Maximum Mean Discrepancy with Unequal Sample Sizes via Generalized U-Statistics

Anonymous authors
Paper under double-blind review

Abstract

Existing two-sample testing techniques, particularly those based on choosing a kernel for the Maximum Mean Discrepancy (MMD), often assume equal sample sizes from the two distributions. Applying these methods in practice can require discarding valuable data, unnecessarily reducing test power. We address this long-standing limitation by extending the theory of generalized U-statistics and applying it to the usual MMD estimator, resulting in new characterization of the asymptotic distributions of the MMD estimator with unequal sample sizes (particularly outside the proportional regimes required by previous partial results). This generalization also provides a new criterion for optimizing the power of an MMD test with unequal sample sizes. Our approach preserves all available data, enhancing test accuracy and applicability in realistic settings. Along the way, we give much cleaner characterizations of the variance of MMD estimators, revealing something that might be surprising to those in the area: while zero MMD implies a degenerate estimator, it is sometimes possible to have a degenerate estimator with nonzero MMD as well. We give a construction of such a case, and a proof that it does not happen in common situations.

1 Introduction

Two-sample testing is a fundamental problem in statistical inference, where the objective is to determine whether two arbitrary distributions, P and Q, differ, based only on samples drawn from each. Applications include distinguishing treatment and control groups (e.g. Kobayashi et al., 2017), evaluating and training generative models (e.g. Li et al., 2015; Dziugaite et al., 2015; Bińkowski et al., 2018; Jayasumana et al., 2024), and domain adaptation (e.g. Long et al., 2013), among many others. Given a dataset $\mathbf{S} := (\mathbf{S}_P, \mathbf{S}_Q)$, where $\mathbf{S}_P := (x_i)_{i\in[n_X]} \sim P^{n_X}$ and $\mathbf{S}_Q = (y_i)_{i\in[n_Y]} \sim Q^{n_Y}$, the goal is to test the null hypothesis $H_0 : P = Q$ versus the alternative $H_1 : P \neq Q$. Typically, we do this by comparing a test statistic $\tau(\mathbf{S})$ to a threshold c_α for a given significance level α, say 0.05. The null hypothesis is rejected if $\tau(\mathbf{S}) > c_\alpha$. The threshold should be set such that if H_0 holds, the probability of $\tau(\mathbf{S})$ exceeding c_α (the false rejection rate) is at most α.

In recent years, tests based on the kernel maximum mean discrepancy or MMD (Gretton, Borgwardt, et al., 2012), or equivalently (Sejdinovic et al., 2013) tests based on energy statistics (Székely and Rizzo, 2013), have gained popularity due to their flexibility, adaptivity to various settings, ease of implementation, computational simplicity, and desirable theoretical properties. Setting the threshold c_α is now usually addressed by permutation testing (Sutherland et al., 2017; Hemerik and Goeman, 2018) rather than the asymptotic distribution, for both computational and statistical reasons. Methods for choosing a kernel that will perform well on the particular task at hand, however, are mostly based on an estimate of the asymptotic behavior of the test statistic under both the null and the alternative distributions (Gretton, Sejdinovic, et al., 2012; Sutherland et al., 2017; Liu et al., 2020; Kübler et al., 2022; Deka and Sutherland, 2023).

When the number of samples from both distributions is equal, an MMD estimator can be constructed as a U-statistic. This allows for the application of well-established asymptotic results for U-statistics (see e.g. Lee, 2019) in designing the testing procedure, as has been done in practice for the line of work cited above (excepting that of Kübler et al. (2022), who use a different and more limited framework).

In practice, however, sample sizes are often unequal, leading to an MMD estimator that is not a U-statistic. Consequently, the theoretical framework developed for the equal sample size case cannot be directly applied. (Even when sample sizes are equal, the U-statistic estimator is slightly different from the usual, lower-variance, unbiased estimator.) Liu et al. (2020) addressed this by training on equal-sized subsamples, finding a kernel which has roughly the best test power for a test between samples of equal size, then perhaps applying that kernel to a test set with differing sizes. While this procedure works, it is wasteful.

This paper addresses this limitation by deriving the asymptotic distributions and variance estimators for the MMD estimator under unequal sample sizes. As noted by Kim et al. (2022), the typical MMD estimator can be viewed as a *generalized* U-statistic, but the asymptotics of this more general framework are as yet underdeveloped. We find the required general results and apply them to the MMD estimator to find its asymptotic distribution under both null and alternative hypotheses, allowing us to choose kernel tests to maximize their power (Sutherland et al., 2017; Liu et al., 2020; Deka and Sutherland, 2023). Along the way, we also correct a common misconception in the literature about MMD under the alternative hypothesis.

Furthermore, we demonstrate that our generalized testing procedure reduces to the previous approach when sample sizes are equal, thus establishing it as a natural extension of the existing framework. This unification of methods provides a comprehensive approach to two-sample testing using MMD, accommodating a wider range of practical scenarios while maintaining theoretical rigor.

2 Preliminaries

The Maximum Mean Discrepancy (MMD) is a widely used metric to measure the distance between probability distributions. Its versatility and theoretical properties make it particularly useful in various statistical applications, including two-sample testing.

2.1 Kernel mean embeddings and Hilbert-Schmidt operators

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the kernel of a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of functions $\mathcal{X} \to \mathbb{R}$; the notation $k(\cdot, x)$ denotes the element $t \mapsto k(t, x)$, which satisfies the reproducing property $\langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$ for all $f \in \mathcal{H}$, and hence $k(x, y) = \langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}}$. The *mean embedding* of a distribution P is

$$\mu_P = \mathop{\mathbb{E}}_{X \sim P}[k(X, \cdot)] \in \mathcal{H}.$$

The mean embedding exists and is well-defined when $\mathbb{E}\|k(X, \cdot)\| = \mathbb{E}\sqrt{k(X, X)} < \infty$; it satisfies a reproducing property for distributions, $\langle f, \mu_P\rangle_{\mathcal{H}} = \mathbb{E}_{X\sim P} f(X)$ for any $f \in \mathcal{H}$. A related concept is the (centered) *covariance operator* of a distribution P, given by

$$C_P = \mathop{\mathbb{E}}_{X \sim P}[k(X, \cdot) \otimes k(X, \cdot)] - \mu_P \otimes \mu_P,$$

which exists as long as $\mathbb{E}\, k(X, X) < \infty$. Here the outer product $f \otimes g$ for $f, g \in \mathcal{H}$ is viewed as an $\mathcal{H} \to \mathcal{H}$ operator with $[f \otimes g](g') = f\langle g, g'\rangle$, so that $\langle f, C_P g\rangle = \mathrm{Cov}_{X\sim P}(f(X), g(X))$ for any $f, g \in \mathcal{H}$. The elements $f \otimes g$ and C_P are Hilbert-Schmidt operators from $\mathcal{H} \to \mathcal{H}$, denoted by $f \otimes g \in \mathrm{HS}(\mathcal{H}, \mathcal{H})$; this is itself a Hilbert space with inner product $\langle f \otimes f', g \otimes g'\rangle_{\mathrm{HS}} = \langle f, g\rangle_{\mathcal{H}}\langle f', g'\rangle_{\mathcal{H}}$. The review of Muandet et al. (2017) describes these two objects further. Throughout this paper, we will assume that they are well-defined:

Setting A. *We assume that $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel which induces a separable RKHS $\mathcal{H}$, and that the covariance operator C_P and mean embedding μ_P exist for all considered distributions, i.e. $\mathbb{E}_{X\sim P}\, k(X, X) < \infty$.*

For separability, it suffices to have a separable $\mathcal{X}$ and continuous k (Steinwart and Christmann, 2008, Lemma 4.33). This gives us a variety of results, such as the following.

Proposition 2.1. *In Setting A, for independent random variables $X \sim P, Y \sim Q$, we have $\mathbb{E}[k(X, Y)^2] < \infty$.*

Proof. Since μ_P, C_P exist, the uncentered covariance operator $\tilde{C}_P = C_P + \mu_P \otimes \mu_P = \mathbb{E}[k(\cdot, X) \otimes k(\cdot, X)]$ is Hilbert-Schmidt, and likewise for $\tilde{C}_Q = \mathbb{E}[k(\cdot, Y) \otimes k(\cdot, Y)]$. By Bochner integrability, we can move expectations in and out of inner products, obtaining that the following inner product must be finite:

$$\langle \tilde{C}_P, \tilde{C}_Q \rangle_{\mathrm{HS}} = \left\langle \mathop{\mathbb{E}}_X k(\cdot, X) \otimes k(\cdot, X), \mathop{\mathbb{E}}_Y k(\cdot, Y) \otimes k(\cdot, Y) \right\rangle_{\mathrm{HS}} = \mathbb{E}\big[\langle k(\cdot, X), k(\cdot, Y)\rangle_{\mathcal{H}}^2\big] = \mathbb{E}\big[k(X, Y)^2\big]. \qquad \square$$

2.2 Maximum Mean Discrepancy

Definition 2.2. *Given a reproducing kernel Hilbert space $\mathcal{H}$, the MMD between distributions P and Q is defined as*

$$\mathrm{MMD}(P, Q) := \sup_{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1} \mathop{\mathbb{E}}_{X \sim P} f(X) - \mathop{\mathbb{E}}_{Y \sim Q} f(Y).$$

The MMD always satisfies all properties of a metric except that it may have $\mathrm{MMD}(P, Q) = 0$ for some $P \neq Q$; this does not happen for *characteristic* kernels (Sriperumbudur et al., 2011). None of our results will require a characteristic kernel.

Using the reproducing property, under Setting A MMD can be rewritten as (Gretton, Borgwardt, et al., 2012)

$$\mathrm{MMD}^2(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}}^2 = \mathop{\mathbb{E}}_{X, X' \sim P}[k(X, X')] + \mathop{\mathbb{E}}_{Y, Y' \sim P}[k(Y, Y')] - 2 \mathop{\mathbb{E}}_{X \sim P, Y \sim Q}[k(X, Y)].$$

Estimators for the MMD are often based on this last form. Assuming independent samples $\mathbf{S}_P := \{x_i\}_{i \in [n_X]} \sim P^{n_X}$ and $\mathbf{S}_Q = \{y_i\}_{i \in [n_Y]} \sim Q^{n_Y}$, the typical unbiased estimator is

$$\widehat{\mathrm{MMD}}^2 = \frac{2}{n_X(n_X - 1)} \sum_{i=1}^{n_X} \sum_{j=i+1}^{n_X} k(x_i, x_j) + \frac{2}{n_Y(n_Y - 1)} \sum_{i=1}^{n_Y} \sum_{j=i+1}^{n_Y} k(y_i, y_j) - \frac{2}{n_X n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} k(x_i, y_j). \quad (1)$$

If $n_X = n_Y = n$, we can obtain a simpler, nearly-equivalent estimator. Let $z_i = (x_i, y_i)$, and define

$$h(z_i, z_j) = k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i) \quad (2)$$

$$\widehat{\mathrm{MMD}}_{\mathrm{U}}{}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} h(z_i, z_j). \quad (3)$$

This estimator differs from $\widehat{\mathrm{MMD}}^2(\mathbf{S}_P, \mathbf{S}_Q)$ only in that it omits the $k(x_i, y_i)$ terms; this remains unbiased for MMD^2, but has very slightly higher variance (compare Proposition 2.6 and Theorem 3.4) and unlike $\widehat{\mathrm{MMD}}^2$ depends on the order of the two samples. The difference can be directly bounded by McDiarmid's inequality (details in Appendix A).

Theorem 2.3. *When the number of samples from P and Q are the same, we have that*

$$\Pr_{\substack{X_1, \ldots, X_n \sim P \\ Y_1, \ldots, Y_n \sim Q}} \left(|\widehat{\mathrm{MMD}}^2 - \widehat{\mathrm{MMD}}_{\mathrm{U}}{}^2| \leq \frac{8 \sup_{x \in \mathcal{X}} k(x, x)}{n^{3/2}} \sqrt{\log \frac{2}{\delta}} \right) \geq 1 - \delta.$$

2.3 U-Statistics

The estimator $\widehat{\mathrm{MMD}}_{\mathrm{U}}{}^2$ is an instance of a class of statistics known as U-statistics, introduced by Hoeffding (1948). We will only need U-statistics of order two here.

Definition 2.4. *Let $Z_1, \ldots, Z_n$ be i.i.d. random variables with support in $\mathcal{Z}$, and let $h : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be symmetric in the sense that $h(z_1, z_2) = h(z_2, z_1)$. A* U-statistic of order two *is defined as*

$$U_n = \frac{1}{n(n-1)} \sum_{i \neq j} h(Z_i, Z_j).$$

The function h is called the kernel of the U-statistic (not to be confused with the RKHS kernel k); for $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$ it is (2). These statistics have many general properties; we will be particularly interested in their variance.

Proposition 2.5. *Let U_n be a U-statistic of order two with kernel h. Then, when $n \geq 2$,*

$$\begin{aligned}\mathrm{Var}(U_n) &= \frac{4(n-2)}{n(n-1)} \operatorname*{Var}_{Z_1}\big[\operatorname*{\mathbb{E}}_{Z_2}[h(Z_1, Z_2) \mid Z_1]\big] + \frac{2}{n(n-1)} \operatorname*{Var}_{Z_1,Z_2}[h(Z_1, Z_2)] \\ &= \frac{4n-6}{n(n-1)} \operatorname*{Var}_{Z_1}[\operatorname*{\mathbb{E}}_{Z_2}[h(Z_1, Z_2) \mid Z_1]] + \frac{2}{n(n-1)} \operatorname*{\mathbb{E}}_{Z_1} \operatorname*{Var}_{Z_2}[h(Z_1, Z_2) \mid Z_1].\end{aligned}$$

The first form is a textbook result, shown e.g. by Lee (2019, Section 1.3) or Serfling (1980, Section 5.2). The second follows immediately from the first by the law of total variance. While this first form is well-known and can be used to give an explicit form for the variance of $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$ (Sutherland and Deka, 2019), we can actually simplify the form considerably. The following result, shown in Appendix B, uses the approach of He et al. (2025, Theorem 6.1).

Proposition 2.6. *In Setting A, we have that*

$$\operatorname*{Var}_{\substack{\mathbf{S}_P \sim P^n \\ \mathbf{S}_Q \sim Q^n}} \left(\widehat{\mathrm{MMD}}_{\mathrm{U}}^2(\mathbf{S}_P, \mathbf{S}_Q)\right) = \frac{4}{n} \langle \mu_P - \mu_Q, (C_P + C_Q)(\mu_P - \mu_Q) \rangle_{\mathcal{H}} + \frac{2}{n(n-1)} \|C_P + C_Q\|_{\mathrm{HS}}^2. \tag{4}$$

For a general U-statistic of order two, $\mathrm{Var}(U_n)$ has three possible asymptotic behaviours: if $\mathrm{Var}(U_n) = \Theta(1/n)$, we call U_n *non-degenerate* or zeroth-order degenerate. If $\mathrm{Var}(U_n) = \Theta(1/n^2)$, we call U_n *first-order degenerate*. Otherwise, $\mathrm{Var}(U_n) = 0$, which we term *infinitely degenerate*. Proposition 2.6 allows us to almost fully characterize the degeneracy of $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$. One of its results needs the following stronger assumptions:

Setting B. *In Setting A, further assume that $\mathcal{X} = \mathbb{R}^d$, k is real-analytic, $\sup_{x \in \mathcal{X}} k(x, x)$ is finite, and the supports of P and Q each have positive Lebesgue measure.*

In Setting B, which encompasses many common kernel choices such as the Gaussian, every function in $\mathcal{H}$ is real-analytic (Chwialkowski et al., 2015, Lemma 1). The following result is proved in Appendix C.

Theorem 2.7. *In Setting A, $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$ is infinitely degenerate (the variance is zero) if and only if $C_P = C_Q = \mathbf{0}$. Note that an infinitely degenerate MMD estimate may still be nonzero, such as if P and Q are distinct point masses.*

Now suppose that at least one of C_P, C_Q is nonzero, so $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$'s order of degeneracy is either zero or one. Then:

- *(i) If $\mu_P = \mu_Q$, $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$ is first-order degenerate.*
- *(ii) When $\mu_P \neq \mu_Q$, $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$ may be either non-degenerate or first-order degenerate.*
- *(iii) Suppose $\mathcal{X}$ is a topological space, $k(x, \cdot)$ is continuous for each x, and $\sup_{x \in \mathcal{X}} k(x, x)$ is finite. Further assume the supports of P and Q are not disjoint. Then $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$ is degenerate if and only if $\mu_P = \mu_Q$.*
- *(iv) In Setting B, $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$ is degenerate if and only if $\mu_P = \mu_Q$.*
- *(v) In Setting B, $\langle \mu_P - \mu_Q, C_P(\mu_P - \mu_Q) \rangle > 0$ if and only if $\langle \mu_P - \mu_Q, C_Q(\mu_P - \mu_Q) \rangle > 0$.*

It is well-known that if $\mu_P = \mu_Q$, the $1/n$ term is zero. To the best of our knowledge, however, it has not been previously recognized in the literature that even when $\mu_P \neq \mu_Q$, the estimator may be first-order degenerate, and several papers make (informal) claims to the contrary. In many situations of interest, however, this is impossible, as shown in parts (iii) and (iv). We also note that (iv) remains true if only one of P, Q has a support with positive Lebesgue measure.

The asymptotic behavior of U-statistics is also highly relevant, and determined by the degree of degeneracy. Here we need only textbook results, as shown by Lee (2019, Section 3.2) or Serfling (1980, Section 5.5).

Theorem 2.8. *Let U_n be a U-statistic of order two with kernel h. Suppose that U_n has mean $\theta = \mathbb{E}\, h(X_1, X_2)$, and let $\sigma_1^2 = 4 \operatorname{Var}_{X_1}[\mathbb{E}_{X_2}[h(X_1, X_2) \mid X_1]]$ be the leading term in the variance decomposition.*

It holds as $n \to \infty$ that $\sqrt{n}\,(U_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma_1^2)$, where if $\sigma_1 = 0$ this convergence is to the point mass at 0.

If $\sigma_1 = 0$, it also holds as $n \to \infty$ that $n\,(U_n - \theta) \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j (V_j^2 - 1)$, where the $\{V_j\}_{j=1}^{\infty}$ are independent standard normal variables, and $(\lambda_j)_{j=1}^{\infty}$ are the eigenvalues of the integral equation $\mathbb{E}_{X_2}[h(x_1, X_2) f(X_2)] = \lambda f(x_1)$.

