Notes on resubmitting revision made to manuscript BIOMTRKA-24-558

The authors would like to thank the editor and reviewers for their constructive comments and suggestions titled General Frameworks for Conditional Two-Sample Testing that have helped improve the quality of this manuscript. The resubmitted manuscript has undergone a thorough revision according to the editor and reviewers' comments. Please see our responses below. For the reviewers' convenience, we have highlighted significant changes in the revised manuscript in blue.

Associate Editor Comments

Associate Editor Comment — There is confusion and inconsistency in the setup and interpretation of the hardness results (Theorem 1 and 2) across conditional independence and conditional two-sample testing frameworks.

Reply: We thank the Associate Editor for this comment. We have addressed this issue in the revised manuscript as follows:

- Added a detailed analysis of user study results.
- Improved the evaluation of counterfactual explanations for user preferences.

Associate Editor Comment — The distinction between the two proposed testing settings is not always clearly maintained, leading to confusion (especially around Algorithm 1 and Theorem 2). The theoretical guarantees are sometimes unclear or possibly misstated.

Associate Editor Comment — Reviewers criticize key assumptions (particularly Assumptions 1(b) and 2(b)) as overly strong or unrealistic in practical scenarios.

Associate Editor Comment — There is a lack of guidance on test selection in practice, and insufficient discussion of stability, estimation effects, and computational costs. Reviewers request clarification on whether the proposed methods can be extended or adapted to other modern testing settings (e.g., local conditional tests).

Reviewer A

Reviewer Comment A.1 — I found the treatment of the relationship between conditional two-sample tests and conditional independence tests somewhat confusing and arbitrary at some places.

For example, the actual hardness result for conditional two-sample testing (Theorem 1) is already set up and formulated in the framework of a conditional independence test (in particular it includes the random variable Z taking values in $\{1, 2\}$). Only in the subsequent remark, it is argued that the type I error in the setting of the theorem is equivalent to the type I error when considering the conditional two-sample test. Would it not be more straightforward to formulate Theorem 1 really as a hardness result for conditional two-sample testing?

Reply: Thank you for the insightful comment. We agree that Theorem 1 is formally stated as a conditional independence testing problem with a binary variable $Z \in \{1, 2\}$, even though subsequent remark connects it to conditional two-sample testing. To clarify, under the assumption that the joint distribution of (X, Y, Z)admits a density, the two hypotheses

$$Y \perp\!\!\!\perp Z \mid X \quad \Longleftrightarrow \quad (Y \mid X, Z = 1) \stackrel{d}{=} (Y \mid X, Z = 2),$$

are statistically equivalent. Let $Z \in \{1, 2\}$ be a binary variable and suppose the joint distribution of (X, Y, Z) admits a density. And suppose $Y \perp\!\!\!\perp Z \mid X$, then by the definition of conditional independence via density functions, we have

$$p_{Y,Z \mid X}(y, z \mid x) = p_{Y \mid X}(y \mid x) \cdot p_{Z \mid X}(z \mid x).$$

Thus,

$$p_{Y|X,Z}(y \mid x, z) = \frac{p_{Y,Z|X}(y, z \mid x)}{p_{Z|X}(z \mid x)} = p_{Y|X}(y \mid x),$$

which implies

$$p_{Y|X,Z=1}(y \mid x) = p_{Y|X,Z=2}(y \mid x).$$

Conversely, suppose that $(Y|X, Z = 1) \stackrel{d}{=} (Y|X, Z = 2)$ which implies $p_{Y|X,Z=1}(y \mid x) = p_{Y|X,Z=2}(y \mid x) =: p_{Y|X}(y \mid x)$. This means the conditional density of Y given X and Z does not depend on Z, so

$$p_{Y|X,Z}(y \mid x, z) = p_{Y|X}(y \mid x).$$

Therefore

$$p_{Y,Z|X}(y, z \mid x) = p_{Y|X,Z}(y \mid x, z) \cdot p_{Z|X}(z \mid x) = p_{Y|X}(y \mid x) \cdot p_{Z|X}(z \mid x),$$

which re-establishes $Y \perp\!\!\!\perp Z \mid X$. Therefore, testing conditional independence and testing for equality of conditional distributions are equivalent when Z is binary and the densities exist. The reason we state Theorem 1 in the conditional independence framework is that it treats Z as a random variable, which is standard in the literature. In contrast, conditional two-sample testing typically assumes fixed group sizes. By conditioning on the sample sizes of $Z = 1$ and $Z = 2$, the two settings become identical. Hence, Theorem 1 can also be interpreted as a hardness result for conditional two sample testing without loss of generality.

