2)\bar{\mathsf{K}}(X_1, X_3)|X_2, X_3\right]^2\right]}{\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2\right]^2} & \text{if } 0 \le \gamma < 1 \\ \frac{1}{\log n} & \text{if } \gamma = 1. \end{cases} \quad (A.6)
$$

Then under $\boldsymbol{H}_0$,

$$
\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(n^{2-\gamma} \frac{T_{n,\gamma}}{\sigma_{n,\gamma}} \le x \right) - \Phi(x) \right| \lesssim (R_{n,\gamma,1} + R_{n,\gamma,2})^{1/5},
$$

where $\sigma_{n,\gamma}^2 = \sum_{i=2}^n \left(\frac{1}{i^\gamma} \sum_{j=1}^{i-1} \mathsf{H}_n(\boldsymbol{Z}_i, \boldsymbol{Z}_j) \right)^2$ and $\Phi(\cdot)$ is the CDF of $\mathrm{N}(0, 1)$.

We note that the terms $R_{n,\gamma,1}$ and $R_{n,\gamma,2}$ are well-defined, as established in Lemma 21. The proof of Theorem 13 is presented in Section C.1. The key idea is to recognize $T_{n,\gamma}$ as a sum of martingale differences and then apply the Berry–Esseen bounds for self-normalized martingales from Fan and Shao (2018). To simplify the above result and to obtain conditions under which the asymptotic null distribution of $T_{n,\gamma}$ is standard Gaussian, we introduce the following assumptions.

Assumption A.1 *For the centered kernel $\bar{\mathsf{K}}_n$ and the sequence of distributions $\{P_n : n \ge 1\}$, we assume that*

$$
\lim_{n \to \infty} \frac{\mathbb{E}\left[\mathbb{E}_{P_n}\left[\bar{\mathsf{K}}_n(X_1, X_2)\bar{\mathsf{K}}_n(X_1, X_3) \mid X_2, X_3 \right]^2 \right]}{\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2\right]^2} = 0,
$$

where X_1, X_2, X_3 are generated independently from P_n.

Remark 14 *The condition in Assumption A.1 indeed holds true when $\mathsf{K}_n(x, y) = e^{-\nu_n \|x-y\|_2^2}$ is the Gaussian kernel with bandwidth $\nu_n \to \infty$. This follows from condition (9) of Li and Yuan (2024). When the kernel K_n and distribution P_n do not change with n, then it is immediately clear that the limit exists but it might not be 0.*

Assumption A.2 *For a kernel K_n and the sequence of distributions $\{P_n : n \geq 1\}$, assume that there exists some $\gamma \in [0, 1]$ such that*

$$\lim_{n\to\infty} t_{n,\gamma,1} \frac{\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^4\right]}{\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2\right]^2} = 0 \text{ and } \lim_{n\to\infty} t_{n,\gamma,2} \frac{\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2 \bar{\mathsf{K}}_n(X_1, X_3)^2\right]}{\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2\right]^2} = 0,$$

where X_1, X_2, X_3 are generated independently from P_n.

One can easily verify that this condition is satisfied whenever the kernel $\mathsf{K} = \mathsf{K}_n$ and the null distribution $P = P_n$ remain fixed as n increases. Under the assumptions stated above, we now obtain the following corollary to Theorem 13, which establishes the convergence of the null distribution of $T_{n,\gamma}$ to the standard Gaussian.

Corollary 15 *Let $\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^4\right] < \infty$ and consider some $\gamma \in [0, 1]$ such that Assumption A.2 holds. Furthermore, adopt Assumption A.1 if $\gamma \in [0, 1)$. Then under $\boldsymbol{H}_0$,*

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(n^{2-\gamma} \frac{T_{n,\gamma}}{\sigma_{n,\gamma}} \leq x \right) - \Phi(x) \right| \to 0.$$

Note that when $\gamma = 1$ the Assumption A.1 is not required. Corollary 15 follows immediately from Theorem 13, and its proof is therefore omitted. With the result from Corollary 15 we can now define the test $\phi_{n,\gamma}$ for $\boldsymbol{H}_0 : P_n = Q_n$ as,

$$\phi_{n,\gamma} = \mathbf{1}\left\{ \frac{\sum_{i=2}^n \frac{1}{i^\gamma} \sum_{j=1}^{i-1} \mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)}{\sqrt{\sum_{i=2}^n \left(\frac{1}{i^\gamma} \sum_{j=1}^{i-1} \mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_j) \right)^2}} > z_\alpha \right\} \text{ for all } 0 \leq \alpha \leq 1. \tag{A.7}$$

Then under the assumptions of Corollary 15, the test $\phi_{n,\gamma}$ is asymptotically valid at level α.

A.2. Consistency against fixed alternatives

In this section we establish the consistency of the test $\phi_{n,\gamma}$ for any $0 \leq \gamma \leq 1$. Recall the general setting where we allow the kernel K_n and the sample space $\mathcal{X}_n = \mathbb{R}^{d_n}$ to change with n and also consider $(P_n, Q_n) \in \mathcal{P}_n \subseteq \mathcal{M}_{\mathsf{K}_n}^{1/2} \times \mathcal{M}_{\mathsf{K}_n}^{1/2}$ which can also change with n. In this setting we identify sufficient conditions under which the test $\phi_{n,\gamma}$ is consistent uniformly over $\mathcal{P}_n$. Let $\delta_n = \mathrm{MMD}[P_n, Q_n, \mathsf{K}_n]$ which can approach 0 in the limit.

Theorem 16 *Define $s_{n,\gamma} = \sum_{i=2}^n (i-1)/i^\gamma$. If $\delta_n > 0$ and,*

$$\lim_{n\to\infty} \sup_{(P_n,Q_n)\in\mathcal{P}_n} \frac{\mathbb{E}\left[\frac{1}{s_{n,\gamma}^2} \sum_{i=2}^n \left(\frac{1}{i^\gamma} \sum_{j=1}^{i-1} \mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_j) \right)^2 \right]}{\delta_n^4} + \frac{\mathrm{Var}\left(\frac{1}{s_{n,\gamma}} \sum_{i=2}^n \frac{1}{i^\gamma} \sum_{j=1}^{i-1} \mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_j) \right)}{\delta_n^4} = 0,$$

(A.8)

then $\lim_{n\to\infty} \sup_{(P_n,Q_n)\in\mathcal{P}_n} \mathbb{E}\left[1 - \phi_{n,\gamma} \right] = 0$.

The proof of Theorem 16 is similar to the proof of Theorem 8 from Shekhar et al. (2022); however, for the sake of completeness, we provide a proof in Appendix C.2.

A.3. Minimax optimality against smooth local alternatives

In this section, we apply the above general consistency result to distributions P_n and Q_n which admit densities p_n and q_n lying in the order β Sobolev ball for some $\beta > 0$ denoted as

$$\mathcal{W}^{\beta,2}(M) = \left\{ f : \mathcal{X} \to \mathbb{R} \mid f \text{ is a.s. continuous, and } \int (1+\omega^2)^{\beta/2} \left\| \mathcal{F}(f)(\omega) \right\|^2 \mathrm{d}\omega < M < \infty \right\}.$$

Take $\mathcal{X} = \mathbb{R}^d$, that is the sample space (in particular the dimension) does not change with n. Formally the null and alternative class of distributions are defined as

$$\mathcal{P}_n^{(0)} := \left\{ P_n \text{ with density } p_n : p_n \in \mathcal{W}^{\beta,2}(M) \right\}$$

$$\mathcal{P}_n^{(1)} := \left\{ (P_n, Q_n) \text{ with densities } (p_n, q_n) \in \mathcal{W}^{\beta,2}(M) \text{ respectively} : \|p_n - q_n\|_2 \geq \Delta_n \right\}$$

for some sequence Δ_n decaying to 0. In particular, under $\boldsymbol{H}_0$, $P_n = Q_n \in \mathcal{P}_n^{(0)}$ and under the alternative we assume $(P_n, Q_n) \in \mathcal{P}_n^{(1)}$. The following result shows that the test $\phi_{n,\gamma}$ from (A.7) with certain choices of γ and the Gaussian kernel (with a suitably chosen scale parameter) is minimax rate optimal for the above class of local alternatives.

Theorem 17 *Let Δ_n be a sequence such that $\lim_{n\to\infty} \Delta_n n^{\frac{2\beta}{d+4\beta}} = \infty$. Then for the test $\phi_{n,\gamma}$ from (A.7) with the Gaussian kernel $\mathsf{K}_n(x,y) = \exp(-\nu_n \|x-y\|^2)$ having bandwidth $\nu_n = n^{4/(d+4\beta)}$ we have*

$$\lim_{n\to\infty} \mathbb{P}_{(P_n=Q_n)\in\mathcal{P}_n^{(0)}} \left(\phi_{n,\gamma} = 1 \right) \leq \delta \text{ and } \lim_{n\to\infty} \sup_{(P_n,Q_n)\in\mathcal{P}_n^{(1)}} \mathbb{P}(\phi_{n,\gamma} = 1) = 1,$$

whenever $0 \leq \gamma < \frac{d+8\beta}{2d+8\beta}$.

The proof of Theorem 17 is provided in Section C.3 and it follows by verifying the conditions of Theorem 16.

Remark 18 *We note that Theorem 17 establishes only one direction of the minimax optimality claim for $\phi_{n,\gamma}$. The complementary lower bound is provided by Li and Yuan (2024), who show that if $\Delta_n n^{\frac{2\beta}{d+4\beta}} = O(1)$, then there exists some $\alpha \in (0,1)$ such that no asymptotically level-α test can achieve power tending to one. This implies that the sequence Δ_n in Theorem 17 is indeed the smallest separation between P_n and Q_n (in terms of L_2 distance between densities) that any test can reliably detect. Consequently, $\phi_{n,\gamma}$ is minimax optimal for $0 \leq \gamma < \frac{d+8\beta}{2d+8\beta}$. The status of γ above this range is presently unknown.*

Appendix B. Additional Experiments

In this section we include additional experiments, including validity of mMMD tests under the null, comparative performance of $T_{n,\gamma}$ for different values of γ and comparative performance of the tests from Section 6 in the minimax regime.

B.1. Comparing performance of $T_{n,\gamma}$ for different γ

In this section, we compare the performance of tests based on $T_{n,\gamma}$ for $\gamma \in \{0, 0.25, 0.5, 0.75, 1\}$. We recall the experimental setting from Section 6.3 and apply the $T_{n,\gamma}$-based tests to distinguish between the distributions P and Q. To ensure a fair comparison across all values of γ, we use a permutation-based calibration procedure for every test. The entire experiment is repeated 100 times, and the resulting empirical performance is summarized in Figure 8.

As noted previously (see Section 6.3), the testing problem becomes increasingly difficult as the group index grows. Consistent with this, the power of the $T_{n,\gamma}$-based tests decreases for all values of γ. Moreover, the
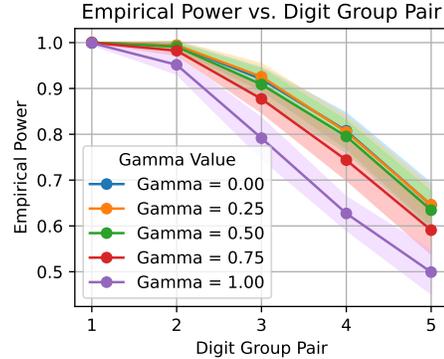
Figure 8: Empirical power comparison.

rate of decline becomes steeper as γ increases. This behavior matches the intuition that mMMD-based tests (corresponding to $\gamma = 1$) trade off statistical power for computational efficiency relative to MMD (corresponding to $\gamma = 0$), and that the family $\{T_{n,\gamma}\}_{\gamma \in [0,1]}$ provides a continuous interpolation between these two extremes.

