General Frameworks for Conditional Two-Sample Testing

Seongchan Lee[1*] Suman Cha[1*] Ilmun Kim[2]

September 27, 2025

Abstract

We study the problem of conditional two-sample testing, which aims to determine whether two populations have the same distribution after accounting for confounding factors. This problem commonly arises in various applications, such as domain adaptation and algorithmic fairness, where comparing two groups is essential while controlling for confounding variables. We begin by establishing a hardness result for conditional two-sample testing, demonstrating that no valid test can have significant power against any single alternative without proper assumptions. We then introduce two general frameworks that implicitly or explicitly target specific classes of distributions for their validity and power. Our first framework allows us to convert any conditional independence test into a conditional two-sample test in a black-box manner, while preserving the asymptotic properties of the original conditional independence test. The second framework transforms the problem into comparing marginal distributions with estimated density ratios, which allows us to leverage existing methods for marginal two-sample testing. We demonstrate this idea in a concrete manner with classification and kernel-based methods. Finally, simulation studies are conducted to illustrate the proposed frameworks in finite-sample scenarios.

Contents

1 Introduction **2**
1.1 Problem Setup . . . 2
1.2 An Overview of Our Results . . . 3
1.3 Literature Review . . . 4
1.4 Organization . . . 5

2 Hardness Result **6**

3 Approach via Conditional Independence Testing **8**

4 Approach via Density Ratio Estimation **11**
4.1 Classifier-based Approach . . . 13
4.2 Linear-time MMD . . . 14
4.3 Block-wise MMD . . . 16
4.4 Quadratic-time MMD . . . 17

5 Numerical Experiments **20**
5.1 Overview of Testing Methods . . . 20
5.2 Synthetic Data Examples . . . 21
5.3 Real Data Analysis . . . 24

[*] These authors contributed equally to this work.
[1] Department of Statistics and Data Science, Yonsei University, Seoul, South Korea.
[2] Department of Statistics and Data Science, Department of Applied Statistics, Yonsei University, Seoul, South Korea.

6 Conclusion **25**

A Proofs **30**
A.1 Proof of Theorem 3 . . . 30
A.2 Proof of Corollary 1 . . . 31
A.3 Proof of Theorem 4 . . . 31
A.4 Proof of Corollary 2 . . . 34
A.5 Proof of Theorem 5 . . . 34
A.6 Proof of Corollary 3 . . . 41
A.7 Proof of Theorem 6 . . . 41
A.8 Proof of Proposition 1 . . . 46
A.9 Proof of Example 1 . . . 47

B Supporting Lemmas **49**

C Additional Numerical Experiments **56**
C.1 Experimental Details . . . 56
C.2 Impact of Density Ratio Estimation Errors on DRT Methods . . . 58
C.3 Real Data Analysis for CIT Methods . . . 58
C.4 Sensitivity Analysis of CIT Methods to Algorithm 1 . . . 59
C.5 Sensitivity Analysis of Algorithm 1 to ε . . . 60
C.6 Additional Simulations . . . 60
C.7 Computational Cost Analysis . . . 62

1 Introduction

This paper addresses the problem of testing for equivalence between two conditional distributions, namely conditional two-sample testing. Statistical methods for this problem have important applications across diverse fields such as domain adaptation and algorithmic fairness. In domain adaptation, for instance, this methodology can serve as a formal framework to validate the covariate shift assumption, where the conditional distribution of Y given X remains unchanged, while the marginal distributions of X may differ. By confirming this assumption, practitioners can effectively re-weight the training data according to the marginal density ratio regarding X, which potentially leads to improved predictive performance and better adaptation to new domains (Shimodaira, 2000; Sugiyama et al., 2007a,b). Moreover, in algorithmic fairness, conditional two-sample testing plays a role in detecting and mitigating biases. In particular, it helps identify whether a certain machine learning model unfairly favors or disadvantages specific groups based on demographic characteristics such as age, gender, or ethnicity (Hardt et al., 2016; Barocas et al., 2023). Conditional two-sample testing also finds applications beyond machine learning. In genomics, for example, scientists seek to identify differences in genetic distributions conditional on various factors such as disease status and environmental exposures (Virolainen et al., 2022; Wu et al., 2023). This methodology aids scientists in understanding the genetic basis of diseases and in developing strategies for personalized medicine by providing a rigorous framework for comparing conditional distributions.

1.1 Problem Setup

With the practical motivation in mind, we now formally set up the problem. Given $n_1, n_2 \in \mathbb{N}$, suppose we observe two mutually independent samples

$$\{(X_i^{(1)}, Y_i^{(1)})\}_{i=1}^{n_1} \overset{\text{i.i.d.}}{\sim} P_{XY}^{(1)} \quad \text{and} \quad \{(X_i^{(2)}, Y_i^{(2)})\}_{i=1}^{n_2} \overset{\text{i.i.d.}}{\sim} P_{XY}^{(2)},$$

where $P_{XY}^{(1)}$ and $P_{XY}^{(2)}$ are joint distributions supported on some generic product space $\mathcal{X} \times \mathcal{Y}$. Let $P_{Y|X}^{(1)}$ and $P_{Y|X}^{(2)}$ denote the conditional distributions of $Y^{(1)} \,|\, X^{(1)}$ and $Y^{(2)} \,|\, X^{(2)}$, respectively. Similarly, let $P_X^{(1)}$ and $P_X^{(2)}$ denote the marginal distributions of $X^{(1)}$ and $X^{(2)}$, respectively. Given these two samples, our goal is to test the equality of two conditional distributions

$$H_0 : P_X^{(1)}\{P_{Y|X}^{(1)}(\cdot \,|\, X) = P_{Y|X}^{(2)}(\cdot \,|\, X)\} = 1 \quad \text{versus} \quad H_1 : P_X^{(1)}\{P_{Y|X}^{(1)}(\cdot \,|\, X) \neq P_{Y|X}^{(2)}(\cdot \,|\, X)\} > 0, \tag{1}$$

where $P_X^{(j)}(\cdot \,|\, x)$ denotes the conditional distribution of $Y^{(j)}$ given $X^{(j)} = x$ for $j \in \{1, 2\}$. In other words, we are interested in determining whether two populations have the same distribution after controlling for potential confounding variables. Throughout this paper, we assume that $P_X^{(1)}$ and $P_X^{(2)}$ have the same support, satisfying $P_X^{(1)} \ll P_X^{(2)}$ and $P_X^{(2)} \ll P_X^{(1)}$ where the symbol $\ll$ denotes absolute continuity. Since $P_X^{(1)}$ and $P_X^{(2)}$ have the same support, the above hypotheses (1) for conditional two-sample testing can be equivalently defined using $P_X^{(2)}$ instead of $P_X^{(1)}$.

As pointed out by Boeken and Mooij (2021) and Yan and Zhang (2022), conditional two-sample testing is closely connected to conditional independence testing. To illustrate this connection, we introduce a binary variable $Z \in \{1, 2\}$, and see that the conditional independence between Y and Z given X is equivalently expressed as

$$Y \perp\!\!\!\perp Z \,|\, X \iff (Y \,|\, X, Z = 1) \overset{d}{=} (Y \,|\, X, Z = 2), \tag{2}$$

where the symbol $\overset{d}{=}$ denotes equality in distribution. This equivalence enables us to convert the problem of conditional two-sample testing to that of conditional independence testing based on the datasets $\{(Y_i, X_i) : Z_i = 1\}$ and $\{(Y_i, X_i) : Z_i = 2\}$. Consequently, we can leverage various existing methods for conditional independence testing to tackle conditional two-sample testing. However, prior work has not rigorously explored this approach, and indeed Yan and Zhang (2022) claim that it is not a sensible approach as the variable Z in the conditional two-sample problem is deterministic. Specifically, letting $n = n_1 + n_2$, $\sum_{i=1}^n \mathbb{1}(Z_i = 1)$ and $\sum_{i=1}^n \mathbb{1}(Z_i = 2)$ correspond to the sample sizes for two populations (i.e., n_1 and n_2), which are fixed in advance for the conditional two-sample problem. Therefore, a gap remains in rigorously connecting these seemingly similar, yet distinct, problems.

1.2 An Overview of Our Results

In this work, we make several contributions to the field of conditional two-sample testing. First, we reaffirm that comparing conditional distributions is intrinsically more difficult than comparing marginal distributions. For marginal two-sample testing, one can design permutation tests that control the type I error, while being powerful against certain alternatives (e.g., Kim et al., 2022a). However, we show that this is not the case for conditional two-sample testing. Our result (Theorem 1) proves that any valid conditional two-sample test has power at most equal to its size against any single alternative if the type of a conditional random vector is continuous. This is reminiscent of the negative result for conditional independence testing proved in Shah and Peters (2020). It is worth highlighting, however, that their negative result does not directly imply our Theorem 1. The proof of Shah and Peters (2020) relies on the assumption that the data $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ are i.i.d., which does not hold in our setup as $\sum_{i=1}^n \mathbb{1}(Z_i = 1)$ and $\sum_{i=1}^n \mathbb{1}(Z_i = 2)$ are deterministic numbers. We handle this distinction through a concentration argument and show that conditional two-sample testing is as difficult as conditional independence testing. This negative result naturally motivates additional assumptions that make the problem feasible.

Our next contribution is to introduce two general frameworks for conditional two-sample testing. The first framework effectively addresses the issue pointed out by Yan and Zhang (2022). In particular, we develop a generic method that converts any conditional independence test into a conditional two-sample test. This general method directly transfers the asymptotic properties of a conditional independence test computed using $\{(X_i, Y_i, Z_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XYZ}$ to the setting of conditional two-sample testing (Theorem 2). At the heart of this approach is the concentration property of a Binomial random variable to its mean, which facilitates the effective construction of i.i.d. samples drawn from P_{XYZ} (see Algorithm 1). This development paves way to leverage any existing methods for conditional independence testing in the literature, thereby expanding the range of tools available to practitioners for conducting two-sample tests.

The second framework that we introduce is based on density ratio estimation. To elaborate, let us assume that $P_X^{(1)}$ and $P_X^{(2)}$ have density functions $f_X^{(1)}$ and $f_X^{(2)}$ with respect to some base measure, and similarly $P_{Y|X}^{(1)}(\cdot \,|\, x)$ and $P_{Y|X}^{(2)}(\cdot \,|\, x)$ have density functions $f_{Y|X}^{(1)}(\cdot \,|\, x)$ and $f_{Y|X}^{(2)}(\cdot \,|\, x)$, respectively. Then for all $x, y \in \mathcal{X} \times \mathcal{Y}$, we have the identity:

$$f_{Y|X}^{(1)}(y \,|\, x) = f_{Y|X}^{(2)}(y \,|\, x) \iff f_{YX}^{(1)}(y, x) = \frac{f_X^{(1)}(x)}{f_X^{(2)}(x)} f_{YX}^{(2)}(y, x) \coloneqq f_{YX}(y, x), \tag{3}$$

where $f_{YX}^{(1)}$ is the joint density function of $(Y^{(1)}, X^{(1)})$ such that $f_{YX}^{(1)}(y, x) = f_{Y|X}^{(1)}(y \,|\, x) f_X^{(1)}(x)$, and $f_{YX}^{(2)}$ is similarly defined for $(Y^{(2)}, X^{(2)})$. The above equivalence (3) allows us to transform the problem of testing for conditional distributions into the one that compares marginal distributions with densities $f_{YX}^{(1)}$ and f_{YX}. The latter problem has been extensively studied with various methods, ranging from classical approaches such as Hotelling's test to modern methods such as kernel maximum mean discrepancy (Gretton et al., 2012; Liu et al., 2020; Schrab et al., 2023) and machine learning-based approaches (e.g., Lopez-Paz and Oquab, 2017; Kim et al., 2019, 2021; Hediger et al., 2022). The issue, however, is that we do not observe samples from f_{YX} but from $f_{YX}^{(2)}$. Therefore, the success of this framework relies on how accurate one can estimate the density ratio

$$r_X(x) \coloneqq \frac{f_X^{(1)}(x)}{f_X^{(2)}(x)}, \tag{4}$$

and incorporate it into the procedure to fill the gap between f_{YX} and $f_{YX}^{(2)}$. We demonstrate this methodology focusing on a classification-based test in Section 4.1 and a kernel-based test in Section 4.2.

1.3 Literature Review

As mentioned earlier, conditional two-sample testing has a wide range of applications in various fields, including machine learning, genetics and economics, where it is important to compare two samples controlling for confounding variables. Despite its broad range of applications and significance, there has been limited research dedicated to tackling this fundamental problem. Similar problems, on the other hand, have been explored in the literature such as testing for the equality of conditional moments (Hall and Hart, 1990; Kulasekera, 1995; Kulasekera and Wang, 1997; Fan and Lin, 1998; Neumeyer and Dette, 2003; Pardo-Fernández et al., 2015) and goodness-of-fit testing for pre-specified conditional distributions (Andrews, 1997; Zheng, 2000; Fan et al., 2006). These methods aim to facilitate the comparison of specific aspects of a distribution such as the conditional mean or second moments, rather than the entire distribution. Our research, however, is centered on nonparametric comparisons of two conditional distributions. This approach is of great importance as it enables a more comprehensive comparison of distributions, capturing differences that may not be evident when only specific moments or pre-specified models are compared.

It is only in recent years that conditional two-sample testing has gained attention, with several novel methods being proposed. Yan and Zhang (2022), for instance, proposed a method that extends unconditional energy distance to its conditional counterpart. They demonstrated that many key properties of the unconditional energy distance are retained in the conditional version. Moreover, they proposed a bootstrap procedure to calibrate their test statistic. To the best of our knowledge, however, the validity of their test remains unexplored, and the $O(n^4)$ time complexity of their algorithm poses a bottleneck to its practical application.

As another example, Hu and Lei (2024) built on the idea of conformal prediction and introduced a nonparametric conditional two-sample test using a weighted rank-sum statistic. This approach involves estimating both marginal and conditional density ratios, and the validity of their method depends on the quality of these ratio estimators. As explained in Example 4, their test statistic can be viewed as an example of our general framework based on density ratio estimation. A more recent work by Chen and Lei (2024) extended the idea of Hu and Lei (2024), leveraging Neyman orthogonality to reduce the first-order bias for the asymptotic normality. As another closely related work, Chatterjee et al. (2024) introduced a kernel-based conditional two-sample test using nearest neighbors. They considered the setting where a random sample $\{(X_i, Y_i^{(1)}, Y_i^{(2)})\}_{i=1}^n$ is generated from a joint distribution, i.e., the response variables $Y^{(1)}$ and $Y^{(2)}$ are conditioned on the same set of covariates X. This setting is notably different from that considered in the prior work (Yan and Zhang, 2022; Hu and Lei, 2024; Chen and Lei, 2024) as well as in our study, which consider potentially different covariates. Hence, the methods proposed by Chatterjee et al. (2024) are not directly comparable to ours.

As explained before, the first framework that we propose can be constructed based on essentially any conditional independence tests from the literature. The problem of testing for conditional independence has been extensively studied, resulting in a variety of methods to handle different scenarios and challenges. Shah and Peters (2020) proposed the Generalized Covariance Measure (GCM) whose validity depends on the performance of regression methods. Recent improvements to this method include the strategies such as weighting (Scheidegger et al., 2022) and applying GCM to a projected random vector (Lundborg et al., 2022; Chakraborty et al., 2024). Other notable methodologies for conditional independence testing include kernel-based tests (Zhang et al., 2011; Doran et al., 2014; Strobl et al., 2019; Pogodin et al., 2024), binning-based tests (Neykov et al., 2021; Kim et al., 2022b; Neykov et al., 2023), regression-based tests (Dai et al., 2022; Williamson et al., 2023) and tests under the model-X framework (Candes et al., 2018; Berrett et al., 2020; Liu et al., 2022; Tansey et al., 2022). Our method can leverage these developments to effectively solve the problem of conditional two-sample testing.

Our second framework can benefit from extensive research done on density ratio estimation in the literature. A straightforward way of estimating density ratio is to first estimate individual density functions, and take their ratio as an estimate. However, this method tends to become unstable, especially in high-dimensional settings. To overcome this issue, Sugiyama et al. (2007b) and Tsuboi et al. (2009) developed methods that directly estimate density ratio without involving density estimation. Kanamori et al. (2010) compared different methods of density ratio estimation, and discussed their theoretical properties. Kanamori et al. (2009) reformulated the problem as a least-squares problem to provide a closed-form solution, whereas Liu et al. (2017) proposed trimmed density ratio estimation to improve stability and robustness by trimming extreme values. More recent advancements in density ratio estimation include Choi et al. (2021); Rhodes et al. (2020); Choi et al. (2022). As explained in Section 4, our approach uses density ratio estimation to deal with discrepancies between $f_{YX}^{(1)}$ and f_{YX}, and transforms the problem of comparing conditional distributions into that of comparing marginal distributions.

1.4 Organization

The rest of this paper is organized as follows. We begin with a hardness result for conditional two-sample testing in Section 2, which shows that no test can have power greater than its size against any alternative without additional assumptions. Section 3 presents our framework that converts tests for conditional independence into those for the equality of conditional distributions. Section 4 introduces another framework based on density ratio estimation. Numerical results illustrating the finite-sample performance of our methods are presented in Section 5, followed by the conclusion in Section 6. The proofs of the results omitted in the main text can be found in the appendix.

2 Hardness Result

Before introducing our frameworks, we present a fundamental hardness result for conditional two-sample testing. Specifically, for a continuous random vector X, our result demonstrates that any valid conditional two-sample test has no power against any alternative. This finding parallels the negative result established by Shah and Peters (2020) for conditional independence testing, and our proof builds crucially on their work. Given the connection established in (2), one might argue that their negative result directly applies to the two-sample problem. However, additional effort is required to make this connection concrete since the sample sizes n_1 and n_2 are deterministic in our setting, which violates the i.i.d. assumption required in Shah and Peters (2020).

To state the result, let $\mathcal{E}$ denote the set of all pairs of distributions $(P_{XY}^{(1)}, P_{XY}^{(2)})$ defined on a generic product space $\mathcal{X} \times \mathcal{Y}$. For each $j \in \{1, 2\}$, assume that the marginals $P_X^{(j)}$ and $P_Y^{(j)}$ are absolutely continuous with respect to the Lebesgue measures on $\mathbb{R}^{d_X}$ and $\mathbb{R}^{d_Y}$, respectively. Let $\mathcal{P}_0 \subset \mathcal{E}$ denote the set of distribution pairs satisfying the null hypothesis H_0 in (1), and let $\mathcal{P}_1 = \mathcal{E} \setminus \mathcal{P}_0$ denote the corresponding alternative class. For $M \in (0, \infty]$, let $\mathcal{E}_M \subseteq \mathcal{E}$ be the subset of pairs with support contained strictly with in an ℓ_∞ ball of radius M. We then define $\mathcal{P}_{0,M} = \mathcal{P}_0 \cap \mathcal{E}_M$ and $\mathcal{P}_{1,M} = \mathcal{P}_1 \cap \mathcal{E}_M$. The following theorem establishes that no valid test ϕ for conditional two-sample testing can attain power exceeding its size.

Theorem 1. *Let $n_1, n_2 \in \mathbb{N}$, $\alpha \in (0, 1)$ and $M \in (0, \infty]$. Suppose we observe two independent samples*

$$\{(X_i^{(1)}, Y_i^{(1)})\}_{i=1}^{n_1} \overset{\text{i.i.d.}}{\sim} P_{XY}^{(1)} \quad \text{and} \quad \{(X_i^{(2)}, Y_i^{(2)})\}_{i=1}^{n_2} \overset{\text{i.i.d.}}{\sim} P_{XY}^{(2)},$$

where $(P_{XY}^{(1)}, P_{XY}^{(2)}) =: P \in \mathcal{E}_M$. Consider a test $\phi : \{(X_i^{(1)}, Y_i^{(1)})\}_{i=1}^{n_1} \cup \{(X_i^{(2)}, Y_i^{(2)})\}_{i=1}^{n_2} \to \{0, 1\}$. Suppose ϕ controls the type I error at level α, i.e.,

$$\sup_{P \in \mathcal{P}_{0,M}} \mathbb{E}_P[\phi] \leq \alpha.$$

Then the power of ϕ is at most α for any $P \in \mathcal{P}_{1,M}$, that is,

$$\sup_{P \in \mathcal{P}_{1,M}} \mathbb{E}_P[\phi] \leq \alpha.$$

Remark. We note that Theorem 1 only focuses on non-randomized tests for simplicity of presentation, but our proof also holds for randomized tests. The proof of Theorem 1 is provided below.

Proof of Theorem 1. Proving Theorem 1 directly is challenging, so we instead rely on the known hardness result for conditional independence testing established by Shah and Peters (2020). To connect the two settings, we embed the conditional two-sample problem into the conditional independence framework by introducing a group indicator $Z \in \{1, 2\}$, as in (2). Here, Z simply indicates which sample an observation comes from, thereby allowing the two conditional distributions to be expressed within a single joint distribution over (X, Y, Z). To make this precise, suppose the two samples have deterministic sizes n_1 and n_2, and write $n := n_1 + n_2$. Formally, the joint distribution of (X, Y, Z) is defined as

$$P_{XYZ} := \tfrac{n_1}{n}, P_{XY}^{(1)} \otimes \delta_{Z=1} + \tfrac{n_2}{n}, P_{XY}^{(2)} \otimes \delta_{Z=2}, \tag{5}$$

where $\delta_{Z=z}$ denotes the Dirac measure at $Z = z$, so that (X, Y) follows $P_{XY}^{(1)}$ when $Z = 1$ and $P_{XY}^{(2)}$ when $Z = 2$. Given a two-sample test

$$\phi_{\text{two}} : \{(X_i^{(1)}, Y_i^{(1)})\}_{i=1}^{n_1} \cup \{(X_i^{(2)}, Y_i^{(2)})\}_{i=1}^{n_2} \to \{0, 1\},$$

we define the associated conditional independence test by

$$\phi_{\text{indep}}\big(\{(X_i, Y_i, Z_i)\}_{i=1}^{n}\big) := \phi_{\text{two}}\big(\{(X_i, Y_i) : Z_i = 1\} \cup \{(X_i, Y_i) : Z_i = 2\}\big),$$

where $\{(X_i, Y_i, Z_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XYZ}$. Under this representation, the two-sample null and alternative classes $\mathcal{P}_{0,M}$ and $\mathcal{P}_{1,M}$ coincide with the conditional independence classes defined by $Y \perp\!\!\!\perp Z \mid X$ and $Y \not\perp\!\!\!\perp Z \mid X$, respectively. The main subtlety is that in the conditional two-sample setting the group sizes $N_1 = \sum_{i=1}^n \mathbb{1}(Z_i = 1)$ and $N_2 = \sum_{i=1}^n \mathbb{1}(Z_i = 2)$ are fixed, whereas the hardness result of Shah and Peters (2020) assumes i.i.d. draws from P_{XYZ}. Thus, the Z_i's are not i.i.d., and a direct application of their result is not immediate. Nevertheless, conditioning on $N_1 = n_1$ and $N_2 = n_2$, the type I error and power of ϕ_{indep} coincide exactly with those of ϕ_{two}, so the reduction remains valid. Moreover, following Neykov et al. (2021), the specific marginal distribution of Z (e.g., $\mathbb{P}(Z = 1) = n_1/n$) does not affect the hardness argument that is the essential difficulty stems from the continuous nature of the conditioning variable X. A detailed justification of this claim is provided in Lemma 8. For a constant $N \in \mathbb{N}$ greater than $2n$, we work with N i.i.d. random samples $\{(X_i, Y_i, Z_i)\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} P$ and define $N_1' = \sum_{i=1}^N \mathbb{1}(Z_i = 1)$ and $N_2' = \sum_{i=1}^N \mathbb{1}(Z_i = 2)$, which follow $N_1' \sim \text{Binomial}(N, \lambda_n)$ and $N_2' \sim \text{Binomial}(N, 1 - \lambda_n)$, respectively. For $n_1, n_2 \in \mathbb{N}$ given in the theorem statement, define a good event $\mathcal{A} \coloneqq \{N_1' \geq n_1, N_2' \geq n_2\}$, whose probability satisfies $\mathbb{P}(\mathcal{A}) \geq 1 - \mathbb{P}(N_1' < n_1) - \mathbb{P}(N_2' < n_2)$ by the union bound. Since $N \geq 2n$, we can ensure that $n_1 - N\lambda_n \leq -\frac{1}{2}N\lambda_n$ and thus

$$
\begin{aligned}
\mathbb{P}(N_1' < n_1) &= \mathbb{P}(N_1' - N\lambda_n < n_1 - N\lambda_n) \leq \mathbb{P}\Big(N_1' - N\lambda_n < -\frac{1}{2}N\lambda_n\Big) \\
&\leq \mathbb{P}\Big(\big|N_1' - N\lambda_n\big| > \frac{1}{2}N\lambda_n\Big) \leq \frac{4(1-\lambda_n)}{N\lambda_n},
\end{aligned}
$$

where the last inequality uses Chebyshev's inequality along with $N_1' \sim \text{Binomial}(N, \lambda_n)$. We can similarly obtain that $\mathbb{P}(N_2' < n_2) \leq \frac{4\lambda_n}{N(1-\lambda_n)}$. Therefore, the probability of the good event $\mathcal{A}$ is lower bounded as

$$
\mathbb{P}(\mathcal{A}) \geq 1 - \frac{4(1-\lambda_n)^2 + 4\lambda_n^2}{N\lambda_n(1-\lambda_n)} \overset{\text{set}}{=} 1 - \varepsilon_N, \tag{6}
$$

where $\varepsilon_N \to 0$ as $N \to \infty$.

We now consider a test ϕ_{indep} that only uses $n_1 + n_2$ data points out of N samples. Importantly, this sample consists of n_1 observations from $\{(Y_i, X_i) : Z_i = 1\}$ and n_2 observations from $\{(Y_i, X_i) : Z_i = 2\}$, whenever $\mathbb{1}(\mathcal{A}) = 1$. If $\mathbb{1}(\mathcal{A}) = 0$, this test simply returns 0 (i.e., accept H_0). Moreover, we assume that this test satisfies $\sup_{P \in \mathcal{P}_{0,M}} \mathbb{E}_P[\phi_{\text{indep}} \mid \mathcal{A}] \leq \alpha$. In fact, since ϕ_{indep} only uses n_1 observations with $Z = 1$ and n_2 observations with $Z = 2$, the previous inequality implies $\sup_{P \in \mathcal{P}_{0,M}} \mathbb{E}_P\left[\phi_{\text{indep}} \mid N_1 = n_1, N_2 = n_2\right] = \sup_{P \in \mathcal{P}_{0,M}} \mathbb{E}_P\left[\phi_{\text{two}}\right] \leq \alpha$, i.e., ϕ_{two} is a valid level α test for conditional two-sample testing. Based on the previous results, the type I error of ϕ_{indep} constructed based on $\{(X_i, Y_i, Z_i)\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} P$ fulfills

$$
\sup_{P \in \mathcal{P}_{0,M}} \mathbb{E}_P[\phi_{\text{indep}}] = \sup_{P \in \mathcal{P}_{0,M}} \mathbb{E}_P[\phi_{\text{indep}} \mathbb{1}(\mathcal{A})] \leq \sup_{P \in \mathcal{P}_{0,M}} \mathbb{E}_P[\phi_{\text{indep}} | \mathcal{A}] \leq \alpha.
$$

This implies that ϕ is a valid test for conditional independence with size α. Therefore, for any $P \in \mathcal{P}_{1,M}$,

$$
\begin{aligned}
&\mathbb{E}_P[\phi_{\text{indep}} \mid \mathcal{A}]\,(1 - \varepsilon_N) \overset{\text{(i)}}{\leq} \mathbb{E}_P[\phi_{\text{indep}} \mid \mathcal{A}]\mathbb{E}_P[\mathbb{1}(\mathcal{A})] \leq \mathbb{E}_P[\phi_{\text{indep}}] \overset{\text{(ii)}}{\leq} \alpha \\
\Longleftrightarrow\; &\mathbb{E}_P[\phi_{\text{indep}} \mid \mathcal{A}] \overset{\text{(iii)}}{=} \mathbb{E}_P\left[\phi_{\text{two}}\right] \leq \frac{\alpha}{1 - \varepsilon_N},
\end{aligned}
$$

where step (i) uses the inequality in 6, step (ii) holds by (Shah and Peters, 2020, Theorem 2 and Remark 4), and step (iii) uses the fact that ϕ_{indep} only uses $n_1 + n_2$ observations as described before and $(\phi_{\text{indep}} \mid N_1 = n_1,\, N_2 = n_2) = \phi_{\text{two}}$. Since N was an arbitrary number greater than or equal to $2n$ and $\varepsilon_N \to 0$ as $N \to \infty$, we can conclude that $\mathbb{E}_P\left[\phi_{\text{two}}\right] \leq \alpha$ for any $P \in \mathcal{P}_{1,M}$. ☐

Theorem 1 clearly explains that it is necessary to impose additional assumptions (e.g., smoothness for distributions) in order to make the conditional two-sample problem feasible. In the next two sections, we explore two general frameworks, which implicitly or explicitly incorporate reasonable assumptions to address

Algorithm 1 Converting a Conditional Independence Test into a Conditional Two-Sample Test

Require: Data $\{(Y_i^{(1)}, X_i^{(1)})\}_{i=1}^{n_1}$ and $\{(Y_i^{(2)}, X_i^{(2)})\}_{i=1}^{n_2}$, a conditional independence test ϕ for $H_0 : Y \perp\!\!\!\perp Z \mid X$ of (asymptotic) size $\alpha \in (0,1)$, and an adjustment parameter $\varepsilon \in (0,1)$

Let $n = n_1 + n_2$. Draw $\tilde{n}_1 \sim \text{Binomial}(\tilde{n}, n_1/n)$ where $\tilde{n} = k^* n$ and $k^* = 1 - 3\log(\varepsilon)/(2n_1) - \sqrt{(1 - 3\log(\varepsilon)/(2n_1))^2 - 1}$. Set $\tilde{n}_2 = \tilde{n} - \tilde{n}_1$.
if $\tilde{n}_1 > n_1$ or $\tilde{n}_2 > n_2$ **then** Accept H_0.
else
    Merge $\{(X_i^{(1)}, Y_i^{(1)}, Z_i = 1)\}_{i=1}^{\tilde{n}_1}$ and $\{(X_i^{(2)}, Y_i^{(2)}, Z_i = 2)\}_{i=1}^{\tilde{n}_2}$, yielding $\mathcal{D}_{\tilde{n}} := \{(X_i, Y_i, Z_i)\}_{i=1}^{\tilde{n}}$.
    Run a conditional independence test ϕ on $\mathcal{D}_{\tilde{n}}$ at level α, and denote the resulting test as $\phi_{\tilde{n}}$.
    if $\phi_{\tilde{n}} = 1$ **then** Reject H_0 **else** Accept H_0.
    end if
end if

this problem. The first framework utilizes any conditional independence test and considers scenarios where this test performs well for verifying conditional independence. Conversely, the second framework assumes that the marginal density ratio r_X is well-behaved and can be estimated with high accuracy.

3 Approach via Conditional Independence Testing

In this section, we introduce our first framework that converts a conditional independence test to a conditional two-sample test, while maintaining the same asymptotic guarantees. The key idea is to construct a dataset $\mathcal{D}_{\tilde{n}}$ consisting of i.i.d. random vectors (Y, Z, X) of size $\tilde{n}$ based on the given two samples $\{(Y_i^{(1)}, X_i^{(1)})\}_{i=1}^{n_1}$ and $\{(Y_i^{(2)}, X_i^{(2)})\}_{i=1}^{n_2}$. To achieve this, letting $n = n_1 + n_2$, we first draw a random variable $\tilde{n}_1$ from Binomial($\tilde{n}$, n_1/n) where $\tilde{n}$ is set to be smaller than n and $\tilde{n}/n \to 1$. Since a Binomial random variable is highly concentrated around its mean, we can guarantee that $\tilde{n}_1 \leq n_1$ and $\tilde{n}_2 := \tilde{n} - \tilde{n}_1 \leq n_2$ with high probability. If a bad event happens where either $\tilde{n}_1 > n_1$ or $\tilde{n}_2 > n_2$, making the construction of $\mathcal{D}_{\tilde{n}}$ infeasible, we simply accept the null hypothesis. This slightly inflates the type II error in finite-sample scenarios, but it is asymptotically negligible. A similar idea has been utilized in Neykov et al. (2021) in a different context to eliminate Poissonization for conditional independence testing.

Having constructed $\mathcal{D}_{\tilde{n}}$ consisting of i.i.d. random samples drawn from the joint distribution of (Y, Z, X), we can now implement a conditional independence test based on $\mathcal{D}_{\tilde{n}}$, while retaining the same theoretical guarantees for conditional two-sample testing. Algorithm 1 summarizes this procedure, and the following theorem formally establishes its theoretical guarantees.