Chan : Would it be clearer to state before Theorem 1 that conditional two-sample testing is equivalent to conditional independence testing, conditional on the sample sizes for $Z = 1$ and $Z = 2$?

Reviewer Comment A.2 — Also Theorem 2 considers the setting of conditional independence tests (with the random variable Z) but at the same time makes a statement about the output of Algorithm 1 But Algorithm 1 is formulated in the setting of conditional two-sample testing (where there is no random variable Z a priori but two different datasets), so there is a mismatch and Algorithm 1 not well-defined in the setting of Theorem 2. Would it not be better to formulate Theorem 2 in terms of conditional two-sample tests, i.e. a statement about the validity of Algorithm 1 for conditional two-sample testing?

In my opinion, the distinction between the two settings should be made with more care. For me it helped to look at Theorem 2 and Algorithm 1 in the following way (I do not claim that this is the way it should be done, but I hope it brings across my point)

Setting 1 (cond. two-sample test): $n_1, n_2 \in \mathbb{N}$ fixed. Observe two mutually independent samples $\left\{\left(X_i^{(1)}, Y_i^{(1)}\right)\right\}_{i=1}^{n_1} \sim P_{XY}^{(1)}$ and $\left\{\left(X_i^{(2)}, Y_i^{(2)}\right)\right\}_{i=1}^{n_2} \sim P_{XY}^{(2)}$. Sample random $\tilde{n}_1$ and $\tilde{n}_2$ as in Algorithm 1. If $\tilde{n}_1 > n_1$ or $\tilde{n}_2 > n_2$, accept H_0 (output 0). Else, set $\mathcal{D}_{\tilde{n}} = \left\{\left(X_i^{(1)}, Y_i^{(1)}, 1\right)\right\}_{i=1}^{\tilde{n}_1} \cup \left\{\left(X_i^{(2)}, Y_i^{(2)}, 2\right)\right\}_{i=1}^{\tilde{n}_2}$ and run the conditional independence test ϕ (that has output either 0 or 1) using $\mathcal{D}_{\tilde{n}}$ at level α.

Setting 2 (cond. independence test): Let $n = n_1 + n_2$ and consider a random vector $(X, Y, Z) \in \mathbb{R}^{d_X + d_Y} \times \{1, 2\}$, where Z takes values 1 and 2 with probability n_1/n and n_2/n, respectively. Let $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ be i.i.d. copies of (X, Y, Z) and let $N_1 = \sum_{i=1}^n 1\,(Z_i = 1)$ and $N_2 = \sum_{i=1}^n 1\,(Z_i = 2)$. Sample random $\tilde{n}_1$ and $\tilde{n}_2$ as in Algorithm 1. If $\tilde{n}_1 > N_1$ or $\tilde{n}_2 > N_2$, accept H_0. Else, let I_1 be the first $\tilde{n}_1$ indices i with $Z_i = 1$ and I_2 be the first $\tilde{n}_2$ indices i with $Z_i = 2$, set $\mathcal{D}_{\tilde{n}} = \{(X_i, Y_i, Z_i)\}_{i \in I_1 \cup I_2}$ and run the conditional independence test ϕ using $\mathcal{D}_{\tilde{n}}$ at level α.

Denote the output of setting 1 as ϕ_1 and the output of setting 2 as ϕ_2. Then, the distribution of ϕ_1 is equal to the distribution of $\phi_2 \mid N_1 = n_1, N_2 = n_2$. Note that setting 1 is the same setting as the problem setup in Section 1.1. and Algorithm 1. Setting 2 is the setup of Theorem 2 and what is meant with "the output of Algorithm 1" in Theorem 2. Maybe I am missing something, but I think that with the distinction into setting 1 and setting 2, one could formulate Theorem 2 and Algorithm 1 both in terms of setting 1 and then use the equality of the distributions of ϕ_1 and $\phi_2 \mid N_1 = n_1, N_2 = n_2$ for the proof.

