**Contents**

# A Background

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

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

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

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

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

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

This implies the following

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

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

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

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

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

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

*Then, we have*

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

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

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

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

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

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

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

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

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

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

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

## B Gaussian limiting distribution of $\widehat{\mathrm{xMMD}}^2$

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

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

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

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

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

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

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

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

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

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

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

We now state the main result of this section.

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

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

We now present the proof of this result.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Using this, we obtain the following:

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

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

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

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

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

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

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

Next, we note that

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

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

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

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

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

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

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

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

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

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

as required. This completes the proof. □

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We will carry over the notations used in the proof of Theorem 15, and in particular, we will use $\bar{U}_X = \frac{1}{n_1}\sum_{i=1}^{n_1} W_i$ and $\bar{U}_Y = \frac{1}{m_1}\sum_{j=1}^{m_1} Z_j$. Since $W_i$ and $Z_j$ are identically distributed under the null, we have $\mathbb{E}_P[W_i^2|\mathbb{X}_2, \mathbb{Y}_2] = \mathbb{E}_P[Z_j^2|\mathbb{X}_2, \mathbb{Y}_2]$, and we will use $\sigma_2^2$ to denote this conditional variance. Then, note the following:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

This, in turn, implies

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

as required by Theorem 5.

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

## C Consistency against fixed and local alternatives (Section 4)

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

*Proof.* We begin by noting that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

In the above display:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We first check the condition on the variance of $\mathrm{x}\widehat{\mathrm{MMD}}^2$. Note that we have the following:
$$\begin{aligned}\bar{U}_X &= \langle \widehat{\mu}_1, \widehat{\mu}_2 - \widehat{\nu}_2\rangle_k = \langle \widetilde{\mu}_1 + \mu, \widetilde{g}_2 + \mu - \nu\rangle_k \\ &= \langle \widetilde{\mu}_1, \widetilde{g}_2\rangle_k + \langle \widetilde{\mu}_1, \mu - \nu\rangle_k + \langle \mu, \widetilde{g}_2\rangle_k + \langle \mu, \mu - \nu\rangle_k\end{aligned}$$
Recall that we use $\widetilde{\mu}_1$ to denote $\widehat{\mu}_1 - \mu$, and similarly use $\widetilde{\mu}_2, \widetilde{\nu}_1, \widetilde{\nu}_2$ and $\widetilde{g}_2$ to denote $\widehat{\mu}_2 - \mu, \widehat{\nu}_1 - \nu, \widehat{\nu}_2 - \nu$ and $g_2 - (\mu - \nu)$ respectively. Similarly, on expanding the term $\bar{U}_Y$, we get
$$\begin{aligned}\bar{U}_Y &= \langle \widehat{\nu}_1, \widehat{\mu}_2 - \widehat{\nu}_2\rangle_k = \langle \widetilde{\nu}_1 + \nu, \widetilde{g}_2 + \mu - \nu\rangle_k \\ &= \langle \widetilde{\nu}_1, \widetilde{g}_2\rangle_k + \langle \widetilde{\nu}_1, \mu - \nu\rangle_k + \langle \nu, \widetilde{g}_2\rangle_k + \langle \nu, \mu - \nu\rangle_k\end{aligned}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## D Gaussian Limit for General Two-Sample U-Statistic

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We now state the main result of this section.

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

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

### D.1 Proof of Theorem 21

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

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

We now present the details of the steps outlined above.

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

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

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

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

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

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

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

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

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

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

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

$\square$

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

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

*Proof.* We begin by noting the following

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

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

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

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

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

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

# E Additional Experiments

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

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

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

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

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

**Details for Figure 3.** The two kernels used in this figure are the Gaussian and Quadratic kernels. The Gaussian kernel with scale parameter $s > 0$ is defined as $k_s(x, y) = \exp(-s\|x - y\|_2^2)$, while the Quadratic kernel with scale $s > 0$ is defined as $k_Q(x, y) = \left(1 + s(x^T y)\right)^2$. With $w$ denoting the median of the pairwise distance between all the observations, we set $s = 1/(2w^2)$ for the Gaussian kernel and $s = 1/w$ for the Quadratic kernel.

