Testing conditional independence under isotonicity

Rohan Hore[1], Jake A. Soloff[1], Rina Foygel Barber[1] and Richard J. Samworth[2]

[1]*Department of Statistics, University of Chicago*

[2]*Statistical Laboratory, University of Cambridge*

November 10, 2024

Abstract

In this paper, we investigate the problem of testing conditional independence (CI) between two random variables X and Y given Z, under the assumption that X is stochastically increasing in Z. The hardness of testing the CI hypothesis is well documented in the literature. While existing approaches often rely on parametric models, smoothness assumptions, or approximations to the conditional distribution of X given Z and/or Y given Z, our test requires no such knowledge beyond a shape constraint. Inspired by the standard permutation method for unconditional independence testing, our procedure determines the significance of a statistic by randomly swapping the X values within ordered pairs of Z samples. The matched pairs and the test statistic may depend on both Y and Z, providing the analyst with significant flexibility in constructing a powerful test. Our test not only achieves finite-sample Type I error control, but also has non-trivial asymptotic power against alternatives that are not too close to the null models. We validate our theoretical findings through a series of simulations and real data experiments.

1 Introduction

Consider the problem of testing the conditional independence (CI) hypothesis

$$H_0^{\mathrm{CI}} : X \perp\!\!\!\perp Y \mid Z,$$

where X and Y are variables of interest (such as a treatment X and an outcome Y), while Z represents a (potentially high-dimensional) confounder. Our available data consist of a sample $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n) \overset{\text{iid}}{\sim} P$, where P is an unknown distribution on $(X, Y, Z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Throughout, we write $P_{X|Z}$ and $P_{Y|Z}$ to denote the conditional distributions of X given Z and Y given Z respectively.

In the case where the distribution of Z is continuous, Shah and Peters (2020) established that, without further assumptions, there is no universally valid test of H_0^{CI} that achieves non-trivial power for *any* alternative distribution $P \notin H_0^{\mathrm{CI}}$; see also Neykov, Balakrishnan and Wasserman (2021) and Kim et al. (2022). Existing approaches to testing conditional independence have therefore sought to guarantee validity (Type I error control) over restricted classes of null distributions that impose one of the following additional structures:

(a) a parametric model, such as joint Gaussianity of (X, Y, Z), or a Gaussian linear model for $Y \mid (X, Z)$ (Kalisch and Bühlmann, 2007);

(b) a known or well-estimated conditional distribution $P_{X|Z}$ (Candès et al., 2018; Barber, Candès and Samworth, 2020; Berrett et al., 2020; Niu et al., 2024); or

(c) smoothness of the conditional distribution $P_{X|Z}$ (Shah and Peters, 2020; Lundborg, Shah and Peters, 2022; Kim et al., 2022; Lundborg et al., 2024+).

1.1 Our contributions

In this work, we introduce a nonparametric structure under which we can test conditional independence: we assume a shape constraint—specifically, a form of stochastic monotonicity—for the conditional distribution of $X \mid Z$. Such a constraint is motivated by several applications, particularly in biomedicine, where for instance incidence of diabetes becomes more prevalent with age (Yan et al., 2023), and left ventricular wall thickness is a known risk factor for hypertrophic cardiomyopathy (O'Mahony et al., 2014). We observe a similar trend in economics, where greater professional experience is often linked to higher salaries. In agriculture, crop yields typically increase with optimal rainfall levels. Drawing insights from these real-world examples, we consider the following constraint:

Assumption 1 (Monotonicity of the conditional distribution $P_{X|Z}$)**.** *Let $\mathcal{X} \subseteq \mathbb{R}$ and let $\preceq$ be a partial order on $\mathcal{Z}$. We assume X is stochastically increasing in Z, meaning that*

$$\text{if } z \preceq z' \text{ then } \mathbb{P}_P\{X \geq x \mid Z = z\} \leq \mathbb{P}_P\{X \geq x \mid Z = z'\} \text{ for all } x.$$

This assumption does not fall into any of the categories (a)–(c) above. We will often consider the case where the control variable Z is univariate, $\mathcal{Z} \subseteq \mathbb{R}$, under the usual total order $\leq$. Our framework also allows for multivariate $Z \in \mathbb{R}^d$, in which case the most common partial order is the coordinatewise order.

Our main contribution is to introduce a broad strategy for testing the isotonic conditional independence (ICI) null hypothesis

$$H_0^{\mathrm{ICI}} : X \perp\!\!\!\perp Y \mid Z, \text{ and } P_{X|Z} \text{ satisfies Assumption 1.} \tag{1}$$

Naturally, this test should only be applied in settings where the monotonicity condition of Assumption 1 is well-motivated, so that a rejection of H_0^{ICI} can reasonably be interpreted as evidence that $X \not\perp\!\!\!\perp Y \mid Z$. However, we emphasize again that some additional assumption beyond H_0^{CI} is essential for any valid test to have non-trivial power at any alternative.

1.2 Background: testing independence

To set the stage for some of the notation and ideas underlying our methodology, we briefly review the simpler framework of permutation testing of the null hypothesis of marginal independence, $X \perp\!\!\!\perp Y$.

Given a joint distribution P on $\mathcal{X} \times \mathcal{Y}$, let $(X_i, Y_i)_{i\in[n]} \overset{\text{iid}}{\sim} P$, and write $\mathbf{X} = (X_i)_{i=1}^n$ and $\mathbf{Y} = (Y_i)_{i=1}^n$. We can reframe the problem of testing marginal independence as testing

whether the entries of $\mathbf{X}$ are i.i.d. given $\mathbf{Y}$. Specifically, permutation tests look for violations of exchangeability of $\mathbf{X}$ given $\mathbf{Y}$. The general approach proceeds as follows: based on $\mathbf{Y}$, the analyst chooses any statistic $T : \mathcal{X}^n \to \mathbb{R}$, with larger values of $T(\mathbf{X})$ indicating evidence against the independence null.[1] Write $\mathcal{S}_n$ for the set of permutations of $[n]$, and for $\sigma \in \mathcal{S}_n$, let $T_\sigma = T(\mathbf{X}^\sigma)$ denote the value of the statistic when the entries of $\mathbf{X}$ are permuted according to σ—that is, $\mathbf{X}^\sigma = (X_{\sigma(1)}, \ldots, X_{\sigma(n)})$. Finally, define a p-value

$$p = \frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} \mathbb{1}\left\{T_\sigma \geq T\right\}.$$

This construction produces a valid p-value for *any* choice of test statistic T, and the statistic T can be tailored to have power against certain specific alternatives. Indeed, this strategy has been successfully employed to construct independence tests via nearest neighbour distances and mutual information (Berrett and Samworth, 2019), moment methods (Kim, Balakrishnan and Wasserman, 2022), kernels (Pfister et al., 2018), the Hilbert–Schmidt independence criterion (Albert et al., 2022) and U-statistics (Berrett and Samworth, 2021; Berrett, Kontoyiannis and Samworth, 2021),

In our work, since we are interested in conditional (rather than marginal) independence, we will follow a similar strategy, but will use a restricted class of functions T and a (data-dependent) subgroup of permutations $\sigma \in \mathcal{S}_n$, both of which respect the stochastic monotonicity Assumption 1. Our framework still affords the analyst a great deal of flexibility in designing their test, while controlling Type I error across the more challenging null class H_0^{ICI}.