Reviewer Comment A.3 — The discussion of why one cannot in general apply a conditional independence test to the full data, but needs to reduce the sample size to $\tilde{n}$ (discussion between lines 266 and 285), seems important to me. Perhaps I am missing something, but the presented coupling argument is a bit unclear to me as it compares two datasets of the same size. How exactly is this connected to the question of why one cannot apply the conditional independence test to the full dataset? Why can't one simply argue that since the number of discarded samples is $O\left((n \log(1/\epsilon))^{1/2}\right)$, if the test is not changed too much by $O\left(n^{1/2}\right)$ additional samples, it should be fine to use the full data? Also, Example 2 (unstable case) seems quite artificial. Is it possible to give another example or a better intuition? For example, from Section 3.5. in the Supplementary Material, one can see that RCIT (in contrast to GCM and PCM) is very sensitive to this issue. Is there an intuition why this is the case?

Reviewer Comment A.4 — Section 1.4., line 191, undefined reference.

Reply: We thank the reviewer for pointing this out. We have corrected the citation error in Section 1.4, line 191, by properly inserting the missing reference.

Reviewer Comment A.5 — In Algorithm 1 , shouldn't $\tilde{n} = k^* n$ be rounded to a whole number?

Reply: We appreciate the reviewer's observation. Indeed, in Algorithm 1, the expression $\tilde{n} = k^* n$ may not yield an integer value. In our implementation, we apply the floor function, i.e., $\tilde{n} = \lfloor k^* n \rfloor$, to ensure the subsample size is a valid integer. We believe this choice is more natural and practical, especially in simulation and empirical settings where the sample size must be an integer. All experiments reported in the paper were conducted using this floored version of $\tilde{n}$. We will clarify this detail in the revised version of the manuscript.

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

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

Chan : floor vs round?

Reviewer Comment A.6 — Line 272 , I don't really understand what is meant with "meaning the conditional testing errors of ϕ are asymptotically equivalent to its marginal errors".

Reply: Thank you for raising this important point. The phrase "the conditional testing errors of ϕ are asymptotically equivalent to its marginal errors" refers to a situation in which the Type I (or Type II) error rate of the testing function ϕ, when conditioned on the realized group sizes (i.e., the number of samples with $Z = 1$ and $Z = 2$), converges asymptotically to the marginal error rate, where group sizes are random and not fixed in advance.

This distinction is central to our framework because the conditional two-sample testing procedure is implemented via conditional independence testing, which typically assumes that the group sizes (e.g., n_1 samples with $Z = 1$ and n_2 with $Z = 2$) are fixed and known. To satisfy this requirement, Algorithm 1 introduces randomized subsampling to fix the number of samples from each group, thereby ensuring the validity of conditional inference. However, this subsampling inevitably leads to a reduced effective sample size, which can be a significant issue in finite-sample settings.

The natural question that arises is: Is such subsampling always necessary? The answer depends on the stability of the testing function ϕ—specifically, whether its behavior is sensitive to random fluctuations in group sizes.

When the test statistic underlying ϕ is stable, in the sense that small deviations in the sample size do not substantially affect its distribution under the null, the conditional validity of the test can be preserved even without subsampling. This is particularly relevant for methods like the Generalized Covariance Measure (GCM), where the performance of the test heavily relies on the quality of the estimated regression functions. In an oracle scenario—where these regression functions are known—the resulting test statistic exhibits stable behavior, and the conditional test behaves as though the group sizes were fixed, despite using all available data.

Therefore, in such cases, the conditional testing errors (i.e., conditional on group sizes) and the marginal testing errors (without subsampling) coincide in the asymptotic regime. Our formulation of stability is motivated by this oracle perspective, and the associated experiments (e.g., Section 3.5) are designed to illustrate how this equivalence manifests when the underlying components of the test are sufficiently accurate.

We will revise the manuscript to clarify this conceptual point and explicitly distinguish between the roles of the testing function ϕ, the test statistic it depends on, and the implications of stability in deciding whether Algorithm 1 is required.

Reviewer Comment A.7 — Line 302, shouldn't it be $\sum_{i=1}^{n} 1\,(Z_i = 1) = n_1$ instead of $\sum_{i=1}^{n} Z_i = n_1$ etc. in the definition of $\hat{f}$ and $\hat{g}$? (Z_i takes values in $\{1, 2\}$ and not in $\{0, 1\}$)

Reviewer Comment A.8 — The expression on line 345 is always ≥ 0. How can it have expectation zero under the null hypothesis?