Theorem 2. *Let $n_1, n_2 \in \mathbb{N}$ with $n = n_1 + n_2$. Suppose we observe two independent samples*

$$\{(X_i^{(1)}, Y_i^{(1)})\}_{i=1}^{n_1} \overset{\text{i.i.d.}}{\sim} P_{XY}^{(1)} \quad \textit{and} \quad \{(X_i^{(2)}, Y_i^{(2)})\}_{i=1}^{n_2} \overset{\text{i.i.d.}}{\sim} P_{XY}^{(2)},$$

where $(P_{XY}^{(1)}, P_{XY}^{(2)}) \in \mathcal{P}$ for some class $\mathcal{P}$ of distribution pairs. Let $\mathcal{P}_0 \subset \mathcal{P}$ and $\mathcal{P}_1 \subset \mathcal{P}$ denote disjoint subclasses corresponding to the null and alternative distribution classes, respectively. Construct the joint distribution P_{XYZ} from $(P_{XY}^{(1)}, P_{XY}^{(2)})$ as in Equation (5). For $\{(X_i, Y_i, Z_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} P_{XYZ}$, consider a conditional independence test $\phi : \{(X_i, Y_i, Z_i)\}_{i=1}^{n} \to \{0,1\}$ such that for any $\alpha \in (0,1)$,

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} \mathbb{E}_P[\phi] \leq \alpha, \qquad \lim_{n\to\infty} \sup_{P\in\mathcal{P}_1} \mathbb{E}_P[1-\phi] = 0.$$

Let $\tilde{\phi}$ denote the test obtained by applying Algorithm 1 with adjustment parameter $\varepsilon = o(1)$, where $\tilde{\phi} = 1$ if and only if H_0 is rejected. Then $\tilde{\phi}$ preserves the same asymptotic guarantees:

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} \mathbb{E}_P[\tilde{\phi}] \leq \alpha, \qquad \lim_{n\to\infty} \sup_{P\in\mathcal{P}_1} \mathbb{E}_P[1-\tilde{\phi}] = 0.$$

Remark. The guarantees of Theorem 2 should be interpreted as asymptotic. Our analysis relies on approximating the test statistic via Algorithm 1, which leaves an $o(1)$ remainder. In addition, the adjustment parameter ε contributes explicit additive errors to the bounds for both type I and type II errors. Both effects vanish only in the limit as $n \to \infty$ and $\varepsilon = \varepsilon_n \to 0$, so exact finite-sample guarantees cannot be provided.

Proof of Theorem 2. We may write $\tilde{\phi} = \mathbb{1}(\tilde{n}_1 \leq n_1)\mathbb{1}(\tilde{n}_2 \leq n_2)\phi_{\tilde{n}}$, where $\phi_{\tilde{n}}$ is defined as ϕ based on $\mathcal{D}_{\tilde{n}} = \{(X_i, Y_i, Z_i)\}_{i=1}^{\tilde{n}}$ in Algorithm 1. For convenience, we denote $P := P_{XYZ}$. Now generate new i.i.d. samples $\tilde{\mathcal{D}}_{\tilde{n}} = \{(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i)\}_{i=1}^{\tilde{n}} \overset{\text{i.i.d.}}{\sim} P$ independent of $\mathcal{D}_{\tilde{n}}$ and define as

$$\tilde{\phi}^\dagger = \mathbb{1}\Bigg\{\sum_{i=1}^{\tilde{n}} \mathbb{1}(\tilde{Z}_i = 1) \leq n_1\Bigg\}\mathbb{1}\Bigg\{\sum_{i=1}^{\tilde{n}} \mathbb{1}(\tilde{Z}_i = 2) \leq n_2\Bigg\}\phi_{\tilde{n}}^\dagger,$$

where $\phi_{\tilde{n}}^\dagger$ is defined as ϕ but based on $\tilde{\mathcal{D}}_{\tilde{n}}$. Although the distributions of $\tilde{\phi}$ (from subsampling the original dataset) and $\tilde{\phi}^\dagger$ (from new i.i.d. samples) are not identical, the difference vanishes as $\tilde{n}/n \to 1$. Hence, for every P,

$$\mathbb{E}_P[\tilde{\phi}] = \mathbb{E}_P[\tilde{\phi}^\dagger] + o(1).$$

Moreover, by construction $\mathbb{E}_P[\tilde{\phi}^\dagger] \leq \mathbb{E}_P[\phi_{\tilde{n}}^\dagger]$ holds. Therefore

$$\mathbb{E}_P[\tilde{\phi}] \leq \mathbb{E}_P[\phi_{\tilde{n}}^\dagger] + o(1),$$

for all P. In other words, we can analyze $\tilde{\phi}$ by working with $\tilde{\phi}^\dagger$ up to an asymptotically negligible $o(1)$ term, which suffices for type I error control. Moving to the type II error, observe that

$$\begin{aligned}\mathbb{1}(\tilde{n}_1 \leq n_1)\mathbb{1}(\tilde{n}_2 \leq n_2)\phi_{\tilde{n}} &= \phi_{\tilde{n}} - \mathbb{1}(\tilde{n}_1 > n_1 \text{ or } \tilde{n}_2 > n_2)\phi_{\tilde{n}} \\ &\geq \phi_{\tilde{n}} - \mathbb{1}(\tilde{n}_1 > n_1) - \mathbb{1}(\tilde{n}_2 > n_2),\end{aligned}$$

by the union bound, which yields

$$\mathbb{E}_P[1 - \tilde{\phi}] \leq \mathbb{E}_P[1 - \phi_{\tilde{n}}^\dagger] + \mathbb{E}_P[\mathbb{1}(\tilde{n}_1 > n_1)] + \mathbb{E}_P[\mathbb{1}(\tilde{n}_2 > n_2)] + o(1). \tag{7}$$

In addition, letting $p = n_1/n$ in Lemma 1 of Appendix B, take $(1+\delta)\tilde{n}n_1/n = k(1+\delta)n_1 = n_1$, which gives $1 + \delta = k^{-1} \iff \delta = k^{-1} - 1$. Thus, by Lemma 1, we have

$$\mathbb{E}_P[\mathbb{1}(\tilde{n}_1 > n_1)] \leq \exp\bigg(-\frac{n_1 k(k^{-1} - 1)^2}{3}\bigg).$$

Letting the right-hand side equal ε and solving for $k \in (0,1)$, we derive the form of k^* as presented in Algorithm 1, which shows that $\mathbb{E}_P[\mathbb{1}(\tilde{n}_1 > n_1)] \leq \varepsilon$. By symmetry, the same analysis holds for the inequality $\mathbb{E}_P[\mathbb{1}(\tilde{n}_2 > n_2)] \leq \varepsilon$. As a result, continuing from the inequality (7), we can upper bound the type II error of $\tilde{\phi}$ (the level-α test obtained from Algorithm 1) as

$$\mathbb{E}_P[1 - \tilde{\phi}] \leq \mathbb{E}_P[1 - \phi_{\tilde{n}}^\dagger] + 2\varepsilon + o(1).$$

Since $\varepsilon = o(1)$, the above display proves the second claim on type II error control. This completes the proof of Theorem 2. □

Our analysis in Theorem 2 is not limited to conditional two-sample testing, and it can be applied to marginal two-sample testing as well. Indeed, the problem of conditional two-sample testing becomes equivalent to the unconditional counterpart when X is degenerate (e.g., $X = 0$ with probability one). Thus, our algorithm serves as a generic method to convert unconditional independence tests to unconditional two-sample tests as well. We also mention that a specific form of k^* in Algorithm 1 is derived from the multiplicative Chernoff bound for a Binomial random variable (Lemma 1), which can be refined by numerically computing the tail probability of a Binomial random variable.

Despite its generality, one obvious drawback of Algorithm 1 is that it does not take the datasets $\{(Y_i^{(1)}, X_i^{(1)})\}_{i=\tilde{n}_1+1}^{n_1}$ and $\{(Y_i^{(2)}, X_i^{(2)})\}_{i=\tilde{n}_2+1}^{n_2}$ into account in the procedure when $\tilde{n}_1 < n_1$ and $\tilde{n}_2 < n_2$. It can be seen that the expected number of discarded samples, i.e., $\mathbb{E}[n - \tilde{n}_1 - \tilde{n}_2]$, is $O\big(\sqrt{n \log(1/\varepsilon)}\big)$. This loss might degrade the performance in small-sample size regimes, but it can be negligible when n is sufficiently large and ε decreases slowly (see Appendices C.4 and C.5 for empirical support). Nevertheless, when a test statistic is sufficiently stable (formally defined in Definition 1) the conclusion of Theorem 2 may hold without further modification of ϕ, meaning the conditional testing errors of ϕ are asymptotically equivalent to its marginal errors.

To illustrate this, we build upon the coupling argument presented by Chung and Romano (2013). First, draw $\bar{n}_1 \sim \text{Binomial}(n, n_1/n)$ and set $\bar{n}_2 = n - \bar{n}_1$. If $\bar{n}_1 > n_1$, draw $\bar{n}_1 - n_1$ additional samples $\{(Y_i^{(1)}, X_i^{(1)})\}_{i=n_1+1}^{\bar{n}_1}$ from $P_{XY}^{(1)}$. Otherwise, draw $\bar{n}_2 - n_2$ additional samples $\{(Y_i^{(2)}, X_i^{(2)})\}_{i=n_2+1}^{\bar{n}_2}$ from $P_{XY}^{(2)}$. In either case, set $\{(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i)\}_{i=1}^n = \{X_i^{(1)}, Y_i^{(1)}, 1\}_{i=1}^{\bar{n}_1} \cup \{X_i^{(2)}, Y_i^{(2)}, 2\}_{i=1}^{\bar{n}_2}$, which can be viewed as i.i.d. draws from a joint distribution of (X, Y, Z), after randomly permuting indices. When this newly constructed dataset is compared with the original dataset $\{(X_i, Y_i, Z_i)\}_{i=1}^n = \{X_i^{(1)}, Y_i^{(1)}, 1\}_{i=1}^{n_1} \cup \{X_i^{(2)}, Y_i^{(2)}, 2\}_{i=1}^{n_2}$, there are $|\bar{n}_1 - n_1|$ distinct data points with the expectation $\mathbb{E}[|\bar{n}_1 - n_1|] \leq \sqrt{n/4}$.

Definition 1 (Stability of a test statistic). *Let T_n be a test statistic computed on the original dataset $\{(X_i, Y_i, Z_i)\}_{i=1}^n$, and let $\tilde{T}_n$ denote its analogue computed on the perturbed dataset $\{(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i)\}_{i=1}^n$ constructed by the above coupling scheme. We say that T_n is* (asymptotically) stable *if*

$$\lim_{n \to \infty} \mathbb{P}\big(|T_n - \tilde{T}_n| > \epsilon\big) = 0 \quad \textit{for every } \epsilon > 0.$$

In this case, T_n and $\tilde{T}_n$ are said to be asymptotically equivalent*.*

This suggests that if a test statistic T_n is asymptotically invariant to $\sqrt{n}$-data perturbations, its asymptotic behavior remains consistent across both the original and the newly constructed datasets. Formally, let $\tilde{T}_n$ denote the analogous test statistic computed on the perturbed dataset $\{(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i)\}_{i=1}^n$. However, whether a given T_n is stable must be assessed on a case-by-case basis. Below, we provide examples illustrating both stable and unstable cases.

Example 1 (Stable case)**.** To simplify our presentation, consider a univariate case of $Y \in \mathbb{R}$, and assume $f(x) \coloneqq \mathbb{E}[Y \,|\, X = x]$ and $g(x) \coloneqq \mathbb{E}[Z \,|\, X = x]$ are known. Letting $R_i \coloneqq \{Y_i - f(X_i)\}\{Z_i - g(X_i)\}$, the generalized covariance measure introduced by Shah and Peters (2020) is

$$T_n = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n R_i}{\big\{\frac{1}{n}\sum_{i=1}^n R_i^2 - \big(\frac{1}{n}\sum_{r=1}^n R_r\big)^2\big\}^{1/2}},$$

and let $\tilde{T}_n$ be similarly defined as T_n based on $\{(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i)\}_{i=1}^n$. Focusing on the numerators of T_n and $\tilde{T}_n$, it can be seen that their difference is

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n R_i - \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{R}_i = \frac{1}{\sqrt{n}}\sum_{i=n_1+1}^{\bar{n}_1} (R_i - \tilde{R}_i) \cdot \mathbb{1}(\bar{n}_1 > n_1) + \frac{1}{\sqrt{n}}\sum_{i=\bar{n}_1+1}^{n_1} (R_i - \tilde{R}_i) \cdot \mathbb{1}(\bar{n}_1 \leq n_1).$$

Under the null hypothesis, the expectation of the difference is zero and the variance is bounded above by $1/\sqrt{n}$ up to a constant, provided that each Y_i has a finite second moment. Therefore, the difference of the numerators is asymptotically negligible. We can show similarly that the difference of the denominators is also asymptotically negligible as detailed in Appendix A.9. Putting thing together concludes that T_n is stable in the sense of Definition 1.

In the above example, we assumed that the conditional expectations f and g are known. In practice, f and g are estimated from the data, and the generalized covariance measure can become highly unstable when the estimators of f and g are themselves unstable.

Example 2 (Unstable case)**.** Consider an extreme case where the estimators $\hat{f}$ and $\hat{g}$ are defined as follows: $\hat{f}(X_i) \coloneqq Y_i \mathbb{1}(\sum_{i=1}^n Z_i = n_1) + f(X_i)\mathbb{1}(\sum_{i=1}^n Z_i \neq n_1)$ and $\hat{g}(X_i) \coloneqq Z_i \mathbb{1}(\sum_{i=1}^n Z_i = n_1) + g(X_i)\mathbb{1}(\sum_{i=1}^n Z_i \neq n_1)$. In this case, T_n is not well-defined as it takes the form $0/0$ deterministically. On the other hand, when n_1 and n_2 are well-balanced (e.g., $n_1/n = 1/2$), the probability of the event $\sum_{i=1}^n \tilde{Z}_i = n_1$ converges to zero as n increases. Under such condition and assuming suitable moment conditions, the test statistic $\tilde{T}_n$ based on $\{(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i)\}_{i=1}^n$ can still converge to a Gaussian limit. This example illustrates that the limiting behavior of T and $\tilde{T}_n$ can differ significantly, which is attributed to the instability of the estimators $\hat{f}$ and $\hat{g}$. In fact, T_n is *not stable* in the sense of Definition 1.

The previous examples highlight the need for caution when converting a conditional independence test to a conditional two-sample test, and also justify our generic approach to converting a conditional independence test to a conditional two-sample test in Algorithm 1. The next section introduces another general framework for conditional two-sample testing based on density ratio estimation.

4 Approach via Density Ratio Estimation

In this section, we present our second framework, which transforms the problem of conditional two-sample testing into one that involves comparing marginal distributions via density ratio estimation. Concretely, we recall from (3) that the null hypothesis of equality of two conditional distributions holds if and only if $f_{YX}^{(1)} = f_{YX}$ where $f_{YX} = r_X \cdot f_{YX}^{(2)}$ and r_X is the density ratio defined in (4). A challenge when applying this approach is that we only have samples from $f_{YX}^{(2)}$, and not from f_{YX}, which makes it impossible to directly compare samples from $f_{YX}^{(1)}$ with those from f_{YX}. However, once the density ratio is known or accurately estimated, we can effectively correct the bias arising from the difference between $f_{YX}^{(2)}$ and f_{YX} in various test statistics, frequently used for marginal two-sample testing. To facilitate our discussion, we first assume that the density ratio r_X is known and provide a detailed analysis on how to deal with the unknown case by focusing on a few cases.

At the core of our idea is importance weighting (Kimura and Hino, 2024, for a survey), a technique that assigns different levels of importance to data points to correct biases and prioritize relevant data. For instance, suppose we would like to estimate the expectation of X under the distribution P with density p, while we only observe data $X_1, \dots, X_n$ from another distribution Q with density q. Then by re-weighting data points using the density ratio p/q, we can obtain an unbiased estimator of the expectation under P as

$$\frac{1}{n}\sum_{i=1}^n \frac{p(X_i)}{q(X_i)} X_i \quad \text{such that} \quad \mathbb{E}_Q\left[\frac{1}{n}\sum_{i=1}^n \frac{p(X_i)}{q(X_i)} X_i\right] = \mathbb{E}_P[X].$$

This idea can be applied to a range of marginal two-sample test statistics as we demonstrate below. Throughout this section, we use the shorthand $V^{(1)} \coloneqq (X^{(1)}, Y^{(1)})$ and $V^{(2)} \coloneqq (X^{(2)}, Y^{(2)})$ to simplify the notation.

Example 3. (Mean comparison) We start with a simple case of comparing the mean of transformed samples. Given a feature map $\psi : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$, one can consider

$$\frac{1}{n_1}\sum_{i=1}^{n_1} \psi\bigl(V_i^{(1)}\bigr) - \frac{1}{n_2}\sum_{i=1}^{n_2} r_X\bigl(X_i^{(2)}\bigr)\psi\bigl(V_i^{(2)}\bigr)$$

as a test statistic for the hypotheses in (1). The expectation of this statistic is equal to zero under the null hypothesis. Moreover, since the test statistic is simply a linear combination of independent random variables, it can be calibrated using the Gaussian approximation.

Example 4. (Rank sum statistic) Instead of comparing the mean, one can compare the stochastic order of two distributions using ranks. Specifically, given a feature map $\psi : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$, a rank sum statistic based on the transformed samples can be computed as

$$\frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} r_X\big(X_j^{(2)}\big) \mathbb{1}\{\psi\big(V_j^{(2)}\big) < \psi\big(V_i^{(1)}\big)\}.$$

Under the null hypothesis, and assuming no ties among transformed samples, it can be seen that the expectation is equal to 1/2. As in Hu and Lei (2024), the test statistic can be shown to be asymptotically Gaussian using the asymptotic theory of U-statistics under conditions. Therefore, the critical value can be determined based on this Gaussian approximation. The power, however, changes depending on ψ. Hu and Lei (2024) takes ψ as an estimate of $f^{(1)}_{Y|X}(\cdot\,|\,\cdot)/f^{(2)}_{Y|X}(\cdot\,|\,\cdot)$.

Example 5. (Classifier-based approach) Let $\mathcal{H}$ be a class of classifiers. Given a binary classifier $h : \mathcal{X} \times \mathcal{Y} \mapsto \{1, 2\}$ where $h \in \mathcal{H}$, $\ell : \mathbb{R} \times \{1, 2\} \mapsto \mathbb{R}$ is a loss function that measures the difference between the predicted value and the true output. The core idea behind classifier-based two-sample tests (Lopez-Paz and Oquab, 2017; Kim et al., 2021; Hediger et al., 2022) is that when the null hypothesis of equality of distributions is true, any classifier will return a random guess. On the other hand, when two distributions are significantly different, the accuracy of a reasonable classifier would be greater than chance level. Therefore, empirical classification accuracy can serve as an effective test statistic. However, since we do not observe a sample from f_{XY} but a sample from $f^{(2)}_{YX}$, we need to take this into consideration when we train a classifier. Specifically, we compute a classifier

$$\widehat{h} = \underset{h \in \mathcal{H}}{\arg\min} \bigg\{ \frac{1}{n_1} \sum_{i=1}^{n_1} \ell\big(h\big(V_i^{(1)}\big), 1\big) + \frac{1}{n_2} \sum_{i=1}^{n_2} r_X\big(X_i^{(2)}\big) \ell\big(h\big(V_i^{(2)}\big), 2\big) \bigg\}, \tag{8}$$

and use the empirical classification accuracy of $\widehat{h}$, again corrected by the density ratio, as our test statistic. When training and testing are performed on independent datasets, the asymptotic null distribution of the classification accuracy is approximately Gaussian (Kim et al., 2021; Hediger et al., 2022); thereby the critical value can be determined based on this Gaussian approximation. We provide a detailed analysis of this approach in Section 4.1, and present numerical results in Section 5.

Example 6. (Kernel MMD) The last example is a kernel MMD statistic (Gretton et al., 2012). Given a kernel k, the population MMD compares the kernel mean embeddings of two distributions with density functions $f^{(1)}_{YX}$ and f_{YX}, respectively. In a kernel form, the squared MMD can be written as

$$\mathrm{MMD}^2 = \mathbb{E}[k(V_1^{(1)}, V_2^{(1)})] + \mathbb{E}[r_X\big(X_1^{(2)}\big) r_X\big(X_2^{(2)}\big) k(V_1^{(2)}, V_2^{(2)})] - 2\mathbb{E}[r_X\big(X_2^{(2)}\big) k(V_1^{(1)}, V_2^{(2)})],$$

where the bias is corrected via importance weighting. The squared MMD can be estimated as

$$\frac{1}{n_1(n_1-1)} \sum_{1 \le i \neq j \le n_1} k\big(V_i^{(1)}, V_j^{(1)}\big) + \frac{1}{n_2(n_2-1)} \sum_{1 \le i \neq j \le n_2} r_X\big(X_i^{(2)}\big) r_X\big(X_j^{(2)}\big) k\big(V_i^{(2)}, V_j^{(2)}\big) - \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} r_X\big(X_j^{(2)}\big) k\big(V_i^{(1)}, V_j^{(2)}\big),$$

which is an unbiased estimator of the population MMD. Unlike the other three test statistics mentioned earlier, this estimator converges to an infinite sum of weighted chi-squared distributions, whose weights are unknown in practice. This was not a major issue for marginal two-sample testing, where permutation tests could calibrate any test statistic in a nonparametric way. However, the standard permutation approach is no longer valid for conditional two-sample testing, which presents a challenge. Therefore, we focus on another estimator, a linear-time MMD statistic, in Section 4.2, which offers advantages in both the tractability of the asymptotic distribution and computational efficiency.

The preceding discussion assumes that the marginal density ratio r_X is known. As mentioned before, the success of this approach hinges on accurately estimating the density ratio r_X. To this end, one can draw upon a wide range of existing techniques in the literature for density ratio estimation (e.g., Sugiyama et al., 2010, 2012) to obtain a reliable testing result. Using the same dataset for both density ratio estimation and other parts of a statistic often results in plug-in bias. Hence, we recommend using auxiliary dataset obtained through, e.g., sample splitting, to estimate the density ratio. We concretely illustrate this approach in Section 4.1 and Section 4.2, using a classifier-based test statistic and a linear-time MMD statistic.

While various existing tools for density ratio estimation offer flexibility, an inherent drawback of this approach is that the behavior of the test statistic could be erratic when the density ratio is irregular and potentially unbounded. This issue can be mitigated by clipping the density ratio estimate at a certain value or shrinking it to zero (e.g., Shimodaira, 2000). Nevertheless, when prior knowledge indicates that the density ratio behaves poorly or is difficult to estimate, the testing performance may degrade (see Section 5 for numerical results) and thus this approach should be used with caution.

4.1 Classifier-based Approach

This subsection illustrates a classifier-based test for conditional two-sample testing. To simplify the presentation, we assume that $n_1 = n_2 = 2n$ and split the dataset into two: $D_a \coloneqq \{V_i^{(1)}\}_{i=1}^n \cup \{V_i^{(2)}\}_{i=1}^n$ and $D_b \coloneqq \{V_i^{(1)}\}_{i=n+1}^{2n} \cup \{V_i^{(2)}\}_{i=n+1}^{2n}$. For some positive integer $m < n$, we further divide D_a as $D_a^* \coloneqq \{V_i^{(1)}\}_{i=1}^m \cup \{V_i^{(2)}\}_{i=1}^m$ and $D_a^{**} \coloneqq D_a \setminus D_a^*$, and let $\widehat{r}_X$ denote an estimator of r_X formed on D_a^{**}. Additionally, let $\widehat{h}$ be a classifier trained as in (8) based on D_b. Let us write $\widehat{A}_{1,i} \coloneqq \mathbb{1}\{\widehat{h}(V_i^{(1)}) = 1\}$ and $\widehat{A}_{2,i} \coloneqq \widehat{r}_X(X_i^{(2)})\mathbb{1}\{\widehat{h}(V_i^{(2)}) = 2\}$ for $i \in \{1, \ldots, m\}$, and define $\overline{A}_1 \coloneqq m^{-1}\sum_{i=1}^m \widehat{A}_{1,i}$ and $\overline{A}_2 \coloneqq m^{-1}\sum_{i=1}^m \widehat{A}_{2,i}$. The population-level classification accuracy of $\widehat{h}$ is $\mathbb{P}\{\widehat{h}(V^{(1)}) = 1\}/2 + \mathbb{E}[r_X(X^{(2)})\mathbb{1}\{\widehat{h}(V^{(2)}) = 2\}]/2$, which is $1/2$ for any classifier $\widehat{h}$ under the null hypothesis. This observation leads to a classifier-based test statistic for conditional two-sample testing given as

$$\widehat{\mathrm{Acc}} \coloneqq \frac{\sqrt{m}(\overline{A}_1 + \overline{A}_2 - 1)}{\sqrt{\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2}}, \tag{9}$$

where $\widehat{\sigma}_1^2 \coloneqq (m-1)^{-1}\sum_{i=1}^m (\widehat{A}_{1,i} - \overline{A}_1)^2$ and $\widehat{\sigma}_2^2 \coloneqq (m-1)^{-1}\sum_{i=1}^m (\widehat{A}_{2,i} - \overline{A}_2)^2$. To formally establish the limiting distribution of $\widehat{\mathrm{Acc}}$, we consider the following assumptions.

Assumption 1. *Let $m_n \coloneqq m$ be an increasing sequence of positive integers with $\lim_{n\to\infty} m_n = \infty$. Consider a class of null distributions $\mathcal{P}_0$ such that*

(a) There are constants $c_1, c_2 \in (0,1)$ such that $c_1 \leq \inf_{P\in\mathcal{P}_0} \mathbb{P}_P\{\widehat{h}(V^{(1)}) = 1 \mid \widehat{h}\} \leq \sup_{P\in\mathcal{P}_0} \mathbb{P}_P\{\widehat{h}(V^{(1)}) = 1 \mid \widehat{h}\} \leq c_2$ for all sufficiently large n. Moreover, assume that there exist constants $C, \delta > 0$ such that $\sup_{P\in\mathcal{P}_0} \mathbb{E}_P[\{\widehat{r}_X(X^{(2)})\}^{2+\delta}] \leq C$ for all sufficiently large n.

(b) For any $\epsilon > 0$, the density ratio estimator satisfies

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} \mathbb{P}_P\big(m\mathbb{E}_P[\{\widehat{r}_X(X^{(2)}) - r_X(X^{(2)})\}^2 \mid \widehat{r}_X] \geq \epsilon\big) = 0.$$

Assumption 1(a) is imposed to establish the (conditional) central limit theorem for the test statistic with the true density ratio, and it excludes a deterministic classifier, which would return the same prediction value regardless of inputs, under the null hypothesis. Assumption 1(b) ensures that the approximation error from $\widehat{r}_X$ is asymptotically negligible, which is similarly assumed in Hu and Lei (2024). In order to theoretically justify this condition, one needs to take m much smaller than the sample size used for training $\widehat{r}_X$. However, our empirical results in Section 5 illustrate that $\widehat{\mathrm{Acc}}$ approximates $N(0,1)$ closely even under balanced splitting. Therefore, echoing Hu and Lei (2024), we suggest taking $m = \lfloor n/2 \rfloor$ in practice.

Under Assumption 1, the classifier-based test statistic in (9) converges to $N(0,1)$ uniformly over $\mathcal{P}_0$.

Theorem 3. *For the class of null distributions $\mathcal{P}_0$ satisfying Assumption 1, $\widehat{\mathrm{Acc}}$ in (9) converges to $N(0,1)$:*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} \sup_{t\in\mathbb{R}} \big|\mathbb{P}_P(\widehat{\mathrm{Acc}} \le t) - \Phi(t)\big| = 0.$$

The proof of Theorem 3 can be found in Appendix A.1. According to Theorem 3, the classifier-based test rejects the null hypothesis when $\widehat{\mathrm{Acc}} > \Phi^{-1}(1-\alpha)$, which has asymptotic validity over $\mathcal{P}_0$ satisfying Assumption 1. We next improve the efficiency of this procedure via K-fold cross-validation. To describe the procedure, we begin by considering K disjoint subsets of D_a, denoted as $D_{a,1}, D_{a,2}, \ldots, D_{a,K}$, of equal size $m := \lfloor n/K \rfloor$ for simplicity. For $j \in \{1,\ldots,K\}$, let $\overline{A}_{1,j} + \overline{A}_{2,j} - 1$ and $\widehat{\sigma}_{1,j}^2 + \widehat{\sigma}_{2,j}^2$ denote the quantities analogous to $\overline{A}_1 + \overline{A}_2 - 1$ and $\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2$, respectively, by letting $D_a^* = D_{a,j}$ and $D_a^{**} = D_a \backslash D_a^*$. We then define the cross-validated classification accuracy statistic as

$$\widehat{\mathrm{Acc}}_{\mathrm{cv}} := \frac{1}{\sqrt{K}} \sum_{j=1}^{K} \frac{\sqrt{m}(\overline{A}_{1,j} + \overline{A}_{2,j} - 1)}{\sqrt{\widehat{\sigma}_{1,j}^2 + \widehat{\sigma}_{2,j}^2}}. \tag{10}$$

The next corollary proves that the cross-validated accuracy statistic is asymptotically normally distributed under the null hypothesis.

Corollary 1. *Consider the same setting as in Theorem 3. For any fixed $K \ge 2$, it holds that*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} \sup_{t\in\mathbb{R}} \big|\mathbb{P}_P(\widehat{\mathrm{Acc}}_{\mathrm{cv}} \le t) - \Phi(t)\big| = 0.$$

According to Corollary 1, the test that rejects the null when $\widehat{\mathrm{Acc}}_{\mathrm{cv}} > \Phi^{-1}(1-\alpha)$ maintains the asymptotic type I error under control. In terms of power, the cross-validated version is generally more powerful than the accuracy test without cross-validation as it uses the sample more efficiently. We numerically demonstrate this point in Section 5.

Many practical classifiers attempt to mimic the Bayes optimal classifier. For the balanced-sample setting, the Bayes optimal classifier is given as $h^\star(x,y) := \mathbb{1}\big(f_{YX}^{(1)}(y,x)/\{f_{YX}^{(1)}(y,x) + f_{YX}(y,x)\} > 1/2\big)$ whose classification accuracy can be explicitly computed in terms of the total variation (TV) distance. Specifically, twice the classification accuracy can be computed as

$$\mathbb{P}\{h^\star(V^{(1)}) = 1\} + \mathbb{E}[r_X(X^{(2)})\mathbb{1}\{h^\star(V^{(2)}) = 2\}] = 1 + \mathrm{TV}(f_{YX}^{(1)}, f_{YX}),$$

where $\mathrm{TV}(f_{YX}^{(1)}, f_{YX})$ denotes the TV distance between two distributions with densities $f_{YX}^{(1)}$ and f_{YX}, respectively. Since the TV distance becomes zero if and only if two distributions are identical, our classifier-based test can be powerful against general alternatives when the classifier in use approximates the Bayes classifier.

The next subsection develops parallel results using a linear-time MMD statistic.

4.2 Linear-time MMD

In this subsection, we provide a detailed treatment of our second framework by focusing on a linear-time MMD statistic (Gretton et al., 2012, Lemma 14) with a kernel k. As in Section 4.1, we assume that $n_1 = n_2 = 2n$ and split the dataset into two subsets: $D_a := \{V_i^{(1)}\}_{i=1}^n \cup \{V_i^{(2)}\}_{i=1}^n$ and $D_b := \{V_i^{(1)}\}_{i=n+1}^{2n} \cup \{V_i^{(2)}\}_{i=n+1}^{2n}$. Letting $\widehat{r}_X$ be an estimator of r_X formed on D_b and $m = \lfloor n/2 \rfloor$, define

$$\begin{aligned}\widehat{S}_i := {} & k(V_i^{(1)}, V_{i+m}^{(1)}) + \widehat{r}_X(X_i^{(2)})\widehat{r}_X(X_{i+m}^{(2)})k(V_i^{(2)}, V_{i+m}^{(2)}) \\ & - \widehat{r}_X(X_i^{(2)})k(V_i^{(2)}, V_{i+m}^{(1)}) - \widehat{r}_X(X_{i+m}^{(2)})k(V_i^{(1)}, V_{i+m}^{(2)}).\end{aligned}$$

The test statistic that we analyze is a t-statistic based on $\widehat{S}_1, \ldots, \widehat{S}_m$. Specifically, letting $\overline{S} \coloneqq m^{-1} \sum_{i=1}^m \widehat{S}_i$ and $\widehat{\sigma}^2 \coloneqq (m-1)^{-1} \sum_{i=1}^m (\widehat{S}_i - \overline{S})^2$, the (studentized) linear-time MMD statistic is given as

$$\widehat{\mathrm{MMD}}_\ell^2 \coloneqq \frac{\sqrt{m}\overline{S}}{\widehat{\sigma}}. \tag{11}$$

In order to establish the asymptotic normality of $\widehat{\mathrm{MMD}}_\ell^2$, we make the following assumptions. Below, let S_i denote the quantity defined similarly as $\widehat{S}_i$ by replacing $\widehat{r}_X$ with the population counterpart r_X.

Assumption 2. *Consider a class of null distributions $\mathcal{P}_0$ and assume that*

(a) There exist constants $c, C > 0$ such that $\inf_{P \in \mathcal{P}_0} \mathbb{E}_P[S_1^2] \geq c$ and $\sup_{P \in \mathcal{P}_0} \mathbb{E}_P[S_1^{2+\delta}] \leq C$ for some $\delta > 0$.

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

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

Assumption 2(a) is about the moment condition for the population counterpart of $\widehat{S}_i$. This assumption is required to apply the uniform central limit theorem. Assumption 2(b) is, on the other hand, required to prove that the difference between $\widehat{\mathrm{MMD}}_\ell^2$ using $\{\widehat{S}_i\}_{i=1}^m$ and that using $\{S_i\}_{i=1}^m$ are asymptotically negligible. This is similar to Assumption 1(b) and the assumption made in Hu and Lei (2024), but this condition is considerably weaker in terms of convergence rate. Assumption 2(c) assumes that the kernel k is uniformly bounded. While this assumption is met for many practical kernels (e.g., Gaussian kernel), it can be relaxed by adopting more complex moment or convergence assumptions.