**Details for Figure 4.** Given observations $X_1, X_2, \ldots, X_n$i.i.d.$P$, consider the problem of one-sample mean-testing, that is, testing $H_0 : \mathbb{E}[X_i] = \mathbf{0}$ versus $H_1 : \mathbb{E}[X_i] = a \neq \mathbf{0}$. When the distribution $P$ is a multivariate Gaussian, Kim and Ramdas (2020) showed that power of their test using a one-sample studentized U-statistic based on a bi-linear kernel is asymptotically $\Phi\left(z_\alpha + \frac{a^T a}{2\sqrt{\text{tr}(\Sigma^2)}}\right)$. The power achieved by the test using the full U-statistic is $\Phi\left(z_\alpha + \frac{a^T a}{\sqrt{2\text{tr}(\Sigma^2)}}\right)$, which differs from the previous expression by a factor of $\sqrt{2}$. A similar relation also holds for the problem of Gaussian covariance testing. Our heuristic in (9) is based on these two observations.

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

## E.2 Additional Figures

**Null Distribution.** Figure 6 denotes the null distribution of our proposed statistic ($\widehat{\text{xMMD}}^2$) along with that of the usual MMD normalized by its empirical standard deviation. The null distribution in Figure 6 is `Dirichlet` with parameter $2 \times \mathbf{1} \in \mathbb{R}^d$ for $d \in \{10, 500\}$.

**Power Curves.** In Figure 7, we plot the power curves for the different tests using a Gaussian Kernel, and we report the results of the same experiment with a polynomial kernel of degree 5 in Figure 8. Recall that the polynomial kernel of degree $r$ and scale parameter $s > 0$ is defined as $k(x, y) = \left(1 + (x^T y)/s\right)^r$. In both instances, we selected the scale parameter using the median heuristic.

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

**ROC curves.** In Figure 9, we plot some additional ROC curves for the different statistics. As before, we used 1000 'null trials' and another 1000 'alternative trials' with sample sizes $n = 200$ and $m = 200$. The data generating distributions $P$ and $Q$ were both Dirichlet with parameters $\mathbf{1} \in \mathbb{R}^d$ and $(1 + \epsilon) \times \mathbf{1} \in \mathbb{R}^d$ for $(d, \epsilon) \in \{(10, 0.4), (100, 0.2), (500, 0.15)\}$.

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

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

- Fix $J$, and choose points $\{v_1, \ldots, v_J\}$ from $\mathbb{R}^d$, where $d$ is the dimension of the observation space.

