A permutation-free kernel two-sample test

Shubhanshu Shekhar
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15213
shubhan2@andrew.cmu.edu

Ilmun Kim
Department of Statistics and Data Science
Department of Applied Statistics
Yonsei University
ilmun@yonsei.ac.kr

Aaditya Ramdas
Department of Statistics and Data Science
Machine Learning Department
Carnegie Mellon University
aramdas@stat.cmu.edu

Abstract

The kernel Maximum Mean Discrepancy (MMD) is a popular multivariate distance metric between distributions that has found utility in two-sample testing. The usual kernel-MMD test statistic is a degenerate U-statistic under the null, and thus it has an intractable limiting distribution. Hence, to design a level-α test, one usually selects the rejection threshold as the $(1-\alpha)$-quantile of the permutation distribution. The resulting nonparametric test has finite-sample validity but suffers from large computational cost, since every permutation takes quadratic time. We propose the cross-MMD, a new quadratic-time MMD test statistic based on sample-splitting and studentization. We prove that under mild assumptions, the cross-MMD has a limiting standard Gaussian distribution under the null. Importantly, we also show that the resulting test is consistent against any fixed alternative, and when using the Gaussian kernel, it has minimax rate-optimal power against local alternatives. For large sample sizes, our new cross-MMD provides a significant speedup over the MMD, for only a slight loss in power.

1 Introduction

We study the two-sample testing problem; given $\mathbb{X} = (X_1, \ldots, X_n) \overset{\text{i.i.d.}}{\sim} P$ and $\mathbb{Y} = (Y_1, \ldots, Y_m) \overset{\text{i.i.d.}}{\sim} Q$, we test the null hypothesis $H_0 : P = Q$ against the alternative $H_1 : P \neq Q$. This is a nonparametric hypothesis problem with a composite null hypothesis and a composite alternative hypothesis. It finds applications in diverse areas such as testing microarray data, clinical diagnosis and database attribute matching (Gretton et al., 2012a).

A popular approach to solving this problem is based on the *kernel-MMD* distance between the two empirical distributions (Gretton et al., 2006). Given a positive definite kernel k, the kernel-MMD distance between two distributions P and Q on $\mathcal{X}$, denoted by $\mathrm{MMD}(P, Q)$, is defined as

$$\mathrm{MMD}(P, Q) = \|\mu - \nu\|_k, \text{ where } \mu(\cdot) = \int_{\mathcal{X}} k(x, \cdot) dP(x), \text{ and } \nu(\cdot) = \int_{\mathcal{X}} k(x, \cdot) dQ(x). \quad (1)$$

Above, μ and ν are commonly called "kernel mean maps", and denote the kernel mean embeddings of the distributions P and Q into the reproducing kernel Hilbert space (RKHS) associated with the positive-definite kernel k, and $\|\cdot\|_k$ denotes the corresponding RKHS norm. Under mild conditions on the positive definite kernel k (Sriperumbudur et al., 2011), MMD is a metric on the space of

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

probability distributions. Gretton et al. (2006) suggested using an empirical estimate of the squared distance as the test statistic. In particular, given $\mathbb{X}$ and $\mathbb{Y}$, define the test statistic

$$\widehat{\mathrm{MMD}}^2 := \frac{1}{n(n-1)m(m-1)} \sum_{1 \le i \ne i' \le n} \sum_{1 \le j \ne j' \le m} h(X_i, X_{i'}, Y_j, Y_{j'}),$$

where $h(x, x', y, y') := k(x, x') - k(x, y') - k(y, x') + k(y, y')$. The above statistic has an alternative form that only takes quadratic time to calculate.

The MMD test rejects the null if $\widehat{\mathrm{MMD}}^2$ exceeds a suitable threshold $\tau \equiv \tau(\alpha)$ that ensures the false positive rate is at most α. For "characteristic kernels", this test is *consistent* against fixed alternatives, meaning the power (the probability of rejecting the null when $P \ne Q$) increases to one as $m, n \to \infty$.

The difficulty in practically determining τ will play a key role in this paper. It is well known that when $P = Q$, $\widehat{\mathrm{MMD}}^2$ is an instance of a "degenerate two-sample U-statistic", meaning that:

$$\text{Under } H_0, \quad \mathbb{E}_P[h(x, X', y, Y')] = \mathbb{E}_Q[h(X, x', Y, y')] = 0.$$

(Above, x, y, x', y' are fixed, and the expectations are over $X, Y, X', Y' \overset{\text{i.i.d.}}{\sim} P$.) As a consequence, its (limiting) null distribution is unwieldy; it is an infinite sum of independent χ^2 random variables weighted by the eigenvalues of an operator that depends on the kernel k and the underlying distribution P (see equation (10) in Appendix A). Since P is unknown, one cannot explicitly calculate τ.

In practice, a permutation-based approach is commonly used, where τ is set as the $(1-\alpha)$-quantile of the kernel-MMD statistic computed on B permuted versions of the aggregate data $(\mathbb{X}, \mathbb{Y})$. The resulting test has finite-sample validity, but its practical applicability is reduced due to the high computational complexity; if $B = 200$ permutations are used, the (permuted) test statistic must be recomputed 201 times, rather than once (usually, B is chosen between 100 and 1000).

Due to the high computational complexity of the permutation test, some permutation-free alternatives for selecting τ have been proposed. However, as we discuss in Section 1.2, these alternatives are either too conservative in practice (using concentration inequalities), or heuristics with no theoretical guarantees (Pearson curves and Gamma approximation) or are only shown to be consistent in the setting where the kernel k does not vary with n (spectral approximation). We later recap some computationally efficient alternatives to $\widehat{\mathrm{MMD}}^2$, but these have significantly lower power.

As far as we are aware, there exists no method in literature based on the kernel-MMD that is (i) permutation-free (does not require permutations), (ii) consistent against any fixed alternative, (iii) achieves minimax rate-optimality against local alternatives, and (iv) is correct for both the fixed kernel setting (k is fixed as $m, n \to \infty$) and the changing kernel setting (k changes as a function of m, n, for instance, by selecting the scale parameter of a Gaussian kernel in a data-driven manner).

Our work delivers a novel and simple test satisfying all four desirable properties. We propose a new variant of the kernel-MMD statistic that (after studentization) has a standard Gaussian limiting distribution under the null in both the fixed and changing kernel settings, in low- and high-dimensional settings. There is a computation-statistics tradeoff: our permutation-free test loses about a $\sqrt{2}$ factor in power compared to the standard kernel-MMD test, but it is hundreds of times faster.

Remark 1. Let $\mathcal{P}(\mathcal{X})$ denote the set of all probability measures on the observation space $\mathcal{X}$, where we often use $\mathcal{X} = \mathbb{R}^d$ for some $d \ge 1$. For simplicity, in the above presentation, the distributions P, Q, kernel k and dimension d did not change with sample size, and this is the setting considered in the majority of the literature. Later, we prove several of our results in a significantly more general setting where P, Q, d, k can vary with n, m. Under the null, this provides a much more robust type-I error control in high-dimensional settings, even with data-dependent kernels. Under the alternative, this provides a more fine-grained power result. To elaborate on the latter, we assume that for every n, m, the pair $(P, Q) = (P_n, Q_n) \in \mathcal{P}_n^{(1)} \subset \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$ for some sequence $\{\mathcal{P}_n^{(1)} : n, m \ge 2\}$. The class $\mathcal{P}_n^{(1)}$ is such that with increasing n and m, it contains pairs (P', Q') that are increasingly closer in some distance measure ϱ; thus the alternatives can approach the null and be equal in the limit. That is, $\Delta_{n,m} := \inf_{(P',Q') \in \mathcal{P}_n^{(1)}} \varrho(P', Q')$ decreases with n, m, and such alternatives are called *local* alternatives (as opposed to *fixed* alternatives). This framework allows us to characterize the *detection boundary* of a test, that is, the smallest perturbation from the null (in terms of $\Delta_{n,m}$) that can be consistently detected by a test.

Paper outline. We present an overview of our main results in Section 1.1 and discuss related work in Section 1.2. In Section 2, we present the cross-MMD statistic and obtain its limiting null distribution in Section 2.1. We demonstrate its consistency against fixed alternatives and minimax rate-optimality against smooth local alternatives in Section 2.2. Section 3 contains numerical experiments that demonstrate our theoretical claims. All our proofs are in the supplement.

1.1 Overview of our main results

We propose a variant of the quadratic time kernel-MMD statistic of (1) that relies on two key ideas: (i) *sample splitting* and (ii) *studentization*. In particular, we split the sample $\mathbb{X}$ of size $n \geq 2$ into $\mathbb{X}_1$ and $\mathbb{X}_2$ of sizes $n_1 \geq 1$ and $n_2 \geq 1$, respectively (and $\mathbb{Y}$ of size $m \geq 2$ into $\mathbb{Y}_1$ and $\mathbb{Y}_2$ of sizes $m_1 \geq 1$ and $m_2 \geq 1$), and define the *two-sample cross kernel-MMD* statistic $\widehat{\mathrm{cMMD}}^2$ as follows:

$$\widehat{\mathrm{cMMD}}^2 := \frac{1}{n_1 m_1 n_2 m_2} \sum_{i=1}^{n_1} \sum_{i'=1}^{n_2} \sum_{j=1}^{m_1} \sum_{j'=1}^{m_2} h(X_i, X_{i'}, Y_j, Y_{j'}). \tag{2}$$

Our final test statistic is $\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2 := \widehat{\mathrm{cMMD}}^2 / \widehat{\sigma}$, where $\widehat{\sigma}$ is an empirical variance introduced in (4).

Our first set of results show that quite generally, $\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2$ has an $N(0,1)$ asymptotic null distribution. Theorem 4 obtains this result in the setting where both the kernel k and null distribution P are fixed. This is then generalized to deal with changing kernels (for instance, Gaussian kernels with data-driven bandwidth choices) in Theorem 5. Finally, in Theorem 15 in Appendix A, we significantly expand the scope of these results by also allowing the null distribution to change with n, and also weakening the moment conditions required by Theorem 5.

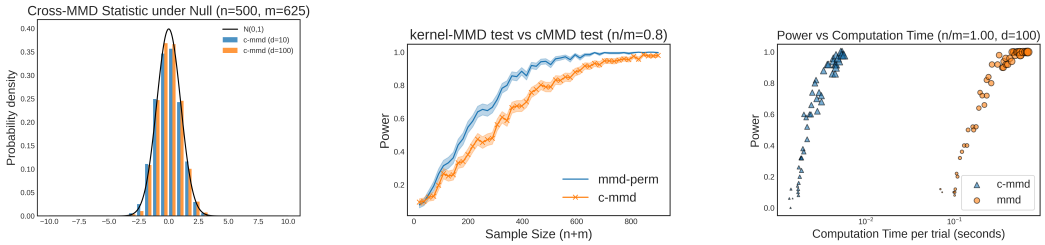

Figure 1: The first figure shows the distribution of our proposed statistic $\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2$ predicted by Theorem 5 under the null for dimensions $d = 10$ and $d = 100$. The statistic is computed with Gaussian kernel ($k_{s_n}(x, y) = \exp(-s_n \|x - y\|_2^2)$) with scale parameter s_n chosen by the *median* heuristic for different choices of n, m and d, and with samples $\mathbb{X}$ and $\mathbb{Y}$ drawn from a multivariate Gaussian distribution with identity covariance matrix. The second figure compares the power curves of the two-sample test using the $\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2$ statistic with the kernel-MMD permutation test (with 200 permutations). The final figure plots the power vs computation time for the two tests. The size of the markers are proportional to the sample-size used in the test.

Our main methodological contribution is the "cMMD test", denoted Ψ, which rejects the null if $\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2$ exceeds $z_{1-\alpha}$, which is the $(1-\alpha)$-quantile of $N(0,1)$. Formally,

$$\text{cMMD test: } \Psi(\mathbb{X}, \mathbb{Y}) = \mathbb{1}_{\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2 \geq z_{1-\alpha}}. \tag{3}$$

By the previous results, Ψ has type-I error at most α, meaning that $\mathbb{E}[\Psi(\mathbb{X}, \mathbb{Y})] \leq \alpha$ under the null.

We next study the power of the cMMD test Ψ in Section 2.2. First, in the fixed alternative case, i.e., when the distributions $P \neq Q$ do not change with n, we show in Theorem 7, that the cMMD test implemented with any characteristic kernel is consistent under a bounded fourth moment condition. Next, we consider the more challenging case of local alternatives, i.e., when the distributions, $P_n \neq Q_n$, change with n. In Theorem 8, we first identify general sufficient conditions for the cMMD test to be uniformly consistent over a class of alternatives. Then, we specialize this to the case when P_n and Q_n admit densities p_n and q_n with $\|p_n - q_n\|_{L^2} \geq \Delta_n$ for some $\Delta_n \to 0$. We

show in Theorem 9, that the cMMD test with a Gaussian kernel $k_{s_n}(x, y) = \exp(-s_n\|x - y\|_2^2)$, with scale parameter s_n increasing at an appropriate rate can consistently detect the local alternatives $\{\Delta_n : n \geq 1\}$ decaying at the minimax rate.

Finally, we note that while our primary focus in the paper is on the special case of kernel-MMD statistic, the ideas involved in defining the cMMD statistic can be extended to the case of general two-sample U-statistics. We describe this in Appendix D.1, and obtain sufficient conditions for asymptotic Gaussian limit of the resulting statistic, possibly of independent interest.

1.2 Comparisons to related work

Attempts to avoid permutations. There have been some prior attempts to avoid permutations, but they are either heuristics (no provable type-I error control) or have poor power (higher type-II error).

The first approach to obtaining a rejection threshold is based on large deviation bounds for the MMD statistic (Gretton et al., 2006, § 3) or the permuted MMD statistic (Kim, 2021, § 5). The resulting tests are distribution-free, but they tend to be too conservative (type-I error much less than α, resulting in low power). Another approach involves choosing the threshold based on parametric estimates of the limiting null distribution. For example, Gretton et al. (2006) suggested fitting to the Pearson family of densities based on the first four moments, while Gretton et al. (2009) introduced a more computationally efficient method using a two-parameter Gamma approximation. However, both of these methods are heuristic and do not have any consistency guarantees.

Gretton et al. (2009) introduced a spectral method for approximating the null distribution using the eigendecomposition of the gram matrix. They showed that the resulting distribution converges to the true null distribution as long as the square roots of the eigenvalues associated with the kernel operator are summable. While this method is asymptotically consistent, the conditions imposed on the kernel are more stringent than that used in our work. Furthermore, this method was shown to be consistent only in the fixed kernel (or low-dimensional) setting. Hence, it is unknown whether the results carry over to the case of kernels varying with sample size or high-dimensional settings. This method is also computationally nontrivial due to the need for a full eigendecomposition. Keeping only the top few eigenvectors is another heuristic, but this introduces an extra hyperparameter and loses theoretical guarantees; as a result this method is rarely used in practice. Our methods are simpler (no extra hyperparameter), faster (closed-form threshold), and more robust (type I error guarantees also hold for changing kernels, and in high-dimensional settings).

Changing the statistic: block-MMD and linear-MMD statistics. An idea more closely related to ours is that changing the test statistic itself would help make it cheaper to compute and also yield a tractable limiting distribution. One approach splits the observations into disjoint blocks, compute the kernel-MMD statistic on every block, and the final test statistic averages over all the blocks. If the size of each block is fixed, we get a linear-time kernel-MMD (Gretton et al., 2012a,b). The case of block sizes increasing with n, m was studied by Zaremba et al. (2013); Ramdas et al. (2015). Depending on the block size (b), the computational complexity of block-MMD statistic varies from linear (constant b) to quadratic ($b = \Omega(n)$). Further, if $b = o(n)$, then one gets a Gaussian null distribution as well.

Our proposed statistic is fundamentally different from the block-MMD statistics, despite both being incomplete U-statistics (Lee, 1990). In particular, the block-MMD statistics can be understood as building a block-diagonal approximation of the gram matrix. On the other hand, our proposed cross-MMD statistic uses the *off-diagonal blocks* of the gram matrix, exactly the blocks that the block-MMD with two blocks ($b = n/2$) excludes! The reason that this is a sensible thing to do is nontrivial, and our test is motivated quite differently from the block-MMD. In fact, when $b = n/2$, the block-MMD does not have a Gaussian null, but the cross-MMD does.

For the block-MMD with $b = o(n)$, the Gaussian null distribution is achieved at the cost of suboptimal power, as observed empirically in Zaremba et al. (2013), and proved by Reddi et al. (2015) for the case of linear-MMD and Ramdas et al. (2015) for general block-MMD statistics. In particular, their power is worse by factors scaling with n, which means that they are not minimax rate optimal. In contrast, our test uses exactly half the elements of the gram matrix, and its power is about a $\sqrt{2}$ worse than the MMD test, independent of n and d, and we prove explicitly that it achieves the minimax rate.

Beyond the kernel-MMD. The literature on two-sample testing is vast, and one can move even further away from the kernel-MMD (than the block-MMD) while retaining some of its intuition. For

example, Chwialkowski et al. (2015) proposed a linear-time test statistic by computing the average squared-distance between the empirical kernel embeddings at J randomly drawn points. Jitkrittum et al. (2016) proposed a variant of this statistic in which the J points are selected to maximize a lower bound on the power. In both cases, when the kernel k being used is analytic, in addition to being characteristic and integrable, the authors showed that the limiting distribution under null for this statistic is a combination of J independent χ^2 random variables. However, similar to the spectral method of Gretton et al. (2009), the high-dimensional behavior of these statistics are unknown. In fact, some preliminary experiments with $d \approx n$ in Appendix E.3 suggest that these linear-time statistics have a different null distribution in this regime. Further, the authors only proved consistency of these tests against fixed alternatives, but their power is not known to be minimax rate optimal. In contrast, our statistic has the same limiting distribution in low- and high-dimensional settings even with changing kernels, and is provably minimax rate optimal for smooth alternatives, and it is a much more direct tweak of the usual MMD test.

One-sample (goodness-of-fit) testing. Kim and Ramdas (2020) proposed and analyzed a similar studentized cross U-statistic in the simpler one-sample setting. Our work has different motivations: our primary goal in this paper is to design a permutation-free kernel-MMD test that does not significantly sacrifice the power, while Kim and Ramdas (2020) pursued the related but different goal of *dimension-agnostic* inference, which means having the same limiting distribution in low-dimensional and high-dimensional settings. Nevertheless, our results can be seen as an extension of their methods to two-sample testing. Our proofs also build on their advances, but we require a more involved analysis since in their case the second distribution is known (making it a point null).

2 Deriving the cross-MMD test

In this section, we present our test statistic and investigate its limiting distribution. First note that the squared kernel-MMD distance between two probability measures P and Q can be expressed as an inner product, namely $\langle \mu - \nu, \mu - \nu \rangle_k$. The usual kernel-MMD statistic is obtained by plugging the empirical kernel embeddings into this inner product expression and removing the diagonal terms to make it unbiased. Our proposal instead considers pairs of empirical estimates $(\widehat{\mu}_1, \widehat{\mu}_2)$ and $(\widehat{\nu}_1, \widehat{\nu}_2)$ constructed via sample splitting, and use the inner product between $\widehat{\mu}_1 - \widehat{\nu}_1$ and $\widehat{\mu}_2 - \widehat{\nu}_2$ instead. This careful construction allows us to obtain a Gaussian limiting distribution after studentization.

To elaborate, recall from Section 1.1 that we partition $\mathbb{X}$ into $\mathbb{X}_1$ and $\mathbb{X}_2$, and similarly $\mathbb{Y}$ into $\mathbb{Y}_1$ and $\mathbb{Y}_2$. We then compute empirical kernel embeddings based on each partition, yielding $\widehat{\mu}_1 := n_1^{-1} \sum_{i=1}^{n_1} k(X_i, \cdot)$, $\widehat{\mu}_2 := n_2^{-1} \sum_{i'=1}^{n_2} k(X_{i'}, \cdot)$, $\widehat{\nu}_1 := m_1^{-1} \sum_{j=1}^{m_1} k(Y_j, \cdot)$ and $\widehat{\nu}_2 := m_2^{-1} \sum_{j'=1}^{m_2} k(Y_{j'}, \cdot)$. Using these embeddings coupled with the kernel trick, the cross U-statistic (2) can be written as $\widehat{\mathrm{cMMD}}^2 = \langle \widehat{\mu}_1 - \widehat{\nu}_1,\ \widehat{\mu}_2 - \widehat{\nu}_2 \rangle_k$. To further motivate our test statistic, denote $U_{X,i} := \langle k(X_i, \cdot),\ \widehat{\mu}_2 - \widehat{\nu}_2 \rangle_k$ for $i = 1, \ldots, n_1$ and $U_{Y,j} := \langle k(Y_j, \cdot),\ \widehat{\mu}_2 - \widehat{\nu}_2 \rangle_k$ for $j = 1, \ldots, m_1$. Then the cross U-statistic can be viewed as the difference between two sample means: $\widehat{\mathrm{cMMD}}^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} U_{X,i} - \frac{1}{m_1} \sum_{j=1}^{m_1} U_{Y,j}$. Since the summands are independent *conditional on* $\mathbb{X}_2$ *and* $\mathbb{Y}_2$, one may expect that $\widehat{\mathrm{cMMD}}^2$ is approximately Gaussian after studentization. Our results in Section 2.1 formalize this intuition under standard moment conditions, where it takes some care to remove the above conditioning, since we care about the unconditional distribution.