B.2. Type-I validity of mMMD

In this section, we revisit the experimental setup from Section 6.2, fix $d = 10$, and examine Type-I error control for the quadratic-time MMD, mMMD, xMMD, the linear-time, and the block-based tests in both small- and large-sample regimes.
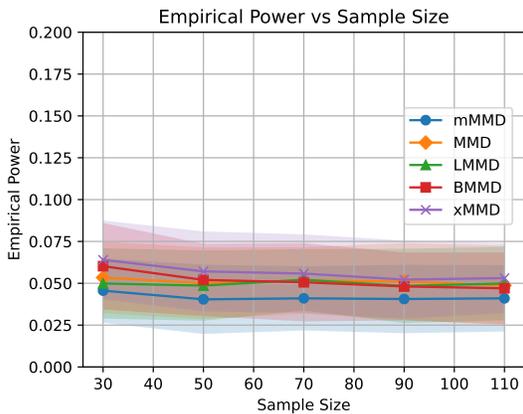
(a) Type-I error control at small samples.
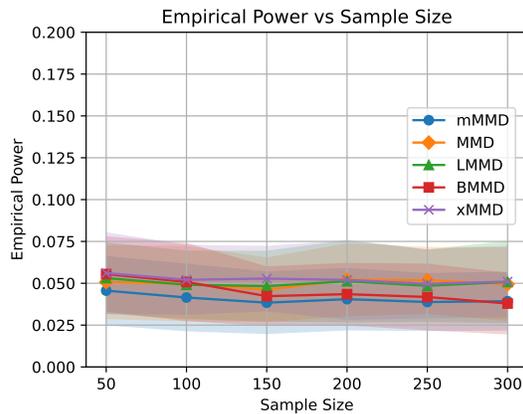
(b) Type-I error control at large samples.

Figure 9: Type-I error control for the original quadratic time MMD, mMMD, xMMD and the linear time LMMD and block-based BMMD.

Figure 9 illustrates the Type-I error control of all the aforementioned tests in both small- and large-sample settings. While all methods maintain the nominal Type-I error level $\alpha = 0.05$ in large-sample scenarios, in the small-sample setting with $n = 10$, the xMMD and block-based test (`BMMD`) fail. This is because both rely on sample splitting and block-based approximations, which are less accurate in such small-sample regimes.

B.3. Performance in minimax regime

Building on Remark 11 and Theorem 17, we observe that our current proof technique does not establish the rate-optimality of the mMMD test. Nevertheless, the experiments in Section 6 indicate that, empirically, the mMMD test performs on par with the xMMD test of Shekhar et al. (2022), which is known to be minimax optimal. To determine whether the absence of a rate-optimality guarantee for the mMMD test reflects a limitation of our analytical approach or an inherent limitation of the test itself, we design in this section a set of experiments that closely mirror the conditions of Theorem 17. In particular we take

$$p_n = \mathrm{N}(0, 1) \text{ and } q_n = \mathrm{N}\left(2\sqrt{-\log\left(1 - \Delta_n^2 \sqrt{\pi}\right)}, 1\right).$$

With these choices, it immediately follows that $\|p_n - q_n\|_2 = \Delta_n$. In Figure 10, we present the performance of the quadratic-time MMD, mMMD, and xMMD tests, as well as the linear-time and block-based MMD tests, in detecting the difference between p_n and q_n. All tests use a Gaussian kernel with bandwidth $\nu_n = n^{4/5}$ (see the choice of bandwidth in Theorem 17), we vary $n \in \{100, 200, 300, 400, 500\}$ and we choose $\Delta_n = n^{-1/5}, n^{-1/4}$. Note that this choice of Δ_n satisfies the constraint from Theorem 17 with $\beta = 1$.

(a) Type-I Error   (b) Power with $\Delta_n = n^{-\frac{1}{5}}$ (c) Power wth $\Delta_n = n^{-\frac{1}{4}}$
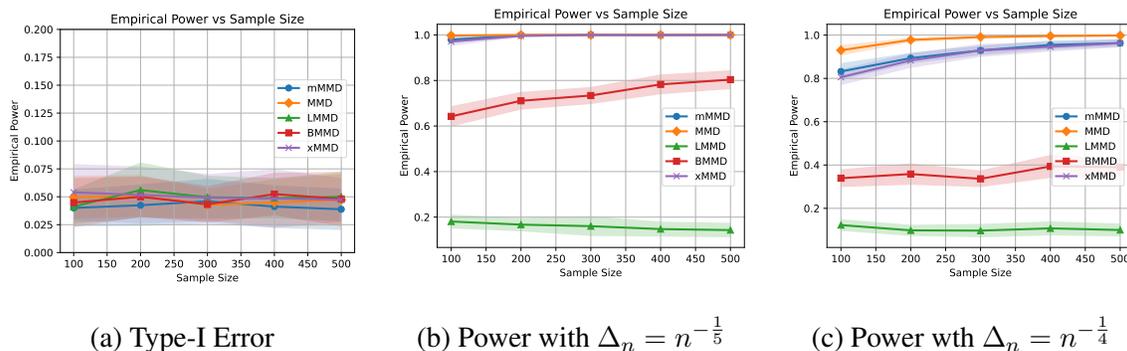
Figure 10: Performance of MMD, mMMD, xMMD (Shekhar et al., 2022) and the linear time `LMMD` and the block-based `BMMD` in the setting from Section B.3.

Figure 10 suggests that, in the minimax regime, the mMMD test performs comparably to both the xMMD and the quadratic-time MMD tests in terms of type-I error control and statistical power. This indicates that the limitation from Remark 11 may arise from the proof technique rather than the test itself, and it is possible that mMMD also achieves rate-optimality.

Appendix C. Proof of results from Section A

In this section we collected the deferred proofs of results from Section A.

C.1. Proof of Theorem 13

The proof proceeds by recognizing $T_{n,\gamma}$ as a martingale and applying the Berry–Esseen bounds for self-normalized martingales from Fan and Shao (2018). To that end, we begin by introducing the necessary notation. Define

$$c_\gamma := \begin{cases} \frac{2}{1-\gamma} & \text{if } 0 \leq \gamma < 1 \\ 4 & \text{if } \gamma = 1. \end{cases} \tag{C.1}$$

Take $M_n = \mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2\right]$ where X_1, X_2 are generated independently from P_n. To identify $T_{n,\gamma}$ as a martingale, define $W_{n,1,\gamma} = 0, 0 \leq \gamma \leq 1$ and

$$W_{n,i,\gamma} = \begin{cases} \frac{1}{n^{1-\gamma}\sqrt{c_\gamma M_n} i^\gamma} \sum_{j=1}^{i-1} \mathsf{H}_n(\boldsymbol{Z}_i, \boldsymbol{Z}_j) & \text{if } 0 \leq \gamma < 1 \\ \frac{1}{\sqrt{\log n}\sqrt{c_1 M_n} i} \sum_{j=1}^{i-1} \mathsf{H}_n(\boldsymbol{Z}_i, \boldsymbol{Z}_j) & \text{if } \gamma = 1. \end{cases} \quad \text{for all } 2 \leq i \leq n,$$

where we take $\boldsymbol{Z}_i = (X_i, Y_i), 1 \leq i \leq n$ and H_n is defined in (A.1). Moreover, consider the sigma algebras $\mathcal{F}_{n,1} = \mathcal{F}_1 = \{\emptyset, \Omega\}$ and $\mathcal{F}_{n,i} := \mathcal{F}_i = \sigma\left(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_i\right)$ for all $2 \leq i \leq n$. Finally, let

$$S_{n,i,\gamma} = \sum_{j=1}^{i} W_{n,j,\gamma} \text{ for all } 0 \leq \gamma \leq 1.$$

It is easy to see that under the null, for $n \geq 1$, $\{(S_{n,i,\gamma}, \mathcal{F}_{n,i}) : 1 \leq i \leq n\}$ forms a zero-mean, square integrable martingale. Finally, notice that,

$$n^{2-\gamma} \frac{T_{n,\gamma}}{\sigma_{n,\gamma}} = \frac{\sum_{i=2}^{n} \frac{1}{i^\gamma} \sum_{j=1}^{i-1} \mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)}{\sqrt{\sum_{i=2}^{n}\left(\frac{1}{i^\gamma}\sum_{j=1}^{i-1}\mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)\right)^2}} = \frac{S_{n,n,\gamma}}{\sqrt{\sum_{i=2}^{n} W_{n,i,\gamma}^2}}.$$

Then by Theorem 2.1 from Fan and Shao (2018) (the required moment condition is satisfied by following computations similar to (C.7), (C.10) and the moment assumption on $\bar{\mathsf{K}}_n$) we get,

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(n^{2-\gamma}\frac{T_{n,\gamma}}{\sigma_{n,\gamma}} \leq x\right) - \Phi(x) \right| \lesssim N_{n,\gamma}^{1/5},$$

where

$$N_{n,\gamma} = \sum_{i=2}^{n} \mathbb{E}\left[W_{n,i,\gamma}^4\right] + \mathbb{E}\left[\left|\sum_{i=2}^{n} \mathbb{E}\left[W_{n,i,\gamma}^2 \mid \mathcal{F}_{n,i-1}\right] - 1\right|^2\right]. \tag{C.2}$$

To complete the proof it is now enough to upper bound $N_{n,\gamma}$. We begin by a simple upper bound on $N_{n,\gamma}$. Note that

$$\mathbb{E}\left[\left(\sum_{i=2}^{n} W_{n,i,\gamma}^2 - \mathbb{E}\left[W_{n,i,\gamma}^2 \mid \mathcal{F}_{n,i-1}\right]\right)^2\right] = \sum_{i=2}^{n} \mathbb{E}\left[\left(W_{n,i,\gamma}^2 - \mathbb{E}\left[W_{n,i,\gamma}^2\right]\right)^2\right] \lesssim \sum_{i=2}^{n} \mathbb{E}\left[W_{n,i,\gamma}^4\right]. \tag{C.3}$$

Then recalling definition of $N_{n,\gamma}$ from (C.2) and by (C.3) we get,

$$N_{n,\gamma} \lesssim \sum_{i=2}^{n} \mathbb{E}\left[W_{n,i,\gamma}^4\right] + \mathbb{E}\left[\left|\sum_{i=2}^{n} W_{n,i,\gamma}^2 - 1\right|^2\right] \tag{C.4}$$

To complete the proof, we proceed by deriving upper bounds for the terms on the right-hand side of (C.4). A key step in this analysis is the following identity, which holds under the null hypothesis for all $1 \le i \ne j \le n$:

$$\mathsf{H}_n(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = \bar{\mathsf{K}}_n(X_i, X_j) - \bar{\mathsf{K}}_n(X_i, Y_j) - \bar{\mathsf{K}}_n(Y_i, X_j) + \bar{\mathsf{K}}_n(Y_i, Y_j). \tag{C.5}$$

This expression will serve as a crucial tool in establishing the bounds from Theorem 13. In the following lemma we first establish the upper bound on the first term from (C.4).

Lemma 19 *Under the assumptions of Theorem 13,*

$$\sum_{i=2}^{n} \mathbb{E}\left[W_{n,i,\gamma}^4\right] \lesssim t_{n,\gamma,1} \frac{\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^4\right]}{M_n^2} + t_{n,\gamma,2} \frac{\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2 \bar{\mathsf{K}}_n(X_1, X_3)^2\right]}{M_n^2} + t_{n,\gamma,2}. \tag{C.6}$$

We postpone the proof Lemma 19 to Section C.1.1. In the following lemma we establish an upper bound on the second term in (C.4).

Lemma 20 *Under the assumptions on Theorem 13,*

$$\mathbb{E}\left[\left|\sum_{i=2}^{n} W_{n,i,\gamma}^2 - 1\right|^2\right] \lesssim R_{n,\gamma,1} + R_{n,\gamma,2},$$

where $R_{n,\gamma,1}$ and $R_{n,\gamma,2}$ are defined in (A.5) and (A.6) respectively.

The proof of Lemma 20 is postponed to Section C.1.2. The proof of Theorem 13 is now completed by recalling the bound on $N_{n,\gamma}$ from (C.4) and substituting the upper bounds on the individual terms from Lemma 19 and Lemma 20.