Having stated the assumptions, we now present the asymptotic normality of $\widehat{\mathrm{MMD}}_\ell^2$ under the null hypothesis in (1).

Theorem 4. *For the class of null distributions $\mathcal{P}_0$ satisfying Assumption 2, $\widehat{\mathrm{MMD}}_\ell^2$ converges to $N(0,1)$ as*

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

We defer the proof of Theorem 4 to Appendix A.3. Based on Theorem 4, the test that rejects the null when $\widehat{\mathrm{MMD}}_\ell^2 > \Phi^{-1}(1-\alpha)$ controls the size uniformly over the class of distributions that satisfy Assumption 2.

Similarly to the classification-based test, we can improve the efficiency of $\widehat{\mathrm{MMD}}_\ell^2$ via K-fold cross-validation. To describe this process, we begin by partitioning the dataset of size $2n$ into K-folds, denoted as $D_1, D_2, \ldots, D_K$, of equal size for simplicity. For $j \in \{1, \ldots, K\}$, let $\overline{S}_j$ and $\widehat{\sigma}_j^2$ denote the quantities similarly defined as $\overline{S}$ and $\widehat{\sigma}^2$, respectively, by letting $D_a = D_j$ and $D_b = \cup_{i=1}^K D_i \setminus D_j$. We then define the cross-validated MMD statistic as

$$\widehat{\mathrm{MMD}}_{\mathrm{cv}}^2 \coloneqq \frac{1}{K} \sum_{j=1}^K \frac{\sqrt{n}\overline{S}_j}{\widehat{\sigma}_j}. \tag{12}$$

The next corollary shows that $\widehat{\mathrm{MMD}}_{\mathrm{cv}}^2$ converges to $N(0,1)$ under the same conditions for Theorem 4.

Corollary 2. *Consider the same setting as in Theorem 4. Then for a fixed $K \geq 2$, it holds that*

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

Again, the test that rejects the null when $\widehat{\mathrm{MMD}}_{\mathrm{cv}}^2 > \Phi^{-1}(1-\alpha)$ is asymptotically level α based on Corollary 2. When the kernel k is a characteristic kernel (e.g., Fukumizu et al., 2007), the population MMD becomes equal to zero if and only if two distributions coincide. Thus as for the classifier-based tests in Section 4.1, the MMD-based tests can be powerful against general alternatives, provided that the density ratio r_X can be accurately estimated.

4.3 Block-wise MMD

In this subsection, we develop our second framework, building on the block-wise MMD statistic proposed by Zaremba et al. (2013) with kernel k. This statistic admits asymptotic normality and, depending on the choice of block size B, it encompasses the linear-time MMD statistic (Gretton et al., 2012, Lemma 14) as a special case while achieving greater statistical power when B is chosen appropriately. Unlike the classifier-based approach described in Section 4.1, we partition the data into two subsets: $D_a := \{V_i^{(1)}\}_{i=1}^{n-N} \cup \{V_i^{(2)}\}_{i=1}^{n-N}$ and $D_b := \{V_i^{(1)}\}_{i=n-N+1}^{n} \cup \{V_i^{(2)}\}_{i=n-N+1}^{n}$. Let $\widehat{r}_X$ be an estimator of r_X constructed on D_b, and denote $M = n - N$. Letting $S := \lfloor M/B \rfloor$ as the number of blocks, we define for each $b = 1, \ldots, S$ the index set

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

For each i, define $W_i := (V_i^{(1)}, V_i^{(2)})$. Then, for each pair $(i,j) \in I_b$, we define the pairwise kernel contribution as

$$\begin{aligned} \widehat{H}_{ij} \; &= \; k\Big(V_i^{(1)}, V_j^{(1)}\Big) + \widehat{r}_X(X_i^{(2)})\, \widehat{r}_X(X_j^{(2)})\, k\Big(V_i^{(2)}, V_j^{(2)}\Big) \\ &\quad - \widehat{r}_X(X_i^{(2)})\, k\Big(V_i^{(2)}, V_j^{(1)}\Big) - \widehat{r}_X(X_j^{(2)})\, k\Big(V_i^{(1)}, V_j^{(2)}\Big) . \end{aligned} \tag{13}$$

For notational simplicity, we write $\widehat{H}_{ij} := \widehat{H}(W_i, W_j)$. For each block b, define the block statistic and its normalized aggregate as follows:

$$\widehat{\eta}_b := B \cdot \binom{B}{2}^{-1} \sum_{(i,j)\in I_b} \widehat{H}_{ij}, \quad \bar{\eta} := \frac{1}{S}\sum_{b=1}^{S} \widehat{\eta}_b, \quad \widehat{\sigma}_B^2 := \frac{1}{S-1}\sum_{b=1}^{S}\big(\widehat{\eta}_b - \bar{\eta}\big)^2.$$

The studentized block-wise MMD statistic is then

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

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

(a) There exist constants $c, C > 0$ such that $\inf_{P\in\mathcal{P}_0} \mathbb{E}_P\big[H_{12}^2\big] \geq c$ and $\sup_{P\in\mathcal{P}_0} \mathbb{E}_P\big[|H_{12}|^{2+\delta}\big] \leq C$ for some $\delta > 0$.

(b) Suppose that $\sup_{P\in\mathcal{P}_0} \mathbb{E}_P\big[r_X(X^{(2)})^2\big] < \infty$ 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 $\|k\|_\infty \leq K$.

(d) The block size B satisfies $c_1 M^\gamma \leq B \leq c_2 M^\gamma$ for some constants $c_1, c_2 > 0$ and $0 \leq \gamma < 1$, for all sufficiently large M.

(e) Suppose the subsample size satisfies $M \wedge N \to \infty$, with growth rates such that $M^{1+\gamma}/N \to c' \in [0,\infty)$ for some $0 \leq \gamma < 1$ and constant c'. Here, $M \wedge N := \min(M,N)$.

Assumption 3(a) imposes a moment condition on the population counterpart of $\widehat{H}_{ij}$ to ensure that the block-wise averages η_b have uniformly bounded moments, a key requirement for applying the uniform central limit theorem. Assumption 3(b) guarantees that replacing the population kernel terms H_{ij} with their estimates $\widehat{H}_{ij}$ has only a negligible effect on $\widehat{\mathrm{MMD}}_B^2$. A related condition in Hu and Lei (2024) is strictly stronger. Under equal sample splitting, they require $\sqrt{n_{11}}\|\widehat{G}_{11} - G_{11}\|_{2,*} = o_P(1)$ (with $\|\cdot\|_{2,*}$ denoting the conditional ℓ_2 norm, $\widehat{G}_{11}$ corresponding to the estimator based on the density ratio $\widehat{r}_X$, G_{11} to its population counterpart r_X, and $o_P(1)$ indicating convergence in probability). This condition forces the estimation error to decay at the $n_{11}^{-1/2}$ rate. In contrast, our assumption directly bounds the squared error with sample size N, which corresponds to only an $N^{-1/4}$ rate in root form, and is therefore weaker. Assumption 3(c) assumes that the kernel k is uniformly bounded. While this assumption is satisfied by many practical kernels (e.g., the Gaussian kernel), it can be relaxed by adopting more complex moment or convergence assumptions. Assumption 3(d) requires the block size B to grow in a balanced manner: if blocks are too small ($\gamma = 0$), the statistic reduces to the linear-time MMD, coinciding exactly with it when $B = 2$, whereas if they are too large ($\gamma = 1$), then B is of order M, so the number of blocks S does not increase, preventing the CLT from holding. To ensure $S \geq 2$, we restrict to $0 \leq \gamma < 1$. Assumption 3(e) requires M and N to grow at compatible rates so that density ratio error is negligible and the block-wise statistic admits a Gaussian limit. As noted by Bordino and Berrett (2025, Example 1), density ratio estimation and statistic computation may use different effective sample sizes, motivating our choice to adopt distinct splitting ratios, though the exact rates differ from theirs.

Theorem 5. *For the class of null distributions $\mathcal{P}_0$ satisfying Assumption 3, $\widehat{\mathrm{MMD}}_B^2$ converges to $N(0,1)$ as*

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

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 n be evenly partitioned into K disjoint folds, denoted by $D_1, \ldots, D_K$, each containing n/K observations. For each $j \in \{1, \ldots, K\}$, define $D_a := D_j$ as 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 allocates a larger portion of the data, specifically a $(1 - 1/K)$ fraction, to density-ratio estimation, which can improve both accuracy and stability.

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 3. *Under the conditions stated in Theorem 5, and for a fixed $K \geq 2$, it holds that*

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

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.

4.4 Quadratic-time MMD

In this subsection, we analyze our second framework, which is based on a quadratic-time MMD statistic with kernel k, introduced in Example 6 and formalized in Gretton et al. (2012, Theorem 12). As in Section 4.3, we split the sample of size n into two disjoint parts: D_a, containing the first $M = n - N$ observations used to compute the test statistic, and D_b, containing the remaining N observations used to estimate the density ratio. Using D_b, we then construct the estimator $\widehat{r}_X$ of r_X.

Based on the kernel $\widehat{H}$ introduced in (13), the quadratic-time MMD estimator takes the U-statistic form

$$\widehat{\mathrm{MMD}}_u^2 = \frac{1}{M(M-1)} \sum_{i \neq j} \widehat{H}(W_i, W_j).$$

In order to study the asymptotic behavior of $\widehat{\mathrm{MMD}}_u^2$, we introduce several notational conventions. First, let $H(W_i, W_j)$ denote the population version of the kernel, obtained by replacing $\widehat{r}_X$ with r_X. Define $\varphi(W_i) := \psi(V_i^{(1)}) - r_X(X_i^{(2)})\,\psi(V_i^{(2)})$, so that $H(W_i, W_j) = \langle \varphi(W_i), \varphi(W_j) \rangle_{\mathcal{H}_k}$.

Next, for each sample size n, let $P_n := P_{XY,n}^{(1)} \otimes P_{XY,n}^{(2)}$ be the joint distribution of $W = (V^{(1)}, V^{(2)})$, with mean embedding $\mu_{P_n} := \mathbb{E}_{P_n}[\varphi(W)]$. This triangular-array formulation allows us to capture settings where the underlying distribution may vary with n. Using this embedding, we define the centered kernel

$$\begin{aligned} \widetilde{H}_n(W_i, W_j) &:= \langle \varphi(W_i) - \mu_{P_n}, \varphi(W_j) - \mu_{P_n} \rangle_{\mathcal{H}_k} \\ &= H(W_i, W_j) - \mathbb{E}_{W \sim P_n}[H(W_i, W)] - \mathbb{E}_{W \sim P_n}[H(W, W_j)] + \mathbb{E}_{W, W' \sim P_n}[H(W, W')]. \end{aligned} \tag{14}$$

Finally, for any fixed n, let $\{(\lambda_{\ell,n}^{(P_n)}, \Psi_{\ell,n}^{(P_n)}) : \ell \geq 1\}$ denote the eigenvalue–eigenfunction pairs of the operator

$$g \mapsto \int \widetilde{H}_n(\cdot, w)\, g(w)\, dP_n(w).$$

When $\widetilde{H}_n$ is square-integrable, it admits the spectral expansion

$$\widetilde{H}_n(w, w') = \sum_{\ell=1}^{\infty} \lambda_{\ell,n}^{(P_n)}\, \Psi_{\ell,n}^{(P_n)}(w) \Psi_{\ell,n}^{(P_n)}(w').$$

Let $\{\mathcal{E}_n\}_{n \geq 1}$ denote a sequence of distribution classes, where each $\mathcal{E}_n$ consists of pairs of distributions for sample size n. Define $\mathcal{P}_n^{(0)} \subset \mathcal{E}_n$ as the set of distribution pairs satisfying the null hypothesis H_0 in (1). For notational simplicity, we omit the dependence on P_n and simply write $\lambda_{\ell,n}$ and $\Psi_{\ell,n}$ in place of $\lambda_{\ell,n}^{(P_n)}$ and $\Psi_{\ell,n}^{(P_n)}$.

Assumption 4. *Consider a class of null distributions $\mathcal{P}_n^{(0)}$. For each n, let*

$$\beta_n := \mathbb{E}_{P_n}[\widetilde{H}_n(W_1, W_2)^4], \qquad d_n := \sum_{\ell \geq 1} |\lambda_{\ell,n}|,$$

where $(\lambda_{\ell,n})_{\ell \geq 1}$ are the eigenvalues of the Hilbert–Schmidt operator associated with $\widetilde{H}_n$, ordered so that $|\lambda_{1,n}| \geq |\lambda_{2,n}| \geq \cdots$, and satisfying $|\lambda_{1,n}| \geq |\lambda_{2,n}| > 0$ for all n.

- *(a) Suppose that*

$$\frac{\beta_n}{\sqrt{|\lambda_{1,n}\lambda_{2,n}|}\, n^{1/6}} \to 0, \qquad \frac{d_n}{\sqrt{|\lambda_{1,n}\lambda_{2,n}|}\, n^{1/2}} \to 0.$$

- *(b) Suppose that $\sup_{P_n \in \mathcal{P}_n^{(0)}} \mathbb{E}_{P_n}\big[r_X(X^{(2)})^2\big] < \infty$ and $\sup_{P_n \in \mathcal{P}_n^{(0)}} \mathbb{E}_{P_n}\big[\{\widehat{r}_X(X^{(2)}) - r_X(X^{(2)})\}^2\big] = o\big(N^{-1/2}\big)$.*
- *(c) The kernel is uniformly bounded as $\|k\|_\infty \leq K$.*
- *(d) Suppose the subsample size satisfies $M \wedge N \to \infty$, with growth rates such that $M^2/N \to c' \in [0, \infty)$ for some constant c'.*

Assumption 4(a) is crucial for applying the Berry–Esseen bound in Lemma 7, as it controls the fourth moment and the spectral sum relative to the leading eigenvalues, ensuring that the quadratic-time U-statistic is well approximated by its non-Gaussian quadratic-form limit. The remaining conditions, Assumption 4(b)–(d), play the same technical roles as those introduced in Section 4.3: Assumption 4(b) ensures that the difference between the U-statistics based on $\widehat{H}$ and H is asymptotically negligible, Assumption 4(c) requires the kernel to be uniformly bounded, and Assumption 4(d) specifies compatible growth rates of the two sample sizes so that density ratio error is negligible and the quadratic-time statistic is well defined. In fact, Assumption 4(d) coincides exactly with the case discussed in Assumption 3(e) when the parameter γ is equal to one.

Theorem 6. *For the class of null distributions $\mathcal{P}_n^{(0)}$ satisfying Assumption 4, the quadratic-time MMD U-statistic satisfies the uniform distributional convergence*

$$\lim_{M\wedge N\to\infty} \sup_{P_n\in\mathcal{P}_n^{(0)}} \sup_{t\in\mathbb{R}} \left| \mathbb{P}_{P_n}\left(M\cdot\widehat{\mathrm{MMD}}_u^2 \le t\right) - \mathbb{P}(G_n^{(P_n)} \le t)\right| = 0,$$

where

$$G_n^{(P_n)} \sim \sum_{\ell=1}^{\infty} \lambda_{\ell,n}^{(P_n)}(a_\ell^2 - 1), \qquad \{a_\ell\}_{\ell\ge1} \overset{i.i.d.}{\sim} N(0,1),$$

and $\{\lambda_{\ell,n}^{(P_n)}\}_{\ell\ge1}$ are the eigenvalues in the spectral expansion of $\widetilde{H}_n$.

Since the limiting distribution involves an infinite sum of shifted chi-squared variables, which is intractable in practice, we employ a multiplier bootstrap approach for hypothesis testing. Given n i.i.d. Rademacher random variables $\xi := (\xi_1, \ldots, \xi_M)$ with values in $\{-1,+1\}^n$, independent of the sample $\mathbb{W} = (W_1, \ldots, W_M)$, define the wild bootstrap statistic

$$\widehat{\mathrm{MMD}}_{\mathrm{wild}}^2 := \frac{1}{M(M-1)} \sum_{i\ne j} \xi_i\xi_j\, \widehat{H}(W_i, W_j),$$

where $\widehat{H}$ is the kernel defined in (13).

Proposition 1. *For the class of null distributions $\mathcal{P}_n^{(0)}$ satisfying Assumption 4, the wild bootstrap procedure is uniformly consistent in the sense that*

$$\lim_{M\wedge N\to\infty} \sup_{P_n\in\mathcal{P}_n^{(0)}} \sup_{x\in\mathbb{R}} \left| \mathbb{P}_{P_n}\left(M\cdot\widehat{\mathrm{MMD}}_{\mathrm{wild}}^2 \le x \mid \mathbb{W}\right) - \mathbb{P}_{P_n}\left(G_n^{(P_n)} \le x\right)\right| = 0,$$

where $G_n^{(P_n)}$ is defined in Theorem 6. Equivalently, the conditional distribution of the wild bootstrap statistic converges to the asymptotic null distribution of the quadratic-time MMD U-statistic, uniformly over $\mathcal{P}_n^{(0)}$.

Let $\{\widehat{T}^{(b)} : b = 1, \ldots, B\}$ denote the wild bootstrap replicates of the quadratic-time MMD statistic, generated using independent Rademacher weights $\xi^{(b)} := (\xi_1^{(b)}, \ldots, \xi_M^{(b)})$ as in Proposition 1. Define the augmented set

$$\{\widehat{T}^{(1)}, \ldots, \widehat{T}^{(B)}, \widehat{T}\},$$

where $\widehat{T}$ denotes the original statistic $M\cdot\widehat{\mathrm{MMD}}_u^2$. Let

$$\widehat{T}_{(1)} \le \widehat{T}_{(2)} \le \cdots \le \widehat{T}_{(B+1)}$$

denote its order statistics. For $\alpha \in (0,1)$, the empirical $(1-\alpha)$ wild bootstrap quantile is

$$\begin{aligned}\widehat{q}_{1-\alpha}^{B}\left(\{\widehat{T}^{(b)}\}_{1\le b\le B} \mid \mathbb{W}\right) &:= \inf\left\{u\in\mathbb{R} : \frac{1}{B+1}\sum_{b=1}^{B+1} \mathbb{1}\left(\widehat{T}^{(b)} \le u\right) \ge 1-\alpha\right\} \\ &= \widehat{T}_{(\lceil (B+1)(1-\alpha)\rceil)}.\end{aligned}$$

For notational simplicity, and whenever no ambiguity arises, we write the threshold as $\widehat{q}_{1-\alpha}$. By Proposition 1, the test that rejects the null when $M\cdot\widehat{\mathrm{MMD}}_u^2 > \widehat{q}_{1-\alpha}$ controls the size uniformly over the class of distributions that satisfy Assumption 4.

The next section illustrates the numerical performance of the proposed tests in comparison to existing methods.

5 Numerical Experiments

In this section, we evaluate the empirical performance of the proposed tests, alongside existing methods from the literature, across various scenarios. Within each scenario, we compare the conditional independence testing (CIT) approach described in Section 3 with the density ratio-based testing (DRT) approach described in Section 4. A brief overview of these methods is provided in Section 5.1, with additional implementation details available in Appendix C.

We empirically evaluate the type I error and power of these methods using both synthetic datasets in Section 5.2 and two real-world datasets in Section 5.3. Our simulation studies with synthetic datasets cover three distinct scenarios, each featuring both bounded and unbounded marginal density ratios. This setup allows us to explore a range of situations from relatively simple cases (with bounded density ratios) to more complex and challenging ones (with unbounded density ratios).

In all simulation studies, the dimension of the covariates X is fixed at $p = 10$. The simulation results are averaged over 500 repetitions with a significance level of $\alpha = 0.05$. For DRT methods, we employ a probabilistic classification approach using linear logistic regression for density ratio estimation (Sugiyama et al., 2010, Section 3). For CIT methods relying on regression estimation, such as the Generalized Covariance Measure (Shah and Peters, 2020) and the Projected Covariance Measure (Lundborg et al., 2022), we use a Random Forest model as the underlying regression method. On the other hand, for the WGSC (Williamson et al., 2023), we utilize XGBoost as it appears to perform best in our simulation scenarios.

It is crucial to note that the efficacy and validity of these methods depend on specific assumptions, which vary across different approaches. For example, DRT methods rely heavily on accurate density ratio estimation, while CIT methods depend on reliable estimation of conditional operators, such as regression functions. We aim to emphasize this distinction by providing a comprehensive evaluation of these approaches for conditional two-sample testing. This analysis sheds light on how each method performs under different conditions, guiding practitioners in choosing the most suitable approach for their specific applications.

5.1 Overview of Testing Methods

This subsection outlines the testing methods employed in our experiments. Further details on the implementation of these methods can be found in Appendix C. We denote the single-split classifier-based test in Section 4.1 as CLF and its cross-fit version as †CLF. Both classifier-based test statistics are built on a specific form of classifiers detailed in Appendix C.1. Moreover, we denote the linear-time MMD in Section 4.2 as MMD-ℓ and its cross-fit version as †MMD-ℓ. The block-wise MMD test introduced in Section 4.3 is denoted as MMDb, and its cross-fit version as †MMDb. Additional conditional two-sample testing methods included in our experiments are as follows:

CP: The conformal prediction (CP) test utilizes a conformity score to produce a weighted rank sum test statistic. This statistic is constructed by estimating both marginal and conditional density ratios, which can be approached using various density ratio estimation methods. For further details, please refer to Hu and Lei (2024).

DCP: The debiased conformal prediction (DCP) test refines the CP test by reducing bias through the use of Neyman orthogonality and using cross-fitting to improve efficiency. This enhancement guarantees asymptotic normality under certain conditions. Further technical details and theoretical guarantees are described in Chen and Lei (2024).

For CIT methods, we employ one kernel-based and three regression-based testing approaches. All of these CIT methods are implemented via Algorithm 1. We empirically observe in Appendix C.4 that the performance of the CIT methods remains largely consistent regardless of whether Algorithm 1 is applied or not, especially when the sample size is large. The following methods are included in our experiments:

RCIT: The randomized conditional independence test (RCIT) approximates the kernel conditional independence test by leveraging random Fourier features, allowing it to scale linearly with sample size. We use the default options in its implementation, and further details are provided in Strobl et al. (2019).

GCM: The generalized covariance measure (GCM) by Shah and Peters (2020) utilizes the normalized covariance between residuals from regression models as a test statistic. This approach provides a flexible framework that can be adapted to various settings by selecting appropriate regression techniques.

PCM: The projected covariance measure (PCM) is a variation of the GCM applied to a transformed version of X. For our simulations, we follow Algorithm 1 from Lundborg et al. (2022). This method retains the general structure of the GCM while introducing a projection step, which enhances power, particularly when the conditional covariance is zero or near zero.

WGSC: This testing procedure proposes a general framework for nonparametric inference on interpretable, algorithm-agnostic variable importance. In our simulation, we follow the approach outlined in Williamson et al. (2023, Algorithm 3), which utilizes sample splitting and cross-fitting.

The code that reproduces all simulation results is available at: `https://github.com/suman-cha/Cond2ST`.

5.2 Synthetic Data Examples

We design three synthetic data scenarios to evaluate the performance of conditional two-sample testing methods under different conditions. Each scenario is implemented with both unbounded (U) and bounded (B) density ratios to assess how the difficulty of density ratio estimation affects the performance of each method. The marginal distributions of X remain consistent across scenarios with unbounded density ratios, and similarly, across scenarios with bounded density ratios.

For the unbounded case (U), we employ Gaussian distributions for marginal distributions of X. Specifically, for $j = 1$, samples are drawn from a standard Gaussian distribution, i.e., $x^{(1)} \sim N(0, I_p)$, where I_p is the identity matrix of dimension p. For $j = 2$, we introduce a covariate shift by sampling from a Gaussian distribution with mean vector $\mu = (1, 1, -1, -1, 0, \ldots, 0)^\top$ and the same covariance structure, i.e., $x^{(2)} \sim N(\mu, I_p)$. In the bounded case (B), we truncate the support of both distributions to $[-0.5, 0.5]$ in each dimension, resulting in truncated Gaussian distributions, $x^{(1)} \sim TN(0, I_p)$ and $x^{(2)} \sim TN(\mu, I_p)$, where μ is the same as in the unbounded case.

Scenario 1: Linear Model with Mean Shift. Inspired by the work of Hu and Lei (2024), this scenario investigates the efficacy of testing methods in detecting the mean difference between two linear models. For each $j \in \{1, 2\}$, we set $y^{(j)} \,|\, x^{(j)} = \delta^{(j)} + x^{(j)\top}\beta + \epsilon^{(j)}$, where $\epsilon^{(j)}$ follows a t-distribution with 2 degrees of freedom. The regression coefficient β is set to $(1, -1, -1, 1, 0, \ldots, 0)^\top$. Under the null hypothesis, we set $\delta^{(1)} = \delta^{(2)} = 0$, while for the alternative hypothesis, we introduce a mean shift by setting $\delta^{(1)} = 0$ and $\delta^{(2)} = 0.5$, thereby creating a difference in the two conditional distributions.

Scenario 2: High Variability in Conditional Distribution. We also investigate the effect of high variability in the conditional distribution, slightly modifying the example outlined in Chatterjee et al. (2024, Section 6.2). Under the null hypothesis, we model the conditional distributions as $y^{(j)} \,|\, x^{(j)} \sim N\left(x^{(j)\top}\beta^{(j)}, (\sigma^{(j)})^2\right)$, where $\beta^{(j)} = \mathbf{1}_p$ defined as a p-dimensional vector of ones and $(\sigma^{(j)})^2 = 10^2$ for both $j \in \{1, 2\}$. This implies that $\beta^{(1)}$ and $\beta^{(2)}$ are identical under the null hypothesis. For the alternative hypothesis, we modify $\beta^{(2)}$ to $(1, \ldots, 1, 0)^\top$ and introduce heteroscedasticity by varying the variance for $j = 2$ as $(\sigma^{(2)})^2 = 10\left(1 + \exp\left(-\|x^{(2)} - 0.5\mathbf{1}_p\|_2^2/64\right)\right)$.

Scenario 3: Post-Nonlinear Model. Our final scenario considers a post-nonlinear (PNL) model, which is widely used in causal predictive inference (Zhang et al., 2017; Li et al., 2023). It tests the capability of

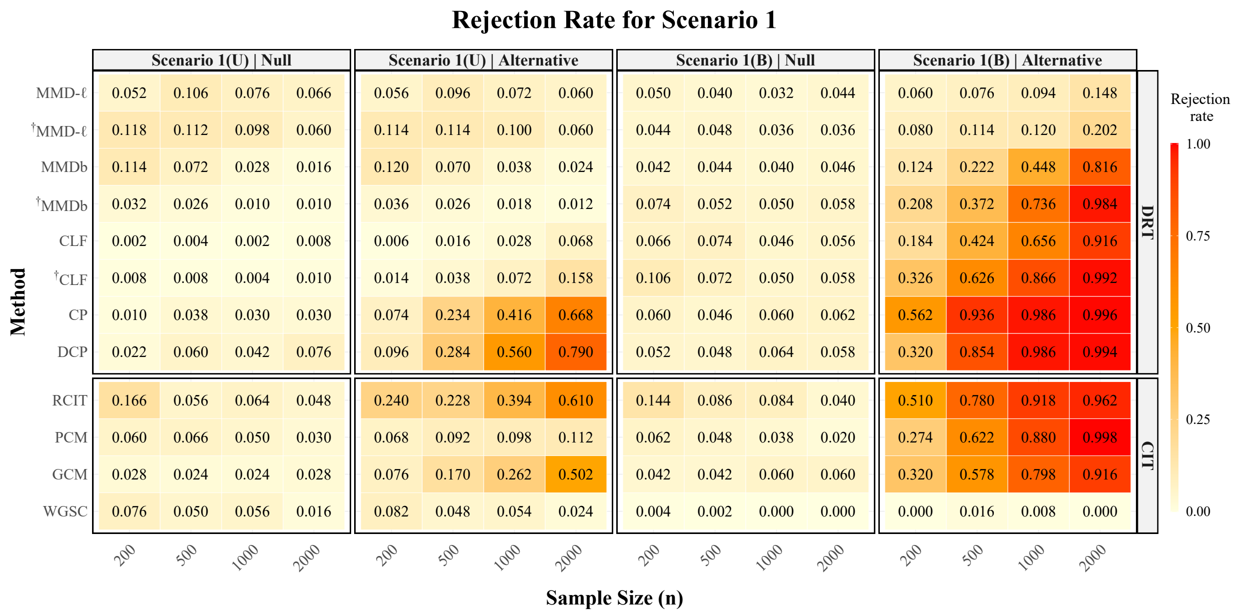

Figure 1: Rejection rates for Scenario 1 under null and alternative hypotheses, shown for both unbounded (U) and bounded (B) settings. Results are averaged over 500 repetitions with significance level $\alpha = 0.05$.

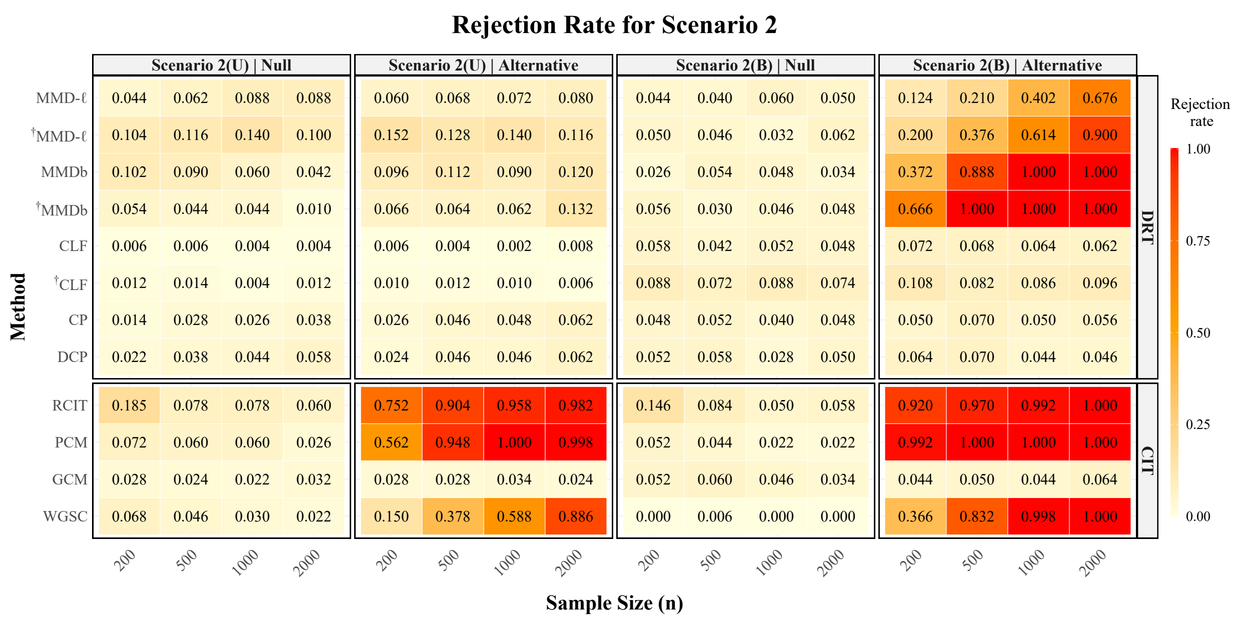

Figure 2: Rejection rates for Scenario 2 under null and alternative hypotheses, shown for both unbounded (U) and bounded (B) settings. Results are averaged over 500 repetitions with significance level $\alpha = 0.05$.

the methods to detect differences in non-linear relationships between variables. We model the conditional distributions as $y^{(j)} \mid x^{(j)} = f^{(j)}(x^{(j)\top}\mathbf{1}_p + 2\epsilon)$, where $\epsilon \sim N(0, 1)$ and $j \in \{1, 2\}$. Under the null hypothesis, we set $f^{(j)}(x) = \cos(x)$ for both $j \in \{1, 2\}$, while for the alternative hypothesis, $f^{(2)}(x)$ is randomly sampled from the set $\{x, x^2, x^3, \sin(x), \tanh(x)\}$.

