General Frameworks for Conditional Two-Sample Testing

Yonsei University

July 8, 2025

1 Block-wise MMD Framework

In this section, we provide a detailed account of our second framework, which is based on the block-wise MMD introduced by Zaremba et al. (2013). This statistic converges to a Gaussian distribution asymptotically and, with a suitably chosen block size B, can achieve higher power compared to linear-time alternatives. Throughout, we consider balanced samples with $n_1 = n_2 = 2n$ and divide the data into two equal parts:

$$\begin{aligned} D_a &:= \big\{V_i^{(1)}\big\}_{i=1}^{n} \cup \big\{V_i^{(2)}\big\}_{i=1}^{n}, \\ D_b &:= \big\{V_i^{(1)}\big\}_{i=n+1}^{2n} \cup \big\{V_i^{(2)}\big\}_{i=n+1}^{2n}. \end{aligned}$$

An estimator $\widehat{r}_X$ of the density ratio $r_X := f_X^{(1)}/f_X^{(2)}$ is fitted using D_b only. Let $S := \lfloor n/B \rfloor$ for the number of blocks. For $b = 1, \ldots, S$ define the index set

$$I_b := \big\{(i,j) \in \mathbb{N}^2 : (b-1)B < i < j \leq bB\big\}.$$

The pairwise kernel contribution for $(i,j) \in I_b$ is

$$\begin{aligned} \widehat{H}_{ij} := {} & k\big(V_i^{(1)}, V_j^{(1)}\big) + \widehat{r}_X(X_i^{(2)})\, \widehat{r}_X(X_j^{(2)})\, k\big(V_i^{(2)}, V_j^{(2)}\big) \\ & - \widehat{r}_X(X_i^{(2)})\, k\big(V_i^{(2)}, V_j^{(1)}\big) - \widehat{r}_X(X_j^{(2)})\, k\big(V_i^{(1)}, V_j^{(2)}\big). \end{aligned}$$

Let $b = 1, \ldots, S$. The corresponding block statistics and their overall average are defined as follows:

$$\widehat{\eta}_b := \binom{B}{2}^{-1} \sum_{(i,j) \in I_b} \widehat{H}_{ij}.$$

Letting $\bar{\eta} := \frac{1}{S}\sum_{b=1}^{S} \widehat{\eta}_b$. and $\widehat{\sigma}_B^2 := \frac{1}{S-1}\sum_{b=1}^{S}\big(\widehat{\eta}_b - \bar{\eta}\big)^2$, and the studentized block-wise MMD statistic is given as

$$\widehat{\mathrm{MMD}}_B^2 := \frac{\sqrt{S}\,\bar{\eta}}{\widehat{\sigma}_B}.$$

To establish the asymptotic normality of $\widehat{\mathrm{MMD}}_B^2$, we make the following assumptions. Let H_{ij} denote the quantity defined similarly to $\widehat{H}_{ij}$, except that $\widehat{r}_X$ is replaced by the population counterpart r_X. Moreover, denote η_b as the counterpart of $\hat{\eta}_b$ using H_{ij}.

Assumption 1. *Consider a class of null distributions $\mathcal{P}_0$ and assume that*

(a) There exist constants $c, C > 0$ such that $\inf_{P \in \mathcal{P}_0} \mathbb{E}_P\left[\eta_1^2\right] \geq c$ and $\sup_{P \in \mathcal{P}_0} \mathbb{E}_P\left[\eta_1^{2+\delta}\right] \leq C$ for some $\delta > 0$.

(b) $\sup_{P \in \mathcal{P}_0} \mathbb{E}_P\big[r_X(X^{(2)})^2\big] < \infty$ and $\sup_{P \in \mathcal{P}_0} \mathbb{E}_P\big[\{\widehat{r}_X(X^{(2)}) - r_X(X^{(2)})\}^2\big] = o\big(n^{-1/2}\big)$.

(c) The kernel is uniformly bounded as $\|k\|_\infty \le K$.

(d) The block size B satisfies $B = o(n^\gamma)$ for some $0 < \gamma < 1$.