C.1.1. PROOF OF LEMMA 19

By definition, if $0 \le \gamma < 1$,

$$\sum_{i=2}^{n} \mathbb{E}\left[W_{n,i,\gamma}^4\right] = \sum_{i=2}^{n} \frac{1}{n^{4(1-\gamma)} M_n^2 i^{4\gamma} c_\gamma^2} \sum_{j_1,j_2,j_3,j_4=1}^{i-1} \mathbb{E}\left[\prod_{\ell=1}^{4} \mathsf{H}_n(\boldsymbol{Z}_i, \boldsymbol{Z}_{j_\ell})\right]$$

$$= \sum_{i=2}^{n} \frac{1}{n^{4(1-\gamma)} M_n^2 i^{4\gamma} c_\gamma^2} \left[\sum_{j=1}^{i-1} \mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_i, \boldsymbol{Z}_j)^4\right] + \sum_{j_1 \ne j_2=1}^{i-1} \mathbb{E}\left[\mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_{j_1})^2 \mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_{j_2})^2\right]\right],$$

where the last equality follows by recalling that under null $\mathbb{E}\left[\mathsf{H}(\boldsymbol{Z}_1, \boldsymbol{Z}_2) \mid \boldsymbol{Z}_1\right] = 0$. Continuing with the chain of equalities

$$\sum_{i=2}^{n} \mathbb{E}\left[W_{n,i,\gamma}^4\right] = \sum_{i=2}^{n} \frac{1}{n^{4(1-\gamma)} M_n^2 i^{4\gamma} c_\gamma^2} \left[(i-1)\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^4\right]\right.$$

$$\left. + (i-1)(i-2)\mathbb{E}\left[\mathsf{H}(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2 \mathsf{H}(\boldsymbol{Z}_1, \boldsymbol{Z}_3)^2\right]\right]$$

$$\lesssim \frac{\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^4\right]}{M_n^2} \sum_{i=2}^{n} \frac{i^{1-4\gamma}}{n^{4(1-\gamma)}} + \frac{\mathbb{E}\left[\mathsf{H}(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2 \mathsf{H}(\boldsymbol{Z}_1, \boldsymbol{Z}_3)^2\right]}{M_n^2} \sum_{i=2}^{n} \frac{i^{2-4\gamma}}{n^{4(1-\gamma)}}. \tag{C.7}$$

Now by Lemma 22 we get

$$\sum_{i=2}^{n} \frac{i^{1-4\gamma}}{n^{4(1-\gamma)}} = \begin{cases} \Theta(n^{-2}) & \text{if } 0 \le \gamma < \frac{1}{2} \\ \Theta\left(\frac{\log n}{n^2}\right) & \text{if } \gamma = \frac{1}{2} \\ \Theta\left(n^{-4(1-\gamma)}\right) & \text{if } \frac{1}{2} < \gamma < 1. \end{cases} \quad \text{and} \quad \sum_{i=2}^{n} \frac{i^{2-4\gamma}}{n^{4(1-\gamma)}} = \begin{cases} \Theta(n^{-1}) & \text{if } 0 \le \gamma < \frac{3}{4} \\ \Theta\left(\frac{\log n}{n}\right) & \text{if } \gamma = \frac{3}{4} \\ \Theta\left(n^{-4(1-\gamma)}\right) & \text{if } \frac{3}{4} < \gamma < 1. \end{cases}.$$

Hence recalling the bound from (C.7) and notation from (A.4) we get

$$\sum_{i=2}^{n} \mathbb{E}\left[W_{n,i,\gamma}^4\right] \lesssim t_{n,\gamma,1} \frac{\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^4\right]}{M_n^2} + t_{n,\gamma,2} \frac{\mathbb{E}\left[\mathsf{H}(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2 \mathsf{H}(\boldsymbol{Z}_1, \boldsymbol{Z}_3)^2\right]}{M_n^2}. \tag{C.8}$$

Furthermore, if $\gamma = 1$, then similar computations as above shows

$$\sum_{i=2}^{n} \mathbb{E}\left[W_{n,i,\gamma}^4\right] \lesssim \frac{\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^4\right]}{M_n^2} \sum_{i=2}^{n} \frac{i^{-3}}{(\log n)^2} + \frac{\mathbb{E}\left[\mathsf{H}(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2 \mathsf{H}(\boldsymbol{Z}_1, \boldsymbol{Z}_3)^2\right]}{M_n^2} \sum_{i=2}^{n} \frac{i^{-2}}{(\log n)^2}$$

$$\lesssim \frac{\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^4\right]}{(\log n)^2 M_n^2} + \frac{\mathbb{E}\left[\mathsf{H}(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2 \mathsf{H}(\boldsymbol{Z}_1, \boldsymbol{Z}_3)^2\right]}{(\log n)^2 M_n^2}. \tag{C.9}$$

The proof is now completed by noting that under null, $\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^4\right] \lesssim \mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^4\right]$ and

$$\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2 \mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_3)^2\right] \lesssim \mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2 \bar{\mathsf{K}}_n(X_1, X_3)^2\right] + M_n^2. \tag{C.10}$$

C.1.2. PROOF OF LEMMA 20

By expanding the square,

$$\mathbb{E}\left[\left|\sum_{i=2}^{n} W_{n,i,\gamma}^2 - 1\right|^2\right] = \mathbb{E}\left[\left(\sum_{i=2}^{n} W_{n,i,\gamma}^2\right)^2\right] - 2\sum_{i=2}^{n} \mathbb{E}\left[W_{n,i,\gamma}^2\right] + 1. \tag{C.11}$$

Note that if $0 \le \gamma < 1$, then

$$\sum_{i=2}^{n} \mathbb{E}\left[W_{n,i,\gamma}^2\right] = \sum_{i=2}^{n} \frac{1}{n^{2(1-\gamma)} M_n i^{2\gamma} c_\gamma} \mathbb{E}\left[\sum_{j_1, j_2 = 1}^{i-1} \mathsf{H}_n(\boldsymbol{Z}_i, \boldsymbol{Z}_{j_1}) \mathsf{H}_n(\boldsymbol{Z}_i, \boldsymbol{Z}_{j_2})\right]$$

$$= \frac{\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2\right]}{M_n c_\gamma} \frac{1}{n^{2(1-\gamma)}} \sum_{i=2}^{n} \frac{i-1}{i^{2\gamma}}. \tag{C.12}$$

Moreover, expanding the term $\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2\right]$ note that

$$\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2\right] = 4\mathbb{E}\left[\bar{\mathsf{K}}(X_1, X_2)^2\right] - 8\mathbb{E}\left[\bar{\mathsf{K}}(X_1, X_2)\bar{\mathsf{K}}(X_1, X_3)\right] = 4\mathbb{E}\left[\bar{\mathsf{K}}(X_1, X_2)^2\right] \tag{C.13}$$

where the last equality follows by noticing that $\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2) \mid X_1\right] = 0$. Hence for $0 \le \gamma < 1$ from (C.12) and (C.13) we conclude

$$\sum_{i=2}^{n} \mathbb{E}\left[W_{n,i,\gamma}^2\right] = \frac{4}{c_\gamma} \frac{1}{n^{2(1-\gamma)}} \sum_{i=2}^{n} \frac{i-1}{i^{2\gamma}}.$$

Similar computations for $\gamma = 1$ shows

$$\sum_{i=2}^{n} \mathbb{E}\left[W_{n,i,1}^2\right] = \frac{1}{c_1 \log n} \sum_{i=2}^{n} \frac{i-1}{i^2}.$$

Then substituting the values in (C.11) we get

$$\mathbb{E}\left[\left|\sum_{i=2}^{n} W_{n,i,\gamma}^2 - 1\right|^2\right] = \begin{cases} \mathbb{E}\left[\left(\sum_{i=2}^{n} W_{n,i,\gamma}^2\right)^2\right] - \frac{8}{c_\gamma} \frac{1}{n^{2(1-\gamma)}} \sum_{i=2}^{n} \frac{i-1}{i^{2\gamma}} + 1 & \text{if } 0 \le \gamma < 1 \\ \mathbb{E}\left[\left(\sum_{i=2}^{n} W_{n,i,\gamma}^2\right)^2\right] - \frac{2}{c_1 \log n} \sum_{i=2}^{n} \frac{i-1}{i^2} + 1 & \text{if } \gamma = 1. \end{cases} \tag{C.14}$$

Expanding the first term shows

$$\mathbb{E}\left[\left(\sum_{i=2}^{n} W_{n,i,\gamma}^2\right)^2\right] = \sum_{i=2}^{n} \mathbb{E}\left[W_{n,i,\gamma}^4\right] + 2 \sum_{i_1 < i_2} \mathbb{E}\left[W_{n,i_1,\gamma}^2 W_{n,i_2,\gamma}^2\right]. \tag{C.15}$$

In the above expansion, note that applying Lemma 19, the first term can be upper bounded by $R_{n,\gamma,1}$. Consequently in the following we further expand the second term $\sum_{i_1 < i_2} \mathbb{E}\left[W_{n,i_1,\gamma}^2 W_{n,i_2,\gamma}^2\right]$. To that end for $1 \le m \le 4$ we begin by defining the set of ordered tuples,

$$\mathcal{S}_{i_1,i_2,m} = \{(k_1, k_2, \ell_1, \ell_2) : |\{j_1, j_2, \ell_1, \ell_2\}| = m, 1 \le j_1, j_2 \le i_1 - 1, 1 \le \ell_1, \ell_2 \le i_2 - 1\},$$

where $\mathcal{S}_{i_1,i_2,m}$ collects all the ordered four-tuples such that the first two coordinated can take values at most $i_1 - 1$, the second two coordinates can take values at most $i_2 - 1$ and there are exactly m many distinct values in the tuple. Finally for $1 \le m \le 4$ define,

$$L_{m,n,\gamma} = \begin{cases} \frac{1}{n^{4(1-\gamma)} M_n^2} \sum_{i_1 < i_2} \frac{1}{i_1^{2\gamma} i_2^{2\gamma}} \sum_{\mathcal{S}_{i_1,i_2,m}} \mathbb{E}\left[\prod_{t=1}^{2} \mathsf{H}_n(\boldsymbol{Z}_{i_1}, \boldsymbol{Z}_{j_t}) \mathsf{H}_n(\boldsymbol{Z}_{i_2}, \boldsymbol{Z}_{\ell_t})\right] & \text{if } 0 \le \gamma < 1 \\ \frac{1}{(M_n \log n)^2} \sum_{i_1 < i_2} \frac{1}{i_1^2 i_2^2} \sum_{\mathcal{S}_{i_1,i_2,m}} \mathbb{E}\left[\prod_{t=1}^{2} \mathsf{H}_n(\boldsymbol{Z}_{i_1}, \boldsymbol{Z}_{j_t}) \mathsf{H}_n(\boldsymbol{Z}_{i_2}, \boldsymbol{Z}_{\ell_t})\right] & \text{if } \gamma = 1. \end{cases} \tag{C.16}$$

Then, by definition we have the following expansion.

$$\sum_{i_1 < i_2 = 2}^{n} \mathbb{E}\left[W_{n,i_1,\gamma}^2 W_{n,i_2,\gamma}^2\right] = \frac{1}{c_\gamma^2} \sum_{m=1}^{4} L_{m,n,\gamma} = \frac{1}{c_\gamma^2} \sum_{m=1}^{2} L_{m,n,\gamma}, \tag{C.17}$$

where the final equality follows by noticing that under $\boldsymbol{H}_0$, $\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2) \mid \boldsymbol{Z}_1\right] = 0$ almost surely. Next, we further decompose the term $L_{2,n,\gamma}$. Define

$$\mathcal{S}_{i_1,i_2,2}^{\circ} = \{(j_1, j_2, \ell_1, \ell_2) : j_1 = j_2 = j, \ell_1 = \ell_2 = \ell, 1 \le k \le i_1 - 1, 1 \le \ell \le i_2 - 1, k \ne \ell\},$$

Then, by construction $\mathcal{S}_{i_1,i_2,2}^{\circ} \subseteq \mathcal{S}_{i_1,i_2,2}$. Let

$$L_{2,n,\gamma}^{\circ} = \begin{cases} \frac{1}{n^{4(1-\gamma)} M_n^2} \sum_{i_1 < i_2} \frac{1}{i_1^{2\gamma} i_2^{2\gamma}} \sum_{\mathcal{S}_{i_1,i_2,2}^{\circ}} \mathbb{E}\left[\prod_{t=1}^{2} \mathsf{H}_n(\boldsymbol{Z}_{i_1}, \boldsymbol{Z}_{j_t}) \mathsf{H}_n(\boldsymbol{Z}_{i_2}, \boldsymbol{Z}_{\ell_t})\right] & \text{if } 0 \le \gamma < 1 \\ \frac{1}{(M_n \log n)^2} \sum_{i_1 < i_2} \frac{1}{i_1^2 i_2^2} \sum_{\mathcal{S}_{i_1,i_2,2}^{\circ}} \mathbb{E}\left[\prod_{t=1}^{2} \mathsf{H}_n(\boldsymbol{Z}_{i_1}, \boldsymbol{Z}_{j_t}) \mathsf{H}_n(\boldsymbol{Z}_{i_2}, \boldsymbol{Z}_{\ell_t})\right] & \text{if } \gamma = 1. \end{cases} \tag{C.18}$$