2.4 MMD-based tests

The eigenvalues λ_j in the degenerate case, as they depend on the kernel and the distribution, are often difficult to find. Because $\widehat{\mathrm{MMD}}_\mathrm{U}{}^2$ is degenerate under the null hypothesis, where $\mathrm{MMD}(P, Q) = 0$, the test threshold must be set based on the more complex distribution. The eigenvalues λ_j can be estimated based on eigendecomposition of the sample kernel matrix (Gretton et al., 2009), but it is usually faster and more effective to choose a threshold based on permutation testing (e.g. Sutherland et al., 2017). A variant of this procedure also achieves finite-sample valid test thresholds (Hemerik and Goeman, 2018), rather than the only asymptotic validity achieved from consistent estimates of λ_j.

Theorems 2.7 and 2.8 imply that if $\mu_P = \mu_Q$, then the $(1-\alpha)$th quantile of $\widehat{\mathrm{MMD}}_\mathrm{U}{}^2$ will be either $\Theta(1/\sqrt{n})$ or simply 0. On the other hand, if $\mu_P \neq \mu_Q$ so that $\mathrm{MMD}^2 > 0$, then $\widehat{\mathrm{MMD}}^2$ will be one of $\mathrm{MMD}^2 + \mathcal{O}_p(1/\sqrt{n})$, $\mathrm{MMD}^2 + \mathcal{O}_p(1/n)$, or simply MMD^2, depending on the degeneracy behaviour. Thus as $n \to \infty$, any test whose asymptotic level is controlled will be consistent, regardless of the degree of degeneracy. That is, whenever $\mu_P \neq \mu_Q$, an MMD test will eventually reject as $n \to \infty$.

To do so, however, may require a very large number of samples when the kernel is a poor match to the problem at hand; for example, a Gaussian kernel based on image pixels does a reasonable job at identifying pixel-level shifts on simple aligned image datasets like MNIST (see Sutherland et al., 2017), but would require huge numbers of samples to identify "semantic" shifts in more complex natural image distributions.

To address this issue, Gretton, Sejdinovic, et al. (2012), Sutherland et al. (2017), Liu et al. (2020), and Kübler et al. (2022) maximize the asymptotic *power* of a given MMD test with equal sample sizes. Assuming that $\sigma_1(P, Q)^2$ as defined in Theorem 2.8 is nonzero,

$$\Pr\left(n\,\widehat{\mathrm{MMD}}_\mathrm{U}{}^2(\mathbf{S}_P, \mathbf{S}_Q) > c_\alpha\right) = \Pr\left(\sqrt{n}\frac{\widehat{\mathrm{MMD}}_\mathrm{U}{}^2(\mathbf{S}_P, \mathbf{S}_Q) - \mathrm{MMD}^2(P, Q)}{\sigma_1(P, Q)} > \frac{\frac{c_\alpha}{\sqrt{n}} - \sqrt{n}\,\mathrm{MMD}^2(P, Q)}{\sigma_1(P, Q)}\right),$$

When $\sigma_1 > 0$, the left-hand side of the inequality converges in distribution to a standard normal, and so

$$\Pr\left(n\,\widehat{\mathrm{MMD}}_\mathrm{U}{}^2(\mathbf{S}_P, \mathbf{S}_Q) > c_\alpha\right) \sim 1 - \Phi\left(\frac{\frac{c_\alpha}{\sqrt{n}} - \sqrt{n}\,\mathrm{MMD}^2(P, Q)}{\sigma_1(P, Q)}\right) = \Phi\left(\sqrt{n}\frac{\mathrm{MMD}^2(P, Q)}{\sigma_1(P, Q)} - \frac{c_\alpha}{\sqrt{n}\sigma_1(P, Q)}\right),$$

where $a \sim b$ denotes $a/b \to 1$ and Φ is the cdf of the standard normal distribution. Since $\mathrm{MMD}(P, Q)$, $\sigma_1(P, Q)$, and c_α are population quantities that do not depend on n, for large sample sizes the asymptotic power expression is dominated by the signal-to-noise ratio $\mathrm{MMD}^2(P, Q)/\sigma_1(P, Q)$. Thus, we can choose a kernel by maximizing a finite-sample estimate of this quantity on a training set, then run a test with that kernel on an independent test set.

3 Asymptotic distribution of the MMD estimate

To generalize this approach to the case where $n_X \neq n_Y$, we will derive the asymptotic distributions of the estimator $\widehat{\mathrm{MMD}}^2$, rather than $\widehat{\mathrm{MMD}}_\mathrm{U}{}^2$. To do so, we fill in results about the theory of generalized U-statistics (Serfling, 1980, Section 5.5.1), of which the $\widehat{\mathrm{MMD}}^2$ estimator is an instance even when $n_X \neq n_Y$.

Gretton, Borgwardt, et al. (2012, Theorem 12) previously considered the asymptotics of $\widehat{\mathrm{MMD}}^2$ when $n_X \neq n_Y$, showing that if n_X/n_Y converges to a positive, finite constant and $\mathrm{MMD}(P, Q) = 0$, then $(n_X +$

$n_Y)\,\widehat{\mathrm{MMD}}^2$ converges in distribution to a slightly different sum of shifted chi-squared variates. Our results, by contrast, will also allow $n_X/n_Y \to 0$ or $n_X/n_Y \to \infty$; to do this, instead of scaling by $n_X + n_Y$, we scale by $\min(n_X, n_Y)$. In the proportional setting, this only differs by a constant, but we also allow for non-proportional settings such as $n_Y = n_X^2$.

Before giving our results, we first empirically demonstrate in Figures 1 and 2 that $\min(n_X, n_Y)$ is indeed the correct scaling. In the proportional regime, either scaling works, while in a non-proportional setting only the $\min(n_X, n_Y)$ scaling leads to convergence; it does so to the distributions predicted by our theorems.

3.1 Generalized U-Statistics

Definition 3.1 (Generalized U-statistic)**.** *For $j \in [c] = \{1, 2, \ldots, c\}$, let $(X_{ij})_{i\in[n_j]} \sim \mu_j^{n_j}$ be mutually independent random variables. Let h be a real-valued function of $m_1+\cdots+m_c$ arguments which is symmetric in the sense that the value of $h(x_{11}, \ldots, x_{m_1 1}; \ldots; x_{1c}, \ldots, x_{m_c c})$ remains unchanged if we permute any block of arguments $(x_{1j}, \ldots, x_{m_j j})$. The c-sample generalized U-statistic associated with kernel h is defined by*

$$U_n = \prod_{j=1}^{c} \binom{n_j}{m_j}^{-1} \sum_{\sigma_1} \cdots \sum_{\sigma_c} h(X_{\sigma_1(1),1}, \ldots, X_{\sigma_1(m_j),1}; \ldots; X_{\sigma_c(1),k}, \ldots, X_{\sigma_c(m_c),c}), \tag{5}$$

where $n = \min\{n_1, \ldots, n_c\}$, and σ_j varies over each injection from $[m_j]$ to $[n_j]$.

In particular, as has been previously pointed out by Kim et al. (2022) and Schrab et al. (2023), $\widehat{\mathrm{MMD}}^2$ can be viewed as a generalized U-statistic with $c = 2$, $m_1 = m_2 = 2$, using the kernel

$$h(x_i, x_j; y_i, y_j) := k(x_i, x_j) + k(y_i, y_j) - \frac{1}{2}\left[k(x_i, y_j) + k(x_j, y_i) + k(x_i, y_i) + k(x_j, y_j)\right]. \tag{6}$$

While a direct implementation of (5) would sum over $\mathcal{O}(n_X^2 n_Y^2)$ evaluations of the function h, in fact each term in this h considers at most two elements; thus many terms in the average are irrelevant. An implementation taking this into account becomes exactly that of $\widehat{\mathrm{MMD}}^2$ in (1), with $\mathcal{O}((n_X + n_Y)^2)$ kernel evaluations.

Let $\mathbf{X} = (X_{11}, \ldots, X_{m_1 1}; \ldots; X_{1c}, \ldots, X_{m_c c}) \sim \prod_{j=1}^{c} \mu_j^{m_j}$. Sen (1974) derived the variance of a generalized U-statistic as

$$\begin{aligned} \mathrm{Var}(U_n) &= \sum_{d_1=0}^{m_1} \cdots \sum_{d_c=0}^{m_c} \left(\prod_{j=1}^{c} \binom{n_j}{m_j}^{-1} \binom{m_j}{d_j} \binom{n_j - m_j}{m_j - d_j} \right) \zeta_{d_1 \ldots d_c} \\ \text{where} \quad \zeta_{d_1 \ldots d_c} &= \mathrm{Var}(\mathbb{E}(h(\mathbf{X}) \mid X_{11}, \ldots, X_{d_1 1}, \ldots, X_{1c}, \ldots, X_{d_c c})). \end{aligned} \tag{7}$$

This motivates the following definition.

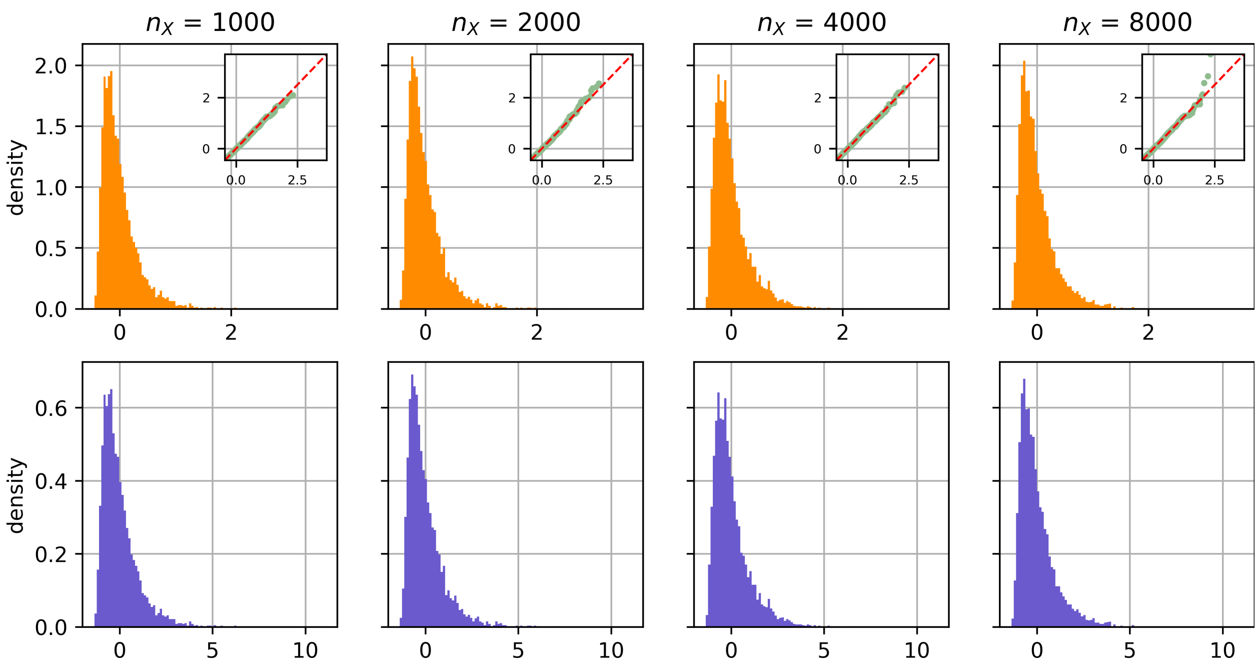
Definition 3.2. *We say that a generalized U-statistic U_n has* order of degeneracy *r if $\zeta_{d_1 \ldots d_c} = 0$ for all $d_1 + \cdots + d_c \le r$, and $\zeta_{d_1 \ldots d_c} > 0$ for at least one $d_1 + \cdots + d_c = r + 1$, where the ζs are as in (7). If all $\zeta_{d_1 \ldots d_c} = 0$, we say its order of degeneracy is infinite. If U_n has order of degeneracy zero, we say U_n is non-degenerate.*

The following result is then immediate from (7), since $\binom{n_j}{m_j}^{-1}\binom{m_j}{d_j}\binom{n_j - m_j}{m_j - d_j} = \Theta(n_j^{-d_j})$:

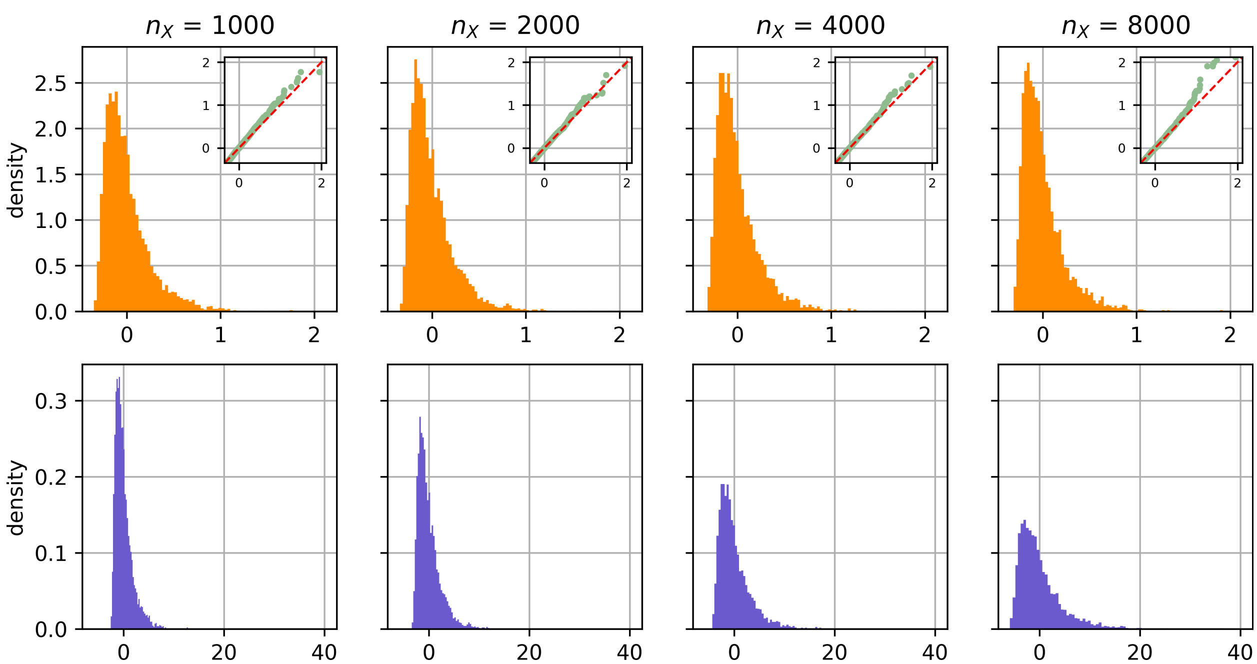
Proposition 3.3. *For a generalized U-statistic with finite order of degeneracy r,*

$$\mathrm{Var}(U_n) = \sum_{d_1=0}^{m_1} \cdots \sum_{d_c=0}^{m_c} \Theta\left(\prod_{j=1}^{c} n_j^{-d_j} \right) \zeta_{d_1 \ldots d_c} = \mathcal{O}\left(n^{-(r+1)} \right). \tag{8}$$

For the MMD in particular, we can express the variance in a form similar in spirit to (4).

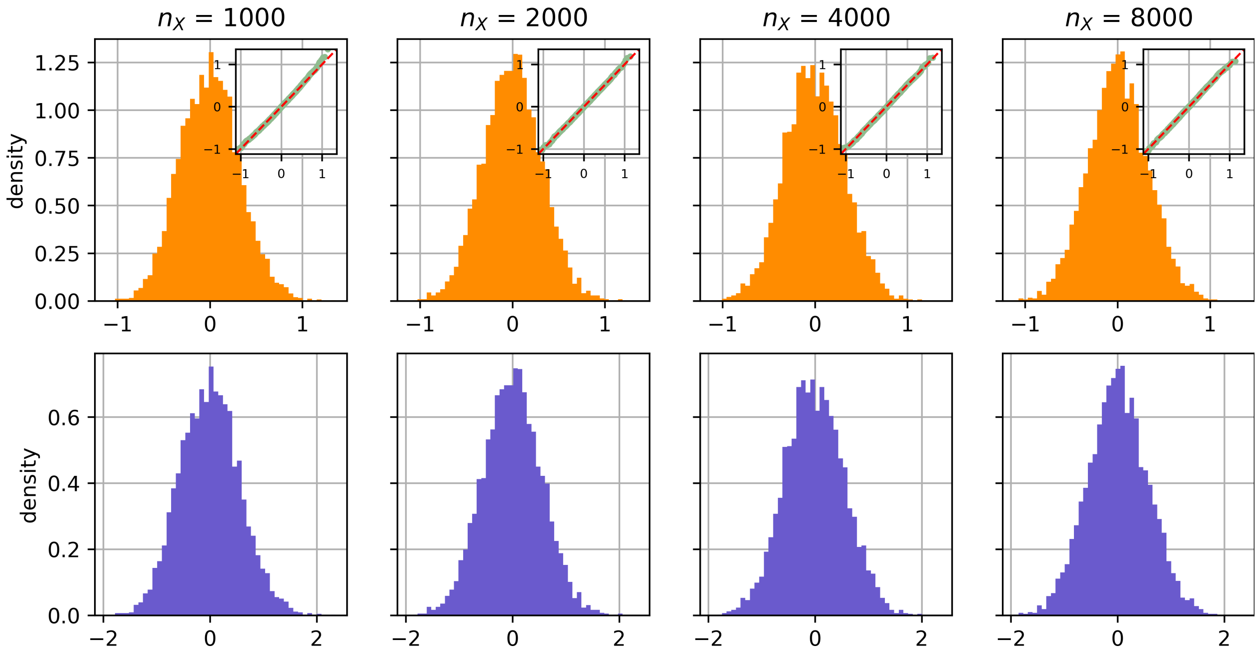


(a) Proportional scaling, $n_Y = n_X/2$.