Assumption 1(a) specifies a moment condition for the population counterpart η_b. This condition is critical for applying the uniform central limit theorem to the block-wise statistics. Assumption 1(b) ensures that the difference between the test statistic using the estimated kernel entries $\{\hat{H}_{ij}\}_{(i,j)\in I}$ and the one using the population counterparts $\{H_{ij}\}_{(i,j)\in I}$ (where $I := \{(i,j) \in [n]^2 \mid i \neq j\}$) becomes asymptotically negligible. While this assumption is similar to one discussed by Hu and Lei (2024), it imposes a considerably weaker convergence rate requirement. Assumption 1(c) requires the kernel k to be uniformly bounded, which is typically satisfied by commonly used kernels such as the Gaussian kernel. Lastly, Assumption 1(d) sets a condition on the block size B, noting that using larger blocks incorporates more samples, which helps reduce variance. However, this also leads to increased computational cost. Importantly, having at least two blocks ($S \ge 2$) is crucial for the central limit theorem to hold for the block means, ensuring that the normalized statistic converges to a Gaussian distribution.

Theorem 1. *For the class of null distributions $\mathcal{P}_0$ satisfying 1, $\widehat{\mathrm{MMD}}_B^2$ converges to $N(0,1)$ as*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} \sup_{t\in\mathbb{R}} \left| \mathbb{P}_P\left(\widehat{\mathrm{MMD}}_B^2 \le t\right) - \Phi(t)\right| = 0.$$

Proof)

Analysis of the numerator. Using the fact that a kernel can be expressed as an inner product of feature maps, $k(x,y) = \langle \psi(x), \psi(y)\rangle$, we can rewrite $\bar{\eta}$ as

$$\bar{\eta} = \frac{1}{S}\sum_{b=1}^{S}\binom{B}{2}^{-1}\sum_{(i,j)\in I_b}\left\langle \psi(V_i^{(1)}) - \widehat{r}_X(X_i^{(2)})\psi(V_i^{(2)}), \psi(V_j^{(1)}) - \widehat{r}_X(X_j^{(2)})\psi(V_j^{(2)})\right\rangle.$$

By adding and subtracting $r_X(X_i^{(2)})\psi(V_i^{(2)})$ and $r_X(X_j^{(2)})\psi(V_j^{(2)})$, $\sqrt{S}\bar{\eta}$ can be written as the sum of the four terms given as:

$$(\mathbb{I}) := \frac{1}{\sqrt{S}}\sum_{b=1}^{S}\underbrace{\binom{B}{2}^{-1}\sum_{(i,j)\in I_b}\langle \psi(V_i^{(1)}) - r_X(X_i^{(2)})\psi(V_i^{(2)}), \psi(V_j^{(1)}) - r_X(X_j^{(2)})\psi(V_j^{(2)})\rangle}_{:=\eta_b}$$

$$(\mathbb{II}) := \frac{1}{\sqrt{S}}\sum_{b=1}^{S}\underbrace{\binom{B}{2}^{-1}\sum_{(i,j)\in I_b}\langle \psi(V_i^{(1)}) - r_X(X_i^{(2)})\psi(V_i^{(2)}), \psi(V_j^{(2)})\rangle\cdot\{\hat{r}_X(X_j^{(2)}) - r_X(X_j^{(2)})\}}_{:=\hat{\eta}_{b,1}}$$

$$(\mathbb{III}) := \frac{1}{\sqrt{S}}\sum_{b=1}^{S}\underbrace{\binom{B}{2}^{-1}\sum_{(i,j)\in I_b}\langle \psi(V_i^{(2)}), \psi(V_j^{(1)}) - r_X(X_j^{(2)})\psi(V_j^{(2)})\rangle\cdot\{\hat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\}}_{:=\hat{\eta}_{b,2}}$$

$$(\mathbb{IV}) := \frac{1}{\sqrt{S}}\sum_{b=1}^{S}\underbrace{\binom{B}{2}^{-1}\sum_{(i,j)\in I_b}\langle \psi(V_i^{(2)}), \psi(V_j^{(2)})\rangle\{\hat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\}\cdot\{\hat{r}_X(X_j^{(2)}) - r_X(X_j^{(2)})\}}_{:=\hat{\eta}_{b,3}}$$

Our goal is to show that, asymptotically, $\sqrt{S}\,\bar{\eta} = (\mathbb{I}) + o_{\mathcal{P}_0}(1)$ holds.