29

We now decompose $L_{2,n,\gamma}$ as $L_{2,n,\gamma} = L_{2,n,\gamma}^\circ + (L_{2,n,\gamma} - L_{2,n,\gamma}^\circ)$. With the notations in place, combining (C.14), (C.15) and (C.17) we have

$$\mathbb{E}\left[\left|\sum_{i=2}^n W_{n,i,\gamma}^2 - 1\right|^2\right] = \sum_{i=2}^n \mathbb{E}\left[W_{n,i,\gamma}^4\right] + \frac{2}{c_\gamma^2} L_{1,n,\gamma} + \frac{2}{c_\gamma^2}(L_{2,n,\gamma} - L_{2,n,\gamma}^\circ) + \zeta_{n,\gamma}, \quad \text{(C.19)}$$

where

$$\zeta_{n,\gamma} = \begin{cases} \frac{2}{c_\gamma^2} L_{2,n,\gamma}^\circ - \frac{8}{c_\gamma} \frac{1}{n^{2(1-\gamma)}} \sum_{i=2}^n \frac{i-1}{i^{2\gamma}} + 1 & \text{if } 0 \le \gamma < 1 \\ \frac{2}{c_\gamma^2} L_{2,n,\gamma}^\circ - \frac{2}{c_1 \log n} \sum_{i=2}^n \frac{i-1}{i^2} + 1 & \text{if } \gamma = 1. \end{cases}$$

In the following we complete the proof by establishing bound on the individual terms in (C.19). Note that by Lemma 19 we already know that $\sum_{i=2}^n \mathbb{E}\left[W_{n,i,\gamma}^4\right] \lesssim R_{n,\gamma,1}$. Next, we bound the term $L_{1,n,\gamma}$.

Bound on $L_{1,n,\gamma}$: For $L_{1,n,\gamma}$, note that for any $i_1 < i_2$ with $(j_1, j_2, \ell_1, \ell_2) \in \mathcal{S}_{i_1, i_2, 1}$,

$$\mathbb{E}\left[\prod_{t=1}^2 \mathsf{H}_n(\boldsymbol{Z}_{i_1}, \boldsymbol{Z}_{j_t}) \mathsf{H}_n(\boldsymbol{Z}_{i_2}, \boldsymbol{Z}_{\ell_t})\right] = \mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2 \mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_3)^2\right],$$

and $|\mathcal{S}_{i_1, i_2, 1}| = i_1 - 1$. Then for $0 \le \gamma < 1$,

$$\begin{aligned} L_{1,n,\gamma} &\lesssim \frac{\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2 \mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_3)^2\right]}{M_n^2} \frac{1}{n^{4(1-\gamma)}} \sum_{i_1 < i_2} \frac{i_1 - 1}{i_1^{2\gamma} i_2^{2\gamma}} \\ &\lesssim \frac{\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2 \mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_3)^2\right]}{M_n^2} \frac{1}{n^{4(1-\gamma)}} \sum_{i_2=2}^n \frac{1}{i_2^{2\gamma}} \sum_{i_1=1}^{i_2-1} i_1^{1-2\gamma} \\ &\lesssim \frac{\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2 \mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_3)^2\right]}{M_n^2} t_{n,\gamma,2}, \end{aligned} \quad \text{(C.20)}$$

where the final bound follows from Lemma 22. Similarly for $\gamma = 1$ we get

$$\begin{aligned} L_{1,n,\gamma} &\lesssim \frac{\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2 \mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_3)^2\right]}{M_n^2} \frac{1}{(\log n)^2} \sum_{i_2=2}^n \frac{1}{i_2^2} \sum_{i_1=1}^{i_2-1} i_1^{-1} \\ &\lesssim \frac{\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2 \mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_3)^2\right]}{M_n^2} t_{n,1,2}. \end{aligned} \quad \text{(C.21)}$$

Hence recalling (C.10) and using the bounds (C.20) and (C.21) we conclude

$$L_{1,n,\gamma} \lesssim \frac{\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2 \mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_3)^2\right]}{M_n^2} t_{n,\gamma,2} \lesssim \frac{\mathbb{E}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2 \bar{\mathsf{K}}_n(X_1, X_3)^2\right]}{M_n^2} t_{n,\gamma,2} + t_{n,\gamma,2}. \quad \text{(C.22)}$$

In the following we bound the term $\zeta_{n,\gamma}$.

Bound on $\zeta_{n,\gamma}$:   To bound $\zeta_{n,\gamma}$ we first simplify the term $L_{2,n,\gamma}^{\circ}$. If $0 \leq \gamma < 1$ then by construction,

$$
\begin{aligned}
L_{2,n,\gamma}^{\circ} &= \frac{1}{n^{4(1-\gamma)} M_n^2} \sum_{i_1 < i_2} \frac{1}{i_1^{2\gamma} i_2^{2\gamma}} \sum_{j=1}^{i_1-1} \sum_{\ell=1}^{i_2-1} \mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_j, \boldsymbol{Z}_\ell)^2\right]^2 \mathbf{1}\{j \neq \ell\} \\
&= \frac{16}{n^{4(1-\gamma)}} \sum_{i_1 < i_2} \frac{(i_1 - 1)(i_2 - 2)}{i_1^{2\gamma} i_2^{2\gamma}},
\end{aligned}
\tag{C.23}
$$

where the last equality follows from (C.13). Similarly if $\gamma = 1$, then

$$
L_{2,n,\gamma}^{\circ} = \frac{16}{(\log n)^2} \sum_{i_1 < i_2} \frac{(i_1 - 1)(i_2 - 2)}{i_1^2 i_2^2}.
\tag{C.24}
$$

Now for $0 \leq \gamma < 1$ with the identity from (C.23) notice,

$$
\begin{aligned}
\zeta_{n,\gamma} &= \frac{2}{c_\gamma^2} L_{2,n,\gamma}^{\circ} - \frac{8}{c_\gamma} \frac{1}{n^{2(1-\gamma)}} \sum_{i=2}^{n} \frac{i-1}{i^{2\gamma}} + 1 \\
&= \left(\frac{4}{c_\gamma n^{2(1-\gamma)}} \sum_{i=2}^{n} \frac{i-1}{i^{2\gamma}} - 1\right)^2 - \frac{32}{c_\gamma^2 n^{4(1-\gamma)}} \sum_{i<j} \frac{(i-1)}{i^{2\gamma} j^{2\gamma}} - \frac{16}{c_\gamma^2 n^{4(1-\gamma)}} \sum_{i=2}^{n} \left(\frac{i-1}{i^{2\gamma}}\right)^2.
\end{aligned}
$$

By Lemma 22 we get

$$
\frac{1}{n^{4(1-\gamma)}} \sum_{i=2}^{n} \left(\frac{i-1}{i^{2\gamma}}\right)^2 \lesssim \begin{cases} 1/n & \text{if } 0 \leq \gamma < 3/4 \\ \log n / n & \text{if } \gamma = 3/4 \\ 1/n^{4(1-\gamma)} & \text{if } 3/4 < \gamma < 1. \end{cases}
$$

Similarly applying Lemma 22 repeatedly we find

$$
\frac{1}{n^{4(1-\gamma)}} \sum_{i<j} \frac{(i-1)}{i^{2\gamma} j^{2\gamma}} \lesssim \begin{cases} 1/n & \text{if } 0 \leq \gamma < 3/4 \\ \log n / n & \text{if } \gamma = 3/4 \\ 1/n^{4(1-\gamma)} & \text{if } 3/4 < \gamma < 1. \end{cases}
$$

Furthermore, from Lemma 23 and recalling definition of c_γ from (C.1) we know

$$
\left| \frac{4}{c_\gamma n^{2(1-\gamma)}} \sum_{i=2}^{n} \frac{i-1}{i^{2\gamma}} - 1 \right| \lesssim \begin{cases} \frac{1}{n} & \text{if } 0 \leq \gamma < 1/2 \\ \frac{\log n}{n} & \text{if } \gamma = 1/2 \\ \frac{1}{n^{2(1-\gamma)}} & \text{if } 1/2 < \gamma < 1 \end{cases}.
$$

Combining we conclude

$$
\left| \frac{2}{c_\gamma^2} L_{2,n,\gamma}^{\circ} - \frac{8}{c_\gamma} \frac{1}{n^{2(1-\gamma)}} \sum_{i=2}^{n} \frac{i-1}{i^{2\gamma}} + 1 \right| \lesssim t_{n,\gamma,2} = \begin{cases} 1/n & \text{if } 0 \leq \gamma < 3/4 \\ \log n / n & \text{if } \gamma = 3/4 \\ 1/n^{4(1-\gamma)} & \text{if } 3/4 < \gamma < 1 \end{cases}.
$$

Similarly for $\gamma = 1$, using the fact $\sum_{i=1}^{n} \frac{1}{i} = \log n + O(1)$ (see Theorem 1 from Apostol (1999)) and using the identity from (C.24) we get,

$$\left| \frac{2}{c_\gamma^2} L_{2,n,\gamma}^\circ - \frac{8}{c_\gamma} \frac{1}{\log n} \sum_{i=2}^{n} \frac{i-1}{i^2} + 1 \right| \lesssim \frac{1}{(\log n)^2} = t_{n,1,2}.$$

Hence we conclude that for all $0 \le \gamma \le 1$,

$$\zeta_{n,\gamma} \lesssim t_{n,\gamma,2}. \tag{C.25}$$

Finally in the following we bound the difference $L_{2,n,\gamma} - L_{2,n,\gamma}^\circ$.

Bound on $L_{2,n,\gamma} - L_{2,n,\gamma}^\circ$ To bound the difference $L_{2,n,\gamma} - L_{2,n,\gamma}^\circ$ note that,

$$|L_{2,n,\gamma} - L_{2,n,\gamma}^\circ|$$
$$\lesssim \begin{cases} \frac{1}{n^{4(1-\gamma)} M_n^2} \sum_{i_1 < i_2} \frac{1}{i_1^{2\gamma} i_2^{2\gamma}} \sum_{j \ne \ell = 1}^{i_1 - 1} |\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2) \mathsf{H}_n(\boldsymbol{Z}_2, \boldsymbol{Z}_3) \mathsf{H}_n(\boldsymbol{Z}_3, \boldsymbol{Z}_4) \mathsf{H}_n(\boldsymbol{Z}_4, \boldsymbol{Z}_1) \right]| & 0 \le \gamma < 1 \\ \frac{1}{(\log n)^2 M_n^2} \sum_{i_1 < i_2} \frac{1}{i_1^2 i_2^2} \sum_{j \ne \ell = 1}^{i_1 - 1} |\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2) \mathsf{H}_n(\boldsymbol{Z}_2, \boldsymbol{Z}_3) \mathsf{H}_n(\boldsymbol{Z}_3, \boldsymbol{Z}_4) \mathsf{H}_n(\boldsymbol{Z}_4, \boldsymbol{Z}_1) \right]| & \gamma = 1. \end{cases}$$

The inner expectation evaluates to

$$\mathbb{E}\left[\mathsf{H}(\boldsymbol{Z}_1, \boldsymbol{Z}_2) \mathsf{H}(\boldsymbol{Z}_2, \boldsymbol{Z}_3) \mathsf{H}(\boldsymbol{Z}_3, \boldsymbol{Z}_4) \mathsf{H}(\boldsymbol{Z}_4, \boldsymbol{Z}_1) \right] = 16 \mathbb{E}\left[\mathbb{E}\left[\bar{\mathsf{K}}(X_1, X_2) \bar{\mathsf{K}}(X_1, X_3) \mid X_2, X_3 \right]^2 \right]. \tag{C.26}$$