2 Methodology

In this section we give a general procedure, called the `PairSwap-ICI` test, for testing the isotonic conditional independence null H_0^{ICI}. Intuitively, it is plausible that we should be able to construct powerful tests against some alternatives. For example, if $Z_i \preceq Z_j$, then the shape constraint ensures that $X_i \leq X_j$ should hold *at least* half of the time; if we instead observe $X_i \gg X_j$, then this may be due to the influence of Y. Our test builds on and formalizes this intuition: after observing $\mathbf{Y}$ and $\mathbf{Z}$, the analyst specifies pairs (i, j) such that $Z_i \preceq Z_j$ and then may use large differences $X_i - X_j$ as evidence against the null.

More formally, based on $\mathbf{Y}$ and $\mathbf{Z}$, and without access to $\mathbf{X}$, the analyst chooses:

(i) A sequence of ordered pairs

$$(i_1, j_1), \ldots, (i_L, j_L)$$

of indices in $[n] = \{1, \ldots, n\}$, where all $2L$ entries are distinct. We require the pairs to be ordered in the sense that

$$Z_{i_\ell} \preceq Z_{j_\ell} \tag{2}$$

for each $\ell \in [L] = \{1, \ldots, L\}$. We refer to such a choice of ordered pairs (with any $L \leq \lfloor n/2 \rfloor$) as a *matching*, and denote the set of all possible matchings in $[n]$ satisfying (2) as $\mathcal{M}_n(\mathbf{Z})$.

[1] We emphasize that $T(\mathbf{X})$ is allowed to depend on both $\mathbf{X}$ and $\mathbf{Y}$—for instance we may define the function as $T(\mathbf{x}) = |\sum_{i=1}^n x_i Y_i|$, though we suppress the dependence on $\mathbf{Y}$ in our notation.

(ii) A sequence of functions $\psi_1, \ldots, \psi_L$, where each $\psi_\ell : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfies the *anti-monotonicity* property

$$\psi_\ell(x + \Delta, x' - \Delta') - \psi_\ell(x' - \Delta', x + \Delta) \geq \psi_\ell(x, x') - \psi_\ell(x', x), \tag{3}$$

for all $\Delta, \Delta' \geq 0$. An example of a class of functions ψ that satisfy anti-monotonicity is $\psi(x, x') = f(x - x')$ for any anti-symmetric and monotone nondecreasing function f, such as $f(x) = x$ or $f(x) = \text{sign}(x)$.

With these choices in place, our test statistic $T : \mathcal{X}^n \to \mathbb{R}$ is defined as[2]

$$T(\mathbf{x}) = \sum_{\ell=1}^{L} \psi_\ell(x_{i_\ell}, x_{j_\ell}). \tag{4}$$

In order to calibrate the test, the analyst compares the observed test statistic $T = T(\mathbf{X})$ with versions of T where indices within pairs (i_ℓ, j_ℓ) are randomly swapped. Specifically, for $\mathbf{s} \in \{\pm 1\}^L$, define $T_\mathbf{s} = T(\mathbf{X}^\mathbf{s})$, where $\mathbf{X}^\mathbf{s}$ is a swapped version of the data vector $\mathbf{X}$, with entries

$$\begin{cases} (X^\mathbf{s}_{i_\ell}, X^\mathbf{s}_{j_\ell}) = (X_{i_\ell}, X_{j_\ell}) & s_\ell = +1, \\ (X^\mathbf{s}_{i_\ell}, X^\mathbf{s}_{j_\ell}) = (X_{j_\ell}, X_{i_\ell}) & s_\ell = -1. \end{cases}$$

That is, $s_\ell = -1$ indicates that the random variables X_{i_ℓ} and X_{j_ℓ} are swapped, while $s_\ell = +1$ indicates no swap. Informally, the two constraints (2) and (3) ensure that, under the null, each $\psi_\ell(X_{i_\ell}, X_{j_\ell})$ is likely to be no larger than its swapped version, $\psi_\ell(X_{j_\ell}, X_{i_\ell})$—and thus, the statistic $T = T(\mathbf{X})$ is likely to be no larger than its swapped copies, $T_\mathbf{s} = T(\mathbf{X}^\mathbf{s})$. If instead $T > T_\mathbf{s}$ for many swaps $\mathbf{s}$, this indicates evidence against the null. To formalize this intuition, we define p-value for `PairSwap-ICI` test as

$$p := \frac{1}{2^L} \sum_{\mathbf{s} \in \{\pm 1\}^L} \mathbb{1}\left\{T_\mathbf{s} \geq T\right\}. \tag{5}$$

In words, we are comparing the observed value T of the statistic, against all possible permuted statistic values $T_\mathbf{s}$ that we would obtain by swapping indices within matched pairs of our observed data $\mathbf{X}$.

Our method is quite flexible, in that the analyst may select any pairs (i_ℓ, j_ℓ) subject to monotonicity (2), and any functions ψ_ℓ satisfying (3), to construct a valid test. In particular, they can decide on these aspects of the test *after* exploring the data $\mathbf{Y}, \mathbf{Z}$, choosing to include a pair (i_ℓ, j_ℓ) if the data observed in $\mathbf{Y}$ indicates that $X_{i_\ell} > X_{j_\ell}$ would be likely under the alternative. However, the quality of the matches (i_ℓ, j_ℓ) and functions ψ_ℓ affects the power of our test; we discuss effective strategies for designing a statistic T in Section 3.

Example: a linear test statistic. Before proceeding, we give a simple example of a test statistic that we might choose to use: consider a linear test statistic,

$$T(\mathbf{x}) = \sum_{i=1}^{n} \beta_i x_i$$

[2]Again, as for marginal permutation tests in Section 1.2, here we suppress dependence on $\mathbf{Y}$ and $\mathbf{Z}$ in the notation $T(\mathbf{x})$, even though this statistic does depend on $\mathbf{Y}$ and $\mathbf{Z}$ through the choices of the matched pairs (i_ℓ, j_ℓ) and functions ψ_ℓ.

for some coefficients $\beta_i \in \mathbb{R}$. This function can be used as the test statistic for the `PairSwap-ICI` test, as long as the coefficients satisfy

$$\beta_{i_\ell} \geq \beta_{j_\ell} \text{ for each } \ell \in [L]. \tag{6}$$

To see why, first note that without loss of generality we can take $\beta_i = 0$ for all $i \in [n] \setminus \{i_1, j_1, \ldots, i_L, j_L\}$, i.e., all data points not belonging to any of the L pairs. This is because the indicator $\mathbb{1}\{T_{\mathbf{s}} \geq T\}$, appearing in the computation of the p-value, is invariant to these terms. Next, define

$$\psi_\ell(x, x') = \beta_{i_\ell} x + \beta_{j_\ell} x',$$

which satisfies (3) because for any $\Delta, \Delta' \geq 0$ with $x, x', x + \Delta, x' - \Delta' \in \mathcal{X}$, we have

$$\psi_\ell(x + \Delta, x' - \Delta') - \psi_\ell(x' - \Delta', x + \Delta) - \psi_\ell(x, x') + \psi_\ell(x', x) = (\beta_{i_\ell} - \beta_{j_\ell})(\Delta + \Delta') \geq 0$$

by our assumption that $\beta_{i_\ell} \geq \beta_{j_\ell}$. We then have $T(\mathbf{x})$ equal to the test statistic defined in (4).