1. **Term** $(\mathbb{I})$**.** In term $(\mathbb{I})$, the population density ratio is known, implying that the sequence $\eta_1, \ldots, \eta_S$ can be treated as independent and identically distributed. Under Assumption 1(a), Lemma 18 of Shah and Peters (2020) ensures that

$$\sup_{P \in \mathcal{P}_0} \sup_{t \in \mathbb{R}} \left| \mathbb{P}_P\left(\sigma_P^{-1}(\mathbb{I}) \le t\right) - \Phi(t) \right| \to 0.$$

2. **Terms** $(\mathbb{II})$ **and** $(\mathbb{III})$**.** We only analyze Term $(\mathbb{II})$ by symmetry, Term $(\mathbb{III})$ can be treated in the same manner. Let $\widehat{\eta}_{b,1}(i,j)$ denote the (i,j)-th pairwise component of the block estimator $\widehat{\eta}_{b,1}$, capturing the interaction between samples i and j within the b-th block. Under H_0, we first observe that each pairwise term has zero conditional mean given the density ratio estimate data D_b:

$$\begin{aligned}
\mathbb{E}\big[\widehat{\eta}_{b,1}(i,j) \mid D_b\big] &= \mathbb{E}\Big[\big\langle \psi(V_i^{(1)}) - r_X(X_i^{(2)})\, \psi(V_i^{(2)}),\, \psi(V_j^{(2)}) \big\rangle \big\{ \widehat{r}_X(X_j^{(2)}) - r_X(X_j^{(2)}) \big\} \mid D_b \Big] \\
&= \big\{ \widehat{r}_X(X_j^{(2)}) - r_X(X_j^{(2)}) \big\} \underbrace{\mathbb{E}\big[\langle \psi(V_i^{(1)}) - r_X(X_i^{(2)})\, \psi(V_i^{(2)}),\, \psi(V_j^{(2)}) \rangle \mid D_b \big]}_{=0 \text{ under } H_0} \\
&= 0,
\end{aligned}$$

Since $\hat{\eta}_{b,1}$ is an average of these terms and D_b is fixed within each block, we have $\mathbb{E}[\hat{\eta}_{b,1} \mid D_b] = 0$. Taking the expectation over D_b then gives $\mathbb{E}[\hat{\eta}_{b,1}] = \mathbb{E}_{D_b}[\mathbb{E}[\hat{\eta}_{b,1} \mid D_b]] = 0$, establishing the unconditional mean zero. Let $A_{ij} := \big\langle \psi(V_i^{(1)}) - r_X(X_i^{(2)})\, \psi(V_i^{(2)}),\, \psi(V_j^{(2)}) \big\rangle$ and $B_j := \widehat{r}_X(X_j^{(2)}) - r_X(X_j^{(2)})$, without loss of generality let $b = 1$,

$$\begin{aligned}
\mathbb{E}[\hat{\eta}_{1,1}^2 \mid D_b] &= \mathbb{E}\Bigg[\Big(\frac{2}{B(B-1)} \sum_{i<j}^{B} A_{ij} B_j \Big)^2 \mid D_b \Bigg] \\
&= \frac{4}{B^2(B-1)^2} \sum_{i<j} \sum_{k<\ell} \mathbb{E}\big[A_{ij} B_j\, A_{k\ell} B_\ell \mid D_b\big] \\
&\overset{(i)}{=} \frac{4}{B^2(B-1)^2} \Bigg[\sum_{i<j} \mathbb{E}\big[A_{ij}^2 B_j^2 \mid D_b\big] + \sum_{i<j} \sum_{\substack{\ell > i \\ \ell \ne j}} \mathbb{E}\big[A_{ij} B_j\, A_{i\ell} B_\ell \mid D_b\big] \\
&\quad + \sum_{i<j} \sum_{\substack{k<j \\ k \ne i}} \mathbb{E}\big[A_{ij} B_j\, A_{kj} B_j \mid D_b\big] + \sum_{i<j} \sum_{\substack{k<i \\ k \ne j}} \mathbb{E}\big[A_{ij} B_j\, A_{ki} B_i \mid D_b\big] \\
&\quad + \sum_{i<j} \sum_{\substack{\ell > j \\ \ell \ne i}} \mathbb{E}\big[A_{ij} B_j\, A_{j\ell} B_\ell \mid D_b\big] \Bigg] \\
&= \frac{2}{B(B-1)} \mathbb{E}\big[A_{12}^2 B_2^2 \mid D_b\big] \\
&\quad + \frac{4(B-2)}{3B(B-1)} \Big\{ \mathbb{E}\big[A_{12} B_2\, A_{13} B_3 \mid D_b\big] + \mathbb{E}\big[A_{13} B_3\, A_{23} B_3 \mid D_b\big] + \mathbb{E}\big[A_{23} B_3\, A_{12} B_2 \mid D_b\big] \Big\} \\
&\quad + \frac{2(B-5)}{3B(B-1)} \mathbb{E}\big[A_{12} B_2\, A_{23} B_3 \mid D_b\big] \\
&:= T_1 + T_2 + T_3,
\end{aligned}$$