Let us further denote the sample means of $U_{X,i}$'s and $U_{Y,j}$'s by $\bar{U}_X$ and $\bar{U}_Y$, respectively, and define

$$\widehat{\sigma}_X^2 := \frac{1}{n_1} \sum_{i=1}^{n_1} \left(U_{X,i} - \bar{U}_X \right)^2, \ \widehat{\sigma}_Y^2 := \frac{1}{m_1} \sum_{j=1}^{m_1} \left(U_{Y,j} - \bar{U}_Y \right)^2 \text{ and } \widehat{\sigma}^2 := \frac{1}{n_1} \widehat{\sigma}_X^2 + \frac{1}{m_1} \widehat{\sigma}_Y^2. \tag{4}$$

Now we have completed the description of our studentized cross U-statistic $\widehat{\overline{\mathrm{cMMD}}}^2 = \widehat{\mathrm{cMMD}}^2 / \widehat{\sigma}$, and the resulting test Ψ in (3). The asymptotic validity of the cMMD test is guaranteed by Theorem 15 that establishes the asymptotic normality of $\widehat{\overline{\mathrm{cMMD}}}^2$ under the null.

Remark 2 (Computational Complexity)**.** The overall cost of computing the statistic $\widehat{\overline{\mathrm{cMMD}}}^2$ is $\mathcal{O}\left((n+m)^2\right)$, and in particular, both $\widehat{\mathrm{cMMD}}^2$ and $\widehat{\sigma}$ have quadratic complexity. To see this, note that $\widehat{\mathrm{cMMD}}^2$ can be expanded into $\langle \widehat{\mu}_1, \widehat{\mu}_2 \rangle_k + \langle \widehat{\nu}_1, \widehat{\nu}_2 \rangle_k - \langle \widehat{\mu}_1, \widehat{\nu}_2 \rangle_k - \langle \widehat{\nu}_1, \widehat{\mu}_2 \rangle_k$. Each of these

terms can be computed in $\mathcal{O}\left((n+m)^2\right)$. Similarly, each term in the summations defining $\widehat{\sigma}_X^2$ and $\widehat{\sigma}_Y^2$ also require $\mathcal{O}\left((n+m)^2\right)$ computation, implying that the $\widehat{\sigma}$ also has $\mathcal{O}((n+m)^2)$ complexity.

Remark 3. To simplify notation in what follows, we denote m as m_n, where m_n is some unknown nondecreasing sequence such that $\lim_{n\to\infty} m_n = \infty$. This still permits m, n to be separate quantities growing to infinity at potentially different rates, but it allows us to index the sequence of problems with the single index n (rather than m, n). We will use k_n, d_n, $\mathcal{X}_n$, P_n and Q_n to indicate that quantities could (but do not have to) change as n increases, and drop the subscript when they are fixed. Furthermore, unless explicitly stated, we will focus on the balanced splitting scheme, i.e., $n_1 = \lfloor n/2 \rfloor$ and $m_1 = \lfloor m/2 \rfloor$ in what follows, because we currently see no apriori reason to split asymmetrically.

2.1 Gaussian limiting distribution under the null hypothesis

As shown in Figure 1, the empirical distribution of $\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2$ resembles a standard normal distribution for various choices of m, n and dimension d under the null. In this section, we formally prove this statement. Recalling the mean embedding μ from (1), define

$$\bar{k}(x, y) := \langle k(x, \cdot) - \mu, k(y, \cdot) - \mu \rangle_k. \tag{5}$$

Theorem 4. *Suppose that k and P do not change with n. If $0 < \mathbb{E}_P[\bar{k}(X, X')^4] < \infty$ for $X, X' \overset{i.i.d.}{\sim} P$, then $\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2 \xrightarrow{d} N(0, 1)$.*

We next present a more general result that implies Theorem 4.

Theorem 5. *Suppose P is fixed, but the kernel k_n changes with n. If*

$$\lim_{n\to\infty} \frac{\mathbb{E}_P[\bar{k}_n(X_1, X_2)^4]}{\mathbb{E}_P[\bar{k}_n(X_1, X_2)^2]^2} \left(\frac{1}{n} + \frac{1}{m_n}\right) = 0, \quad \textit{and} \quad \lim_{n\to\infty} \frac{\lambda_{1,n}^2}{\sum_{l=1}^{\infty} \lambda_{l,n}^2} \textit{ exists}, \tag{6}$$

where $(\lambda_{l,n})_{l=1}^{\infty}$ denote the eigenvalues of $\bar{k}$ introduced in (16)*, then we have $\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2 \xrightarrow{d} N(0, 1)$.*

It is easy to check that condition (6) is trivially satisfied if the kernels $\{k_n : n \geq 1\}$ are uniformly bounded by some constant; prominent examples are the Gaussian or Laplace kernel with a sample size dependent bandwidth. Thus, the above condition really exists to handle unbounded kernels and heavy-tailed distributions. To motivate this requirement, we recall Bentkus and Götze (1996) (see Fact 11 in Appendix A) who proved a studentized CLT for i.i.d. random variables in a triangular array setup: $W_{1,n}, W_{2,n}, \ldots, W_{n,n} \overset{\text{i.i.d.}}{\sim} P_n$. Define $V_n = \sqrt{n} \sum_{i=1}^n W_{i,n} / \sqrt{\sum_i (W_{i,n} - \bar{W}_n)^2}$ where $\bar{W}_n = (\sum_i W_{i,n})/n$. They showed that a sufficient condition for the asymptotic normality of V_n is that $\lim_{n\to\infty} \mathbb{E}_{P_n}[W_{1,n}^3] / \sqrt{\mathbb{E}_{P_n}[W_{1,n}^2]^3 n} = 0$. (This last condition is trivially true if P_n does not change with n, meaning that the triangular array setup is irrelevant and $W_{1,n}$ can be replaced by W_1.)

Our requirement is slightly stronger: condition (6) with $\bar{k}_n(X_1, X_2)$ replaced by $W_{1,n}$ implies the previous condition of Bentkus and Götze (1996) (details in Remark 12 in Appendix A). We need this stronger condition, because the terms in the definition of $\widehat{\mathrm{c}\mathrm{MMD}}^2$ are not i.i.d. (indeed, not even independent), and thus we cannot directly apply the result of Bentkus and Götze (1996). Instead, we take a different route by first conditioning on the second half of data $(\mathbb{X}_2, \mathbb{Y}_2)$, then showing the conditional asymptotic normality of the standardized $\widehat{\mathrm{c}\mathrm{MMD}}^2$ (i.e., divided by conditional standard deviation instead of empirical), and finally showing that the ratio of conditional and empirical standard deviations converge in probability to 1 (see Appendix B).

Finally, we note that the result of Theorem 5 can be further generalized in several ways: **(i)** instead of a fixed P and changing k_n, we can consider a sequence of pairs $\{(P_n, k_n) : n \geq 1\}$ changing with n, **(ii)** we can let $P_n \in \mathcal{P}_n^{(0)}$, for a class of distributions changing with n, and obtain the Gaussian limit uniformly over all elements of $\mathcal{P}_n^{(0)}$, and finally, **(iii)** the moment requirements on $\bar{k}_n$ stated in condition in (6) can also be slightly weakened. We state and prove this significantly more general version of Theorem 5 in Appendix B.

Remark 6. In the statement of the two theorems of this section, the splits $(\mathbb{X}_1, \mathbb{Y}_1)$ and $(\mathbb{X}_2, \mathbb{Y}_2)$ are assumed to be drawn i.i.d. from the same distribution P. However, a closer look at the proof

of Theorem 5 indicates that the conclusions of the above two theorems hold even when the two splits are independent and drawn i.i.d. from possibly different distributions; that is $(\mathbb{X}_1, \mathbb{Y}_1)$ and $(\mathbb{X}_2, \mathbb{Y}_2)$ are independent of each other and drawn i.i.d. from distributions P_1 and P_2 respectively, with $P_1 \neq P_2$. In particular, under this more general condition, the asymptotic normality of $\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2$ still holds, and the resulting test Ψ still controls the type-1 error at the desired level. This may be useful for two-sample testing in settings where the entire set of data is not i.i.d., but two different parts of the data were collected in two different situations. The usual MMD can also handle such scenarios by using a subset of permutations that do not exchange the data across the two situations.

2.2 Consistency against fixed and local alternatives

Here, we show that the cMMD test Ψ introduced in (3) is consistent against a fixed alternative and also has minimax rate-optimal power against smooth local alternatives separated in L^2 norm.

We first show that analogous to Theorem 4, cMMD is consistent against fixed alternatives.

Theorem 7. *Suppose P, Q, k do not change with n, and $P \neq Q$. If k is a characteristic kernel satisfying $0 < \mathbb{E}_P[\bar{k}(X_1, X_2)^4] < \infty$, and $0 < \mathbb{E}_Q[\bar{k}(Y_1, Y_2)^4] < \infty$, then the* cMMD *test is consistent, meaning it has asymptotic power* 1.

The moment conditions required above are mild, and are satisfied trivially, for instance, by bounded kernels such as the Gaussian kernel. The "characteristic" condition is also needed for the consistency of the usual MMD test (Gretton et al., 2012a), and is also satisfied by the Gaussian kernel.

Recalling Remark 1, we next consider the more challenging setting where d_n, k_n can change with n, and (P_n, Q_n) can vary within a class $\mathcal{P}_n^{(1)} \subset \mathcal{P}(\mathcal{X}_n) \times \mathcal{P}(\mathcal{X}_n)$ that can also change with n. We present a sufficient condition under which the cMMD test Ψ is consistent *uniformly* over $\mathcal{P}_n^{(1)}$. Define $\gamma_n \coloneqq \mathrm{MMD}(P_n, Q_n)$, which is assumed nonzero for each n but could approach zero in the limit.

Theorem 8. *Let $\{\delta_n : n \geq 2\}$ denote any positive sequence converging to zero. If*

$$\lim_{n\to\infty} \sup_{(P_n,Q_n)\in\mathcal{P}_n^{(1)}} \frac{\mathbb{E}_{P_n,Q_n}[\widehat{\sigma}^2]}{\delta_n \gamma_n^4} + \frac{\mathbb{V}_{P_n,Q_n}(\widehat{\mathrm{cMMD}}^2)}{\gamma_n^4} = 0, \text{ where } \mathbb{V} \text{ denotes variance,} \tag{7}$$

then $\lim_{n\to\infty} \sup_{(P_n,Q_n)\in\mathcal{P}_n^{(1)}} \mathbb{E}_{P_n,Q_n}[1 - \Psi(\mathbb{X}, \mathbb{Y})] = 0$, meaning the cMMD *test is consistent.*

Note that while *any* sequence $\{\delta_n\}$ converging to zero suffices for the general statement above, the condition (7) is easiest to satisfy for slowly decaying δ_n, such as $\delta_n = 1/\log\log n$ for instance. The sufficient conditions for consistency of Ψ stated in terms of $\widehat{\sigma}$ and $\widehat{\mathrm{cMMD}}^2$ in (7) can also be translated into equivalent conditions on the kernel function k_n, similar to (6), and we present the details in Appendix C. Importantly, if P_n, Q_n, d_n are fixed and k_n is bounded, then both $\mathbb{E}[\widehat{\sigma}^2]$ and $\mathbb{V}(\widehat{\mathrm{cMMD}}^2)$ are $O(1/n)$, and γ_n is a constant, so the condition is trivially satisfied, and in fact the above condition is even weaker than the fourth-moment condition of the previous theorem.

2.3 Minimax rate optimality against smooth local alternatives

We now apply the general result of Theorem 8 to the case where the distributions P_n and Q_n admit Lebesgue densities p_n and q_n that lie in the order β Sobolev ball for some $\beta > 0$, defined as $\mathcal{W}^{\beta,2}(M) \coloneqq \{f : \mathcal{X} \to \mathbb{R} \mid f \text{ is a.s. continuous, and } \int (1+\omega^2)^{\beta/2} \|\mathcal{F}(f)(\omega)\|^2 dw < M < \infty\}$. Formally, we define the null and alternative class of distributions as follows:

$$\mathcal{P}_n^{(0)} = \{P \text{ with density } p : p \in \mathcal{W}^{\beta,2}(M)\}, \quad \text{and}$$
$$\mathcal{P}_n^{(1)} = \{(P, Q) \text{ with densities } p, q \in \mathcal{W}^{\beta,2}(M) : \|p - q\|_{L^2} \geq \Delta_n\},$$

for some sequence Δ_n decaying to zero. In particular, we assume that under H_0, $P_n = Q_n$ and $P_n \in \mathcal{P}_n^{(0)}$, while under H_1, we assume that $(P_n, Q_n) \in \mathcal{P}_n^{(1)}$.

Our next result shows that for suitably chosen scale parameter, the cMMD test Ψ with the Gaussian kernel is minimax rate-optimal for the above class of local alternatives. For simplicity, we state this result with $n = m$, noting that the result easily extends to the case when there exist constants $0 < c \leq C$, such that $c \leq n/m \leq C$.

Theorem 9. *Consider the case when $n = m$, and let $\{\Delta_n : n \geq 1\}$ be a sequence such that $\lim_{n\to\infty} \Delta_n n^{2\beta/(d+4\beta)} = \infty$. On applying the* cMMD *test Ψ with the Gaussian kernel $k_{s_n}(x, y) = \exp(-s_n\|x - y\|_2^2)$, if we choose the scale as $s_n \asymp n^{4/(d+4\beta)}$, then we have*

$$\lim_{n\to\infty} \sup_{P_n \in \mathcal{P}_n^{(0)}} \mathbb{E}_{P_n}[\Psi(\mathbb{X}, \mathbb{Y})] \leq \alpha \quad \textit{and} \quad \lim_{n\to\infty} \inf_{(P_n,Q_n)\in\mathcal{P}_n^{(1)}} \mathbb{E}_{P_n,Q_n}[\Psi(\mathbb{X}, \mathbb{Y})] = 1. \tag{8}$$

The proof of this statement is in Appendix C, and it follows by verifying that the conditions required by Theorem 8 are satisfied for the above choices of Δ_n and s_n.

Remark 10. Li and Yuan (2019, Theorem 5 (ii)) showed a converse of the above statement: if $\lim_{n\to\infty} \Delta_n n^{2\beta/(d+4\beta)} < \infty$, then there exists an $\alpha \in (0, 1)$ such that any asymptotically level α test $\widetilde{\Psi}$ must have $\lim_{n\to\infty} \inf_{(P,Q)\in\mathcal{P}(\Delta_n)} \mathbb{E}_{P,Q}[\widetilde{\Psi}(\mathbb{X}, \mathbb{Y})] < 1$. Hence, the sequence of $\{\Delta_n : n \geq 1\}$ used in Theorem 9 represents the smallest L_2-deviations that can be detected by any test, and (8) shows that our cMMD test Ψ can detect such changes, establishing its minimax rate-optimality.

3 Experiments

We now present experimental validation of the theoretical claims of the previous section. In particular, our experiments demonstrate that **(i)** the limiting null distribution of $\widehat{\overline{\mathrm{cMMD}}}^2$ is $N(0, 1)$ under a wide range of choices of dimension d, sample sizes n, m and the kernel k, and **(ii)** the power of our cMMD test is competitive with the kernel-MMD permutation test. We now describe the experiments in more detail. Additional experimental results are reported in Appendix E.

Limiting null distribution of $\widehat{\overline{\mathrm{cMMD}}}^2$. We showed in Theorem 15 that the statistic $\widehat{\overline{\mathrm{cMMD}}}^2$ has a limiting normal distribution under some mild assumptions. We empirically test this result when $\mathbb{X}$ and $\mathbb{Y}$ are drawn from $N(\mathbf{0}, I_d)$ with $\mathbf{0}$ denoting the all-zeros vector in $\mathbb{R}^d$, and in particular, study the effects of (i) dimension: $d = 10$ versus $d = 500$, (ii) skewness of the samples: $n/m = 1$ versus $n/m = 0.1$, and (iii) choice of kernel: Gaussian versus Quadratic, both with data-dependent scale parameters using *median* heuristic.

As shown in the first row of Figure 2, the distribution of $\widehat{\overline{\mathrm{cMMD}}}^2$ is robust to all these effects, and is close to $N(0, 1)$ in all cases. In contrast, the distribution of the kernel-MMD statistic scaled by its empirical standard deviation (obtained using 200 bootstrap samples) in the bottom row of Figure 2 shows strong changes with these parameters. We present additional figures and details of the implementation in Appendix E.

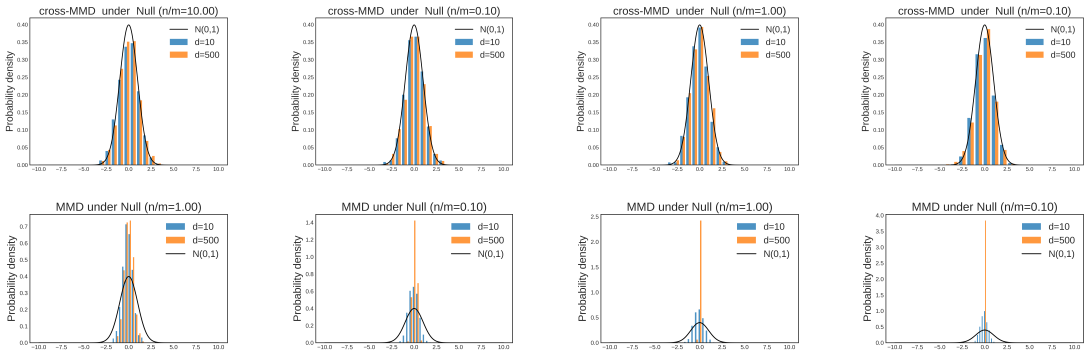

Figure 2: The first two columns show the null distribution of the $\widehat{\overline{\mathrm{cMMD}}}^2$ statistic (top row) and the $\widehat{\mathrm{MMD}}^2$ statistic scaled by its empirical standard deviation (bottom row) using the Gaussian kernel with scale-parameter chosen using the median heuristic. The last two columns show the null distribution for the two statistics using the Quadratic kernel with scale parameter chosen using the median heuristic. The figures demonstrate that the null distribution of $\widehat{\mathrm{MMD}}^2$ changes significantly with dimension (d), the ratio n/m and the choice of the kernel, unlike our proposed statistic.

Evaluation of the power of Ψ. For $d \geq 1$ and $j \leq d$, let $a_{\epsilon,j}$ denote the element of $\mathbb{R}^d$ with first j coordinates equal to ϵ, and others equal to 0. We consider the two-sample testing problem with $P = N(\mathbf{0}, I_d)$ $Q = N(a_{\epsilon,j}, I_d)$ for different choices of ϵ and d and j. We compare the performance of our proposed test Ψ with the kernel-MMD permutation test, implemented with $B = 200$ permutations, and plot the power-curves (over 400 trials) in Figure 3. We also propose a heuristic for predicting the power of the permutation test (denoted by ρ_{perm}) using the power of Ψ (denoted by ρ_Ψ) as follows (with Φ denoting the standard normal cdf, and z_α its α-quantile):

$$\widehat{\rho}_{\text{perm}} = \Phi\left(z_\alpha + \sqrt{2}\left(\Phi^{-1}(\rho_\Psi) - z_\alpha\right)\right). \tag{9}$$

This heuristic is motivated by the power expressions derived by Kim and Ramdas (2020) for the problems of one-sample Gaussian mean and covariance testing (we discuss this further in Appendix E). The term $\sqrt{2}$ in the above expression quantifies the price to pay for sample-splitting. As shown in Figure 3, this heuristic gives us an accurate estimate of the power of the kernel-MMD permutation test, without incurring the computational burden.

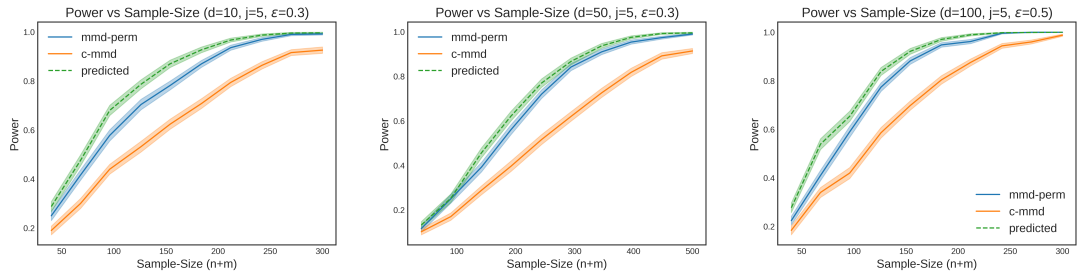

Figure 3: Curves showing the variation in power versus sample-size for the cMMD test and the kernel-MMD permutation test. $\mathbb{X}$ are drawn from $N(\mathbf{0}, I_d)$ i.i.d. and $\mathbb{Y}$ is drawn from $N(a_{\epsilon,j}, I_d)$ where $a_{\epsilon,j}$ is obtained by perturbing the first $j \leq d$ coordinates of $\mathbf{0}$ by ϵ, the kernel used is the Gaussian kernel with scale parameter chosen via the median heuristic. The dashed curve shows the predicted power of the kernel-MMD permutation test using the heuristic defined in (9).