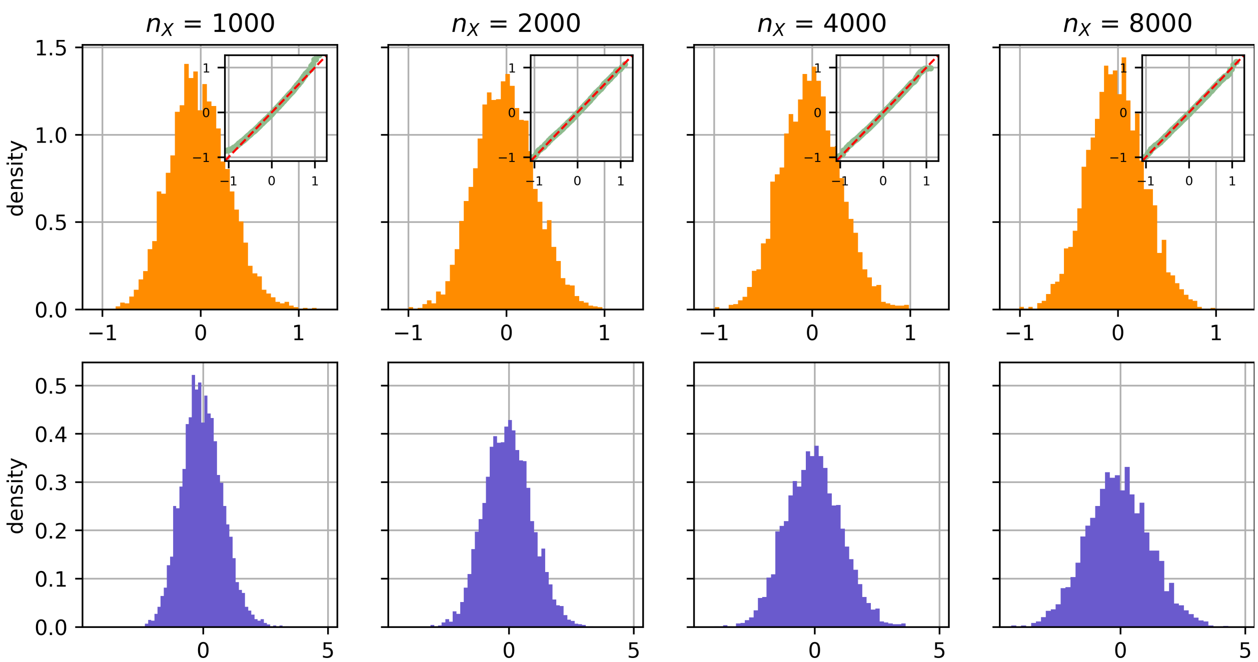


(b) Non-proportional scaling, $n_Y = 5\sqrt{n_X}$.

Figure 1: Histograms of $n\,\widehat{\mathrm{MMD}}^2$ for $P = Q = \mathrm{Laplace}(0, 1/\sqrt{2})$ and a unit-lengthscale Gaussian kernel; orange (top) rows use $n = \min(n_X, n_Y)$, while blue (bottom) rows use $n = n_X + n_Y$. In the proportional setting (panel a), both converge; the only difference is a constant scaling, since $n_X + n_X/2 = 3\min(n_X, n_X/2)$. In the non-proportional setting (panel b), however, it is clear that the $n_X + n_Y$ scaling is not converging in distribution, while the $\min(n_X, n_Y)$ scaling is. The $\min(n_X, n_Y)$ results include inset Q-Q plots, comparing empirical quantiles to those predicted by the limiting distributions of Theorem 3.7; eigenvalues in that distribution are estimated based on a sample with $n_X = n_Y = 5000$.



(a) Proportional scaling, $n_Y = n_X/2$.



(b) Non-proportional scaling, $n_Y = 5\sqrt{n_X}$.

Figure 2: Histograms of $\sqrt{n}(\widehat{\mathrm{MMD}}^2 - \mathrm{MMD})$ for $P = \mathrm{Laplace}(0, 1/\sqrt{2})$, $Q = \mathrm{Laplace}(0, 3)$, and k is a unit-lengthscale Gaussian kernel; here $\mathrm{MMD}(P, Q) > 0$ is estimated based on 5000 samples. As in Figure 1, orange (top) rows use $n = \min(n_X, n_Y)$, while blue (bottom) use $n = n_X + n_Y$. We again see that the $n_X + n_Y$ scaling is clearly not converging in distribution in the non-proportional setting (panel b), while $\min(n_X, n_Y)$ converges in accordance with the distribution from Corollary 3.9. Quantiles of the limiting distribution are computed based on a variance estimated based on a sample with $n_X = n_Y = 5000$.

Theorem 3.4. *Under Setting A, it holds for any $n_X, n_Y \geq 2$ that*

$$\begin{aligned}
\operatorname*{Var}_{\substack{\mathbf{S}_P \sim P^{n_X} \\ \mathbf{S}_Q \sim Q^{n_Y}}}(\widehat{\mathrm{MMD}}^2(\mathbf{S}_P, \mathbf{S}_Q)) &= \frac{4}{n_X}\left(1 - \frac{4}{n_Y} \cdot \frac{n_X - 2}{n_X - 1} \cdot \frac{n_Y - 2}{n_Y - 1}\right) \langle \mu_P - \mu_Q, C_P(\mu_P - \mu_Q)\rangle \\
&+ \frac{4}{n_Y}\left(1 - \frac{4}{n_X} \cdot \frac{n_X - 2}{n_X - 1} \cdot \frac{n_Y - 2}{n_Y - 1}\right) \langle \mu_P - \mu_Q, C_Q(\mu_P - \mu_Q)\rangle \\
&+ \frac{2}{n_X(n_X - 1)}\|C_P\|_{\mathrm{HS}}^2 + \frac{2}{n_Y(n_Y - 1)}\|C_Q\|_{\mathrm{HS}}^2 + \frac{4}{n_X n_Y}\langle C_P, C_Q\rangle_{\mathrm{HS}} \\
&+ \frac{16}{n_X n_Y}\left(\frac{n_X - 2}{n_X - 1} \cdot \frac{n_Y - 2}{n_Y - 1}\right)\left[\langle \mu_P, C_P \mu_P\rangle + \langle \mu_Q, C_Q \mu_Q\rangle\right].
\end{aligned}$$

Notice that each of the factors in large parentheses has limit 1 as $n_X, n_Y \to \infty$, regardless of their relative rate. The proof of Theorem 3.4, which follows from (7), is in Appendix D. The computation of conditional variances also yields the following result about degeneracy.

Proposition 3.5. $\widehat{\mathrm{MMD}}^2$ *has the same order of degeneracy as* $\widehat{\mathrm{MMD}}_{\mathrm{U}}{}^2$*; thus Theorem 2.7 applies to* $\widehat{\mathrm{MMD}}^2$ *as well.*

Remark 3.6. In the proportional regime $n_i = \Theta(n)$, the rate of Proposition 3.3 becomes $\Theta(n^{-(r+1)})$, i.e. the variance characterization is tight for any generalized U-statistic. The same holds for $\widehat{\mathrm{MMD}}^2$ in Setting B via Theorem 2.7 part (v). In general, however, we only have $\mathcal{O}(n^{-(r+1)})$, not Θ: for instance, if $C_P = 0$, $\langle \mu_P - \mu_Q, C_Q(\mu_P - \mu_Q)\rangle > 0$, and $\langle \mu_Q, C_Q \mu_Q\rangle > 0$, $\widehat{\mathrm{MMD}}^2$ is non-degenerate but its variance in Equation (4) is $\Theta\big(0 + \frac{1}{n_Y} + 0 + \frac{1}{n_Y^2} + 0 + \frac{1}{n_X n_Y}\big) = \Theta\big(\frac{1}{n_Y}\big)$, which for $n_Y = n_X^{10}$ is $\Theta(n^{-10})$, not $\Theta(n^{-1})$.

3.2 Distribution when the MMD is zero

We first consider the distribution of the estimator when $\mu_P = \mu_Q$, which implies first-order degeneracy. (Per Theorem 2.7, first-order degeneracy is also possible when $\mu_P \neq \mu_Q$, but relatively uncommon; since it makes the proof more difficult, we do not handle that case here.)

In Theorem E.5, stated in Appendix E, we derive the asymptotic distribution of a general U-statistic with first order degeneracy and a square-integrable kernel. To our knowledge, the generality of this result (in particular to $c > 2$ and to kernels which potentially treat the types of samples asymmetrically) is novel; it extends the approach of Serfling (1980, Section 5.5.2) and Anderson et al. (1994), which was later used specifically for $\widehat{\mathrm{MMD}}^2$ in the proportional setting by Gretton, Borgwardt, et al. (2012, Theorem 12).

Since our assumptions are minimal, the resulting distribution is somewhat complicated to describe. As such, we leave the details to Appendix E, and only give the particular case for $\widehat{\mathrm{MMD}}^2$ here.

Theorem 3.7. *Assume Setting A and that* $\mathrm{MMD}^2(P, Q) = 0$*. Assume* $\min\{n_X, n_Y\}/n_X \to \rho_X$ *and* $\min\{n_X, n_Y\}/n_Y \to \rho_Y$ *for some* ρ_X, ρ_Y *in* $[0, 1]$*.* $\widehat{\mathrm{MMD}}^2$ *converges in distribution as*

$$\min\{n_X, n_Y\}\,\widehat{\mathrm{MMD}}^2 \xrightarrow{d} (\rho_X + \rho_Y)\sum_{l=1}^{\infty} \lambda_l\left(Z_l^2 - 1\right),$$

where each Z_l *is independently* $\mathcal{N}(0, 1)$*, and the* λ_l *are the eigenvalues of the integral equation*

$$\mathop{\mathbb{E}}_{X}\left[\langle \phi(X) - \mu_P, \phi(y) - \mu_P\rangle g(X)\right] = \lambda g(y).$$

For comparison's sake, in our notation Theorem 12 of Gretton, Borgwardt, et al. (2012) says that[1] if $n_X/(n_X+n_Y) \to \ell_X$ and $n_Y/(n_X+n_Y) \to \ell_Y = 1-\ell_X$ for $\ell_X, \ell_Y \in (0,1)$, then

$$(n_X+n_Y)\,\widehat{\mathrm{MMD}}^2 \xrightarrow{d} \frac{1}{\ell_X \ell_Y} \sum_{l=1}^{\infty} \lambda_l \left(Z_l^2 - 1\right).$$

To confirm our results agree, suppose $n_Y = \frac{1}{\rho_Y} n_X$, with $\rho_Y \in (0,1]$, so $\rho_X = 1$. Thus $\ell_X = \frac{1}{1+\frac{1}{\rho_Y}} = \frac{\rho_Y}{1+\rho_Y}$ and $\ell_Y = \frac{1}{1+\rho_Y}$. Let $D = \sum_l \lambda_l (Z_l^2 - 1)$. Then our result is $n_X\,\widehat{\mathrm{MMD}}^2 \xrightarrow{d} (1+\rho_Y)\,D = \frac{1}{\ell_Y} D$, while since $n_X + n_Y = (1+\frac{1}{\rho_Y})n_X = \frac{1}{\ell_X} n_X$, theirs equivalently says that $\frac{1}{\ell_X} n_X\,\widehat{\mathrm{MMD}}^2 \xrightarrow{d} \frac{1}{\ell_X \ell_Y} D$. As demonstrated by Figures 1 and 2, however, their result cannot be generalized to non-proportional asymptotics.

3.3 Non-degenerate asymptotic normality

Our derivation of the asymptotic distribution under the alternative follows from a result on generalized U-statistics, which is simpler to state than the corresponding result for the degenerate case. A similar result was presented as an exercise, without details, by Serfling (1980, Section 5.5). Lehmann (1951) also stated asymptotic normality for the special case $c = 2$, $m_1 = m_2$, but gave a proof only when $n_1 = n_2$.

Theorem 3.8. *Let U_n be a c-sample U-statistic with kernel h, where $n = \min\{n_1, \dots, n_c\}$. If $n/n_j \to \rho_j \in [0,1]$ for each $j \in [c]$, then*

$$\sqrt{n}(U_n - \mathbb{E}[h(\mathbf{X})]) \xrightarrow{d} \mathcal{N}\left(0, \sum_{j=1}^{c} \rho_j m_j^2 \operatorname{Var}\left(\mathbb{E}\left[h(\mathbf{X}) \mid X_{1j}\right]\right)\right),$$

where $\mathcal{N}(0,0)$ is interpreted as a point mass at 0. In particular if U_n is non-degenerate and each $\rho_j > 0$ (the proportional regime), the distribution above is normal with positive variance.

The following application to MMD is immediate from Theorem 3.8 and Theorem 2.7 part (v).

Corollary 3.9. *Assume $\min\{n_X, n_Y\}/n_X \to \rho_X$ and $\min\{n_X, n_Y\}/n_Y \to \rho_Y$ for some $\rho_X, \rho_Y \in [0,1]$. The estimator $\widehat{\mathrm{MMD}}^2$ is asymptotically normal, following*

$$\sqrt{\min\{n_X, n_Y\}}(\widehat{\mathrm{MMD}}^2 - \mathrm{MMD}^2) \xrightarrow{d} \mathcal{N}(0, 4\rho_X \zeta_X + 4\rho_Y \zeta_Y),$$

where $\zeta_X = \langle \mu_P - \mu_Q, C_P(\mu_P - \mu_Q) \rangle$ and $\zeta_Y = \langle \mu_P - \mu_Q, C_P(\mu_P - \mu_Q) \rangle$.

In Setting B, when $\widehat{\mathrm{MMD}}^2$ is non-degenerate the distribution above has positive variance.

Estimating the variance

In order to maximize the power of an $\widehat{\mathrm{MMD}}^2$ test, notice that paralleling the case for $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$, Corollary 3.9 implies that when $\sigma^2 = 4\rho_X \zeta_X + 4\rho_Y \zeta_Y > 0$,

$$\Pr\left(n\,\widehat{\mathrm{MMD}}^2(\mathbf{S}_P, \mathbf{S}_Q) > c_\alpha\right) \sim \Phi\left(\sqrt{n}\frac{\mathrm{MMD}^2}{\sigma} - \frac{c_\alpha}{\sqrt{n}\,\sigma}\right).$$

Thus we can choose the asymptotically best kernel for a test with a given sample size ratio by maximizing the signal-to-noise ratio MMD^2/σ, which we can estimate by $\widehat{\mathrm{MMD}}^2/(\hat\sigma + \lambda)$ for some small $\lambda > 0$. To find an estimator $\hat\sigma$, notice that

$$\zeta_X = \operatorname*{Var}_X\big(\operatorname*{\mathbb{E}}_{X'}[k(X,X')]\big) + \operatorname*{Var}_X\big(\operatorname*{\mathbb{E}}_{Y}[k(X,Y)]\big) - 2\operatorname*{Cov}_X\big(\operatorname*{\mathbb{E}}_{X'}[k(X,X')], \operatorname*{\mathbb{E}}_{Y}[k(X,Y)]\big)$$
$$\zeta_Y = \operatorname*{Var}_Y\big(\operatorname*{\mathbb{E}}_{Y'}[k(Y,Y')]\big) + \operatorname*{Var}_Y\big(\operatorname*{\mathbb{E}}_{X}[k(X,Y)]\big) - 2\operatorname*{Cov}_Y\big(\operatorname*{\mathbb{E}}_{Y'}[k(Y,Y')], \operatorname*{\mathbb{E}}_{X}[k(X,Y)]\big).$$

[1]They wrote their limiting distribution as $\sum_{l=1}^{\infty} \lambda_l \left((\ell_X^{-1/2} A_l - \ell_Y^{-1/2} B_l)^2 - (\ell_X \ell_Y)^{-1}\right)$ for A_l, B_l standard normal; this implies the above result by noting that $\ell_X^{-1/2} A_l - \ell_Y^{-1/2} B_l \stackrel{d}{=} \sqrt{\frac{1}{\ell_X} + \frac{1}{\ell_Y}} Z_l = \sqrt{\frac{\ell_Y + \ell_X}{\ell_X \ell_Y}} Z_l = \sqrt{\frac{1}{\ell_X \ell_Y}} Z_l$.

We use simple plug-in estimators of the following form:

$$\operatorname*{Var}_{X}[\operatorname*{\mathbb{E}}_{X'}[k(X,X')]] \approx \frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{m}\sum_{i=1}^{m}k(x_i,x_j)\right)^2 - \left(\frac{1}{m^2}\sum_{j=1}^{m}\sum_{i=1}^{m}k(x_i,x_j)\right)^2$$

$$\operatorname*{Var}_{X}(\operatorname*{\mathbb{E}}_{Y}[k(X,Y)]) \approx \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n}k(x_i,y_j)\right)^2 - \left(\frac{1}{nm}\sum_{j=1}^{n}\sum_{i=1}^{m}k(x_i,y_j)\right)^2$$

$$\operatorname*{Cov}_{X}(\operatorname*{\mathbb{E}}_{X'}[k(X,X')],\operatorname*{\mathbb{E}}_{Y}[k(X,Y)]) \approx \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{m}\sum_{a=1}^{m}k(x_i,x_a)\right)\left(\frac{1}{n}\sum_{b=1}^{n}k(x_i,y_b)\right) - \left(\frac{1}{m^2}\sum_{i=1}^{m}\sum_{a=1}^{m}k(x_i,x_a)\right)\left(\frac{1}{mn}\sum_{j=1}^{m}\sum_{b=1}^{n}k(x_j,y_b)\right).$$

These terms, as various means of the X-to-X, X-to-Y, and Y-to-Y kernel matrices, can be written straightforwardly in modern automatic differentiation libraries. Thus $\widehat{\mathrm{MMD}}^2/(\hat{\sigma}+\lambda)$ can be easily written as an objective function for gradient-based optimization of kernel parameters. These plug-in estimators are biased, both for simplicity and because biased estimators tend to work better in the denominator here. Unbiased estimators (perhaps also incorporating the lower-order terms of Theorem 3.4), however, could also be derived as was done for $\widehat{\mathrm{MMD}_\mathrm{U}}^2$ by Sutherland and Deka (2019).

4 Experimental comparisons of tests

In this section, we evaluate the performance of two-sample tests based on our estimators when $n_X \neq n_Y$, showing that power is improved when we use as many samples as are available even when that number is asymmetric. In all cases, we use a variant of permutation tests such that the Type-I error rate is exactly controlled under the null (Hemerik and Goeman, 2018).