Our experimental results provide several key insights into the performance of conditional two-sample testing methods across diverse scenarios. A consistent pattern observed throughout all scenarios is the superior performance of DRT methods in bounded settings compared to unbounded settings. This improvement

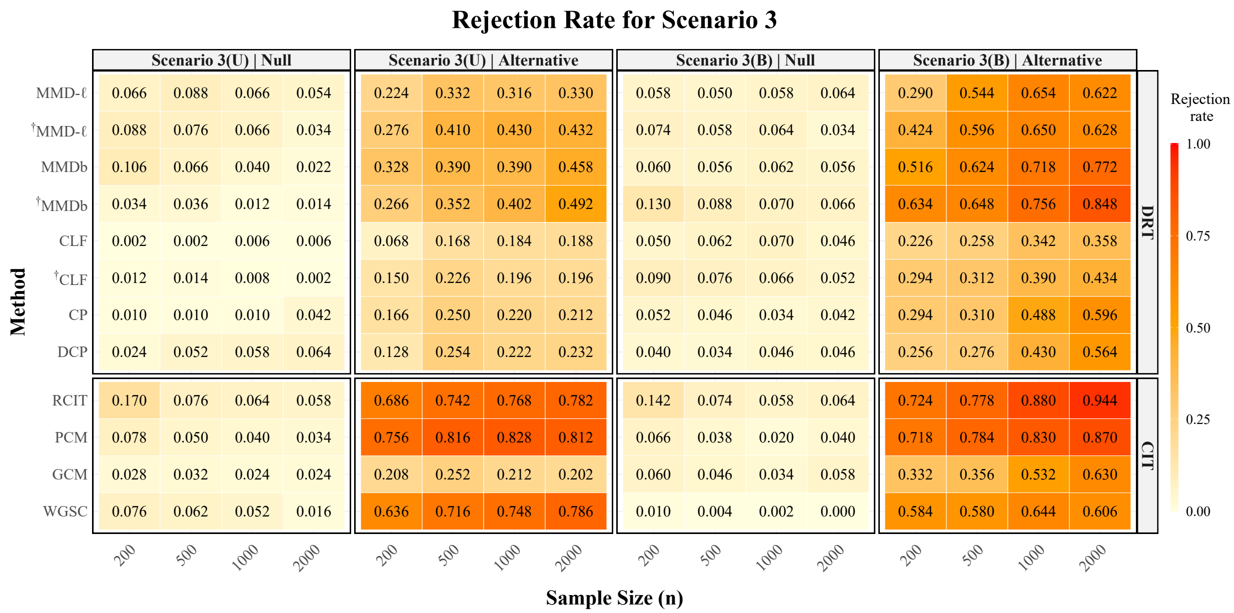

Figure 3: Rejection rates for Scenario 3 under null and alternative hypotheses, shown for both unbounded (U) and bounded (B) settings. Results are averaged over 500 repetitions with significance level $\alpha = 0.05$.

can be attributed to the relative ease of density ratio estimation when the density ratio is bounded, leading to more stable results. In contrast, CIT methods exhibit relatively consistent performance regardless of whether the density ratios are bounded or unbounded.

In Scenario 1, the classifier-based test shows the most significant improvement in performance when transitioning from unbounded to bounded cases. However, MMD-ℓ shows lower sensitivity in detecting mean shifts compared to other DRT methods. Among CIT methods, RCIT and GCM exhibit the best performance in this scenario. Scenario 2 highlights the strengths of MMD-ℓ, which only considers the marginal density ratio of X, in comparison to other DRT methods that account for conditional density ratios. MMD-ℓ shows a distinct advantage in this scenario. CIT methods also generally perform well under these conditions, showing their robustness to complex distributional changes. In Scenario 3, which tests the ability of methods to detect non-linear relationships, all DRT methods improve performance in the bounded case. Among CIT methods, there is no significant difference in performance, except for GCM and WGSC. GCM shows improved performance in the bounded case, while WGSC shows degraded performance.

These results underscore the critical role of accurate density ratio estimation in determining the performance of DRT methods. While CIT methods demonstrate consistent performance across both bounded and unbounded cases, suggesting their utility in a wide range of practical scenarios, they also have limitations. CIT methods, particularly regression-based approaches like GCM, PCM, and WGSC, can be sensitive to the choice of regression model, as we demonstrate in Appendix C. Notably, RCIT exhibits high type I error rates in all scenarios when sample sizes are relatively small, suggesting that caution is needed when applying RCIT to limited datasets. On the other hand, some methods show overly conservative behavior in certain scenarios. The cross-validated versions of DRT methods (†MMD-ℓ and †CLF) consistently show power gains compared to their non-cross-validated counterparts as discussed in Section 4.1 and Section 4.2. Overall, our findings offer important insights into the strengths and limitations of different conditional two-sample testing methods.

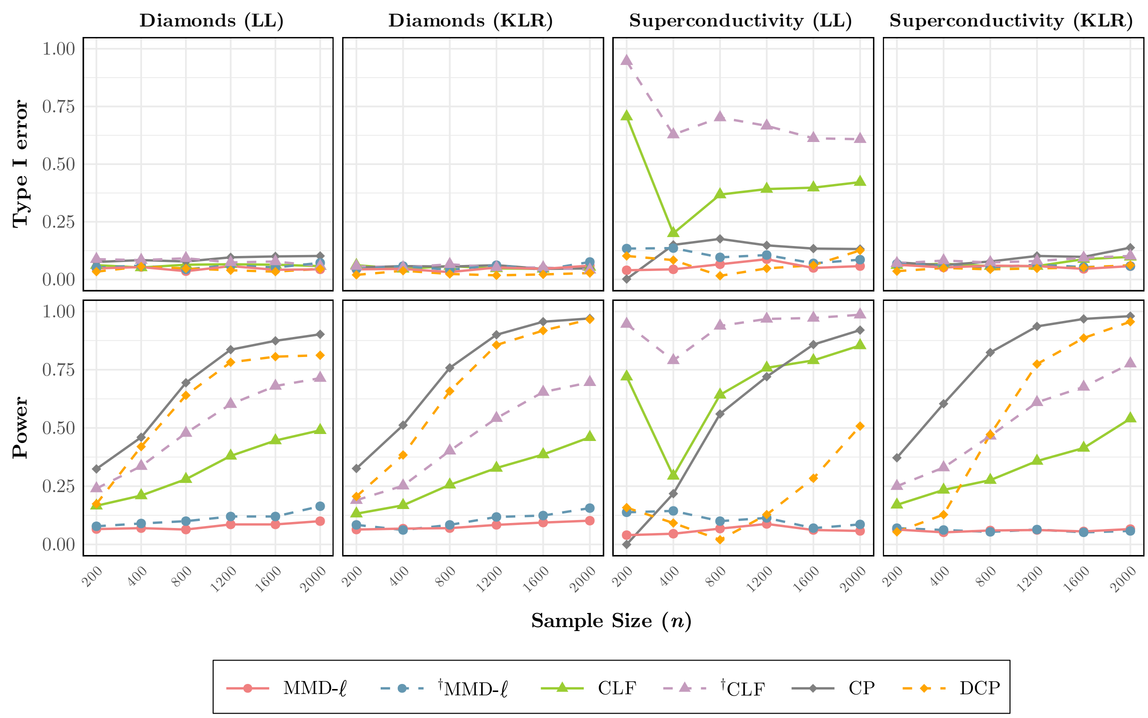

Figure 4: Performance comparison of DRT methods on diamonds and superconductivity datasets using LLR and KLR for density ratio estimation. Rejection rates are averaged over 500 repetitions with $\alpha = 0.05$, under null (*top*) and alternative (*bottom*) hypotheses.

5.3 Real Data Analysis

We further evaluate the performance of our proposed approaches on two real-world datasets: the diamonds dataset and the superconductivity dataset. Following Kim et al. (2023), we treat each dataset as a population from which we draw samples, allowing for controlled experiments with known ground truth. Prior to analysis, we apply standard scaling to both X and Y variables. To introduce covariate shift, we implement biased sampling procedures. Specifically, we sample $X^{(1)}$ uniformly from the original feature space, while $X^{(2)}$ is sampled with probability proportional to $\exp(-x_1^2)$, where x_1 denotes the first feature of X. For the response variable Y, under the null hypothesis, we employ uniform sampling for both $Y^{(1)}$ and $Y^{(2)}$, breaking the original X-Y dependence to ensure identical conditional distributions. Under the alternative hypothesis, $Y^{(1)}$ is sampled uniformly, while $Y^{(2)}$ is sampled with probability proportional to $\exp(-y)$, where y represents the values of Y in the dataset. Figure 4 illustrates the performance of the DRT methods on both datasets, using linear logistic (LLR) and kernel logistic regression (KLR) for density ratio estimation.

Diamonds dataset. The diamonds dataset, available in the R package `ggplot2`, consists of 53,490 observations and 10 features, including price, carat, clarity and color. In our analysis, we set the price variable as Y, and use the 6 numerical variables `(carat, depth, table, x, y, z)` as X. As illustrated in Figure 4, most DRT methods exhibit good type I error control under both LLR and KLR, with rejection rates generally close to the significance level α. Under the alternative hypothesis, we observe a clear trend of increasing power with sample size for all methods. Particularly, the cross-validated versions (†MMD-ℓ and †CLF) exhibit improved power, consistent with our observations in the synthetic data examples.

Superconductivity dataset. The superconductivity dataset, obtained from the UCI Machine Learning Repository and compiled by Hamidieh (2018), presents a more complex and high-dimensional challenge compared to the diamonds dataset. It comprises 81 features extracted from 21,263 superconductors, with the critical temperature at which the material transitions to a superconducting state serving as the response variable Y. The results reveal a significant contrast between density ratio estimation methods based on LLR and KLR. Under LLR, several DRT methods, especially the classifier-based tests, struggle to control the type I error, with rejection rates far exceeding the significance level. Conversely, when using KLR for density ratio estimation, DRT methods show improved type I error control.

These empirical findings emphasize the importance of carefully considering the nature of the data and the choice of density ratio estimation techniques when applying DRT methods for conditional two-sample testing. The performance of different methods can vary significantly, indicating the need for careful method selection and, potentially, more advanced approaches when handling complex and high-dimensional data. While we focus on DRT methods in this section, experimental results for CIT methods are presented in Appendix C.3 for completeness.

6 Conclusion

In this paper, we shed new light on the relatively underexplored problem of conditional two-sample testing. We begin by characterizing the fundamental difficulty of the problem and highlighting the importance of assumptions to make it feasible. We then introduce two general frameworks: (1) converting conditional independence tests into conditional two-sample tests and (2) transforming the problem of comparing conditional distributions into marginal distributions based on density ratio estimation. Both approaches offer significant flexibility, allowing one to leverage well-developed tools to effectively tackle the problem.

Our work opens up several interesting directions for future work. One promising avenue is to extend our framework to conditional K-sample testing with a general $K \geq 2$. Such an extension would expand the applicability of our framework beyond the comparison of just two groups. This setting is related to conditional independence testing where Z is a categorical random variable taking values in $\{1, 2, \ldots, K\}$. We expect our results established in Section 3 to serve as a cornerstone for this extension. Another direction worth exploring is establishing a framework for conditional two-sample testing based on resampling methods. One promising approach is the Sampling Importance Resampling (SIR) algorithm (Givens and Hoeting, 2012, Chapter 6.3), which allows us to obtain an approximate sample from the distribution with density f_{YX}. Future work can focus on methods that compare the sample from $P_{XY}^{(1)}$ with the approximate sample obtained from the SIR algorithm. Finally, one can explore other two-sample test statistics beyond those listed in Section 4. Of particular interest is the block-wise MMD statistic (Zaremba et al., 2013). This statistic has a tractable limiting distribution, while achieving lower variance with a slight increase in computational cost compared to the linear-time MMD statistic. We leave all these interesting topics for future work.

Acknowledgments We acknowledge support from the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2022R1A4A1033384), and the Korea government (MSIT) RS-2023-00211073. We are grateful to Gyumin Lee for the careful proofreading.

(update the reference)

References

Andrews, D. W. K. (1997). A Conditional Kolmogorov Test. *Econometrica*, 65(5):1097–1128.

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. The MIT Press.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):175–197.

Boeken, P. A. and Mooij, J. M. (2021). A bayesian nonparametric conditional two-sample test with an application to local causal discovery. In de Campos, C. and Maathuis, M. H., editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1565–1575. PMLR.

Bordino, A. and Berrett, T. B. (2025). Density ratio permutation tests with connections to distributional shifts and conditional two-sample testing. *arXiv preprint arXiv:2505.24529*.

Candes, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):551–577.

Chakraborty, A., Zhang, J., and Katsevich, E. (2024). Doubly robust and computationally efficient high-dimensional variable selection. *arXiv preprint arXiv:2409.09512*.

Chatterjee, A., Niu, Z., and Bhattacharya, B. B. (2024). A kernel-based conditional two-sample test using nearest neighbors (with applications to calibration, regression curves, and simulation-based inference). *arXiv preprint arXiv:2407.16550*.

Chen, Y. and Lei, J. (2024). De-Biased Two-Sample U-Statistics With Application To Conditional Distribution Testing. *arXiv preprint arXiv:2402.00164*.

Choi, K., Liao, M., and Ermon, S. (2021). Featurized density ratio estimation. In *Uncertainty in Artificial Intelligence*, pages 172–182.

Choi, K., Meng, C., Song, Y., and Ermon, S. (2022). Density ratio estimation via infinitesimal classification. In *International Conference on Artificial Intelligence and Statistics*, pages 2552–2573.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507.

Dai, B., Shen, X., and Pan, W. (2022). Significance tests of feature relevance for a black-box learner. *IEEE transactions on neural networks and learning systems*, 35(2):1898–1911.

de la Peña, V. H. and Giné, E. (1999). *Decoupling: From Dependence to Independence*. Probability and Its Applications. Springer, New York.

de la Peña, V. H. and Montgomery-Smith, S. J. (1995). Decoupling inequalities for the tail probabilities of multivariate u-statistics. *The Annals of Probability*, 23(2):806–816.

Doran, G., Muandet, K., Zhang, K., and Schölkopf, B. (2014). A permutation-based kernel conditional independence test. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, page 132–141, Arlington, Virginia, USA. AUAI Press.

Fan, J. and Lin, S.-K. (1998). Test of significance when data are curves. *Journal of the American Statistical Association*, 93(443):1007–1021.

Fan, Y., Li, Q., and Min, I. (2006). A nonparametric bootstrap test of conditional distributions. *Econometric Theory*, 22(4):587–613.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2007). Kernel measures of conditional dependence. *Advances in Neural Information Processing Systems*, 20:489–496.

Giné, E., Latała, R., and Zinn, J. (2000). Exponential and moment inequalities for u-statistics. In *High Dimensional Probability II*, pages 13–38. Springer.

Givens, G. H. and Hoeting, J. A. (2012). *Computational statistics*. John Wiley & Sons, Hoboken, NJ, USA, 2 edition.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.

Hall, P. and Hart, J. D. (1990). Bootstrap test for difference between means in nonparametric regression. *Journal of the American Statistical Association*, 85(412):1039–1049.

Hamidieh, K. (2018). A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354.

Hardt, M., Price, E., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29, page 3323–3331. Curran Associates, Inc.

Hediger, S., Michel, L., and Näf, J. (2022). On the use of random forest for two-sample testing. *Computational Statistics & Data Analysis*, 170:107435.

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.

Kanamori, T., Hido, S., and Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445.

Kanamori, T., Suzuki, T., and Sugiyama, M. (2010). Theoretical analysis of density ratio estimation. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 93(4):787–798.

Kim, I., Balakrishnan, S., and Wasserman, L. (2022a). Minimax optimality of permutation tests. *The Annals of Statistics*, 50(1):225–251.

Kim, I., Lee, A. B., and Lei, J. (2019). Global and local two-sample tests via regression. *Electronic Journal of Statistics*, 13(2):5253–5305.

Kim, I., Neykov, M., Balakrishnan, S., and Wasserman, L. (2022b). Local permutation tests for conditional independence. *The Annals of Statistics*, 50(6):3388–3414.

Kim, I., Neykov, M., Balakrishnan, S., and Wasserman, L. (2023). Conditional Independence Testing for Discrete Distributions: Beyond χ^2-and G-tests. *arXiv preprint arXiv:2308.05373*.

Kim, I., Ramdas, A., Singh, A., and Wasserman, L. (2021). Classification accuracy as a proxy for two-sample testing. *The Annals of Statistics*, 49(1):411–434.

Kimura, M. and Hino, H. (2024). A short survey on importance weighting for machine learning. *arXiv preprint arXiv:2403.10175*.

Kulasekera, K. (1995). Comparison of regression curves using quasi-residuals. *Journal of the American Statistical Association*, 90(431):1085–1093.

Kulasekera, K. and Wang, J. (1997). Smoothing parameter selection for power optimality in testing of regression curves. *Journal of the American Statistical Association*, 92(438):500–511.

Lehmann, E. L. (2004). *Elements of Large-Sample Theory*. Springer Science & Business Media.

Li, S., Zhang, Y., Zhu, H., Wang, C., Shu, H., Chen, Z., Sun, Z., and Yang, Y. (2023). K-nearest-neighbor local sampling based conditional independence testing. *Advances in Neural Information Processing Systems*, 36:23321–23344.

Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. J. (2020). Learning deep kernels for non-parametric two-sample tests. In *International Conference on Machine Learning*, pages 6316–6326.

Liu, M., Katsevich, E., Janson, L., and Ramdas, A. (2022). Fast and powerful conditional randomization testing via distillation. *Biometrika*, 109(2):277–293.

Liu, S., Takeda, A., Suzuki, T., and Fukumizu, K. (2017). Trimmed density ratio estimation. *Advances in Neural Information Processing Systems*, 30:4521–4531.

Lopez-Paz, D. and Oquab, M. (2017). Revisiting Classifier Two-Sample Tests. In *International Conference on Learning Representations*.

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)*.

Mulzer, W. (2018). Five proofs of Chernoff's bound with applications. *arXiv preprint arXiv:1801.03365*.

Neumeyer, N. and Dette, H. (2003). Nonparametric comparison of regression curves: an empirical process approach. *The Annals of Statistics*, 31(3):880–920.

Neykov, M., Balakrishnan, S., and Wasserman, L. (2021). Minimax optimal conditional independence testing. *The Annals of Statistics*, 49(4):2151–2177.

Neykov, M., Wasserman, L., Kim, I., and Balakrishnan, S. (2023). Nearly Minimax Optimal Wasserstein Conditional Independence Testing. *arXiv preprint arXiv:2308.08672*.

Pardo-Fernández, J. C., Jiménez-Gamero, M. D., and El Ghouch, A. (2015). Tests for the equality of conditional variance functions in nonparametric regression. *Electronic Journal of Statistics*, 9(2).

Pogodin, R., Schrab, A., Li, Y., Sutherland, D. J., and Gretton, A. (2024). Practical Kernel Tests of Conditional Independence. *arXiv preprint arXiv:2402.13196*.

Rhodes, B., Xu, K., and Gutmann, M. U. (2020). Telescoping density-ratio estimation. *Advances in Neural Information Processing Systems*, 33:4905–4916.

Scheidegger, C., Hörrmann, J., and Bühlmann, P. (2022). The weighted generalised covariance measure. *Journal of Machine Learning Research*, 23(273):1–68.

Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. (2023). MMD aggregated two-sample test. *Journal of Machine Learning Research*, 24(194):1–81.

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.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244.

Strobl, E. V., Zhang, K., and Visweswaran, S. (2019). Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1):20180017.

Sugiyama, M., Krauledat, M., and Müller, K.-R. (2007a). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(35):985–1005.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P., and Kawanabe, M. (2007b). Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in Neural Information Processing Systems*, 20.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2010). Density ratio estimation: A comprehensive review. *RIMS Kokyuroku*, pages 10–31.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge University Press.

Tansey, W., Veitch, V., Zhang, H., Rabadan, R., and Blei, D. M. (2022). The holdout randomization test for feature selection in black box models. *Journal of Computational and Graphical Statistics*, 31(1):151–162.

Tsuboi, Y., Kashima, H., Hido, S., Bickel, S., and Sugiyama, M. (2009). Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155.

Virolainen, S. J., VonHandorf, A., Viel, K. C. M. F., Weirauch, M. T., and Kottyan, L. C. (2022). Gene–environment interactions and their impact on human health. *Genes & Immunity*, 24(1):1–11.

Williamson, B. D., Gilbert, P. B., Simon, N. R., and Carone, M. (2023). A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 118(543):1645–1658.

Wu, H., Eckhardt, C. M., and Baccarelli, A. A. (2023). Molecular mechanisms of environmental exposures and human disease. *Nature Reviews Genetics*, 24(5):332–344.

Yan, J. and Zhang, X. (2022). A nonparametric two-sample conditional distribution test. *arXiv preprint arXiv:2210.08149*.

Yanushkevichiene, O., Januškevičius, R., and Gesevičienė, V. (2016). On bounds for some u-statistics under the two-eigenvalues assumption. *Journal of Mathematical Sciences*, 214.

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.

Zhang, K., Huang, B., Zhang, J., Glymour, C., and Schölkopf, B. (2017). Causal discovery from nonstationary/heterogeneous data: skeleton estimation and orientation determination. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1347–1353.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-Based Conditional Independence Test and Application in Causal Discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 804–813, Arlington, Virginia, USA. AUAI Press.

Zheng, J. X. (2000). A consistent test of conditional parametric distributions. *Econometric Theory*, 16(5):667–691.

Zhu, J. and Hastie, T. (2005). Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, 14(1):185–205.

Overview of Appendices. In Appendix A, we present the proofs omitted in the main paper. Appendix B gathers several lemmas that support these proofs. Finally, Appendix C provides implementation details of numerical experiments and additional simulation results.

A Proofs

Notation. For real sequences (a_n) and (b_n), we say that $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n \leq Cb_n$ for all n. Let $(X_{P,n})_{n \in \mathbb{N}, P \in \mathcal{P}}$ be a family of sequences of random variables determined by $P \in \mathcal{P}$. We say that $X_{P,n} = X_n = o_{\mathcal{P}}(n^{-a})$ and $X_n = O_{\mathcal{P}}(n^{-a})$ to mean respectively for all $\epsilon > 0$,

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P(n^a |X_{P,n}| > \epsilon) \to 0, \quad \text{and}$$

$$\text{there exists } M > 0 \text{ such that } \sup_{n \in \mathbb{N}} \sup_{P \in \mathcal{P}} \mathbb{P}_P(n^a |X_{P,n}| > M) < \epsilon.$$

For a positive integer n, we use the shorthand $[n]$ to denote the set $\{1, \ldots, n\}$.

A.1 Proof of Theorem 3

We analyze the numerator and the denominator of $\widehat{\text{Acc}}$, separately. In particular, we first show that the numerator converges to a Gaussian distribution and the denominator is ratio-consistent to the population-level standard deviation under Assumption 1.

Analysis of the numerator. Starting with the numerator, let us rewrite

$$\begin{aligned}
\overline{A}_1 + \overline{A}_2 - 1 &= \frac{1}{m} \sum_{i=1}^{m} \big[\mathbb{1}\{\widehat{h}(V_i^{(1)}) = 1\} + r_X(X_i^{(2)}) \mathbb{1}\{\widehat{h}(V_i^{(2)}) = 2\} - 1\big] \\
&\quad + \frac{1}{m} \sum_{i=1}^{m} \{\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\} \mathbb{1}\{\widehat{h}(V_i^{(2)}) = 2\} \\
&= \frac{1}{m} \sum_{i=1}^{m} \underbrace{\big[\mathbb{1}\{\widehat{h}(V_i^{(1)}) = 1\} + r_X(X_i^{(2)}) \mathbb{1}\{\widehat{h}(V_i^{(2)}) = 2\} - 1\big]}_{:= L_i(\widehat{h})} + o_{\mathcal{P}_0}(m^{-1/2}),
\end{aligned}$$

where the last approximation holds since

$$\left| \frac{1}{m} \sum_{i=1}^{m} \{\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\} \mathbb{1}\{\widehat{h}(V_i^{(2)}) = 2\} \right| \leq \sqrt{\frac{1}{m} \sum_{i=1}^{m} \{\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\}^2},$$

and the upper bound is $o_{\mathcal{P}_0}(m^{-1/2})$ due to Assumption 1(b). Thus $\overline{A}_1 + \overline{A}_2 - 1$ is dominated by the average of $L_i(\widehat{h})$ values. Given this and Slutsky's theorem, it suffices to study the limiting distribution of the sample average of $L_i(\widehat{h})$. Indeed, under Assumption 1(a), the conditional central limit theorem (Lemma 3) yields that

$$\lim_{n \to \infty} \sup_{P \in \mathcal{P}_0} \sup_{t \in \mathbb{R}} \left| \mathbb{P}_P\left(\frac{\frac{1}{\sqrt{m}} \sum_{i=1}^{m} L_i(\widehat{h})}{\{\text{Var}[L(\widehat{h}) \,|\, \widehat{h}]\}^{1/2}} \leq t \right) - \Phi(t) \right| = 0.$$

Consistency of the variance estimate. We next show the ratio-consistency of the variance estimator. Observe that

$$\text{Var}[L(\widehat{h}) \,|\, \widehat{h}] = \underbrace{\text{Var}[\mathbb{1}\{\widehat{h}(V_i^{(1)}) = 1\} \,|\, \widehat{h}]}_{:= \sigma_1^2} + \underbrace{\text{Var}\big[r_X(X_i^{(2)}) \mathbb{1}\{\widehat{h}(V_i^{(2)}) = 2\}\big]}_{:= \sigma_2^2},$$

and

$$\left| \frac{\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2}{\sigma_1^2 + \sigma_2^2} - 1 \right| \leq \left| \frac{\widehat{\sigma}_1^2 - \sigma_1^2}{\sigma_1^2} \right| + \left| \frac{\widehat{\sigma}_2^2 - \sigma_2^2}{\sigma_2^2} \right|.$$

Therefore, in order to show the ratio consistency of $\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2$, it suffices to show the ratio consistency of $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$, individually. To this end, we use conditional Chebyshev's inequality and show

$$\sup_{P \in \mathcal{P}_0} \mathbb{P}_P\big(|\widehat{\sigma}_1^2/\sigma_1^2 - 1| \geq t \,|\, \widehat{h}\big) \leq \frac{1}{t^2} \sup_{P \in \mathcal{P}_0} \mathrm{Var}_P(\widehat{\sigma}_1^2/\sigma_1^2 \,|\, \widehat{h}) \leq \frac{1}{t^2 c_1(1-c_2)m},$$

for sufficiently large n and for all $t > 0$, under Assumption 1(a). Hence $\widehat{\sigma}_1^2/\sigma_1^2$ converges to one in probability uniformly over $\mathcal{P}_0$. On the other hand, letting $A_{2,i} \coloneqq \widehat{r}_X(X_i^{(2)})\mathbb{1}\{\widehat{h}(V_i^{(2)}) = 2\}$, we have

$$\begin{aligned}
\widehat{\sigma}_2^2 &= \frac{1}{m-1}\sum_{i=1}^m \bigg[A_{2,i} - \frac{1}{m}\sum_{j=1}^m A_{2,j}\bigg]^2 \\
&+ \frac{1}{m-1}\sum_{i=1}^m \bigg[\big(\widehat{A}_{2,i} - A_{2,i}\big) - \frac{1}{m}\sum_{j=1}^m \big(\widehat{A}_{2,j} - A_{2,j}\big)\bigg]^2 \\
&+ \frac{2}{m-1}\sum_{i=1}^m \bigg[A_{2,i} - \frac{1}{m}\sum_{j=1}^m A_{2,j}\bigg] \cdot \bigg[\big(\widehat{A}_{2,i} - A_{2,i}\big) - \frac{1}{m}\sum_{j=1}^m \big(\widehat{A}_{2,j} - A_{2,j}\big)\bigg] \\
&\coloneqq (\mathbb{I}) + (\mathbb{II}) + (\mathbb{III}).
\end{aligned}$$

Similarly as before, the term $(\mathbb{I})/\sigma_2^2$ converges to one in probability uniformly over $\mathcal{P}_0$ under Assumption 1(a). It can be further shown that the term $(\mathbb{II})/\sigma_2^2$ is $o_{\mathcal{P}_0}(1)$ by Markov's inequality combined with Assumption 1(b). Lastly, the term $(\mathbb{III})$ satisfies $(\mathbb{III}) \leq \sqrt{(\mathbb{I}) \times (\mathbb{II})}$, which is again $o_{\mathcal{P}_0}(1)$. Therefore, $\widehat{\sigma}_2^2/\sigma_2^2$ converges to one in probability uniformly over $\mathcal{P}_0$. This further proves that $\sqrt{(\sigma_1^2 + \sigma_2^2)/(\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2)} = 1 + o_{\mathcal{P}_0}(1)$ by Lundborg et al. (2022, Lemma S7).

Putting all pieces together with Lemma 4(b) proves the claim.

A.2 Proof of Corollary 1

For each $j \in \{1, \ldots, K\}$, the proof of Theorem 3 shows that $\sqrt{\sigma_1^2 + \sigma_2^2}/\sqrt{\widehat{\sigma}_{1,j}^2 + \widehat{\sigma}_{2,j}^2} = 1 + o_{\mathcal{P}_0}(1)$. Thus, by Lemma 4(b), it is enough to show the asymptotic normality of

$$\frac{1}{\sqrt{K}}\sum_{j=1}^K \frac{\sqrt{m}(\overline{A}_{1,j} + \overline{A}_{2,j} - 1)}{\sqrt{\sigma_1^2 + \sigma_2^2}}.$$

Without loss of generality, denote the sample indices of $D_1, D_2, \ldots, D_K$ as

$$I_1 = \big\{1, \ldots, m\big\},\ I_2 = \big\{m+1, \ldots,\ 2m\big\}, \ldots, I_K = \big\{m(K-1)+1, \ldots, mK\big\}.$$

Then the proof of Theorem 3 establishes that

$$\frac{1}{\sqrt{K}}\sum_{j=1}^K \frac{\sqrt{m}(\overline{A}_{1,j} + \overline{A}_{2,j} - 1)}{\sqrt{\sigma_1^2 + \sigma_2^2}} = \sqrt{\frac{1}{mK}}\sum_{j=1}^K \underbrace{\Bigg\{\sum_{i \in I_j} \frac{A_{1,i} + A_{2,i} - 1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\Bigg\}}_{\coloneqq B_j} + o_{\mathcal{P}_0}(1),$$

where $A_{1,i} + A_{2,i} - 1 \coloneqq \mathbb{1}\{\widehat{h}(V_i^{(1)}) = 1\} + r_X(X_i^{(2)})\mathbb{1}\{\widehat{h}(V_i^{(2)}) = 2\} - 1$. Note that $B_1, \ldots, B_K$ are mutually independent conditional on $\widehat{h}$. As in the proof of Theorem 3, we apply the conditional central limit theorem (Lemma 3) to the average of $B_1, \ldots, B_K$ conditional on $\widehat{h}$ under Assumption 1, which completes the proof of Corollary 1.

A.3 Proof of Theorem 4

The proof consists of two parts as in the proof of Theorem 3. In the first part, we investigate the numerator of $\widehat{\mathrm{MMD}}_\ell^2$, i.e., $\overline{S}$, whereas in the second part, we show the consistency of the denominator to the population variance under Assumption 2. The proof is then completed by applying Lemma 4(b).

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 $\overline{S}$ as

$$\overline{S} = \frac{1}{m}\sum_{i=1}^{m} \big\langle \psi\big(V_i^{(1)}\big) - \widehat{r}_X\big(X_i^{(2)}\big)\psi\big(V_i^{(2)}\big),\ \psi\big(V_{i+m}^{(1)}\big) - \widehat{r}_X\big(X_{i+m}^{(2)}\big)\psi\big(V_{i+m}^{(2)}\big)\big\rangle.$$

By adding and subtracting $r_X(X_i^{(2)})\psi(V_i^{(2)})$ and $r_X(X_{i+m}^{(2)})\psi(V_{i+m}^{(2)})$, $\overline{S}$ can be written as the sum of the four terms given as:

$$(\mathbb{I}) \coloneqq \frac{1}{m}\sum_{i=1}^{m} \underbrace{\big\langle \psi\big(V_i^{(1)}\big) - r_X\big(X_i^{(2)}\big)\psi\big(V_i^{(2)}\big), \psi\big(V_{i+m}^{(1)}\big) - r_X\big(X_{i+m}^{(2)}\big)\psi\big(V_{i+m}^{(2)}\big)\big\rangle}_{\coloneqq S_i},$$

$$(\mathbb{II}) \coloneqq \frac{1}{m}\sum_{i=1}^{m} \underbrace{\big\langle \psi\big(V_i^{(1)}\big) - r_X\big(X_i^{(2)}\big)\psi\big(V_i^{(2)}\big), \psi\big(V_{i+m}^{(2)}\big)\big\rangle \cdot \big\{\widehat{r}_X\big(X_{i+m}^{(2)}\big) - r_X\big(X_{i+m}^{(2)}\big)\big\}}_{\coloneqq \widehat{S}_{i,a}},$$

$$(\mathbb{III}) \coloneqq \frac{1}{m}\sum_{i=1}^{m} \underbrace{\big\langle \psi\big(V_i^{(2)}\big), \psi\big(V_{i+m}^{(1)}\big) - r_X\big(X_{i+m}^{(2)}\big)\psi\big(V_{i+m}^{(2)}\big)\big\rangle \cdot \big\{\widehat{r}_X\big(X_i^{(2)}\big) - r_X\big(X_i^{(2)}\big)\big\}}_{\coloneqq \widehat{S}_{i,b}},$$

$$(\mathbb{IV}) \coloneqq \frac{1}{m}\sum_{i=1}^{m} \underbrace{\big\langle \psi\big(V_i^{(2)}\big), \psi\big(V_{i+m}^{(2)}\big)\big\rangle \big\{\widehat{r}_X\big(X_i^{(2)}\big) - r_X\big(X_i^{(2)}\big)\big\} \cdot \big\{\widehat{r}_X\big(X_{i+m}^{(2)}\big) - r_X\big(X_{i+m}^{(2)}\big)\big\}}_{\coloneqq \widehat{S}_{i,c}}.$$

The first term $(\mathbb{I})$ does not involve an estimate of the density ratio and will be asymptotically Gaussian since it is the sum of i.i.d. random variables under the null hypothesis. The other terms $(\mathbb{II})$, $(\mathbb{III})$, and $(\mathbb{IV})$ are asymptotically negligible under the conditions of the theorem. Hence $\overline{S}$ will be dominated by $(\mathbb{I})$. Let us analyze each term separately.