The proof of (C.26) is postponed to Section C.1.3. Then,

$$|L_{2,n,\gamma} - L_{2,n,\gamma}^\circ| \lesssim \begin{cases} \frac{1}{n^{4(1-\gamma)} M_n^2} \sum_{i_1 < i_2} \frac{i_1^2}{i_1^{2\gamma} i_2^{2\gamma}} \mathbb{E}\left[\mathbb{E}\left[\bar{\mathsf{K}}(X_1, X_2) \bar{\mathsf{K}}(X_1, X_3) \mid X_2, X_3 \right]^2 \right] & 0 \le \gamma < 1 \\ \frac{1}{(\log n)^2 M_n^2} \sum_{i_1 < i_2} \frac{1}{i_2^2} \mathbb{E}\left[\mathbb{E}\left[\bar{\mathsf{K}}(X_1, X_2) \bar{\mathsf{K}}(X_1, X_3) \mid X_2, X_3 \right]^2 \right] & \gamma = 1. \end{cases}$$

Applying Lemma 22 we can simplify the above bound as,

$$|L_{2,n,\gamma} - L_{2,n,\gamma}^\circ| \lesssim \begin{cases} \frac{\mathbb{E}\left[\mathbb{E}\left[\bar{\mathsf{K}}(X_1,X_2) \bar{\mathsf{K}}(X_1,X_3) | X_2, X_3 \right]^2 \right]}{M_n^2} & \text{if } 0 \le \gamma < 1 \\ \frac{\mathbb{E}\left[\mathbb{E}\left[\bar{\mathsf{K}}(X_1,X_2) \bar{\mathsf{K}}(X_1,X_3) | X_2, X_3 \right]^2 \right]}{M_n^2 \log n} & \text{if } \gamma = 1 \end{cases}$$
$$\lesssim \begin{cases} \frac{\mathbb{E}\left[\mathbb{E}\left[\bar{\mathsf{K}}(X_1,X_2) \bar{\mathsf{K}}(X_1,X_3) | X_2, X_3 \right]^2 \right]}{M_n^2} & \text{if } 0 \le \gamma < 1 \\ \frac{1}{\log n} & \text{if } \gamma = 1, \end{cases} \tag{C.27}$$

where the final bound follows from Cauchy-Schwartz inequality and recalling that $M_n = \mathbb{E}\left[\bar{\mathsf{K}}(X_1, X_2)^2 \right]$. The proof of Lemma 20 is now completed by recalling the decomposition from (C.19) and combining the individual bounds from Lemma 19, (C.22), (C.25) and (C.27).

C.1.3. PROOF OF (C.26):

Let

$$\boldsymbol{Z}_i^{(0)} = X_i \text{ and } \boldsymbol{Z}_i^{(1)} = Y_i \text{ for all } 1 \le i \le 4.$$

Recalling (C.5) we get

$$\mathsf{H}_n(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = \sum_{\varepsilon_i, \varepsilon_j \in \{0,1\}} (-1)^{\varepsilon_i + \varepsilon_j} \bar{\mathsf{K}}_n \left(\boldsymbol{Z}_i^{(\varepsilon_i)}, \boldsymbol{Z}_j^{(\varepsilon_j)} \right).$$

Then,

$$\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)\mathsf{H}_n(\boldsymbol{Z}_2, \boldsymbol{Z}_3)\mathsf{H}_n(\boldsymbol{Z}_3, \boldsymbol{Z}_4)\mathsf{H}_n(\boldsymbol{Z}_4, \boldsymbol{Z}_1)$$
$$= \sum_{\substack{\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4 \\ \varepsilon_1', \varepsilon_2', \varepsilon_3', \varepsilon_4' \in \{0,1\}}} (-1)^{\sum_{\ell=1}^4 \varepsilon_\ell + \varepsilon_\ell'} \bar{\mathsf{K}}_n \left(\boldsymbol{Z}_1^{(\varepsilon_1)}, \boldsymbol{Z}_2^{(\varepsilon_2)} \right) \bar{\mathsf{K}}_n \left(\boldsymbol{Z}_2^{(\varepsilon_2')}, \boldsymbol{Z}_3^{(\varepsilon_3)} \right) \bar{\mathsf{K}}_n \left(\boldsymbol{Z}_3^{(\varepsilon_3')}, \boldsymbol{Z}_4^{(\varepsilon_4)} \right) \bar{\mathsf{K}}_n \left(\boldsymbol{Z}_4^{(\varepsilon_4')}, \boldsymbol{Z}_1^{(\varepsilon_1')} \right).$$

Now recall that under null $\mathbb{E} \left[\bar{\mathsf{K}}_n \left(X_1, X_2 \right) \mid X_1 \right] = 0$. Hence it follows that

$$\mathbb{E} \left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)\mathsf{H}_n(\boldsymbol{Z}_2, \boldsymbol{Z}_3)\mathsf{H}_n(\boldsymbol{Z}_3, \boldsymbol{Z}_4)\mathsf{H}_n(\boldsymbol{Z}_4, \boldsymbol{Z}_1) \right]$$
$$= \sum_{\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4 \in \{0,1\}} \mathbb{E} \left[\bar{\mathsf{K}}_n \left(\boldsymbol{Z}_1^{(\varepsilon_1)}, \boldsymbol{Z}_2^{(\varepsilon_2)} \right) \bar{\mathsf{K}}_n \left(\boldsymbol{Z}_2^{(\varepsilon_2)}, \boldsymbol{Z}_3^{(\varepsilon_3)} \right) \bar{\mathsf{K}}_n \left(\boldsymbol{Z}_3^{(\varepsilon_3)}, \boldsymbol{Z}_4^{(\varepsilon_4)} \right) \bar{\mathsf{K}}_n \left(\boldsymbol{Z}_4^{(\varepsilon_4)}, \boldsymbol{Z}_1^{(\varepsilon_1)} \right) \right]$$
$$= 16\mathbb{E} \left[\bar{\mathsf{K}}_n(X_1, X_2)\bar{\mathsf{K}}_n(X_2, X_3)\bar{\mathsf{K}}_n(X_3, X_4)\bar{\mathsf{K}}_n(X_4, X_1) \right]$$
$$= 16\mathbb{E} \left[\mathbb{E} \left[\bar{\mathsf{K}}_n(X_1, X_2)\bar{\mathsf{K}}_n(X_1, X_3) \mid X_2, X_3 \right]^2 \right].$$

C.2. Proof of Theorem 16

For notational convenience define $s_n = \sum_{i=2}^n (i-1)/i^\gamma$ and

$$R_{n,\gamma} = \frac{1}{s_n} \sum_{i=2}^n \frac{1}{i^\gamma} \sum_{j=1}^{i-1} \mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_j), \ \hat{\sigma}_{n,\gamma}^2 = \frac{1}{s_n^2} \sum_{i=2}^n \left(\frac{1}{i^\gamma} \sum_{j=1}^{i-1} \mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_j) \right)^2.$$

Now we begin by noting that,

$$\mathbb{E} \left[1 - \phi_{n,\gamma} \right] = \mathbb{P} \left(R_{n,\gamma} \le z_\delta \hat{\sigma}_{n,\gamma} \right).$$

Fix $\varepsilon > 0$ and consider the event $\mathcal{E} = \left\{ \hat{\sigma}_{n,\gamma}^2 \le \mathbb{E} \left[\hat{\sigma}_{n,\gamma}^2 \right] / \varepsilon \right\}$. A simple application of Markov inequality shows that $\mathbb{P}(\mathcal{E}^c) \le \varepsilon$. Hence, by decomposing the event $\{R_{n,\gamma} \le z_\delta \hat{\sigma}_{n,\gamma}\}$ into $\mathcal{E}$ and $\mathcal{E}^c$ we get,

$$\mathbb{P} \left(R_{n,\gamma} \le z_\delta \hat{\sigma}_{n,\gamma} \right) \le \mathbb{P} \left(R_{n,\gamma} \le z_\delta \sqrt{\mathbb{E} \left[\hat{\sigma}_{n,\gamma}^2 \right] / \varepsilon} \right) + \varepsilon.$$

Now using the assumption (A.8) from we know that for large enough n,

$$z_\delta \sqrt{\mathbb{E} \left[\hat{\sigma}_{n,\gamma}^2 \right] / \varepsilon} \le \frac{\delta_n^2}{2}.$$

Furthermore by definition,

$$\mathbb{E}\left[R_{n,\gamma}\right] = \mathrm{MMD}^2[P_n, Q_n, \mathsf{K}_n] = \delta_n^2.$$

Combining these observations, for large enough n we now get

$$\mathbb{P}\left(R_{n,\gamma} \le z_\delta \sqrt{\mathbb{E}\left[\hat{\sigma}_{n,\gamma}^2\right]/\varepsilon}\right) \le \mathbb{P}\left(R_{n,\gamma} - \mathbb{E}\left[R_{n,\gamma}\right] \le \delta_n^2/2 - \delta_n^2\right) \le 4\frac{\mathrm{Var}\left(R_{n,\gamma}\right)}{\delta_n^4},$$

where the final inequality follows from Chebyshev's inequality. This implies that

$$\sup_{(P_n, Q_n) \in \mathcal{P}_n} \mathbb{P}\left(R_{n,\gamma} \le z_\delta \hat{\sigma}_{n,\gamma}\right) \le \sup_{(P_n, Q_n) \in \mathcal{P}_n} 4\frac{\mathrm{Var}\left(R_{n,\gamma}\right)}{\delta_n^4} + \varepsilon.$$

The proof is now completed by recalling the assumption (A.8) and then taking $n \to \infty$ and $\varepsilon \to 0$ respectively.

C.3. Proof of Theorem 17

We divide the proof in two parts. In Section C.3.1 we verify the type-I error control and in Section C.3.2 we verify the uniform consistency of the tests $\phi_{n,\gamma}$.

C.3.1. TYPE-I ERROR:

To verify the Type-I error control we verify the conditions of Corollary 15. Recall that here the scale parameter is given by $\nu_n = n^{\frac{4}{d+4\beta}}$. Now recalling the expression of moments of Gaussian kernel derived by Li and Yuan (2024) we know

$$\mathbb{E}_{P_n}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2\right] = \Omega(\nu_n^{-d/2}), \ \mathbb{E}_{P_n}\left[\bar{\mathsf{K}}_n(X_1, X_2)^4\right] \lesssim \nu_n^{-d/2}, \tag{C.28}$$

and

$$\mathbb{E}_{P_n}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2\bar{\mathsf{K}}_n(X_1, X_3)^2\right] \lesssim \nu_n^{-3d/4}. \tag{C.29}$$

By condition (9) in Li and Yuan (2024) one can immediately verify that Assumption A.1 is satisfied. Now we need to verify that Assumption A.2 also holds true. We divide this into two parts. First note that

$$t_{n,\gamma,1}\frac{\mathbb{E}_{P_n}\left[\bar{\mathsf{K}}_n(X_1, X_2)^4\right]}{\mathbb{E}_{P_n}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2\right]^2} \lesssim t_{n,\gamma,1}\nu_n^{d/2} = t_{n,\gamma,1}n^{\frac{2d}{d+4\beta}},$$

where the upper bound follows from (C.28). Now recalling the expression of $t_{n,\gamma,1}$ from (A.4) we get

$$t_{n,\gamma,1}n^{\frac{2d}{d+4\beta}} = o(1) \text{ for all } 0 \le \gamma < \frac{d+8\beta}{2(d+4\beta)}.$$

Now for the second condition in (A.4) note that

$$t_{n,\gamma,2}\frac{\mathbb{E}_{P_n}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2\bar{\mathsf{K}}_n(X_1, X_3)^2\right]}{\mathbb{E}_{P_n}\left[\bar{\mathsf{K}}_n(X_1, X_2)^2\right]^2} \lesssim t_{n,\gamma,2}\nu_n^{d/4} = t_{n,\gamma,2}n^{\frac{d}{d+4\beta}}.$$

where the upper bound now follows from (C.29). Now recalling the expression of $t_{n,\gamma,2}$ from (A.4) we get

$$t_{n,\gamma,2} n^{\frac{d}{d+4\beta}} = o(1) \text{ for all } 0 \leq \gamma < \frac{3d + 16\beta}{4(d + 4\beta)}.$$

Combining we find that Assumption A.2 is satisfied whenever $0 \leq \gamma < \frac{d+8\beta}{2(d+4\beta)}$. Hence, by applying Corollary 15 we conclude that $\phi_{n,\gamma}$ has asymptotic type-I error control for $0 \leq \gamma < \frac{d+8\beta}{2(d+4\beta)}$.