Synthetic problem: normals with different variances We first consider the behaviour of a simple setting where $P = \mathcal{N}(0,1)$ and $Q = \mathcal{N}(0,1.2)$, and keep the number of samples $n_Y = 50$ from Q fixed while varying $n_X \in \{50, 100, 200, 400, 800\}$. We use a unit-lengthscale Gaussian kernel, and set the threshold for a significance level of 0.05 based on $5,000$ permutations; we estimate the rejection rate using $5,000$ repetitions. The resulting plots for power and type-1 error as functions of the sample size m are shown in Figure 3.

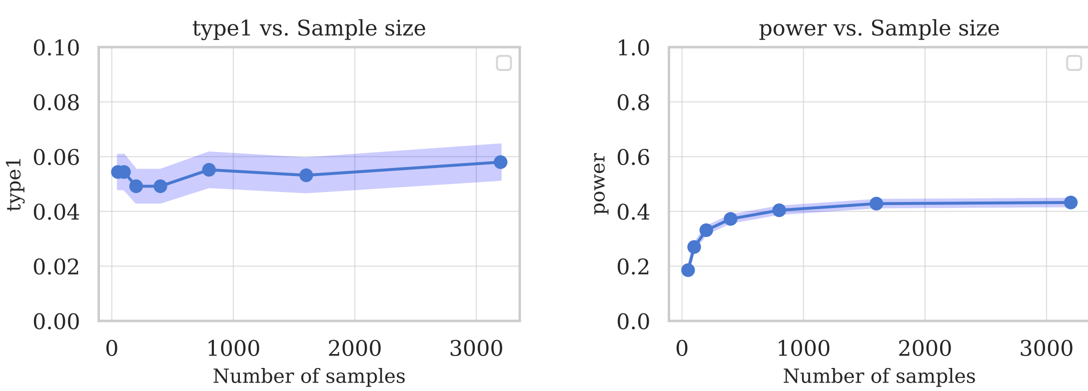


Figure 3: Type-I error rate stays controlled as n_X changes, while the power increases.

***CIFAR-10* vs. *CIFAR-10.1*.** In this experiment, we compare the test subset of *CIFAR-10* (Krizhevsky, 2009), consisting of 10,000 images, with the *CIFAR-10.1* dataset (Recht et al., 2019) of 2,031 images. Following the experimental setup of Liu et al. (2020), we extend the framework to accommodate unequal sample sizes, taking advantage of the flexibility provided by our proposed method.

The experiment is carried out in the following steps:

1. **Data Preparation**: We randomly sample $N_{tr} = 1000$ images from *CIFAR-10.1* and rN_{tr} images from *CIFAR-10* for training, where $r \in \{1, 2, 4\}$. For testing, we use the remaining $N_{te} = 1031$ samples from *CIFAR-10.1* and rN_{te} samples from *CIFAR-10*.
2. **Kernel Selection**: We employ a deep kernel, as described in Liu et al. (2020), learned from the training data to capture the data distributions in a high-dimensional feature space.
3. **Testing Method**: We apply our proposed test to measure the divergence between the distributions P and Q. The rejection rates are calculated by averaging the results over multiple trials, each with newly sampled test sets. We evaluate the test power as the sample size ratio r varies.

The results demonstrate that our method successfully distinguishes between the two distributions. For $r = 1$, the test achieves a power of 0.71, for $r = 2$ a power of 0.96, and for $r = 4$ the test achieves a perfect power of 1. This indicates that as the sample size disparity increases, our method becomes more effective at identifying differences between the distributions.

5 Conclusion

In this paper we have shown that the generalized U-statistic estimator $\widehat{\mathrm{MMD}}^2$ provides an effective framework for two-sample testing with unequal sample sizes. $\widehat{\mathrm{MMD}}^2$ is compatible with the classical setting where the samples sizes are proportional, but our results extend to the situation where one sample is asymptotically dominant by changing the scaling from $n_X + n_Y$ to $\min\{n_X, n_Y\}$. To derive the asymptotic distributions of $\widehat{\mathrm{MMD}}^2$, we characterized generalized U-statistics based on the degeneracy of their kernel and matched this order of degeneracy with the hypotheses H_0 and H_1 under appropriate assumptions. Even when $n_X = n_Y$, while $\widehat{\mathrm{MMD}}^2$ and $\widehat{\mathrm{MMD}}_{\mathrm{U}}{}^2$ are asymptotically equivalent in probability, $\widehat{\mathrm{MMD}}^2$ has strictly lower variance on finite samples if $\langle C_P, C_Q \rangle \neq 0$; when $n_X \neq n_Y$, the variance gap can be large.

Leveraging our theoretical results, we derived an approximation to the asymptotic test power, which is (nearly) monotonic in the signal-to-noise ratio MMD^2 / σ. This yields a power-optimization scheme based on maximizing $\widehat{\mathrm{MMD}}^2 / (\hat{\sigma} + \lambda)$, with which we demonstrated improved performance in distinguishing synthetic Gaussians and CIFAR image data.

The main theoretical aspect we did not fully characterize is determining the conditions under which ζ_X and ζ_Y are positive, or have matching signs. Theorem 2.7 makes significant progress towards this, in the case of continuous kernels (with overlapping support) and/or analytic kernels (with potentially non-overlapping support); typical kernels defined by neural networks will be continuous but not analytic, leaving the disjoint-support question with such kernels unknown. These questions, however, seem to be of particular interest only when the supports are very complex; distributions with “simple” disjoint supports are easy to distinguish.

References

N.H. Anderson, P. Hall, and D.M. Titterington (1994). “Two-Sample Test Statistics for Measuring Discrepancies Between Two Multivariate Probability Density Functions Using Kernel-Based Density Estimates.” In *Journal of Multivariate Analysis* 50.1, pp. 41–54.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton (2018). “Demystifying MMD GANs.” In *ICLR*. arXiv: `1801.01401`.

Kacper Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton (2015). “Fast Two-Sample Testing with Analytic Representations of Probability Measures.” In *NeurIPS*. arXiv: `1506.04725`.

Namrata Deka and Danica J. Sutherland (2023). "MMD-B-Fair: Learning Fair Representations with Statistical Testing." In *AISTATS*. arXiv: 2211.07907.

Nelson Dunford and Jacob T. Schwartz (1988). *Linear Operators, Part 2: Spectral Theory, Self Adjoint Operators in Hilbert Space*. Wiley Classics Library. Wiley.

Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani (2015). "Training generative neural networks via Maximum Mean Discrepancy optimization." In *UAI*. arXiv: 1505.03906.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola (2012). "A kernel two-sample test." In *Journal of Machine Learning Research* 13.Mar, pp. 723–773.

Arthur Gretton, Kenji Fukumizu, Zaïd Harchaoui, and Bharath K. Sriperumbudur (2009). "A Fast, Consistent Kernel Two-Sample Test." In *NeurIPS*.

Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K. Sriperumbudur (2012). "Optimal kernel choice for large-scale two-sample tests." In *NeurIPS*.

Zheng He, Roman Pogodin, Yazhe Li, Namrata Deka, Arthur Gretton, and Danica J. Sutherland (2025). "On the Hardness of Conditional Independence Testing in Practice." In *NeurIPS*. arXiv: 2512.14000.

Jesse Hemerik and Jelle Goeman (2018). "Exact testing with random permutations." In *TEST* 27, pp. 811–825.

Wassily Hoeffding (1948). "A Class of Statistics with Asymptotically Normal Distribution." In *Annals of Mathematical Statistics* 19, pp. 308–334.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar (2024). "Rethinking FID: Towards a Better Evaluation Metric for Image Generation." In *CVPR*.

Ilmun Kim, Sivaraman Balakrishnan, and Larry Wasserman (2022). "Minimax optimality of permutation tests." In *Annals of Statistics* 50.1, pp. 225–251.

Hirofumi Kobayashi, Cheng Lei, Yi Wu, Ailin Mao, Yiyue Jiang, Baoshan Guo, Yasuyuki Ozeki, and Keisuke Goda (2017). "Label-free detection of cellular drug responses by high-throughput bright-field imaging and machine learning." In *Scientific Reports* 7.1. DOI: 10.1038/s41598-017-12378-4.

Alex Krizhevsky (2009). "Learning Multiple Layers of Features from Tiny Image." URL: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Jonas M. Kübler, Vincent Stimper, Simon Buchholz, Krikamol Muandet, and Bernhard Schölkopf (2022). "AutoML Two-Sample Test." In *NeurIPS*. arXiv: 2206.08843.

A.J. Lee (2019). *U-Statistics: Theory and Practice*. CRC Press.

E. L. Lehmann (1951). "Consistency and Unbiasedness of Certain Nonparametric Tests." In *The Annals of Mathematical Statistics* 22.2, pp. 165–179.

Yujia Li, Kevin Swersky, and Richard Zemel (2015). "Generative Moment Matching Networks." In *ICML*. arXiv: 1502.02761.

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J. Sutherland (2020). "Learning Deep Kernels for Non-Parametric Two-Sample Tests." In *ICML*. arXiv: 2002.09116.

Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S. Yu (2013). "Transfer Feature Learning with Joint Distribution Adaptation." In *ICCV*. DOI: 10.1109/ICCV.2013.274.

Boris Mityagin (2015). "The Zero Set of a Real Analytic Function." arXiv: 1512.07276.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf (2017). "Kernel Mean Embedding of Distributions: A Review and Beyond." In *Foundations and Trends® in Machine Learning* 10.1–2, pp. 1–141. DOI: 10.1561/2200000060.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar (2019). "Do ImageNet Classifiers Generalize to ImageNet?" In *ICML*. arXiv: 1902.10811.

Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, and Arthur Gretton (2023). "MMD Aggregated Two-Sample Test." In *Journal of Machine Learning Research* 24.194, pp. 1–81.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu (2013). "Equivalence of distance-based and RKHS-based statistics in hypothesis testing." In *The Annals of Statistics* 41.5, pp. 2263–2291.

Pranab Kumar Sen (1974). "Weak Convergence of Generalized U-Statistics." In *The Annals of Probability* 2.1, pp. 90–102.

Robert J. Serfling (1980). *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Mathematical Statistics. Hoboken, New Jersey: Wiley. DOI: 10.1002/9780470316481.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R.G. Lanckriet (2011). “Universality, Characteristic Kernels and RKHS Embedding of Measures.” In *Journal of Machine Learning Research* 12.70, pp. 2389–2410.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert Lanckriet, and Bernhard Schölkopf (2008). “Injective Hilbert Space Embeddings of Probability Measures.” In *COLT*.

Ingo Steinwart and Andreas Christmann (2008). *Support Vector Machines*. Springer New York.

Danica J. Sutherland and Namrata Deka (2019). “Unbiased estimators for the variance of MMD estimators.” arXiv: `1906.02104`.

Danica J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton (2017). “Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy.” In *ICLR*. arXiv: `1611.04488`.

Gábor J. Székely and Maria L. Rizzo (2013). “Energy statistics: A class of statistics based on distances.” In *Journal of Statistical Planning and Inference* 143.8, pp. 1249–1272.

A Proof of Theorem 2.3

Theorem 2.3. *When the number of samples from P and Q are the same, we have that*

$$\Pr_{\substack{X_1,\dots,X_n\sim P\\Y_1,\dots,Y_n\sim Q}}\left(|\widehat{\mathrm{MMD}}^2-\widehat{\mathrm{MMD}}_{\mathrm{U}}{}^2|\le\frac{8\sup_{x\in\mathcal{X}}k(x,x)}{n^{3/2}}\sqrt{\log\frac{2}{\delta}}\right)\ge 1-\delta.$$

Proof. Letting $K=\sup_{x\in\mathcal{X}}k(x,x)$, we have that

$$k(x,y)=\langle k(x,\cdot),k(y,\cdot)\rangle\le\|k(x,\cdot)\|\|k(y,\cdot)\|=\sqrt{k(x,x)k(y,y)}\le K.$$

The difference between the estimators is

$$\begin{aligned}\widehat{\mathrm{MMD}}^2-\widehat{\mathrm{MMD}}_{\mathrm{U}}{}^2&=-\frac{2}{n^2}\sum_{i,j}k(x_i,y_j)+\frac{2}{n(n-1)}\sum_{i\neq j}k(x_i,y_j)\\&=-\frac{2}{n^2}\sum_{i=1}^n k(x_i,y_i)+\frac{2}{n}\left(\frac{1}{n-1}-\frac{1}{n}\right)\sum_{i\neq j}k(x_i,y_j)\\&=\frac{2}{n}\left(-\frac{1}{n}\sum_{i=1}^n k(x_i,y_i)+\frac{1}{n(n-1)}\sum_{i\neq j}k(x_i,y_j)\right).\end{aligned}$$

This sum satisfies bounded differences: if we consider changing a single x_i, it changes a single term in the first sum by at most $2K$, meaning that the first average changes by at most $2K/n$. At most $n-1$ terms change in the second sum, again each by up to $2K$, giving again a total change of at most $2K/n$. Thus the overall difference in estimators changes by at most $8K/n^2$. The same holds for changing a single y_i. This difference is also mean zero, so since all of the $2n$ arguments x_i, y_i are mutually independent, we can apply McDiarmid's inequality to get

$$\Pr\left(|\widehat{\mathrm{MMD}}^2-\widehat{\mathrm{MMD}}_{\mathrm{U}}{}^2|\le\frac{8K}{n^2}\sqrt{\frac{1}{2}(2n)\log\frac{2}{\delta}}\right)\ge 1-\delta.$$ ☐

B Proof of U-statistic variance decomposition (Proposition 2.6)

Proposition 2.6. *In Setting A, we have that*

$$\operatorname*{Var}_{\substack{\mathbf{S}_P\sim P^n\\\mathbf{S}_Q\sim Q^n}}\left(\widehat{\mathrm{MMD}}_{\mathrm{U}}{}^2(\mathbf{S}_P,\mathbf{S}_Q)\right)=\frac{4}{n}\langle\mu_P-\mu_Q,(C_P+C_Q)(\mu_P-\mu_Q)\rangle_{\mathcal{H}}+\frac{2}{n(n-1)}\|C_P+C_Q\|_{\mathrm{HS}}^2. \tag{4}$$

Proof. This result follows directly from Theorem 6.1 of He et al. (2025), but we reproduce their approach here for completeness.

Define $\delta_{xy}=k(x,\cdot)-k(y,\cdot)\in\mathcal{H}$, for any $x,y\in\mathcal{X}$. Then h of (2) is $h((x,y),(x',y'))=\langle\delta_{xy},\delta_{x'y'}\rangle_{\mathcal{H}}$. We also have that $\mathbb{E}\,\delta_{XY}=\mathbb{E}\,k(X,\cdot)-\mathbb{E}\,k(Y,\cdot)=\mu_P-\mu_Q$ and

$$\begin{aligned}\mathbb{E}\,\delta_{XY}\otimes\delta_{XY}&=\mathbb{E}\,k(X,\cdot)\otimes k(X,\cdot)-\mathbb{E}\,k(X,\cdot)\otimes k(Y,\cdot)-\mathbb{E}\,k(Y,\cdot)\otimes k(X,\cdot)+\mathbb{E}\,k(Y,\cdot)\otimes k(Y,\cdot)\\&=(C_P+\mu_P\otimes\mu_P)-\mu_P\otimes\mu_Q-\mu_Q\otimes\mu_P+(C_Q+\mu_Q\otimes\mu_Q)\\&=C_P+C_Q+(\mu_P-\mu_Q)\otimes(\mu_P-\mu_Q).\end{aligned}$$

For the first term in the variance decomposition of Proposition 2.5,

$$\operatorname*{\mathbb{E}}_{X',Y'}[h((x,y),(X',Y'))]=\langle\delta_{xy},\mu_P-\mu_Q\rangle_{\mathcal{H}}$$

$$\operatorname*{Var}_{X,Y}\left[\operatorname*{\mathbb{E}}_{X',Y'}[h((X,Y),(X',Y'))\mid X,Y]\right]=\langle\mu_P-\mu_Q,(C_P+C_Q)(\mu_P-\mu_Q)\rangle_{\mathcal{H}}.$$

For the second term,

$$\begin{aligned}
\operatorname*{Var}_{X',Y'}[h((x,y),(X',Y'))] &= \operatorname*{Var}_{X',Y'}[\langle \delta_{xy}, \delta_{X'Y'}\rangle] \\
&= \mathbb{E}\left[\langle \delta_{xy}, \delta_{X'Y'}\rangle^2\right] - \left(\mathbb{E}\langle \delta_{xy}, \delta_{X'Y'}\rangle\right)^2 \\
&= \mathbb{E}\langle \delta_{xy}, \delta_{X'Y'}\rangle\langle \delta_{X'Y'}, \delta_{xy}\rangle - \langle \delta_{xy}, \mu_P - \mu_Q\rangle^2 \\
&= \mathbb{E}\langle \delta_{xy}, (\delta_{X'Y'} \otimes \delta_{X'Y'})\delta_{xy}\rangle - \langle \delta_{xy}, ((\mu_P - \mu_Q) \otimes (\mu_P - \mu_Q))\, \delta_{xy}\rangle \\
&= \big\langle \delta_{xy}, \big(\mathbb{E}(\delta_{X'Y'} \otimes \delta_{X'Y'}) - (\mu_P - \mu_Q) \otimes (\mu_P - \mu_Q)\big)\delta_{xy}\big\rangle \\
&= \big\langle \delta_{xy}, \big(C_P + C_Q\big)\delta_{xy}\big\rangle \\
&= \langle \delta_{xy} \otimes \delta_{xy}, C_P + C_Q\rangle_{\mathrm{HS}} \\
\mathop{\mathbb{E}}_{X,Y}\left[\operatorname*{Var}_{X',Y'}[h((X,Y),(X',Y')) \mid X,Y]\right] &= \left\langle \mathop{\mathbb{E}}_{X,Y}\left[\delta_{XY} \otimes \delta_{XY}\right], C_P + C_Q \right\rangle_{\mathrm{HS}} \\
&= \|C_P + C_Q\|_{\mathrm{HS}}^2 + \langle (\mu_P - \mu_Q) \otimes (\mu_P - \mu_Q), C_P + C_Q\rangle_{\mathrm{HS}} \\
&= \|C_P + C_Q\|_{\mathrm{HS}}^2 + \langle \mu_P - \mu_Q, (C_P + C_Q)(\mu_P - \mu_Q)\rangle_{\mathcal{H}}.
\end{aligned}$$

Letting $\nu = \langle \mu_P - \mu_Q, (C_P + C_Q)(\mu_P - \mu_Q)\rangle_{\mathcal{H}}$ for brevity, we have obtained (4):

$$\operatorname{Var}\left(\widehat{\mathrm{MMD}}_{\mathrm{U}}^2\right) = \frac{4n-6}{n(n-1)}\nu + \frac{2}{n(n-1)}\left(\|C_P + C_Q\|_{\mathrm{HS}}^2 + \nu\right) = \frac{4}{n}\nu + \frac{2}{n(n-1)}\|C_P + C_Q\|_{\mathrm{HS}}^2. \qquad \square$$

C Proof of degeneracy characterization (Theorem 2.7)

Theorem 2.7. *In Setting A, $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$ is infinitely degenerate (the variance is zero) if and only if $C_P = C_Q = \mathbf{0}$. Note that an infinitely degenerate MMD estimate may still be nonzero, such as if P and Q are distinct point masses.*

Now suppose that at least one of C_P, C_Q is nonzero, so $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$'s order of degeneracy is either zero or one. Then:

(i) If $\mu_P = \mu_Q$, $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$ is first-order degenerate.