Reviewer Comment A.9 — In equation (4), shouldn't 1 and 2 be swapped, since the true label for $V_i^{(1)}$ is 1 and the the true label for $V_i^{(2)}$ is 2?

Reply: We thank the reviewer for catching this. Indeed, the true label for $V_i^{(1)}$ should be 1, and that for $V_i^{(2)}$ should be 2. The current expression in Equation (4) mistakenly swaps the labels. We will revise the equation as follows:

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

We appreciate the reviewer's careful reading.

Reviewer Comment A.10 — I was first confused by the sampling mechanism in Section 5.2. Do I understand correctly that under the null hypothesis, the conditional distribution of $Y^{(1)}$ and $Y^{(2)}$ given $X^{(1)}$ and $X^{(2)}$, respectively is just the marginal distribution of Y ? And for the alternative hypothesis, there is introduced some artificial dependence of the conditional distribution of $Y^{(2)}$ given $X^{(2)}$? If this is correct, it would perhaps be helpful to the reader to explicitly state that the dependence between X and Y in the original dataset is completely destroyed and some artificial dependence is introduced.

Reviewer B

Reviewer Comment B.1 — The hardness result is unsurprising, as it has already been discussed in Remark 4 of Shah and Peters (2020). Nevertheless, Theorem 1 and its proof require refinement. In the proof, instead of considering n i.i.d. samples $\{(X_i, Y_i, Z_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} P$, the authors actually work with N i.i.d. samples $\{(X_i, Y_i, Z_i)\}_{i=1}^{N} \overset{\text{i.i.d.}}{\sim} P$. Then, under the event $\{N_1' = \sum_{i=1}^{N} \mathbb{1}(Z_i = 1) \geq n_1, N_2' = \sum_{i=1}^{N} \mathbb{1}(Z_i = 2) \geq n_2\}$, the authors use only n_1 observations with $Z = 1$ and n_2 observations with $Z = 2$ out of the N samples. Therefore, the resulting $n = n_1 + n_2$ data points are not i.i.d. from P, contrary to what is stated in Theorem 1. Moreover, the interpretation of the conditioning event $\{N_1 = \sum_{i=1}^{n} \mathbb{1}(Z_i = 1) = n_1, N_2 = \sum_{i=1}^{n} \mathbb{1}(Z_i = 2) = n_2\}$ in Theorem 1 is unclear, since the test ϕ is based on $n = n_1 + n_2$ data points out of N samples rather than n i.i.d. samples. In a nutshell, the test ϕ in the proof is a conditional independence test based on $\{(X_i, Y_i, Z_i)\}_{i=1}^{N} \overset{\text{i.i.d.}}{\sim} P$, which is not the same as that in Theorem 1. Additionally, the proof crucially relies on Theorem 2 and Remark 4 of Shah and Peters (2020), where no assumption is made about the marginal distribution of $Z \in \{1, 2\}$, i.e., $P(Z = 1)$. However, the authors assume $P(Z = 1) = n_1/n$ for n_1, n_2 as given in Theorem 1, thereby imposing a restriction on the null distributions $\mathcal{P}_{0,M}$. This raises the question of whether the hardness result in Shah and Peters (2020) remains valid for conditional independence testing when the marginal distribution of $Z \in \{1, 2\}$ is specified. While the answer appears to be affirmative, at least when X is continuous, a discussion on the implications of this restriction would be valuable.

- Check the proof of Neykov et al. (2021)

Reviewer Comment B.2 — The statements in Theorem 2 are somewhat confusing. In Algorithm 1, the authors construct $\mathcal{D}_{\widetilde{n}}$ of size $\widetilde{n}$ and then apply a conditional independence test ϕ using $\mathcal{D}_{\widetilde{n}}$. However, in Theorem 2, the test ϕ is applied to $\{(X_i, Y_i, Z_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} P$, which consists of n samples rather than $\widetilde{n}$. Besides, $\widetilde{\phi}$ is a conditional two-sample test based on the given two samples $\{(Y_i^{(1)}, X_i^{(1)})\}_{i=1}^{n_1}$ and $\{(Y_i^{(2)}, X_i^{(2)})\}_{i=1}^{n_2}$, rather than n i.i.d. samples $\{(X_i, Y_i, Z_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} P$. As shown in the proof, the test $\widetilde{\phi}$ can be expressed as $\widetilde{\phi} = \mathbb{1}(\widetilde{n}_1 \leq n_1)\, \mathbb{1}(\widetilde{n}_2 \leq n_2)\, \phi_{\widetilde{n}}$, where n_1 and n_2 are given. Consequently, the interpretation of the conditioning event $\{N_1 = \sum_{i=1}^{n} \mathbb{1}(Z_i = 1) = n_1, N_2 = \sum_{i=1}^{n} \mathbb{1}(Z_i = 2) = n_2\}$ with respect to $\widetilde{\phi}$ is unclear. Further clarification is necessary to ensure correctness.