C.3.2. CONSISTENCY

To verify consistency we need to show that conditions of Theorem 16 are satisfied. Recalling (A.8) we need to show that,

$$\lim_{n\to\infty} \sup_{(P_n,Q_n)\in\mathcal{P}_n} \frac{\mathbb{E}\left[\frac{1}{s_{n,\gamma}^2} \sum_{i=2}^{n} \left(\frac{1}{i^\gamma} \sum_{j=1}^{i-1} \mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)\right)^2\right]}{\delta_n^4} + \frac{\operatorname{Var}\left(\frac{1}{s_{n,\gamma}} \sum_{i=2}^{n} \frac{1}{i^\gamma} \sum_{j=1}^{i-1} \mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)\right)}{\delta_n^4} = 0$$

where $s_{n,\gamma} = \sum_{i=2}^{n} \frac{i-1}{i^\gamma}$. We divide the proof into two parts to show separately,

$$\lim_{n\to\infty} \sup_{(P_n,Q_n)\in\mathcal{P}_n} \frac{\mathbb{E}\left[\frac{1}{s_{n,\gamma}^2} \sum_{i=2}^{n} \left(\frac{1}{i^\gamma} \sum_{j=1}^{i-1} \mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)\right)^2\right]}{\delta_n^4} = 0 \tag{C.30}$$

and,

$$\lim_{n\to\infty} \sup_{(P_n,Q_n)\in\mathcal{P}_n} \frac{\operatorname{Var}\left(\frac{1}{s_{n,\gamma}} \sum_{i=2}^{n} \frac{1}{i^\gamma} \sum_{j=1}^{i-1} \mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)\right)}{\delta_n^4} = 0. \tag{C.31}$$

We begin with the proof of (C.30).

Proof of (C.30): We begin by expanding the numerator from (C.30).

$$\mathbb{E}\left[\frac{1}{s_{n,\gamma}^2} \sum_{i=2}^{n} \left(\frac{1}{i^\gamma} \sum_{j=1}^{i-1} \mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)\right)^2\right]$$

$$= \frac{1}{s_{n,\gamma}^2} \sum_{i=2}^{n} \frac{1}{i^{2\gamma}} \sum_{j=1}^{i-1} \mathbb{E}\left[\mathsf{H}_n^2(\boldsymbol{Z}_i, \boldsymbol{Z}_j)\right] + \frac{1}{s_{n,\gamma}^2} \sum_{i=2}^{n} \frac{1}{i^{2\gamma}} \sum_{j_1 \neq j_2} \mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_i, \boldsymbol{Z}_{j_1}) \mathsf{H}_n(\boldsymbol{Z}_i, \boldsymbol{Z}_{j_2})\right]$$

$$= \frac{1}{s_{n,\gamma}^2} \sum_{i=2}^{n} \frac{i-1}{i^{2\gamma}} \mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2\right] + \frac{1}{s_{n,\gamma}^2} \sum_{i=2}^{n} \frac{(i-1)(i-2)}{i^{2\gamma}} \mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2) \mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_3)\right]. \tag{C.32}$$

Now by definition,

$$\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2\right] \lesssim \mathbb{E}\left[\mathsf{K}_n(X_1, X_2)^2\right] + 2\mathbb{E}\left[\mathsf{K}_n(X_1, Y_2)^2\right] + \mathbb{E}\left[\mathsf{K}_n(Y_1, Y_2)^2\right].$$

Recall that K_n is a Gaussian kernel and $\|p\|, \|q\| \leq M$. Hence applying Cauchy-Schwarz inequality on Lemma 25 we get,

$$\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2\right] \lesssim \nu_n^{-d/2}.$$

35

Furthermore,

$$
\begin{aligned}
\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_3)\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2) \mid \boldsymbol{Z}_1\right]^2\right] \\
&\lesssim \mathbb{E}\left[\left(\mathbb{E}\left[\mathsf{K}_n(X_1, X_2) \mid X_1\right] - \mathbb{E}\left[\mathsf{K}_n(X_1, Y_2) \mid X_1\right]\right)^2\right] \\
&\qquad + \mathbb{E}\left[\left(\mathbb{E}\left[\mathsf{K}_n(X_2, Y_1) \mid Y_1\right] - \mathbb{E}\left[\mathsf{K}_n(Y_1, Y_2) \mid Y_1\right]\right)^2\right] \\
&= \mathbb{E}\left[f(X)^2\right] + \mathbb{E}\left[f(Y)^2\right],
\end{aligned}
$$

where $f(\cdot) = \mathbb{E}\left[\mathsf{K}_n(X, \cdot)\right] - \mathbb{E}\left[\mathsf{K}_n(Y, \cdot)\right]$ with $X \sim p$ and $Y \sim q$. Then quoting Lemma 20 from Shekhar et al. (2022) we get

$$
\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_3)\right] \lesssim \|p - q\|_{L_2}^2 \, \nu_n^{-3d/4}.
$$

Now from Lemma 15 in Li and Yuan (2024) there exists $C > 0$ such that for $g = p - q \in \mathcal{W}^{\beta,2}(M)$,

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

given that $\nu_n \geq C\|g\|_{L_2}^{-2/\beta}$. Now note that by Lemma 25 and that $\delta_n = \mathrm{MMD}[P_n, Q_n, \mathsf{K}_n]$,

$$
\int \exp\left(-\frac{\|\omega\|^2}{4\nu_n}\right) |\mathcal{F}g(\omega)|^2 \, \mathrm{d}\omega = \left(\frac{\nu_n}{\pi}\right)^{d/2} \delta_n^2,
$$

implying $\|p - q\|_{L_2}^2 \lesssim \nu_n^{d/2}\delta_n^2$. Then,

$$
\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)\mathsf{H}_n(\boldsymbol{Z}_1, \boldsymbol{Z}_3)\right] \lesssim \|p - q\|_{L_2}^2 \, \nu_n^{-3d/4} \lesssim \nu_n^{-d/4}\delta_n^2. \tag{C.33}
$$

Substituting the bounds back in (C.32)

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{s_{n,\gamma}^2}\sum_{i=2}^n\left(\frac{1}{i^\gamma}\sum_{j=1}^{i-1}\mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)\right)^2\right] &\lesssim \frac{1}{s_{n,\gamma}^2}\sum_{i=2}^n\frac{i-1}{i^{2\gamma}}\nu_n^{-d/2} + \frac{1}{s_{n,\gamma}^2}\sum_{i=2}^n\frac{(i-1)(i-2)}{i^{2\gamma}}\nu_n^{-d/4}\delta_n^2 \\
&\lesssim \frac{1}{n^2}\nu_n^{-d/2} + \frac{1}{n}\nu_n^{-d/4}\delta_n^2, \tag{C.34}
\end{aligned}
$$

where the final bound follows from Lemma 22. Then,

$$
\frac{\mathbb{E}\left[\frac{1}{s_{n,\gamma}^2}\sum_{i=2}^n\left(\frac{1}{i^\gamma}\sum_{j=1}^{i-1}\mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)\right)^2\right]}{\delta_n^4} \lesssim \frac{\nu_n^{-d/2}}{n^2\delta_n^4} + \frac{\nu_n^{-d/4}}{n\delta_n^2} \lesssim \frac{1}{\left(n^{2\beta/(d+4\beta)}\Delta_n\right)^4} + \frac{1}{\left(n^{2\beta/(d+4\beta)}\Delta_n\right)^2}, \tag{C.35}
$$

where the final bound follows by recalling that in $\mathcal{P}_n^{(1)}$, $\Delta_n \leq \|p - q\|_2 \leq \nu_n^{d/4}\delta_n$. The proof is now completed by recalling that $n^{2\beta/(d+4\beta)}\Delta_n \to \infty$.

Proof of (C.31): Once again we begin by expanding the numerator from (C.31).

$$
\text{Var}\left(\frac{1}{s_{n,\gamma}}\sum_{i=2}^{n}\frac{1}{i^\gamma}\sum_{j=1}^{i-1}\mathsf{H}(\boldsymbol{Z}_i,\boldsymbol{Z}_j)\right)
$$

$$
=\mathbb{E}\left[\frac{1}{s_{n,\gamma}^2}\sum_{i=2}^{n}\left(\frac{1}{i^\gamma}\sum_{j=1}^{i-1}\mathsf{H}_n(\boldsymbol{Z}_i,\boldsymbol{Z}_j)\right)^2\right]+2\mathbb{E}\left[\frac{1}{s_{n,\gamma}^2}\sum_{i<j}\frac{1}{i^\gamma j^\gamma}\sum_{a=1}^{i-1}\sum_{b=1}^{j-1}\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_i,\boldsymbol{Z}_a)\mathsf{H}_n(\boldsymbol{Z}_j,\boldsymbol{Z}_b)\right]\right]-\delta_n^4.
$$

Notice that we already know the bound on the first term from (C.34). In the following we analyse the second term only.

$$
\mathbb{E}\left[\frac{1}{s_{n,\gamma}^2}\sum_{i<j}\frac{1}{i^\gamma j^\gamma}\sum_{a=1}^{i-1}\sum_{b=1}^{j-1}\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_i,\boldsymbol{Z}_a)\mathsf{H}_n(\boldsymbol{Z}_j,\boldsymbol{Z}_b)\right]\right]
$$

$$
=\frac{1}{s_{n,\gamma}^2}\sum_{i<j}\frac{1}{i^\gamma j^\gamma}\left[\sum_{a=1}^{i-1}\sum_{b=1}^{j-1}\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_i,\boldsymbol{Z}_a)\mathsf{H}_n(\boldsymbol{Z}_j,\boldsymbol{Z}_b)\right]\mathbf{1}\left\{a\neq b\right\}+\sum_{a=1}^{i-1}\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_i,\boldsymbol{Z}_a)\mathsf{H}_n(\boldsymbol{Z}_j,\boldsymbol{Z}_a)\right]\right]
$$

$$
=\frac{1}{s_{n,\gamma}^2}\sum_{i<j}\frac{(i-1)(j-2)}{i^\gamma j^\gamma}\delta_n^4+\frac{1}{s_{n,\gamma}^2}\sum_{i<j}\frac{i-1}{i^\gamma j^\gamma}\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1,\boldsymbol{Z}_2)\mathsf{H}_n(\boldsymbol{Z}_1,\boldsymbol{Z}_3)\right].
$$

Now by Lemma 24 we know that,

$$
\lim_{n\to\infty}\sup_{(P_n,Q_n)\in\mathcal{P}_n^{(1)}}\frac{\text{Var}\left(\frac{1}{s_{n,\gamma}}\sum_{i=2}^{n}\frac{1}{i^\gamma}\sum_{j=1}^{i-1}\mathsf{H}(\boldsymbol{Z}_i,\boldsymbol{Z}_j)\right)}{\delta_n^4}
$$

$$
=\lim_{n\to\infty}\sup_{(P_n,Q_n)\in\mathcal{P}_n^{(1)}}\frac{\mathbb{E}\left[\frac{1}{s_{n,\gamma}^2}\sum_{i=2}^{n}\left(\frac{1}{i^\gamma}\sum_{j=1}^{i-1}\mathsf{H}(\boldsymbol{Z}_i,\boldsymbol{Z}_j)\right)^2\right]+\frac{2}{s_{n,\gamma}^2}\sum_{i<j}\frac{i-1}{i^\gamma j^\gamma}\mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_1,\boldsymbol{Z}_2)\mathsf{H}_n(\boldsymbol{Z}_1,\boldsymbol{Z}_3)\right]}{\delta_n^4}.
$$

$$(\text{C.36})$$

Now to complete the proof note that by Lemma 22,

$$
\frac{2}{s_{n,\gamma}^2}\sum_{i<j}\frac{i-1}{i^\gamma j^\gamma}\lesssim\frac{1}{n}.
$$

Hence continuing the trail of equalities from (C.36) and recalling bounds from (C.35) and (C.33) we get,

$$
\lim_{n\to\infty}\sup_{(P_n,Q_n)\in\mathcal{P}_n^{(1)}}\frac{\text{Var}\left(\frac{1}{s_{n,\gamma}}\sum_{i=2}^{n}\frac{1}{i^\gamma}\sum_{j=1}^{i-1}\mathsf{H}(\boldsymbol{Z}_i,\boldsymbol{Z}_j)\right)}{\delta_n^4}
$$

$$
\lesssim\frac{1}{\left(n^{2\beta/(d+4\beta)}\Delta_n\right)^4}+\frac{1}{\left(n^{2\beta/(d+4\beta)}\Delta_n\right)^2}+\frac{\nu_n^{-d/4}}{n\delta_n^2}\to0,
$$

which completes the proof.