(ii) When $\mu_P \neq \mu_Q$, $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$ may be either non-degenerate or first-order degenerate.

(iii) Suppose $\mathcal{X}$ is a topological space, $k(x,\cdot)$ is continuous for each x, and $\sup_{x\in\mathcal{X}} k(x,x)$ is finite. Further assume the supports of P and Q are not disjoint. Then $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$ is degenerate if and only if $\mu_P = \mu_Q$.

(iv) In Setting B, $\widehat{\mathrm{MMD}}_{\mathrm{U}}^2$ is degenerate if and only if $\mu_P = \mu_Q$.

(v) In Setting B, $\langle \mu_P - \mu_Q, C_P(\mu_P - \mu_Q)\rangle > 0$ if and only if $\langle \mu_P - \mu_Q, C_Q(\mu_P - \mu_Q)\rangle > 0$.

Proof. Because C_P and C_Q are each positive semi-definite, $C_P + C_Q = 0$ if and only if $C_P = 0 = C_Q$. Proposition 2.6 then immediately gives both the result about infinite degeneracy and (i).

Proof of (ii) Most usual settings are non-degenerate; for an explicit example, consider $\mathcal{X} = \mathbb{R}$, $P = \mathcal{N}(0,1)$, $Q = \mathcal{N}(1,1)$, and $k(x,y) = xy$. Then $\mu_P = (t \mapsto 0) = \mathbf{0}$, $\mu_Q = (t \mapsto t)$, $C_P = C_Q = \mathrm{Id}$, and so $\langle \mu_P - \mu_Q, (C_P + C_Q)(\mu_P - \mu_Q)\rangle = 2\|\mu_Q\|_{\mathcal{H}}^2 = 2\|k(\cdot,1)\|_{\mathcal{H}}^2 = 2k(1,1) = 2$.

For a degenerate case, consider $\mathcal{X} = \mathbb{R}$, $k(x,y) = \max(1 - |x-y|, 0)$, $P = \mathrm{Uniform}(\{1,2\})$, and $Q = \mathrm{Uniform}(\{3,4\})$. These have nonzero kernel covariance operators: for instance,

$$\langle k(1,\cdot), C_P k(1,\cdot)\rangle = \operatorname*{Var}_{X\sim P}[k(1,X)] = \tfrac{1}{2}\cdot 1^2 + \tfrac{1}{2}\cdot 0^2 - \left(\tfrac{1}{2}\cdot 1 + \tfrac{1}{2}\cdot 0\right)^2 = \tfrac{1}{2} - \tfrac{1}{4} = \tfrac{1}{4} > 0.$$

We also have $\mu_P \neq \mu_Q$, as can be seen by considering their inner products with e.g. $k(2.2, \cdot)$, or because $P \neq Q$ and k is characteristic (Sriperumbudur et al., 2008, Corollary 8). Notice, however, that $\Pr_{X \sim P, Y \sim Q}(k(X, Y) = 0) = 1$. Thus, if we write

$$\widehat{\mathrm{MMD}}_{\mathrm{U}}{}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j) - \frac{2}{n^2} \sum_{i \neq j} k(x_i, y_j),$$

the last sum is identically zero, leaving us with the sum of two independent U-statistics with kernel k. Applying Proposition 2.5 for the first such sum, $\mathrm{Var}[\mathbb{E}[k(X, X') \mid X]] = \mathrm{Var}_{X' \sim P}[\mu_P(X')] = 0$ because $\mu_P(1) = \mu_P(2) = \frac{1}{2}$, and $P = \mathrm{Uniform}(\{1, 2\})$; however, for any $x \in \{1, 2\}$, $\mathrm{Var}[k(x, X)] = \frac{1}{4}$ as seen above, and hence the variance is $\frac{2}{n(n-1)} \cdot \frac{1}{4} = \frac{1}{2n(n-1)}$. The other term is independent and has the same variance. Thus the variance of $\widehat{\mathrm{MMD}}_{\mathrm{U}}{}^2$ is $1/(n(n-1))$, so it must be first-order degenerate: we have a situation where $\mu_P \neq \mu_Q$, $C_P \neq 0$, $C_Q \neq 0$, and yet $\widehat{\mathrm{MMD}}_{\mathrm{U}}{}^2$ is degenerate.

Proof of (iii) and (iv) If $\mu_P = \mu_Q$, then Proposition 2.6 immediately implies that $\widehat{\mathrm{MMD}}_{\mathrm{U}}{}^2$ is degenerate. Thus, suppose that $\widehat{\mathrm{MMD}}_{\mathrm{U}}{}^2$ is degenerate for C_P, C_Q not both zero; we will show that $\mu_P = \mu_Q$, or equivalently that $\Delta = \mu_P - \mu_Q$ is the zero function. Since C_P and C_Q are positive semi-definite and we have assumed degeneracy, Proposition 2.6 implies that we must have

$$\langle \mu_P - \mu_Q, C_P(\mu_P - \mu_Q) \rangle = \operatorname*{Var}_{X \sim P}(\Delta(X)) = 0 \quad \text{and} \quad \langle \mu_P - \mu_Q, C_Q(\mu_P - \mu_Q) \rangle = \operatorname*{Var}_{Y \sim Q}(\Delta(Y)) = 0.$$

Thus, there must be $c_P, c_Q \in \mathbb{R}$ such that

$$\Pr_{X \sim P}(\Delta(X) = c_P) = 1 \qquad \text{and} \qquad \Pr_{Y \sim Q}(\Delta(Y) = c_Q) = 1.$$

But notice that

$$\|\mu_P - \mu_Q\|^2 = \langle \mu_P, \Delta \rangle - \langle \mu_Q, \Delta \rangle = \mathop{\mathbb{E}}_{X \sim P}[\Delta(X)] - \mathop{\mathbb{E}}_{Y \sim Q}[\Delta(Y)] = c_P - c_Q,$$

and so if $c_P = c_Q$ then $\mu_P = \mu_Q$.

For (iii), let x be in the support of P, and suppose that $\Delta = \mu_P - \mu_Q$ has $\Delta(x) \neq c_P$. But then by continuity there is an open neighbourhood $N_x \ni x$ for which $|\Delta(x') - c_P| > \epsilon$ for all $x' \in N_x$; since x is in the support of P, $P(N_x) > 0$, contradicting that $\Delta(x) = c_P$ P-almost surely. We similarly have $\Delta(y) = c_Q$ for all y in the support of Q; thus, if there is a point in both supports, we must have $c_P = c_Q$.

For (iv), k being bounded and real-analytic implies that each $f \in \mathcal{H}$ is as well (Chwialkowski et al., 2015, Lemma 1). As before, degeneracy implies $\Delta(x) = c_P$ for all x in the support of P, and $\Delta(y) = c_Q$ for all y in the support of Q. Mityagin (2015) shows that if a real-analytic function is zero on a set of positive Lebesgue measure, the function is identically zero; since $\Delta - c_P$ is analytic, it must be zero on all of $\mathcal{X}$. But since $\mathbb{R}^d$ is Hausdorff, the support of Q is nonempty, and hence $c_Q = c_P$.

Proof of (v): Reusing our previous notation, positive semi-definiteness implies $\langle \Delta, C_P \Delta \rangle, \langle \Delta, C_Q \Delta \rangle$ are either zero or positive. Suppose $\langle \Delta, C_P \Delta \rangle = 0$; then following the proof of (iv) we know that the function Δ is almost surely a constant c_P on the support of P, and since it is also real-analytic and P's support has positive Lebesgue measure, it is equal to c_P everywhere. Thus $\langle \Delta, C_Q \Delta \rangle = \mathrm{Var}_{Y \sim Q} \Delta(Y) = 0$. The same reasoning applies for Q, and so $\langle \Delta, C_Q \Delta \rangle = 0$ implies $\langle \Delta, C_P \Delta \rangle = 0$. □

D Two Sample MMD Variance

We now decompose the variance of $\widehat{\mathrm{MMD}}^2$ based on the formula (7).

Theorem 3.4. *Under Setting A, it holds for any $n_X, n_Y \geq 2$ that*

$$\begin{aligned}
\operatorname*{Var}_{\substack{\mathbf{S}_P \sim P^{n_X} \\ \mathbf{S}_Q \sim Q^{n_Y}}}(\widehat{\mathrm{MMD}}^2(\mathbf{S}_P, \mathbf{S}_Q)) &= \frac{4}{n_X}\left(1 - \frac{4}{n_Y} \cdot \frac{n_X - 2}{n_X - 1} \cdot \frac{n_Y - 2}{n_Y - 1}\right)\langle \mu_P - \mu_Q, C_P(\mu_P - \mu_Q)\rangle \\
&+ \frac{4}{n_Y}\left(1 - \frac{4}{n_X} \cdot \frac{n_X - 2}{n_X - 1} \cdot \frac{n_Y - 2}{n_Y - 1}\right)\langle \mu_P - \mu_Q, C_Q(\mu_P - \mu_Q)\rangle \\
&+ \frac{2}{n_X(n_X - 1)}\|C_P\|_{\mathrm{HS}}^2 + \frac{2}{n_Y(n_Y - 1)}\|C_Q\|_{\mathrm{HS}}^2 + \frac{4}{n_X n_Y}\langle C_P, C_Q\rangle_{\mathrm{HS}} \\
&+ \frac{16}{n_X n_Y}\left(\frac{n_X - 2}{n_X - 1} \cdot \frac{n_Y - 2}{n_Y - 1}\right)\left[\langle \mu_P, C_P \mu_P\rangle + \langle \mu_Q, C_Q \mu_Q\rangle\right].
\end{aligned}$$

Proof. We will need to compute various conditional expectations of the MMD estimator kernel function $h(x, x'; y, y') = \langle k(x, \cdot) - k(y, \cdot), k(x', \cdot) - k(y', \cdot)\rangle$. For brevity, in this proof we will use $\phi(x)$ to denote the feature map $k(x, \cdot)$. We can omit some calculations because of the additional symmetry $h(x, x'; y, y') = h(y, y'; x, x')$.

As C_P, C_Q are self-adjoint, we have the useful identity

$$\begin{aligned}
\langle \mu_P, C_P \mu_P\rangle + \langle \mu_Q, C_P \mu_Q\rangle - 2\langle \mu_Q, C_P \mu_P\rangle &= [\langle \mu_P, C_P \mu_P\rangle - \langle \mu_P, C_P \mu_Q\rangle] + [\langle \mu_Q, C_P \mu_Q\rangle - \langle \mu_Q, C_P \mu_P\rangle] \\
&= \langle \mu_P, C_P(\mu_P - \mu_Q)\rangle + \langle \mu_Q, C_P(\mu_Q - \mu_P)\rangle \\
&= \langle \mu_P - \mu_Q, C_P(\mu_P - \mu_Q)\rangle,
\end{aligned}$$

and similarly $\langle \mu_Q, C_Q \mu_Q\rangle + \langle \mu_P, C_Q \mu_P\rangle - 2\langle \mu_P, C_Q \mu_Q\rangle = \langle \mu_P - \mu_Q, C_Q(\mu_P - \mu_Q)\rangle$.

Order 1 terms.

$$\begin{aligned}
\mathbb{E}(h(X, X'; Y, Y') \mid X) &= \mu_P(X) + \mathbb{E}\, k(Y, Y') - \mu_Q(X) - \mathbb{E}\, k(X', Y) \\
\zeta_X &= \mathrm{Var}(\mu_P(X) - \mu_Q(X) + \mathrm{const}) \\
&= \langle \mu_P - \mu_Q, C_P(\mu_P - \mu_Q)\rangle
\end{aligned}$$

Order 2 terms.

$$\begin{aligned}
\mathbb{E}(h(X, X'; Y, Y') | X, X') &= k(X, X') + \mathbb{E}\, k(Y, Y') - \mu_Q(X) - \mu_Q(X') \\
\mathrm{Var}(k(X, X')) &= \mathbb{E}[k(X, X')^2] - (\mathbb{E}[k(X, X')])^2 \\
&= \mathbb{E}\langle \phi(X) \otimes \phi(X), \phi(X') \otimes \phi(X')\rangle_{\mathrm{HS}} - \|\mu_P\|^4 \\
&= (\|C_P\|_{\mathrm{HS}}^2 + 2\langle C_P, \mu_P \otimes \mu_P\rangle_{\mathrm{HS}} + \|\mu_P\|^4) - \|\mu_P\|^4 \\
&= \|C_P\|_{\mathrm{HS}}^2 + 2\langle \mu_P, C_P \mu_P\rangle \\
\mathrm{Cov}(\mu_Q(X), k(X, X')) &= \mathbb{E}[\mu_Q(X) k(X, X')] - \mathbb{E}[k(X, Y)]\, \mathbb{E}[k(X, X')] \\
&= \langle \mu_Q, C_P \mu_P\rangle \\
\zeta_{XX'} &= \mathrm{Var}(k(X, X')) + \mathrm{Var}(\mu_Q(X)) + \mathrm{Var}(\mu_Q(X')) \\
&\quad - 2\,\mathrm{Cov}(\mu_Q(X), k(X, X')) - 2\,\mathrm{Cov}(\mu_Q(X'), k(X, X')) \\
&= \|C_P\|_{\mathrm{HS}}^2 + 2\langle \mu_P, C_P \mu_P\rangle + 2\langle \mu_Q, C_P \mu_Q\rangle - 4\langle \mu_Q, C_P \mu_P\rangle \\
&= \|C_P\|_{\mathrm{HS}}^2 + 2\langle \mu_P - \mu_Q, C_P(\mu_P - \mu_Q)\rangle
\end{aligned}$$

Let $\tilde{C}_P = \mathbb{E}[\phi(X) \otimes \phi(X)]$ and define $\tilde{C}_Q$ analogously.

$$
\mathbb{E}(h(X, X'; Y, Y')|X, Y) = \mu_P(X) + \mu_Q(Y) - \frac{1}{2}\left[\mu_Q(X) + \mu_P(Y) + k(X, Y) + \mathbb{E}\, k(X, Y)\right]
$$

$$
\begin{aligned}
\langle C_P, C_Q \rangle_{\mathrm{HS}} &= \langle \tilde{C}_P, \tilde{C}_Q \rangle_{\mathrm{HS}} - \langle \tilde{C}_P, \mu_Q \otimes \mu_Q \rangle_{\mathrm{HS}} - \langle \tilde{C}_Q, \mu_P \otimes \mu_P \rangle_{\mathrm{HS}} + \langle \mu_P, \mu_Q \rangle^2 \\
&= \mathbb{E}[k(X, Y)^2] - \langle C_P, \mu_Q \otimes \mu_Q \rangle_{\mathrm{HS}} - \langle C_Q, \mu_P \otimes \mu_P \rangle_{\mathrm{HS}} - \langle \mu_P, \mu_Q \rangle^2 \\
&= \mathrm{Var}(k(X, Y)) - \langle \mu_Q, C_P \mu_Q \rangle - \langle \mu_P, C_Q \mu_P \rangle
\end{aligned}
$$

$$
\begin{aligned}
\zeta_{XY} &= \mathrm{Var}\left((\mu_P(X) - \frac{1}{2}\mu_Q(X) \right) + \mathrm{Var}\left(\mu_Q(Y) - \frac{1}{2}\mu_P(Y) \right) + \frac{1}{4}\mathrm{Var}(k(X, Y)) \\
&\quad + \mathrm{Cov}\left(\mu_P(X) - \frac{1}{2}\mu_Q(X), k(X, Y) \right) + \mathrm{Cov}\left(\mu_Q(Y) - \frac{1}{2}\mu_P(Y), k(X, Y) \right) \\
&= \langle \mu_P - \frac{1}{2}\mu_Q, C_P(\mu_P - \frac{1}{2}\mu_Q) \rangle + \langle \mu_Q - \frac{1}{2}\mu_P, C_Q(\mu_Q - \frac{1}{2}\mu_P) \rangle \\
&\quad + \frac{1}{4}\langle C_P, C_Q \rangle_{\mathrm{HS}} + \frac{1}{4}\langle \mu_Q, C_P \mu_Q \rangle + \frac{1}{4}\langle \mu_P, C_Q \mu_P \rangle + \langle \mu_P - \frac{1}{2}\mu_Q, C_P \mu_Q \rangle \\
&\quad + \langle \mu_Q - \frac{1}{2}\mu_P, C_Q \mu_P \rangle \\
&= \frac{1}{4}\langle C_P, C_Q \rangle_{\mathrm{HS}} + \langle \mu_P, C_P \mu_P \rangle + \langle \mu_Q, C_Q \mu_Q \rangle,
\end{aligned}
$$

where the last line follows by

$$
\begin{aligned}
&\langle \mu_P - \frac{1}{2}\mu_Q, C_P(\mu_P - \frac{1}{2}\mu_Q) \rangle + \frac{1}{4}\langle \mu_Q, C_P \mu_Q \rangle + \langle \mu_P - \frac{1}{2}\mu_Q, C_P \mu_Q \rangle \\
&= \left[\langle \mu_P - \frac{1}{2}\mu_Q, C_P(\mu_P - \frac{1}{2}\mu_Q) \rangle + \langle \mu_P - \frac{1}{2}\mu_Q, C_P \frac{1}{2}\mu_Q \rangle \right] + \left[\frac{1}{4}\langle \mu_Q, C_P \mu_Q \rangle + \langle \frac{1}{2}\mu_P - \frac{1}{4}\mu_Q, C_P \mu_Q \rangle \right] \\
&= \langle \mu_P - \frac{1}{2}\mu_Q, C_P \mu_P \rangle + \langle \frac{1}{2}\mu_P, C_P \mu_Q \rangle = \langle \mu_P - \frac{1}{2}\mu_Q, C_P \mu_P \rangle + \langle \frac{1}{2}\mu_Q, C_P \mu_P \rangle = \langle \mu_P, C_P \mu_P \rangle.
\end{aligned}
$$