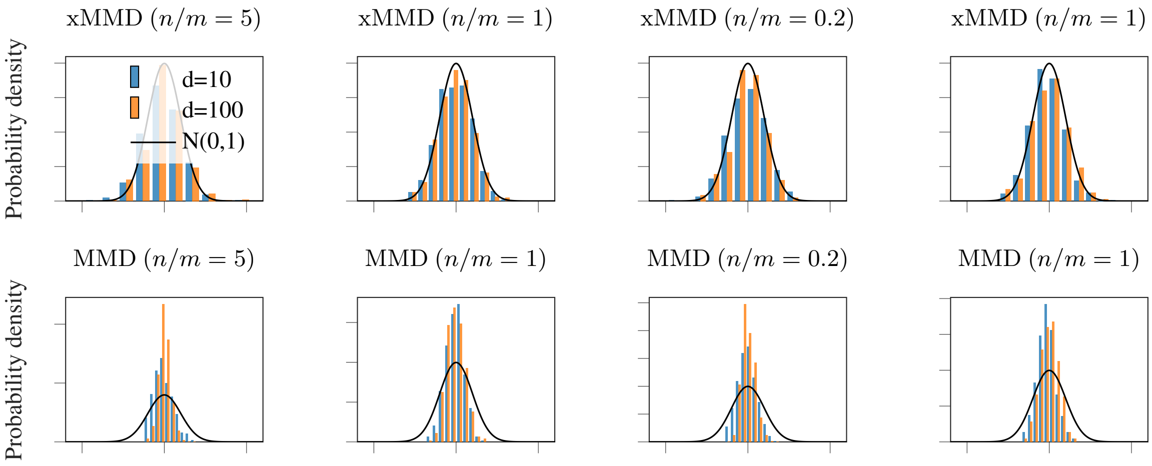

Figure 6: The first two columns show the null distribution of the $\widehat{\mathrm{xMMD}}^2$ statistic (top row) and the $\widehat{\mathrm{MMD}}^2$ statistic scaled by its empirical standard deviation (bottom row) using the Gaussian kernel with scale-parameter chosen using the median heuristic. The last two columns show the null distribution for the two statistics using the Polynomial kernel of degree $5$ with scale parameter chosen using the median heuristic. The figures demonstrate that the null distribution of $\widehat{\mathrm{MMD}}^2$ changes significantly with dimension ($d$), the ratio $n/m$ and the choice of the kernel, unlike our proposed statistic.

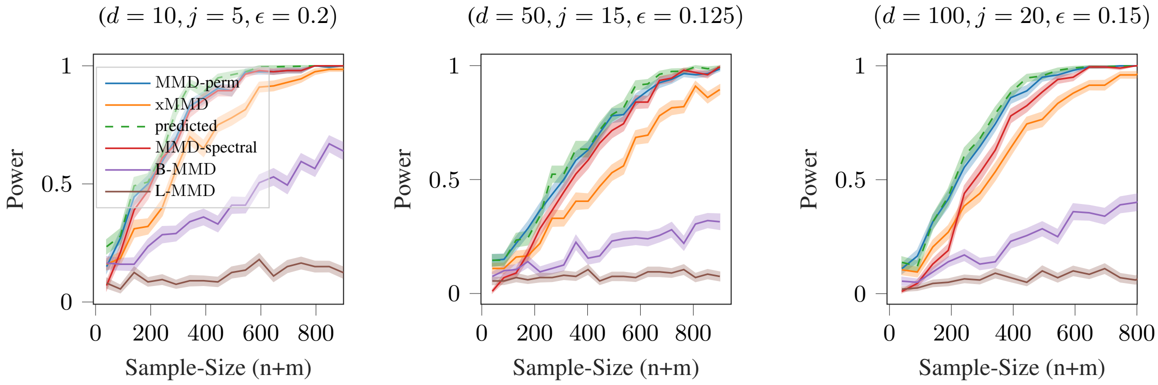

Figure 7: Power Curves for the different tests using Gaussian kernel with scale parameter chosen via median heuristic. The two distributions are $P = N(\mathbf{0}, I_d)$ and $Q = N(a_{\epsilon,j}, I_d)$ where $a_{\epsilon,j}$ is obtained by setting the first $j \le d$ coordinates of $\mathbf{0} \in \mathbb{R}^d$ equal to $\epsilon$. The figures demonstrate that the xMMD test is competitive with more computationally expensive tests (MMD-perm and MMD-spectral), while performing significantly better than the low complexity alternatives (B-MMD and L-MMD). The batch-size used in the B-MMD test was $\sqrt{n}$.