1. **Term $(\mathbb{I})$.** Define $\mathrm{Var}_P[S_1] = \sigma_P^2$. Then under Assumption 2(a), Lemma 2 yields

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

2. **Terms $(\mathbb{II})$ and $(\mathbb{III})$.** We only analyze the term $(\mathbb{II})$ since $(\mathbb{III})$ can be handled in exactly the same way by symmetry. Under the null hypothesis, by the law of total expectation, it can be seen that the expectation of the summands of $(\mathbb{II})$ is equal to zero:

$$\mathbb{E}\big[\big\langle \psi\big(V_i^{(1)}\big) - r_X\big(X_i^{(2)}\big)\psi\big(V_i^{(2)}\big), \psi\big(V_{i+m}^{(2)}\big)\big\rangle \cdot \big\{\widehat{r}_X\big(X_{i+m}^{(2)}\big) - r_X\big(X_{i+m}^{(2)}\big)\big\}\big] = 0,$$

which leads to $\mathbb{E}[(\mathbb{II})] = 0$. On the other hand, the conditional second moment (or the conditional variance) of $(\mathbb{II})$ given D_b satisfies

$$\begin{aligned}\mathbb{E}\big[(\mathbb{II})^2 \,|\, D_b\big] &= \frac{1}{m}\mathbb{E}\big[\big\langle \psi\big(V_1^{(1)}\big) - r_X\big(X_1^{(2)}\big)\psi\big(V_1^{(2)}\big), \psi\big(V_{1+m}^{(2)}\big)\big\rangle^2 \cdot \big\{\widehat{r}_X\big(X_{1+m}^{(2)}\big) - r_X\big(X_{1+m}^{(2)}\big)\big\}^2 \,|\, D_b\big] \\ &\lesssim \frac{K^2}{m}\big(1 + \mathbb{E}\big[r_X\big(X_1^{(2)}\big)^2\big]\big) \cdot \mathbb{E}\big[\big\{\widehat{r}_X\big(X_{1+m}^{(2)}\big) - r_X\big(X_{1+m}^{(2)}\big)\big\}^2 \,|\, D_b\big],\end{aligned}$$

where we use the fact that $\langle \psi(x), \psi(y) \rangle = k(x,y)$, whose ℓ_∞ norm is uniformly bounded by the constant K. Therefore, under the condition that

$$\sup_{P\in\mathcal{P}_0} \mathbb{E}_P\big[\big\{\widehat{r}_X\big(X_{1+m}^{(2)}\big) - r_X\big(X_{1+m}^{(2)}\big)\big\}^2\big] = o(m^{-1/2}) \quad \text{and} \quad \sup_{P\in\mathcal{P}_0} \mathbb{E}_P\big[r_X\big(X_1^{(2)}\big)^2\big] < \infty,$$

Chebyshev's inequality yields $(\mathbb{II}) = o_{\mathcal{P}_0}(m^{-1/2})$ and similarly $(\mathbb{III}) = o_{\mathcal{P}_0}(m^{-1/2})$.

3. **Term** ($\mathbb{IV}$)**.** The fourth term ($\mathbb{IV}$) can be written as

$$(\mathbb{IV}) = \frac{1}{m}\sum_{i=1}^{m} k\big(V_i^{(2)}, V_{i+m}^{(2)}\big)\big\{\widehat{r}_X\big(X_i^{(2)}\big) - r_X\big(X_i^{(2)}\big)\big\} \cdot \big\{\widehat{r}_X\big(X_{i+m}^{(2)}\big) - r_X\big(X_{i+m}^{(2)}\big)\big\}.$$

Since the kernel is uniformly bounded and by the Cauchy–Schwarz inequality, we have

$$\begin{aligned}|(\mathbb{IV})| &\leq K\left(\frac{1}{m}\sum_{i=1}^{m}\big\{\widehat{r}_X\big(X_i^{(2)}\big) - r_X\big(X_i^{(2)}\big)\big\}^2\right)^{1/2} \cdot \left(\frac{1}{m}\sum_{i=1}^{m}\big\{\widehat{r}_X\big(X_{i+m}^{(2)}\big) - r_X\big(X_{i+m}^{(2)}\big)\big\}^2\right)^{1/2} \\ &\overset{(\star)}{=} o_{\mathcal{P}_0}(m^{-1/4})o_{\mathcal{P}_0}(m^{-1/4}) = o_{\mathcal{P}_0}(m^{-1/2}),\end{aligned}$$

which follows by Markov's inequality along with the condition that

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

and step $(\star)$ holds by Lundborg et al. (2022, Lemma S5). Therefore it holds that $(\mathbb{IV}) = o_{\mathcal{P}_0}(m^{-1/2})$.

Now combining the results establishes that

$$\sup_{P\in\mathcal{P}_0}\sup_{t\in\mathbb{R}} |\mathbb{P}_P(\sqrt{m}\sigma_P^{-1}\overline{S} \leq t) - \Phi(t)| \to 0.$$

Consistency of the variance estimate. Denoting

$$\widehat{\sigma}_P^2 := \frac{1}{m-1}\sum_{i=1}^{m}(\widehat{S}_i - \overline{S})^2,$$

we would like to show that $\widehat{\sigma}_P^2/\sigma_P^2$ converges to one in probability, which further implies $\sigma_P/\widehat{\sigma}_P = 1 + o_{\mathcal{P}_0}(1)$ by Lundborg et al. (2022, Lemma S7). Since the test statistic $\widehat{\mathrm{MMD}}_\ell^2$ is scale-invariant, we may assume that $\sigma_P^2 = 1$ without loss of generality. Moreover, the preceding analysis ensures that $\overline{S} = o_{\mathcal{P}_0}(1)$. Therefore we only need to show $\frac{1}{m}\sum_{i=1}^{m}\widehat{S}_i^2$ converges to one in probability. To this end, observe

$$\begin{aligned}\left|\frac{1}{m}\sum_{i=1}^{m}\widehat{S}_i^2 - 1\right| &= \left|\frac{1}{m}\sum_{i=1}^{m}\big(S_i + \widehat{S}_{i,a} + \widehat{S}_{i.b} + \widehat{S}_{i,c}\big)^2 - 1\right| \\ &\leq \left|\frac{1}{m}\sum_{i=1}^{m}S_i^2 - 1\right| + \left|\frac{1}{m}\sum_{i=1}^{m}\big(\widehat{S}_{i,a} + \widehat{S}_{i.b} + \widehat{S}_{i,c}\big)^2\right| + 2\left|\frac{1}{m}\sum_{i=1}^{m}S_i\big(\widehat{S}_{i,a} + \widehat{S}_{i.b} + \widehat{S}_{i,c}\big)\right| \\ &\leq \left|\frac{1}{m}\sum_{i=1}^{m}S_i^2 - 1\right| + \left|\frac{1}{m}\sum_{i=1}^{m}\big(\widehat{S}_{i,a} + \widehat{S}_{i.b} + \widehat{S}_{i,c}\big)^2\right| \\ &\qquad + 2\sqrt{\frac{1}{m}\sum_{i=1}^{m}S_i^2}\sqrt{\frac{1}{m}\sum_{i=1}^{m}\big(\widehat{S}_{i,a} + \widehat{S}_{i.b} + \widehat{S}_{i,c}\big)^2},\end{aligned}$$

where the last inequality follows by the Cauchy–Schwarz inequality. By the law of large numbers, $\frac{1}{m}\sum_{i=1}^{m}S_i^2$ converges to one in probability. Thus the proof amounts to showing that $\frac{1}{m}\sum_{i=1}^{m}\big(\widehat{S}_{i,a}+\widehat{S}_{i.b}+\widehat{S}_{i,c}\big)^2 = o_{\mathcal{P}_0}(1)$, which is implied by

$$\frac{1}{m}\sum_{i=1}^{m}\widehat{S}_{i,a}^2 = o_{\mathcal{P}_0}(1), \quad \frac{1}{m}\sum_{i=1}^{m}\widehat{S}_{i,b}^2 = o_{\mathcal{P}_0}(1) \quad \text{and} \quad \frac{1}{m}\sum_{i=1}^{m}\widehat{S}_{i,c}^2 = o_{\mathcal{P}_0}(1).$$

This can be done as the way how ($\mathbb{II}$), ($\mathbb{III}$), and ($\mathbb{IV}$) are handled earlier along with Markov's inequality. This completes the proof.

A.4 Proof of Corollary 2

For each $j \in \{1, \dots, K\}$, the proof of Theorem 4 shows that $\sigma_P/\widehat{\sigma}_j = 1 + o_{\mathcal{P}_0}(1)$. Thus, by Lemma 4(b), it is enough to show the asymptotic normality of

$$\frac{1}{K}\sum_{j=1}^{K}\frac{\sqrt{n}\overline{S}_j}{\sigma}.$$

Without loss of generality, denote the sample indices of $D_1, D_2, \dots, D_K$ as

$$I_1 = \left\{1, \dots, \frac{2n}{K}\right\},\ I_2 = \left\{\frac{2n}{K}+1, \dots,\ \frac{4n}{K}\right\}, \dots, I_K = \left\{2n - \frac{2n}{K}+1, \dots, 2n\right\},$$

and let the first n/K elements of I_j as I_j' (e.g., $I_1' = \{1, \dots, n/K\}$). Then with $m' = n/K$, the proof of Theorem 4 establishes that

$$\sum_{j=1}^{K}\overline{S}_j = \frac{K}{n}\sum_{j=1}^{K}\sum_{i \in I_j'} S_i + o_{\mathcal{P}_0}(n^{-1/2}),$$

where

$$S_i = \big\langle \psi\big(V_i^{(1)}\big) - r_X\big(X_i^{(2)}\big)\psi\big(V_i^{(2)}\big), \psi\big(V_{i+m'}^{(1)}\big) - r_X\big(X_{i+m'}^{(2)}\big)\psi\big(V_{i+m'}^{(2)}\big)\big\rangle.$$

Notably, $\sum_{i \in I_1'} S_i, \dots, \sum_{i \in I_K'} S_i$ are mutually independent. Hence

$$\mathrm{Var}\left(\frac{K}{n}\sum_{j=1}^{K}\sum_{i \in I_j'}\sigma_P^{-1}S_i\right) = \frac{K^2}{n^2}\cdot K \cdot \frac{n}{K} = \frac{K^2}{n}.$$

By the central limit theorem (Lemma 2),

$$\frac{1}{K}\sum_{j=1}^{K}\sum_{i \in I_j'}\frac{\sqrt{n}S_i}{\sigma_P} = \frac{1}{K}\sum_{j=1}^{K}\frac{\sqrt{n}\overline{S}_j}{\sigma} + o_{\mathcal{P}_0}(1)$$

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

A.5 Proof of Theorem 5

(To make unified version) The proof consists of two parts as in the proof of Theorem 3. In the first part, we investigate the numerator of $\widehat{\mathrm{MMD}}_B^2$, i.e., $\bar{\eta}$, whereas in the second part, we show the consistency of the denominator to the population variance under Assumption 3. The proof is then completed by applying Lemma 4(b).

Analysis of the numerator. All inner products and norms below are taken with respect to the reproducing kernel Hilbert space $\mathcal{H}_k$, but for notational simplicity we omit explicit reference to $\mathcal{H}_k$. Using the feature map representation $k(x,y) = \langle \psi(x), \psi(y)\rangle$, the block-averaged statistic $\bar{\eta}$ can be written as

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

For each block $b = 1, \dots, S$, define

$$V_{B,(b)} := \left\|\frac{1}{B}\sum_{i=(b-1)B+1}^{bB}\psi(V_i^{(1)}) - \frac{1}{B}\sum_{i=(b-1)B+1}^{bB}\widehat{r}_X(X_i^{(2)})\,\psi(V_i^{(2)})\right\|,$$

which is the Hilbert–space norm of the block-wise difference between the block-wise average embedding of the first sample and the reweighted block-wise average embedding of the second sample. With this notation, Recall that

$$\bar{\eta} = \frac{1}{S}\sum_{b=1}^{S}\hat{\eta}_b, \qquad \hat{\eta}_b := \Big(\underbrace{B\,V_{B,(b)}^2 - \tfrac{1}{B}\sum_{i=(b-1)B+1}^{bB}\widehat{H}_{ii}}_{=:T_b}\Big)\frac{B}{B-1},$$

so that each block contribution $\hat{\eta}_b$ is expressed in terms of T_b up to the factor $\frac{B}{B-1}$. Each block term $V_{B,(b)}$ admits the decomposition

$$\begin{aligned}\sqrt{B}\,V_{B,(b)} &= \Bigg\|\underbrace{\frac{1}{\sqrt{B}}\sum_{i=(b-1)B+1}^{bB}\big\{\psi(V_i^{(1)}) - r_X(X_i^{(2)})\,\psi(V_i^{(2)})\big\}}_{=:(\mathbb{I})_b} + \underbrace{\frac{1}{\sqrt{B}}\sum_{i=(b-1)B+1}^{bB}\psi(V_i^{(2)})\{r_X(X_i^{(2)}) - \widehat{r}_X(X_i^{(2)})\}}_{=:(\mathbb{II})_b}\Bigg\| \\ &= \sqrt{\|(\mathbb{I})_b\|^2 + \|(\mathbb{II})_b\|^2 + 2\langle(\mathbb{I})_b,(\mathbb{II})_b\rangle}.\end{aligned}$$

By construction of $\widehat{H}$ and H, the diagonal correction can be written as

$$\begin{aligned}\frac{1}{B}\sum_{i=(b-1)B+1}^{bB}\widehat{H}_{ii} = \frac{1}{B}&\sum_{i=(b-1)B+1}^{bB}H_{ii} \\ &+ \frac{1}{B}\sum_{i=(b-1)B+1}^{bB}\big(\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\big)^2\,k(V_i^{(2)},V_i^{(2)}) \\ &- \frac{2}{B}\sum_{i=(b-1)B+1}^{bB}\big(\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\big)\,k(V_i^{(1)},V_i^{(2)}).\end{aligned}$$

Combining these expressions, T_b admits the decomposition

$$\begin{aligned}T_b &= \|(\mathbb{I})_b\|^2 - \frac{1}{B}\sum_{i=(b-1)B+1}^{bB}H_{ii} \\ &\quad + \|(\mathbb{II})_b\|^2 + 2\langle(\mathbb{I})_b,(\mathbb{II})_b\rangle \\ &\quad - \frac{1}{B}\sum_{i=(b-1)B+1}^{bB}\{\,\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\,\}^2 k(V_i^{(2)},V_i^{(2)}) \\ &\quad + \frac{2}{B}\sum_{i=(b-1)B+1}^{bB}\{\,\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\,\}\,k(V_i^{(1)},V_i^{(2)}) \\ &=: T_{1,b} + T_{2,b} + T_{3,b} + T_{4,b}.\end{aligned}$$

For convenience, we define block-averaged quantities

$$T_j := \frac{1}{\sqrt{S}}\sum_{b=1}^{S}T_{j,b}, \quad j \in \{1,2,3,4\},$$

so that

$$\sqrt{S}\,\bar{\eta} = \frac{B}{B-1}\cdot(T_1 + T_2 + T_3 + T_4).$$

Since $\hat{\eta}_b = T_b \cdot \frac{B}{B-1}$, the difference $\hat{\eta}_b - T_b = \frac{T_b}{B-1}$ is $o_{\mathcal{P}_0}(1)$ as $B \to \infty$, provided that $T_b = O_{\mathcal{P}_0}(1)$ (as will be shown below). Thus $\hat{\eta}_b$ and T_b are asymptotically equivalent, and it suffices to analyze T_b. In particular, establishing that $T_2, T_3, T_4 = o_{\mathcal{P}_0}(1)$ will imply

$$\sqrt{S}\,\bar{\eta} = T_1 + o_{\mathcal{P}_0}(1),$$

so that the asymptotic distribution of $\sqrt{S}\,\bar{\eta}$ is fully determined by the leading term T_1.

1. **Term T_1.** Recall that

$$\begin{aligned} T_{1,b} &= \|(\mathbb{I})_b\|^2 - \frac{1}{B}\sum_{i=1}^{B} H_{ii} \\ &= (B-1) \cdot \frac{2}{B(B-1)} \sum_{(i,j)\in I_b} H_{ij} \end{aligned}$$

Hence $T_{1,b}$ coincides with the order-two symmetric U-statistic based on the kernel H, scaled by a factor of $(B-1)$. Under the null, we have $\mathbb{E}_P[T_{1,1}] = 0$. Standard U-statistic theory then gives

$$\mathbb{E}_P[T_{1,1}^2] = \frac{2(B-1)}{B}\,\mathbb{E}_P[H_{12}^2].$$

For $B \geq 2$, the prefactor satisfies $(B-1)/B \geq 1/2$. Hence, by Assumption 3(a),

$$\inf_{P\in\mathcal{P}_0} \mathbb{E}_P[T_{1,1}^2] \;\geq\; c.$$

This provides a uniform positive lower bound on the variance of $T_{1,1}$, independent of B.

For higher moments, a direct expansion of $T_{1,1}$ would introduce explicit dependence on B. Instead, by Lemma 6, the decoupled statistic

$$T_{1,1}^{\mathrm{dec}} = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B} H(Z_i^{(1)}, Z_j^{(2)})$$

satisfies

$$\mathbb{E}\big[|T_{1,1}^{\mathrm{dec}}|^{2+\delta}\big] \;\leq\; C_\delta\,\mathbb{E}[|H_{12}|^{2+\delta}],$$

where C_δ is independent of B. Classical decoupling inequalities (see de la Peña and Montgomery-Smith, 1995; de la Peña and Giné, 1999) then ensure that the same bound holds for the original U-statistic $T_{1,1}$, up to a universal constant. By Assumption 3(a), there exist constants $c', C' > 0$, depending only on (c, C, δ), such that

$$\inf_{P\in\mathcal{P}_0} \mathbb{E}_P[T_{1,1}^2] \geq c', \qquad \sup_{P\in\mathcal{P}_0} \mathbb{E}_P[|T_{1,1}|^{2+\delta}] \leq C'.$$

Defining $\mathrm{Var}_P[T_{1,1}] = \sigma_P^2$, Lemma 2 applies and yields

$$\lim_{N\wedge M\to\infty}\ \sup_{P\in\mathcal{P}_0}\sup_{t\in\mathbb{R}} \Big|\mathbb{P}_P\big(\sigma_P^{-1} T_1 \leq t\big) - \Phi(t)\Big| = 0.$$

Since the number of blocks $S = \lfloor M/B \rfloor$ diverges as $N \wedge M \to \infty$, the central limit theorem follows.

2. **Term T_2.** Recall that

$$T_2 = \frac{1}{\sqrt{S}}\sum_{b=1}^{S} T_{2,b} = \frac{1}{\sqrt{S}}\sum_{b=1}^{S} \|(\mathbb{III})_b\|^2 + \frac{2}{\sqrt{S}}\sum_{b=1}^{S}\langle(\mathbb{I})_b, (\mathbb{III})_b\rangle.$$

By the Cauchy–Schwarz inequality,

$$\left|\frac{1}{\sqrt{S}}\sum_{b=1}^{S}\langle(\mathbb{I})_b,(\mathbb{III})_b\rangle\right| \le \sqrt{S}\left(\frac{1}{S}\sum_{b=1}^{S}\|(\mathbb{I})_b\|^2\right)^{1/2}\left(\frac{1}{S}\sum_{b=1}^{S}\|(\mathbb{III})_b\|^2\right)^{1/2}.$$

Since $\|(\mathbb{I})_b\|^2$ is centered under the null hypothesis (i.e. $\mathbb{E}[\|(\mathbb{I})_b\|^2]=0$) and $\|(\mathbb{I})_b\|^2$ has finite variance under Assumption 3(a), Chebyshev's inequality implies

$$\frac{1}{S}\sum_{b=1}^{S}\|(\mathbb{I})_b\|^2 = o_{\mathcal{P}_0}(S^{-1/2}).$$

For each block,

$$(\mathbb{III})_b = \frac{1}{\sqrt{B}}\sum_{i=(b-1)B+1}^{bB}\psi(V_i^{(2)})\left\{r_X(X_i^{(2)})-\hat{r}_X(X_i^{(2)})\right\},$$

and by Cauchy–Schwarz,

$$\|(\mathbb{III})_b\|_{\mathcal{H}_k}^2 \le K^2\cdot B\cdot\left(\frac{1}{B}\sum_{i=(b-1)B+1}^{bB}\{r_X(X_i^{(2)})-\hat{r}_X(X_i^{(2)})\}^2\right),$$

where we use the fact that $\langle\psi(x),\psi(y)\rangle = k(x,y)$ with $\|k\|_\infty \le K$. Moreover, by Assumption 3(b),

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

It then follows that the cross term satisfies

$$\begin{aligned}\left|\frac{1}{\sqrt{S}}\sum_{b=1}^{S}\langle(\mathbb{I})_b,(\mathbb{III})_b\rangle\right| &\overset{(i)}{=} o_{\mathcal{P}_0}(S^{1/4})\cdot o_{\mathcal{P}_0}(B^{1/2}N^{-1/4})\\ &= o_{\mathcal{P}_0}\left(\frac{M^{(1+\gamma)/4}}{N^{1/4}}\right),\end{aligned}$$

where (i) follows from Lundborg et al. (2022, Lemma S5) together with Markov's inequality. Similarly, for the quadratic term,

$$\frac{1}{\sqrt{S}}\sum_{b=1}^{S}\|(\mathbb{III})_b\|_{\mathcal{H}_k}^2 = o_{\mathcal{P}_0}\left(\frac{M^{(1+\gamma)/2}}{N^{1/2}}\right).$$

Moreover, under Assumption 3(d), the block count satisfies $S=\lfloor M/B\rfloor$, and hence Markov's inequality implies

$$T_2 = o_{\mathcal{P}_0}\left(\frac{M^{(1+\gamma)/4}}{N^{1/4}}\right).$$

Finally, Assumption 3(e) guarantees that $T_2 = o_{\mathcal{P}_0}(1)$.

3. **Terms T_3 and T_4.** Recall that

$$T_3 = -\frac{1}{\sqrt{S}}\sum_{b=1}^{S}\frac{1}{B}\sum_{i=(b-1)B+1}^{bB}\left\{\widehat{r}_X(X_i^{(2)})-r_X(X_i^{(2)})\right\}^2 k(V_i^{(2)},V_i^{(2)}),$$

$$T_4 = \frac{1}{\sqrt{S}}\sum_{b=1}^{S}\frac{2}{B}\sum_{i=(b-1)B+1}^{bB}\left\{\widehat{r}_X(X_i^{(2)})-r_X(X_i^{(2)})\right\}k(V_i^{(1)},V_i^{(2)}).$$

Since the kernel is uniformly bounded, each block contribution satisfies

$$|T_{3,b}| \leq K \cdot \frac{1}{B} \sum_{i=(b-1)B+1}^{bB} \big\{\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\big\}^2,$$
$$|T_{4,b}| \leq K \cdot \Big(\frac{1}{B} \sum_{i=(b-1)B+1}^{bB} \big\{\widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)})\big\}^2\Big)^{1/2}.$$

By Markov's inequality and the condition

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

we obtain

$$T_3 = o_{\mathcal{P}_0}\left(\frac{\sqrt{S}}{N^{1/2}}\right), \qquad T_4 = o_{\mathcal{P}_0}\left(\frac{\sqrt{S}}{N^{1/4}}\right).$$

Finally, by Assumption 3(e), both T_3 and T_4 are asymptotic negligible, i.e.

$$T_3 = o_{\mathcal{P}_0}(1), \qquad T_4 = o_{\mathcal{P}_0}(1).$$

Combining the bounds for all four terms, we conclude that

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

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 $\hat{\sigma}_P/\sigma_P = 1+o_{\mathcal{P}_0}(1)$ 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 $\frac{1}{S}\sum_{b=1}^{S}\hat{\eta}_b^2$ converges to one in probability. Recalling the block-level definition $\hat{\eta}_b = \frac{B}{B-1}T_b$, this reduces to

$$\frac{1}{S}\sum_{b=1}^{S}\hat{\eta}_b^2 = \Big(\frac{B}{B-1}\Big)^2 \cdot \frac{1}{S}\sum_{b=1}^{S} T_b^2.$$

so it suffices to show that $\frac{1}{S}\sum_{b=1}^{S} T_b^2$ converges to one in probability. To establish this, note that:

$$\begin{aligned}
\left|\frac{1}{S}\sum_{b=1}^{S} T_b^2 - 1\right| &= \left|\frac{1}{S}\sum_{b=1}^{S}(T_{1,b}+T_{2,b}+T_{3,b}+T_{4,b})^2 - 1\right| \\
&\leq \left|\frac{1}{S}\sum_{b=1}^{S} T_{1,b}^2 - 1\right| + \left|\frac{1}{S}\sum_{b=1}^{S}(T_{2,b}+T_{3,b}+T_{4,b})^2\right| + 2\left|\frac{1}{S}\sum_{b=1}^{S} T_{1,b}(T_{2,b}+T_{3,b}+T_{4,b})\right| \\
&\leq \left|\frac{1}{S}\sum_{b=1}^{S} T_{1,b}^2 - 1\right| + \left|\frac{1}{S}\sum_{b=1}^{S}(T_{2,b}+T_{3,b}+T_{4,b})^2\right| \\
&\qquad + 2\sqrt{\frac{1}{S}\sum_{b=1}^{S} T_{1,b}^2}\sqrt{\frac{1}{S}\sum_{b=1}^{S}(T_{2,b}+T_{3,b}+T_{4,b})^2},
\end{aligned}$$

where the last inequality follows from the Cauchy–Schwarz inequality. Applying the law of large numbers, we see that $\frac{1}{S}\sum_{b=1}^{S} T_{1,b}^2$ converges to one in probability. Consequently, the proof reduces to establishing that

$$\frac{1}{S}\sum_{b=1}^{S}(T_{2,b}+T_{3,b}+T_{4,b})^2 = o_{\mathcal{P}_0}(1).$$

This in turn follows from showing

$$\frac{1}{S}\sum_{b=1}^{S}T_{2,b}^2 = o_{\mathcal{P}_0}(1), \quad \frac{1}{S}\sum_{b=1}^{S}T_{3,b}^2 = o_{\mathcal{P}_0}(1), \quad \frac{1}{S}\sum_{b=1}^{S}T_{4,b}^2 = o_{\mathcal{P}_0}(1).$$

These results can be established using the similar techniques as for T_2, T_3, and T_4:

1. **Step 1.** Recall that

$$T_{2,b} = \|(\mathbb{III})_b\|^2 + 2\langle(\mathbb{I})_b, (\mathbb{III})_b\rangle.$$

By the inequality $(a+b)^2 \le 2(a^2+b^2)$ for $a, b \ge 0$, it follows that

$$T_{2,b}^2 \;\le\; 2\|(\mathbb{III})_b\|^4 + 8\langle(\mathbb{I})_b, (\mathbb{III})_b\rangle^2.$$

(1) The $\|(\mathbb{III})_b\|^4$ term. We first control the term $\|(\mathbb{III})_b\|^4$. From the definition of $(\mathbb{III})_b$ and the boundedness of the kernel, we obtain

$$\|(\mathbb{III})_b\|_{\mathcal{H}_k}^4 \le K^4 \cdot B^2\biggl(\frac{1}{B}\sum_{i=(b-1)B+1}^{bB}\{r_X(X_i^{(2)}) - \hat{r}_X(X_i^{(2)})\}^2\biggr)^2.$$

Applying Lundborg et al. (2022, Lemma S5) and Markov's inequality, together with the condition

$$\sup_{P\in\mathcal{P}_0}\mathbb{E}_P\Bigl[(\hat{r}_X(X^{(2)}) - r_X(X^{(2)}))^2\Bigr] = o(N^{-1/2}),$$

we deduce

$$\|(\mathbb{III})_b\|_{\mathcal{H}_k}^4 = o_{\mathcal{P}_0}(B^2/N).$$

Under Assumption 3(e), this implies $\|(\mathbb{III})_b\|^4 = o_{\mathcal{P}_0}(1)$. Moreover by Markov's inequality $\frac{1}{S}\sum_{b=1}^{S}\|(\mathbb{III})_b\|^4 = o_{\mathcal{P}_0}(1)$.

(2) The $\langle(\mathbb{I})_b, (\mathbb{III})_b\rangle^2$ term. For the second contribution, we first apply the Cauchy–Schwarz inequality:

$$\langle(\mathbb{I})_b, (\mathbb{III})_b\rangle^2 \;\le\; \|(\mathbb{I})_b\|^2\|(\mathbb{III})_b\|^2.$$

Since $\|(\mathbb{III})_b\|^2 \le \sum_{b=1}^{S}\|(\mathbb{III})_b\|^2$, it follows that

$$\begin{aligned}\frac{1}{S}\sum_{b=1}^{S}\langle(\mathbb{I})_b, (\mathbb{III})_b\rangle^2 &\le \frac{1}{S}\sum_{b=1}^{S}\|(\mathbb{I})_b\|^2\|(\mathbb{III})_b\|^2 \\ &\le \Bigl(\frac{1}{S}\sum_{b=1}^{S}\|(\mathbb{I})_b\|^2\Bigr)\cdot\Bigl(\sum_{b=1}^{S}\|(\mathbb{III})_b\|^2\Bigr).\end{aligned}$$

By Assumption 3(a), $\|(\mathbb{I})_b\|^2$ has mean zero and finite variance. Hence, by Chebyshev's inequality,

$$\frac{1}{S}\sum_{b=1}^{S}\|(\mathbb{I})_b\|^2 = o_{\mathcal{P}_0}(S^{-1/2}).$$

Moreover, Assumption 3(b) gives

$$\sum_{b=1}^{S}\|(\mathbb{III})_b\|^2 = S\cdot\frac{1}{S}\sum_{b=1}^{S}\|(\mathbb{III})_b\|^2 = S\cdot o_{\mathcal{P}_0}(BN^{-1/2}).$$

Combining these bounds together with Assumption 3(e), we conclude that the entire cross term is of order $o_{\mathcal{P}_0}(1)$. Combining (1) and (2), we conclude that

$$\frac{1}{S}\sum_{b=1}^{S}T_{2,b}^2 = o_{\mathcal{P}_0}(1).$$

2. **Step 2.** For $\frac{1}{S}\sum_{b=1}^{S}T_{3,b}^2$ and $\frac{1}{S}\sum_{b=1}^{S}T_{4,b}^2$, we proceed in the same way as in **Step 1**. In particular, for $\frac{1}{S}\sum_{b=1}^{S}T_{4,b}^2$ we apply Jensen's inequality, while for $\frac{1}{S}\sum_{b=1}^{S}T_{3,b}^2$ we directly use the same bounding technique as in **Step 1**. Together with Markov's inequality, these bounds imply that both averages vanish:

$$\frac{1}{S}\sum_{b=1}^{S}T_{3,b}^2 = o_{\mathcal{P}_0}(1), \qquad \frac{1}{S}\sum_{b=1}^{S}T_{4,b}^2 = o_{\mathcal{P}_0}(1).$$

Combining these results with Markov's inequality completes the proof.

A.6 Proof of Corollary 3

For each $j \in \{1, \dots, K\}$, the proof of Theorem 5 establishes that $\widehat{\sigma}_j/\sigma_P = 1 + o_{\mathcal{P}_0}(1)$. Therefore, by Lemma 4, 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 n be evenly divided into $S := \frac{n}{B}$ disjoint blocks of size B, and assume that S is divisible by K. We assign the blocks to K disjoint subsets, denoted $\mathcal{B}_1, \dots, \mathcal{B}_K$, such that each fold j contains exactly S/K blocks, i.e., $|\mathcal{B}_j| = S/K$. Each block $b \in \{1, \dots, S\}$ 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

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

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, $\sum_{b\in\mathcal{B}_1}\eta_b, \dots, \sum_{b\in\mathcal{B}_K}\eta_b$ 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 (Lemma 2), 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.

A.7 Proof of Theorem 6

By direct expansion, we have the exact relation

$$\widehat{\mathrm{MMD}}_u^2 = \frac{M}{M-1}V_M^2 - \frac{1}{M(M-1)}\sum_{i=1}^{M}\widehat{H}(Z_i, Z_i),$$

where

$$V_M = \left\| \frac{1}{M}\sum_{i=1}^{M}\psi(V_i^{(1)}) - \frac{1}{M}\sum_{i=1}^{M}\widehat{r}_X(X_i^{(2)})\,\psi(V_i^{(2)}) \right\|_{\mathcal{H}_k}.$$

Multiplying both sides by M yields

$$M \cdot \widehat{\mathrm{MMD}}_u^2 = MV_M^2 + \left(\frac{M}{M-1} - 1\right) \cdot MV_M^2 - \frac{M}{M-1} \cdot \frac{1}{M} \sum_{i=1}^{M} \widehat{H}(W_i, W_i).$$

This decomposition highlights that the main stochastic contribution arises from MV_M^2, while the diagonal term provides a deterministic centering. Accordingly, we proceed as follows to establish the asymptotic distribution of $M \cdot \widehat{\mathrm{MMD}}_u^2$.

- **Step 1.** the fluctuation of MV_M^2, which converges to an infinite sum of chi-squared random variables and is therefore stochastically bounded.
- **Step 2.** the diagonal average $\frac{1}{M}\sum_{i=1}^{M} \widehat{H}(W_i, W_i)$, which converges to a finite deterministic value depending on the kernel operator.