Order 3 terms.

$$
\mathbb{E}(h(X, X'; Y, Y') \mid X, X', Y) = k(X, X') + \mu_Q(Y) - \frac{1}{2}[k(X, Y) + \mu_Q(X') + k(X', Y) + \mu_Q(X)]
$$

Let $R := k(X, X') - \frac{1}{2}[k(X, Y) + k(X', Y)]$, then we may write

$$
\begin{aligned}
\zeta_{XX'Y} &= \mathrm{Var}(\mu_Q(Y)) + \frac{1}{4}\mathrm{Var}(\mu_Q(X')) + \frac{1}{4}\mathrm{Var}(\mu_Q(X)) + \mathrm{Var}(R) \\
&\quad + 2\,\mathrm{Cov}(\mu_Q(Y), R) - \mathrm{Cov}(\mu_Q(X'), R) - \mathrm{Cov}(\mu_Q(X), R) \\
&= \langle \mu_Q, C_Q \mu_Q \rangle + \frac{1}{2}\langle \mu_Q, C_P \mu_Q \rangle + \mathrm{Var}(R) + 2\,\mathrm{Cov}(\mu_Q(Y), R) \\
&\quad - 2\,\mathrm{Cov}(\mu_Q(X'), R).
\end{aligned}
$$

We may simplify the terms including R

$$\begin{aligned}\operatorname{Var}(R) &= \operatorname{Var}(k(X,X')) + \frac{1}{4}\operatorname{Var}(k(X,Y)) + \frac{1}{4}\operatorname{Var}(k(X',Y)) - \operatorname{Cov}(k(X,X'),k(X,Y)) \\ &\quad - \operatorname{Cov}(k(X,X'),k(X',Y)) + \frac{1}{2}\operatorname{Cov}(k(X,Y),k(X',Y)) \\ &= \|C_P\|_{\mathrm{HS}}^2 + 2\langle \mu_P, C_P\mu_P\rangle + \frac{1}{2}\left[\langle C_P, C_Q\rangle_{\mathrm{HS}} + \langle \mu_Q, C_P\mu_Q\rangle + \langle \mu_P, C_Q\mu_P\rangle\right] \\ &\quad - 2\langle \mu_P, C_P\mu_Q\rangle + \frac{1}{2}\langle \mu_P, C_Q\mu_P\rangle \\ &= \|C_P\|_{\mathrm{HS}}^2 + 2\langle \mu_P, C_P\mu_P\rangle + \frac{1}{2}\langle C_P, C_Q\rangle_{\mathrm{HS}} + \frac{1}{2}\langle \mu_Q, C_P\mu_Q\rangle + \langle \mu_P, C_Q\mu_P\rangle - 2\langle \mu_P, C_P\mu_Q\rangle\end{aligned}$$

$$\operatorname{Cov}(\mu_Q(Y),R) = \operatorname{Cov}\left(\mu_Q(Y), -\frac{1}{2}[k(X,Y)+k(X',Y)]\right) = -\operatorname{Cov}(\mu_Q(Y),k(X,Y)) = -\langle \mu_Q, C_Q\mu_P\rangle$$

$$\operatorname{Cov}(\mu_Q(X'),R) = \operatorname{Cov}\left(\mu_Q(X'), k(X,X') - \frac{1}{2}k(X',Y)\right) = \langle \mu_Q, C_P\mu_P\rangle - \frac{1}{2}\langle \mu_Q, C_P\mu_Q\rangle$$

Plugging the above into the variance equation gives

$$\begin{aligned}\zeta_{XX'Y} &= \langle \mu_Q, C_Q\mu_Q\rangle + \frac{1}{2}\langle \mu_Q, C_P\mu_Q\rangle + \|C_P\|_{\mathrm{HS}}^2 + 2\langle \mu_P, C_P\mu_P\rangle \\ &\quad + \frac{1}{2}\langle C_P, C_Q\rangle_{\mathrm{HS}} + \frac{1}{2}\langle \mu_Q, C_P\mu_Q\rangle + \langle \mu_P, C_Q\mu_P\rangle - 2\langle \mu_P, C_P\mu_Q\rangle \\ &\quad - 2\langle \mu_Q, C_Q\mu_P\rangle - 2\langle \mu_Q, C_P\mu_P\rangle + \langle \mu_Q, C_P\mu_Q\rangle \\ &= \|C_P\|_{\mathrm{HS}}^2 + \frac{1}{2}\langle C_P, C_Q\rangle_{\mathrm{HS}} + 2[\langle \mu_P, C_P\mu_P\rangle + \langle \mu_Q, C_P\mu_Q\rangle - 2\langle \mu_P, C_P\mu_Q\rangle] \\ &\quad + [\langle \mu_Q, C_Q\mu_Q\rangle + \langle \mu_P, C_Q\mu_P\rangle - 2\langle \mu_Q, C_Q\mu_P\rangle] \\ &= \|C_P\|_{\mathrm{HS}}^2 + \frac{1}{2}\langle C_P, C_Q\rangle_{\mathrm{HS}} + 2\langle \mu_P - \mu_Q, C_P(\mu_P - \mu_Q)\rangle \\ &\quad + \langle \mu_P - \mu_Q, C_Q(\mu_P - \mu_Q)\rangle \\ &= \|C_P\|_{\mathrm{HS}}^2 + \frac{1}{2}\langle C_P, C_Q\rangle_{\mathrm{HS}} + \langle \mu_P - \mu_Q, (2C_P + C_Q)(\mu_P - \mu_Q)\rangle\end{aligned}$$

Order 4 terms. Let $R' := k(Y,Y') - \frac{1}{2}[k(X',Y') + k(X,Y')]$, we have $h(X,X';Y,Y') = R + R'$. Similar to the calculation for R we have

$$\operatorname{Var}(R') = \|C_Q\|_{\mathrm{HS}}^2 + 2\langle \mu_Q, C_Q\mu_Q\rangle + \frac{1}{2}\langle C_P, C_Q\rangle_{\mathrm{HS}} + \frac{1}{2}\langle \mu_Q, C_P\mu_Q\rangle + \langle \mu_P, C_Q\mu_P\rangle - 2\langle \mu_Q, C_Q\mu_P\rangle.$$

Next we have

$$\begin{aligned}2\operatorname{Cov}(R,R') &= 2\operatorname{Cov}\left(k(X,X') - \frac{1}{2}[k(X,Y)+k(X',Y)], k(Y,Y') - \frac{1}{2}[k(X',Y')+k(X,Y')]\right) \\ &= -\operatorname{Cov}(k(X,X'),k(X',Y')+k(X,Y')) - \operatorname{Cov}(k(Y,Y'),k(X,Y)+k(X',Y)) \\ &\quad + \frac{1}{2}\operatorname{Cov}(k(X,Y)+k(X',Y),k(X',Y')+k(X,Y')) \\ &= \\ &= -2\langle \mu_P, C_P\mu_Q\rangle - 2\langle \mu_Q, C_Q\mu_P\rangle + \langle \mu_Q, C_P\mu_Q\rangle\end{aligned}$$

Finally we have

$$\begin{aligned}\zeta_{XX'YY'} &= \operatorname{Var}(R) + \operatorname{Var}(R') + 2\operatorname{Cov}(R,R') \\ &= \|C_P\|_{\mathrm{HS}}^2 + \|C_Q\|_{\mathrm{HS}}^2 + \langle C_P, C_Q\rangle_{\mathrm{HS}} + 2[\langle \mu_P, C_P\mu_P\rangle + \langle \mu_Q, C_P\mu_Q\rangle - 2\langle \mu_P, C_P\mu_Q\rangle] \\ &\quad + 2[\langle \mu_Q, C_Q\mu_Q\rangle + \langle \mu_P, C_Q\mu_P\rangle - 2\langle \mu_Q, C_Q\mu_P\rangle] \\ &= \|C_P\|_{\mathrm{HS}}^2 + \|C_Q\|_{\mathrm{HS}}^2 + \langle C_P, C_Q\rangle_{\mathrm{HS}} + 2\langle \mu_P - \mu_Q, (C_P + C_Q)(\mu_P - \mu_Q)\rangle\end{aligned}$$

Simplification Plugging the ζ values into (7) yields the variance for $\widehat{\mathrm{MMD}}^2$ of

$$\begin{aligned}
&\frac{4}{n_X}\left(\frac{n_X-2}{n_X-1}\cdot\frac{n_Y-2}{n_Y}\cdot\frac{n_Y-3}{n_Y-1}\right)\zeta_X+\frac{4}{n_Y}\left(\frac{n_X-2}{n_X}\cdot\frac{n_X-3}{n_X-1}\cdot\frac{n_Y-2}{n_Y-1}\right)\zeta_Y\\
&+\frac{2}{n_X(n_X-1)}\left(\frac{n_Y-2}{n_Y}\cdot\frac{n_Y-3}{n_Y-1}\right)\zeta_{XX'}+\frac{2}{n_Y(n_Y-1)}\left(\frac{n_X-2}{n_X}\cdot\frac{n_X-3}{n_X-1}\right)\zeta_{YY'}\\
&+\frac{16}{n_Xn_Y}\left(\frac{n_X-2}{n_X-1}\cdot\frac{n_Y-2}{n_Y-1}\right)\zeta_{XY}\\
&+\frac{8}{n_X(n_X-1)n_Y}\left(\frac{n_Y-2}{n_Y-1}\right)\zeta_{XX'Y}+\frac{8}{n_Xn_Y(n_Y-1)}\left(\frac{n_X-2}{n_X-1}\right)\zeta_{XYY'}\\
&+\frac{4}{n_X(n_X-1)n_Y(n_Y-1)}\zeta_{XX'YY'},
\end{aligned}$$

where

$$\begin{aligned}
\zeta_X &= \langle\mu_P-\mu_Q, C_P(\mu_P-\mu_Q)\rangle \qquad \zeta_Y = \langle\mu_P-\mu_Q, C_Q(\mu_P-\mu_Q)\rangle\\
\zeta_{XX'} &= \|C_P\|_{\mathrm{HS}}^2+2\langle\mu_P-\mu_Q, C_P(\mu_P-\mu_Q)\rangle \qquad \zeta_{YY'} = \|C_Q\|_{\mathrm{HS}}^2+2\langle\mu_P-\mu_Q, C_Q(\mu_P-\mu_Q)\rangle\\
\zeta_{XY} &= \frac{1}{4}\langle C_P, C_Q\rangle_{\mathrm{HS}}+\langle\mu_P, C_P\mu_P\rangle+\langle\mu_Q, C_Q\mu_Q\rangle\\
\zeta_{XX'Y} &= \|C_P\|_{\mathrm{HS}}^2+\frac{1}{2}\langle C_P, C_Q\rangle_{\mathrm{HS}}+\langle\mu_P-\mu_Q,(2C_P+C_Q)(\mu_P-\mu_Q)\rangle\\
\zeta_{XYY'} &= \|C_Q\|_{\mathrm{HS}}^2+\frac{1}{2}\langle C_P, C_Q\rangle_{\mathrm{HS}}+\langle\mu_P-\mu_Q,(C_P+2C_Q)(\mu_P-\mu_Q)\rangle\\
\zeta_{XX'YY'} &= \|C_P\|_{\mathrm{HS}}^2+\|C_Q\|_{\mathrm{HS}}^2+\langle C_P, C_Q\rangle_{\mathrm{HS}}+2\langle\mu_P-\mu_Q,(C_P+C_Q)(\mu_P-\mu_Q)\rangle.
\end{aligned}$$

Let $\nu_P=\langle\mu_P-\mu_Q, C_P(\mu_P-\mu_Q)\rangle$ and $\nu_Q=\langle\mu_P-\mu_Q, C_Q(\mu_P-\mu_Q)\rangle$. Then

$$\begin{aligned}
\zeta_X &= \nu_P \qquad \zeta_Y=\nu_Q\\
\zeta_{XX'} &= \|C_P\|_{\mathrm{HS}}^2+2\nu_P \qquad \zeta_{YY'}=\|C_Q\|_{\mathrm{HS}}^2+2\nu_Q\\
\zeta_{XY} &= \frac{1}{4}\langle C_P, C_Q\rangle_{\mathrm{HS}}+\langle\mu_P, C_P\mu_P\rangle+\langle\mu_Q, C_Q\mu_Q\rangle\\
\zeta_{XX'Y} &= \|C_P\|_{\mathrm{HS}}^2+\frac{1}{2}\langle C_P, C_Q\rangle_{\mathrm{HS}}+2\nu_P+\nu_Q\\
\zeta_{XYY'} &= \|C_Q\|_{\mathrm{HS}}^2+\frac{1}{2}\langle C_P, C_Q\rangle_{\mathrm{HS}}+\nu_P+2\nu_Q\\
\zeta_{XX'YY'} &= \|C_P\|_{\mathrm{HS}}^2+\|C_Q\|_{\mathrm{HS}}^2+\langle C_P, C_Q\rangle_{\mathrm{HS}}+2\nu_P+2\nu_Q.
\end{aligned}$$

This means that $\mathrm{Var}(\widehat{\mathrm{MMD}}^2)$ is

$$\begin{aligned}
&\frac{4}{n_X}\left(\frac{n_X-2}{n_X-1}\cdot\frac{n_Y-2}{n_Y}\cdot\frac{n_Y-3}{n_Y-1}\right)\nu_P+\frac{4}{n_Y}\left(\frac{n_X-2}{n_X}\cdot\frac{n_X-3}{n_X-1}\cdot\frac{n_Y-2}{n_Y-1}\right)\nu_Q\\
&+\frac{2}{n_X(n_X-1)}\left(\frac{n_Y-2}{n_Y}\cdot\frac{n_Y-3}{n_Y-1}\right)(\|C_P\|_{\mathrm{HS}}^2+2\nu_P)+\frac{2}{n_Y(n_Y-1)}\left(\frac{n_X-2}{n_X}\cdot\frac{n_X-3}{n_X-1}\right)(\|C_Q\|_{\mathrm{HS}}^2+2\nu_Q)\\
&+\frac{4}{n_Xn_Y}\left(\frac{n_X-2}{n_X-1}\cdot\frac{n_Y-2}{n_Y-1}\right)[\langle C_P, C_Q\rangle_{\mathrm{HS}}+4\langle\mu_P, C_P\mu_P\rangle+4\langle\mu_Q, C_Q\mu_Q\rangle]\\
&+\frac{8}{n_X(n_X-1)n_Y}\left(\frac{n_Y-2}{n_Y-1}\right)\left[\|C_P\|_{\mathrm{HS}}^2+\frac{1}{2}\langle C_P, C_Q\rangle_{\mathrm{HS}}+2\nu_P+\nu_Q\right]\\
&+\frac{8}{n_Xn_Y(n_Y-1)}\left(\frac{n_X-2}{n_X-1}\right)\left[\|C_Q\|_{\mathrm{HS}}^2+\frac{1}{2}\langle C_P, C_Q\rangle_{\mathrm{HS}}+\nu_P+2\nu_Q\right]\\
&+\frac{4}{n_X(n_X-1)n_Y(n_Y-1)}\left[\|C_P\|_{\mathrm{HS}}^2+\|C_Q\|_{\mathrm{HS}}^2+\langle C_P, C_Q\rangle_{\mathrm{HS}}+2\nu_P+2\nu_Q\right].
\end{aligned}$$

Gathering ν_P terms, we get a coefficient of $4/n_X$ times

$$\begin{aligned}
&\frac{(n_X-2)(n_Y-2)(n_Y-3)+(n_Y-2)(n_Y-3)+4(n_Y-2)+2(n_X-2)+2}{(n_X-1)n_Y(n_Y-1)} \\
&= \frac{(n_X-1)(n_Y-2)(n_Y-3)+4(n_Y-2)+2(n_X-1)}{(n_X-1)n_Y(n_Y-1)} \\
&= \frac{(n_Y-2)(n_Y-3)+2}{n_Y(n_Y-1)} + \frac{4(n_Y-2)}{(n_X-1)n_Y(n_Y-1)} \\
&= \frac{n_Y^2-5n_Y+8}{n_Y(n_Y-1)} + \frac{4(n_Y-2)}{(n_X-1)n_Y(n_Y-1)} \\
&= \frac{n_Y(n_Y-1)-4(n_Y-2)}{n_Y(n_Y-1)} + \frac{4(n_Y-2)}{(n_X-1)n_Y(n_Y-1)} \\
&= 1-\frac{4(n_Y-2)}{n_Y(n_Y-1)}\left(1-\frac{1}{n_X-1}\right) \\
&= 1-\frac{4}{n_Y}\cdot\frac{n_X-2}{n_X-1}\cdot\frac{n_Y-2}{n_Y-1};
\end{aligned}$$

the ν_Q terms are symmetric, so that term's coefficient is

$$\frac{4}{n_Y}\left(1-\frac{4}{n_X}\cdot\frac{n_X-2}{n_X-1}\cdot\frac{n_Y-2}{n_Y-1}\right).$$

Next gathering the $\|C_P\|_{\mathrm{HS}}^2$ terms, we find a coefficient of

$$\frac{2}{n_X(n_X-1)}\cdot\frac{(n_Y-2)(n_Y-3)+4(n_Y-2)+2}{n_Y(n_Y-1)} = \frac{2}{n_X(n_X-1)}\frac{n_Y^2-n_Y}{n_Y(n_Y-1)} = \frac{2}{n_X(n_X-1)},$$

and the $\|C_Q\|_{\mathrm{HS}}^2$ terms are again symmetric for a coefficient of $2/(n_Y(n_Y-1))$. Finally, the $\langle C_P, C_Q\rangle_{\mathrm{HS}}$ terms gather to a coefficient of $4/(n_X n_Y)$, since

$$\frac{(n_X-2)(n_Y-2)+(n_Y-2)+(n_X-2)+1}{(n_X-1)(n_Y-1)} = \frac{(n_X-1)(n_Y-2)+(n_X-1)}{(n_X-1)(n_Y-1)} = \frac{(n_X-1)(n_Y-1)}{(n_X-1)(n_Y-1)} = 1.$$

The result follows. ☐

E Asymptotics of Generalized U-statistics

This section characterizes the asymptotic behaviour of generalized U-statistics by filling in and generalizing the approach of Serfling (1980, Chapter 5). Throughout the section we fix a generalized U-statistic U_n constructed from the kernel $h(x_{11},\dots,x_{m_1 1},\dots,x_{1c},\dots,x_{m_c c})$ and random samples $\mathbf{S}=\{X_{ij}: j\in[c], i\in[n_j]\}$, where $(X_{ij})_{i\in[n_j]}\sim\mu_j^{n_j}$ for each j. Additionally define

$$\mathbf{X}=(X_{ij}: j\ \in[c],\ i\in[m_j])\sim\prod_{j=1}^{c}\mu_j^{m_j},\qquad \theta=\mathbb{E}[h(\mathbf{X})],\qquad \tilde{h}=h-\theta.$$

To make our presentation more concise, we introduce an abbreviation to our expression in Definition 3.1 with $U_n=\prod_{j=1}^c\binom{n_j}{m_j}^{-1}\sum_{\mathbb{X}}h(\mathbb{X})$, $\mathbb{X}$ varying over each distinct collection of arguments in $\mathbf{S}$.