- Using $\mathbb{X}$ and $\mathbb{Y}$ with $n = m$, compute $\{z_i : 1 \le i \le n\}$, where $z_i = [k(v_J, X_i) - k(v_J, Y_i)]_{j=1}^J \in \mathbb{R}^J$ for ME test, and $z_i = [\hat{l}(X_i)\sin(X_i^T v_j) - \hat{l}(Y_i)\sin(Y_i^T v_j), \hat{l}(X_i)\cos(X_i^T v_j) - \hat{l}(Y_i)\cos(Y_i^T v_j)]_{j=1}^J \in \mathbb{R}^{2J}$ for the SCF test. Define $\bar{z}_n = \frac{1}{n}\sum_{i=1}^n z_i$, and $S_n = \frac{1}{n-1}(z_i - \bar{z}_n)(z_i - \bar{z}_n)^T$.
- Using the above, define the test statistic

$$\hat{\lambda}_n \coloneqq \bar{z}_n^T \left(S_n + \gamma_n I\right)^{-1} \bar{z}_n,$$

where $\gamma_n$ is some regularization parameter that converges to $0$ with $n$, and $I$ denotes the identity matrix. For a fixed $d$ and $J$, Jitkrittum et al. (2016) show that the above statistic has

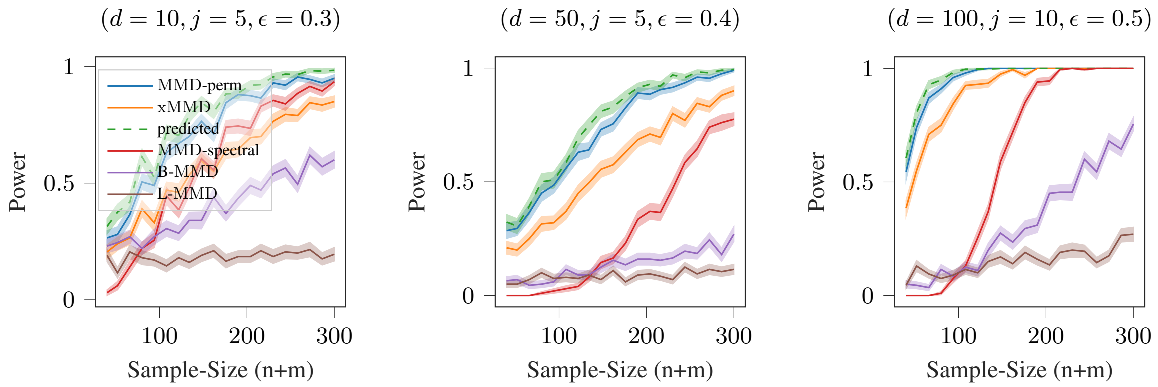

Figure 8: Power curves of the different kernel-based tests using a polynomial kernel of degree 5, i.e., $k(x, y) = \left(1 + (x^T y)/s\right)^5$ with $s$ chosen via the median heuristic.