Step 1. To analyze the first step, it is convenient to study $\sqrt{M}V_M$, since

$$MV_M^2 = \left(\sqrt{M}V_M\right)^2.$$

Instead of analyzing MV_M^2 directly, we first consider $\sqrt{M}V_M$ to isolate the effect of density ratio estimation. This linearization shows that the additional error from $\widehat{r}_X - r_X$ is asymptotically negligible under Assumption 4(c)–(d). Equivalently, our goal is to establish that, asymptotically,

$$\sqrt{M}\, V_M = \left\| \frac{1}{\sqrt{M}} \sum_{i=1}^{M} \psi(V_i^{(1)}) - \frac{1}{\sqrt{M}} \sum_{i=1}^{M} r_X(X_i^{(2)})\, \psi(V_i^{(2)}) \right\|_{\mathcal{H}_k},$$

so that the contribution from the density ratio estimator disappears in the limit. To make this precise, we decompose

$$\begin{aligned}\sqrt{M}\, V_M &= \Bigg\| \underbrace{\frac{1}{\sqrt{M}} \sum_{i=1}^{M} \{\psi(V_i^{(1)}) - r_X(X_i^{(2)})\, \psi(V_i^{(2)})\}}_{(\mathbb{I})} + \underbrace{\frac{1}{\sqrt{M}} \sum_{i=1}^{M} \psi(V_i^{(2)}) \{r_X(X_i^{(2)}) - \widehat{r}_X(X_i^{(2)})\}}_{(\mathbb{III})} \Bigg\|_{\mathcal{H}_k} \\ &= \sqrt{\|(\mathbb{I})\|_{\mathcal{H}_k}^2 + \|(\mathbb{III})\|_{\mathcal{H}_k}^2 + 2\langle(\mathbb{I}), (\mathbb{III})\rangle_{\mathcal{H}_k}}.\end{aligned}$$

Hence the problem reduces to showing

1. $\|(\mathbb{I})\|_{\mathcal{H}_k}$ is stochastically bounded.
2. $\|(\mathbb{III})\|_{\mathcal{H}_k} = o_{\mathcal{P}_n^{(0)}}(1)$.

The first condition guarantees that the main term does not diverge, while the second ensures that the error term vanishes. Together, they also imply that the cross term $2\langle(\mathbb{I}), (\mathbb{III})\rangle_{\mathcal{H}_k}$ is asymptotically negligible, Consequently, we obtain

$$MV_M^2 = \|(\mathbb{I})\|_{\mathcal{H}_k}^2 + o_{\mathcal{P}_n^{(0)}}(1),$$

which provides the desired reduction for analyzing the asymptotic distribution of MV_M^2.

1. **Term** $(\mathbb{I})$**.** Recall that $\varphi(W) = \psi(V^{(1)}) - r_X(X^{(2)})\, \psi(V^{(2)})$ and that $\widetilde{H}_n$ denotes the centered kernel introduced in (14). Define $(\mathbb{I}) = \frac{1}{\sqrt{M}} \sum_{i=1}^{M} \varphi(W_i)$. Then

$$\begin{aligned}\|(\mathbb{I})\|_{\mathcal{H}_k}^2 &= \frac{1}{M} \Big\langle \sum_{i=1}^{M} \varphi(W_i), \sum_{j=1}^{M} \varphi(W_j) \Big\rangle_{\mathcal{H}_k} \\ &= \frac{1}{M} \sum_{i,j=1}^{M} H(W_i, W_j),\end{aligned}$$

where $H(W_i, W_j) = \langle \varphi(W_i), \varphi(W_j) \rangle_{\mathcal{H}_k}$. Using the definition of $\widetilde{H}_n$, under the null hypothesis we may rewrite

$$\|(\mathbb{I})\|^2_{\mathcal{H}_k} = \frac{1}{M} \sum_{i,j=1}^{M} \widetilde{H}_n(W_i, W_j),$$

so that $\frac{1}{M}\|(\mathbb{I})\|^2$ is a V-statistic with kernel $\widetilde{H}_n$. By the spectral decomposition, we may write

$$\begin{aligned}
\|(\mathbb{I})\|^2_{\mathcal{H}_k} &= \frac{1}{M} \sum_{i,j=1}^{M} \sum_{\ell=1}^{\infty} \lambda_{\ell,n}^{(P_n)} \, \Psi_{\ell,n}^{(P_n)}(W_i) \Psi_{\ell,n}^{(P_n)}(W_j) \\
&= \sum_{\ell=1}^{\infty} \lambda_{\ell,n}^{(P_n)} \Big(\frac{1}{\sqrt{M}} \sum_{i=1}^{M} \Psi_{\ell,n}^{(P_n)}(W_i) \Big)^2.
\end{aligned}$$

By Assumption 4(a), the sequence of eigenvalues is uniformly square–summable, so that the series above is well defined. Next, under Assumption 4(b), for each fixed ℓ and fixed distribution P_n the central limit theorem implies that $\frac{1}{\sqrt{M}} \sum_{i=1}^{M} \Psi_{\ell,n}^{(P_n)}(W_i)$ converges in distribution to a standard Gaussian. By the continuous mapping theorem, the squared terms converge in distribution to $\chi^2(1)$ random variables, and hence

$$\sum_{\ell=1}^{\infty} \lambda_{\ell,n}^{(P_n)} \Big(\frac{1}{\sqrt{M}} \sum_{i=1}^{M} \Psi_{\ell,n}^{(P_n)}(W_i) \Big)^2$$

converges to $\sum_{\ell \geq 1} \lambda_{\ell,n}^{(P_n)} a_\ell^2$, where $a_\ell \sim N(0,1)$. For each fixed P_n, this establishes pointwise convergence in distribution of $\|(\mathbb{I})\|^2_{\mathcal{H}_k}$. To strengthen this result to uniform convergence of the distribution functions over $t \in \mathbb{R}$, we invoke Polya's theorem (Lehmann, 2004, Theorem 2.6.1), which states that convergence in distribution implies uniform convergence of the associated distribution functions when the limit distribution is continuous. Consequently, for any sequence $\{P_n : n \geq 1\}$ with $P_n \in \mathcal{P}_n^{(0)}$, we obtain

$$\lim_{n \to \infty} \sup_{t \in \mathbb{R}} \Big| \mathbb{P}_{P_n} \big(\|(\mathbb{I})\|^2_{\mathcal{H}_k} \leq t \big) - \mathbb{P} \big(\widetilde{G}_n^{(P_n)} \leq t \big) \Big| = 0,$$

where

$$\widetilde{G}_n^{(P_n)} \sim \sum_{\ell=1}^{\infty} \lambda_{\ell,n}^{(P_n)} \, a_\ell^2, \qquad \{a_\ell\}_{\ell \geq 1} \overset{\text{i.i.d.}}{\sim} N(0,1).$$

2. **Term** $(\mathbb{II})$**.** By the definition of the Hilbert norm and the Cauchy–Schwarz inequality,

$$\begin{aligned}
\Big\| \frac{1}{\sqrt{M}} \sum_{i=1}^{M} \psi(V_i^{(2)}) \cdot \big\{ r_X(X_i^{(2)}) - \hat{r}_X(X_i^{(2)}) \big\} \Big\| \leq \sqrt{M} \Big(\frac{1}{M} \sum_{i=1}^{M} \|\psi(V_i^{(2)})\|^2 \Big)^{1/2} \\
\cdot \Big(\frac{1}{M} \sum_{i=1}^{M} \big\{ r_X(X_i^{(2)}) - \hat{r}_X(X_i^{(2)}) \big\}^2 \Big)^{1/2}.
\end{aligned}$$

Under Assumption 4(d), the kernel $k(\cdot,\cdot) = \langle \psi(\cdot), \psi(\cdot) \rangle$ is uniformly bounded in $\mathcal{H}_k$, so the first factor is bounded by a constant $K > 0$. For the second factor, the convergence rate of the density ratio estimator implies

$$\sqrt{M} \left(\frac{1}{M} \sum_{i=1}^{M} \big(r_X(X_i^{(2)}) - \hat{r}_X(X_i^{(2)}) \big)^2 \right)^{1/2} = o_{\mathcal{P}_n^{(0)}} \Big(\frac{M^{1/2}}{N^{1/4}} \Big).$$

Hence,

$$\Big\| \frac{1}{\sqrt{M}} \sum_{i=1}^{M} \psi(V_i^{(2)}) \cdot \big\{ r_X(X_i^{(2)}) - \hat{r}_X(X_i^{(2)}) \big\} \Big\| = o_{\mathcal{P}_n^{(0)}}\Big(\frac{M^{1/2}}{N^{1/4}} \Big).$$

By Assumption 4(e) ($M^2 \ll N$), it follows that

$$\|(\mathbb{III})\|^2 = o_{\mathcal{P}_0}(1).$$

3. **Term $\langle (\mathbb{I}), (\mathbb{III}) \rangle_{\mathcal{H}_k}$.** By the Cauchy–Schwarz inequality,

$$\big| \langle (\mathbb{I}), (\mathbb{III}) \rangle \big| \le \|(\mathbb{I})\| \cdot \|(\mathbb{III})\|.$$

Since $\|(\mathbb{I})\| = O_{\mathcal{P}_0}(1)$ and $\|(\mathbb{III})\| = o_{\mathcal{P}_0}(1)$, we conclude

$$\langle (\mathbb{I}), (\mathbb{III}) \rangle = o_{\mathcal{P}_0}(1).$$

Combining the above results, we establish that for any sequence $\{P_n : n \ge 1\}$ with $P_n \in \mathcal{P}_n^{(0)}$,

$$\lim_{n\to\infty} \sup_{t \in \mathbb{R}} \Big| \mathbb{P}_{P_n}\big(M V_M^2 \le t \big) - \mathbb{P}\big(\widetilde{G}_n^{(P_n)} \le t \big) \Big| = 0,$$

where

$$\widetilde{G}_n^{(P_n)} \;\sim\; \sum_{\ell=1}^{\infty} \lambda_{\ell,n}^{(P_n)}\, a_\ell^2, \qquad \{a_\ell\}_{\ell \ge 1} \overset{\text{i.i.d.}}{\sim} N(0,1).$$

Step 2. By the construction of $\widehat{H}$ and H, we can write

$$\frac{1}{M} \sum_{i=1}^{M} \widehat{H}(Z_i, Z_i) = \frac{1}{M} \sum_{i=1}^{M} H(Z_i, Z_i) + \frac{1}{M} \sum_{i=1}^{M} \big\{ \widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)}) \big\}^2 k(V_i^{(2)}, V_i^{(2)}) - \frac{2}{M} \sum_{i=1}^{M} \big\{ \widehat{r}_X(X_i^{(2)}) - r_X(X_i^{(2)}) \big\} k(V_i^{(1)}, V_i^{(2)}).$$

Under the null hypothesis, and under Assumption 4(c) and (d), and using the above analysis, we obtain

$$\frac{1}{M} \sum_{i=1}^{M} \widehat{H}(W_i, W_i) = \frac{1}{M} \sum_{i=1}^{M} \widetilde{H}_n(Z_i, Z_i) + o_{\mathcal{P}_n^{(0)}}(1).$$

Furthermore, by the spectral decomposition, we have

$$\frac{1}{M} \sum_{i=1}^{M} \widehat{H}(W_i, W_i) = \sum_{\ell=1}^{\infty} \lambda_{\ell,n}^{(P_n)} \Big(\frac{1}{M} \sum_{i=1}^{M} \big(\Psi_{\ell,n}^{(P_n)}(W_i) \big)^2 \Big) + o_{\mathcal{P}_n^{(0)}}(1).$$

Finally, under Assumption 4(a) and (b), this expression converges to $\sum_{\ell=1}^{\infty} \lambda_{\ell,n}^{(P_n)}$. Recall that

$$M \cdot \widehat{\mathrm{MMD}}_u^2 = M V_M^2 + \underbrace{\Big(\frac{M}{M-1} - 1 \Big) \cdot M V_M^2}_{(a)} - \underbrace{\frac{M}{M-1} \cdot \frac{1}{M} \sum_{i=1}^{M} \widehat{H}(W_i, W_i)}_{(b)}.$$

Combining the results from **Step 1** and **Step 2**, we observe the following. Since MV_M^2 converges to a limiting distribution, it is stochastically bounded, which implies that $(a) = o_{\mathcal{P}_n^{(0)}}(1)$. Moreover, (b) converges to $\sum_{\ell=1}^{\infty} \lambda_{\ell,n}^{(P_n)}$. Therefore, we conclude that

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \Big| \mathbb{P}_{P_n}\big(M \cdot \widehat{\mathrm{MMD}}_u^2 \le t\big) - \mathbb{P}\big(G_n^{(P_n)} \le t\big) \Big| = 0,$$

where

$$G_n^{(P_n)} \sim \sum_{\ell=1}^{\infty} \lambda_{\ell,n}^{(P_n)} (a_\ell^2 - 1), \qquad \{a_\ell\}_{\ell\ge1} \overset{\text{i.i.d.}}{\sim} N(0,1).$$

This is sufficient to establish the uniform result

$$\lim_{n\to\infty} \sup_{P_n\in\mathcal{P}_n^{(0)}} \sup_{t\in\mathbb{R}} \Big| \mathbb{P}_{P_n}\big(M \cdot \widehat{\mathrm{MMD}}_u^2 \le t\big) - \mathbb{P}\big(G_n^{(P_n)} \le t\big) \Big| = 0.$$

To see this, select a sequence P_n' such that, for all n, we have

$$\begin{aligned} \sup_{t\in\mathbb{R}} \Big| \mathbb{P}_{P_n'}\big(M \cdot \widehat{\mathrm{MMD}}_u^2 \le t\big) - \mathbb{P}\big(G_n^{(P_n)} \le t\big) \Big| &\le \sup_{P_n\in\mathcal{P}_n^{(0)}} \sup_{t\in\mathbb{R}} \Big| \mathbb{P}_{P_n}\big(M \cdot \widehat{\mathrm{MMD}}_u^2 \le t\big) - \mathbb{P}\big(G_n^{(P_n)} \le t\big) \Big| \\ &\le \sup_{t\in\mathbb{R}} \Big| \mathbb{P}_{P_n'}\big(M \cdot \widehat{\mathrm{MMD}}_u^2 \le t\big) - \mathbb{P}\big(G_n^{(P_n)} \le t\big) \Big| + \frac{1}{n}. \end{aligned}$$

Since both the leftmost and rightmost terms converge to zero, it follows that the middle term also converges to zero, as required. This completes the proof of Theorem 6.

A.8 Proof of Proposition 1

Proof. Define the wild bootstrap statistic with the known density ratio, denoted by $\mathrm{MMD}^2_{\mathrm{wild}}$, as

$$\mathrm{MMD}^2_{\mathrm{wild}} := \frac{1}{M(M-1)} \sum_{i \neq j} \xi_i \xi_j \, H(W_i, W_j),$$

where H is the counterpart of $\widehat{H}$ obtained by using the true density ratio r_X instead of its estimator $\widehat{r}_X$. For comparison, we also define the oracle U-statistic

$$U_M(H) := \frac{1}{M(M-1)} \sum_{i \neq j} H(W_i, W_j).$$

We decompose the difference between the wild bootstrap statistic with the estimated kernel and the oracle U-statistic:

$$M \cdot \widehat{\mathrm{MMD}}^2_{\mathrm{wild}} - M \cdot U_M(H) = \underbrace{\Big(M \cdot \widehat{\mathrm{MMD}}^2_{\mathrm{wild}} - M \cdot \mathrm{MMD}^2_{\mathrm{wild}}\Big)}_{(\mathbb{I})} + \underbrace{\Big(M \cdot \mathrm{MMD}^2_{\mathrm{wild}} - M \cdot U_M(H)\Big)}_{(\mathbb{II})}.$$

We show that ($\mathbb{I}$) is asymptotically negligible, while ($\mathbb{II}$) is asymptotically consistent. so that $M \cdot \widehat{\mathrm{MMD}}^2_{\mathrm{wild}}$ and $M \cdot U_M(H)$ share the same asymptotic distribution under the null. This establishes the consistency of the wild bootstrap for approximating the null distribution.

1. **Term** ($\mathbb{II}$)**.** We first analyze term ($\mathbb{II}$), which plays a key role in establishing consistency. Recall the definition of the wild bootstrap statistic:

$$\mathrm{MMD}^2_{\mathrm{wild}} = \frac{1}{M(M-1)} \sum_{i \neq j} \xi_i \xi_j \, \widetilde{H}_n(W_i, W_j),$$

where $\{W_i\}_{i=1}^M$ are i.i.d. samples and $\{\xi_i\}_{i=1}^M$ are independent bootstrap multipliers. Under the null hypothesis we may replace H with $\widetilde{H}_n$. By the spectral decomposition of $\widetilde{H}_n$, we obtain

$$M \cdot \mathrm{MMD}^2_{\mathrm{wild}} = \frac{M}{M-1} \sum_{\ell=1}^{\infty} \lambda_{\ell,n}^{(P_n)} \left\{ \Big(\frac{1}{\sqrt{M}} \sum_{i=1}^{M} \xi_i \Psi_{\ell,n}^{(P_n)}(W_i) \Big)^2 - \frac{1}{M} \sum_{i=1}^{M} \xi_i^2 \, (\Psi_{\ell,n}^{(P_n)}(W_i))^2 \right\}.$$

Conditional on $\mathbb{W} = \{W_1, \ldots, W_M\}$, the linear terms have mean zero and variance $\frac{1}{M}\sum_{i=1}^M (\Psi_{\ell,n}^{(P_n)}(W_i))^2$. By the conditional central limit theorem for multipliers, for each fixed ℓ the normalized sum converges in distribution to a centered Gaussian random variable, so the leading quadratic form converges to a weighted sum of independent χ_1^2 variables with weights $\{\lambda_{\ell,n}^{(P_n)}\}$. The diagonal correction term converges in probability to 1 by the law of large numbers.

Therefore, conditionally on $\mathbb{W}$ and for any sequence $\{P_n\}_{n\geq 1}$ with $P_n \in \mathcal{P}_n^{(0)}$, the asymptotic distribution of $M \cdot \mathrm{MMD}^2_{\mathrm{wild}}$ is given by

$$\sum_{\ell=1}^{\infty} \lambda_{\ell,n}^{(P_n)} (a_\ell^2 - 1), \qquad a_\ell \overset{\text{i.i.d.}}{\sim} N(0,1),$$

which coincides with the null limiting distribution of the quadratic-time MMD U-statistic. Since the limiting distribution is continuous, Polya's theorem implies uniform convergence of the corresponding distribution functions. In particular,

$$\lim_{n\to\infty} \sup_{P_n \in \mathcal{P}_n^{(0)}} \sup_{x \in \mathbb{R}} \Big| \mathbb{P}_{P_n}\big(M \cdot \mathrm{MMD}^2_{\mathrm{wild}} \leq x \mid \mathbb{W}\big) - \mathbb{P}_{P_n}\big(M \cdot U_M(H) \leq x\big) \Big| = 0.$$

This establishes the conditional and uniform consistency of the wild bootstrap.

2. **Term** ($\mathbb{I}$). Similar to Appendix A.7, we define

$$V_M^{(\text{wild})} := \left\| \frac{1}{M} \sum_{i=1}^{M} \xi_i \psi(V_i^{(1)}) - \frac{1}{M} \sum_{i=1}^{M} \xi_i \widehat{r}_X(X_i^{(2)}) \psi(V_i^{(2)}) \right\|_{\mathcal{H}_k}.$$

A direct expansion yields

$$M \cdot \widehat{\text{MMD}}^2_{\text{wild}} = M\big(V_M^{(\text{wild})}\big)^2 + \Big(\frac{M}{M-1} - 1\Big) \cdot M\big(V_M^{(\text{wild})}\big)^2 - \frac{M}{M-1} \cdot \frac{1}{M} \sum_{i=1}^{M} \xi_i^2 \widehat{H}(W_i, W_i).$$

As before, the main stochastic contribution comes from $M(V_M^{(\text{wild})})^2$, while the last two terms act as centering and vanish asymptotically. Decomposing $\sqrt{M} V_M^{(\text{wild})}$ into

$$T_1 := \frac{1}{\sqrt{M}} \sum_{i=1}^{M} \xi_i \big\{ \psi(V_i^{(1)}) - r_X(X_i^{(2)}) \psi(V_i^{(2)}) \big\}, \quad T_2 := \frac{1}{\sqrt{M}} \sum_{i=1}^{M} \xi_i\, \psi(V_i^{(2)}) \{ r_X(X_i^{(2)}) - \widehat{r}_X(X_i^{(2)}) \},$$

we obtain

$$M(V_M^{(\text{wild})})^2 = \|T_1\|^2_{\mathcal{H}_k} + o_{\mathcal{P}_n^{(0)}}(1).$$

Conditional on $\mathbb{Z}$, the structure of T_1 coincides with that of Term ($\mathbb{III}$) analyzed in the earlier step. Hence, the same spectral decomposition and CLT arguments apply, and $\|T_1\|^2_{\mathcal{H}_k}$ converges in distribution to an infinite weighted sum of independent χ_1^2 random variables. The remainder T_2 has mean zero, and its conditional variance is negligible under Assumption 4(c)–(e). By Chebyshev's inequality, it follows that $T_2 = o_{\mathcal{P}_n^{(0)}}(1)$. Finally, since the diagonal correction term satisfies $\xi_i^2 = 1$, the similar argument as in Appendix A.7 applies. Therefore, the difference between the wild bootstrap statistic based on $\widehat{H}$ and its oracle counterpart based on H is asymptotically negligible conditional on $\mathbb{W}$.

Combining the analyses of Term ($\mathbb{I}$) and Term ($\mathbb{III}$), we conclude that for any sequence $\{P_n\}_{n \geq 1}$ with $P_n \in \mathcal{P}_n^{(0)}$,

$$\lim_{n \to \infty} \sup_{P_n \in \mathcal{P}_n^{(0)}} \sup_{x \in \mathbb{R}} \Big| \mathbb{P}_{P_n}\big(M \cdot \widehat{\text{MMD}}^2_{\text{wild}} \leq x \mid \mathbb{W}\big) - \mathbb{P}_{P_n}\big(G_n^{(P_n)} \leq x\big) \Big| = 0,$$

where

$$G_n^{(P_n)} \sim \sum_{\ell=1}^{\infty} \lambda_{\ell,n}^{(P_n)} (a_\ell^2 - 1), \qquad a_\ell \overset{\text{i.i.d.}}{\sim} N(0,1).$$

That is, the plug-in wild bootstrap statistic $M \cdot \widehat{\text{MMD}}^2_{\text{wild}}$ converges, conditional on $\mathbb{W}$, to the same sum of infinitely weighted chi-squared random variables as the oracle statistic. Consequently, the wild bootstrap procedure based on the estimated density ratio is conditionally consistent for approximating the null distribution of the quadratic-time MMD statistic.

□

A.9 Proof of Example 1

In this section, we aim to present a detailed analysis of the asymptotic equivalence between the GCM statistic T_n and its counterpart $\tilde{T}_n$ constructed using $\{(\tilde{X}_i, \tilde{Y}_i, \tilde{Z}_i)\}_{i=1}^n$. Throughout this section, we assume that Y has a finite second moment and that $Y \perp\!\!\!\perp Z \,|\, X$, i.e., the null hypothesis holds. Let

$$T_n = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n R_i}{\big\{ \frac{1}{n} \sum_{i=1}^n R_i^2 - \big(\frac{1}{n} \sum_{r=1}^n R_r \big)^2 \big\}^{1/2}} := \frac{\nu_R}{\hat{\sigma}_R},$$

and $\tilde{T}_n \coloneqq \nu_{\tilde{R}}/\hat{\sigma}_{\tilde{R}}$. Let $\sigma^2_{\tilde{R}} > 0$ denote the variance of $\{\tilde{Y} - f(\tilde{X})\}\{\tilde{Z} - g(\tilde{X})\}$ where $(\tilde{X}, \tilde{Y}, \tilde{Z})$ is a random draw from the joint distribution P_{XYZ}. We begin with an upper bound for $|T_n - \tilde{T}_n|$:

$$\begin{aligned}\Big|\frac{\nu_R}{\hat{\sigma}_R} - \frac{\nu_{\tilde{R}}}{\hat{\sigma}_{\tilde{R}}}\Big| &\leq \Big|\frac{\nu_R}{\hat{\sigma}_R} - \frac{\nu_{\tilde{R}}}{\hat{\sigma}_R}\Big| + \Big|\frac{\nu_{\tilde{R}}}{\hat{\sigma}_R} - \frac{\nu_{\tilde{R}}}{\hat{\sigma}_{\tilde{R}}}\Big| \\ &\leq \frac{1}{\hat{\sigma}_R}|\nu_R - \nu_{\tilde{R}}| + \frac{|\nu_{\tilde{R}}|}{(\hat{\sigma}_R + \hat{\sigma}_{\tilde{R}})\hat{\sigma}_R\hat{\sigma}_{\tilde{R}}}|\hat{\sigma}^2_R - \hat{\sigma}^2_{\tilde{R}}|,\end{aligned}$$

from which the proof boils down to showing the convergence of the following four terms to zero in probability: (a) $\nu_R - \nu_{\tilde{R}}$, (b) $\hat{\sigma}^2_R - \hat{\sigma}^2_{\tilde{R}}$, (c) $\hat{\sigma}^2_R - \sigma^2_{\tilde{R}}$ and (d) $\hat{\sigma}^2_{\tilde{R}} - \sigma^2_{\tilde{R}}$. Under these convergence results, the asymptotic equivalence follows by the continuous mapping theorem along with the fact that $\nu_{\tilde{R}}$ is stochastically bounded by the central limit theorem. In what follows, we establish convergence of (a), (b), (c), and (d) to zero in probability in order.

1. **Term (a):** $\nu_R - \nu_{\tilde{R}}$**.** Starting with the term (a), the difference between ν_R and $\nu_{\tilde{R}}$ can be written as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n R_i - \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{R}_i = \underbrace{\frac{1}{\sqrt{n}}\sum_{i=n_1+1}^{\bar{n}_1}(R_i - \tilde{R}_i)\cdot \mathbb{1}(\bar{n}_1 > n_1)}_{\coloneqq \Delta_1} + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=\bar{n}_1+1}^{n_1}(R_i - \tilde{R}_i)\cdot \mathbb{1}(\bar{n}_1 \leq n_1)}_{\coloneqq \Delta_2}.$$

Remark that Z_i is a fixed constant for a given index i, which allows us to show that $\mathbb{E}[R_i] = 0$ for any $i \in [n]$. For example, when $i = 1$, Z_1 equals 1 (since $X_1 = X_1^{(1)}$) and thus the law of total expectation yields

$$\mathbb{E}[R_1] = \mathbb{E}[\{1 - g(X_1)\}\{Y_1 - f(X_1)\}] = \mathbb{E}\big[\{1 - g(X_1)\}\,\mathbb{E}[\{Y_1 - f(X_1)\}\,|\,X_1]\big] = 0,$$

where we recall $f(X_1) = \mathbb{E}[Y_1\,|\,X_1]$. It also follows that $\mathbb{E}[\tilde{R}_i] = 0$ for any $i \in [n]$ under the null hypothesis. This together with the law of total expectation shows that

$$\mathbb{E}[\Delta_1] = \mathbb{E}\left[\frac{1}{\sqrt{n}}\sum_{i=n_1+1}^{\bar{n}_1}\mathbb{1}(\bar{n}_1 > n_1)\mathbb{E}\big[R_i - \tilde{R}_i\,|\,\bar{n}_1\big]\right] = 0,$$

and similarly $\mathbb{E}[\Delta_2] = 0$. Thus, we have $\mathbb{E}[\nu_R - \nu_{\tilde{R}}] = 0$.

Now consider the variance of $\nu_R - \nu_{\tilde{R}}$. Since $\mathbb{E}[\nu_R - \nu_{\tilde{R}}] = 0$ and $\mathbb{E}[\Delta_1\Delta_2] = 0$, we have

$$\mathrm{Var}[\nu_R - \nu_{\tilde{R}}] = \mathrm{Var}[\Delta_1] + \mathrm{Var}[\Delta_2].$$

For $\mathrm{Var}[\Delta_1]$, we have

$$\begin{aligned}\mathrm{Var}[\Delta_1] &= \mathbb{E}\big[\mathrm{Var}\,\{\Delta_1\,|\,\bar{n}_1\}\big] + \mathrm{Var}\,\big[\underbrace{\mathbb{E}\{\Delta_1\,|\,\bar{n}_1\}}_{=0}\big] \\ &= \mathbb{E}\left[\mathrm{Var}\left\{\frac{1}{\sqrt{n}}\sum_{i=n_1+1}^{\bar{n}_1}(R_i - \tilde{R}_i)\cdot\mathbb{1}(\bar{n}_1 > n_1)\;\Big|\;\bar{n}_1\right\}\right] \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{i=n_1+1}^{\bar{n}_1}\mathbb{1}(\bar{n}_1 > n_1)\,\mathbb{E}\big\{\big(R_i - \tilde{R}_i\big)^2\;|\;\bar{n}_1\big\}\right] \\ &\overset{\text{(i)}}{\leq} 2\mathbb{E}\left[\frac{\bar{n}_1 - n_1}{n}\cdot\frac{1}{\bar{n}_1 - n_1}\sum_{i=n_1+1}^{\bar{n}_1}\mathbb{E}\big(R_i^2 + \tilde{R}_i^2\,|\,\bar{n}_1\big)\right] \overset{\text{(ii)}}{\leq} \frac{4}{n}\mathbb{E}\left[|\bar{n}_1 - n_1|\right]\mathrm{Var}(Y_1) \overset{\text{(iii)}}{\leq} \frac{2}{\sqrt{n}}\mathrm{Var}(Y_1),\end{aligned}$$

where (i) follows from the inequality $(x-y)^2 \leq 2x^2 + 2y^2$ and (ii) uses the law of total variance along with the fact that $R_i^2 \leq \{Y_i - f(X_i)\}^2$ and $\tilde{R}_i^2 \leq \{\tilde{Y}_i - f(\tilde{X}_i)\}^2$ since $Z_i, \tilde{Z}_i \in \{1, 2\}$. For the last inequality (iii), we use $\mathbb{E}[|\bar{n}_1 - n_1|] \leq \sqrt{n}/2$. The same bound holds for $\mathrm{Var}[\Delta_2]$ and thus

$$\mathrm{Var}[\nu_R - \nu_{\tilde{R}}] \leq \frac{4}{\sqrt{n}}\mathrm{Var}(Y_1).$$

Combining the results with Chebyshev's inequality now shows that $\nu_R - \nu_{\tilde{R}}$ converges to zero in probability.

2. **Term (b): $\hat{\sigma}_R^2 - \hat{\sigma}_{\tilde{R}}^2$.** We next aim to show that

$$\hat{\sigma}_R^2 - \hat{\sigma}_{\tilde{R}}^2 = \left\{\frac{1}{n}\sum_{i=1}^n R_i^2 - \Big(\frac{1}{n}\sum_{r=1}^n R_r\Big)^2\right\} - \left\{\frac{1}{n}\sum_{i=1}^n \tilde{R}_i^2 - \Big(\frac{1}{n}\sum_{r=1}^n \tilde{R}_r\Big)^2\right\}$$

converges to zero in probability. We decompose this into two terms

$$(\mathbb{I}) := \frac{1}{n}\sum_{i=1}^n R_i^2 - \frac{1}{n}\sum_{i=1}^n \tilde{R}_i^2 \quad \text{and} \quad (\mathbb{II}) := \Big(\frac{1}{n}\sum_{r=1}^n R_r\Big)^2 - \Big(\frac{1}{n}\sum_{r=1}^n \tilde{R}_r\Big)^2,$$

and show each of them converges to zero in probability. For the first term $(\mathbb{I})$, we have

$$\frac{1}{n}\sum_{i=1}^n R_i^2 - \frac{1}{n}\sum_{i=1}^n \tilde{R}_i^2 = \underbrace{\frac{1}{n}\sum_{n_1+1}^{\bar{n}_1}(R_i^2 - \tilde{R}_i^2)\cdot \mathbb{1}(\bar{n}_1 > n_1)}_{:=\tilde{\Delta}_1} + \underbrace{\frac{1}{n}\sum_{\bar{n}_1+1}^{n_1}(R_i^2 - \tilde{R}_i^2)\cdot \mathbb{1}(\bar{n}_1 \le n_1)}_{:=\tilde{\Delta}_2}.$$

Using the law of total expectation, we obtain

$$\begin{aligned}
\mathbb{E}[|\tilde{\Delta}_1|] &= \mathbb{E}\left[\Big|\frac{1}{n}\sum_{i=n_1+1}^{\bar{n}_1}(R_i^2 - \tilde{R}_i^2)\cdot \mathbb{1}(\bar{n}_1 > n_1)\Big|\right] \\
&\le \mathbb{E}\left[\frac{\bar{n}_1 - n_1}{n}\cdot \frac{1}{\bar{n}_1 - n_1}\sum_{i=n_1+1}^{\bar{n}_1} \mathbb{1}(\bar{n}_1 > n_1)\cdot \mathbb{E}\big\{|R_i^2 - \tilde{R}_i^2| \,|\, \bar{n}_1\big\}\right] \\
&\le \frac{2}{n}\mathbb{E}[|\bar{n}_1 - n_1|]\mathrm{Var}(Y_1) \le \frac{\mathrm{Var}(Y_1)}{\sqrt{n}},
\end{aligned}$$

where the first inequality is derived from Jensen's inequality, and the remaining steps follow from the previous results. A similar argument applies to $\tilde{\Delta}_2$. By Markov's inequality, $(\mathbb{I})$ converges to zero. For the second term $(\mathbb{II})$, we have

$$\left|\Big(\frac{1}{n}\sum_{r=1}^n R_r\Big)^2 - \Big(\frac{1}{n}\sum_{r=1}^n \tilde{R}_r\Big)^2\right| \le \left|\frac{1}{n}\sum_{r=1}^n R_r + \frac{1}{n}\sum_{r=1}^n \tilde{R}_r\right| \cdot \left|\frac{1}{n}\sum_{r=1}^n R_r - \frac{1}{n}\sum_{r=1}^n \tilde{R}_r\right|,$$

which can be shown to converge to zero in probability using the previous results.