We remark that choosing such a test statistic is by no means implying an assumption that the dependence between X and Y follows a linear model —it may be the case that the statistic $T(\mathbf{x}) = \beta^\top \mathbf{x}$ has good power for distinguishing the null from the alternative even if a linear model is only a coarse approximation to the true model.

2.1 Validity

Our first main result is that our method yields a valid test of H_0^{ICI}.

Theorem 1. *Under H_0^{ICI}, the conditional Type I error of the* `PairSwap-ICI` *test satisfies $\mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} \leq \alpha$ for all $\alpha \in [0, 1]$. In particular, the test enjoys marginal error control: $\mathbb{P}\{p \leq \alpha\} \leq \alpha$ for all α.*

Our proof of Theorem 1 formalizes our intuition at the start of this section, making use of the fact that, under the null, $\psi_\ell(X_{i_\ell}, X_{j_\ell})$ tends to be smaller than its swapped version $\psi_\ell(X_{j_\ell}, X_{i_\ell})$.

Proof of Theorem 1. Our proof is split into three steps. First we derive some deterministic properties of the p-value p. Next, we compare to the *sharp null*, where the pair X_{i_ℓ}, X_{j_ℓ} are identically distributed (rather than stochastically ordered). Finally, we examine the validity of the test under the sharp null.

Step 1: some deterministic properties of the p-value. First, fix $\alpha \in [0, 1]$ and define a function $p : \mathbb{R}^n \to [0, 1]$ as

$$p(\mathbf{x}) = \frac{1}{2^L} \sum_{\mathbf{s} \in \{\pm 1\}^L} \mathbb{1}\{T(\mathbf{x}^{\mathbf{s}}) \geq T(\mathbf{x})\}, \tag{7}$$

so that the p-value p in (5) can be written as $p = p(\mathbf{X})$. For each $\mathbf{s} \in \{\pm 1\}^L$, we can observe that the value $p(\mathbf{x}^{\mathbf{s}})$ is simply computing the quantile of $T(\mathbf{x}^{\mathbf{s}})$ among all possible swapped

statistics, $\big(T(\mathbf{x}^{\mathbf{s}'}) : \mathbf{s}' \in \{\pm 1\}^L\big)$. Consequently, it holds deterministically that

$$\frac{1}{2^L} \sum_{\mathbf{s} \in \{\pm 1\}^L} \mathbb{1}\left\{p(\mathbf{x}^{\mathbf{s}}) \leq \alpha\right\} \leq \alpha. \tag{8}$$

(Lemma D.24 in the Appendix verifies this bound, for completeness.)

In addition, we claim that $p(\cdot)$ is monotone in its coordinates, namely, $p(\mathbf{x})$ is nonincreasing in each x_{i_ℓ}, and nondecreasing in each x_{j_ℓ}. To see why, for each $\mathbf{s} \in \{\pm 1\}^L$, we can calculate

$$\begin{aligned}
\mathbb{1}\left\{T(\mathbf{x}^{\mathbf{s}}) \geq T(\mathbf{x})\right\} &= \mathbb{1}\left\{\sum_{\ell: s_\ell = +1} \psi_\ell(x_{i_\ell}, x_{j_\ell}) + \sum_{\ell: s_\ell = -1} \psi_\ell(x_{j_\ell}, x_{i_\ell}) \geq \sum_{\ell=1}^{L} \psi_\ell(x_{i_\ell}, x_{j_\ell})\right\} \\
&= \mathbb{1}\left\{\sum_{\ell: s_\ell = -1} \big(\psi_\ell(x_{i_\ell}, x_{j_\ell}) - \psi_\ell(x_{j_\ell}, x_{i_\ell})\big) \leq 0\right\}.
\end{aligned}$$

By the anti-monotonicity condition (3) on ψ_ℓ, this function is nonincreasing in each x_{i_ℓ}, and nondecreasing in each x_{j_ℓ}, and therefore the same is true for $p(\mathbf{x})$ as well.

Step 2: compare to the sharp null. For each $i \in [n]$, let $P_i = P_{X|Z}(\cdot \mid Z_i)$ denote the null distribution of X_i (after conditioning on $\mathbf{Y}, \mathbf{Z}$). By Assumption 1, we know that $P_{i_\ell} \preceq_{\text{st}} P_{j_\ell}$ for each pair $\ell \in [L]$, where $\preceq_{\text{st}}$ denotes the stochastic ordering on distributions. Next, we also define distributions $\bar{P}_\ell$ for each $\ell \in [L]$, given by the mixture

$$\bar{P}_\ell = \frac{1}{2} P_{i_\ell} + \frac{1}{2} P_{j_\ell}.$$

In particular, then,

$$P_{i_\ell} \preceq_{\text{st}} \bar{P}_\ell \preceq_{\text{st}} P_{j_\ell}, \quad \ell \in [L]. \tag{9}$$

We will now compare the observed data values, whose distribution (conditional on $\mathbf{Y}, \mathbf{Z}$) is given by

$$\mathbf{X} = (X_1, \ldots, X_n) \ \sim \ P_1 \times \cdots \times P_n,$$

against a different distribution,

$$\mathbf{X}_\sharp = \big((X_\sharp)_1, \ldots, (X_\sharp)_n\big) \ \sim \ (P_\sharp)_1 \times \cdots \times (P_\sharp)_n,$$

where the distributions $(P_\sharp)_i$ are defined by setting

$$(P_\sharp)_{i_\ell} = (P_\sharp)_{j_\ell} = \bar{P}_\ell$$

for each $\ell \in [L]$ (and, for any index $i \in [n] \setminus \{i_1, j_1, \ldots, i_L, j_L\}$ that does not belong to any of the L matched pairs, we simply take $(P_\sharp)_i = P_i$). We can think of this alternative vector of observations as being drawn from a *sharp null*, because for each pair ℓ, the random variables $(X_\sharp)_{i_\ell}, (X_\sharp)_{j_\ell}$ are identically distributed (rather than stochastically ordered, as for X_{i_ℓ}, X_{j_ℓ}). In particular, this implies that for any $\mathbf{s} \in \{\pm 1\}^L$,

$$(\mathbf{X}_\sharp)^{\mathbf{s}} \stackrel{\text{d}}{=} \mathbf{X}_\sharp \tag{10}$$

after conditioning on $\mathbf{Y}, \mathbf{Z}$.

In Step 1, we verified that the function $p(\mathbf{x})$ is nonincreasing in each x_{i_ℓ}, and nondecreasing in each x_{j_ℓ}. In particular, combined with the stochastic ordering (9), this means that $p(\mathbf{X}_\sharp) \preceq_{\text{st}} p(\mathbf{X})$ (conditional on $\mathbf{Y}, \mathbf{Z}$). We therefore have

$$\mathbb{P}\left\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\right\} = \mathbb{P}\left\{p(\mathbf{X}) \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\right\} \leq \mathbb{P}\left\{p(\mathbf{X}_\sharp) \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\right\}.$$

From this point on, then, we only need to verify the validity of the p-value $p(\mathbf{X}_\sharp)$ computed under the sharp null.