Definition E.1. *The order r projection of U_n is given by*

$$\hat{U}_{n,r}=\theta+\sum_{(i_1j_1),\dots,(i_rj_r)}\left(\mathbb{E}[U_n\mid X_{i_1j_1},\dots,X_{i_rj_r}]-\theta\right),$$

where the sum is over each subset of $\mathbf{S}$ of size r.

Often we will simplify the projection of U_n using its order of degeneracy. This is made precise by the following proposition.

Proposition E.2. *Let $\mathbb{X}, \mathbb{X}'$ be subsets of the sample* $\mathbf{S}$ *where $\mathbb{X} \sim \prod_{j=1}^{c} \mu_j^{m_j}$, then $\mathbb{E}[h(\mathbb{X}) \mid \mathbb{X}'] = \mathbb{E}[h(\mathbb{X}) \mid \mathbb{X} \cap \mathbb{X}']$ where we interpret $\mathbb{E}[h(\mathbb{X}) \mid \emptyset] = \mathbb{E}[h(\mathbb{X})]$. In particular if U_n is order r degenerate and $|\mathbb{X} \cap \mathbb{X}'| \leq r$ then $\mathbb{E}[\tilde{h}(\mathbb{X}) \mid \mathbb{X}'] = 0$.*

Proof. Since $\mathbf{S}$ consists of independent variables, $\mathbb{E}[h(\mathbb{X}) \mid \mathbb{X}']$ is obtained by integrating out the arguments in $h(\mathbb{X})$ which are not shared with $\mathbb{X}'$. It then follows that $\mathbb{E}[h(\mathbb{X}) \mid \mathbb{X}'] = \mathbb{E}[h(\mathbb{X}) \mid \mathbb{X} \cap \mathbb{X}']$. Now if U_n is order r degenerate and $|\mathbb{X} \cap \mathbb{X}'| \leq r$, the symmetries of h imply $\mathrm{Var}(\mathbb{E}[\tilde{h}(\mathbb{X}) | \mathbb{X} \cap \mathbb{X}']) = 0$ and in turn $\mathbb{E}[\tilde{h}(\mathbb{X}) | \mathbb{X}'] = \mathbb{E}[\tilde{h}(\mathbb{X})] = 0$. □

The following lemma shows that if U_n has order-r degeneracy (identifying $r = 0$ with non-degeneracy) then its asymptotic distribution can be approximated by $\hat{U}_{n,r+1}$ in the mean-square norm.

Lemma E.3. *If U_n is a generalized U-statistic with order of degeneracy at least r, then $\mathbb{E}\,|U_n - \hat{U}_{n,r+1}|^2 = \mathcal{O}(n^{-(r+2)})$.*

Proof. Expanding $\mathbb{E}[U_n \mid X_{i_1 j_1}, \ldots, X_{i_{r+1} j_{r+1}}]$ yields

$$U_n - \hat{U}_{n,r+1} = \prod_{j=1}^{c} \binom{n_j}{m_j}^{-1} \sum_{\mathbb{X}} \left[\tilde{h}(\mathbb{X}) - \sum_{(i_1 j_1), \ldots, (i_{r+1} j_{r+1})} \mathbb{E}[\tilde{h}(\mathbb{X}) \mid X_{i_1 j_1}, \ldots, X_{i_{r+1} j_{r+1}}] \right].$$

Turning to the inner sum, Proposition E.2 implies $\mathbb{E}[\tilde{h}(\mathbb{X}) \mid X_{i_1 j_1}, \ldots, X_{i_{r+1} j_{r+1}}] = 0$ if $\mathbb{X}$ and $X_{i_1 j_1}, \ldots, X_{i_{r+1} j_{r+1}}$ share less than $r+1$ terms. Thus removing these zero-valued terms leaves us with

$$U_n - \hat{U}_{n,r} = \prod_{j=1}^{c} \binom{n_j}{m_j}^{-1} \sum_{\mathbb{X}} \left[\tilde{h}(\mathbb{X}) - \sum_{(i_1 j_1), \ldots, (i_{r+1} j_{r+1}) \in \mathbb{X}} \mathbb{E}[\tilde{h}(\mathbb{X}) \mid X_{i_1 j_1}, \ldots, X_{i_{r+1} j_{r+1}}] \right], \tag{9}$$

where $(i_1 j_1), \ldots, (i_{r+1} j_{r+1}) \in \mathbb{X}$ denotes that $\mathbb{X}$ contains $X_{i_1 j_1}, \ldots, X_{r_{r+1}, j_{r+1}}$. The above puts $U_n - \hat{U}_{n,r}$ in the form a generalized U-statistic and furthermore conditioning on any $r+1$ terms in $\mathbf{X}$ gives

$$\begin{aligned} &\mathbb{E}\left[\tilde{h}(\mathbf{X}) - \sum_{(i_1 j_1), \ldots, (i_{r+1} j_{r+1}) \in \mathbf{X}} \mathbb{E}[\tilde{h}(\mathbf{X}) \mid X_{i_1 j_1}, \ldots, X_{i_{r+1} j_{r+1}}] \,\middle|\, X_{s_1 t_1}, \ldots, X_{s_{r+1} t_{r+1}} \right] \\ &\quad = \mathbb{E}[\tilde{h}(\mathbf{X}) \mid X_{s_1 t_1}, \ldots, X_{s_{r+1} t_{r+1}}] - \mathbb{E}[\tilde{h}(\mathbf{X}) \mid X_{s_1 t_1}, \ldots, X_{s_{r+1} t_{r+1}}] = 0 \quad \text{by Proposition E.2.} \end{aligned}$$

If we had conditioned on fewer than $r+1$ variables, the term above would be zero again due to U_n's order of degeneracy. Therefore $U_n - \hat{U}_{n,r}$'s order of degeneracy is at least $r+1$. Since $U_n - \hat{U}_{n,r}$ has zero mean, it follows from the variance formula (8) that $\mathbb{E}\,|U_n - \hat{U}_{n,r}|^2 = \mathcal{O}(n^{-(r+2)})$. □

The next two results builds on our lemma and deduces the asymptotic distribution of generalized U-statistics based on their first and second order projections.

Theorem 3.8. *Let U_n be a c-sample U-statistic with kernel h, where $n = \min\{n_1, \ldots, n_c\}$. If $n/n_j \to \rho_j \in [0,1]$ for each $j \in [c]$, then*

$$\sqrt{n}(U_n - \mathbb{E}[h(\mathbf{X})]) \xrightarrow{d} \mathcal{N}\left(0, \sum_{j=1}^{c} \rho_j m_j^2 \,\mathrm{Var}\left(\mathbb{E}\left[h(\mathbf{X}) \mid X_{1j}\right]\right)\right),$$

where $\mathcal{N}(0,0)$ is interpreted as a point mass at 0. In particular if U_n is non-degenerate and each $\rho_j > 0$ (the proportional regime), the distribution above is normal with positive variance.

Proof. By Lemma E.3 we have that

$$\mathbb{E}[|\sqrt{n}(U_n - \hat{U}_{n,1})|^2] = n\,\mathbb{E}[|U_n - \hat{U}_{n,1}|^2] = n\,\mathcal{O}(n^{-2}) = \mathcal{O}(n^{-1}),$$

so it suffices to deduce the distribution of $\sqrt{n}(\hat{U}_{n,1} - \theta)$. Define $\tilde{h}_j = \mathbb{E}[\tilde{h}(\mathbf{X}) \mid X_{1j} = \cdot]$, we may write

$$\hat{U}_{n,1} - \theta = \prod_{j=1}^{c} \binom{n_j}{m_j}^{-1} \sum_{j=1}^{c} \sum_{i=1}^{n_j} \sum_{\mathbb{X}} \mathbb{E}[\tilde{h}(\mathbb{X}) \mid X_{ij}]$$

Note that if X_{ij} is not contained in $\mathbb{X}$ then Proposition E.2 implies $\mathbb{E}[\tilde{h}(\mathbb{X}) \mid X_{ij}] = 0$, thus counting the number of $\mathbb{X}$ containing X_{ij} leaves us with

$$\hat{U}_{n,1} - \theta = \prod_{j=1}^{c} \binom{n_j}{m_j}^{-1} \sum_{j=1}^{c} \sum_{i=1}^{n_j} \binom{n_j - 1}{m_j - 1} \prod_{j' \neq j} \binom{n_{j'}}{m_{j'}} \tilde{h}_j(X_{ij}) = \sum_{j=1}^{c} \frac{m_j}{n_j} \sum_{i=1}^{n_j} \tilde{h}_j(X_{ij}).$$

If $\mathrm{Var}(\mathbb{E}[h(\mathbf{X}) \mid X_{1j}]) > 0$, we may combine the central limit theorem and Slutsky's theorem to get

$$\sqrt{n}\left(\frac{m_j}{n_j} \sum_{i=1}^{n_j} \tilde{h}_j(X_{ij})\right) = \sqrt{\frac{n}{n_j}} m_j \left(\frac{1}{\sqrt{n_j}} \sum_{i=1}^{n_j} \tilde{h}_j(X_{ij})\right) \xrightarrow{d} \mathcal{N}\left(0, \rho_j m_j^2 \,\mathrm{Var}(\mathbb{E}[h(\mathbf{X}) \mid X_{1j}])\right).$$

Otherwise if $\mathrm{Var}(\mathbb{E}[h(\mathbf{X}) \mid X_{1j}]) = 0$ then $\sum_{i=1}^{n_j} \tilde{h}_j(X_{ij})$ is almost surely zero. The desired limiting distribution follows by applying the continuous mapping theorem to $\sqrt{n}(\hat{U}_{n,1} - \theta)$. If U_n is non-degenerate then $\mathrm{Var}(\mathbb{E}[h(\mathbf{X}) \mid X_{1j}]) > 0$ for some j, thus proportionality guarantees that the variance is positive. ☐

To deal with the asymmetric kernel associated with second order projections of generalized U-statistics, we have a preliminary result analogous to a singular value decomposition (SVD) for square-integrable kernels.

Lemma E.4 (Kernel SVD)**.** *Let $k \in L_2(\mathcal{X} \times \mathcal{Y}, \mu \times \nu)$ and $L_2(\mathcal{X}, \mu), L_2(\mathcal{Y}, \nu)$ be separable. Then k has the representation as an L_2 limit*

$$k(x, y) = \sum_n \sigma_n v_n(x) u_n(y),$$

where σ_n are the singular values of the operator $T : L_2(\mathcal{X}, \mu) \to L_2(\mathcal{Y}, \nu)$ defined by

$$T : f \mapsto \int k(x, \cdot) f(x)\; d\mu(x),$$

and $\{v_n\} \subset L_2(\mathcal{X}, \mu), \{u_n\} \subset L_2(\mathcal{Y}, \nu)$ are orthonormal.

Proof. Since k is square integrable, T is Hilbert-Schmidt and hence compact. Letting T^* denote the adjoint, T^*T is a compact self-adjoint operator and hence provides a countable orthonormal basis of eigenvectors $\{v_n\}$ of T^*T with corresponding eigenvectors $\{\lambda_n\}$ arranged so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. For all non-zero λ_n, put $\sigma_n = \sqrt{\lambda_n}$ and $u_n = \sigma_n^{-1} T v_n$. Notice that

$$\langle u_n, u_m \rangle = \frac{1}{\sigma_n \sigma_m} \langle T v_n, T v_m \rangle = \frac{1}{\sigma_n \sigma_m} \langle T^* T v_n, v_m \rangle = \frac{\sigma_n}{\sigma_m} \langle v_n, v_m \rangle$$

so $\{u_n\}$ is orthonormal in $L_2(\mathcal{Y}, \nu)$. Thus we have

$$\begin{aligned}
&\iint \left| k(x, y) - \sum_{n=1}^{L} \sigma_n v_n(x) u_n(y) \right|^2 \; d\mu(x) d\nu(y) \\
&= \iint \left(|k(x, y)|^2 - 2 \sum_{n=1}^{L} \sigma_n k(x, y) v_n(x) u_n(y) + \sum_{n=1}^{L} \sigma_n^2 v_n^2(x) u_n^2(y) \right) \; d\mu(x) d\nu(y) \\
&= \|k\|_2^2 - 2 \sum_{n=1}^{L} \sigma_n \int u_n(y) (T v_n)(y)\; d\nu(y) + \sum_{n=1}^{L} \sigma_n^2 \\
&= \|k\|_2^2 - \sum_{n=1}^{L} \sigma_n^2 = \|k\|_2^2 - \sum_{n=1}^{L} \langle T^* T v_n, v_n \rangle_2 = \|k\|_2^2 - \sum_{n=1}^{L} \|T v_n\|_2^2.
\end{aligned}$$

It is a standard result that $\|k\|_2 = \|T\|_{\text{HS}}$. Therefore the final term vanishes as $L \to \infty$. $\square$

Observe that $\hat{U}_{n,2}$ is a linear combination of the conditional kernels

$$\tilde{h}_{jj} := \mathbb{E}(\tilde{h}(\mathbf{X}) \mid X_{1j} = \cdot, X_{2j} = \cdot), \quad 1 \le j \le c$$
$$\tilde{h}_{st} := \mathbb{E}(\tilde{h}(\mathbf{X}) \mid X_{1s} = \cdot, X_{1t} = \cdot), \quad 1 \le s < t \le c.$$

Now assuming each $\tilde{h}_{jj} \in L_2(\mu_j^2)$ and $\tilde{h}_{st} \in L_2(\mu_s \times \mu_t)$ leads to a variety of kernel decompositions. For the symmetric kernel $\tilde{h}_{jj}$, Dunford and Schwartz (1988, Section VI, Exercises 44 and 56) give the decomposition

$$\tilde{h}_{jj}(x_1, x_2) = \sum_{l=1}^{\infty} \lambda_{jl} \psi_{jl}(x_1) \psi_{jl}(x_2), \tag{10}$$

where the limit is taken with in L_2 norm, and $\{\psi_{jl}\}, \{\lambda_{jl}\}$ are the (orthonormal) eigenfunction-eigenvalue pairs of $T_j := f \mapsto \mathbb{E}_{X \sim \mu_j}[\tilde{h}_{jj}(\cdot, X) f(X)]$. The aforementioned Lemma E.4 applies to $\tilde{h}_{st}$ giving

$$\tilde{h}_{st}(x_1, x_2) = \sum_{l=1}^{\infty} \sigma_{stl} v_{stl}(x_1) u_{stl}(x_2). \tag{11}$$

These results are applicable to the coming proof.