We now use ROC curves to compare the tradeoff between type-I and type-II errors for the usual MMD, linear and block-MMDs with our $\widehat{\overline{\text{cMMD}}}^2$. We use the same distributions $P = N(\mathbf{0}, I_d)$ and $Q = N(a_{\epsilon,j}, I_d)$ as before, and plot results for $(d, j, \epsilon) \in \{(10, 5, 0.2), (100, 20, 0.15), (500, 100, 0.1)\}$ in Figure 4. Due to sample splitting, the tradeoff achieved by our proposed statistic is slightly worse than that of $\widehat{\text{MMD}}^2$, but significantly better than other computationally efficient variants of kernel-MMD statistics. More details about the implementation, and additional figures are presented in Appendix E.

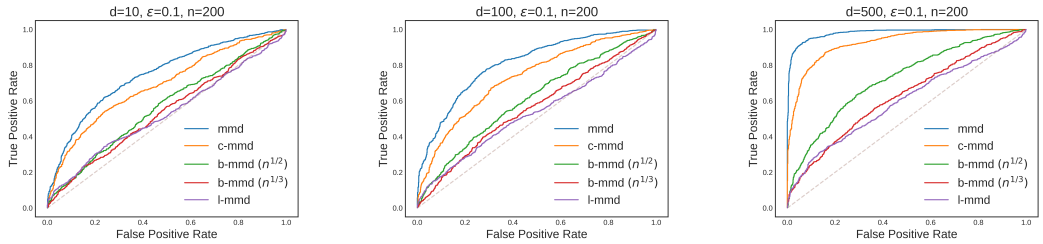

Figure 4: ROC curves highlighting the trade-off between type-I and type-II errors achieved by the MMD, cross-MMD, batch-MMD with batch sizes $n^{1/2}$ and $n^{1/3}$, and linear-MMD statistics.

4 Conclusion and future work

We proposed a variant of the kernel-MMD statistic, called cross-MMD, based on the ideas of sample-splitting and studentization, and showed that it has a standard normal limiting null distribution. Using

this key result, we introduced a permutation-free (and hence computationally efficient) MMD test for the two-sample problem. Experiments indicate that the power achieved by our test is within a $\sqrt{2}$-factor of the power of the kernel-MMD permutation test (that requires recomputing the statistic hundreds of times). In other words, our results achieve the following favorable tradeoff: *we get a significant reduction in computation at the price of a small reduction in power*.

Sejdinovic et al. (2013) establish in some generality that distance-based two-sample tests (like the energy distance (Székely and Rizzo, 2013)) can be viewed as kernel-MMD tests with a particular choice of kernel k. Hence our results are broadly applicable to distance-based statistics as well.

Since two-sample testing and independence testing can be reduced to each other, it is an interesting direction for future work to see if the ideas developed in our paper can be used for designing permutation-free versions of kernel-based independence tests like HSIC (Gretton et al., 2007) or distance covariance (Székely and Rizzo, 2009; Lyons, 2013).

Our techniques seem to rely on the specific structure of two-sample U-statistics of degree 2. Extending these to more general U-statistics of higher degrees is another important question for future work.

A final question is to figure out whether it is possible to achieve minimax optimal power using a sub-quadratic time test statistic. One potential approach would be to work with a kernel approximated by random Fourier features (Rahimi and Recht, 2007; Zhao and Meng, 2015). Depending on the number of random features, our test statistic can be computed in sub-quadratic time and it would be interesting to see whether the resulting test can still be minimax optimal in power. We leave this important question for future work.

References

Z. Bai and H. Saranadasa. Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, pages 311–329, 1996.

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

K. P. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. *Advances in Neural Information Processing Systems*, 28, 2015.

R. D'Agostino and E. S. Pearson. Tests for departure from normality. empirical results for the distributions of b_2 and $\sqrt{b}_1$. *Biometrika*, 60(3):613–622, 1973.

R. Durrett. *Probability: theory and examples*, volume 49. Cambridge University Press, 2019.

J. H. Friedman and L. C. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, pages 697–717, 1979.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems*, 19, 2006.

A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. *Advances in Neural Information Processing Systems*, 20, 2007.

A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur. A fast, consistent kernel two-sample test. *Advances in Neural Information Processing Systems*, 22, 2009.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012a.

A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. *Advances in Neural Information Processing Systems*, 25, 2012b.

N. Henze and M. D. Penrose. On the multivariate runs test. *Annals of statistics*, pages 290–298, 1999.

W. Jitkrittum, Z. Szabó, K. P. Chwialkowski, and A. Gretton. Interpretable distribution features with maximum testing power. *Advances in Neural Information Processing Systems*, 29, 2016.

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

I. Kim and A. Ramdas. Dimension-agnostic inference using cross U-statistics. *arXiv preprint arXiv:2011.05068*, 2020.

J. Lee. *U-statistics: Theory and Practice*. Citeseer, 1990.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer Science & Business Media, 2006.

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

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

A. Rahimi and B. Recht. Random Features for Large-Scale Kernel Machines. *Advances in Neural Information Processing systems*, 20, 2007.

A. Ramdas, S. J. Reddi, B. Poczos, A. Singh, and L. Wasserman. Adaptivity and computation-statistics tradeoffs for kernel and distance based high dimensional two sample testing. *arXiv preprint arXiv:1508.00655*, 2015.

S. Reddi, A. Ramdas, B. Póczos, A. Singh, and L. 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.

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

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research*, 12(7), 2011.

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

G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

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

W. Zaremba, A. Gretton, and M. Blaschko. B-test: low variance kernel two-sample test. *Advances in Neural Information Processing Systems*, 2013.

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In cases where computation cost is not an issue, the power achieved by our test will be slightly smaller (approximately by a factor of $\sqrt{2}$) as compared to the kernel-MMD permutation test.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] All the proofs are in the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We will include the code with the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We include all the details of the experiments in the appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] The experiments were run on a workstation, whose specifications are provided in Appendix E
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Contents

A Background **14**

B Gaussian limiting distribution of $\widehat{\overline{\mathrm{cMMD}}}^2$ **15**

B.1 Statement of the general result (both k_n and P_n changing with n) 16

B.1.1 Proof of the general result with changing k_n and P_n 16

B.2 Fixed P, changing k_n (Theorem 5) 21

B.3 Fixed k, and fixed P (Theorem 4) 22

C Consistency against fixed and local alternatives (Section 4) **23**

C.1 Proof of Theorem 8 (General conditions for consistency) 23

C.2 Proof of Theorem 7 (Consistency against fixed alternative) 23

C.3 Proof of Theorem 9 (Type-I error control and consistency against local alternative) 25

D Gaussian Limit for General Two-Sample U-Statistic **28**

D.1 Proof of Theorem 21 29

E Additional Experiments **30**

E.1 Implementation details of experiments reported in the main text 30

E.2 Additional Figures 31

E.3 Comparison with ME and SCF tests of Jitkrittum et al. (2016) 33

E.4 Type-I Error and goodness-of-fit test of null distribution 34

E.5 Comparison with Friedman-Rafsky test 35

A Background

To keep the paper self-contained, we collect the relevant definitions and theorems from prior work that are used in proving the main results of our paper.

Central Limit Theorems. We first recall a central limit theorem for studentized statistics by Bentkus and Götze (1996).

Fact 11 (Berry Esseen CLT)**.** *For some i.i.d. $\sim P$ random variables $W_1, \ldots, W_n$, define the statistic $\widehat{\overline{\mathrm{cMMD}}}^2 = \frac{\sum_{i=1}^n W_i}{\sqrt{\frac{1}{n}\sum_{i=1}^n (W_i - \bar{W}_n)^2}}$. If $\mathbb{E}_P[W_i] = 0$ and $0 < \mathbb{E}_P[W_i^2] < \infty$, then there exists a universal constant $C < \infty$ such that*

$$\sup_{x\in\mathbb{R}} \left|\mathbb{P}_P\left(T \le x\right) - \Phi(x)\right| \le C \frac{\mathbb{E}_P[|W_1^3|]}{\mathbb{E}_P[W_1^2]^{3/2}\sqrt{n}}.$$

Remark 12. Note that by Cauchy–Schwarz inequality, we have

$$\mathbb{E}_P[W_1^3] = \mathbb{E}_P[W_1^2 \times W_1] \le \sqrt{\mathbb{E}_P[W_1^4]\mathbb{E}_P[W_1^2]}.$$

This implies the following

$$\frac{\mathbb{E}_P[W_1^3]}{\mathbb{E}_P[W_1^2]^{3/2}} \le \left(\frac{\mathbb{E}_P[W_1^4]}{\mathbb{E}_P[W_1^2]^2}\right)^{1/2}.$$

Thus a sufficient condition for applying Fact 11 to show the convergence in distribution to $N(0,1)$ for a triangular sequence $\{W_{i,n} : 1 \leq i \leq n,\ n \geq 1\}$, with $\{W_{i,n} : 1 \leq i \leq n\}$ drawn i.i.d. from some distribution P_n is

$$\lim_{n\to\infty} \frac{\mathbb{E}_{P_n}[W_{1,n}^4]}{\mathbb{E}_{P_n}[nW_{1,n}^2]^2} = 0.$$

We next recall a consequence of Lindeberg's Central Limit Theorem (CLT), as stated in (Lehmann and Romano, 2006, Lemma 11.3.3).

Fact 13. *Let $Z_1, Z_2, \ldots$ be a sequence of i.i.d. zero-mean random variables with finite variance σ^2. Let $c_1, c_2, \ldots$ be a real-valued sequence, satisfying:*

$$\lim_{n\to\infty} \max_{1\leq i\leq n} \frac{c_i^2}{\sum_{j=1}^n c_j^2} = 0.$$

Then, we have

$$\frac{\sum_{i=1}^n c_i Y_i}{\sqrt{\sum_{j=1}^n c_j^2}} \xrightarrow{d} N(0,\sigma^2).$$

Null distribution of MMD statistic. Assuming that (n,m) are such that $n/m \to c$ for some $c > 0$, and let $\{u_l : l \geq 1\}$ and $\{v_l : l \geq 1\}$ denote two independent sequences of i.i.d. $N(0,1)$ random variables. Furthermore, let $\{\lambda_l : l \geq 1\}$ denote the eigenvalues of the kernel operator $f(\cdot) \mapsto \int_{\mathcal{X}} f(x)k(\cdot,x)dP(x)$. Using techniques from the theory of U-statistics, Gretton et al. (2012a) showed that

$$(n+m)\widehat{\mathrm{MMD}}^2 \xrightarrow{d} \sum_{l=1}^{\infty} \lambda_l \left(\frac{\left(c^{1/2}u_l - v_l\right)^2}{1+c} - \frac{(1+c)^2}{c} \right). \tag{10}$$

(10) shows that the null distribution of $\widehat{\mathrm{MMD}}$ is an infinite combination of chi-squared random variables, weighted by the eigenvalues of the kernel operator. Due to this form, the null distribution has a complex dependence on the kernel and the null distribution P.

Gaussian kernel calculations. Next, we recall some facts derived byLi and Yuan (2019), about the the Gaussian kernel $k_s(x,y) \coloneqq \exp\left(-s\|x-y\|_2^2\right)$, and probability distributions that admit density functions lying in the Sobolev ball $\in \mathcal{W}^{\beta,2}(M)$.

Fact 14. *Consider a Gaussian kernel that varies with sample size, $k_n(x,y) = \exp(-s_n\|x-y\|_2^2)$. Let $\bar{k}_n$ be as defined in* (15)*, $\mathcal{X} = \mathbb{R}^d$ and $X_1, X_2, X_3, X_4 \sim P_n$ i.i.d., $Y_1, Y_2 \sim Q_n$, where P_n and Q_n have densities p_n and q_n in $\mathcal{W}^{\beta,2}(M)$ and $\|p_n - q_n\|_{L^2} = \Delta_n$, for some real valued sequence $\{\Delta_n : n \geq 1\}$ converging to* 0*. Then, we have the following:*

$$\mathbb{E}_{P_n}[\bar{k}_n^2(X_1,X_2)] \asymp s^{-d/2}, \quad \textit{and} \quad \mathbb{E}_{Q_n}[\bar{k}_n^2(Y_1,Y_2)] \asymp s^{-d/2} \tag{11}$$

$$\mathbb{E}_{P_n}[\bar{k}_n^4(X_1,X_2)] \lesssim s^{-d/2}, \tag{12}$$

$$\mathbb{E}_{P_n}[\bar{k}_n^2(X_1,X_2)\bar{k}_n^2(X_1,X_3)] \lesssim s^{-3d/4}, \tag{13}$$

$$\gamma_n(P_n,Q_n) = \mathrm{MMD}(P_n,Q_n) \gtrsim s_n^{-d/2}\Delta_n. \tag{14}$$

Additional Notation. We use $U = o_P(u_n)$ and $U = O_P(u_n)$ to denote that $U/u_n \xrightarrow{p} 0$ and that U/u_n is stochastically bounded. For real valued sequences, we use $a_n \lesssim b_n$ if there exists a constant C such that $a_n \leq Cb_n$ for all n. We use $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

B Gaussian limiting distribution of $\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2$

In this section, we present the results about the limiting null distribution of the statistic $\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2$. The general outline of the section is as follows:

- In Appendix B.1, we state the most general version of the result on the limiting distribution of $\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2$ (Theorem 15), that we alluded to in Section 2.1. We then prove this result in Appendix B.1.1.
- In Appendix B.3, we show how the general result can be used to prove Theorem 5, where the kernel is allowed to change with n while the distribution P is fixed.
- Finally, in Appendix B.3, we show how Theorem 5 can be used to conclude the result for the case when both the kernel k and null distribution P are fixed with n.

B.1 Statement of the general result (both k_n and P_n changing with n)

As stated in Remark 3, we assume that $m \equiv m_n$ is some non-decreasing function of n. We consider a sequence of positive-definite kernels $\{k_n : n \geq 2\}$, and probability distributions $\{P_n : n \geq 1, 2\}$, and define

$$\bar{k}(x, y) \equiv \bar{k}_n(x, y) = \langle k_n(x, \cdot) - \mu_{P_n}, k_n(y, \cdot) - \mu_{P_n} \rangle_k, \tag{15}$$

where μ_{P_n} denotes the embedding of the distribution P_n into the RKHS associated with the kernel k_n. For any fixed values of n, we use $\{(\lambda_{l,n}, \varphi_{l,n}) : l \geq 1\}$ to denote the eigenvalue-eigenfunction sequence associated with the integral operator $g \mapsto \int \bar{k}(\cdot, x) g(x) dP_n(x)$. If $\bar{k}$ happens to be square-integrable (in addition to being symmetric), it has the following representation:

$$\bar{k}_n(x, y) = \sum_{l=1}^{\infty} \lambda_{l,n} \varphi_{l,n}(x) \varphi_{l,n}(y). \tag{16}$$

We now state the assumption required to prove the limiting normal distribution of the statistic $\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2$. As we will see in Appendix B.2, in the special case of fixed P, the condition in (17) is a weaker version of that used in Theorem 5.

Assumption 1. *For $\bar{k}$ introduced in* (5)*, $\{(\lambda_{l,n}, \varphi_{l,n}) : l \geq 1\}$ introduced in* (16) *and for a sequence $\{P_n : n \geq 1\}$, we assume that*

$$\frac{\mathbb{E}_{P_n}[\bar{k}^4(X_1, X_2)](n^{-1} + m_n^{-1}) + \mathbb{E}_{P_n}[\bar{k}^2(X_1, X_3)\bar{k}^2(X_2, X_3)]}{\mathbb{E}_{P_n}[\bar{k}^2(X_1, X_2)]^2 \left(\frac{1}{n^{-1}+m_n^{-1}}\right)} \to 0, \quad \textit{and} \tag{17}$$

$$\lim_{n\to\infty} \frac{\lambda_{1,n}^2}{\sum_{l=1}^{\infty} \lambda_{l,n}^2} \quad \textit{exists.}$$

We now state the main result of this section.

Theorem 15. *Suppose the sequence $\{m_n : n \geq 1\}$ satisfies $\lim_{n\to\infty} n/m_n$ exists and is non-zero. Let $\{k_n : n \geq 1\}$ be a sequence of positive definite kernels, and let $\mathcal{P}_n^{(0)}$ denote a family of distributions such that, for every $n \geq 1$ and $P_n \in \mathcal{P}_n^{(0)}$, Assumption 1 is satisfied by the pair $(\bar{k}_n, P_n)$ with $\bar{k}_n$ defined in* (15)*. Then, we have that*

$$\lim_{n\to\infty} \sup_{P_n \in \mathcal{P}_n^{(0)}} \sup_{x\in\mathbb{R}} |\mathbb{P}_{P_n}(\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2 \leq x) - \Phi(x)| = 0.$$

We now present the proof of this result.

B.1.1 Proof of the general result with changing k_n and P_n

To simplify the notation, we will drop the subscripts from k_n, $\bar{k}_n$, P_n, $\lambda_{l,n,m}$ and $\varphi_{l,n,m}$ in this proof outline. Furthermore, note that as mentioned in Remark 3, we assume that $n_1 = n/2$ and $n_1 = m/2$.

For any $x \in \mathcal{X}$, introduce the term $\widetilde{k}(x, \cdot)$ to denote $k(x, \cdot) - \mu$. Next, we define the following terms

$$S_X = \langle \widehat{\mu}_1 - \mu, \overbrace{(\widehat{\mu}_2 - \mu) - (\widehat{\nu}_2 - \mu)}^{:= g_2} \rangle_k, \quad \text{and} \quad S_Y = \langle \widehat{\nu}_1 - \mu, g_2 \rangle_k,$$

and note that we can write $\widehat{\mathrm{cMMD}}^2 = \bar{U}_X - \bar{U}_Y = S_X - S_Y$ (S_X differs from $\bar{U}_X$ due to the extra μ term in the first argument of the inner product). Recall that we use μ and ν to denote the kernel embeddings of the distributions P and Q.

We can further rewrite S_X and S_Y in terms of $\{W_i : 1 \leq i \leq n_1\}$ and $\{Z_j : 1 \leq j \leq m_1\}$ as follows:

$$S_X = \frac{1}{n_1}\sum_{i=1}^{n_1} \overbrace{\langle \widetilde{k}(X_i, \cdot),\, g_2 \rangle_k}^{:= W_i}, \quad \text{and} \quad S_Y = \frac{1}{m_1}\sum_{j=1}^{m_1} \overbrace{\langle \widetilde{k}(Y_j, \cdot),\, g_2 \rangle_k}^{:= Z_j}. \tag{18}$$

With these terms defined, we proceed in the following steps:

- **Step 1:** First, we consider the standardized random variables $T_{s,X}$ and $T_{s,Y}$, defined as
$$T_{s,X} := \frac{\sqrt{n_1} S_X}{\mathbb{E}_{P_n}[W_i^2 | \mathbb{X}_2, \mathbb{Y}_2]}, \quad \text{and} \quad T_{s,Y} := \frac{\sqrt{m_1} S_Y}{\mathbb{E}_{P_n}[Z_j^2 | \mathbb{X}_2, \mathbb{Y}_2]},$$
and prove that they converge in distribution to $N(0,1)$ conditioned on $(\mathbb{X}_2, \mathbb{Y}_2)$. To prove that the limiting distribution is Gaussian, we verify that $\frac{\mathbb{E}_{P_n}[W_i^4 | \mathbb{X}_2, \mathbb{Y}_2]}{n_1 \mathbb{E}_{P_n}[W_i^2 | \mathbb{X}_2, \mathbb{Y}_2]^2} \xrightarrow{p} 0$ and $\frac{\mathbb{E}_{P_n}[Z_j^4 | \mathbb{X}_2, \mathbb{Y}_2]}{m_1 \mathbb{E}_{P_n}[Z_j^2 | \mathbb{X}_2, \mathbb{Y}_2]^2} \xrightarrow{p} 0$. This is formally shown in Lemma 16 below.
- **Step 2:** Next, building upon the previous result, and using the conditional independence of $T_{s,X}$ and $T_{s,Y}$, we show in Lemma 17 below, that the standardized statistic $T_s = (S_X - S_Y)/\sqrt{n_1^{-1}\mathbb{E}_{P_n}[W_1^2 | \mathbb{X}_2, \mathbb{Y}_2] + m_1^{-1}\mathbb{E}_{P_n}[Z_1^2 | \mathbb{X}_2, \mathbb{Y}_2]}$ also converges in distribution to $N(0,1)$.
- **Step 3:** We then prove in Lemma 18 below that the ratio $\frac{n_1^{-1}\mathbb{E}_{P_n}[W_1^2 | \mathbb{X}_2, \mathbb{Y}_2] + m_1^{-1}\mathbb{E}_{P_n}[Z_1^2 | \mathbb{X}_2, \mathbb{Y}_2]}{n_1^{-1}\widehat{\sigma}_X^2 + m_1^{-1}\widehat{\sigma}_Y^2}$ converges in probability to 1.