Step 3: validity under the sharp null. For data $\mathbf{X}_\sharp$ drawn under a sharp null, we have

$$\begin{aligned}\mathbb{P}\left\{p(\mathbf{X}_\sharp) \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\right\} &= \frac{1}{2^L} \sum_{\mathbf{s} \in \{\pm 1\}^L} \mathbb{P}\left\{p\big((\mathbf{X}_\sharp)^{\mathbf{s}}\big) \leq \alpha \;\middle|\; \mathbf{Y}, \mathbf{Z}\right\} \\ &= \mathbb{E}\left[\frac{1}{2^L} \sum_{\mathbf{s} \in \{\pm 1\}^L} \mathbb{1}\left\{p\big((\mathbf{X}_\sharp)^{\mathbf{s}}\big) \leq \alpha\right\} \;\middle|\; \mathbf{Y}, \mathbf{Z}\right] \leq \alpha,\end{aligned}$$

where the first step holds by (10), while the last step holds by the deterministic calculation (8) from Step 1. □

The p-value constructed in (5) requires computing $T_{\mathbf{s}} = T(\mathbf{X}^{\mathbf{s}})$ for all 2^L values of $\mathbf{s} \in \{-1, 1\}^L$, which may be computationally prohibitive for moderate or large L. In practice, it is common to use a Monte Carlo approximation to the p-value: we sample $\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)} \overset{\text{iid}}{\sim} \text{Unif}(\{\pm 1\}^L)$, and then compute

$$\hat{p}_M = \frac{1 + \sum_{m=1}^M \mathbb{1}\left\{T_{\mathbf{s}^{(m)}} \geq T\right\}}{1 + M}$$

The extra '1+' term appearing in the numerator and denominator is necessary to ensure error control for this Monte Carlo version of our test (Davison and Hinkley, 1997; Phipson and Smyth, 2010); in particular, this correction ensures we cannot have $\hat{p}_M = 0$. The following theorem verifies that this version of the test also controls the Type I error.

Theorem 2. *Fix any $M \in \mathbb{N}$. Under H_0^{ICI}, it holds that $\mathbb{P}\left\{\hat{p}_M \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\right\} \leq \alpha$ for all $\alpha \in [0, 1]$, and consequently, $\mathbb{P}\left\{\hat{p}_M \leq \alpha\right\} \leq \alpha$.*

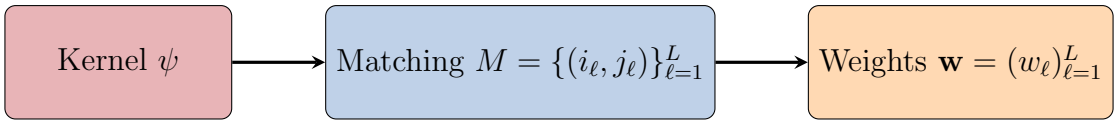
3 Designing a powerful `PairSwap-ICI` test

In this section, we construct a principled, powerful implementation of our test, by designing concrete choices for the pairs (i_ℓ, j_ℓ) and the functions ψ_ℓ introduced in Section 2. Throughout, we will restrict our attention to test statistics $T(\mathbf{x})$ of the form

$$T(\mathbf{x}) = \sum_{\ell=1}^L w_\ell \, \psi(x_{i_\ell}, x_{j_\ell}). \tag{11}$$

That is, in the original definition of the test statistic (4), we take $\psi_\ell(\cdot) = w_\ell \psi(\cdot)$ for some sequence of non-negative weights $\mathbf{w} = (w_\ell)_{\ell=1}^L$ and some *fixed* kernel ψ, which is required to satisfy the anti-monotonicity condition (3).

With this simplification, designing a test statistic now requires specifying the kernel ψ, deciding which pairs (i_ℓ, j_ℓ) are matched, and finally, how much weight w_ℓ to assign to each pair, as depicted in this flowchart:



As guaranteed by Theorem 1, our test controls the Type I error for any choice of ψ, M and $\mathbf{w}$, subject to the conditions (2) and (3) outlined at the start of Section 2. However, for the test to be effective, we need to tailor these choices to the specific application of interest. Of course, all of these choices interact with each other: what constitutes a good matching depends on how we choose the weights, and vice versa.

3.1 Specifying the kernel ψ

We begin by considering several simple options for the kernel ψ. As a first example, consider $\psi(x, x') = x - x'$. This choice of ψ means that $\psi(X_{i_\ell}, X_{j_\ell})$ is likely to be ≤ 0 under the null (since $Z_{i_\ell} \preceq Z_{j_\ell}$), but under the alternative, may be likely to be large (if the pair (i_ℓ, j_ℓ) is chosen wisely). Of course, we also allow for nonlinear test statistics to handle a broader range of settings. If X has heavy tails, then the distribution of a linear statistic T can be very sensitive to extreme values. We can ameliorate this sensitivity by using $\psi(x, x') = \text{sign}(x - x')$, or $\psi(x, x') = (-K) \vee (x - x') \wedge K$ (i.e., the truncation of $x - x'$ to some bounded interval $[-K, K]$) for some constant $K > 0$.

Example: a linear test statistic, revisited. To give more motivation for these simple choices, we will now see that a `PairSwap-ICI` test run with any *linear* test statistic $T_{\text{lin}}(\mathbf{x}) = \sum_{i=1}^n \beta_i x_i$, can always be expressed in the form (11) with the linear kernel $\psi(x, x') = x - x'$. To see why, define

$$w_\ell = \frac{\beta_{i_\ell} - \beta_{j_\ell}}{2},$$

(and note that we must have $w_\ell \geq 0$ due to (6)). Then we can write

$$T_{\text{lin}}(\mathbf{x}) = T(\mathbf{x}) + T_{\text{sym}}(\mathbf{x}),$$

where T is defined as in (11), and where the term

$$T_{\text{sym}}(\mathbf{x}) = \sum_{\ell=1}^L \frac{\beta_{i_\ell} + \beta_{j_\ell}}{2}(x_{i_\ell} + x_{j_\ell}) + \sum_{i \in [n] \setminus \{i_1, j_1, \ldots, i_L, j_L\}} \beta_i x_i$$

is symmetric in the pair (x_{i_ℓ}, x_{j_ℓ}) for each ℓ. Thus $T_{\text{sym}}(\mathbf{x}^{\mathbf{s}}) = T_{\text{sym}}(\mathbf{x})$ for any $\mathbf{x}$ and any $\mathbf{s} \in \{\pm 1\}^L$. It follows that

$$\mathbb{1}\{(T_{\text{lin}})_{\mathbf{s}} \geq T_{\text{lin}}\} = \mathbb{1}\{T_{\mathbf{s}} \geq T\}$$

for every $\mathbf{s}$—that is, the p-value p defined in (5) is *identical* if we use the test statistic T of the form (11) in place of the original linear test statistic T_{lin}.

3.2 Oracle strategies for choosing the matching and weights

We now build intuition for how to choose the matching M and weights $\mathbf{w}$ effectively by sketching the asymptotics of our test, assuming some oracle knowledge (or estimates) of the data distribution. Let us consider any statistic T of the form (11). Throughout this section, the kernel ψ is a fixed function, and we wish to choose the weights $\mathbf{w} = (w_\ell)_{\ell\in[L]}$ and matching $M = \{(i_\ell, j_\ell)\}_{\ell\in[L]}$ to maximize the power of our test. Given the data $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, the reference statistic $T_{\mathbf{s}}$ is a sum of L independent random variables. Under some regularity conditions on the weights $\mathbf{w}$ and the function ψ, a central limit theorem (CLT) approximation gives, for large L, that

$$p \approx \bar{\Phi}(\hat{T}) \qquad \text{where} \qquad \hat{T} = \hat{T}(\mathbf{w}, M) := \frac{\sum_{\ell=1}^{L} w_\ell \psi(X_{i_\ell}, X_{j_\ell})}{\sqrt{\sum_{\ell=1}^{L} w_\ell^2 \psi(X_{i_\ell}, X_{j_\ell})^2}},$$

and where $\bar{\Phi}$ denotes the standard Gaussian survival function, i.e., $\bar{\Phi}(t) = 1 - \Phi(t)$, where Φ is the standard Gaussian distribution function. The above approximation holds for fixed $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ and relies only on the CLT approximation for a weighted sum of L independent signs $s_1, \ldots, s_L \in \{\pm 1\}$.