Theorem E.5. *Let U_n be a c-sample generalized U-statistic with an associated kernel h, where $n = \min\{n_1, \ldots, n_c\}$ and $\rho_j := \lim_{n\to\infty} n/n_j$ exists for each j. Further for assume $\tilde{h}_{jj} \in L_2(\mu_j^2)$ for $j \in [c]$ and $\tilde{h}_{st} \in L_2(\mu_s \times \mu_t)$ for $1 \le s < t \le c$, admitting the decompositions* (10) *and* (11)*. If U_n is first-order degenerate then $n(U_n - \theta)$ converges in distribution to Y, where*

$$Y = \sum_{j=1}^{c} \binom{m_j}{2} \rho_j \sum_{l=1}^{\infty} \lambda_{jl}(a_{jl}^2 - 1) + \sum_{s<t} m_s m_t \sqrt{\rho_s \rho_t} \sum_{l=1}^{\infty} \sigma_{stl} b_{stl} c_{stl},$$

and $\{a_{jl}\}, \{b_{stl}\}, \{c_{stl}\}$ are marginally $\mathcal{N}(0,1)$ variables obtained from the weak limits

$$\frac{1}{\sqrt{n_j}} \sum_{i=1}^{n_j} \psi_{jl}(X_{ij}) \to_d a_{jl}, \quad \frac{1}{\sqrt{n_s}} \sum_{i=1}^{n_s} v_{stl}(X_{is}) \to_d b_{stl}, \quad \frac{1}{\sqrt{n_t}} \sum_{i=1}^{n_t} u_{stl}(X_{it}) \to_d c_{stl}.$$

Proof. By Lemma E.3, $n(U_n - \theta)$ converges in L_2 to $V_n := n(\hat{U}_{n,2} - \theta)$. Writing the projection explicitly

$$n^{-1} V_n = \sum_{j=1}^{c} \sum_{i_1 < i_2} [\mathbb{E}(U_n \mid X_{i_1 j}, X_{i_2 j}) - \theta] + \sum_{s<t} \sum_{i_1, i_2} [\mathbb{E}(U_n \mid X_{i_1 s}, X_{i_2 t}) - \theta],$$

which may be simplified by removing the zero-valued terms according to Proposition E.2

$$
\begin{aligned}
\sum_{i_1<i_2}[\mathbb{E}(U_n \mid X_{i_1j}, X_{i_2j})-\theta] &= \prod_{j'=1}^{c}\binom{n_{j'}}{m_{j'}}^{-1}\sum_{i_1<i_2}\sum_{\mathbb{X}}\mathbb{E}(\tilde{h}(\mathbb{X}) \mid X_{i_1j}, X_{i_2j}) \\
&= \prod_{j'=1}^{c}\binom{n_{j'}}{m_{j'}}^{-1}\sum_{i_1<i_2}\binom{n_j-2}{m_j-2}\prod_{j'\neq j}\binom{n_{j'}}{m_{j'}}\tilde{h}_{jj}(X_{i_1j}, X_{i_2j}) \\
&= \binom{m_j}{2}\binom{n_j}{2}^{-1}\sum_{i_1<i_2}\tilde{h}_{jj}(X_{i_1j}, X_{i_2j}) \\
\sum_{i_1,i_2}[\mathbb{E}(U_n \mid X_{i_1s}, X_{i_2t})-\theta] &= \prod_{j=1}^{c}\binom{n_j}{m_j}^{-1}\sum_{i_1,i_2}\sum_{\mathbb{X}}\mathbb{E}(\tilde{h}(\mathbb{X}) \mid X_{i_1s}, X_{i_2t}) \\
&= \prod_{j=1}^{c}\binom{n_j}{m_j}^{-1}\sum_{i_1,i_2}\binom{n_s-1}{m_s-1}\binom{n_t-1}{m_t-1}\prod_{j\neq s,t}\binom{n_j}{m_j}\tilde{h}_{st}(X_{i_2,s}, X_{i_2t}) \\
&= \frac{m_s m_t}{n_s n_t}\sum_{i_1,i_2}\tilde{h}_{st}(X_{i_2,s}, X_{i_2t}).
\end{aligned}
$$

Thus we're left with

$$
n^{-1}V_n = \sum_{j=1}^{c}\binom{m_j}{2}\binom{n_j}{2}^{-1}\sum_{i_1<i_2}\tilde{h}_{jj}(X_{i_1j}, X_{i_2j}) + \sum_{s<t}\frac{m_s m_t}{n_s n_t}\sum_{i_1,i_2}\tilde{h}_{st}(X_{i_1s}, X_{i_2t}).
$$

Replacing $\tilde{h}_{jj}$ and $\tilde{h}_{st}$ with their L_2 expansions, we define the truncation of V_n to L eigenfunctions

$$
\begin{aligned}
n^{-1}V_{nL} :=& \sum_{j=1}^{c}\binom{m_j}{2}\binom{n_j}{2}^{-1}\sum_{l=1}^{L}\sum_{i_1<i_2}\lambda_{jl}\psi_{jl}(X_{i_1j})\psi_{jl}(X_{i_2j}) \\
&+\sum_{s<t}\frac{m_s m_t}{n_s n_t}\sum_{l=1}^{L}\sum_{i_1,i_2}\sigma_{stl}v_{stl}(X_{i_1s})u_{stl}(X_{i_2t}). \qquad (12)
\end{aligned}
$$

and the truncation of Y by

$$
Y_L := \sum_{j=1}^{c}\binom{m_j}{2}\rho_j\sum_{l=1}^{L}\lambda_{jl}(a_{jl}^2-1) + \sum_{s<t}m_s m_t\sqrt{\rho_s\rho_t}\sum_{l=1}^{L}\sigma_{stl}b_{stl}c_{stl}.
$$

Using V_{nL} and Y_L as points of comparison, we have the following error decomposition in terms of characteristic functions:

$$
\begin{aligned}
|\,\mathbb{E}\exp(ixV_n) - \mathbb{E}\exp(ixY)| \leq& |\,\mathbb{E}\exp(ixV_n) - \mathbb{E}\exp(ixV_{nL})| + |\,\mathbb{E}\exp(ixV_{nL}) - \mathbb{E}\exp(ixY_L)| \\
&+|\,\mathbb{E}\exp(ixY_L) - \mathbb{E}\exp(ixY)|.
\end{aligned}
$$

The rest of the proof is divided in three sections, each focused on bounding a term above.

First term: We first apply the inequality $|\exp(iz)-1| \leq |z|$

$$
\begin{aligned}
&|x|^{-1}\,\mathbb{E}\,|\exp(ixV_n)-\exp(ixV_{nL})| \\
&\quad\leq \mathbb{E}\,|V_n - V_{nL}| \leq [\mathbb{E}\,|V_n - V_{nL}|^2]^{1/2} \qquad \text{by Hölder's inequality} \\
&\quad\leq \sum_{j=1}^{c}\binom{m_j}{2}\underbrace{\left[\mathbb{E}\left|n\binom{n_j}{2}^{-1}\sum_{l=L+1}^{\infty}\sum_{i_1<i_2}\lambda_{jl}\psi_{jl}(X_{i_1 j})\psi_{jl}(X_{i_2 j})\right|^2\right]^{1/2}}_{A_j} \\
&\quad\quad+\sum_{s<t}m_s m_t\underbrace{\left[\mathbb{E}\left|\frac{n}{n_s n_t}\sum_{l=L+1}^{\infty}\sum_{i_1,i_2}\sigma_{stl}v_{stl}(X_{i_1 s})u_{stl}(X_{i_2 t})\right|^2\right]^{1/2}}_{B_{st}}.
\end{aligned}
$$

Since $V_{nL} \to V_n$ in L_2 we have

$$
\begin{aligned}
A_j^2 &= n^2\binom{n_j}{2}^{-2}\mathbb{E}\left|\sum_{l=L+1}^{\infty}\sum_{i_1<i_2}\lambda_{jl}\psi_{jl}(X_{i_1 j})\psi_{jl}(X_{i_2 j})\right|^2 \\
&= n^2\binom{n_j}{2}^{-2}\sum_{l=L+1}^{\infty}\lambda_{jl}^2\,\mathbb{E}\left|\sum_{i_1<i_2}\psi_{jl}(X_{i_1 j})\psi_{jl}(X_{i_2 j})\right|^2 \qquad \text{by orthogonality} \\
&= n^2\binom{n_j}{2}^{-2}\sum_{l=L+1}^{\infty}\lambda_{jl}^2\sum_{i_1<i_2}\sum_{i_3<i_4}\mathbb{E}[\psi_{jl}(X_{i_1 j})\psi_{jl}(X_{i_2 j})\psi_{jl}(X_{i_3 j})\psi_{jl}(X_{i_4 j})].
\end{aligned}
$$

First-order degeneracy implies

$$
\mathbb{E}[\tilde{h}_{jj}(\cdot, X_{ij})] = \sum_{l=1}^{\infty}\lambda_{jl}\,\mathbb{E}[\psi_{jl}(X_{ij})]\psi_{jl}(\cdot) = 0
$$

so by linear independence of the eigenfunctions we must have $\mathbb{E}[\psi_{jl}(X_{ij})] = 0$ if $\lambda_{jl} \neq 0$. It follows by independence that

$$
\mathbb{E}[\psi_{jl}(X_{i_1 j})\psi_{jl}(X_{i_2 j})\psi_{jl}(X_{i_3 j})\psi_{jl}(X_{i_4 j})] = \begin{cases} 1 & \text{if } (i_1,i_2)=(i_3,i_4) \\ 0 & \text{otherwise.} \end{cases}
$$

Hence we get

$$
A_j^2 = n^2\binom{n_j}{2}^{-1}\sum_{l=L+1}^{\infty}\lambda_{jl}^2 = \frac{2n^2}{n_j(n_j-1)}\sum_{l=L+1}^{\infty}\lambda_{jl}^2 \leq 4\sum_{l=L+1}^{\infty}\lambda_{jl}^2.
$$

We may similarly show that $\mathbb{E}[v_{stl}(X_{i_1 s})] = \mathbb{E}[u_{stl}(X_{i_2 t})] = 0$ if $\sigma_{stl} \neq 0$, thus

$$
\begin{aligned}
B_{st}^2 &= \left(\frac{n}{n_s n_t}\right)^2\mathbb{E}\left|\sum_{l=L+1}^{\infty}\sum_{i_1,i_2}\sigma_{stl}v_{stl}(X_{i_1 s})u_{stl}(X_{i_2 t})\right|^2 \\
&= \left(\frac{n}{n_s n_t}\right)^2\sum_{l=L+1}^{\infty}\sigma_{stl}^2\,\mathbb{E}\left|\sum_{i_1,i_2}v_{stl}(X_{i_1 s})u_{stl}(X_{i_2 t})\right|^2 = \frac{n^2}{n_s n_t}\sum_{l=L+1}^{\infty}\sigma_{stl}^2 \leq \sum_{l=L+1}^{\infty}\sigma_{stl}^2.
\end{aligned}
$$

Since the eigenvalues $\{\lambda_{jl}\}$ and singular values $\{\sigma_{stl}\}$ belong to Hilbert-Schmidt operators, their squared series converge. Therefore $\lim_{L\to\infty}\mathbb{E}\,|\exp(ixV_n) - \exp(ixV_{nL})| = 0$ as it is a linear combination of the A_j, B_{st}. Importantly, our choice of L is independent of n.

Second term: Fix any $L \geq 1$, we will focus on the following terms in expression (12) of V_{nL}

$$\alpha_j := \frac{2}{n_j} \sum_{l=1}^{L} \sum_{i_1 < i_2} \lambda_{jl} \psi_{jl}(X_{i_1 j}) \psi_{jl}(X_{i_2 j})$$
$$\beta_{st} := \frac{1}{\sqrt{n_s n_t}} \sum_{l=1}^{L} \sum_{i_1, i_2} \sigma_{stl} v_{stl}(X_{i_1 s}) u_{stl}(X_{i_2 t}),$$

which we may rewrite as

$$\begin{aligned}
\alpha_j &= \frac{1}{n_j} \sum_{l=1}^{L} \lambda_{jl} \left(\left(\sum_{i=1}^{n_j} \psi_{jl}(X_{ij}) \right)^2 - \sum_{i=1}^{n_j} \psi_{jl}^2(X_{ij}) \right) \\
&= \sum_{l=1}^{L} \lambda_{jl} \left(\left(\frac{1}{\sqrt{n_j}} \sum_{i=1}^{n_j} \psi_{jl}(X_{ij}) \right)^2 - \frac{1}{n_j} \sum_{i=1}^{n_j} \psi_{jl}^2(X_{ij}) \right) \\
\beta_{st} &= \sum_{l=1}^{L} \sigma_{stl} \left(\frac{1}{\sqrt{n_s}} \sum_{i_1=1}^{n_s} v_{stl}(X_{i_1 s}) \right) \left(\frac{1}{\sqrt{n_t}} \sum_{i_2=1}^{n_t} u_{stl}(X_{i_2 t}) \right).
\end{aligned}$$

Note that $\frac{1}{n_j} \sum_{i=1}^{n_j} \psi_{jl}^2(X_{ij}) \to 1$ a.s. by the law of large numbers. By the above we may view V_{nL} as a continuous function in terms of the sample means and $(Z_1, \ldots, Z_c)$, where Z_j is the vector concatenating

$$\left(\frac{1}{\sqrt{n_j}} \sum_{i=1}^{n_j} \psi_{jl}(X_{ij}) : 1 \leq l \leq L \right)^\top$$
$$\left(\frac{1}{\sqrt{n_j}} \sum_{i=1}^{n_j} v_{jtl}(X_{ij}) : 1 \leq l \leq L,\ j < t \right)^\top$$
$$\left(\frac{1}{\sqrt{n_j}} \sum_{i=1}^{n_j} u_{sjl}(X_{ij}) : 1 \leq l \leq L,\ s < j \right)^\top.$$

That is, Z_j contains all normalized samples from μ_j. The central limit theorem gives $Z_j \to_d \mathcal{N}(0, \mathbf{\Sigma}_j)$ as $n_j \to \infty$, where $\mathbf{\Sigma}_j$ is the covariance matrix between pairs of $\psi_{jl}(X_{ij}), v_{jtl}(X_{ij}), u_{sjl}(X_{ij})$. Moreover Z_j is independent across each j, so we have the joint convergence $(Z_1, \ldots, Z_c) \to_d \mathcal{N}(0, \mathbf{\Sigma})$ as $n \to \infty$, where $\mathbf{\Sigma}$ is the block diagonal matrix $\text{diag}(\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_c)$. A final application of the continuous mapping theorem and Slutsky's theorem gives

$$\begin{aligned}
V_{nL} &= \sum_{j=1}^{c} \binom{m_j}{2} \frac{n}{n_j - 1} \alpha_j + \sum_{s<t} m_s m_t \frac{n}{\sqrt{n_s n_t}} \beta_{st} \\
&\longrightarrow_d \sum_{j=1}^{c} \binom{m_j}{2} \rho_j \sum_{l=1}^{L} \lambda_{jl} (a_{jl}^2 - 1) + \sum_{s<t} m_s m_t \sqrt{\rho_s \rho_t} \sum_{l=1}^{L} \sigma_{stl} b_{stl} c_{stl} = Y_L.
\end{aligned}$$

Convergence in distribution accordingly gives $|\mathbb{E} \exp(ixV_{nL}) - \mathbb{E} \exp(ixY_L)| \to 0$ as $n \to \infty$.

Third term: Recall that Y is the L_2 limit of Y_L, so again using $|\exp(iz) - 1| \leq |z|$ we get

$$|\mathbb{E} \exp(ixY_L) - \mathbb{E} \exp(ixY)| \leq |x| [\mathbb{E} |Y_L - Y|^2]^{1/2} \to 0 \text{ as } L \to \infty.$$

To complete the proof, pick $\varepsilon > 0$. For L sufficiently large the first and third term are less than $\varepsilon/3$. Fixing this L, we may pick N such that the second term is bounded by $\varepsilon/3$ for all $n \geq N$, giving

$$|\mathbb{E} \exp(ixV_n) - \mathbb{E} \exp(ixY)| < \varepsilon, \quad \forall n \geq N. \qquad \square$$

We can now determine the null distribution of $\widehat{\mathrm{MMD}}^2$ as a specialized case of the preceding theorem. We shall see this distribution has a much cleaner form than above due to the additional symmetries of the MMD kernel.

Theorem 3.7. *Assume Setting A and that* $\mathrm{MMD}^2(P, Q) = 0$. *Assume* $\min\{n_X, n_Y\}/n_X \to \rho_X$ *and* $\min\{n_X, n_Y\}/n_Y \to \rho_Y$ *for some* ρ_X, ρ_Y *in* $[0, 1]$. $\widehat{\mathrm{MMD}}^2$ *converges in distribution as*

$$\min\{n_X, n_Y\} \widehat{\mathrm{MMD}}^2 \xrightarrow{d} (\rho_X + \rho_Y) \sum_{l=1}^{\infty} \lambda_l \left(Z_l^2 - 1\right),$$

where each Z_l *is independently* $\mathcal{N}(0, 1)$*, and the* λ_l *are the eigenvalues of the integral equation*

$$\mathop{\mathbb{E}}_X [\langle \phi(X) - \mu_P, \phi(y) - \mu_P \rangle g(X)] = \lambda g(y).$$

Proof. Theorem 3.4 implies that $\widehat{\mathrm{MMD}}^2$ is first-order degenerate so we seek to use Theorem 3.7. Define

$$\begin{aligned}
\tilde{h}_{XX} &= \mathbb{E}[h(X, X'; Y, Y') \mid X = \cdot, X' = \cdot] \\
\tilde{h}_{YY} &= \mathbb{E}[h(X, X'; Y, Y') \mid Y = \cdot, Y' = \cdot] \\
\tilde{h}_{XY} &= \mathbb{E}[h(X, X'; Y, Y') \mid X = \cdot, Y = \cdot]
\end{aligned}$$

Since $\mathrm{MMD}^2(P, Q) = 0$ implies $\mu_P = \mu_Q$, we calculate the conditional kernels as

$$\begin{aligned}
\tilde{h}_{XX}(X, X') &= k(X, X') + \mathbb{E}\, k(Y, Y') - \mu_Q(X) - \mu_Q(X') \\
&= \langle \phi(X) - \mu_Q, \phi(X) - \mu_Q \rangle \\
&= \langle \phi(X) - \mu_P, \phi(X) - \mu_P \rangle \\
\tilde{h}_{XY}(X, Y) &= \mu_P(X) + \mu_Q(Y) - \frac{1}{2} \left(\mu_Q(X) + \mu_P(Y) + k(X, Y) + \mathbb{E}\, k(X', Y')\right) \\
&= \frac{1}{2} \left(\mu_P(X) + \mu_P(Y) - k(X, Y) - \mathbb{E}\, k(X', Y')\right) \\
&= -\frac{1}{2} \langle \phi(X) - \mu_P, \phi(Y) - \mu_P \rangle
\end{aligned}$$

and by symmetry $\tilde{h}_{YY}(Y, Y') = \langle \phi(X) - \mu_P, \phi(Y) - \mu_P \rangle$ as well. It follows that the integral operators of $\tilde{h}_{XX}, \tilde{h}_{YY}, \tilde{h}_{XY}$ share the same normalized eigenfunctions, and the eigenvalues of $\tilde{h}_{XY}$ carry an extra factor of $-1/2$. Now Proposition 2.1 implies $\tilde{h}_{XX}, \tilde{h}_{YY}, \tilde{h}_{XY}$ are square integrable with respect to $P^2, Q^2, P \times Q$, so following Theorem E.5

$$\min\{n_x, n_y\} \widehat{\mathrm{MMD}}^2 \to_d \rho_X \sum_{l=1}^{\infty} \lambda_l (a_l^2 - 1) + \rho_Y \sum_{l=1}^{\infty} \lambda_l (b_l^2 - 1) - 2\sqrt{\rho_X \rho_Y} \sum_{l=1}^{\infty} \lambda_l a_l b_l,$$

where $a_l, b_l \sim \mathcal{N}(0, 1)$ are independent. Finally we may complete the square, simplifying the above to

$$\sum_{l=1}^{\infty} \lambda_l \left[(\rho_X^{1/2} a_l - \rho_Y^{1/2} b_l)^2 - (\rho_X + \rho_Y)\right] = \sum_{l=1}^{\infty} \lambda_l \left[(\rho_X + \rho_Y) Z_l^2 - (\rho_X + \rho_Y)\right] = (\rho_X + \rho_Y) \sum_{l=1}^{\infty} \lambda_l (Z_l^2 - 1),$$

as desired. □