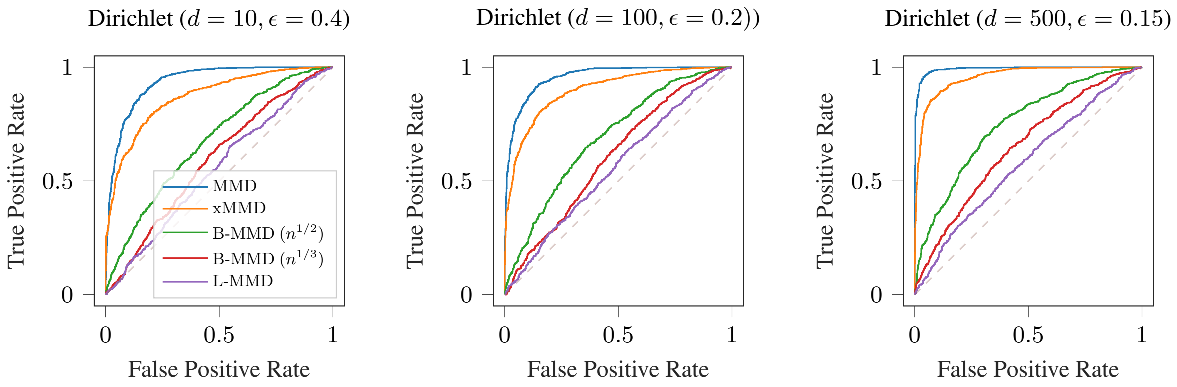

Figure 9: ROC curves using the different statistics with Gaussian kernel for testing two Dirichlet distributions in dimensions $d \in \{10, 100, 500\}$ with sample-size $n = m = 200$. The two distributions are $P = \texttt{Dirichlet}(\mathbf{1})$ and $Q = \texttt{Dirichlet}((1 + \epsilon) \times \mathbf{1})$ where $\mathbf{1} \in \mathbb{R}^d$ is the all-ones vector.

a $\chi^2(J)$ (resp. $\chi^2(2J)$) limiting null distribution in the ME (resp. SCF) case. This result is used to calibrate the test at a given level $\alpha$.

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

The ME and SCF tests are calibrated based on the limiting distribution of their statistic in the low dimensional regime: fixed $d$, and $n \to \infty$. However, the high type-I error of these tests for small $n$ values suggests that their limiting distribution may be different in the high dimensional regime, when both $d$ and $n$ go to infinity. We further observe this in Figure 11 when $d = 100$ and $d/n > 1$.

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

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

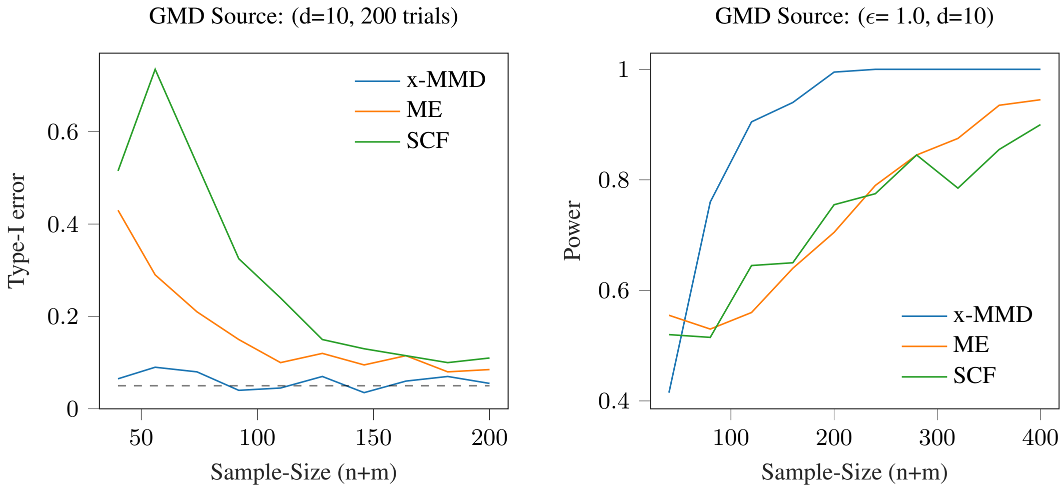

Figure 10: The figures plot the variation of the type-I error (left) and the power (right) with sample-size of the three tests: cross-MMD, and the two linear time tests, ME and SCF, proposed by Jitkrittum et al. (2016).