It only remains to state and prove the three lemmas used above, which we do after this proof. Barring that, combining the above three steps completes the proof of the theorem. □

Before proceeding, we first introduce the terms $a_i = \langle \widetilde{k}(X_i, \cdot), \widehat{\mu}_2 - \mu \rangle_k$ and $b_i = \langle \widetilde{k}(X_i, \cdot), \widehat{\nu}_2 - \mu \rangle_k$, and note that we can further decompose W_i into $a_i - b_i$ for $1 \leq i \leq n_1$. Similarly, for $1 \leq j \leq m_1$, we can write Z_j as $c_j - d_j$ with $c_j = \langle \widetilde{k}(Y_j, \cdot), \widehat{\mu}_2 - \mu \rangle_k$ and $d_j = \langle \widetilde{k}(Y_j, \cdot), \widehat{\nu}_2 - \mu \rangle_k$.

We now state and prove the intermediate results to obtain Theorem 15.

Lemma 16. *Under the conditions of Theorem 15, we have the following:*

$$\frac{\mathbb{E}_{P_n}[W_i^4 | \mathbb{X}_2, \mathbb{Y}_2]}{n_1 \mathbb{E}_{P_n}[W_i^2 | \mathbb{X}_2, \mathbb{Y}_2]^2} \xrightarrow{p} 0, \quad \textit{and} \quad \frac{\mathbb{E}_{P_n}[Z_j^4 | \mathbb{X}_2, \mathbb{Y}_2]}{m_1 \mathbb{E}_{P_n}[Z_j^2 | \mathbb{X}_2, \mathbb{Y}_2]^2} \xrightarrow{p} 0.$$

Hence, as a consequence of the Lyapunov form of CLT (see Fact 11 and Remark 12 in Appendix A), this means that $T_{s,X} \xrightarrow{d} N(0,1)$ and $T_{s,Y} \xrightarrow{d} N(0,1)$ conditioned on $(\mathbb{X}_2, \mathbb{Y}_2)$.

Proof. We describe the steps for proving the first statement (involving W_i), noting that the other statement follows in an entirely analogous manner. Throughout this proof, we will use the shorthand $\mathbb{E}_2[\cdot]$ to denote the $\mathbb{E}_{P_n}[\cdot | \mathbb{X}_2, \mathbb{Y}_2]$.

By two applications of the AM-GM inequality, we observe that $W_i^4 = (a_i - b_i)^4 \leq 16(a_i^4 + b_i^4)$. Hence, we have the following:

$$\begin{aligned}
\frac{\mathbb{E}_2[W_i^4]}{16 n_1 \mathbb{E}_2[W_i^2]^2} &\leq \frac{\mathbb{E}_2[a_i^4 + b_i^4]}{n_1 \mathbb{E}_2[(a_i - b_i)^2]} \\
&= \frac{n_1 \mathbb{E}_2[a_i^4]}{\mathbb{E}_{P_n}[\bar{k}(X_1, X_2)^2]^2} \times \frac{\mathbb{E}_{P_n}[\bar{k}(X_1, X_2)^2]^2}{n_1^2 \mathbb{E}_2[(a_i - b_i)^2]} + \frac{m_1 \mathbb{E}_2[b_i^4]}{\mathbb{E}_{P_n}[\bar{k}(Y_1, Y_2)^2]^2} \times \frac{\mathbb{E}_{P_n}[\bar{k}(Y_1, Y_2)^2]^2}{m_1^2 \mathbb{E}_2[(a_i - b_i)^2]} \qquad (19) \\
&:= A_1 \times A_2 + B_1 \times B_2. \qquad (20)
\end{aligned}$$

Thus, to complete the proof, it suffices to show that $A_1 \times A_2$ and $B_1 \times B_2$ converge in probability to 0. This can be shown in two steps:

- Under the assumptions of Theorem 15, we have $A_1 \xrightarrow{p} 0$ and $B_1 \xrightarrow{p} 0$. To prove this result, it suffices to show that $\mathbb{E}_{P_n}[A_1] \to 0$ and $\mathbb{E}_{P_n}[B_1] \to 0$. The result then follows by an application of Markov's inequality.
- A_2 and B_2 are bounded in probability.

We first show that $\mathbb{E}_{P_n}[A_1] \to 0$. The result for B_1 follows similarly.

$$\begin{aligned}
\mathbb{E}_{P_n}[A_1] &= \frac{n_1}{\mathbb{E}_{P_n}[\bar{k}(X_1, X_2)^2]} \mathbb{E}_{P_n}\left[\mathbb{E}_2\left[a_i^2\right]\right] \\
&\overset{(i)}{=} \frac{n_1}{\mathbb{E}_{P_n}[\bar{k}(X_1, X_2)^2]} \left(\frac{\mathbb{E}_{P_n}\left[\bar{k}^4(X_1, X_2)\right]}{n_1^3} + \frac{3n_1(n_1-1)}{n_1^4} \mathbb{E}[\bar{k}^2(X_1, X_3)\bar{k}^2(X_2, X_3)] \right) \\
&\leq \frac{3}{\mathbb{E}_{P_n}[\bar{k}(X_1, X_2)^2]} \left(\frac{\mathbb{E}_{P_n}\left[\bar{k}^4(X_1, X_2)\right]}{n_1^2} + \frac{1}{n_1} \mathbb{E}[\bar{k}^2(X_1, X_3)\bar{k}^2(X_2, X_3)] \right),
\end{aligned}$$

which goes to 0 as required, by invoking the condition in (17) of Assumption 1. For (i), we used the expression derived by Kim and Ramdas (2020) while proving their Theorem 6.

To complete the proof, we show that A_2 is bounded in probability (the result for B_2 follows similarly). We consider two cases, depending on whether $\rho_1 := \lim_{n,m\to\infty} \frac{\lambda_1^2}{\sum_l \lambda_l^2}$ is equal to 0 or greater than 0 (the existence of this limit is assumed).

Case 1: $\rho_1 > 0$. We first observe that as a consequence of (15) and the orthonormality of the eigenfunctions, we have

$$\mathbb{E}_{P_n}[\bar{k}(X_1, X_2)^2] = \mathbb{E}_{P_n}\left[\sum_{l,l'} \lambda_l \lambda_{l'} \varphi_l(X_1)\varphi_{l'}(X_1)\varphi_l(X_2)\varphi_{l'}(X_2)\right] = \sum_{l=1}^{\infty} \lambda_l^2.$$

Using this, we obtain the following:

$$\frac{1}{(A_2)^{1/2}} = \frac{n_1 \mathbb{E}_2[a_i^2 + b_i^2 - 2a_i b_i]}{\sum_{l=1}^{\infty} \lambda_l^2}.$$

By repeated use of (15), we can show that the following identities hold:

$$\begin{aligned}
\mathbb{E}_2[a_i^2] &= \frac{1}{(n-n_1)^2} \sum_{l=1}^{\infty} \lambda_l^2 \left(\sum_{i'} \varphi_l(X_{i'}) \right)^2, \\
\mathbb{E}_2[b_i^2] &= \frac{1}{(m-m_1)^2} \sum_{l=1}^{\infty} \lambda_l^2 \left(\sum_{j'} \varphi_l(Y_{j'}) \right)^2, \quad \text{and} \\
\mathbb{E}_2[a_i b_i] &= \frac{1}{(n-n_1)(m-m_1)} \sum_{l=1}^{\infty} \lambda_l^2 \left(\sum_{i'} \varphi_l(X_{i'}) \right) \left(\sum_{j'} \varphi_l(Y_{j'}) \right).
\end{aligned}$$

Plugging these equalities in the expression for A_2, and using $\rho_l = \frac{\lambda_l}{\sum_{l'} \lambda_{l'}^2}$, we get

$$\begin{aligned}
(A_2)^{1/2} &= \frac{1}{n_1 \sum_l \rho_l \left(\frac{1}{n-n_1} \sum_{i'} \varphi_l(X_{i'}) - \frac{1}{m-m_1} \sum_{j'} \varphi_l(Y_{j'}) \right)^2} \\
&\leq \frac{1}{\rho_1 \left(\frac{\sqrt{n_1}}{n-n_1} \sum_{i'} \varphi_1(X_{i'}) - \frac{\sqrt{n_1}}{m-m_1} \sum_{j'} \varphi_1(Y_{j'}) \right)^2}
\end{aligned}$$

Since $n_1 = n/2$, $m_1 = m/2$, we have $\sqrt{n_1}/(n-n_1) = \sqrt{2/n}$ and $\sqrt{n_1}/(m-m_1) = \sqrt{2n}/m$. Introduce the notation $u_{i'} = \sqrt{2/n}/\sqrt{1+n/m}$ and $v_{j'} = (\sqrt{2n}/m)/\sqrt{1+n/m}$, and note that

$$
\begin{aligned}
(A_2)^{1/2} &\leq \frac{1}{\left(1+\frac{n}{m}\right)\rho_1\left(\sum_{i'} u_{i'}\varphi_1(X_{i'}) - \sum_{j'} v_{j'}\varphi_1(Y_{j'})\right)^2} \\
&\leq \frac{1}{\rho_1\left(\sum_{i'} u_{i'}\varphi_1(X_{i'}) - \sum_{j'} v_{j'}\varphi_1(Y_{j'})\right)^2}. \qquad (21)
\end{aligned}
$$

Next, we note that

$$
\begin{aligned}
\lim_{n\to\infty} \max_{i',j'} \frac{u_{i'}^2 + v_{j'}^2}{\sum_{i'} u_{i'}^2 + \sum_{j'} v_{j'}^2} &= \lim_{n\to\infty} \frac{2}{n+n^2/m} + \frac{2}{m+m^2/n} \\
&\leq \lim_{n\to\infty} 2\left(\frac{1}{n}+\frac{1}{m}\right) = 0.
\end{aligned}
$$

Thus, by an application of Lindeberg's CLT, we observe that the denominator in (21) converges in distribution to $N(0,\rho_1)^2$. This implies that $A_2 = \mathcal{O}_P(1)$, as required.

Case 2: $\rho_1 = 0$. Again, we observe that

$$
(A_2)^{-1/2} = \frac{n_1\mathbb{E}_2[a_i^2]}{\mathbb{E}_{P_n}[\bar{k}(X_1,X_2)^2]} + \frac{n_1\mathbb{E}_2[b_i^2]}{\mathbb{E}_{P_n}[\bar{k}(X_1,X_2)^2]} - 2\frac{n_1\mathbb{E}_2[a_ib_i]}{\mathbb{E}_{P_n}[\bar{k}(X_1,X_2)^2]}.
$$

The first two terms in the display above are $1+o_P(1)$, as shown in (Kim and Ramdas, 2020, pg 55, Step 2). For the last term, we introduce the notation $g(x,y) = \mathbb{E}_{P_n}[\bar{k}(X,x)\bar{k}(X,y)]$, and note the following:

$$
R := \frac{n_1\mathbb{E}_2[a_ib_i]}{\mathbb{E}_{P_n}[\bar{k}(X_1,X_2)^2]} = \frac{n_1}{(n-n_1)(m-m_1)}\sum_{i',j'} g(X_{i'},Y_{j'}).
$$

Since $X_{i'}$ and $Y_{j'}$ are independent, we observe that $\mathbb{E}_{P_n}[g(X_{i'},Y_{j'})] = 0$, and hence $\mathbb{E}_{P_n}[R] = 0$. Furthermore, the variance of R satisfies

$$
\begin{aligned}
\mathbb{E}_{P_n}[R^2] &= \frac{n_1^2}{(n-n_1)(m-m_1)}\frac{\mathbb{E}_{P_n}[g(X_1,X_2)^2]}{\mathbb{E}_{P_n}[\bar{k}(X_1,X_2)^2]} \\
&= \frac{n_1^2}{(n-n_1)(m-m_1)}\frac{\mathbb{E}_{P_n}[\bar{k}(X_1,X_3)^2\bar{k}(X_2,X_3)^2]}{\mathbb{E}_{P_n}[\bar{k}(X_1,X_2)^2]^2} \\
&= \frac{\sum_l \lambda_l^4}{(\sum_l \lambda_l^2)^2} \leq \frac{\lambda_1^2}{\sum_{l'}\lambda_{l'}^2}\sum_l \frac{\lambda_l^2}{\sum_{l'}\lambda_{l'}^2} = \frac{\lambda_1^2}{\sum_{l'}\lambda_{l'}^2} \to \rho_1 = 0.
\end{aligned}
$$

This implies that the term R is $o_P(1)$, and hence we have

$$
(A_2)^{1/2} = \frac{1}{2+o_P(1)} = \mathcal{O}_P(1),
$$

as required. This completes the proof. □

Next, we show that we can use Lemma 16 to obtain the limiting distribution of the standardized statistic $T_s = \frac{S_X - S_Y}{\sqrt{n_1^{-1}\mathbb{E}[W_1^2|\mathbb{X}_2,\mathbb{Y}_2] + m_1^{-1}\mathbb{E}[Z_1^2|\mathbb{X}_2,\mathbb{Y}_2]}}$.

Lemma 17. *Under the conditions of Theorem 15, the standardized statistic T_s converges in distribution to $N(0,1)$.*

Proof. This statement simply follows from the observation that $\mathbb{E}_2[Z_1^2] = \mathbb{E}_2[W_1^2]$ almost surely under the null hypothesis. Then, the term $\alpha_n := (\sqrt{n_1^{-1}\mathbb{E}_2[W_1^2]})/(\sqrt{n_1^{-1}\mathbb{E}_2[W_1^2] + m_1^{-1}\mathbb{E}_2[Z_1^2]}) = \sqrt{1/(1+n_1m_1^{-1})}$ converges to a constant (say $\alpha \in (0,1)$).

Using the result of Lemma 16, we can then conclude that $\alpha_n T_{s,X} \xrightarrow{d} N(0, \alpha^2)$ and $\sqrt{1-\alpha_n^2} T_{s,Y} \xrightarrow{d} N(0, 1-\alpha^2)$. This implies, due to Lévy's continuity theorem (Durrett, 2019, Theorem 3.3.17. (i)), the pointwise convergence of the characteristic functions of these sequences. In particular, let $\psi_{n,X}$ and $\psi_{n,Y}$ denote the characteristic functions of $\alpha_n T_{s,X}$ and $\sqrt{1-\alpha_n^2} T_{s,Y}$ respectively. Then, due to the conditional independence of $T_{s,X}$ and $T_{s,Y}$ given $(\mathbb{X}_2, \mathbb{Y}_2)$, we note that the characteristic function of $T_s = \alpha_n T_{s,X} = \sqrt{1-\alpha_n^2} T_{s,Y}$, denoted by $\psi_n(t)$, satisfies

$$\begin{aligned}
\psi_n(t) &\coloneqq \mathbb{E}_{P_n}[\exp(itT_s) \,|\mathbb{X}_2, \mathbb{Y}_2] \\
&= \mathbb{E}_{P_n}[\exp(it\,\alpha_n T_{s,X}) \,|\mathbb{X}_2, \mathbb{Y}_2] \times \mathbb{E}_{P_n}\left[\exp\left(-it\sqrt{1-\alpha_n^2} T_{s,Y}\right) |\mathbb{X}_2, \mathbb{Y}_2\right] \\
&= \psi_{n,X}(t) \times \psi_{n,Y}(-t).
\end{aligned}$$

Now, taking the limit $n \to \infty$, we get that

$$\begin{aligned}
\lim_{n\to\infty} \psi_n(t) &= \lim_{n\to\infty} \psi_{n,X}(t) \times \psi_{n,Y}(-t) \\
&= \exp\left(-\frac{1}{2}\left(\alpha^2 t^2\right)\right) \times \exp\left(-\frac{1}{2}\left((1-\alpha^2)t^2\right)\right) \\
&= \exp\left(-\frac{t^2}{2}\right).
\end{aligned}$$

Thus, we have shown that conditioned on $(\mathbb{X}_2, \mathbb{Y}_2)$, the characteristic function, ψ_n of T_s converges pointwise to the characteristic function of a $N(0,1)$ distribution. Hence, by the other direction of Lévy's continuity theorem (Durrett, 2019, Theorem 3.3.17. (ii)), we conclude that $T_s \xrightarrow{d} N(0,1)$.

Finally, we pass from the conditional statement to the unconditional one by noting that $T_s \xrightarrow{d} N(0,1)$ conditioned on $(\mathbb{X}_2, \mathbb{Y}_2)$ implies that $\sup_{x\in\mathbb{R}} |\,\mathbb{P}_{P_n}(T_s \le x) - \Phi(x)| \xrightarrow{p} 0$, because the $N(0,1)$ distribution is continuous. This fact, coupled with the boundedness of $\sup_{x\in\mathbb{R}} |\,\mathbb{P}_{P_n}(T_s \le x) - \Phi(x)|$ implies that it also converges in expectation, as required. Thus, we have shown that the limiting distribution of the standardized statistic T_s is $N(0,1)$ unconditionally. □

We now prove that the studentized statistic also has the same limiting distribution as the standardized statistic T_s by appealing to Slutsky's theorem and the continuous mapping theorem.

Lemma 18. *The ratio of $\widehat{\sigma}^2$ and the conditional variance $n_1^{-1}\mathbb{E}_2[W_1^2] + m_1^{-1}\mathbb{E}_2[Z_1^2]$ converges in probability to 1. Stated formally,*

$$\frac{n_1^{-1}\widehat{\sigma}_X^2 + m_1^{-1}\widehat{\sigma}_Y^2}{n_1^{-1}\mathbb{E}_2[W_1^2] + m_1^{-1}\mathbb{E}_2[Z_1^2]} \xrightarrow{p} 1.$$

Recall that we use the notation $\mathbb{E}_2[\cdot]$ to denote the conditional expectation on the second half of the data, i.e., $\mathbb{E}_{P_n}[\cdot|\mathbb{X}_2, \mathbb{Y}_2]$.

Proof. Since $\mathbb{E}_2[W_1^2] = \mathbb{E}_2[Z_1^2]$ almost surely, it suffices to show the following two statements to conclude the result:

$$\frac{\widehat{\sigma}_X^2}{\mathbb{E}_2[W_1^2]} \xrightarrow{p} 1, \quad \text{and} \quad \frac{\widehat{\sigma}_Y^2}{\mathbb{E}_2[Z_1^2]} \xrightarrow{p} 1.$$

We provide the details of the first statement, since the second can be obtained similarly. Consider the following:

$$\begin{aligned}
\frac{(n_1-1)^{-1}\sum_{i=1}^{n_1}(W_i - \bar{U}_X)^2 - \mathbb{E}_2[W_1^2]}{\mathbb{E}_2[W_1^2]} &= \frac{\sum_{i=1}^{n_1}(W_i - \bar{U}_X)^2 - (n_1-1)\mathbb{E}_2[W_1^2]}{\mathbb{E}_{P_n}[\bar{k}^2(X_1, X_2)]} \times \frac{\mathbb{E}_{P_n}[\bar{k}^2(X_1, X_2)]}{(n_1-1)\mathbb{E}_2[W_1^2]} \\
&= C_1 \times C_2.
\end{aligned}$$

Note that $C_2 = \frac{n_1}{n_1-1}\sqrt{A_2}$, where A_2 was introduced in (20) and shown to be $O_P(1)$ in the proof of Lemma 16. Hence, to complete the proof, we will show that $C_1 \xrightarrow{p} 0$. This can be concluded by

noting that $\mathbb{E}_{P_n}[C_1] = 0$, and that the variance of C_1 satisfies:

$$\begin{aligned}
\mathbb{V}_{P_n}[C_1] &= \mathbb{E}_{P_n}[\mathbb{V}_{P_n}[C_1|\mathbb{X}_2, \mathbb{Y}_2]] + \mathbb{V}_{P_n}[\mathbb{E}_{P_n}[C_1|\mathbb{X}_2, \mathbb{Y}_2]] \\
&= \frac{(n_1-1)^2}{\mathbb{E}_{P_n}[\bar{k}^2(X_1, X_2)]^2} \mathbb{E}_{P_n}\left[\mathbb{V}_{P_n}\left[\frac{1}{n_1-1}\sum_{i=1}^{n_1}(W_i - \bar{U}_X)^2\right]\right] \\
&\leq \frac{(n_1-1)^2}{\mathbb{E}_{P_n}[\bar{k}^2(X_1, X_2)]^2} \frac{\mathbb{E}_{P_n}[W_1^4]}{n_1} \leq \frac{n_1 \mathbb{E}_{P_n}[W_1^4]}{\mathbb{E}_{P_n}[\bar{k}^2(X_1, X_2)]^2} \leq 16 \frac{n_1 \mathbb{E}_{P_n}[a_1^4 + b_1^4]}{\mathbb{E}_{P_n}[\bar{k}^2(X_1, X_2)]^2} \\
&= 16\,(A_1 + B_1)\,,
\end{aligned}$$

where the terms A_1 and B_1 were introduced in (19). As mentioned during the proof of Lemma 16, both of these terms can be shown to converge in probability to 0 as required. □

The previous three lemmas prove that for any sequence $\{P_n : n \geq 1\}$ with $P_n \in \mathcal{P}_n^{(0)}$, we have $\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |\mathbb{P}_{P_n}\left(\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2 \leq x\right) - \Phi(x)| = 0$. This is sufficient to conclude the uniform result

$$\lim_{n\to\infty} \sup_{P_n \in \mathcal{P}_n^{(0)}} \sup_{x\in\mathbb{R}} |\mathbb{P}_{P_n}\left(\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2 \leq x\right) - \Phi(x)| = 0.$$

This is because we can select a sequence P_n' such that for all n, we have

$$\begin{aligned}
\sup_{x\in\mathbb{R}} |\mathbb{P}_{P_n'}\left(\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2\right) - \Phi(x)| &\leq \sup_{P_n \in \mathcal{P}_n^{(0)}} \sup_{x\in\mathbb{R}} |\mathbb{P}_{P_n}\left(\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2\right) - \Phi(x)| \\
&\leq \sup_{x\in\mathbb{R}} |\mathbb{P}_{P_n'}\left(\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2\right) - \Phi(x)| + \frac{1}{n}.
\end{aligned}$$

Since the left and right terms converge to zero, it follows that the middle term does too, as required. This completes the proof of Theorem 15.