3. **Terms (c) and (d): $\hat{\sigma}_R^2 - \sigma_{\tilde{R}}^2$ and $\hat{\sigma}_{\tilde{R}}^2 - \sigma_{\tilde{R}}^2$.** We can see that the term (d) converges to zero in probability by the conventional law of large numbers. The term (c) also converges to zero as well since $\hat{\sigma}_R^2 - \sigma_{\tilde{R}}^2 = (\mathrm{c})' + (\mathrm{c})''$, where $(\mathrm{c})' = \hat{\sigma}_R^2 - \hat{\sigma}_{\tilde{R}}^2$ and $(\mathrm{c})'' = \hat{\sigma}_{\tilde{R}}^2 - \sigma_{\tilde{R}}^2$. In fact, $(\mathrm{c})' = (\mathrm{b})$ and $(\mathrm{c})'' = (\mathrm{d})$, and both are known to converge to zero in probability based on the previous results.

This completes the proof which shows that T_n and $\tilde{T}_n$ are asymptotically equivalent for the stable case. Hence T_n is stable in the sense of Definition 1.

B Supporting Lemmas

In this section, we collect several lemmas from the existing literature for completeness. The proof of the following lemma can be found, for example, in Mulzer (2018).

Lemma 1. *Let $Z_1, \ldots, Z_n$ be i.i.d. Bernoulli random variables with success probability $p \in [0,1]$ and $S_n = \sum_{i=1}^n Z_i$. For any $\delta \in [0,1]$, it holds that*

$$\mathbb{P}\{S_n \ge (1+\delta)np\} \le e^{-\frac{np\delta^2}{3}} \quad \text{and} \quad \mathbb{P}\{S_n \le (1-\delta)np\} \le e^{-\frac{np\delta^2}{3}}.$$

The following is the uniform central limit theorem result in Shah and Peters (2020, Lemma 18).

Lemma 2. *(Shah and Peters, 2020, Lemma 18) Let $\mathcal{P}$ be a family of distributions for a random variable $\zeta \in \mathbb{R}$ and suppose that $\zeta_1 \zeta_2, \ldots$ are i.i.d. copies of ζ. For each $n \in \mathbb{N}$, let $S_n := n^{-1/2} \sum_{i=1}^n \zeta_i$. Suppose that for all $P \in \mathcal{P}$, we have $\mathbb{E}_P(\zeta) = 0$, $\mathbb{E}_P(\zeta^2) = 1$ and $\mathbb{E}_P(|\zeta|^{2+\eta}) < c$ for some $\eta, c > 0$. We have that*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}} \left|\mathbb{P}_P(S_n \le t) - \Phi(t)\right| = 0.$$

The next lemma corresponds to Lundborg et al. (2022, Lemma S8) on conditional uniform central limit theorem.

Lemma 3. *(Lundborg et al., 2022, Lemma S8) Let $(X_{n,i})_{n\in\mathbb{N}, i\in[n]}$ be a triangular array of real-valued random variables and let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be a filtration on $\mathcal{F}$. Assume that*

1. *$X_{n,1}, \ldots, X_{n,n}$ are conditionally independent given $\mathcal{F}_n$, for each $n \in \mathbb{N}$;*
2. *$\mathbb{E}_P(X_{n,i} \,|\, \mathcal{F}_n) = 0$ for all $n \in \mathbb{N}, i \in [n]$;*
3. *$\left|n^{-1} \sum_{i=1}^n \mathbb{E}_P(X_{n,i}^2 \,|\, \mathcal{F}_n) - 1\right| = o_{\mathcal{P}}(1)$;*
4. *there exists $\delta > 0$ such that*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_P\left(|X_{n,i}|^{2+\delta} \,|\, \mathcal{F}_n\right) = o_{\mathcal{P}}(n^{\delta/2}).$$

Then $S_n = n^{-1/2} \sum_{m=1}^n X_{n,m}$ converges uniformly in distribution to $N(0,1)$, i.e.

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{x\in\mathbb{R}} |\mathbb{P}_P(S_n \le x) - \Phi(x)| = 0.$$

The next lemma corresponds to Lundborg et al. (2022, Lemma 20) on uniform Slutsky's theorem.

Lemma 4. *(Shah and Peters, 2020, Lemma 20) Let $\mathcal{P}$ be a family of distributions that determines the law of a sequences $(V_n)_{n\in\mathbb{N}}$ and $(W_n)_{n\in\mathbb{N}}$ of random variables. Suppose*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}} |\mathbb{P}_P(V_n \le t) - \Phi(t)| = 0.$$

Then we have the following.

(a) If $W_n = o_{\mathcal{P}}(1)$, we have $\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}} |\mathbb{P}_P(V_n + W_n \le t) - \Phi(t)| = 0$.

(b) If $W_n = 1 + o_{\mathcal{P}}(1)$, we have $\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}} |\mathbb{P}_P(V_n / W_n \le t) - \Phi(t)| = 0$.

The following two lemmas provide moment bounds for canonical U-statistics. Lemma 5 gives a general inequality for decoupled kernels (Giné et al., 2000), while Lemma 6 specializes this to the block statistic $T_{1,1}$, showing that its $(2+\delta)$-moment is controlled by the kernel moment. Standard decoupling results (de la Peña and Montgomery-Smith, 1995; de la Peña and Giné, 1999) then extend the bound to the original statistic.

Lemma 5. *(Giné et al., 2000, Equation (3.3)) Let $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded canonical kernel, and define $h_{i,j} := h(X_i^{(1)}, X_j^{(2)})$ for $1 \le i, j \le n$, where $\{X_i^{(1)}\}_{i=1}^n$ and $\{X_j^{(2)}\}_{j=1}^n$ are independent sequences of random variables. Then, for any $p \ge 2$, there exists a constant $C := C_p > 0$ such that*

$$\mathbb{E}\Big|\sum_{i=1}^n\sum_{j=1}^n h_{i,j}\Big|^p \le C_p \max\Bigg\{p^p\Bigg(\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\, h_{i,j}^2\Bigg)^{p/2}, p^{3p/2}\,\mathbb{E}_1 \max_{1\le i\le n}\Bigg(\sum_{j=1}^n \mathbb{E}_2\, h_{i,j}^2\Bigg)^{p/2},$$
$$p^{3p/2}\,\mathbb{E}_2 \max_{1\le j\le n}\Bigg(\sum_{i=1}^n \mathbb{E}_1\, h_{i,j}^2\Bigg)^{p/2}, p^{2p}\,\mathbb{E} \max_{1\le i,j\le n} |h_{i,j}|^p\Bigg\}.$$

Here, $\mathbb{E}_1$ and $\mathbb{E}_2$ denote expectations with respect to $X^{(1)}$ and $X^{(2)}$, respectively.

Lemma 6 (Moment bound for $T_{1,1}$ in Appendix A.5)**.** *Let $H : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ be a symmetric canonical kernel. Define the decoupled statistic*

$$T_{1,1}^{\mathrm{dec}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{B} H(W_i^{(1)}, W_j^{(2)}),$$

where $\{W_i^{(1)}\}_{i=1}^{B}$ and $\{W_j^{(2)}\}_{j=1}^{B}$ are independent i.i.d. samples from P. Then, for any $\delta > 0$, there exists a constant $C_\delta > 0$, independent of B, such that

$$\mathbb{E}\big[|T_{1,1}^{\mathrm{dec}}|^{2+\delta}\big] \;\leq\; C_\delta \, \mathbb{E}[|H(W_1, W_2)|^{2+\delta}].$$

Moreover, by classical decoupling inequalities (see de la Peña and Montgomery-Smith, 1995; de la Peña and Giné, 1999), the same bound also holds for the original U-statistic

$$T_{1,1} = \frac{2}{B(B-1)} \sum_{1 \leq i < j \leq B} H(W_i, W_j),$$

up to a universal constant independent of B.

Proof. Fix $p = 2 + \delta$ for some $\delta > 0$. Applying Lemma 5 with $h_{i,j} := H(W_i^{(1)}, W_j^{(2)})$, we obtain

$$\mathbb{E}\Big|\sum_{i=1}^{B} \sum_{j=1}^{B} h_{i,j}\Big|^{p} \;\lesssim\; A_1 + A_2 + A_3 + A_4,$$

where we omit multiplicative constants that do not depend on B. In deriving this inequality, we also used the elementary bound $\max\{a_1, \ldots, a_m\} \leq a_1 + \cdots + a_m$. The four terms are given by

$$A_1 = \Big(\sum_{i=1}^{B} \sum_{j=1}^{B} \mathbb{E}[h_{i,j}^2]\Big)^{p/2}, \qquad A_2 = \mathbb{E}_1\Big[\max_{1 \leq i \leq B} \Big(\sum_{j=1}^{B} \mathbb{E}_2[h_{i,j}^2]\Big)^{p/2}\Big],$$

$$A_3 = \mathbb{E}_2\Big[\max_{1 \leq j \leq B} \Big(\sum_{i=1}^{B} \mathbb{E}_1[h_{i,j}^2]\Big)^{p/2}\Big], \qquad A_4 = \mathbb{E}\Big[\max_{1 \leq i,j \leq B} |h_{i,j}|^p\Big].$$

We now bound each term:

1. **Term A_1.** By the i.i.d. structure,

$$\sum_{i=1}^{B} \sum_{j=1}^{B} \mathbb{E}[h_{ij}^2] = B^2 \, \mathbb{E}[h_{12}^2],$$

and therefore

$$A_1 \;\lesssim\; B^p \cdot (\mathbb{E}[h_{12}^2])^{p/2} \;\leq\; B^p \cdot \mathbb{E}[|h_{12}|^p],$$

where the last inequality follows from Jensen's inequality.

2. **Term A_2. and** A_3 Replacing the maximum with a sum, for term A_2

$$\mathbb{E}_1\Big[\max_{1 \leq i \leq B} \Big(\sum_{j=1}^{B} \mathbb{E}_2[h_{i,j}^2]\Big)^{p/2}\Big] \leq \sum_{i=1}^{B} \mathbb{E}_1\Big(\sum_{j=1}^{B} \mathbb{E}_2[h_{12}^2]\Big)^{p/2}.$$

By symmetry in i, this equals

$$B \cdot \mathbb{E}_1\Big(\sum_{j=1}^{B} \mathbb{E}_2[h_{1j}^2]\Big)^{p/2}.$$

Since $p > 2$ and $x \mapsto x^{p/2}$ is convex, Jensen's inequality yields

$$\Big(\sum_{j=1}^{B} \mathbb{E}_2[h_{1j}^2]\Big)^{p/2} \le B^{p/2-1} \sum_{j=1}^{B} \big(\mathbb{E}_2[h_{1j}^2]\big)^{p/2}.$$

Each summand is identical by i.i.d. structure, hence

$$\mathbb{E}_1\Big(\sum_{j=1}^{B} \mathbb{E}_2[h_{1j}^2]\Big)^{p/2} \le B^{p/2}(\mathbb{E}[h_{12}^2])^{p/2}.$$

Therefore, again by Jensen's inequality

$$A_2 \;\le\; B^{1+p/2} \cdot \mathbb{E}[|h_{12}|^p].$$

The same bound applies to A_3 by symmetry.

3. **Term A_4.** Again replacing the maximum with a sum yields

$$\mathbb{E}\Big[\max_{1\le i,j\le B} |h_{ij}|^p\Big] \le B^2\, \mathbb{E}|h_{12}|^p,$$

thus

$$A_4 \;\le\; B^2\, \mathbb{E}|h_{12}|^p.$$

Combining the above estimates, we obtain

$$\mathbb{E}\Bigg|\sum_{i=1}^{B}\sum_{j=1}^{B} h_{ij}\Bigg|^p \;\lesssim\; B^p \cdot \mathbb{E}[|h_{12}|^p].$$

Normalizing by B^{-p}, we conclude

$$\mathbb{E}\big|T_{1,1}^{\mathrm{dec}}\big|^p \;\lesssim\; \mathbb{E}[|h_{12}|^p].$$

Finally, classical decoupling inequalities (see de la Peña and Montgomery-Smith, 1995; de la Peña and Giné, 1999) ensure that the same bound applies to the original U-statistic $T_{1,1}$ (up to a universal constant which is independent of B). ☐

The following lemma provides a Berry–Esseen type bound for degenerate U-statistics, adapted from Yanushkevichiene et al. (2016, Theorem 1).

Lemma 7. *(Yanushkevichiene et al., 2016, Theorem 1) Let $X_1, \dots, X_n$ be i.i.d. random variables with distribution P on a sample space $\mathcal{X}$. Consider a symmetric kernel $h : \mathcal{X}^2 \to \mathbb{R}$ that is* degenerate*, in the sense that*

$$\mathbb{E}[h(x, X)] = 0 \quad \text{for every } x \in \mathcal{X}.$$

The corresponding second-order U-statistic is

$$U_n \;=\; \frac{1}{2} \cdot \frac{1}{n} \sum_{1\le i<j\le n} h(X_i, X_j).$$

Let Q be the Hilbert–Schmidt operator associated with h, with eigenvalues $(\lambda_\ell)_{\ell\ge1}$ ordered so that $|\lambda_1| \ge |\lambda_2| \ge \cdots$. Assume:

1. $\beta = \big(\mathbb{E}\,|h(X_1, X_2)|^{18/5}\big)^{5/36} < \infty.$

2. $|\lambda_1| \geq |\lambda_2| > 0$.

Define the limiting random variable

$$U_0 = \sum_{\ell=1}^{\infty} \lambda_\ell \, (a_\ell^2 - 1),$$

where $\{a_\ell\}_{\ell \geq 1} \overset{\text{i.i.d.}}{\sim} N(0,1)$. *Let* $F(t) = \mathbb{P}(U_n \leq t)$, $F_0(t) = \mathbb{P}(U_0 \leq t)$ *and*

$$\Delta_n = \sup_{t \in \mathbb{R}} |F(t) - F_0(t)|.$$

Then there exists a constant $c > 0$ *such that*

$$\Delta_n \leq c \frac{1}{\sqrt{|\lambda_1 \lambda_2|}} \left(\frac{\beta}{n^{1/6}} + \frac{\sum_{\ell \geq 1} |\lambda_\ell|}{n^{1/2}} \right).$$

The following lemma is central to establishing the hardness result under the specified marginal distribution. While we adapt the argument to our setting, the main idea is based on Neykov et al. (2021).

Lemma 8. *(Neykov et al., 2021, Modified version of Lemma A.1) Suppose* $(Z, Y, X) \in \{1,2\} \times \mathbb{R}^{d_Y + d_X}$ *has a distribution supported either on* $\{1,2\} \times [-M, M]^{d_Y + d_X}$ *for some* $M \in (0, \infty)$, *or on* $\{1,2\} \times (-\infty, \infty)^{d_Y + d_X}$. *Let* $\{(Z_i, Y_i, X_i)\}_{i \in [n]}$ *be i.i.d. copies of* (Z, Y, X), *with* $\mathbb{P}(Z = 1) = \lambda_n \in (0,1)$ *and* $\mathbb{P}(Z = 2) = 1 - \lambda_n$. *Then, for any* $\delta > 0$, *there exists a constant* $C := C(\delta) > 0$ *such that for any* $\varepsilon > 0$ *and any Borel set*

$$D \subseteq (\{1,2\} \times \mathbb{R}^{d_Y + d_X})^n \times [0,1],$$

it is possible to construct an i.i.d. sequence $\{(\widetilde{Z}_i, \widetilde{Y}_i, \widetilde{X}_i)\}_{i \in [n]}$ *satisfying* $\widetilde{Z}_i \perp \widetilde{Y}_i \mid \widetilde{X}_i$ *for all* $i \in [n]$ *and the following two properties.*

(i) First, with probability at least $1 - \delta$, *the modified sequence is close to the original sample in the sense that*

$$\mathbb{P}\Big(\max_{i \in [n]} \big\| (\widetilde{Y}_i, \widetilde{X}_i) - (Y_i, X_i) \big\| \infty < \varepsilon, \ \widetilde{Z}_i = Z_i \text{ for all } i \Big) > 1 - \delta.$$

(ii) Second, if $U \sim \mathrm{Unif}[0,1]$ *is independent of* $\{(\widetilde{Z}_i, \widetilde{Y}_i, \widetilde{X}_i)\}_{i \in [n]}$, *then the joint probability satisfies the inequality*

$$\mathbb{P}\Big(\big(\{(\widetilde{Z}_i, \widetilde{Y}_i, \widetilde{X}_i)\}_{i \in [n]}, U \big) \in D \Big) \leq C \cdot \mu(D),$$

where μ *denotes the product of counting measure on* $\{1,2\}^n$ *and Lebesgue measure on* $\mathbb{R}^{(d_X + d_Y)n} \times [0,1]$.

Proof of Lemma 8. To simplify the presentation, we focus on the case where $d_X = d_Y = 1$. The general case with higher dimensions can be handled similarly, requiring only minor adjustments. As in the proof of Shah and Peters (2020), it is enough to establish the following key lemma, which forms the core of their argument. Most of the technical arguments and constructions in our proof are directly adapted from Neykov et al. (2021). Our main modification lies in specifying the marginal distribution of Z, thereby restricting the null distribution class. However, this specification does not affect the construction of the lemma, and the core argument remains intact.

Step 1 (Preparation) First consider the case where the support is $\{1,2\} \times (-\infty, \infty)^2$, i.e., $M = \infty$. For any $\delta > 0$, one can select $M' := M'(\delta) < \infty$ such that

$$\mathbb{P}\big(\|(Y,X)\|_\infty > M'\big) < \frac{\delta}{2n}.$$

Define the modified variables $(\bar{Z}, \bar{Y}, \bar{X})$ by setting $\bar{Z} := Z$ always, and setting $(\bar{Y}, \bar{X}) := (Y, X)$ when $\|(Y,X)\|_\infty \leq M'$, while replacing $(\bar{Y}, \bar{X})$ by an independent sample drawn uniformly from $[-M', M']^2$ otherwise. By the union bound,

$$\mathbb{P}\big(\forall i \in [n] : (\bar{Z}_i, \bar{Y}_i, \bar{X}_i) = (Z_i, Y_i, X_i)\big) > 1 - \frac{\delta}{2}.$$

Henceforth, we work with $\{(\bar{Z}_i, \bar{Y}_i, \bar{X}_i)\}_{i \in [n]}$ (denoted by $\{(Z_i, Y_i, X_i)\}_{i \in [n]}$ for convenience), and we will construct $\{(\widetilde{Z}_i, \widetilde{Y}_i, \widetilde{X}_i)\}_{i \in [n]}$ satisfying

$$\mathbb{P}\Big(\max_{i \in [n]} \big\|(\widetilde{Z}_i, \widetilde{Y}_i, \widetilde{X}_i) - (\bar{Z}_i, \bar{Y}_i, \bar{X}_i)\big\|_\infty < \varepsilon\Big) = 1.$$

Next, we assume that the conditional densities $p_{Y,X|Z=z}(y,x)$ are bounded by some constant $L := L(\delta)$, uniformly over $z \in \{1,2\}$. For any $\bar{L} > 0$, define

$$S_{\bar{L}} := \big\{(z,y,x) \mid p_{Y,X|Z=z}(y,x) > \bar{L}\big\}.$$

As $\bar{L} \to \infty$, $S_{\bar{L}} \downarrow \varnothing$. Hence, for any given δ, we can choose $\bar{L}(\delta)$ large enough so that

$$\mathbb{P}\Big((Z,Y,X) \in S^c_{\bar{L}(\delta)}\Big) > 1 - \frac{\delta}{2n}.$$

Construct $(\bar{Z}, \bar{Y}, \bar{X})$ by setting $\bar{Z} := z$, and $(\bar{Y}, \bar{X}) := (Y, X)$ if $(Z,Y,X) \in S^c_{\bar{L}(\delta)}$, while drawing $(\bar{Y}, \bar{X})$ uniformly from $[-M, M]^2$ otherwise. Then the resulting conditional densities are bounded by

$$L(\delta) := \bar{L}(\delta) + \frac{\delta}{2n(2M)^2}.$$

Therefore, we again have

$$\mathbb{P}\big(\forall i \in [n] : (\bar{Z}_i, \bar{Y}_i, \bar{X}_i) = (Z_i, Y_i, X_i)\big) > 1 - \frac{\delta}{2}.$$

Step 2 (Construction) Let $\{A_1, A_2\} = \{\{1\}, \{2\}\}$ denote the (trivial) partition of $\{1,2\}$ corresponding to the value of Z. Similarly, let $\{B_1, \ldots, B_m\}$ and $\{C_1, \ldots, C_m\}$ be equi-partitions of $[-M, M]$ into intervals of length $2M/m$. Divide each C_k further into m^2 sub-intervals of equal length, denoted by C_{ijk}, so that each small interval corresponds to a pair $\big(A_i, B_j\big)$, with $i \in \{1,2\}, j \in [m]$. Given a draw (Z,Y,X), we construct $(\widetilde{Z}, \widetilde{Y}, \widetilde{X})$ as follows. Suppose that $X \in A_i, Y \in B_j$, and $Z \in C_k$. Then we set $\widetilde{Z} := Z$, generate $\widetilde{X}$ uniformly in C_{ijk}, and generate $\widetilde{Y}$ uniformly in B_j. By construction, this guarantees $\widetilde{Z} \perp \widetilde{Y} \mid \widetilde{X}$. Moreover, it is clear that by construction

$$\mathbb{P}\Big(\max_{i \in [n]} \big\|(\widetilde{Y}_i, \widetilde{X}_i) - (Y_i, Z_i)\big\|_\infty < \frac{2M}{m},\ \widetilde{Z}_i = Z_i \text{ for all } i\Big) = 1.$$

Hence if we take m large enough so that $\frac{2M}{m} < \varepsilon$ we guarantee that (i) is satisfied. What is more may write out the density of $(\tilde{Z}, \tilde{Y}, \tilde{X})$ as

$$p_{\widetilde{Y},\widetilde{X}|\widetilde{Z}=i}(\widetilde{y}, \widetilde{x}) = \sum_{j,k} \frac{m^3}{(2M)^2} \mathbb{1}(\widetilde{y} \in B_j, \widetilde{x} \in C_{ijk}) \cdot \mathbb{P}(Y \in B_j, X \in C_k \mid Z = i).$$

Step 3 (showing part (ii)) Recall that we are assuming that the conditional density satisfies $p_{Y,Z|X=x}(y,z) \leq L$ for some constant $L > 0$. It is simple to see that the probability that $(X,Y) \in A_i \times B_j$ is bounded as

$$\sum_{z\in\{1,2\}} \int_{B_j \times [-M,M]} p_{Y,X|X=x}(y,x) \cdot p_z(z) dy dz \leq \frac{L\lambda_n (2M)^2}{m} + \frac{L(1-\lambda_n)(2M)^2}{m} = \frac{L(2M)^2}{m}.$$

It follows that if we have n observations $\{(Z_i, Y_i, X_i)\}_{i\in[n]}$, the probability that at least two points Y_k and Y_ℓ fall into the same interval B_j, for some $j \in [m]$, is bounded by

$$\left(\frac{L(2M)^2}{m}\right)^n \left[m^n - m(m-1)\cdots(m-n+1)\right] = \mathcal{O}\left(\frac{n^2 \left(L(2M)^2\right)^n}{m}\right).$$

Since the number of all possible arrangements with points falling into different intervals B_j is $m(m-1)\cdots(m-n+1)$, while the total number of possible arrangements for the n points is m^n. Denote by S the complement of this event. Note that when S occurs, all Y_i fall into different intervals B_j, and vice versa. Next, suppose that D is an arbitrary fixed Borel set. We have

$$\begin{aligned}\mathbb{P}\Big(\big((\widetilde{Z}_i, \widetilde{Y}_i, \widetilde{X}_i)_{i\in[n]}, U\big) \in D\Big) \leq\ & \mathbb{P}\Big(\big((\widetilde{Z}_i, \widetilde{Y}_i, \widetilde{X}_i)_{i\in[n]}, U\big) \in D \cap (S \times [0,1])\Big) \\ &+ \mathbb{P}\Big(\big((\widetilde{Z}_i, \widetilde{Y}_i, \widetilde{X}_i)_{i\in[n]}, U\big) \notin S \times [0,1]\Big)\end{aligned}$$

We already have a bound on the second term on the RHS above

$$\mathbb{P}\Big(\big((\widetilde{Z}_i, \widetilde{Y}_i, \widetilde{X}_i)_{i\in[n]}, U\big) \notin S \times [0,1]\Big) = \mathcal{O}\left(\frac{n^2\left(L(2M)^2\right)^n}{m}\right).$$

Suppose now that we randomize the assignment on the set C_{jk}. In other words, there is a permutation $\pi : [m] \mapsto [m]$ that assigns each interval B_j to a sub-interval $C_{\pi_j k}$. Denote by $(\widetilde{X}^\pi, \widetilde{Y}^\pi, \widetilde{Z}^\pi)$ the vectors generated in this manner. Clearly all properties described above hold for $\{(\widetilde{Z}_i^\pi, \widetilde{Y}_i^\pi, \widetilde{X}_i^\pi)\}_{i\in[n]}$ for any permutation π. We have that

$$\begin{aligned}&\frac{1}{m!}\sum_{\pi\in S_m} \mathbb{P}\Big((\{(\widetilde{Z}_i^\pi, \widetilde{Y}_i^\pi, \widetilde{X}_i^\pi)\}_{i\in[n]}, U) \in D \cap (S \times [0,1])\Big) \\ &\quad = \frac{1}{m!}\sum_{\pi\in S_m} \int D \cap (S \times [0,1)] \prod_{l\in[n]} \sum_{x_l\in\{1,2\}} \sum_{j_l,k_l} \frac{m^3}{(2M)^2} \mathbb{1}\big(\widetilde{z}_l = z_l, \widetilde{y}_l \in B_{j_l}, \widetilde{x}_l \in C_{\pi j_l k_l}\big) \\ &\quad \times \mathbb{P}\big(Z = z_l, Y \in B_{j_l}, X \in C_{k_l}\big) \\ &\quad \leq \frac{(Lm^2)^n}{m!} \sum_{\pi\in S_m} \int_{D\cap(S\times[0,1])} \sum_{\{x_l\},\{j_l\},\{k_l\}} \mathbb{1}\big(\widetilde{Z}_l^\pi = x_l, \widetilde{Y}_l^\pi \in B_{j_l}, \widetilde{X}_l^\pi \in C_{\pi_{j_l} k_l}\big).\end{aligned}$$

In the above summation, the sequences $\{j_l\}_{l\in[n]}, \{k_l\}_{l\in[n]}$ are sequences of n numbers from $[m]$. Since the integration is over the set $D \cap (S \times [0,1])$, all indices j_l must be distinct; otherwise, the integral is zero. Thus the summation is effectively over all sequences $\{j_l\}_{l\in[n]}, \{k_l\}_{l\in[n]}$ such that no two indices j_l, j_l are equal. Fixing π_{jl} for these n distinct values and permuting the remaining $m - n$ indices, the number of permutations is $(m-n)!$. Since

$$\prod_{l\in[n]} C_{k_l} = \prod_{l\in[n]} \sum_j C_{jk_l},$$

contains all unique permutations of n elements (and more), we obtain the following bound:

$$\begin{aligned}
&\frac{(Lm^2)^n}{m!}\sum_{j_l,k_l}\int_{D\cap(S\times[0,1])}\sum_{\pi\in S_m}\prod_{l\in[n]}\mathbb{1}\big(\widetilde{Y}_l^{\pi}\in B_{j_l},\widetilde{X}_l^{\pi}\in C_{\pi_{j_l}k_l}\big)\\
&\leq\frac{(Lm^2)^n(m-n)!}{m!}\sum_{j_l,k_l}\int_{D\cap(S\times[0,1])}\prod_{l\in[n]}\mathbb{1}\big(\widetilde{Y}_l^{\pi}\in B_{j_l},\widetilde{X}_l^{\pi}\in C_{k_l}\big)\\
&\leq\frac{(Lm^2)^n(m-n)!}{m!}\mu(D\cap(S\times[0,1]))\leq\frac{(Lm^2)^n(m-n)!}{m!}\mu(D).
\end{aligned}$$

Therefore, there exists a permutation π^* such that

$$\mathbb{P}\Big(\big(\{\big(\widetilde{Z}_i^{\pi^*},\widetilde{Y}_i^{\pi^*},\widetilde{X}_i^{\pi^*}\big)\}i\in[n],U\big)\in D\cap(S\times[0,1])\Big)\leq\frac{(Lm^2)^n(m-n)!}{m!}\mu(D).$$

Finally, taking m sufficiently large, we obtain

$$\mathbb{P}\big(\big(\{\big(\widetilde{X}_i^{\pi^*},\widetilde{Y}_i^{\pi^*},\widetilde{Z}_i^{\pi^*}\big)\}_{i\in[n]},U\big)\in D\big)\leq\frac{(Lm^2)^n(m-n)!}{m!}\mu(D)+\mathcal{O}\Big(\frac{n^2\big(L(2M)^2\big)^n}{m}\Big)\leq C\mu(D).$$

C Additional Numerical Experiments

In this section, we present the experimental details and additional numerical experiments not included in the main body of the paper. Specifically, we describe the experimental details in Appendix C.1 and provide additional simulation results, including an empirical analysis of the impact of density ratio estimation errors on DRT methods in Appendix C.2, a real data analysis for CIT methods in Appendix C.3, a sensitivity analysis of the application of Algorithm 1 to the CIT approach in Appendix C.4, and a sensitivity analysis of Algorithm 1 to adjustment parameter ε in Appendix C.5.

C.1 Experimental Details

We begin with the implementation details of our numerical experiments, including density ratio estimation techniques, linear-time MMD test, classifier-based test, and conditional independence testing approaches.

Density Ratio Estimation. We estimate the density ratio $r_X(x)$ defined in (4) using a probabilistic classification-based approach described in Sugiyama et al. (2010, Section 3). Specifically, we focus on two classifiers: linear logistic regression (LLR) and kernel logistic regression (KLR).

Given samples $\{(X_i^{(1)},Y_i^{(1)})\}_{i=1}^{n_1}\overset{\text{i.i.d.}}{\sim}P_{XY}^{(1)}$ and $\{(X_j^{(2)},Y_j^{(2)})\}_{j=1}^{n_2}\overset{\text{i.i.d.}}{\sim}P_{XY}^{(2)}$, consider $\{(X_i,\ell_i)\}_{i=1}^{n}$, where $(X_1,\ldots,X_n)=(X_1^{(1)},\ldots,X_{n_1}^{(1)},X_1^{(2)},\ldots,X_{n_2}^{(2)})$ and $(\ell_1,\ldots,\ell_{n_1},\ell_{n_1+1},\ldots,\ell_n)=(0,\ldots,0,1,\ldots,1)$ with $\ell_i=\mathbb{1}(i\geq n_1+1)$. Further denote $X_i=\big(X_i(1),\ldots,X_i(p)\big)^\top$ where p is the dimension of X_i and let $\boldsymbol{\beta}:=(\beta_0,\beta_1,\ldots,\beta_p)^\top$.

- For LLR method, we model the posterior probability as

  $$\eta(X_i;\boldsymbol{\beta})=\mathbb{P}(\ell=1\,|\,X_i)=\frac{1}{1+\exp\left(-\beta_0+\sum_{j=1}^{p}\beta_jX_i(j)\right)}.$$

  The estimated coefficients $\hat{\boldsymbol{\beta}}$ are obtained by minimizing the negative log-likelihood.

- For KLR method (Zhu and Hastie, 2005), we use $\eta(X_i;\boldsymbol{\beta})=1/\left(1+\exp\left(-\theta(X_i;\boldsymbol{\beta})\right)\right)$, where $\theta(X_i;\boldsymbol{\beta})=\beta_0+\sum_{j=1}^{p}\beta_jk(X_i(j),x)$ and $k(x,y)=\exp(-\|x-y\|^2/\sigma^2)$. The estimated coefficients $\hat{\boldsymbol{\beta}}$ are obtained by minimizing the following penalized negative log-likelihood:

  $$-\sum_{i=1}^{n}\left[\ell_i\theta(X_i;\boldsymbol{\beta})-\log\left(1+\exp\left(\theta(X_i;\boldsymbol{\beta})\right)\right)\right]+\frac{\lambda}{2}\|\theta\|_{\mathcal{H}_k}^2,$$

where $\mathcal{H}_k$ is the reproducing kernel Hilbert space generated by k and λ is a regularization parameter. The density ratio estimate is then:

$$\hat{r}_X(X_i) = \frac{n_2}{n_1} \cdot \frac{\eta(X_i; \hat{\boldsymbol{\beta}})}{1 - \eta(X_i; \hat{\boldsymbol{\beta}})}. \tag{15}$$

For the joint density ratio, we use (X_i, Y_i) instead of X_i alone. We set $\sigma^2 = 200$, following Hu and Lei (2024), and fix $\lambda = 0.0005$ throughout our simulations.