Appendix D. Proof of Theorem 7 and Theorem 9

D.1. Proof of Theorem 7

To prove Theorem 7 we need only verify the conditions of Theorem 8 in this setting. Since $P,Q\in\mathcal{M}_K^1$ by a simple application of Cauchy-Schwartz ineqaulity all expectations in the following exists

and are finite. Note that in this setting $\delta_n = \delta = \mathrm{MMD}[P, Q, \mathsf{K}] > 0$. Then following computations similar to (C.32) we get,

$$\mathbb{E}\left[\sigma_n^2\right] \lesssim \frac{1}{n^2} \sum_{i=1}^{n} \frac{i-1}{i^2} + \frac{1}{n^2} \sum_{i=2}^{n} \frac{(i-1)(i-2)}{i^2} = O(1/n) \to 0. \tag{D.1}$$

Moreover, similar to (C.36) we get the following decompositions.

$$\mathrm{Var}(T_n) = \mathbb{E}\left[\frac{1}{n^2} \sum_{i=2}^{n} \left(\frac{1}{i} \sum_{j=1}^{i-1} \mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)\right)^2\right] + 2\mathbb{E}\left[\frac{1}{n^2} \sum_{i<j} \frac{1}{ij} \sum_{a=1}^{i-1}\sum_{b=1}^{j-1} \mathbb{E}\left[\mathsf{H}_n(\boldsymbol{Z}_i, \boldsymbol{Z}_a)\mathsf{H}_n(\boldsymbol{Z}_j, \boldsymbol{Z}_b)\right]\right] - \delta^4.$$

Furthermore,

$$\mathbb{E}\left[\frac{1}{n^2} \sum_{i<j} \frac{1}{ij} \sum_{a=1}^{i-1}\sum_{b=1}^{j-1} \mathbb{E}\left[\mathsf{H}(\boldsymbol{Z}_i, \boldsymbol{Z}_a)\mathsf{H}(\boldsymbol{Z}_j, \boldsymbol{Z}_b)\right]\right]$$
$$= \frac{1}{n^2} \sum_{i<j} \frac{(i-1)(j-2)}{ij}\delta^4 + \frac{1}{n^2} \sum_{i<j} \frac{i-1}{ij}\mathbb{E}\left[\mathsf{H}(\boldsymbol{Z}_1, \boldsymbol{Z}_2)\mathsf{H}(\boldsymbol{Z}_1, \boldsymbol{Z}_3)\right]$$
$$= \delta^4/2(1 + o(1)).$$

Combined with (D.1) we can now easily conclude that $\mathrm{Var}(T_n) \to 0$. This completes checking the conditions of Theorem 8 and subsequenlty proves Theorem 7.

D.2. Proof of Theorem 9

The statistic T_n from (2.5) can be rewritten as $T_n = \frac{1}{2n} \sum_{i \neq j} w_{ij} \mathsf{H}_n(\boldsymbol{Z}_i, \boldsymbol{Z}_j)$ where H_n is defined in (5.1) $w_{ij} = 1/\max\{i, j\}, 1 \leq i \neq j \leq n$ and $\boldsymbol{Z}_i = (X_i, Y_i), 1 \leq i \leq n$. For notational simplicity take $\mu_n = \mathrm{MMD}[P_n, Q_n, \mathsf{K}_n]^2$ and write,

$$\mathsf{H}_n(\boldsymbol{z}_1, \boldsymbol{z}_2) = \mu_n + h_n(\boldsymbol{z}_1) + h_n(\boldsymbol{z}_2) + g_n(\boldsymbol{z}_1, \boldsymbol{z}_2),$$

where $g_n(\boldsymbol{z}_1, \boldsymbol{z}_2) = \mathsf{H}_n(\boldsymbol{z}_1, \boldsymbol{z}_2) - h_n(\boldsymbol{z}_1) - h_n(\boldsymbol{z}_2) + \mu_n$. Substituing back this expression in T_n we get,

$$\frac{T_n}{\mu_n} = 1 + \sum_{i=1}^{n} \underbrace{w_i^{(n)} h_n(\boldsymbol{Z}_i)}_{W_{i,n}} + \frac{1}{2n\mu_n} \sum_{i \neq j} w_{ij} g(\boldsymbol{Z}_i, \boldsymbol{Z}_j) + O(\log n/n),$$

where $w_i^{(n)} = \frac{i-1}{n\mu_n i} + \frac{1}{n\mu_n} \sum_{j=i+1}^{n} \frac{1}{j}$. Define $L_n = \sum_{i=1}^{n} W_{i,n}$. Moreover we can show,

$$S_n := n\mu_n^2 \sum_{i=1}^{n} \left(w_i^{(n)}\right)^2 \to 5. \tag{D.2}$$

We postpone the proof of (D.2) to Section D.2.1. Now note that $w_i^{(n)} \lesssim H_n/n\mu_n$ for all $1 \leq i \leq n$ and hence $\sum_{i=1}^{n} \left(w_i^{(n)}\right)^3 \lesssim \frac{(\log n)^3}{n^3 \mu_n^3}$. Now to prove CLT, we will invoke Lyapunov's version on L_n.

To that end note that $\mathbb{E}\left[h_n(\boldsymbol{Z}_1)\right] = 0$ which implies, $\mathbb{E}\left[W_{n,i}\right] = 0$ for all $1 \leq i \leq n$ and,

$$\frac{\sum_{i=1}^n \mathbb{E}\left[|W_{n,i}|^3\right]}{\left(\sum_{i=1}^n \mathbb{E}\left[W_{n,i}^2\right]\right)^{3/2}} = \frac{\mathbb{E}\left[|h_n(\boldsymbol{Z}_1)|^3\right]}{\mathbb{E}\left[h_n(\boldsymbol{Z}_1)^2\right]^{3/2}} \frac{\sum_{i=1}^n \left(w_i^{(n)}\right)^3}{\left[\sum_{i=1}^n \left(w_i^{(n)}\right)^2\right]^{3/2}} = O\left(\frac{(\log n)^3 \mathbb{E}\left[|h_n(\boldsymbol{Z}_1)|^3\right]}{n^{3/2}\mathrm{Var}\left(h_n(\boldsymbol{Z}_1)\right)^{3/2}}\right).$$

Hence by Lyapunov CLT we have $\frac{L_n}{\sqrt{\mathrm{Var}(L_n)}} \xrightarrow{D} N(0,1)$. Now recalling the definition of g_n note that $\mathbb{E}\left[g_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)\right] = 0$ and

$$\mathrm{Var}\left(\frac{1}{2n\mu_n}\sum_{i \neq j} w_{ij} g_n(\boldsymbol{Z}_i, \boldsymbol{Z}_j)\right) \lesssim \frac{\mathbb{E}\left[g_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2\right]}{\mu_n^2}\frac{H_n}{n^2}.$$

Hence we get the bound,

$$\mathbb{E}\left[\frac{\left(\frac{1}{2n\mu_n}\sum_{i \neq j} w_{ij} g_n(\boldsymbol{Z}_i, \boldsymbol{Z}_j)\right)^2}{\mathrm{Var}(L_n)}\right] \lesssim \frac{\mathbb{E}\left[g_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2\right]}{\mu_n^2}\frac{H_n}{n^2}\frac{n\mu_n^2}{\mathbb{E}\left[h_n(\boldsymbol{Z}_1)^2\right]} = O\left(\frac{\log n \mathbb{E}\left[g_n(\boldsymbol{Z}_1, \boldsymbol{Z}_2)^2\right]}{n\mathrm{Var}\left(h_n(\boldsymbol{Z}_1)\right)}\right).$$

Moreover,

$$\frac{\log n}{n\sqrt{\mathrm{Var}(L_n)}} \lesssim \frac{\mu_n \log n}{\sqrt{n\mathrm{Var}\left(h_n(\boldsymbol{Z}_1)^2\right)}}.$$

Combining all the above and recalling the assumptions of Theorem 9 we conclude,

$$\frac{1}{\sqrt{\mathrm{Var}(L_n)}}\left(\frac{T_n}{\mu_n} - 1\right) \xrightarrow{D} N(0,1).$$

Moreover using (D.2) we know $n\mu_n^2 \frac{\mathrm{Var}(L_n)}{\mathrm{Var}(h_n(\boldsymbol{Z}_1)^2)} \to 5$. Hence,

$$\sqrt{\frac{n\mu_n^2}{\mathrm{Var}\left(h_n(\boldsymbol{Z}_1)\right)}}\left(\frac{T_n}{\mu_n} - 1\right) \xrightarrow{D} N(0,5) \implies \sqrt{\frac{n}{\mathrm{Var}\left(h_n(\boldsymbol{Z}_1)\right)}}\left(T_n - \mu_n\right) \xrightarrow{D} N(0,5).$$

D.2.1. PROOF OF (D.2):

By a simple computation,

$$S_n = n\mu_n^2 \sum_{i=1}^n \left(w_i^{(n)}\right)^2 = \frac{1}{n}\left[\sum_{i=1}^n \left(\frac{i-1}{i}\right)^2 + 2\sum_{i=1}^n \frac{i-1}{i}\sum_{j=i+1}^n \frac{1}{j} + \sum_{i=1}^n \left(\sum_{j=i+1}^n \frac{1}{j}\right)^2\right].$$

For the last term note that

$$\frac{1}{n}\sum_{i=1}^n \left(\sum_{j=i+1}^n \frac{1}{j}\right)^2 = \frac{1}{n}\sum_{j_1,j_2} \frac{\min\{j_1, j_2\} - 1}{j_1 j_2} = \frac{1}{n}\sum_{j_1,j_2} \frac{\min\{j_1, j_2\}}{j_1 j_2} - \frac{H_n^2}{n},$$

where H_n is the harmonic sum till n. Now note that,

$$\frac{1}{n} \sum_{j_1, j_2} \frac{\min\{j_1, j_2\}}{j_1 j_2} = \frac{2}{n} \sum_{j_1 \leq j_2} \frac{1}{j_2} - \frac{1}{n} \sum_j \frac{1}{j} = 2 - \frac{H_n}{n}.$$

Hence,

$$\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=i+1}^{n} \frac{1}{j} \right)^2 = 2 - \frac{H_n}{n} + \frac{H_n^2}{n} \to 2.$$

For the first term we have,

$$\frac{1}{n} \sum_{i=1}^{n} \left(\frac{i-1}{i} \right)^2 = 1 - 2\frac{H_n}{n} + \frac{1}{n} \sum_{i=1}^{n} \frac{1}{i^2} \to 1.$$

Finally for the second term,

$$\frac{1}{n} \sum_{i=1}^{n} \frac{i-1}{i} \sum_{j=i+1}^{n} \frac{1}{j} = \frac{1}{n} \sum_{j=2}^{n} \frac{1}{j} \sum_{i=1}^{j-1} \frac{i-1}{i} = \frac{1}{n} \sum_{j=2}^{n} \frac{j-1-H_{j-1}}{j} = 1 - \frac{H_n}{n} - \frac{1}{n} \sum_{j=2}^{n} \frac{H_{j-1}}{j} \to 1.$$

Combining we can conclude that $S_n \to 5$.