B.2 Fixed P, changing k_n (Theorem 5)

We note that the statement of Theorem 5 requires an additional technical assumption on the eigenvalues of the kernel operator, introduced in (15). We repeat the statement of Theorem 5 with this additional requirement below.

Theorem 5'. Suppose P is fixed, but the kernel k_n changes with n. If

$$\lim_{n\to\infty} \frac{\mathbb{E}_P[\bar{k}_n(X_1, X_2)^4]}{\mathbb{E}_P[\bar{k}_n(X_1, X_2)^2]^2}\left(\frac{1}{n} + \frac{1}{m_n}\right) = 0, \quad \text{and} \quad \lim_{n\to\infty} \frac{\lambda_{1,n}^2}{\sum_{l=1}^{\infty} \lambda_{l,n}^2} \text{ exists,} \tag{22}$$

then we have $\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2 \xrightarrow{d} N(0,1)$.

Proof. The proof of this statement will follow the general outline of the proof of Theorem 15. However, in this special case when P is fixed, we can remove the condition that $\lim_{n\to\infty} m_n/n$ exists and is non-zero, that is required by Theorem 15.

We will carry over the notations used in the proof of Theorem 15, and in particular, we will use $\bar{U}_X = \frac{1}{n_1}\sum_{i=1}^{n_1} W_i$ and $\bar{U}_Y = \frac{1}{m_1}\sum_{j=1}^{m_1} Z_j$. Since W_i and Z_j are identically distributed under the null, we have $\mathbb{E}_P[W_i^2|\mathbb{X}_2, \mathbb{Y}_2] = \mathbb{E}_P[Z_j^2|\mathbb{X}_2, \mathbb{Y}_2]$, and we will use σ_2^2 to denote this conditional variance. Then, note the following:

$$\begin{aligned}
\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2 &= \frac{\bar{U}_X - \bar{U}_Y}{\widehat{\sigma}} = \frac{\bar{U}_X - \bar{U}_Y}{\sigma_2\left(\sqrt{n_1^{-1} + m_1^{-1}}\right)} \times \frac{\sigma_2\left(\sqrt{n_1^{-1} + m_1^{-1}}\right)}{\widehat{\sigma}} \\
&:= T_1 \times T_2.
\end{aligned} \tag{23}$$

To complete the proof, we will show that $T_1 \xrightarrow{d} N(0,1)$ and $T_2 \xrightarrow{p} 1$. The result then follows by an application of Slutsky's theorem.

First, we consider the term T_1 in (23). Let $\widetilde{W}_i := W_i/\sigma_2$ and $\widetilde{Z}_j := Z_j/\sigma_2$. Then, conditioned on $(\mathbb{X}_2, \mathbb{Y}_2)$, the terms $\widetilde{W}_i$ and $\widetilde{Z}_j$ are independent and identically distributed. Introducing the constants $u_i = \sqrt{\frac{m_1}{n_1(m_1+n_1)}}$ and $v_j = \sqrt{\frac{n_1}{m_1(m_1+n_1)}}$, we can write

$$T_1 = \sum_{i=1}^{n_1} u_i \widetilde{W}_i - \sum_{j=1}^{m_1} v_j \widetilde{Z}_j.$$

We can check that the constants (u_i) and (v_j) satisfy the property:

$$\lim_{n\to\infty} \max_{i,j} \frac{u_i^2 + v_j^2}{\sum_{i'=1}^{n_1} u_{i'}^2 + \sum_{j'=1}^{m_1} v_{j'}^2} \leq \lim_{n\to\infty} \max_{i,j} \frac{1}{m_1} + \frac{1}{n_1} = 0.$$

Thus, by an application of Lindeberg's CLT, we note that $T_1 \xrightarrow{d} N(0,1)$ conditioned on $(\mathbb{X}_2, \mathbb{Y}_2)$. Since the limiting distribution (in this case, standard normal) is continuous, this also means that the T_1 converges to $N(0,1)$ in the Kolmogorov-Smirnov metric, that is, $\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |\mathbb{P}_P(T_1 \leq x|\mathbb{X}_2, \mathbb{Y}_2) - \Phi(x)| \xrightarrow{p} 0$. Since the random variable $\sup_{x\in\mathbb{R}} |\mathbb{P}_P(T_1 \leq x|\mathbb{X}_2, \mathbb{Y}_2) - \Phi(x)|$ is bounded, convergence in probability implies that $\lim_{n\to\infty} \mathbb{E}_P[\sup_{x\in\mathbb{R}} |\mathbb{P}_P(T_1 \leq x|\mathbb{X}_2, \mathbb{Y}_2) - \Phi(x)|] = 0$, which in turn implies that $\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |\mathbb{E}_P[\mathbb{P}_P(T_1 \leq x|\mathbb{X}_2, \mathbb{Y}_2) - \Phi(x)]| = 0$, as required.

We now consider the second term, T_2, in (23). It remains to show that $T_2 \xrightarrow{p} 1$. We will show that $1/T_2^2 - 1 \xrightarrow{p} 0$, and the result will follow by an application of the continuous mapping theorem.

$$\left|\frac{1}{T_2^2} - 1\right| = \left|\frac{\frac{\widehat{\sigma}_X^2}{n_1} + \frac{\widehat{\sigma}_Y^2}{m_1}}{\sigma_2^2\left(\frac{1}{n_1} + \frac{1}{m_1}\right)} - 1\right| \leq \left|\frac{\widehat{\sigma}_X^2}{\sigma_2^2} - 1\right| + \left|\frac{\widehat{\sigma}_Y^2}{\sigma_2^2} - 1\right|. \tag{24}$$

Thus, it suffices to show that both terms in (24) converge in probability to 0. This is exactly the result that is proved in Lemma 18 under the two conditions listed in Assumption 1. The condition on eigenvalues is already assumed in the statement of Theorem 5', and thus we will show that the condition on the kernels, stated in (22), implies the condition (17). To prove this, we first, we note that

$$\begin{aligned}\mathbb{E}_P\left[\bar{k}_n(X_1, X_2)^2 \bar{k}_n(X_1, X_3)^2\right] &\leq \mathbb{E}_P\left[\bar{k}_n(X_1, X_2)^4\right]^{1/2} \mathbb{E}_P\left[\bar{k}_n(X_1, X_3)^4\right]^{1/2} \\ &= \mathbb{E}_P\left[\bar{k}_n(X_1, X_2)^4\right].\end{aligned}$$

Thus, the term in (17) is upper bounded by

$$\frac{\mathbb{E}_P\left[\bar{k}_n(X_1, X_2)^4\right]}{\mathbb{E}_P\left[\bar{k}_n(X_1, X_2)^2\right]^2}\left(\frac{1}{n} + \frac{1}{m_n}\right)\left(1 + \frac{1}{n} + \frac{1}{m_n}\right).$$

Since, we have assumed that $\lim_{n\to\infty} m_n \to \infty$, there exists and n_0, such that for all $n \geq n_0$, $1 + \frac{1}{n} + \frac{1}{m_n} \leq 2$. This implies that if (22) is satisfied, then (17) in Assumption 1 is also satisfied, as required. □

B.3 Fixed k, and fixed P (Theorem 4)

We prove Theorem 4 by showing that under the bounded fourth moment assumption on $\bar{k}$, both the conditions required by Theorem 5' are satisfied.

Note that since $\mathbb{E}_P[\bar{k}(X_1, X_2)] = 0$, the positive and finite fourth moment also implies that the second moment of $\bar{k}(X_1, X_2)$ is also positive and finite. Hence, we have that

$$\frac{\mathbb{E}_P[\bar{k}(X_1, X_2)^4]}{\mathbb{E}_P[\bar{k}(X_1, X_2)^2]^2} < \infty.$$

This, in turn, implies

$$\lim_{n\to\infty} \frac{\mathbb{E}_P[\bar{k}(X_1, X_2)^4]}{\mathbb{E}_P[\bar{k}(X_1, X_2)^2]^2}\left(\frac{1}{n} + \frac{1}{m_n}\right) = 0,$$

as required by Theorem 5.

For the second part of the condition, we note that as kernel k and probability distribution P are fixed, the term $\frac{\lambda_1^2}{\sum_l \lambda_l^2}$ doesn't change with n, and hence its limit exists. Thus, both the conditions for Theorem 5' are satisfied, as required.

C Consistency against fixed and local alternatives (Section 4)

C.1 Proof of Theorem 8 (General conditions for consistency)

Proof. We begin by noting that

$$\mathbb{E}_{P_n,Q_n}[1-\Psi(\mathbb{X},\mathbb{Y})] = \mathbb{P}_{P_n,Q_n}\left(\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2 \leq z_{1-\alpha}\right) = \mathbb{P}_{P_n,Q_n}\left(\widehat{\mathrm{cMMD}}^2 \leq z_{1-\alpha}\widehat{\sigma}\right).$$

Now, introduce the event $\mathcal{E} = \{\widehat{\sigma}^2 \leq \mathbb{E}[\widehat{\sigma}^2]/\delta_n\}$, where (δ_n) is a positive sequence converging to zero. By an application of Markov's inequality, we have $\mathbb{P}_{P_n,Q_n}(\mathcal{E}^c) \leq \delta_n$, which implies that

$$\begin{aligned}
\mathbb{P}_{P_n,Q_n}\left(\widehat{\mathrm{cMMD}}^2 \leq z_{1-\alpha}\sqrt{\widehat{\sigma}^2}\right) &= \mathbb{P}_{P_n,Q_n}\left(\{\widehat{\mathrm{cMMD}}^2 \leq z_{1-\alpha}\sqrt{\widehat{\sigma}^2}\} \cap \mathcal{E}\right) \\
&\quad + \mathbb{P}_{P_n,Q_n}\left(\{\widehat{\mathrm{cMMD}}^2 \leq z_{1-\alpha}\sqrt{\widehat{\sigma}^2}\} \cap \mathcal{E}^c\right) \\
&\leq \mathbb{P}_{P_n,Q_n}\left(\widehat{\mathrm{cMMD}}^2 \leq z_{1-\alpha}\sqrt{\mathbb{E}_{P_n,Q_n}[\widehat{\sigma}^2]/\delta_n}\right) + \mathbb{P}_{P_n,Q_n}(\mathcal{E}^c) \\
&\leq \mathbb{P}_{P_n,Q_n}\left(\widehat{\mathrm{cMMD}}^2 \leq z_{1-\alpha}\sqrt{\mathbb{E}_{P_n,Q_n}[\widehat{\sigma}^2]/\delta_n}\right) + \delta_n. \qquad (25)
\end{aligned}$$

By the assumption that $\delta_n \to 0$, it suffices to show that the worst-case value of the first term in (25) converges to zero to complete the proof.

To do this, we observe that (7) implies that there exists a finite value of n, say n_0, such that for all $n \geq n_0$ and $m \geq m_{n_0}$, we have

$$\sup_{(P_n,Q_n)\in\mathcal{P}_n^{(1)}} \frac{\mathbb{E}_{P_n,Q_n}[\widehat{\sigma}^2]}{\gamma_n^4\delta_n} \leq \frac{1}{4z_{1-\alpha}^2},$$

which implies that $z_{1-\alpha}\sqrt{\mathbb{E}_{P_n,Q_n}[\widehat{\sigma}^2]/\delta_n} \leq \gamma_n^2/2$. Furthermore, since $\widehat{\mathrm{cMMD}}^2 = \langle \widehat{\mu}_1 - \widehat{\nu}_1, \widehat{\mu}_2 - \widehat{\nu}_2\rangle_k$, it follows that $\mathbb{E}_{P_n,Q_n}[\widehat{\mathrm{cMMD}}^2] = \gamma_n^2$. Combining these two observations, we get for all $n \geq n_0$:

$$\begin{aligned}
\mathbb{P}_{P_n,Q_n}\left(\widehat{\mathrm{cMMD}}^2 \leq z_{1-\alpha}\sqrt{\mathbb{E}_{P_n,Q_n}[\widehat{\sigma}^2]/\delta_n}\right) &\leq \mathbb{P}_{P_n,Q_n}\left(\widehat{\mathrm{cMMD}}^2 - \mathbb{E}_{P_n,Q_n}[\widehat{\mathrm{cMMD}}^2] \leq \frac{\gamma_n^2}{2} - \gamma_n^2\right) \\
&\overset{(i)}{\leq} 4\frac{\mathbb{V}_{P_n,Q_n}(\widehat{\mathrm{cMMD}}^2)}{\gamma_n^4},
\end{aligned}$$

where `(i)` follows from Chebychev's inequality. This implies that

$$\sup_{(P_n,Q_n)\in\mathcal{P}_n^{(1)}} \mathbb{P}_{P_n,Q_n}\left(\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2 < z_{1-\alpha}\right) \leq \sup_{(P_n,Q_n)\in\mathcal{P}_n^{(1)}} 4\frac{\mathbb{V}_{P_n,Q_n}(\bar{U})}{\gamma_n^4}.$$

The required conclusion that $\sup_{(P_n,Q_n)\in\mathcal{P}_n^{(1)}} \mathbb{P}_{P_n,Q_n}(\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2 \leq z_{1-\alpha}) \to 0$ now follows from the second term in (7). □

C.2 Proof of Theorem 7 (Consistency against fixed alternative)

We prove Theorem 7 by showing that the sufficient conditions for consistency, as derived in Theorem 8, are satisfied under the assumptions of Theorem 7.

First, since the kernel is assumed to be characteristic, and $P_n = P \neq Q = Q_n$, it means that the kernel-MMD distance between P and Q must be strictly positive. In other words, we have $\gamma_n = \mathrm{MMD}(P, Q) \coloneqq \gamma > 0$ for all $n \geq 1$. Hence, in order to verify the condition (7), it suffices to show that the following two properties hold:

$$\lim_{n\to\infty} \mathbb{E}_{P,Q}[\widehat{\sigma}^2] = \lim_{n\to\infty} \frac{2}{n}\mathbb{E}_{P,Q}[\widehat{\sigma}_X^2] + \frac{2}{m_n}\mathbb{E}_{P,Q}[\widehat{\sigma}_Y^2] = 0, \quad \text{and} \tag{26}$$

$$\lim_{n\to\infty} \mathbb{V}_{P,Q}\left(\widehat{\mathrm{cMMD}}^2\right) = 0. \tag{27}$$

In the equality in (26), we used the fact that $n_1 = n/2$ and $m_1 = m_n/2$ (see Remark 3).

Verifying (26)**.** We begin by noting that it suffices to show that $\mathbb{E}_{P,Q}[\widehat{\sigma}_X^2] < \infty$ and $\mathbb{E}_{P,Q}[\widehat{\sigma}_Y^2] < \infty$ to conclude (26) (this is because we have assumed in Remark 3 that $\lim_{n\to\infty} m_n = \infty$). We present the details for $\widehat{\sigma}_X^2$ as the same arguments can be used to conclude the result for $\widehat{\sigma}_Y^2$.

Recall that $\widehat{\sigma}_X^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\langle k(X_i, \cdot), g_2\rangle_k - \bar{U}_X\right)^2$, where $g_2 = \widehat{\mu}_2 - \widehat{\nu}_2$. Since $X_1, \ldots, X_{n_1}$ are i.i.d., this implies that

$$\begin{aligned}
\mathbb{E}_{P,Q}[\widehat{\sigma}_X^2] &= \mathbb{E}_{P,Q}\left[\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\langle k(X_i, \cdot), g_2\rangle_k - \bar{U}_X\right)^2\right] = \mathbb{E}_{P,Q}\left[\left(\langle k(X_1, \cdot), g_2\rangle_k - \bar{U}_X\right)^2\right] \\
&= \mathbb{E}_{P,Q}\left[\langle k(X_1, \cdot) - \widehat{\mu}_1, g_2\rangle_k^2\right] = \mathbb{E}_{P,Q}\left[\langle k(X_1, \cdot) - \widehat{\mu}_1, \widehat{\mu}_2 - \widehat{\nu}_2\rangle_k^2\right] && (28) \\
&\leq \mathbb{E}_{P,Q}\left[\|k(X_1, \cdot) - \widehat{\mu}_1\|_k^2 \|\widehat{\mu}_2 - \widehat{\nu}_2\|_k^2\right] && (29) \\
&\leq \mathbb{E}_{P,Q}\left[\|k(X_1, \cdot) - \widehat{\mu}_1\|_k^2\right] \mathbb{E}_{P,Q}\left[\|\widehat{\mu}_2 - \widehat{\nu}_2\|_k^2\right] && (30) \\
&\leq \left(2\mathbb{E}_{P,Q}\left[\|k(X_1, \cdot)\|_k^2 + \|\widehat{\mu}_1\|_k^2\right]\right) \times \left(2\mathbb{E}_{P,Q}\left[\|\widehat{\mu}_2\|_k^2 + \|\widehat{\nu}_2\|_k^2\right]\right) && (31) \\
&\leq \left(4\mathbb{E}_{P,Q}\left[k(X_1, X_1)\right]\right) \times \left(2\mathbb{E}_{P,Q}\left[k(X_2, X_2) + k(Y_1, Y_1)\right]\right) < \infty. && (32)
\end{aligned}$$

In the above display:
(28) uses the fact that $\widehat{\mathrm{cMMD}}_X^2 = \langle \widehat{\mu}_1, g_2\rangle_k = \langle \widehat{\mu}_1, \widehat{\mu}_2 - \widehat{\nu}_2\rangle_k$, and the linearity of inner product,
(29) uses the Cauchy–Schwarz inequality,
(30) uses the fact that the two terms inside the expectation are independent,
(31) uses the fact that $\|a - b\|_k^2 \leq (\|a\|_k + \|b\|_k)^2 \leq 2\left(\|a\|_k^2 + \|b\|_k^2\right)$, and
(32) uses the facts that $\|k(X_1, \cdot)\|_k^2 = k(X_1, X_1)$, $\mathbb{E}_{P,Q}[\|\widehat{\mu}_1\|_k^2] \leq \mathbb{E}_{P,Q}[k(X_1, X_1)]$, $\mathbb{E}_{P,Q}[\|\widehat{\mu}_2\|_k^2] \leq \mathbb{E}_{P,Q}[k(X_2, X_2)]$ for $X_2 \sim P$ independent of X_1 and $\mathbb{E}_{P,Q}[\|\widehat{\mu}_2\|_k^2]$ and $\mathbb{E}_{P,Q}[\|\widehat{\nu}_2\|_k^2] \leq \mathbb{E}_{P,Q}[k(Y_1, Y_1)]$ for $Y_1 \sim Q$. We show the details for the bound for $\mathbb{E}_{P,Q}[\|\widehat{\mu}_1\|_k^2]$ below:

$$\begin{aligned}
\mathbb{E}_{P,Q}[\|\widehat{\mu}_1\|_k^2] &= \mathbb{E}_{P,Q}\left[\frac{4}{n^2}\sum_{i=1}^{n/2}\sum_{l=1}^{n/2}\langle k(X_i, \cdot), k(X_l, \cdot)\rangle_k\right] \\
&\leq \frac{4}{n^2}\sum_{i=1}^{n/2}\sum_{l=1}^{n/2}\left(\mathbb{E}_{P,Q}[k(X_i, X_i)]\mathbb{E}_{P,Q}[k(X_l, X_l)]\right)^{1/2} \\
&= \mathbb{E}_{P,Q}[k(X_1, X_1)],
\end{aligned}$$

where the inequality follows from an application of Cauchy–Schwarz inequality. The bounds for $\mathbb{E}_{P,Q}[\|\widehat{\mu}_2\|_k^2]$ and $\mathbb{E}_{P,Q}[\|\widehat{\nu}_2\|_k^2]$ also follow from the same steps.