Reviewer Comment B.3 — For the classifier-based approach, Assumption 1(b) is extremely restrictive. Even for a correctly specified parametric model $\widehat{r}_X$ (let alone more flexible nonparametric techniques), we have $E_P\left[\{\widehat{r}_X(X^{(2)}) - r_X(X^{(2)})\}^2 \mid \widehat{r}_X\right] = O_p(m^{-1})$, which implies that Assumption 1(b) is not satisfied. Consequently, unbalanced sample splitting becomes necessary, and taking $m = \lfloor n/2 \rfloor$ in practice is not justifiable. Notably, Assumption 2(b) in Hu and Lei (2024) is weaker than Assumption 1(b) in this paper; see the discussions after Theorem 1 in Hu and Lei (2024). A crucial question remains: how can the sample be split in a data-adaptive manner when unbalanced sample splitting is unavoidable?

- Check the technique of Guo and Shah (2024)

Reviewer Comment B.4 — For the MMD-based approach, the authors adopt a linear-time MMD statistic instead of the quadratic-time MMD statistic. In marginal two-sample testing, the quadratictime MMD test is known to achieve minimax optimal power. By contrast, the lineartime test generally exhibits lower power-by more than a constant factor-and has a worse rate (Ramdas et al., 2015), making it less favorable in practice. Although the standard permutation approach is not applicable to conditional two-sample testing, the authors should explore the strategies for calibrating the quadratic-time statistic, given its desirable theoretical properties.

Reviewer Comment B.5 — No single test exhibits consistent performance across all scenarios, often displaying er- ratic size or power behavior in certain cases. A crucial question in practice, therefore, is: which test should one use to draw reliable conclusions? More specifically, which of the two frameworks should be preferred? If opting for the conditional independence-based framework, which conditional independence test should be employed? Alternatively, if selecting the density ratio-based framework, should one use the classifier-based test or the MMD-based test?

Reviewer Comment B.6 — Two Gaussian distributions differing only in their means represent perhaps the easiest case, where a simple linear logistic regression model can correctly specify the density ratio. It would be valuable to assess the performance of the proposed tests when $X^{(1)}$ or $X^{(2)}$ is drawn from heavy-tailed or other non-Gaussian distributions.

Reviewer Comment B.7 — In Section 5.2 , since $X^{(2)}$ is sampled with probability proportional to $\exp\left(-x_1^2\right)$, the linear logistic regression (LLR) model misspecifies the density ratio. This misspecification may explain the poor performance of LLR on the superconductivity dataset. Nonetheless, an explanation of why LLR performs well on the diamonds dataset would be beneficial.

Reviewer Comment B.8 — The medium heuristic is arguably the most popular method for selecting the band- width. Is there a specific reason why the bandwidth for the linear-time MMD test is fixed at 1 across all experiments

Reviewer Comment B.9 — On line 347, why does the expectation of the rank sum statistic equal zero under the null hypothesis?

Reviewer C

Reviewer Comment C.1 — The core idea for using CIT based methods for the conditional two-sample problem is to use subsampling where on average $O(\sqrt{n\log(1/\varepsilon)})$ many samples are discarded, which can affect the finite sample performance. The authors do acknowledge this limitation and provides experiments for choices of ε. I believe this requires a more thorough discussion, for example one can compare the power of the proposed test and the corresponding CIT in an oracle setting where we generate $Z \in \{1, 2\}$ and then generate $(X, Y) \mid Z \sim P_{XY}^{(Z)}$.

Reviewer Comment C.2 — The discussion about stability of the CIT method provided after Algorithm 1 needs more clarity. It would be beneficial for the reader if the authors clarify what they mean by stability when this is first mentioned. To my understanding it appears that the current discussion provides an alternative approach to Algorithm 1 where no samples are discarded.