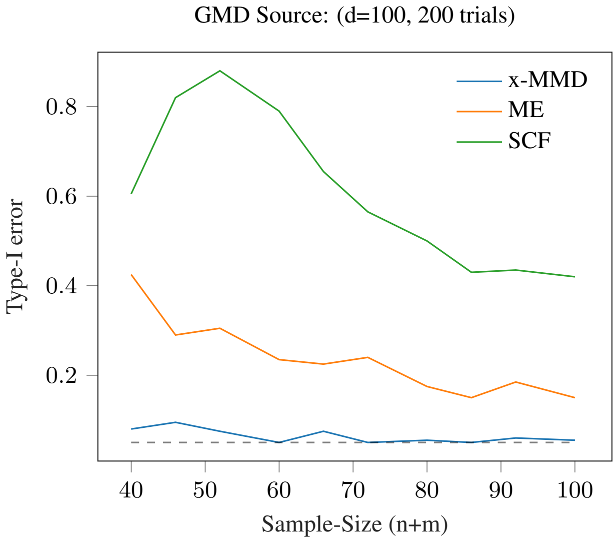

Figure 11: The ME and SCF tests provide poor control over the type-I error in the regime when $d/n$ is large, suggesting that the limiting null distribution is different (or the convergence rate is slow) in this regime.

- More importantly, the ME and SCF tests are only valid in the 'low-dimensional setting': fixed $d$ and $J$, with $n \to \infty$. In the high dimensional setting, when $(d, n) \to \infty$, the limiting null distribution may no longer be $\chi^2(J)$. This is also suggested by the behavior of type-I error of ME and SCF tests in Figure 10 and Figure 11. This results in the following practical issue: *given a problem with* $n = 500$ *and* $d = 200$, *how should one calibrate the threshold for those tests?*

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

- In the regime where the number of features, $J$, is allowed to increase with $n$, we expect that the resulting ME and SCF tests may have low power (for small regularization parameter $\gamma_n$). This is because, the test statistic $\hat{\lambda}_n$ used by ME and SCF tests is similar to Hotelling's $T^2$

statistic, for which Bai and Saranadasa (1996) characterized the asymptotic power in this regime. In particular, their Theorem 2.1 implies that the power of the $T^2$ test grows slowly with $n$, especially when $J/n \approx 1$.

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

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

In this section, we experimentally verify the limiting Gaussian distribution of the $\mathrm{x}\widehat{\mathrm{MMD}}^2$ statistic under the null. We first plot the variation of the type-I error of our cross-MMD test with sample size in Figure 12. We considered the case when $\mathbb{X}$ and $\mathbb{Y}$ are both drawn i.i.d. from a multivariate Gaussian vector in dimension $d \in \{10, 100\}$, and $n = m$.

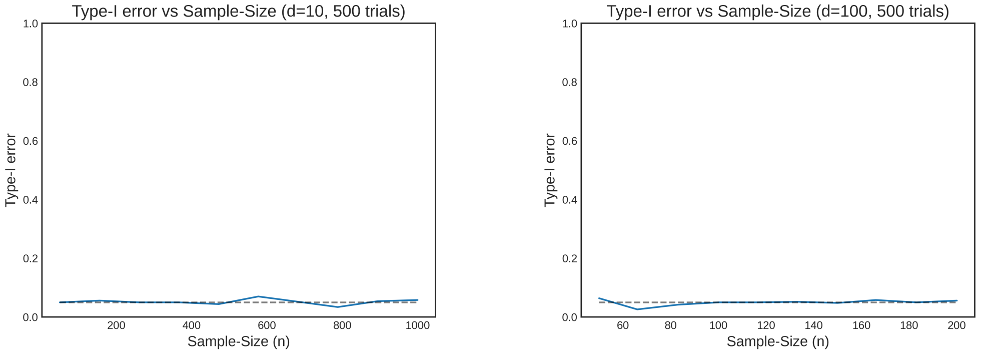

Figure 12: The two figures show the variation of the type-I error of the cross-MMD test with sample-size for dimensions $d \in \{10, 100\}$. The dashed horizontal line denotes the level $\alpha = 0.05$. In summary, these tests do not find evidence against the null hypothesis that the null distribution is Gaussian.