Thus, we have shown that $\mathbb{E}_{P,Q}[\widehat{\sigma}_X^2] < \infty$. The result for $\mathbb{E}_{P,Q}[\widehat{\sigma}_Y^2]$ follows in an analogous manner.

Verifying (27)**.** We begin by noting that the expected value of $\widehat{\mathrm{cMMD}}^2 = \langle \widehat{\mu}_1 - \widehat{\nu}_1, \widehat{\mu}_2 - \widehat{\nu}_2\rangle_k = \bar{U}_X - \bar{U}_Y$ is equal to $\mathrm{MMD}^2(P, Q) = \|\mu - \nu\|_k^2 = \gamma^2$. Thus, we have

$$\begin{aligned}
\mathbb{V}_{P,Q}(\widehat{\mathrm{cMMD}}^2) &= \mathbb{E}_{P,Q}\left[\left(\widehat{\mathrm{cMMD}}^2 - \langle \mu - \nu, \mu - \nu\rangle_k\right)^2\right] \\
&= \mathbb{E}_{P,Q}\left[\left(\left(\bar{U}_X - \langle \mu, \mu - \nu\rangle_k\right) - \left(\bar{U}_Y - \langle \nu, \mu - \nu\rangle_k\right)\right)^2\right] \\
&= 2\mathbb{E}_{P,Q}\left[\left(\bar{U}_X - \langle \mu, \mu - \nu\rangle_k\right)^2\right] + 2\mathbb{E}_{P,Q}\left[\left(\bar{U}_Y - \langle \nu, \mu - \nu\rangle_k\right)^2\right]. && (33)
\end{aligned}$$

We present the details for showing that the first term in (33) converges to 0 with n. The result for the second term can be proved similarly.

Before proceeding, we introduce some notation: we will use $\widetilde{\mu}_1$ to denote $\widehat{\mu}_1 - \mu$, the centered version of $\widehat{\mu}_1$. Similarly, we will use $\widetilde{\mu}_2$, $\widetilde{\nu}_1$, $\widetilde{\nu}_2$ and $\widetilde{g}_2$ to represent $\widehat{\mu}_2 - \mu$, $\widehat{\nu}_1 - \nu$, $\widehat{\nu}_2 - \nu$ and $g_2 - (\mu - \nu)$ respectively. With these notations, note that we can write

$$\begin{aligned}\mathbb{E}_{P,Q}\left[\left(\bar{U}_X - \langle \mu, \nu - \mu \rangle_k\right)^2\right] &= \mathbb{E}_{P,Q}\left[\left(\langle \widetilde{\mu}_1, g_2 \rangle_k + \langle \mu, \widetilde{g}_2 \rangle_k\right)^2\right] \\ &\leq 2\mathbb{E}_{P,Q}\left[\langle \widetilde{\mu}_1, g_2 \rangle_k^2\right] + 2\mathbb{E}_{P,Q}\left[\langle \mu, \widetilde{g}_2 \rangle_k^2\right]. \end{aligned} \tag{34}$$

We now show that the first term of (34) is $\mathcal{O}(1/n)$.

$$\begin{aligned}\mathbb{E}_{P,Q}\left[\langle \widetilde{\mu}_1, g_2 \rangle_k^2\right] &\leq \mathbb{E}_{P,Q}\left[\|\widetilde{\mu}_1\|_k^2 \|\widehat{\mu}_2 - \widehat{\nu}_2\|_k^2\right] \leq \mathbb{E}_{P,Q}[\|\widetilde{\mu}_1\|_k^2]\mathbb{E}_{P,Q}\left[2\left(\|\widehat{\mu}_2\|_k^2 + \|\widehat{\nu}_2\|_k^2\right)\right] \\ &\leq \mathbb{E}_{P,Q}\left[\|\widetilde{\mu}_1\|_k^2\right]\left(2\mathbb{E}_{P,Q}[k(X_1, X_1)] + 2\mathbb{E}_{P,Q}[k(Y_1, Y_1)]\right) \qquad (35) \\ &= \mathcal{O}\left(\mathbb{E}_{P,Q}\left[\frac{4}{n^2}\sum_{i=1}^{n/2}\sum_{l=1}^{n/2}\langle \widetilde{k}(X_i, \cdot), \widetilde{k}(X_l, \cdot)\rangle_k\right]\right) \qquad (36) \\ &= \mathcal{O}\left(\mathbb{E}_{P,Q}\left[\frac{4}{n^2}\sum_{i=1}^{n/2}\langle \widetilde{k}(X_i, \cdot), \widetilde{k}(X_i, \cdot)\rangle_k\right]\right) \qquad (37) \\ &= \mathcal{O}\left(\frac{2}{n}\mathbb{E}_{P,Q}\left[k(X_1, X_1) - \|\mu\|_k^2\right]\right) = \mathcal{O}\left(\frac{1}{n}\right).\end{aligned}$$

In the above display:
(35) bounds $\mathbb{E}_{P,Q}[\|\widehat{\mu}_2\|_k^2]$ with $\mathbb{E}_{P,Q}[k(X_1, X_1)]$ and $\mathbb{E}_{P,Q}[\|\widehat{\nu}_2\|_k^2]$ with $\mathbb{E}_{P,Q}[k(Y_1, Y_1)]$ following the same argument as in (32).
(36) simply expands $\|\widetilde{\mu}_1\|_k^2$, and
(37) uses the fact that for $l \neq i$, we have $\mathbb{E}_{P,Q}[\langle \widetilde{k}(X_i, \cdot), \widetilde{k}(X_l, \cdot)\rangle_k h] = 0$.

We next show that the second term in (34) is $\mathcal{O}(1/n + 1/m_n)$.

$$\begin{aligned}\mathbb{E}_{P,Q}\left[\langle \mu, \widetilde{g}_2 \rangle_k^2\right] &\leq 2\mathbb{E}_{P,Q}[\|\mu\|_k^2]\left(\mathbb{E}_{P,Q}\left[\|\widetilde{\mu}_2\|_k^2 + \|\widetilde{\nu}\|_k^2\right]\right) \\ &\leq 2\mathbb{E}_{P,Q}[\|\mu\|_k^2]\left(\frac{2}{n}\mathbb{E}_{P,Q}[k(X_1, X_1) - \|\mu\|_k^2] + \frac{2}{m_n}\mathbb{E}_{P,Q}[k(Y_1, Y_2) - \|\nu\|_k^2]\right) \\ &= \mathcal{O}\left(\frac{1}{n} + \frac{1}{m_n}\right).\end{aligned}$$

Thus, since $\lim_{n\to\infty} m_n = \infty$, both the terms in (34) converge to 0 as n goes to infinity. This completes the proof that $\lim_{n\to\infty} \mathbb{E}_{P,Q}[(\bar{U}_X - \langle \mu, \mu - \nu \rangle_k)^2] = 0$. We can use the same arguments to show that $\lim_{n\to\infty} \mathbb{E}_{P,Q}[(\bar{U}_Y - \langle \nu, \mu - \nu \rangle_k)^2] = 0$. Together, these two statements imply that $\lim_{n\to\infty} \mathbb{V}_{P,Q}(\widehat{\mathrm{cMMD}}^2) = 0$ following (33).

C.3 Proof of Theorem 9 (Type-I error control and consistency against local alternative)

Type-I error bound. To obtain the bound on the type-I error, we verify the conditions required by Theorem 15, by using the expressions for moments of the Gaussian kernel derived by Li and Yuan (2019), and recalled in Fact 14.

First, we note that the scale parameters $s_n = n^{4/(d+4\beta)}$, satisfies the property:

$$\lim_{n\to\infty} \frac{s_n}{n^{4/d}} = \lim_{n\to\infty} n^{-\frac{4}{d}(1 - \frac{d}{d+4\beta})} = 0.$$

In other words, we have $s_n = o(n^{4/d})$. We now verify the required conditions:

- Since we have assumed $m_n = n$ in this case, $\lim_{n\to\infty} n/m_n = 1$ exists.

- For checking the condition on the eigenvalues, it suffices to show that
$$\lim_{n\to\infty} \frac{\mathbb{E}_{P_n,Q_n}[\mathbb{E}_{P_n,Q_n}[\bar{k}(X_1,X_2)\bar{k}(X_1,X_3)|X_2,X_3]^2]}{\mathbb{E}_{P_n,Q_n}[\bar{k}(X_1,X_2)^2]^2} = 0,$$
since this is equivalent to $\lim_{n\to\infty} \frac{\lambda_1^2}{\sum_l \lambda_l^2} = 0$. This result follows by a combination of (11) and (13).
- We next check the condition (17). We do this in two steps. First we consider the term,
$$\frac{\mathbb{E}_{P_n,Q_n}[\bar{k}_n(X_1,X_2)^4]}{\mathbb{E}_{P_n,Q_n}[\bar{k}_n(X_1,X_2)^2]^2 n^2} \lesssim \frac{s_n^{-d/2}}{(s_n^{-d/2})^2}\frac{1}{n^2} = \frac{s_n^{d/2}}{n^2} \to 0,$$
where the first inequality uses (11) and (12), while the last step uses the fact that $s_n = o(n^{4/d})$. Next, we consider the quantity
$$\frac{\mathbb{E}_{P_n,Q_n}[\bar{k}_n^2(X_1,X_2)\bar{k}_n^2(X_1,X_3)]}{n\mathbb{E}_{P_n,Q_n}[\bar{k}_n^2(X_1,X_2)]^2} \lesssim \frac{1}{n}\frac{s_n^{-3d/4}}{(s_n^{-d/2})^2} = \frac{s_n^{d/4}}{n} = \left(\frac{s_n}{n^{4/d}}\right)^{d/4} \to 0.$$

Together with Theorem 15, the above conditions imply that the statistic $\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2$ computed using Gaussian kernel with scale parameter $s_n = n^{4/(d+4\beta)}$ has a standard normal null distribution uniformly over the class $\mathcal{P}_n^{(0)}$. This implies the required result about asymptotic type-I error of the cMMD test Ψ.

Consistency. To prove the consistency results, we verify that the sufficient conditions established by the general result, Theorem 8, are satisfied by the Gaussian kernel with scale parameter $s_n = n^{4/(d+4\beta)}$.

We first check the condition on the variance of $\widehat{\mathrm{cMMD}}^2$. Note that we have the following:
$$\begin{aligned}\bar{U}_X &= \langle \widehat{\mu}_1, \widehat{\mu}_2 - \widehat{\nu}_2\rangle_k = \langle \widetilde{\mu}_1 + \mu, \widetilde{g}_2 + \mu - \nu\rangle_k \\ &= \langle \widetilde{\mu}_1, \widetilde{g}_2\rangle_k + \langle \widetilde{\mu}_1, \mu-\nu\rangle_k + \langle \mu, \widetilde{g}_2\rangle_k + \langle \mu, \mu-\nu\rangle_k\end{aligned}$$
Recall that we use $\widetilde{\mu}_1$ to denote $\widehat{\mu}_1 - \mu$, and similarly use $\widetilde{\mu}_2, \widetilde{\nu}_1, \widetilde{\nu}_2$ and $\widetilde{g}_2$ to denote $\widehat{\mu}_2 - \mu, \widehat{\nu}_1 - \nu, \widehat{\nu}_2 - \nu$ and $g_2 - (\mu - \nu)$ respectively. Similarly, on expanding the term $\bar{U}_Y$, we get
$$\begin{aligned}\bar{U}_Y &= \langle \widehat{\nu}_1, \widehat{\mu}_2 - \widehat{\nu}_2\rangle_k = \langle \widetilde{\nu}_1 + \nu, \widetilde{g}_2 + \mu - \nu\rangle_k \\ &= \langle \widetilde{\nu}_1, \widetilde{g}_2\rangle_k + \langle \widetilde{\nu}_1, \mu-\nu\rangle_k + \langle \nu, \widetilde{g}_2\rangle_k + \langle \nu, \mu-\nu\rangle_k\end{aligned}$$
Since $\widehat{\mathrm{cMMD}}^2 = \bar{U}_X - \bar{U}_Y$, we get that
$$\widehat{\mathrm{cMMD}}^2 = \langle \widetilde{\mu}_1 - \widetilde{\nu}_1, \widetilde{g}_2\rangle_k + \langle \widetilde{\mu}_1 - \widetilde{\nu}_1, \mu-\nu\rangle_k + \langle \mu-\nu, \widetilde{g}_2\rangle_k + \gamma_n^2.$$
Therefore, the variance of $\widehat{\mathrm{cMMD}}^2$ is
$$\mathbb{V}\left(\widehat{\mathrm{cMMD}}^2\right) = \mathbb{E}_{P_n,Q_n}\left[\langle \widetilde{\mu}_1 - \widetilde{\nu}_1, \widetilde{g}_2\rangle_k^2 + \langle \widetilde{\mu}_1 - \widetilde{\nu}_1, \mu-\nu\rangle_k^2 + \langle \mu-\nu, \widetilde{g}_2\rangle_k^2\right], \tag{38}$$
since all the cross terms are zero in expectation, due to the sample-splitting used in defining $\widehat{\mathrm{cMMD}}^2$. We now obtain upper bounds on the three terms in the right-hand-side of (38).
$$\begin{aligned}\mathbb{E}_{P_n,Q_n}\left[\langle \widetilde{\mu}_1 - \widetilde{\nu}_1, \widetilde{\mu}_2 - \widetilde{\nu}_2\rangle_k^2\right] &\le \mathbb{E}_{P_n,Q_n}\left[\|\widetilde{\mu}_1 - \widetilde{\nu}_1\|_k^2\right]\mathbb{E}_{P_n,Q_n}\left[\|\widetilde{\mu}_2 - \widetilde{\nu}_2\|_k^2\right] \\ &\le 4\left(\mathbb{E}_{P_n,Q_n}\left[\|\widetilde{\mu}_1\|_k^2\right] + \mathbb{E}_{P_n,Q_n}\left[\|\widetilde{\nu}_1\|_k^2\right]\right) \\ &\quad\times\left(\mathbb{E}_{P_n,Q_n}\left[\|\widetilde{\mu}_2\|_k^2\right] + \mathbb{E}_{P_n,Q_n}\left[\|\widetilde{\nu}_2\|_k^2\right]\right) \\ &= 4\left(\frac{\mathbb{E}_{P_n,Q_n}[\bar{k}(X,X)]}{n_1} + \frac{\mathbb{E}_{P_n,Q_n}[\bar{k}(Y,Y)]}{m_1}\right) \\ &\quad\times\left(\frac{\mathbb{E}_{P_n,Q_n}[\bar{k}(X,X)]}{n_2} + \frac{\mathbb{E}_{P_n,Q_n}[\bar{k}(Y,Y)]}{m_2}\right) \\ &\le \frac{32}{n^2}\left(\mathbb{E}_{P_n,Q_n}[\bar{k}(X,X)^2] + \mathbb{E}_{P_n,Q_n}[\bar{k}(Y,Y)^2]\right) \qquad (39) \\ &= \mathcal{O}\left(\frac{M s_n^{-d/2}}{n^2}\right). \qquad (40)\end{aligned}$$

In the above display, (39) follows uses Jensen's inequality, while (40) uses the upper bound on the second moment of $\bar{k}(X,X)$ and $\bar{k}(Y,Y)$ derived by Li and Yuan (2019), and recalled in (11) of Fact 14. For the second term in (38), we proceed as follows:

$$\begin{aligned}\mathbb{E}_{P_n,Q_n}\left[\langle\widetilde{\mu}_1-\widetilde{\nu}_1,\mu-\nu\rangle_k^2\right] &\leq 2\|\mu-\nu\|_k^2\mathbb{E}_{P_n,Q_n}\left[\|\widetilde{\mu}_1\|_k^2+\|\widetilde{\nu}_1\|_k^2\right]\\ &\leq \frac{\gamma_n^2}{n}\left(\sqrt{\mathbb{E}_{P_n,Q_n}[\bar{k}(X,X)^2]}+\sqrt{\mathbb{E}_{P_n,Q_n}[\bar{k}(Y,Y)^2]}\right)\\ &= \mathcal{O}\left(\frac{\gamma_n^2 s_n^{-d/4}}{n}\right). \end{aligned} \tag{41}$$

Similarly, we can get the same bound on the third term of (38)

$$\mathbb{E}_{P_n,Q_n}\left[\langle\widetilde{\mu}_2-\widetilde{\nu}_2,\mu-\nu\rangle_k^2\right] = \mathcal{O}\left(\frac{\gamma_n^2 s_n^{-d/4}}{n}\right). \tag{42}$$

Thus, combining (40) (41) and (42), we get that

$$\begin{aligned}\sup_{(P_n,Q_n)\in\mathcal{P}_n^{(1)}}\frac{\mathbb{V}_{P_n,Q_n}(\widehat{\mathrm{cMMD}}^2)}{\gamma_n^4} &\lesssim \frac{s_n^{-d/2}}{n^2\gamma_n^4}+\frac{s_n^{-d/4}}{n\gamma_n^2} \lesssim \frac{s_n^{-d/2}}{n^2 s_n^{-d}\Delta_n^4}+\frac{s_n^{-d/4}}{n s_n^{-d/2}\Delta_n^2}\\ &= \frac{s_n^{d/2}}{n^2\Delta_n^4}+\frac{s_n^{d/4}}{n\Delta_n^2}.\end{aligned} \tag{43}$$

The second inequality in (43) uses (14) that says $\gamma_n^2 \gtrsim s_n^{-d/2}\Delta_n^2$. Finally, using the fact that the scale parameter $s_n \asymp n^{4/(d+4\beta)}$, we get that

$$\lim_{n\to\infty}\sup_{(P_n,Q_n)\in\mathcal{P}_n^{(1)}}\frac{\mathbb{V}_{P_n,Q_n}(\widehat{\mathrm{cMMD}}^2)}{\gamma_n^4} \lesssim \lim_{n\to\infty}\left(\frac{1}{\left(n^{2\beta/(d+4\beta)}\Delta_n\right)^4}+\frac{1}{\left(n^{2\beta/(d+4\beta)}\Delta_n\right)^2}\right)=0,$$

where the equality follows from the condition imposed on Δ_n in the statement of Theorem 9. Thus, we have verified the condition on the variance of $\widehat{\mathrm{cMMD}}^2$ as required by (7).

It remains to verify the condition on the expected empirical variance in (7).

$$\begin{aligned}\mathbb{E}_{P_n,Q_n}\left[\widehat{\sigma}_X^2\right] &= \mathbb{E}_{P_n,Q_n}\left[\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\langle\widetilde{k}(X_i,\cdot),g_2\rangle_k-\langle\widetilde{\mu}_1,g_2\rangle_k\right)^2\right]\\ &= \mathbb{E}_{P_n,Q_n}\left[\langle\widetilde{k}(X_1,\cdot),g_2\rangle_k^2\right]\left(1-\frac{1}{n_1}\right)\\ &\leq \mathbb{E}_{P_n,Q_n}\left[\langle\widetilde{k}(X_1,\cdot),\widetilde{g}_2\rangle_k^2\right]+\mathbb{E}_{P_n,Q_n}\left[\langle\widetilde{k}(X_1,\cdot),\mu-\nu\rangle_k^2\right]\\ &\leq \mathbb{E}_{P_n,Q_n}[\bar{k}(X_1,X_1)]\left(\frac{\mathbb{E}_{P_n,Q_n}[\bar{k}(X_1,X_1)]}{n_2}+\frac{\mathbb{E}_{P_n,Q_n}[\bar{k}(Y_1,Y_1)]}{m_2}\right)+\gamma_n^2\mathbb{E}_{P_n,Q_n}[\bar{k}(X_1,X_1)]\\ &\lesssim \frac{s_n^{-d/2}}{n}+\gamma_n^2 s_n^{-d/4}.\end{aligned}$$

Similarly, we can get the same upper bound for the term $\mathbb{E}_{P_n,Q_n}[\widehat{\sigma}_Y^2]$. Since $\widehat{\sigma}^2 = n_1^{-1}\widehat{\sigma}_X^2+m_1^{-1}\widehat{\sigma}_Y^2$, we get that

$$\lim_{n\to\infty}\sup_{(P_n,Q_n)\in\mathcal{P}_n^{(1)}}\frac{\mathbb{E}_{P_n,Q_n}[\widehat{\sigma}^2]}{\gamma_n^4} \lesssim \lim_{n\to\infty}\frac{s_n^{-d/2}}{n^2\gamma_n^4}+\frac{s_n^{-d/4}}{n\gamma_n^2}.$$

We saw in (43) that this limit is equal to 0. Thus, the condition on $\widehat{\sigma}$ as required by (7) is also satisfied for $s_n \asymp n^{4/(d+4\beta)}$. Hence, by an application of Theorem 8, the test Ψ with Gaussian kernel and $s_n \asymp n^{4/(d+4\beta)}$ is consistent against the local alternatives with Δ_n satisfying $\lim_{n\to\infty}\Delta_n n^{2\beta/(d+4\beta)}=\infty$. This completes the proof.