Reviewer Comment C.3 — Building on the previous comment, this alternative approach seems underdeveloped. While Algorithm 1 and Theorem 2 provides a general CIT based framework, this alternative approach seems highly dependent on the specific CIT method and its implementation. A case-by-case evaluation seems insufficient for building a general framework and it would be great to have more concrete guidelines for this alternative approach.

Reviewer Comment C.4 — Additionally, Example 1 assume knowledge of conditional expectations and Example 2 is an extreme case instability under estimation. I think a more detailed discussion is required on how estimation affects instability and as a consequence the alternative approach.

Reviewer Comment C.5 — Theorem 1 establishes that conditional two-sample testing is generally impossible (achieving non-trivial power) without further assumptions, implying that structural constraints (for example, smoothness conditions) are necessary. In the density ratio based approaches (henceforth abbreviated as DRT), it is not immediately clear how such structural assumptions are implicitly incorporated. It would be great if the authors can discuss how the assumptions for DRT (Assumptions 1 and 2, in particular to my understanding 1(b) and 2(b)) serve as necessary structural conditions to overcome the hardness from Theorem 1.

Reviewer Comment C.6 — Under the DRT setting, Assumptions 1(b) and 2(b) place restrictions on the convergence rates of the density ratio estimator. It would be beneficial for the reader if there is a discussion about whether these rates are achievable in practice, citing relevant sources and the corresponding estimators which achieve these rates.

Reviewer Comment C.7 — One of the specific DRT approaches is based on the kernel based MMD statistic. However, the authors only discuss the linear MMD statistic. It is well known that the quadratic MMD achieves much better power than the linear one. While the authors acknowledge the difficulty of implementing the quadratic one using permutation tests, is it possible to have a version of the quadratic MMD test and calibrate that using multiplier bootstrap as in Gretton et al. (2009)?

Reviewer Comment C.8 — Building on the previous point, the manuscript's discussion of permutation-based implementations for the quadratic-time MMD feels incomplete. Several permutation strategies already exist for conditional independence tests-for example, Berrett et al. (2020); Candes et al. (2018); Kim et al. (2022). Because Theorems 1 and 2 establish a link between conditional independence tests and conditional two-sample tests, it seems natural to ask whether those existing strategies could be adapted to produce a permutation test in the quadratic-time setting. If such an adaptation is difficult, a detailed explanation of the conceptual or computational obstacles would be helpful.

Reviewer Comment C.9 — Building on the authors' classifier-based conditional two-sample tests, recent work has focused on local conditional comparisons - specifically, testing whether $P^{(1)}_{Y|X}$ equals $P^{(2)}_{Y|X}$ at a given covariate value $X = x_0$ (see Linhart et al. (2023)). Could the proposed framework be adapted to develop such point-wise conditional tests?

Reviewer Comment C.10 — There appears to be some minor inconsistency between Theorem 2 and Algorithm 1 . Specifically, Algorithm 1 defines Z_i for $1 \leq i \leq \tilde{n}_1$ and similarly for $\tilde{n}_2$, but

Theorem 2 defined N_1 and N_2 using the full sample size n. I am not sure what happens for the discarded samples as those Z_i are never defined.

Reviewer Comment C.11 — At first glance, Theorem 2 is framed as an asymptotic result. Yet any finite-sample valid CIT ϕ would appear to satisfy the same assumptions. If that is correct, could the guarantees for Algorithm 1 be reformulated to hold in finite samples as well? A brief remark following Theorem 2 would help clarify this point.

Reviewer Comment C.12 — For completeness, please include the explicit refined bound for $k^\star$ in the supplementary material.

Reviewer Comment C.13 — A concise discussion of the computational costs associated with the different methods would be very helpful for practitioners

Reviewer Comment C.14 — A reference link seems to be broken at the end of page 5.

Reply: We appreciate the reviewer's careful reading. The broken reference link at the end of page 5 has been corrected in the revised version.

Reviewer Comment C.15 — Minor proofreading for typos/grammar is recommended.

References

Berrett, T. B., Wang, Y., Barber, R. J., 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.

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.

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

Guo, F. R. and Shah, R. D. (2024). Rank-transformed subsampling: inference for multiple data splitting and exchangeable p-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(1):256–286.

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.

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

Linhart, J., Gramfort, A., and Rodrigues, P. (2023). L-c2st: Local diagnostics for posterior approximations in simulation-based inference. In *Advances in Neural Information Processing Systems*, volume 36, pages 56384–56410.

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

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

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.