The above calculation tells us that we should aim to choose weights that *maximize* the approximate probability of rejection, $\mathbb{P}\left\{\bar{\Phi}(\hat{T}) \le \alpha \;\middle|\; \mathbf{Y}, \mathbf{Z}\right\}$, in order to achieve the best possible power. Under some conditions this conditional power can be further approximated as

$$\Phi\left(\frac{\sum_{\ell=1}^{L} w_\ell \mathbb{E}\left[\psi(X_{i_\ell}, X_{j_\ell}) \mid \mathbf{Y}, \mathbf{Z}\right]}{\sqrt{\sum_{\ell=1}^{L} w_\ell^2 \mathrm{Var}\left(\psi(X_{i_\ell}, X_{j_\ell}) \mid \mathbf{Y}, \mathbf{Z}\right)}} - \bar{\Phi}^{-1}(\alpha)\right) \tag{12}$$

(see Theorem C.12 in the Appendix for a closer look at this approximation). The following lemma shows how to maximize this approximation over the weights, treating the matching M as fixed.

Lemma 3. *Assume that* $\mathrm{Var}\left(\psi_\ell(X_{i_\ell}, X_{j_\ell}) \mid \mathbf{Y}, \mathbf{Z}\right) > 0$ *for* $\ell \in [L]$. *Considered as a function of* $\mathbf{w} = (w_1, \ldots, w_L) \in [0, \infty)^L$, *the function in* (12) *is maximised by the choice*

$$w_\ell^* = \frac{\max\left\{\mathbb{E}\left[\psi(X_{i_\ell}, X_{j_\ell}) \mid \mathbf{Y}, \mathbf{Z}\right], 0\right\}}{\mathrm{Var}\left(\psi(X_{i_\ell}, X_{j_\ell}) \mid \mathbf{Y}, \mathbf{Z}\right)}, \tag{13}$$

for $\ell \in [L]$.

Using a plug-in estimate for the moments. In Lemma 3, the oracle weight vector $\mathbf{w}^*$ depends on the conditional expected value and variance, $\mathbb{E}\left[\psi(X_{i_\ell}, X_{j_\ell}) \mid \mathbf{Y}, \mathbf{Z}\right]$ and $\mathrm{Var}\left(\psi(X_{i_\ell}, X_{j_\ell}) \mid \mathbf{Y}, \mathbf{Z}\right)$. In practice, we will assume that we have access to estimates $\hat{E}_{ij}$ of $\mathbb{E}\left[\psi(X_{i_\ell}, X_{j_\ell}) \mid \mathbf{Y}, \mathbf{Z}\right]$ and $\hat{V}_{ij} > 0$ of $\mathrm{Var}\left(\psi(X_i, X_j) \mid \mathbf{Y}, \mathbf{Z}\right)$, for each $i, j \in [n]$, constructed from data that are independent of $\mathbf{X}$. For example, these might be obtained from a fitted

model for the conditional distribution of X given Y, Z constructed using a separate data set (e.g. via sample splitting). For the linear kernel $\psi(x, x') = x - x'$, this would amount to estimating the first two moments of $X_i \mid Y_i, Z_i$ for $i \in [n]$, whereas if $\psi(x, x') = \text{sign}(x - x')$, then we would need to estimate $\mathbb{P}\{X_i > X_j \mid Y_i, Z_i, Y_j, Z_j\}$ for distinct indices i, j. With these estimates in place, we can seek to maximize the power of our test by choosing weights

$$\hat{w}_\ell = \frac{\max\{\hat{E}_{i_\ell j_\ell}, 0\}}{\hat{V}_{i_\ell j_\ell}}$$

for $\ell \in [L]$.

Choosing the matching. In the oracle setting, once we fix the choice of weights as in (13), the estimator (12) of the test's conditional power is maximized by solving a maximum-weight matching problem, namely that of finding

$$M^* \in \operatorname*{argmax}_{M \in \mathcal{M}_n(\mathbf{Z})} \sum_{(i,j) \in M} \left(W^*_{ij}\right)^2 \quad \text{where} \quad W^*_{ij} = \frac{\max\left\{\mathbb{E}\left[\psi(X_i, X_j) \mid \mathbf{Y}, \mathbf{Z}\right], 0\right\}}{\text{Var}\left(\psi(X_i, X_j) \mid \mathbf{Y}, \mathbf{Z}\right)}. \tag{14}$$

We then run `PairSwap-ICI` with this oracle matching M^*, and with weights $w_\ell = W^*_{i_\ell j_\ell}$ for each pair $(i_\ell, j_\ell) \in M^*$.

Similarly, using the plug-in estimates, the conditional power is approximately maximized by finding

$$\hat{M} \in \operatorname*{argmax}_{M \in \mathcal{M}_n(\mathbf{Z})} \sum_{(i,j) \in M} \hat{W}^2_{ij} \quad \text{where } \hat{W}_{ij} = \frac{\max\{\hat{E}_{ij}, 0\}}{\hat{V}_{ij}}. \tag{15}$$

To run `PairSwap-ICI`, we then take weights $w_\ell = \hat{W}_{i_\ell j_\ell}$ for each pair $(i_\ell, j_\ell) \in \hat{M}$.

The plug-in matching $\hat{M}$ can be computed in polynomial time (Edmonds, 1965; Duan and Pettie, 2014). Specifically, if $m = |\{(i, j) : Z_i \preceq Z_j\}|$, then $\hat{M}$ can be computed in time $O(mn + n^2 \log n)$ using an algorithm of Gabow (1985).[3]

3.3 Heuristic strategies for choosing the matching and weights

Above, in Section 3.2, the test statistic $T(\mathbf{x})$ was designed with the aim of maximizing the power of our test, but with the assumption that we have access to substantial knowledge about the distribution of the data—for instance, we have been able to fit a model for the distribution of (X, Y, Z) using a separate data set. Of course, an accurate estimate of the true model can enable a very powerful test (since we have a good approximation of the alternative that we are testing against), but in some settings the approach of Section 3.2 may not be practical, either because we are not able to reliably approximate the distribution of the data, or because the required computations are too costly (in particular, the oracle matching strategy).

[3]For our experiments in Sections 5 and 6, we use the Python package `networkx` (Hagberg, Swart and Schult, 2008), which uses the Blossom algorithm (Edmonds, 1965) and runs in time $O(n^3)$.

In this section, therefore, we take a completely different approach: we will propose an extremely simple scheme for designing weights $\mathbf{w}$, and two easy strategies for choosing a matching M, that do not require extensive prior knowledge or costly calculations. Of course, this will come at some cost in terms of the resulting power of the test, since we are no longer mimicking an oracle test—but, as we will see in both our theoretical guarantees and our empirical results below, these simple strategies can often attain high power nonetheless.