Next, we plot the p-values for the test for normality proposed by D'Agostino and Pearson (1973), and implemented in the function `scipy.stats.normaltest` in Python. We performed this test at different sample-sizes ($n$), and for each value of $n$, we calculated the $\mathrm{x}\widehat{\mathrm{MMD}}^2$ statistic on 200 different indpendent sample pairs. The results are shown in Figure 13

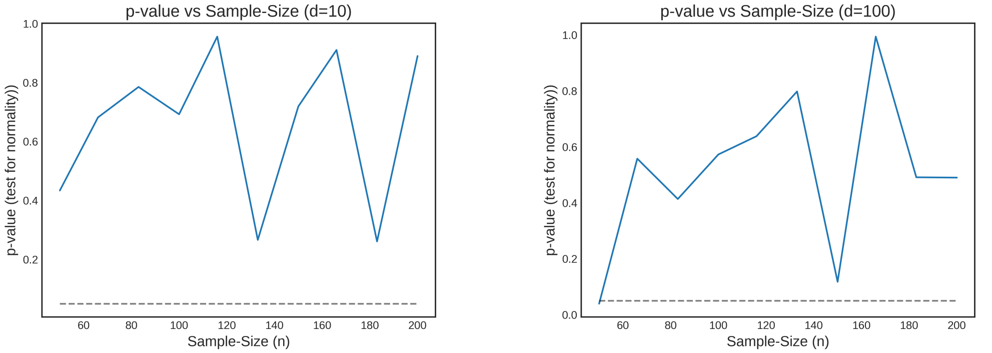

Figure 13: The two figures show p-values for the test for normality proposed by D'Agostino and Pearson (1973) (using the implementation `scipy.stats.normaltest`) of the cross-MMD statistic for dimensions $d \in \{10, 100\}$. In both dimension regimes, the test does not find evidence against the null that the cross-MMD statistic is normally distributed under the null.

### E.5 Comparison with Friedman-Rafsky test

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

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

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

In Figure 14, we compare the power of the FR permutation-test with our cross-MMD test in a low dimensional ($d/n$ small) and a high dimensional ($d/n$ large) problem. In both cases, it is observed that the power of FR test is significantly smaller than that of cross-MMD test.

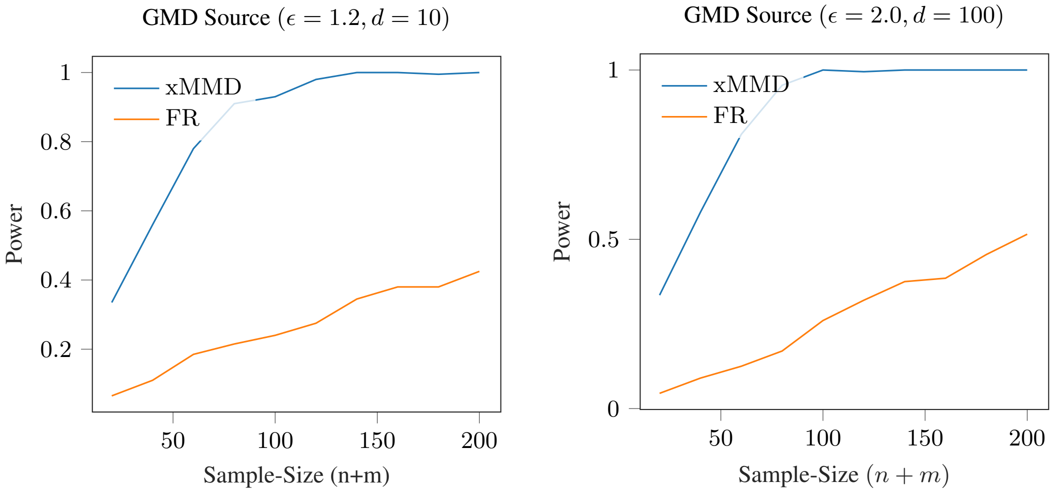

Figure 14: The figures show the power curves for Friedman-Rafsky (FR) test and our cross-MMD test in the low ($d = 10$) and high ($d = 100$) dimensional settings with $m = n$ in both plots. The figures indicate that our cross-MMD test is significantly more powerful than the FR test.