Appendix E. Technical Results

Lemma 21 *Let K be a kernel satisfying Assumption 2.1 and suppose that $X \sim P$ is not almost surely constant. Then, for independent samples $X_1, X_2 \overset{i.i.d.}{\sim} P$ with $P \in \mathcal{M}_{\mathsf{K}}^1(\mathcal{X})$,*

$$\mathbb{E}\left[\bar{\mathsf{K}}(X_1, X_2)^2 \right] > 0.$$

Proof Let $\mathcal{K} \otimes \mathcal{K}$ be the tensor product space (see Berlinet and Thomas-Agnan (2011, Section 4.6)). Let $h(X_i) = \mathsf{K}(X_i, \cdot) - \nu_P, i = 1, 2$. With this notation we have,

$$\mathbb{E}\left[\bar{\mathsf{K}}(X_1, X_2)^2 \right] = \mathbb{E}\left[\langle h(X_1) \otimes h(X_1), h(X_2) \otimes h(X_2) \rangle_{\mathcal{K} \otimes \mathcal{K}} \right] = \| \mathbb{E}\left[h(X) \otimes h(X) \right] \|_{\mathcal{K} \otimes \mathcal{K}}^2,$$

where $X \sim P$. Now for the sake of contradiction assume that $\mathbb{E}\left[\bar{\mathsf{K}}(X_1, X_2)^2 \right] = 0$. Then for any $r \in \mathcal{K}$,

$$\mathbb{E}\left[\langle r, h(X) \rangle_{\mathcal{K}}^2 \right] = \mathbb{E}\left[\langle r \otimes r, h(X) \otimes h(X) \rangle \right] = 0.$$

This implies that $\mathsf{K}(X_1, \cdot) = \mathsf{K}(X_2, \cdot)$ almost surely. Now recalling the characteristic property of K and considering the Dirac measures $\delta_x, x \in \mathcal{X}$ we can immediately conclude that $x \to \mathsf{K}(x, \cdot)$ is an injective function. Hence $\mathsf{K}(X_1, \cdot) = \mathsf{K}(X_2, \cdot)$ almost surely implies that $X_1 = X_2$ almost surely, which leads to a contradiction since X_1, X_2 are generated independently from P. ∎

Lemma 22 *For any p,*

$$\sum_{i=1}^{n} i^p = \begin{cases} \Theta\left(n^{p+1}\right) & \text{if } p > -1 \\ \Theta\left(\log n\right) & \text{if } p = -1 \\ \Theta(1) & \text{if } p < -1 \end{cases}.$$

Proof We use the method of integral bounds by considering the function $f(x) = x^p$.

Case 1: $p > -1$ Let $f(x) = x^p$. If $p > 0$, $f(x)$ is increasing, so

$$\int_1^n x^p \, dx \leq \sum_{i=1}^n i^p \leq \int_1^{n+1} x^p \, dx,$$

which implies

$$\frac{n^{p+1} - 1}{p + 1} \leq \sum_{i=1}^n i^p \leq \frac{(n+1)^{p+1} - 1}{p + 1}.$$

If $-1 < p < 0$, $f(x)$ is decreasing, so

$$\int_1^{n+1} x^p \, dx \leq \sum_{i=1}^n i^p \leq 1^p + \int_1^n x^p \, dx,$$

which implies

$$\frac{(n+1)^{p+1} - 1}{p + 1} \leq \sum_{i=1}^n i^p \leq 1 + \frac{n^{p+1} - 1}{p + 1}.$$

In both subcases, the lower and upper bounds are $\Theta(n^{p+1})$. Thus, $\sum_{i=1}^n i^p = \Theta(n^{p+1})$.

Case 2: $p = -1$ The sum is the n-th harmonic number, $H_n = \sum_{i=1}^n \frac{1}{i}$. The function $f(x) = 1/x$ is decreasing for $x > 0$, so we apply the integral bounds:

$$\int_1^{n+1} \frac{1}{x} \, dx \leq H_n \leq 1 + \int_1^n \frac{1}{x} \, dx \, ,$$

$$\ln(n + 1) \leq H_n \leq 1 + \ln n.$$

Since $\ln(n + 1)$ and $1 + \ln n$ are both $\Theta(\ln n)$, it follows that $H_n = \Theta(\ln n)$.

Case 3: $p < -1$ The sum's terms are positive, so it is bounded below by its first term:

$$\sum_{i=1}^n i^p \geq 1^p = 1 \implies \sum_{i=1}^n i^p = \Omega(1).$$

Since $f(x) = x^p$ is decreasing, the sum is bounded above by the convergent integral:

$$\sum_{i=1}^n i^p \leq 1 + \int_1^n x^p \, dx < 1 + \int_1^\infty x^p \, dx.$$

We evaluate the integral, noting $p + 1 < 0$:

$$\int_1^\infty x^p \, dx = \left[\frac{x^{p+1}}{p+1} \right]_1^\infty = \lim_{b \to \infty} \frac{b^{p+1}}{p+1} - \frac{1}{p+1} = 0 - \frac{1}{p+1} = \frac{-1}{p+1}.$$

The sum is bounded above by the constant $1 - \frac{1}{p+1}$, so $\sum_{i=1}^n i^p = O(1)$. Combining the bounds gives $\sum_{i=1}^n i^p = \Theta(1)$. ■

Lemma 23 *For every $n \geq 2$,*

$$\left| \frac{1}{n^{2(1-\gamma)}} \sum_{i=2}^{n} \frac{i-1}{i^{2\gamma}} - \frac{1}{2(1-\gamma)} \right| \lesssim \begin{cases} \frac{1}{n} & \text{if } 0 \leq \gamma < 1/2 \\ \frac{\log n}{n} & \text{if } \gamma = 1/2 \\ \frac{1}{n^{2(1-\gamma)}} & \text{if } 1/2 < \gamma < 1 \end{cases} .$$

Proof Notice that

$$\frac{1}{n^{2(1-\gamma)}} \sum_{i=2}^{n} \frac{i-1}{i^{2\gamma}} = \frac{1}{n^{2(1-\gamma)}} \sum_{i=1}^{n} \frac{1}{i^{2\gamma-1}} - \frac{1}{n^{2(1-\gamma)}} \sum_{i=1}^{n} \frac{1}{i^{2\gamma}}. \tag{E.1}$$

Invoking Theorem 2 from [Apostol](1999) with $f(x) = x^{1-2\gamma}$ we get

$$\left| \sum_{i=1}^{n} \frac{1}{i^{2\gamma-1}} - \int_{1}^{n} x^{1-2\gamma} \mathrm{d}x \right| \lesssim \int_{1}^{n} (1-2\gamma) x^{-2\gamma} \mathrm{d}x + 1 \lesssim \frac{1}{n^{2\gamma-1}} + 1.$$

Evaluating the integral gives

$$\left| \sum_{i=1}^{n} \frac{1}{i^{2\gamma-1}} - \frac{n^{2(1-\gamma)}}{2(1-\gamma)} \right| \lesssim \frac{1}{n^{2\gamma-1}} + 1 + \frac{1}{2(1-\gamma)}.$$

Hence,

$$\left| \frac{1}{n^{2(1-\gamma)}} \sum_{i=1}^{n} \frac{1}{i^{2\gamma-1}} - \frac{1}{2(1-\gamma)} \right| \lesssim \frac{1}{n} + \frac{1}{n^{2(1-\gamma)}}. \tag{E.2}$$

Also, by Lemma 22 we know,

$$\frac{1}{n^{2(1-\gamma)}} \sum_{i=1}^{n} \frac{1}{i^{2\gamma}} = \begin{cases} \Theta\left(1/n\right) & \text{if } 0 \leq \gamma < 1/2 \\ \Theta\left(\log n/n\right) & \text{if } \gamma = 1/2 \\ \Theta\left(1/n^{2(1-\gamma)}\right) & \text{if } 1/2 < \gamma < 1. \end{cases} \tag{E.3}$$

Now combining (E.1), (E.2) and (E.3) we get,

$$\left| \frac{1}{n^{2(1-\gamma)}} \sum_{i=2}^{n} \frac{i-1}{i^{2\gamma}} - \frac{1}{2(1-\gamma)} \right|$$

$$\lesssim \left| \frac{1}{n^{2(1-\gamma)}} \sum_{i=1}^{n} \frac{1}{i^{2\gamma-1}} - \frac{1}{2(1-\gamma)} \right| + \frac{1}{n^{2(1-\gamma)}} \sum_{i=1}^{n} \frac{1}{i^{2\gamma}} \lesssim \begin{cases} \frac{1}{n} & \text{if } 0 \leq \gamma < 1/2 \\ \frac{\log n}{n} & \text{if } \gamma = 1/2 \\ \frac{1}{n^{2(1-\gamma)}} & \text{if } 1/2 < \gamma < 1 \end{cases}$$

which completes the proof. ∎

Lemma 24 *Let $s_{n,\gamma} = \sum_{i=2}^{n} \frac{i-1}{i^{\gamma}}$. Then for $\gamma < 1$,*

$$\lim_{n \to \infty} \frac{1}{s_{n,\gamma}^2} \sum_{i<j} \frac{(i-1)(j-2)}{i^{\gamma} j^{\gamma}} = \frac{1}{2}.$$

Proof We will analyze the denominator $s_{n,\gamma}$ and the numerator $\sum_{i<j} \frac{(i-1)(j-2)}{i^\gamma j^\gamma}$ separately. We begin with the denominator given by $s_{n,\gamma} = \sum_{i=2}^n \frac{i-1}{i^\gamma}$. We can split this sum:

$$s_{n,\gamma} = \sum_{i=2}^n \left(\frac{i}{i^\gamma} - \frac{1}{i^\gamma} \right) = \sum_{i=2}^n i^{1-\gamma} - \sum_{i=2}^n i^{-\gamma}.$$

For $\gamma < 1$ we consider the two terms separately. For the first term, once again applying the integral approximation from Theorem 2 of Apostol (1999) we can easily show,

$$\frac{1}{n^{2-\gamma}} \left| \sum_{i=2}^n i^{1-\gamma} - \int_2^n x^{1-\gamma} \mathrm{d}x \right| = o(1).$$

Evaluating the inner integral we then conclude

$$\frac{1}{n^{2-\gamma}} \sum_{i=2}^n i^{1-\gamma} - \frac{1}{2-\gamma} = o(1). \tag{E.4}$$

By a similar argument we can show $\frac{1}{n^{2-\gamma}} \sum_{i=2}^n i^{-\gamma} = o(1)$ and hence $s_{n,\gamma}/n^{2-\gamma} \to 1/(2-\gamma)$.

To find the limit of the numerator, note that we can rewrite the numerator as

$$N := \sum_{i<j} \frac{(i-1)(j-2)}{i^\gamma j^\gamma} = \frac{1}{2} \left[\left(\sum_{i=1}^n \frac{i-1}{i^\gamma} \right)^2 - \sum_{i=1}^n \left(\frac{i-1}{i} \right)^2 \right] - \sum_{i<j} \frac{i-1}{i^\gamma j^\gamma}.$$

Then applying Lemma 22 a simple computation yields

$$\frac{N}{n^{4-2\gamma}} = \frac{1}{2} \left(\frac{1}{n^{2-\gamma}} \sum_{i=1}^n \frac{i-1}{i^\gamma} \right)^2 + o(1) = \frac{1}{2} \left(\frac{1}{n^{2-\gamma}} \sum_{i=1}^n i^{1-\gamma} \right)^2 + o(1).$$

Recalling the limit from (E.4) we can conclude that $N/n^{2-4\gamma} \to 1/2(2-\gamma)^2$. The proof is now completed by recalling the limit of $s_{n,\gamma}$. ∎

Lemma 25 *Consider $f, g \in L_2(\mathbb{R}^d)$. Then*

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

where $G_\nu(x,y) = \exp(-\nu \|x - y\|_2^2)$ is the Gaussian kernel with bandwidth ν, $\mathcal{F}f(\cdot)$ and $\mathcal{F}g(\cdot)$ are the corresponding Fourier transforms and $\overline{\mathcal{F}f(\cdot)}$ is the complex conjugate.

Proof Consider Z to be a Gaussian random vector with mean $\mathbf{0}_d$ and covariance matrix $2\nu \mathbf{I}_d$. Then,

$$\int G_\nu(x,y)f(x)g(y)\mathrm{d}x\mathrm{d}y = \int \exp\left(-\nu\left\|x-y\right\|^2\right)f(x)g(y)\mathrm{d}x\mathrm{d}y$$

$$= \int \mathbb{E}\exp\left[\iota Z^\top(x-y)\right]f(x)g(y)\mathrm{d}x\mathrm{d}y$$

$$= \mathbb{E}\int \exp(\iota Z^\top x)f(x)\mathrm{d}x\int \exp(-\iota Z^\top y)g(y)\mathrm{d}y$$

$$= \int \frac{1}{(4\pi\nu)^{d/2}}\exp\left(-\frac{\left\|\omega\right\|^2}{4\nu}\right)\int \exp(\iota\omega^\top x)f(x)\mathrm{d}x\int \exp(-\iota\omega^\top y)g(y)\mathrm{d}y\mathrm{d}\omega$$

$$= \left(\frac{\pi}{\nu}\right)^{d/2}\int \exp\left(-\frac{\left\|\omega\right\|^2}{4\nu}\right)\overline{\mathcal{F}f(\omega)}\mathcal{F}g(\omega)\mathrm{d}\omega.$$

∎