Throughout this section, we will restrict our attention to the one-dimensional setting, $\mathcal{Z} \subseteq \mathbb{R}$, and will also choose the kernel $\psi(x, x') = x - x'$ when defining our test statistic as in (11). We will work in the setting where we hypothesize that, under the alternative, there is a *positive* association between X and Y even after controlling for Z (of course, if our hypothesis is a negative association, we can follow an analogous strategy). The idea is simple: we will take pairs (i_ℓ, j_ℓ) such that

- $Z_{i_\ell} \leq Z_{j_\ell}$ (as required for validity), but $Z_{i_\ell} \approx Z_{j_\ell}$; and
- $Y_{i_\ell} > Y_{j_\ell}$ (so that, under the alternative, we expect $X_{i_\ell} > X_{j_\ell}$).

3.3.1 A simple weighting scheme

We begin by defining a mechanism for choosing the weights: we will take

$$w_\ell = \left(Y_{i_\ell} - Y_{j_\ell}\right)_+,$$

the difference in Y values (if this difference is positive for the pair (i_ℓ, j_ℓ)).

Why is this simple strategy a reasonable choice across broad settings? Our test statistic $T(\mathbf{X})$ is given by

$$T(\mathbf{X}) = \sum_{\ell : Y_{i_\ell} > Y_{j_\ell}} (Y_{i_\ell} - Y_{j_\ell}) \cdot (X_{i_\ell} - X_{j_\ell}).$$

Since under the alternative, we expect $X_{i_\ell} > X_{j_\ell}$, this means that the expected value of $T(\mathbf{X})$ is large and positive under the alternative (but, of course, is non-positive under the null).

To consider another motivation, let us examine a specific model. Suppose that Y has a linear effect on the mean of X, so that

$$\mathbb{E}\left[X \mid Y = y, Z = z\right] = \beta^* y + \mu_Z^*(z),$$

and the conditional variance is constant,

$$\mathrm{Var}\left(X \mid Y = y, Z = z\right) = \sigma^{*2}.$$

Then, following the oracle strategy of Section 3.2,as in (13) we have oracle weights

$$w_\ell^* = \frac{\max\left\{\mathbb{E}\left[X_{i_\ell} - X_{j_\ell} \mid \mathbf{Y}, \mathbf{Z}\right], 0\right\}}{\mathrm{Var}\left(X_{i_\ell} - X_{j_\ell} \mid \mathbf{Y}, \mathbf{Z}\right)} = \frac{\max\left\{\beta^*(Y_{i_\ell} - Y_{j_\ell}) + (\mu_Z^*(Z_{i_\ell}) - \mu_Z^*(Z_{j_\ell})), 0\right\}}{2\sigma^{*2}}$$
$$\approx \frac{\beta^*}{2\sigma^{*2}} \max\left\{(Y_{i_\ell} - Y_{j_\ell}), 0\right\},$$

where the last step holds since we have assumed $Z_{i_\ell} \approx Z_{j_\ell}$ in our choice of the pair (and so $\mu_Z^*(Z_{i_\ell}) \approx \mu_Z^*(Z_{j_\ell})$ likely holds). But crucially, our test is invariant to rescaling the weights—that is, choosing weights $w_\ell = \max\left\{Y_{i_\ell} - Y_{j_\ell}, 0\right\}$ is equivalent to choosing weights $\frac{\beta^*}{2\sigma^{*2}} \max\left\{(Y_{i_\ell} - Y_{j_\ell}), 0\right\}$, and thus is nearly equivalent to the oracle weights.

3.3.2 Two simple matching schemes

Next, we propose two matching strategies that, again, do not require any knowledge or estimate of the model.

Neighbour matching. Our first simple matching strategy is to choose pairs that are nearest neighbours in the list of sorted Z values.

Algorithm 1 neighbour matching

Preliminaries: sort Z values, i.e., find a permutation π of $\{1, \dots, n\}$ such that

$$Z_{\pi(1)} \leq Z_{\pi(2)} \leq \cdots \leq Z_{\pi(n-1)} \leq Z_{\pi(n)}.$$

for $m = 1, \dots, \lfloor n/2 \rfloor$ **do**

If $Y_{\pi(2m-1)} > Y_{\pi(2m)}$, then add a new pair to the matching,

$$(i_\ell, j_\ell) = (\pi(2m-1), \pi(2m)).$$

end for

This strategy ensures that, for all pairs $(i_\ell, j_\ell) = (\pi(2m-1), \pi(2m))$ that are included in the matching, we have $Z_{i_\ell} \leq Z_{j_\ell}$ and $Y_{i_\ell} > Y_{j_\ell}$ by construction, and moreover, it likely holds that $Z_{i_\ell} \approx Z_{j_\ell}$ (since we have chosen two consecutive Z values in the sorted list).

However, an obvious limitation of this naïve matching strategy is that many consecutive pairs $(\pi(2m-1), \pi(2m))$ in the sorted list can fail to have $Y_{\pi(2m-1)} > Y_{\pi(2m)}$ just by chance. In particular, if the Z values in this pair are approximately equal (as we might expect), then the values $Y_{\pi(2m-1)}, Y_{\pi(2m)}$ are expected to be approximately i.i.d.—which means that the event $Y_{\pi(2m-1)} > Y_{\pi(2m)}$ will fail half the time. In other words, we are discarding approximately half of the data—we will expect to have $L \approx n/4$ pairs in this matching (meaning that $2L \approx n/2$ many data points have been assigned to a matched pair).

Cross-bin matching. Our next strategy is cross-bin matching, which aims to avoid the inefficiency of neighbour matching—we aim to use (nearly) all of the data, rather than discarding half the data as is likely the case for neighbour matching. To implement this strategy, we will partition the list of sorted Z values into K bins, and will allow a pair of data points to be matched as long as the Z values are in adjacent bins (rather than requiring consecutive Z values, as for neighbour matching).

Algorithm 2 Cross-bin matching

Preliminaries: sort Z values, i.e., find a permutation π of $\{1, \ldots, n\}$ such that

$$Z_{\pi(1)} \leq Z_{\pi(2)} \leq \cdots \leq Z_{\pi(n-1)} \leq Z_{\pi(n)},$$

and define K bins of indices,

$$\begin{aligned} A_1 &= \{\pi(1), \ldots, \pi(m\}, \\ A_2 &= \{\pi(m+1), \ldots, \pi(2m)\}, \\ &\ldots \\ A_K &= \{\pi((K-1)m+1), \ldots, \pi(Km)\}, \end{aligned}$$

where $m = \lfloor n/K \rfloor$.
for $k = 1, \ldots, K$ **do**
Define $r_{k,1}, \ldots, r_{k,m}$ as a permutation of A_k such that

$$Y_{r_{k,1}} \geq \cdots \geq Y_{r_{k,m}}.$$

end for
for $k = 1, \ldots, K-1$ **do**
for $s = 1, \ldots, \lfloor m/2 \rfloor$ **do**
If $Y_{r_{k,s}} > Y_{r_{k+1,m+1-s}}$, then add a new pair to the matching,

$$(i_\ell, j_\ell) = (r_{k,s}, r_{k+1,m+1-s}).$$

end for
end for

To explain this procedure in words:

- First we group the Z values into K many bins, with each bin A_k containing m many consecutive Z values.
- Then, we attempt to match the largest values of Y in bin k with the smallest values of Y in bin $k+1$—that is, in the inner "for loop", at step $s = 1$ we are attempting to match the largest Y value in bin k (i.e., $Y_{r_{k,1}}$) with the smallest Y value in bin $k+1$ (i.e., $Y_{r_{k+1,m}}$), and then at step $s = 2$ we proceed to matching the second-largest and second-smallest, and so on.