D Gaussian Limit for General Two-Sample U-Statistic

We now generalize the asymptotic normality for kernel-MMD statistic stated in Theorem 15 to a larger class of two-sample U-statistics. As before, given $\mathbb{X} = (X_1, \ldots, X_n)$ and $\mathbb{Y} = (Y_1, \ldots, Y_m)$, we consider the two-sample U-statistic with arbitrary kernel h defined as

$$U = \frac{1}{\binom{n}{2}} \frac{1}{\binom{m}{2}} \sum_{i'<i} \sum_{j'<j} h(X_i, X_{i'}, Y_j, Y_{j'}).$$

We assume that h is a degenerate kernel, similar to the MMD case, and satisfies

$$\mathbb{E}_P[h(X, x', Y, y')] = \mathbb{E}_P[h(x, X', y, Y')] = 0,$$

when X, X', Y, Y' are i.i.d. random variables drawn from any distribution P.

With $\mathbb{X}_1 = (X_1, \ldots, X_{n_1})$ and $\mathbb{X}_2 = (X_{n_1+1}, \ldots, X_n)$ and $\mathbb{Y}_1 = (Y_1, \ldots, Y_{m_1})$ and $\mathbb{Y}_2 = (Y_{m_1+1}, \ldots, Y_m)$, we introduce the following terms:

$$\phi(x, y) \coloneqq \frac{1}{n_2} \frac{1}{m_2} \sum_{X_{i'} \in \mathbb{X}_2} \sum_{Y_{j'} \in \mathbb{Y}_2} h(x, X_{i'}, y, Y_{j'}), \quad \text{with } n_2 = n - n_1, \text{ and } m_2 = m - m_1 \tag{44}$$

$$q(x_1, x_2, y_2) \coloneqq \mathbb{E}[h(x_1, x_2, Y, y_2)] \quad \text{and} \quad \bar{q}(x) \coloneqq \frac{1}{n_2 m_2} \sum_{X_{i'} \in \mathbb{X}_2, Y_{j'} \in \mathbb{Y}_2} q(x, X_{i'}, Y_{j'}), \tag{45}$$

$$r(x_2, y_1, y_2) \coloneqq \mathbb{E}[h(X, x_2, y_1, y_2)] \quad \text{and} \quad \bar{r}(y) \coloneqq \frac{1}{n_2 m_2} \sum_{X_{i'} \in \mathbb{X}_2, Y_{j'} \in \mathbb{Y}_2} r(X_{i'}, y, Y_{j'}). \tag{46}$$

Using the above terms, we can now define the statistic $T = \bar{U}/\widehat{\sigma}$, with

$$\bar{U} = \frac{1}{n_1} \frac{1}{m_1} \sum_{X_i \in \mathbb{X}_1} \sum_{Y_j \in \mathbb{Y}_1} \phi(X_i, Y_j), \quad \text{and} \quad \widehat{\sigma}^2 = \frac{\widehat{\sigma}_X^2}{n_1} + \frac{\widehat{\sigma}_Y^2}{m_1}, \quad \text{where}$$

$$\widehat{\sigma}_X^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} \left(\bar{q}(X_i) - \frac{1}{n_1} \sum_{l=1}^{n_1} \bar{q}(X_l) \right)^2, \quad \widehat{\sigma}_Y^2 = \frac{1}{m_1} \sum_{j=1}^{m_1} \left(\bar{r}(Y_j) - \frac{1}{m_1} \sum_{l=1}^{m_1} \bar{r}(Y_l) \right)^2.$$

Remark 19. Note that the cross U-statistic written above corresponds exactly with the definition of the cross U-statistic for the kernel-MMD case in (2). To motivate the definitions of the empirical variance terms, note that in the case of kernel-MMD statistic, we have $h(x_1, x_2, y_1, y_2) = \langle k(x_1, \cdot) - k(y_1, \cdot), k(x_2, \cdot) - k(y_2, \cdot) \rangle_k$. We can check that in this case, we have $q(x_1, x_2, y_2) = \langle \widetilde{k}(x_1, \cdot), k(x_1, \cdot) - k(x_2, \cdot) \rangle_k$. This implies that $\bar{q}(X_i)$ equals the term W_i introduced (18), and thus $\frac{1}{n_1} \sum_{i=1}^{n_1} \bar{q}(X_i)$ is a centered analog of $\bar{U}_X$. Hence, the term $\widehat{\sigma}_X^2$ defined above reduces exactly to the $\widehat{\sigma}_X^2$ introduced in (4).

We next state the assumptions required to show the limiting Gaussian distribution of the statistic T when $\mathbb{X}$ and $\mathbb{Y}$ are drawn independently from the same distribution.

Assumption 2. *Let (h_n, P_n) be a sequence of kernel and probability distribution pairs, and let $\mathbb{X}$ and $\mathbb{Y}$ be two i.i.d. samples of sizes n and m_n respectively, drawn independently from P_n. With ϕ, $\bar{q}_n$ and $\bar{r}_n$ as defined in* (44)*,* (45) *and* (46) *respectively, we assume the following are true:*

$$\lim_{n \to \infty} \mathbb{E}_{P_n} \left[\frac{\mathbb{E}_{P_n}[\phi^2(X_1, Y_1) | \mathbb{X}_2, \mathbb{Y}_2]}{m_n \mathbb{E}_{P_n}[\bar{q}(X_1)^2 | \mathbb{X}_2, \mathbb{Y}_2] + n \mathbb{E}_{P_n}[\bar{r}(Y_1)^2 | \mathbb{X}_2, \mathbb{Y}_2]} \right] = 0, \quad \textit{and} \tag{47}$$

$$\lim_{n \to \infty} \mathbb{E}_{P_n} \left[\frac{1}{n} \frac{\mathbb{E}_{P_n}[\bar{q}^4(X_1) | \mathbb{X}_2, \mathbb{Y}_2]}{\mathbb{E}_{P_n}[\bar{q}^2(X_1) | \mathbb{X}_2, \mathbb{Y}_2]^2} + \frac{1}{m_n} \frac{\mathbb{E}_{P_n}[\bar{r}^4(Y_1) | \mathbb{X}_2, \mathbb{Y}_2]}{\mathbb{E}_{P_n}[\bar{r}^2(Y_1) | \mathbb{X}_2, \mathbb{Y}_2]^2} \right] = 0. \tag{48}$$

Remark 20. Note that in specific the case of kernel-MMD statistic, we can check that $\mathbb{E}_{P_n}[\phi(X_1, Y_1)^2 | \mathbb{X}_2, \mathbb{Y}_2] = \mathbb{E}_{P_n}[\bar{q}(X_1)^2 + \bar{r}(Y_1)^2 | \mathbb{X}_2, \mathbb{Y}_2]$. Hence (47) always holds. The second condition of Assumption 2, stated in (48), is a stronger version of the moment conditions used by Theorem 5 and Theorem 15.

We now state the main result of this section.

Theorem 21. *For every $n \geq 1$, let $\mathbb{X}$ and $\mathbb{Y}$ denote independent samples of sizes n and m_n respectively, drawn from a distribution P_n. Suppose the sample-sizes are such that $\lim_{n\to\infty} m_n/n$ exists and is non-zero. Let (h_n, P_n) denote a sequence satisfying the conditions of Assumption 2. Then, we have that*

$$\lim_{n,m\to\infty} \sup_{x\in\mathbb{R}} |\mathbb{P}_{P_n}(T \leq x) - \Phi(x)| = 0.$$

D.1 Proof of Theorem 21

Before describing the details, we first present the outline of the proof.

1. We first consider the standardized version of the statistic, defined as $T_s = \bar{U}/\sigma_P$, where $\sigma_P^2 = n_1^{-1}\mathbb{E}_{P_n}[\bar{q}(X_1)^2|\mathbb{X}_2, \mathbb{Y}_2] + m_1^{-1}\mathbb{E}_{P_n}[\bar{r}(Y_1)^2|\mathbb{X}_2, \mathbb{Y}_2]$. In Lemma 22, we show that the difference between T_s and its projected variant, $T_{P,s} = \bar{U}_P/\sigma_P = \left(n_1^{-1}\sum_i \bar{q}(X_i) + m_1^{-1}\sum_j \bar{r}(Y_j)\right)/\sigma_P$, converges in probability to 0. Hence, we can focus on the term $T_{P,s}$. This result uses the condition (47) of Assumption 2.
2. We then show in Lemma 23, that the statistic $T_{P,s}$ converges in distribution to $N(0,1)$. This combined with the previous result implies that $T_s \xrightarrow{d} N(0,1)$.
3. To complete the proof, we show in Lemma 24, that the ratio of the empirical variance $\widehat{\sigma}^2$ and the conditional variance σ_P^2 converge in probability to 1. This fact combined with the continuous mapping theorem and Slutsky's theorem implies the result. The proof of Lemma 24 relies on the condition (48) of Assumption 2.

We now present the details of the steps outlined above.

Consider the standardized statistic, T_s, defined as $\bar{U}/\sigma_P$, where $\sigma_P^2 = \mathbb{V}_{P_n}(\bar{U}_P|\mathbb{X}_2, \mathbb{Y}_2) = n_1^{-1}\mathbb{E}_{P_n}[\bar{q}^2(X_1)|\mathbb{X}_2, \mathbb{Y}_2] + m_1^{-1}\mathbb{E}_{P_n}[\bar{r}^2(Y_1)|\mathbb{X}_2, \mathbb{Y}_2] \coloneqq n_1^{-1}\sigma_{P,X}^2 + m_1^{-1}\sigma_{P,Y}^2$. Introduce the term $T_{P,s} = \frac{\bar{U}_P}{\sigma_P}$.

Lemma 22. *Under the conditions of Assumption 2, we have $T_p - T_{P,s} \xrightarrow{p} 0$.*

Proof. We first show that $T_s - T_{P,s} \xrightarrow{p} 0$, conditioned on the second half of the observations, $(\mathbb{X}_2, \mathbb{Y}_2)$. As a result of this, the conditional limiting distributions of the two random variables T_s and $T_{P,s}$ are the same. Since $\bar{U}_P$ is the projection of $\bar{U}$ on the sum on independent (conditioned on $(\mathbb{X}_2, \mathbb{Y}_2)$) random variables, we have

$$\begin{aligned}\mathbb{V}_{P_n}(T_s - T_{P,S}|\mathbb{X}, \mathbb{Y}_2) &= \mathbb{V}_{P_n}(T_s|\mathbb{X}, \mathbb{Y}_2) + \mathbb{V}_{P_n}(T_{P,s}|\mathbb{X}, \mathbb{Y}_2) - 2\mathbb{E}_{P_n}\left[(T_{P,s} + (T_s - T_{P,s}))\, T_{P,s}|\mathbb{X}_2, \mathbb{Y}_2\right] \\ &= \mathbb{V}_{P_n}(T_s|\mathbb{X}, \mathbb{Y}_2) - \mathbb{V}_{P_n}(T_{P,s}|\mathbb{X}, \mathbb{Y}_2) = \mathbb{V}_{P_n}(T_s|\mathbb{X}, \mathbb{Y}_2) - 1,\end{aligned}$$

using the fact that $(T_s - T_{P,s}) \perp T_{P,s}$ conditioned on $(\mathbb{X}_2, \mathbb{Y}_2)$. Next, using the formula for the variance of two-sample U-statistics, we have

$$\begin{aligned}\mathbb{V}_{P_n}(T_s|\mathbb{X}_2, \mathbb{Y}_2) &= \left(\frac{\sigma_{P,X}^2}{n_1} + \frac{\sigma_{P,Y}^2}{m_1} + \frac{1}{n_1 m_1}\mathbb{E}_{P_n}\left[\phi(X_1, X_2)^2|\mathbb{X}_2, \mathbb{Y}_2\right]\right)/\sigma_P^2 \\ &= 1 + \frac{1}{n_1 m_1}\frac{\mathbb{E}_{P_n}[\phi^2(X_1, X_2)|\mathbb{X}_2, \mathbb{Y}_2]}{\sigma_P^2}.\end{aligned}$$

The result then follows by an application of the condition (47) of Assumption 2, and the fact that $n_1 = n/2$ and $m_1 = m_n/2$. □

Our next result establishes the limiting distribution of the statistic $T_{P,s}$.

Lemma 23. *Under Assumption 2, we have $T_{P,s} \xrightarrow{d} N(0,1)$.*

Proof. Recall that $T_{P,s} = \bar{U}_P/\sigma_P$, where $\bar{U}_P \coloneqq \bar{U}_{P,X} - \bar{U}_{P,Y} = \frac{1}{n_1}\sum_{i=1}^{n_1}\bar{q}(X_i) - \frac{1}{m_1}\sum_{j=1}^{m_1}\bar{r}(Y_j)$, and $\sigma_P^2 = n_1^{-1}\sigma_{P,X}^2 + m_1^{-1}\sigma_{P,Y}^2$. Introduce the terms $T_X = \bar{U}_{P,X}/\sqrt{n_1^{-1}\sigma_{P,X}^2}$ and $T_Y = \bar{U}_{P,Y}/\sqrt{m_1^{-1}\sigma_{P,Y}^2}$. The result then follows in the following two steps:

- We first observe that T_X and T_Y conditioned on $(\mathbb{X}_2, \mathbb{Y}_2)$ converge in distribution to $N(0,1)$. The result follows by applying Lindeberg's CLT.
- Next, using the assumption that $\lim_{n\to\infty} m_n/n$ exists, and is non-zero, we next observe that $T_{P,s} \xrightarrow{d} N(0,1)$. The proof of this result follows from the same argument used in Lemma 17.

□

Together, the previous two lemmas imply that $T_s \xrightarrow{d} N(0,1)$. To complete the proof, we need to show that the ratio of the conditional variance σ_P^2, and the empirical variance $\widehat{\sigma}^2$ converge in probability to 1.

Lemma 24. *Under Assumption 2, we have* $\frac{\widehat{\sigma}^2}{\sigma_P^2} \xrightarrow{p} 1$.

Proof. We begin by noting the following

$$\begin{aligned}\frac{\widehat{\sigma}^2}{\sigma_P^2} - 1 &= \frac{n_1^{-1}\left(\widehat{\sigma}_X^2 - \sigma_{P,X}^2\right) m_1^{-1}\left(\widehat{\sigma}_Y^2 - \sigma_{P,Y}^2\right)}{\sigma_P^2} \\ &\leq \left|\frac{\widehat{\sigma}_X^2}{\sigma_{P,X}^2} - 1\right| + \left|\frac{\widehat{\sigma}_Y^2}{\sigma_{P,Y}^2} - 1\right|. \end{aligned} \tag{49}$$

Thus it suffices to show that the two terms in (49) converge in probability to 0. Since $n_1/(n_1-1)$ converges to 1, it suffices to consider

$$E := \frac{(n_1-1)^{-1}\sum_{i=1}^{n_1}\left(\bar{q}(X_i) - \bar{U}_{P,X}\right)^2 - \mathbb{E}_{P_n}[\bar{q}(X_1)|\mathbb{X}_2, \mathbb{Y}_2]}{\mathbb{E}_{P_n}[\bar{q}^2(X_1)|\mathbb{X}_2, \mathbb{Y}_2]}.$$

First note that $\mathbb{E}_{P_n}[E|\mathbb{X}_2, \mathbb{Y}_2] = 0$. Hence, its variance can be written as

$$\mathbb{V}_{P_n}(E) = \mathbb{E}_{P_n}[\mathbb{V}_{P_n}(E|\mathbb{X}_2, \mathbb{Y}_2)] \leq \frac{1}{n_1}\mathbb{E}_{P_n}\left[\frac{\mathbb{E}_{P_n}[\bar{q}^4(X_1)|\mathbb{X}_2, \mathbb{Y}_2]}{\mathbb{E}_{P_n}[\bar{q}^2(X_1)|\mathbb{X}_2, \mathbb{Y}_2]^2}\right]. \tag{50}$$

The last term in (50) converges to 0 by Assumption 2, implying that $\frac{\widehat{\sigma}_X^2}{\sigma_{P,X}^2}$ converges in the second moment to 1, which in turn implies their convergence in probability to 1. Following the same arguments, we can also show that $\frac{\widehat{\sigma}_Y^2}{\sigma_{P,Y}^2}$ also converge in probability to 1, as required. □

E Additional Experiments

Computing Infrastructure. All the experiments were performed on a workstation with Intel(R) Core(TM) i7-9700K CPU 3.60GHz and 32 GB of RAM with an NVIDIA GTX 1080 GPU.

E.1 Implementation details of experiments reported in the main text

Details for Figure 1. For the null distribution, we set $n = 500$ and $m = 625$ and generated both $\mathbb{X}$ and $\mathbb{Y}$ from $N(\mathbf{0}, I_d)$ for $d = 10$ and 100. In both cases, we computed the $\widehat{\overline{\mathrm{cMMD}}}^2$ statistic 2000 times to plot the histogram.

For the second figure, we obtain the power curves for the cMMD test and the MMD test with 200 permutations for testing $P = N(\mathbf{0}, I_d)$ againt $Q = N(a_{\epsilon,j}, I_d)$. Here $d = 10, j = 5$ and $\epsilon = 0.2$, and recall that $a_{\epsilon,j}$ is the vector in $\mathbb{R}^d$ obtained by setting the first $j \leq d$ coordinates of $\mathbf{0}$ equal to ϵ. We selected n and m from 20 equally spaced points in the intervals $[10, 400]$ and $[10, 500]$ respectively, and ran 200 trials of the tests for every (n, m) pair to obtain the power curves. The error regions in the figure correspond to one bootstrap standard deviation with 200 bootstrap samples.

For the third figure, we set $d = 100$, $j = 20$, $\epsilon = 0.1$, $P = N(\mathbf{0}, I_d)$ and $Q = N(a_{\epsilon,j}, I_d)$. We ran the two tests, cMMD and MMD with 200 permutations, for 20 different (n, m) pairs in the range $[10, 500]$, and repeated the experiment 200 times for every such pair. The figure plots the wall-clock time, measure by Python's `time.time()` function, and plot the power against the average wall-clock time over the 200 trials. The size of the marker is proportional to the sample size (i.e., $n + m$).

Details for Figure 2. The two kernels used in this figure are the Gaussian and Quadratic kernels. The Gaussian kernel with scale parameter $s > 0$ is defined as $k_s(x, y) = \exp(-s\|x - y\|_2^2)$, while the Quadratic kernel with scale $s > 0$ is defined as $k_Q(x, y) = \left(1 + s(x^T y)\right)^2$. With w denoting the median of the pairwise distance between all the observations, we set $s = 1/(2w^2)$ for the Gaussian kernel and $s = 1/w$ for the Quadratic kernel.

Details for Figure 3. Given observations $X_1, X_2, \ldots, X_n$i.i.d.P, consider the problem of one-sample mean-testing, that is, testing $H_0 : \mathbb{E}[X_i] = \mathbf{0}$ versus $H_1 : \mathbb{E}[X_i] = a \neq \mathbf{0}$. When the distribution P is a multivariate Gaussian, Kim and Ramdas (2020) showed that power of their test using a one-sample studentized U-statistic based on a bi-linear kernel is asymptotically $\Phi\left(z_\alpha + \frac{a^T a}{2\sqrt{\mathrm{tr}(\Sigma^2)}}\right)$. The power achieved by the test using the full U-statistic is $\Phi\left(z_\alpha + \frac{a^T a}{\sqrt{2\mathrm{tr}(\Sigma^2)}}\right)$, which differs from the previous expression by a factor of $\sqrt{2}$. A similar relation also holds for the problem of Gaussian covariance testing. Our heuristic in (9) is based on these two observations.

Details for Figure 4. For plotting the ROC curves, we proceed as follows. We fix $n = m = 200$, and then compute the MMD, block-MMD, linear-MMD and cross-MMD statistics for 1000 independent repetitions of 'null' and 'alternative' trials. For every null trial, we calculate all the statistics on independent samples of sizes n and m drawn from $P = N(\mathbf{0}, I_d)$, while for every alternative trial we calculate the statistics on independent samples of size n and m drawn from $P = N(\mathbf{0}, I_d)$ and $Q = N(a_{\epsilon,j}, I_d)$ respectively. Recall that $a_{\epsilon,j}$ is obtained by setting the first j coordinates of $\mathbf{0}$ equal to ϵ. Having obtained 2000 values for every statistic, we then plot the tradeoff between false positives (FP) and true positives (TP) as the rejection threshold is increased. The ability of a statistic to distinguish between the null and the alternative is quantified by the area under the curve. In Figure 4, we used $(d, j, \epsilon) \in \{(10, 5, 0.1), (100, 20, 0.1), (500, 100, 0.1)\}$.