where (i) separate the sum into the complete match case $(i = k,\, j = \ell)$ and the four single-overlap cases: $(i = k,\, j \ne \ell)$, $(i \ne k,\, j = \ell)$, $(i = \ell,\, j \ne k)$, and $(j = k,\, i \ne \ell)$. All remaining terms with fully distinct indices vanish under H_0 due to conditional independence given D_b. Therefore, only the exact match and the four overlapping cases contribute to the variance expansion.

Among these, T_1 vanishes faster than T_2 and T_3 as $B \to \infty$, so we focus on bounding the latter two terms. Using the fact that $\langle \psi(x), \psi(y) \rangle = k(x, y)$ is uniformly bounded by a constant K, and assuming for each $i \in \{1, \ldots, B\}$ that

$$\sup_{P \in \mathcal{P}_0} \mathbb{E}_P \left[\left\{ \hat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)}) \right\}^2 \right] = o(n^{-1/2}) \quad \text{and} \quad \sup_{P \in \mathcal{P}_0} \mathbb{E}_P \left[r_X(X_i^{(2)})^2 \right] < \infty,$$

we obtain the following bounds:

$$\begin{aligned}
T_2 &\lesssim \frac{K^2}{B} \left(1 + \mathbb{E}[r_X(X_2^{(2)})]\right) \left(1 + \mathbb{E}[r_X(X_3^{(2)})]\right) \mathbb{E}[\hat{r}_X(X_2^{(2)}) - r_X(X_2^{(2)})] \, \mathbb{E}[\hat{r}_X(X_3^{(2)}) - r_X(X_3^{(2)})] \\
&\quad + \frac{K^2}{B} \left(1 + \mathbb{E}[r_X(X_2^{(2)})^2]\right) \mathbb{E}\left[\left\{ \hat{r}_X(X_3^{(2)}) - r_X(X_3^{(2)}) \right\}^2 \right] \\
&\overset{(ii)}{=} o_{\mathcal{P}_0}(B^{-1} n^{-1/2}),
\end{aligned}$$

where (ii) uses Cauchy–Schwarz inequality and Lundborg et al. (2022, Lemma S5). Similarly,

$$\begin{aligned}
T_3 &\lesssim \frac{K^2}{B} \left(1 + \mathbb{E}[r_X(X_2^{(2)})]\right) \left(1 + \mathbb{E}[r_X(X_3^{(2)})]\right) \\
&\qquad \times \mathbb{E}[\hat{r}_X(X_2^{(2)}) - r_X(X_2^{(2)})] \, \mathbb{E}[\hat{r}_X(X_3^{(2)}) - r_X(X_3^{(2)})] \\
&= o_{\mathcal{P}_0}(B^{-1} n^{-1/2}).
\end{aligned}$$

Therefore, under Assumption 1(d), the conditional variance of $(\mathbb{III})$ satisfies

$$\mathrm{Var}[(\mathbb{III}) \mid D_b] = \frac{1}{S} \sum_{b=1}^{S} \mathrm{Var}[\hat{\eta}_{b,1} \mid D_b] = o_{\mathcal{P}_0}\Big(\frac{1}{Bn^{1/2}} \Big) = o_{\mathcal{P}_0}\Big(n^{-\frac{1+2\gamma}{2}} \Big),$$

for some $0 < \gamma < 1$. Applying Chebyshev's inequality yields $(\mathbb{III}) = o_{\mathcal{P}_0}(1)$ and similarly $(\mathbb{IIII}) = o_{\mathcal{P}_0}(1)$.