Linear-Time MMD Test. For the linear-time MMD tests, we use a Gaussian kernel with the bandwidth parameter fixed at 1 across all experiments. In the cross-validated version (†MMD-ℓ), we use 2-fold cross validation (i.e., $K = 2$) with an equal splitting ratio.

Classifier-based Test. As mentioned in Section 4.1, under the balanced-sample setting, the Bayes optimal classifier is defined as:

$$h^*(y, x) \coloneqq \mathbb{1}\left(\frac{f_{YX}^{(1)}(y, x)}{f_{YX}^{(1)}(y, x) + f_{YX}(y, x)} > \frac{1}{2}\right).$$

This classifier can be equivalently expressed using density ratios:

$$h^\star(y, x) = \mathbb{1}\left(\frac{r_X(x)}{r_X(x) + r_{YX}(y, x)} > \frac{1}{2}\right),$$

where $r_X(x) = f_X^{(1)}(x)/f_X^{(2)}(x)$ as in (4) and $r_{YX}(y, x) \coloneqq f_{YX}^{(1)}(y, x)/f_{YX}^{(2)}(y, x)$. The empirical classifier is then defined as a plug-in estimator of $h^\star$:

$$\hat{h}(y, x) \coloneqq \mathbb{1}\left(\frac{\hat{r}_X(x)}{\hat{r}_X(x) + \hat{r}_{YX}(y, x)} > \frac{1}{2}\right), \tag{16}$$

where $\hat{r}_X$ and $\hat{r}_{YX}$ are the estimated marginal and joint density ratios, respectively, obtained using the classification-based approach described above.

The classifier $\hat{h}$ is constructed based on the training set, whereas the testing set is further split into two subsets with a ratio of 8:2. A larger subset is used for density ratio estimation, and the other subset is used for calculating the test statistic. In the cross-validated version (†CLF), we use 2-fold cross-validation as in †MMD-ℓ, maintaining 8:2 splitting ratio within each fold.

Randomized Conditional Independence Test. The RCIT method is implemented using the default hyperparameter settings specified in Strobl et al. (2019). Specifically, we use the default approximation method (Lindsay–Pilla–Basak method) for the null distribution. The number of random Fourier features is set to 100 for the conditioning set, and 5 for the non-conditioning sets.

Regression Methods for CIT. In our implementation of CIT approaches, we utilize several standard regression techniques. Table 1 below provides an overview of the key methods and their corresponding hyperparameter settings.
The code for reproducing all of our simulation results (including those in Appendices C.2 to C.5) and for more detailed settings is available on GitHub: `https://github.com/suman-cha/Cond2ST`.

Table 1: Description of regression methods used in CIT approach.

Regression Method	R Implementation	Tuning Parameters	Description
Random Forests	`ranger`	$mtry = \sqrt{p}$	# of variables to split at each node
XGBoost	`xgboost`	max depth $= 6$ $\eta = 0.3$	maximum tree depth learning rate

C.2 Impact of Density Ratio Estimation Errors on DRT Methods

To complement the real data analysis presented in Section 5.3, we conduct experiments to examine the relationship between density ratio estimation errors and type I errors of DRT methods. Figure 5 illustrates the log-scaled mean squared error (MSE) of the marginal density ratio r_X and the conditional density ratio $r_{Y|X}$ estimates for both the LLR and KLR methods across various sample sizes. Our experimental setup involves 500 simulations for each combination of sample size, dataset, and estimation method. We report the median MSE to provide a robust measure of estimation accuracy. For better visualization, the log-scaled MSE values are clipped: marginal density ratio errors above 10 are capped at 10 and conditional density ratio errors are limited to a maximum of 1. Notably, the true error values for the LLR method in high-dimensional settings, significantly exceed these clipped limits.

In the low-dimensional diamonds dataset, both LLR and KLR methods show relatively low MSE values for both marginal and conditional density ratio estimation. The performance gap between LLR and KLR methods diminishes as the sample size increases. This observation aligns with the findings in Section 5.3, where simpler methods like LLR suffice to control the type I error in low-dimensional settings. The low estimation errors explain their similar performance in such scenarios.

In contrast, the high-dimensional superconductivity dataset shows significant differences between the methods. KLR consistently outperforms LLR in both marginal and conditional density ratio estimation, maintaining low and stable MSE values across all sample sizes. On the other hand, LLR shows extremely high MSE values, particularly for small sample sizes. Although LLR shows some improvement as the sample size increases, it remains inferior to KLR in terms of estimation accuracy.

The high estimation errors for LLR in high-dimensional settings, even beyond the clipping applied, account for poor type I error control observed in Figure 4. These results highlight the need for more advanced density ratio estimation techniques to ensure the validity of tests in complex and high-dimensional scenarios.

C.3 Real Data Analysis for CIT Methods

We present the results for CIT methods applied to the diamonds and superconductivity datasets, complementing the analysis discussed in Section 5.3. Figure 6 shows the rejection rates for these methods under both the null and alternative hypotheses across various sample sizes.

For the low-dimensional diamonds dataset, the CIT methods generally exhibit good type I error control, with rejection rates close to the significance level $\alpha = 0.05$ under the null hypothesis. Under the alternative hypothesis, we observe increasing power for all methods except for WGSC as the sample size grows. Notably, RCIT and GCM show superior performance in terms of power.

In the high-dimensional superconductivity dataset, the performance of CIT methods is similar to that observed in the diamonds dataset, with no significant differences compared to DRT methods as shown in Figure 4. In terms of type I error control, GCM exhibits increasing rejection rates under the null hypothesis as the sample size grows. RCIT shows more inflated type I error, especially at small sample size. Regarding power, the CIT methods demonstrate relatively consistent performance across both datasets.

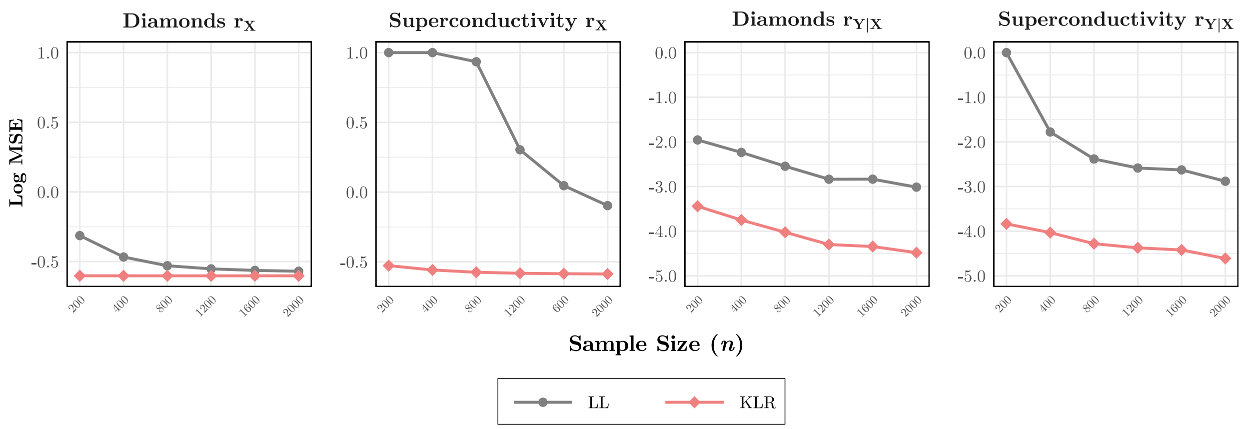

Figure 5: Log-scaled mean squared errors of marginal density ratio $r_X(x)$ (*left*) and conditional density ratio $r_{Y|X}(y|x)$ (*right*) estimates for LLR and KLR methods across various sample sizes. Results are shown for diamonds and superconductivity datasets, based on median values from 500 simulations under the null hypothesis.

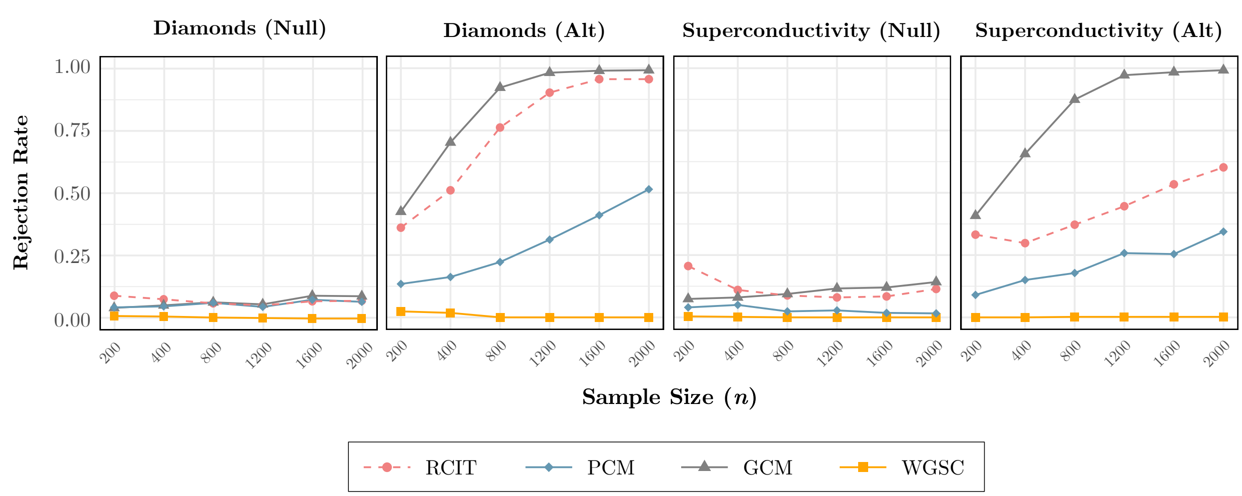

Figure 6: Rejection rates of CIT methods on the diamonds and superconductivity datasets under null and alternative hypotheses. Results are averaged over 500 repetitions with significance level $\alpha = 0.05$.

C.4 Sensitivity Analysis of CIT Methods to Algorithm 1

We examine the impact of Algorithm 1 on CIT methods across scenarios with unbounded marginal density ratios, as outlined in Section 5.2. Our analysis encompasses RCIT and the regression-based methods; PCM, GCM, and WGSC. The latter three are implemented using various regression models, such as linear models (`lm`), Random Forests (`rf`), and XGBoost (`xgb`). Tables 2 to 4 show the results for Scenarios 1(U), 2(U), and 3(U), respectively. In these tables, a checkmark (✓) indicates that Algorithm 1 is applied, whereas a cross (×) indicates it is not.

RCIT exhibits significant sensitivity to the application of Algorithm 1, particularly in Scenario 1(U). In this case, without the algorithm, the type I error rates of RCIT increase with sample size, whereas with the algorithm, these rates decrease as the sample size grows. This behavior highlights the potential stabilizing effect of Algorithm 1 on the performance of RCIT. On the other hand, GCM shows consistent performance across different regression methods, suggesting the robustness to the choice of underlying regression models. In contrast, the performance of PCM varies significantly with the choice of regression method. WGSC shows inconsistent patterns across scenarios, indicating potential sensitivity to specific data properties or model assumptions. This variability underscores the need for careful consideration when applying WGSC for conditional two-sample testing.

C.5 Sensitivity Analysis of Algorithm 1 to ε

We conduct a sensitivity analysis of Algorithm 1 with respect to the adjustment parameter ε. This parameter determines the size of the constructed testing dataset $\mathcal{D}_{\tilde{n}}$, through the equation $\tilde{n} = kn$, where k is a function of ε, as defined in Algorithm 1. The goal of this section is to examine the impact of different ε values on the performance of conditional two-sample testing via CIT methods. We consider 3 candidates for ε: $\{1/n, 1/\log(n), 1/\sqrt{\log(n)}\}$. The analysis covers scenarios with unbounded marginal density ratios, as described in Section 5.2, examining both null and alternative hypotheses across different sample sizes. The settings of these experiments correspond to the settings detailed in Section 5 with $n \in \{200, 500, 1000\}$.

Tables 5 to 7 present the results for each scenario, comparing the performance of RCIT, PCM, GCM, and WGSC across different ε values and sample sizes. Although the performance of the CIT methods remains relatively stable across different choices of ε, some variations in rejection rates are observed, particularly for small sample sizes.

Table 2: Performance comparison of CIT methods for Scenario 1(U) under null and alternative hypotheses. Rejection rates are provided for RCIT and regression-based methods (PCM, GCM, WGSC), implemented using `lm`, `rf`, and `xgb`. Results are shown for various sample sizes, with and without Algorithm 1.

n	Hypothesis	Algorithm 1	RCIT	PCM			GCM			WGSC		
				`lm`	`rf`	`xgb`	`lm`	`rf`	`xgb`	`lm`	`rf`	`xgb`
200	Null	✓	0.166	0.034	0.060	0.104	0.048	0.028	0.082	0.000	0.278	0.076
		×	0.202	0.044	0.064	0.098	0.032	0.036	0.074	0.000	0.062	0.052
	Alternative	✓	0.240	0.038	0.068	0.102	0.164	0.076	0.144	0.000	0.284	0.082
		×	0.362	0.046	0.078	0.104	0.300	0.178	0.268	0.000	0.082	0.074
500	Null	✓	0.056	0.066	0.066	0.074	0.036	0.024	0.060	0.000	0.248	0.050
		×	0.210	0.040	0.060	0.078	0.046	0.050	0.064	0.000	0.070	0.076
	Alternative	✓	0.228	0.070	0.092	0.100	0.356	0.170	0.296	0.000	0.252	0.048
		×	0.536	0.038	0.138	0.126	0.552	0.444	0.602	0.000	0.086	0.078
1000	Null	✓	0.064	0.050	0.050	0.060	0.048	0.024	0.060	0.000	0.270	0.056
		×	0.296	0.056	0.038	0.068	0.044	0.062	0.058	0.000	0.068	0.026
	Alternative	✓	0.394	0.050	0.098	0.124	0.586	0.262	0.516	0.000	0.268	0.054
		×	0.762	0.054	0.178	0.182	0.764	0.650	0.858	0.000	0.054	0.044
2000	Null	✓	0.048	0.054	0.030	0.056	0.046	0.028	0.062	0.000	0.282	0.016
		×	0.500	0.036	0.042	0.056	0.048	0.054	0.048	0.000	0.038	0.022
	Alternative	✓	0.610	0.050	0.112	0.110	0.796	0.502	0.820	0.000	0.282	0.024
		×	0.914	0.036	0.292	0.272	0.890	0.812	0.998	0.000	0.086	0.028

C.6 Additional Simulations

To provide a more comprehensive evaluation of the proposed methods, we extend our synthetic examples beyond the Gaussian covariate distributions to include heavy-tailed and non-Gaussian scenarios. These

Table 3: Performance comparison of CIT methods for Scenario 2(U) under null and alternative hypotheses. Rejection rates are provided for RCIT and regression-based methods (PCM, GCM, WGSC), implemented using `lm`, `rf`, and `xgb`. Results are shown for various sample sizes, with and without Algorithm 1.

n	Hypothesis	Algorithm 1	RCIT	PCM			GCM			WGSC		
				`lm`	`rf`	`xgb`	`lm`	`rf`	`xgb`	`lm`	`rf`	`xgb`
200	Null	✓	0.162	0.036	0.074	0.098	0.072	0.044	0.084	0.000	0.256	0.060
		×	0.116	0.046	0.048	0.090	0.048	0.058	0.110	0.000	0.056	0.040
	Alternative	✓	0.730	0.038	0.466	0.188	0.064	0.044	0.104	0.000	0.284	0.192
		×	0.808	0.050	0.504	0.254	0.054	0.050	0.110	0.000	0.200	0.174
500	Null	✓	0.074	0.056	0.060	0.104	0.044	0.046	0.094	0.000	0.270	0.042
		×	0.094	0.042	0.042	0.076	0.060	0.056	0.072	0.000	0.058	0.056
	Alternative	✓	0.906	0.056	0.864	0.564	0.054	0.050	0.092	0.000	0.320	0.330
		×	0.880	0.040	0.924	0.702	0.044	0.048	0.074	0.000	0.364	0.494
1000	Null	✓	0.086	0.046	0.032	0.070	0.048	0.020	0.056	0.000	0.272	0.030
		×	0.070	0.048	0.040	0.094	0.048	0.044	0.094	0.000	0.046	0.040
	Alternative	✓	0.968	0.046	0.990	0.974	0.052	0.026	0.066	0.000	0.398	0.604
		×	0.954	0.048	1.000	0.998	0.040	0.048	0.080	0.000	0.724	0.832
2000	Null	✓	0.048	0.036	0.030	0.064	0.032	0.018	0.048	0.000	0.272	0.030
		×	0.076	0.040	0.036	0.070	0.042	0.046	0.072	0.000	0.046	0.020
	Alternative	✓	0.980	0.034	1.000	1.000	0.030	0.020	0.044	0.000	0.486	0.900
		×	0.980	0.042	1.000	1.000	0.040	0.026	0.066	0.000	0.960	0.994

Table 4: Performance comparison of CIT methods for Scenario 3(U) under null and alternative hypotheses. Rejection rates are provided for RCIT and regression-based methods (PCM, GCM, WGSC), implemented using `lm`, `rf`, and `xgb`. Results are shown for various sample sizes, with and without Algorithm 1.

n	Hypothesis	Algorithm 1	RCIT	PCM			GCM			WGSC		
				`lm`	`rf`	`xgb`	`lm`	`rf`	`xgb`	`lm`	`rf`	`xgb`
200	Null	✓	0.166	0.048	0.062	0.100	0.052	0.028	0.096	0.000	0.254	0.064
		×	0.142	0.042	0.048	0.086	0.058	0.056	0.098	0.000	0.066	0.056
	Alternative	✓	0.688	0.046	0.726	0.578	0.214	0.204	0.266	0.000	0.406	0.636
		×	0.718	0.040	0.758	0.646	0.236	0.248	0.282	0.000	0.658	0.670
500	Null	✓	0.076	0.058	0.064	0.072	0.046	0.030	0.088	0.000	0.268	0.038
		×	0.066	0.042	0.034	0.070	0.046	0.050	0.080	0.000	0.056	0.060
	Alternative	✓	0.734	0.062	0.804	0.762	0.268	0.250	0.294	0.000	0.426	0.708
		×	0.764	0.046	0.808	0.794	0.264	0.270	0.292	0.000	0.742	0.754
1000	Null	✓	0.064	0.052	0.034	0.072	0.050	0.034	0.066	0.000	0.272	0.020
		×	0.086	0.048	0.046	0.072	0.036	0.036	0.080	0.000	0.052	0.036
	Alternative	✓	0.774	0.056	0.826	0.832	0.244	0.214	0.246	0.000	0.470	0.762
		×	0.818	0.048	0.830	0.838	0.236	0.236	0.268	0.000	0.800	0.806
2000	Null	✓	0.070	0.034	0.024	0.062	0.078	0.050	0.084	0.000	0.274	0.026
		×	0.082	0.046	0.036	0.084	0.076	0.068	0.098	0.000	0.044	0.024
	Alternative	✓	0.790	0.038	0.806	0.816	0.224	0.220	0.250	0.000	0.472	0.790
		×	0.810	0.044	0.814	0.824	0.232	0.232	0.268	0.000	0.816	0.802

Table 5: Sensitivity analysis of Algorithm 1 for Scenario 1(U) under null and alternative hypotheses. The table shows rejection rates of four CIT methods (RCIT, PCM, GCM, WGSC) for ε values and sample sizes.

n	Hypothesis	ε	RCIT	PCM	GCM	WGSC
200	Null	$1/n$	0.164	0.064	0.022	0.072
		$1/\log(n)$	0.166	0.064	0.036	0.076
		$1/\sqrt{\log(n)}$	0.168	0.072	0.018	0.062
	Alternative	$1/n$	0.220	0.070	0.080	0.068
		$1/\log(n)$	0.240	0.068	0.076	0.082
		$1/\sqrt{\log(n)}$	0.224	0.074	0.072	0.066
500	Null	$1/n$	0.090	0.038	0.026	0.052
		$1/\log(n)$	0.056	0.066	0.024	0.050
		$1/\sqrt{\log(n)}$	0.080	0.036	0.036	0.044
	Alternative	$1/n$	0.250	0.050	0.126	0.040
		$1/\log(n)$	0.228	0.092	0.170	0.048
		$1/\sqrt{\log(n)}$	0.244	0.062	0.166	0.046
1000	Null	$1/n$	0.046	0.052	0.030	0.038
		$1/\log(n)$	0.064	0.050	0.024	0.056
		$1/\sqrt{\log(n)}$	0.062	0.040	0.030	0.040
	Alternative	$1/n$	0.376	0.078	0.258	0.048
		$1/\log(n)$	0.394	0.098	0.262	0.054
		$1/\sqrt{\log(n)}$	0.394	0.078	0.288	0.038

additional scenarios are designed to assess performance when the marginal distributions of X deviate from normality.

Scenario 4: Linear Model with Heavy-tailed Covariates. This scenario extends Scenario 1 by replacing the Gaussian covariate distributions with heavy-tailed Student's t-distributions. For group $j \in \{1,2\}$, the covariates are generated as $X^{(j)} = \mu^{(j)} + \tau$, where $\tau \sim t_2$,$\mu^{(1)} = (1,1,-1,-1,0,\ldots,0)^\top$ and $\mu^{(2)} = \mathbf{0}_d$. The conditional distribution of the response follows $y^{(j)} \mid x^{(j)} = \delta^{(j)} + x^{(j)\top}\beta + \epsilon^{(j)}$, where $\beta = (1,-1,1,-1,0,\ldots,0)^\top$ and $\epsilon^{(j)} \sim N(0,1)$. Under the null hypothesis, we set $\delta^{(1)} = \delta^{(2)} = 0$, while for the alternative hypothesis, we introduce a mean shift by setting $\delta^{(1)} = 0$ and $\delta^{(2)} = 0.5$.

C.7 Computational Cost Analysis

This section summarizes the empirical computational costs of the methods considered in Section 5 and provides a brief discussion to guide practitioners. Unless stated otherwise, each result indicates the average elapsed time (in seconds) to run a test end-to-end. Each result is derived from 500 repetitions and is executed in R using a single thread.

Empirical Performance Analysis. Table 8 presents empirical computation time for Scenario 1(U) from Section 5.2 varying in $n \in \{200, 500, 1000, 2000\}$. DRT methods with LLR density ratio estimator show low computational costs, but DCP shows higher computational costs compared to other DRT methods because its cross-fitting and orthogonalization step. CIT methods show relatively higher costs than DRT methods.

Shifting to high-dimensional settings, Table 9 reports times for the superconductivity dataset. Here, DRT costs rise notably with KLR, reflecting its kernel computations in high dimensions, whereas LLR remains efficient. CIT methods show similar patterns, with `rf` again proving costliest. Figure 7 visualizes these trends on a log-scale, highlighting linear growth for most methods while underscoring estimator impacts.

Table 6: Sensitivity analysis of Algorithm 1 for Scenario 2(U) under null and alternative hypotheses. The table shows rejection rates of four CIT methods (RCIT, PCM, GCM, WGSC) for ε values and sample sizes.

n	Hypothesis	ε	RCIT	PCM	GCM	WGSC
200	Null	$1/n$	0.168	0.070	0.030	0.064
		$1/\log(n)$	0.162	0.074	0.044	0.060
		$1/\sqrt{\log(n)}$	0.164	0.068	0.028	0.064
	Alternative	$1/n$	0.694	0.418	0.026	0.200
		$1/\log(n)$	0.730	0.466	0.044	0.192
		$1/\sqrt{\log(n)}$	0.724	0.500	0.038	0.178
500	Null	$1/n$	0.080	0.054	0.020	0.046
		$1/\log(n)$	0.074	0.060	0.046	0.042
		$1/\sqrt{\log(n)}$	0.078	0.040	0.024	0.038
	Alternative	$1/n$	0.874	0.824	0.028	0.276
		$1/\log(n)$	0.906	0.864	0.050	0.330
		$1/\sqrt{\log(n)}$	0.902	0.812	0.030	0.324
1000	Null	$1/n$	0.094	0.044	0.028	0.038
		$1/\log(n)$	0.086	0.032	0.020	0.030
		$1/\sqrt{\log(n)}$	0.064	0.042	0.032	0.040
	Alternative	$1/n$	0.958	0.968	0.022	0.576
		$1/\log(n)$	0.968	0.990	0.026	0.604
		$1/\sqrt{\log(n)}$	0.964	0.980	0.022	0.626

These empirical patterns suggest that for time-sensitive applications, simpler estimators (e.g., LLR or `lm`) paired with efficient tests like MMD-ℓ or RCIT offer computational advantages, especially in high dimensions.

High-level takeaways.

- *Computationally efficient methods.* For a fixed kernel and linear density ratio estimator, the DRT methods MMD-ℓ, CLF, and CP exhibit empirical runtimes that scale approximately linearly in the sample size n. Cross-fitting (denoted † increases the computational time by a constant factor but preserves the linear scaling. DCP incurs a greater computational cost due to cross-fitting and an orthogonalization step.
- *Impact of estimator choice.* The choice of estimator is critical to performance. Within the DRT framework, KLR can be substantially slower than LLR, particularly in high dimensions. Similarly, for CITs, random forests (`rf`) are more computationally intensive than linear models (`lm`). For problems where the signal is approximately linear or low-dimensional, the simpler estimator (LLR, `lm`) offer a favorable balance between computational cost and statistical power.

Table 7: Sensitivity analysis of Algorithm 1 for Scenario 3(U) under null and alternative hypotheses. The table shows rejection rates of four CIT methods (RCIT, PCM, GCM, WGSC) for ε values and sample sizes.

n	Hypothesis	ε	RCIT	PCM	GCM	WGSC
200	Null	$1/n$	0.160	0.070	0.026	0.070
		$1/\log(n)$	0.166	0.062	0.028	0.064
		$1/\sqrt{\log(n)}$	0.112	0.060	0.026	0.076
	Alternative	$1/n$	0.688	0.704	0.216	0.626
		$1/\log(n)$	0.688	0.726	0.204	0.636
		$1/\sqrt{\log(n)}$	0.698	0.722	0.210	0.644
500	Null	$1/n$	0.074	0.038	0.006	0.048
		$1/\log(n)$	0.076	0.064	0.030	0.038
		$1/\sqrt{\log(n)}$	0.084	0.054	0.030	0.038
	Alternative	$1/n$	0.722	0.792	0.250	0.692
		$1/\log(n)$	0.734	0.804	0.250	0.708
		$1/\sqrt{\log(n)}$	0.730	0.782	0.252	0.694
1000	Null	$1/n$	0.076	0.042	0.038	0.048
		$1/\log(n)$	0.064	0.034	0.034	0.020
		$1/\sqrt{\log(n)}$	0.060	0.038	0.032	0.036
	Alternative	$1/n$	0.766	0.822	0.216	0.738
		$1/\log(n)$	0.774	0.826	0.214	0.762
		$1/\sqrt{\log(n)}$	0.768	0.822	0.208	0.760

Table 8: Average computation time (seconds) for Scenario 1(U) across sample sizes(n). Values are means for 500 repetitions. Methods marked † use 2-fold cross-fitting.

Method	**Sample Size(n)**			
	200	**500**	**1000**	**2000**
GCM	0.029	0.037	0.071	0.193
PCM	0.062	0.072	0.104	0.235
RCIT	0.017	0.031	0.049	0.089
WGSC	5.098	6.590	7.526	12.010
MMD-ℓ	0.005	0.008	0.012	0.022
CLF	0.008	0.011	0.017	0.028
†MMD-ℓ	0.010	0.015	0.024	0.044
†CLF	0.017	0.023	0.034	0.058
CP	0.007	0.012	0.021	0.047
DCP	0.071	0.243	1.090	3.895

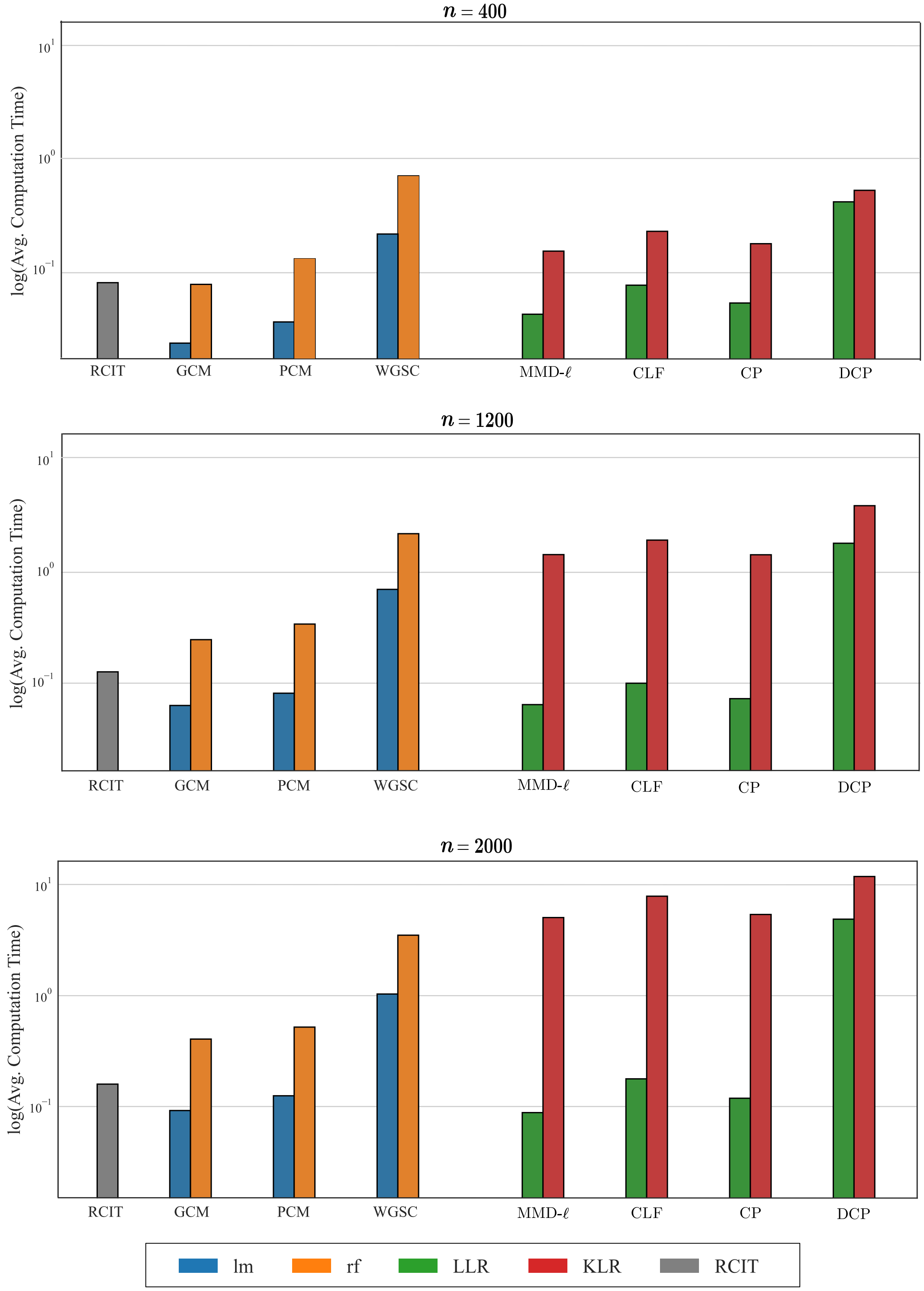

Figure 7: Average computation time (seconds) versus n on the high-dimensional Superconductivity dataset for all methods. Plots reflect the mean over 500 repetitions. Full details are provided in Appendix C.6.

Table 9: Average computation time (seconds) on the high-dimensional Superconductivity dataset. CIT methods are shown with linear models (`lm`) and random forests (`rf`). DRT methods are shown with linear logistic regression (LLR) and kernel logistic regression (KLR). Means are over 500 repetitions and † indicates 2-fold cross-fitting.

Method	**Estimator**	**Sample Size (n)**					
		200	**400**	**800**	**1200**	**1600**	**2000**
RCIT		0.041	0.072	0.106	0.117	0.164	0.159
GCM	`lm`	0.013	0.021	0.037	0.058	0.089	0.092
	`rf`	0.042	0.070	0.145	0.228	0.347	0.406
PCM	`lm`	0.031	0.032	0.051	0.075	0.085	0.125
	`rf`	0.084	0.118	0.197	0.316	0.360	0.520
WGSC	`lm`	0.130	0.198	0.396	0.649	0.837	1.035
	`rf`	0.420	0.662	1.319	2.064	2.748	3.515
MMD-ℓ	LLR	0.030	0.037	0.053	0.059	0.074	0.088
	KLR	0.043	0.139	0.661	1.337	2.862	5.069
†MMD-ℓ	LLR	0.063	0.071	0.111	0.112	0.161	0.195
	KLR	0.079	0.275	1.186	2.674	5.817	9.739
CLF	LLR	0.067	0.068	0.089	0.092	0.127	0.178
	KLR	0.061	0.209	0.799	1.806	4.251	7.902
†CLF	LLR	0.125	0.116	0.159	0.179	0.226	0.266
	KLR	0.117	0.335	1.414	3.633	8.026	13.501
CP	LLR	0.035	0.047	0.065	0.067	0.097	0.119
	KLR	0.046	0.162	0.670	1.332	3.178	5.405
DCP	LLR	0.245	0.387	0.868	1.692	3.200	4.903
	KLR	0.214	0.491	1.884	3.693	7.181	11.895