E.2 Additional Figures

Null Distribution. Figure 5 denotes the null distribution of our proposed statistic ($\widehat{\overline{\mathrm{cMMD}}}^2$) along with that of the usual MMD normalized by its empirical standard deviation. The null distribution in Figure 5 is `Dirichlet` with parameter $2 \times \mathbf{1} \in \mathbb{R}^d$ for $d \in \{10, 500\}$.

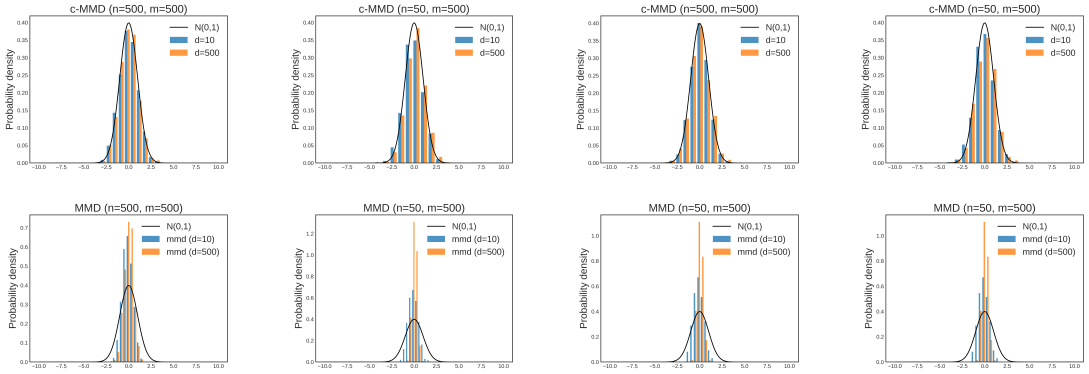

Figure 5: The first two columns show the null distribution of the $\widehat{\overline{\mathrm{cMMD}}}^2$ statistic (top row) and the $\widehat{\mathrm{MMD}}^2$ statistic scaled by its empirical standard deviation (bottom row) using the Gaussian kernel with scale-parameter chosen using the median heuristic. The last two columns show the null distribution for the two statistics using the Polynomial kernel of degree 5 with scale parameter chosen using the median heuristic. The figures demonstrate that the null distribution of $\widehat{\mathrm{MMD}}^2$ changes significantly with dimension (d), the ratio n/m and the choice of the kernel, unlike our proposed statistic.

Power Curves. In Figure 6, we plot the power curves for the different tests using a Gaussian Kernel, and we report the results of the same experiment with a polynomial kernel of degree 5 in Figure 7. Recall that the polynomial kernel of degree r and scale parameter $s > 0$ is defined as $k(x, y) = \left(1 + (x^T y)/s\right)^r$. In both instances, we selected the scale parameter using the median heuristic.

From the figure, we can see that the cMMD test is competitive with the computationally more costly tests, namely the MMD permutation test and the MMD-spectral test of Gretton et al. (2009). Furthermore, the performance of cMMD test is significantly better than the existing computationally efficient tests, namely block-MMD test (with block-size $\sqrt{n}$) and linear-MMD test.

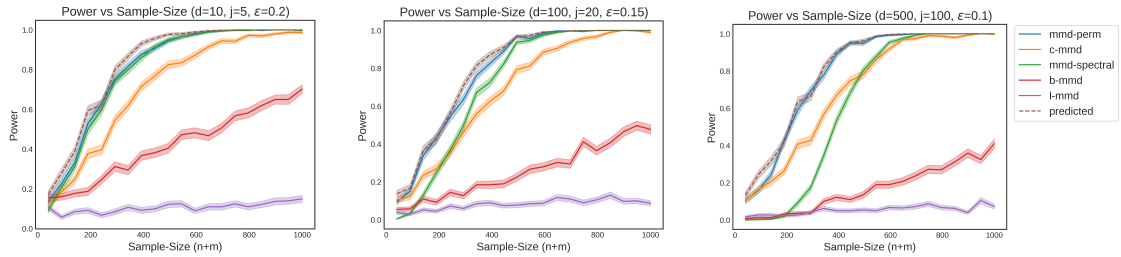

Figure 6: Power Curves for the different tests using Gaussian kernel with scale parameter chosen via median heuristic. The two distributions are $P = N(\mathbf{0}, I_d)$ and $Q = N(a_{\epsilon,j}, I_d)$ where $a_{\epsilon,j}$ is obtained by setting the first $j \leq d$ coordinates of $\mathbf{0} \in \mathbb{R}^d$ equal to ϵ. The figures demonstrate that the cMMD test is competitive with more computationally expensive tests (MMD-perm and MMD-spectral), while performing significantly better than the low complexity alternatives (b-MMD and l-MMD). The batch-size used in the b-MMD test was $\sqrt{n}$.

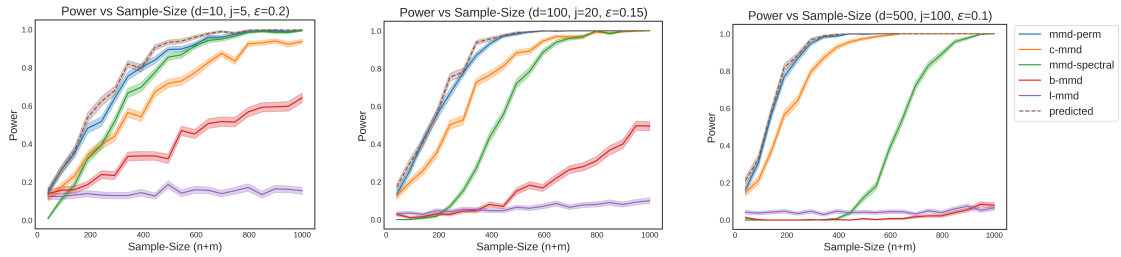

Figure 7: Power curves in the same setting as Figure 6, but with a polynomial kernel of degree 5, i.e., $k(x, y) = \left(1 + (x^T y)/s\right)^5$ with s chosen via the median heuristic.

ROC curves. In Figure 8, we plot some additional ROC curves for the different statistics. As before, we used 1000 'null trials' and another 1000 'alternative trials' with sample sizes $n = 200$ and $m = 200$. The data generating distributions P and Q were both Dirichlet with parameters $\mathbf{1} \in \mathbb{R}^d$ and $(1 + \epsilon) \times \mathbf{1} \in \mathbb{R}^d$ for $(d, \epsilon) \in \{(10, 0.4), (100, 0.2), (500, 0.15)\}$.

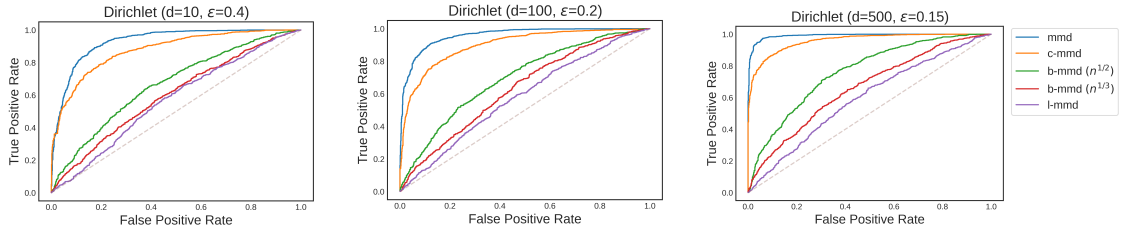

Figure 8: ROC curves using the different statistics with Gaussian kernel for testing two Dirichlet distributions in dimensions $d \in \{10, 100, 500\}$ with sample-size $n = m = 200$. The two distributions are $P = \texttt{Dirichlet}(\mathbf{1})$ and $Q = \texttt{Dirichlet}((1 + \epsilon) \times \mathbf{1})$ where $\mathbf{1} \in \mathbb{R}^d$ is the all-ones vector.

E.3 Comparison with ME and SCF tests of Jitkrittum et al. (2016)

We now present some experimental results comparing the performance of our cross-MDD test with the linear time mean embedding (MD) and smoothed characteristic function (SCF) tests of Jitkrittum et al. (2016). These tests proceed in the following steps:

- Fix J, and choose points $\{v_1, \ldots, v_J\}$ from $\mathbb{R}^d$, where d is the dimension of the observation space.
- Using $\mathbb{X}$ and $\mathbb{Y}$ with $n = m$, compute $\{z_i : 1 \leq i \leq n\}$, where $z_i = [k(v_J, X_i) - k(v_J, Y_i)]_{j=1}^J \in \mathbb{R}^J$ for ME test, and $z_i = [\hat{l}(X_i)\sin(X_i^T v_j) - \hat{l}(Y_i)\sin(Y_i^T v_j), \hat{l}(X_i)\cos(X_i^T v_j) - \hat{l}(Y_i)\cos(Y_i^T v_j)]_{j=1}^J \in \mathbb{R}^{2J}$ for the SCF test. Define $\bar{z}_n = \frac{1}{n}\sum_{i=1}^n z_i$, and $S_n = \frac{1}{n-1}(z_i - \bar{z}_n)(z_i - \bar{z}_n)^T$.
- Using the above, define the test statistic
$$\hat{\lambda}_n := \bar{z}_n^T \left(S_n + \gamma_n I\right)^{-1} \bar{z}_n,$$
where γ_n is some regularization parameter that converges to 0 with n, and I denotes the identity matrix. For a fixed d and J, Jitkrittum et al. (2016) show that the above statistic has a $\chi^2(J)$ (resp. $\chi^2(2J)$) limiting null distribution in the ME (resp. SCF) case. This result is used to calibrate the test at a given level α.

In Figure 9, we plot the variation of type-I error and power with sample-size of the three tests for the Gaussian Mean Difference (GMD) source with $d = 10$. As the figures suggest, the cross-MMD achieves higher power and tighter control over the type-I error than the ME and SCF tests in this regime.

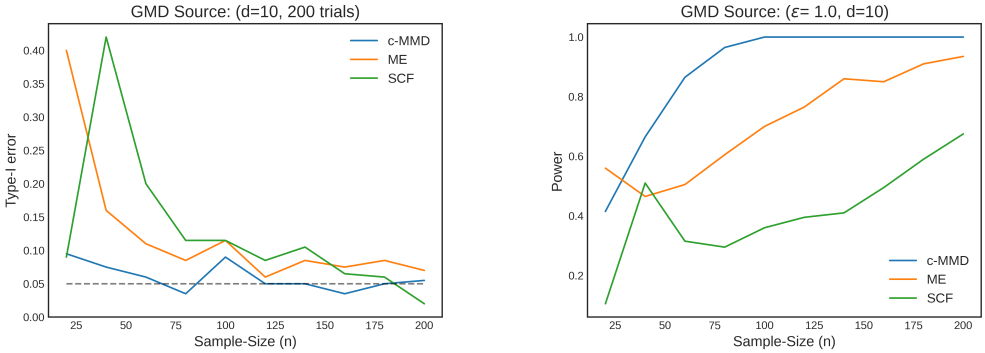

Figure 9: The figures plot the variation of the type-I error (left) and the power (right) with sample-size of the three tests: cross-MMD, and the two linear time tests, ME and SCF, proposed by Jitkrittum et al. (2016).

The ME and SCF tests are calibrated based on the limiting distribution of their statistic in the low dimensional regime: fixed d, and $n \to \infty$. However, the high type-I error of these tests for small n values suggests that their limiting distribution may be different in the high dimensional regime, when both d and n go to infinity. We further observe this in Figure 10 when $d = 100$ and $d/n > 1$.

We end this section with a discussion of some key points of difference between the ME and SCF tests, and our proposed cross-MMD test.

- The ME and SCF tests require the kernel to be uniformly bounded, whereas our test requires only mild moment conditions that are even satisfied by unbounded kernels if the underlying distributions are not too heavy-tailed (formally described in Assumption 1). Furthermore, the ME and SCF tests have several tuning parameters: number of features J, $\{v_1, \ldots, v_J\}$, bandwidth, step-size for gradient ascent etc. In practice, J is usually set to 5, and the other parameters are selected by solving a $Jd + 1$ dimensional optimization problem via gradient ascent. While each step of gradient ascent has linear in n complexity, the number of steps needed may be large for higher dimensions, resulting in a higher computational overhead.

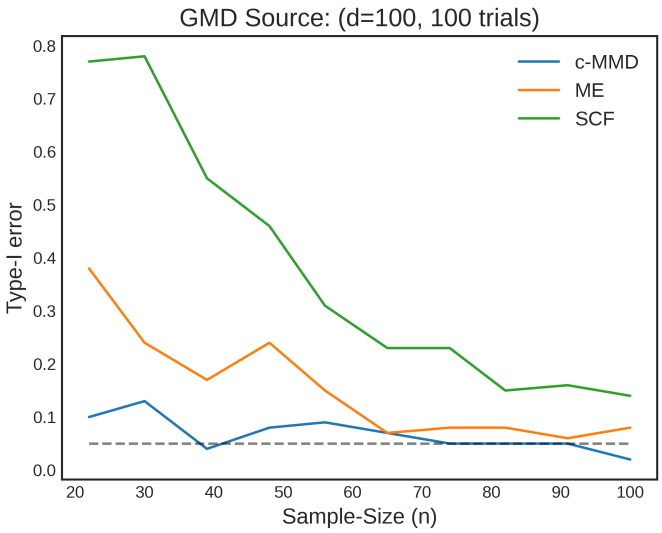

Figure 10: The ME and SCF tests provide poor control over the type-I error in the regime when d/n is large, suggesting that the limiting null distribution is different (or the convergence rate is slow) in this regime.

- More importantly, the ME and SCF tests are only valid in the 'low-dimensional setting': fixed d and J, with $n \to \infty$. In the high dimensional setting, when $(d, n) \to \infty$, the limiting null distribution may no longer be $\chi^2(J)$. This is also suggested by the behavior of type-I error of ME and SCF tests in Figure 9 and Figure 10. This results in the following practical issue: *given a problem with $n = 500$ and $d = 200$, how should one calibrate the threshold for those tests?*

  Our proposed test does not suffer from this, because in both high and low dimensional settings, our statistic has the same limiting distribution. This is a significant practical advantage of our cross-MMD test over ME and SCF tests.

- In the regime where the number of features, J, is allowed to increase with n, we expect that the resulting ME and SCF tests may have low power (for small regularization parameter γ_n). This is because, the test statistic $\hat{\lambda}_n$ used by ME and SCF tests is similar to Hotelling's T^2 statistic, for which Bai and Saranadasa (1996) characterized the asymptotic power in this regime. In particular, their Theorem 2.1 implies that the power of the T^2 test grows slowly with n, especially when $J/n \approx 1$.

  Finally, we note that our ideas also extend to more general degenerate U-statistics (as discussed in Appendix D.1). Hence, they are also applicable in cases beyond MMD distance, where we may not have good linear time alternatives.

E.4 Type-I Error and goodness-of-fit test of null distribution

In this section, we experimentally verify the limiting Gaussian distribution of the $\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2$ statistic under the null. We first plot the variation of the type-I error of our cross-MMD test with sample size in Figure 11. We considered the case when $\mathbb{X}$ and $\mathbb{Y}$ are both drawn i.i.d. from a multivariate Gaussian vector in dimension $d \in \{10, 100\}$, and $n = m$.

Next, we plot the p-values for the test for normality proposed by D'Agostino and Pearson (1973), and implemented in the function `scipy.stats.normaltest` in Python. We performed this test at different sample-sizes (n), and for each value of n, we calculated the $\widehat{\bar{\mathrm{c}}\mathrm{MMD}}^2$ statistic on 200 different indpendent sample pairs. The results are shown in Figure 12

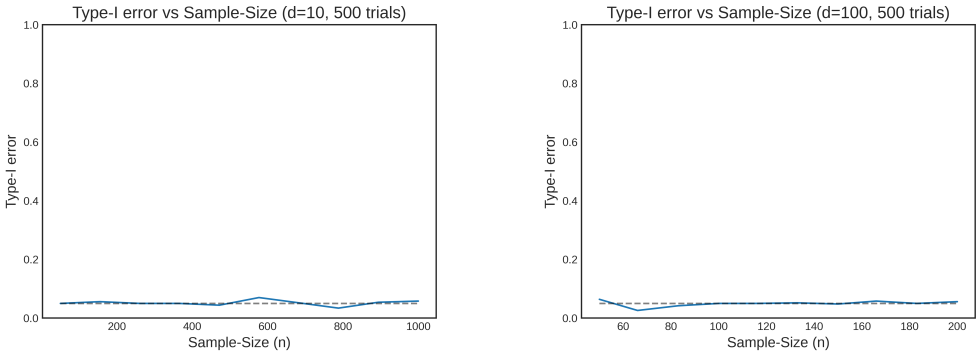

Figure 11: The two figures show the variation of the type-I error of the cross-MMD test with sample-size for dimensions $d \in \{10, 100\}$. The dashed horizontal line denotes the level $\alpha = 0.05$.

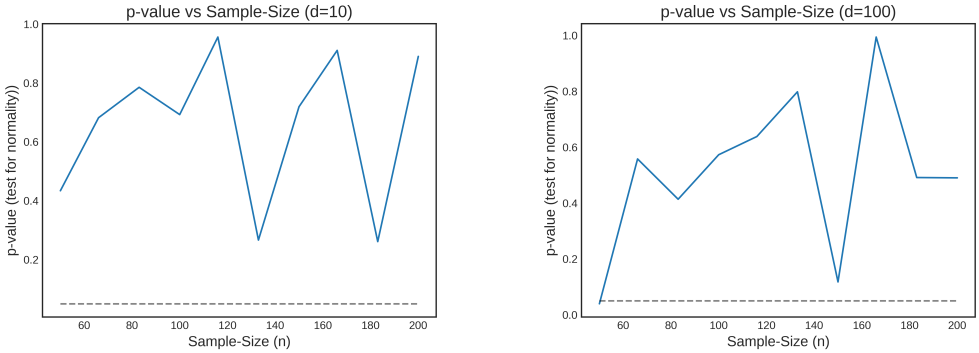

Figure 12: The two figures show p-values for the test for normality proposed by D'Agostino and Pearson (1973) (using the implementation `scipy.stats.normaltest`) of the cross-MMD statistic for dimensions $d \in \{10, 100\}$.

E.5 Comparison with Friedman-Rafsky test

We now compare the performance of our cross-MMD test with the Friedman-Rafsky two-sample test. This test, proposed by Friedman and Rafsky (1979), uses a graph-based statistic that is a multivariate generalization of the Wald-Wolfowitz runs statistic introduced by Wald and Wolfowitz (1940). This statistic, denoted by R, is constructed as follows:

- Pool the samples $\mathbb{X}$ and $\mathbb{Y}$ to get $\mathbb{Z}$ of size $N = n + m$. Construct the complete graph with N nodes, and edge weights equal to the euclidean distance between two end points.
- Construct the minimal spanning tree (MST) of the complete graph G, and denote the 0-1 valued adjacency matrix of this MST by M.
- The statistic R is defined as one more than the number of edges in M with endpoints from different samples.

The statistic R is expected to take a large value under the null when $\mathbb{X}$ and $\mathbb{Y}$ are drawn from the same distribution. Hence, the FR test rejects the null for small values of R. The rejection threshold can be obtained either by the limiting distribution of R characterized by (Henze and Penrose, 1999, Theorem 1), or using the permutation-test.

In Figure 13, we compare the power of the FR permutation-test with our cross-MMD test in a low dimensional (d/n small) and a high dimensional (d/n large) problem. In both cases, it is observed that the power of FR test is significantly smaller than that of cross-MMD test.

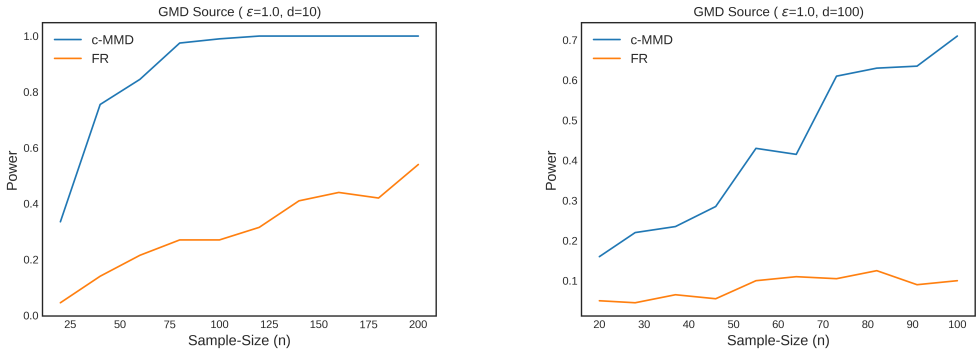

Figure 13: The figures show the power curves for Friedman-Rafsky (FR) test and our cross-MMD test in the low ($d = 10$) and high ($d = 100$) dimensional settings. In both these plots, we set $m = n$.