3. **Term** $(\mathbb{IV})$**.** The final term $(\mathbb{IV})$ is defined as

$$(\mathbb{IV}) := \sqrt{S} \cdot \frac{1}{S} \sum_{b=1}^{S} \hat{\eta}_{b,3}.$$

Since the kernel k is uniformly bounded by K, we have

$$\begin{aligned}
|\hat{\eta}_{b,3}| &\leq K \left| \frac{2}{B(B-1)} \sum_{(i,j) \in I_b} \{\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\} \cdot \{\widehat{r}_X(X_j^{(2)}) - r_X(X_j^{(2)})\} \right| \\
&\leq \frac{2K}{B(B-1)} \sum_{(i,j) \in I_b} \left|\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\right| \cdot \left|\widehat{r}_X(X_j^{(2)}) - r_X(X_j^{(2)})\right| \\
&\overset{(i)}{\leq} \frac{K}{B(B-1)} \Bigg(\sum_{i=(b-1)B+1}^{bB} \left|\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\right| \Bigg)^2 \\
&\overset{(ii)}{\leq} \frac{K \cdot B^2}{B(B-1)} \cdot \frac{1}{B} \sum_{i=(b-1)B+1}^{bB} \left\{ \widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)}) \right\}^2,
\end{aligned}$$

where (i) uses the inequality $\sum_{i<j} |a_i||a_j| \leq \frac{1}{2} \left(\sum_i |a_i| \right)^2$, and (ii) follows from Jensen's inequality.

Thus, for each block b, we obtain the bound

$$|\hat{\eta}_{b,3}| \leq \frac{K \cdot B^2}{B(B-1)} \cdot \frac{1}{B} \sum_{i=(b-1)B+1}^{bB} \left\{\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\right\}^2 =: \Delta_b.$$

Assuming

$$\sup_{P \in \mathcal{P}_0} \mathbb{E}_P\left[\{\widehat{r}_X(X^{(2)}) - r_X(X^{(2)})\}^2\right] = o(n^{-1/2}),$$

we can apply Markov's inequality to conclude that $\Delta_b = o_{\mathcal{P}_0}(n^{-1/2})$ uniformly over b. Since $(\mathbb{IV}) = \sqrt{S} \cdot \frac{1}{S}\sum_{b=1}^{S} \hat{\eta}_{b,3}$, another application of Markov's inequality yields

$$(\mathbb{IV}) = \sqrt{S} \cdot o_{\mathcal{P}_0}(n^{-1/2}) = \left\lfloor \frac{n}{B} \right\rfloor^{1/2} \cdot o_{\mathcal{P}_0}(n^{-1/2}) = o_{\mathcal{P}_0}(B^{-1/2}) = o_{\mathcal{P}_0}(n^{-\gamma/2}),$$

where the final equality holds under Assumption 1(d) and which implies $(\mathbb{IV}) = o_{\mathcal{P}_0}(1)$.

Combining the bounds for all four terms, we conclude that

$$\sup_{P \in \mathcal{P}_0} \sup_{t \in \mathbb{R}} \left| \mathbb{P}_P\left(\sqrt{S}\, \sigma_P^{-1} \bar{\eta} \leq t\right) - \Phi(t) \right| \to 0,$$

Consistency of the variance estimate. Denoting

$$\hat{\sigma}_P^2 := \frac{1}{S-1} \sum_{b=1}^{S} \left(\hat{\eta}_b - \bar{\eta}\right)^2,$$

We aim to demonstrate that the ratio $\hat{\sigma}_P^2/\sigma_P^2$ converges to one in probability, which, in turn, implies $\hat{\sigma}_P/\sigma_P = 1 + o_{\mathcal{P}_0}$ following (Lundborg et al., 2022, Lemma 7). Given that the test statistic $\widehat{\mathrm{MMD}}_B^2$ is scale-invariant, we can assume $\sigma_P^2 = 1$ without any loss of generality. Additionally, the preceding analysis confirms that $\bar{\eta} = o_{\mathcal{P}_0}(1)$. Consequently, it suffices to show that $\frac{1}{S}\sum_{b=1}^{S} \hat{\eta}_b$ converges to one in probability. To establish this, note that:

$$\begin{aligned}
\left| \frac{1}{S}\sum_{b=1}^{S} \hat{\eta}_b^2 - 1 \right| &= \left| \frac{1}{S}\sum_{b=1}^{S} (\eta_b + \hat{\eta}_{b,1} + \hat{\eta}_{b,2} + \hat{\eta}_{b,3})^2 - 1 \right| \\
&\leq \left| \frac{1}{S}\sum_{b=1}^{S} \hat{\eta}_b^2 - 1 \right| + \left| \frac{1}{S}\sum_{b=1}^{S} (\hat{\eta}_{b,1} + \hat{\eta}_{b,2} + \hat{\eta}_{b,3})^2 \right| + 2\left| \frac{1}{S}\sum_{b=1}^{S} \eta_b(\hat{\eta}_{b,1} + \hat{\eta}_{b,2} + \hat{\eta}_{b,3}) \right| \\
&\leq \left| \frac{1}{S}\sum_{b=1}^{S} \hat{\eta}_b^2 - 1 \right| + \left| \frac{1}{S}\sum_{b=1}^{S} (\hat{\eta}_{b,1} + \hat{\eta}_{b,2} + \hat{\eta}_{b,3})^2 \right| \\
&\qquad + 2\sqrt{\frac{1}{S}\sum_{b=1}^{S} \eta_b^2} \sqrt{\frac{1}{S}\sum_{b=1}^{S} (\hat{\eta}_{b,1} + \hat{\eta}_{b,2} + \hat{\eta}_{b,3})^2},
\end{aligned}$$

where the last inequality is a consequence of the Cauchy–Schwarz inequality. Applying the law of large numbers, we see that $\frac{1}{S}\sum_{b=1}^{S} \eta_b^2$ converges to one in probability. Consequently, the proof reduces to establishing that $\frac{1}{S}\sum_{b=1}^{S} (\hat{\eta}_{b,1} + \hat{\eta}_{b,2} + \hat{\eta}_{b,3})^2 = o_{\mathcal{P}_0}(1)$. This follows from showing that

$$\frac{1}{S}\sum_{b=1}^{S} \hat{\eta}_{b,1} = o_{\mathcal{P}_0}(1), \quad \frac{1}{S}\sum_{b=1}^{S} \hat{\eta}_{b,2} = o_{\mathcal{P}_0}(1), \quad \frac{1}{S}\sum_{b=1}^{S} \hat{\eta}_{b,3} = o_{\mathcal{P}_0}(1).$$

These results follow from the same techniques used for $(\mathbb{II})$, $(\mathbb{III})$, and $(\mathbb{IV})$, together with Markov's inequality, thereby concluding the proof.

2 Cross-fitted Extension of MMD Statistic

To further enhance sample efficiency while preserving the favorable asymptotic properties of the block-wise MMD statistic, we adopt a K-fold cross-validation strategy, similar in spirit to the approach used in linear-time MMD tests.

Let the full dataset of size $2n$ be evenly partitioned into K disjoint folds, denoted by $D_1, \ldots, D_K$, each of size $2n/K$. For each $j \in \{1, \ldots, K\}$, let $D_a := D_j$ denote the held-out block used to compute the test statistic, and let $D_b := \cup_{i=1}^{K} D_i \setminus D_j$ be the union of the remaining $K-1$ folds used to estimate r_X. This strategy is theoretically justified because it leverages a larger portion of the data—specifically, a $(1-1/K)$ fraction—for estimating the density ratio, potentially leading to more accurate and stable estimation.

We compute $\bar{\eta}_j$ and $\widehat{\sigma}_j$ based on D_a and D_b, following the block-wise MMD construction. The cross-fitted statistic is then defined as

$$
{}^{\dagger}\widehat{\mathrm{MMD}}_{\mathrm{B}}^{2} := \frac{1}{K} \sum_{j=1}^{K} \frac{\sqrt{S}\bar{\eta}_j}{\widehat{\sigma}_j}.
$$

The following corollary extends the central limit theorem result to the cross-fitted statistic:

Corollary 1 (Asymptotic Normality of Cross-fitted Statistic)**.** *Under the conditions stated in Theorem 1, and for a fixed $K \geq 2$, the cross-validated test statistic satisfies:*

$$
\lim_{n \to \infty} \sup_{P \in \mathcal{P}_0} \sup_{t \in \mathbb{R}} \left| \mathbb{P}_P \left({}^{\dagger}\widehat{\mathrm{MMD}}_{\mathrm{B}}^{2} \leq t \right) - \Phi(t) \right| = 0.
$$

This result justifies the use of a test that rejects H_0 whenever ${}^{\dagger}\widehat{\mathrm{MMD}}_{\mathrm{B}}^{2} > \Phi^{-1}(1-\alpha)$ as an asymptotically level-α procedure. Moreover, since each fold contributes to both estimation and evaluation in a balanced manner, this cross-fitting strategy yields a more efficient use of the data, reducing variance while maintaining theoretical guarantees.