This scheme is illustrated in Figure 1.

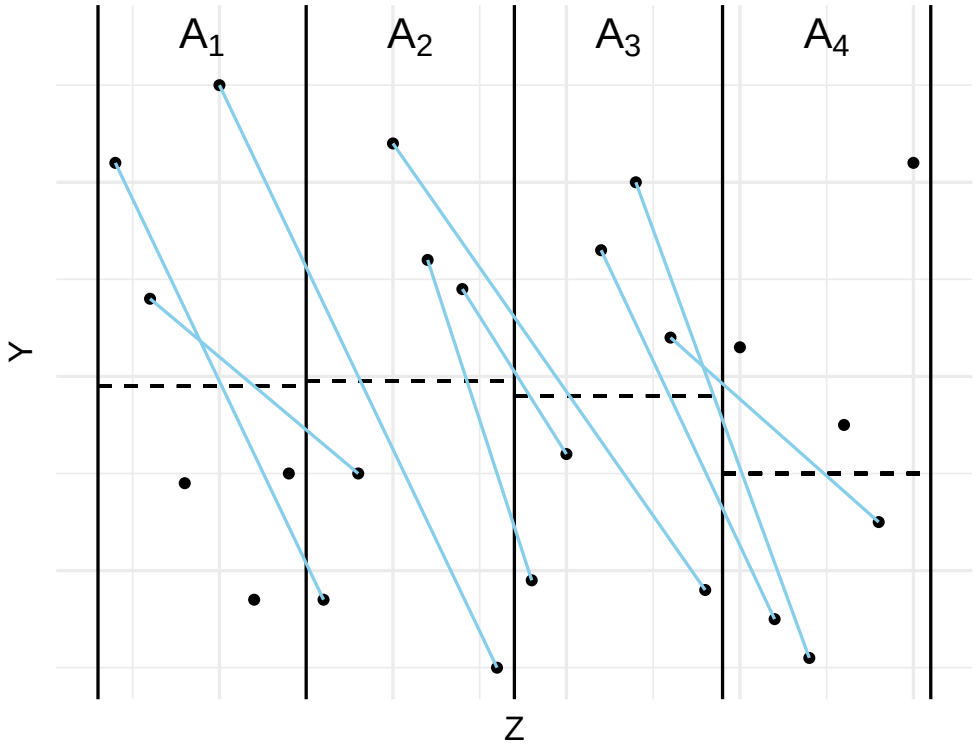


Figure 1: Demonstration of the cross-bin matching scheme described in Algorithm 2.

Why do we expect that this strategy will be more powerful than the simpler neighbour matching strategy? At a high level, while neighbour matching is expected to discard around half of the data, the cross-bin matching strategy can potentially assign nearly all data points to a matched pair. However, there is a potential tradeoff: while the pairs produced by both strategies will likely satisfy $Z_{i_\ell} \approx Z_{j_\ell}$, this approximation will be closer to equality for neighbour matching (where pairs consist of consecutive Z values) than cross-bin matching (where pairs consist of Z values in neighbouring bins, i.e., they may be up to $2m$ positions apart in the sorted list). If m is not too large (i.e., the number of bins K is not too small), though, this difference is hopefully negligible. We will examine both methods theoretically in the following section and will see these tradeoffs in more detail.

4 Power analysis

In this section, we study the power of `PairSwap-ICI` test under the following general model. We assume that the data $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = (X_i, Y_i, Z_i)_{i\in[n]} \overset{\text{iid}}{\sim} P$ is drawn according to

$$\mathbf{X} = \mu(\mathbf{Y}, \mathbf{Z}) + \boldsymbol{\zeta}, \tag{16}$$

where $\mu : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ (applied componentwise) is a measurable function, and with $(Y_i, Z_i)_{i\in[n]} \overset{\text{iid}}{\sim} P_{Y,Z}$ drawn independently from $(\zeta_i)_{i\in[n]} \overset{\text{iid}}{\sim} P_\zeta$. We suppose that P_ζ has mean 0 and unknown variance $\sigma^2 > 0$. Throughout this section, we also assume that the statistic T admits the form in (11) with $\psi(x, x') = x - x'$, and we restrict our attention to the setting $\mathcal{Y} = \mathcal{Z} = \mathbb{R}$; the partial ordering $\preceq$ for Z will simply be the usual ordering $\leq$ on $\mathbb{R}$.

Defining isotonic signal strength (ISS). Before stating the results on power under this signal plus noise model, we first introduce some notation. We denote by $\mathcal{C}_{\mathrm{ISO}}$ the set of non-decreasing functions on $\mathbb{R}$ and then, we define

$$\mathrm{ISS}_n = \inf_{g \in \mathcal{C}_{\mathrm{ISO}}} \mathbb{E}_{P^n_{Y,Z}} \left[\|\mu(\mathbf{Y}, \mathbf{Z}) - g(\mathbf{Z})\|_2 \right], \quad \widehat{\mathrm{ISS}}_n = \inf_{g \in \mathcal{C}_{\mathrm{ISO}}} \|\mu(\mathbf{Y}, \mathbf{Z}) - g(\mathbf{Z})\|_2, \tag{17}$$

referring to them as the oracle and empirical *isotonic signal strength* respectively (here $g(\mathbf{Z})$ is applied componentwise). In particular, under the null, we would have $\mathrm{ISS}_n = \widehat{\mathrm{ISS}}_n = 0$ (since we can simply take g to be the true mean function, which does not depend on Y), while under the alternative these quantities might be large.

Preview of results. To analyze the performance of `PairSwap-ICI` testing procedure, we study the power conditional on $\mathbf{Y}$ and $\mathbf{Z}$, i.e. $\mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\}$, and we demonstrate how this power is characterized by the isotonic signal strength quantities ISS_n and $\widehat{\mathrm{ISS}}_n$. The organization of this section is as follows.

- In Section 4.1 we study the asymptotic upper and lower bound on the power of oracle matching (as defined in Section 3.2), conditional on $\mathbf{Y}$ and $\mathbf{Z}$, and establish that the empirical quantity $\widehat{\mathrm{ISS}}_n$ governs their behaviour. The same guarantees also hold for the plug-in matching strategy defined in Section 3.2, as long as the estimate $\hat{\mu}$ is consistent. Qualitatively, this implies that we can have non-trivial power guarantees against the alternatives with large isotonic signal strength.
- Next, in Section 4.2, we show that the converse also holds, i.e., one can not distinguish the null class from the alternatives with small isotonic signal strength. More precisely, if the oracle quantity ISS_n is too small, then no valid testing procedure for H_0^{ICI} can have non-trivial power.
- Finally, in Section 4.3, we specialize our power guarantees to the special case of partially linear Gaussian models. We show that even without the knowledge of μ or an approximation for the same, we can achieve near-optimal power guarantees with some of the more practical matching algorithms from Section 3.3.

4.1 ISS dictates the power of oracle matching

Below, we study the best-case performance of `PairSwap-ICI` procedure, specifically the asymptotic behaviour of the conditional power for oracle matching. In particular Theorem 4 provides asymptotic upper and lower bounds on the conditional power $\mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\}$ of oracle matching under the following framework: we assume that the distributions $P_{Y,Z}$ and P_ζ in (16) are independent of the sample size n, while the regression function $\mu(\cdot, \cdot)$ does depend on n; however, we suppress the dependence of n in our notation for simplicity.