2.1 Proof of Corollary 1

For each $j \in \{1, \ldots, K\}$, the proof of Theorem 1 establishes that $\widehat{\sigma}_j / \sigma_P = 1 + o_{\mathcal{P}_0}(1)$. Therefore, by Lemma 20 of Shah and Peters (2020), it suffices to show the asymptotic normality of the standardized statistic

$$
\frac{1}{K} \sum_{j=1}^{K} \frac{\sqrt{S}\, \bar{\eta}_j}{\sigma_P},
$$

where $\bar{\eta}_j$ denotes the average of block-level statistics in fold j. Let the full dataset of size $2n$ be evenly divided into $S := \frac{2n}{B}$ disjoint blocks of size B, and assume that S is divisible by K. We assign the blocks to K disjoint subsets, denoted $\mathcal{B}_1, \ldots, \mathcal{B}_K$, such that each fold j contains exactly S/K blocks, i.e., $|\mathcal{B}_j| = S/K$. Each block $b \in \{1, \ldots, S\}$ consists of a set of index pairs (i, j), denoted $I_b \subset [B] \times [B]$, and the corresponding block-level statistic is defined as

$$
\eta_b := \binom{B}{2}^{-1} \sum_{(i,j) \in I_b} \left\langle \psi(V_i^{(1)}) - r_X(X_i^{(2)})\psi(V_i^{(2)}),\ \psi(V_j^{(1)}) - r_X(X_j^{(2)})\psi(V_j^{(2)}) \right\rangle.
$$

Then, for each fold j, we define the block-averaged statistic as

$$
\bar{\eta}_j := \frac{K}{S} \sum_{b \in \mathcal{B}_j} \eta_b.
$$

Summing over all folds yields

$$\sum_{j=1}^{K} \bar{\eta}_j = \frac{K}{S} \sum_{j=1}^{K} \sum_{b \in \mathcal{B}_j} \eta_b = \frac{K}{S} \sum_{b=1}^{S} \eta_b + o_{\mathcal{P}_0}(S^{-1/2}),$$

where we use the fact that the collection $\{\mathcal{B}_j\}$ partitions the S blocks. Importantly, $\sum_{b \in \mathcal{B}_1} \eta_b, \ldots, \sum_{b \in \mathcal{B}_K} \eta_b$ are mutually independent. Hence

$$\mathrm{Var}\left(\frac{K}{S} \sum_{j=1}^{K} \sum_{b \in \mathcal{B}_j} \frac{\eta_b}{\sigma_P} \right) = \frac{K^2}{S^2} \cdot K \cdot \frac{S}{K} = \frac{K^2}{S},$$

Finally, applying the central limit theorem (Shah and Peters, 2020, Lemma 18), we obtain

$$\frac{1}{K} \sum_{j=1}^{K} \frac{\sqrt{S}\, \bar{\eta}_j}{\sigma_P} = \frac{1}{K} \sum_{j=1}^{K} \sum_{b \in \mathcal{B}_j} \frac{\sqrt{S}\, \eta_b}{\sigma_P} + o_{\mathcal{P}_0}(1),$$

converges to $N(0,1)$ as desired.

References

Hu, X. and Lei, J. (2024). A two-sample conditional distribution test using conformal prediction and weighted rank sum. *Journal of the American Statistical Association*, 119(546):1136–1154.

Lundborg, A. R., Kim, I., Shah, R. D., and Samworth, R. J. (2022). The Projected Covariance Measure for assumption-lean variable significance testing. *arXiv preprint arXiv:2211.02039 (accepted to the Annals of Statistics)*.

Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538.

Zaremba, W., Gretton, A., and Blaschko, M. (2013). B-test: A non-parametric, low variance kernel two-sample test. *Advances in Neural Information Processing Systems*, 26.