Theorem 4. *Suppose that* $\widehat{\mathrm{ISS}}_n > 0$ *and that* $\|\mu(\mathbf{Y}, \mathbf{Z})\|_\infty \vee \|\mu(\mathbf{Y}, \mathbf{Z})\|_\infty^4 = \mathrm{o}_P(\widehat{\mathrm{ISS}}_n)$. *Then, under the model* (16)*, the conditional power of oracle matching with* $\alpha \in (0, 1/2)$ *satisfies*

$$\Phi\left(\frac{\widehat{\mathrm{ISS}}_n}{\sqrt{2}\sigma} - \bar{\Phi}^{-1}(\alpha)\right) - \mathrm{o}_P(1) \leq \mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} \leq \Phi\left(\frac{\widehat{\mathrm{ISS}}_n}{\sigma} - \bar{\Phi}^{-1}(\alpha)\right) + \mathrm{o}_P(1). \tag{18}$$

We observe that $\widehat{\mathrm{ISS}}_n/\sigma$, which can be interpreted here as the *signal-to-noise ratio*, dictates both the upper and lower bound. We also notice that the upper and lower bounds on conditional power match up to a factor of 2—in fact, this factor is unavoidable without imposing further model assumptions (e.g., symmetry of $\mu(Y,Z)$ conditional on Z, as we will discuss in more detail in Appendix C.4.3).

Remark 1. *While the asymptotic lower and upper bounds require that* $\big(\|\mu(\mathbf{Y},\mathbf{Z})\|_\infty \vee \|\mu(\mathbf{Y},\mathbf{Z})\|_\infty^4\big)/\widehat{\mathrm{ISS}}_n = \mathrm{o}_P(1)$*, we remark that this is a natural assumption on the quality of matching. For example, if* X *is bounded, then* $\|\mu(\mathbf{Y},\mathbf{Z})\|_\infty \vee \|\mu(\mathbf{Y},\mathbf{Z})\|_\infty^4 = \mathrm{O}(1)$*—and so for this assumption to hold, it is sufficient to require* $\widehat{\mathrm{ISS}}_n \to \infty$*, i.e., the amount of signal in the data increases with* n*.*

Power guarantees for plug-in matching with an estimate of μ. Now, we shift our attention to the more practical setting, where we do not have the oracle knowledge of μ. A natural solution is data splitting—i.e., we learn an estimate $\hat{\mu}$ on one random split of the data, and then implement `PairSwap-ICI` test with the plug-in matching $\hat{M}$. While the aforementioned data splitting approach is more accurate and practical, for the sake of simplicity, in order to state the following result, we assume that we have an independent data where we can learn $\hat{\mu}$. Finally, under suitable consistency assumptions on $\hat{\mu}$, we can recover the power guarantees in Theorem 4.

Theorem 5. *Consider the setting and assumptions of Theorem 4. Suppose we use plug-in matching* $\hat{M}$ *with an estimate* $\hat{\mu}$ *constructed based on independent data, where* $\hat{\mu}$ *satisfies*

$$\|\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \mu(\mathbf{Y},\mathbf{Z})\|_2 = \mathrm{o}_P(\widehat{\mathrm{ISS}}_n), \quad \|\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \mu(\mathbf{Y},\mathbf{Z})\|_\infty^4 = \mathrm{o}_P(\widehat{\mathrm{ISS}}_n).$$

Then the conditional power of the `PairSwap-ICI` *satisfies (18) for* $\alpha \in (0, 1/2)$*.*

4.2 ISS characterizes hardness of testing the null H_0^{ICI}

While $\widehat{\mathrm{ISS}}_n$ governs the asymptotic upper and lower bounds on power in 4 and 5, the relationship between ISS_n and power extends beyond the `PairSwap-ICI` test. Here, we establish this connection formally. In particular, we will see that the quantity ISS_n determines which alternative models are distinguishable from the class of null models, via *any* valid testing procedure. Towards this goal, we start with a simple total-variation calculation to give an upper bound on the power function of any valid test.

Proposition 6. *Fix* $\alpha \in (0,1)$*. Fix any test* ϕ *that controls false positives at level* α*, i.e.,*[4]

$$\phi : (\mathcal{X} \times \mathbb{R} \times \mathbb{R})^n \to [0,1] \text{ such that } \sup_{P \in H_0^{\mathrm{ICI}}} \mathbb{E}_P\left[\phi(\mathbf{X},\mathbf{Y},\mathbf{Z})\right] \le \alpha.$$

Then for any distribution $P_{X,Y,Z}$*,*

$$\mathbb{E}_{P_{X,Y,Z}}\left[\phi(\mathbf{X},\mathbf{Y},\mathbf{Z})\right] \le \alpha + \inf_{Q_{X,Y,Z} \in H_0^{\mathrm{ICI}}} \mathrm{d_{TV}}\left(P_{X,Y,Z}^n, Q_{X,Y,Z}^n\right).$$

[4]Traditionally, we think of a hypothesis test ϕ as a map from data to a decision, i.e., $\phi(\mathbf{X},\mathbf{Y},\mathbf{Z}) \in \{0,1\}$. Why, then, do we define tests ϕ as mapping to the space $[0,1]$? This is because, in some settings, we may want to consider randomized tests—for instance, in the notation above where ϕ maps to $[0,1]$, an outcome $\phi(\mathbf{X},\mathbf{Y},\mathbf{Z}) = 0.75$ represents that, given the data, our randomized test rejects the null with probability 0.75. Of course, a nonrandomized test is simply a special case, obtained by restricting the output of ϕ to lie in $\{0,1\}$.

This last total variation term is the distance of $P_{X,Y,Z}$ from the null class H_0^{ICI}, meaning the closer P is to null models, the harder it will be to get non-trivial power against P. While in general it is hard to derive exact expressions for the total variation term, under some additional model assumptions on P, we can come up with interpretable upper bounds for the same. Two such examples are listed below.

- **Gaussian setting:** In the first example, we consider a special class of Gaussian alternatives, which is given by (16) with $P_\zeta = \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$. In this special case, ISS_n gives a meaningful upper bound on the offset total-variation term, up to a constant.

  Corollary 7. *Suppose, $P_\zeta = \mathcal{N}(0, \sigma^2)$ and $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ satisfy (16). Then*

  $$\inf_{Q_{X,Y,Z} \in H_0^{\mathrm{ICI}}} \mathrm{d_{TV}}\left(P^n_{X,Y,Z}, Q^n_{X,Y,Z}\right) \leq \frac{\mathrm{ISS}_n}{2\sigma},$$

  and consequently, for any test ϕ that controls Type I error at level α,

  $$\mathbb{E}_P\left[\phi(\mathbf{X}, \mathbf{Y}, \mathbf{Z})\right] \leq \alpha + \frac{\mathrm{ISS}_n}{2\sigma}.$$

- **Binary setting:** Now, suppose $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are generated from the model $P_{X,Y,Z}$ given by

  $$X_i \sim \mathrm{Ber}(\mu(Y_i, Z_i)), \quad (Y_1, Z_1), \cdots, (Y_n, Z_n) \overset{\mathrm{iid}}{\sim} P_{Y,Z}. \tag{19}$$

  Here as well, ISS_n leads to a very interpretable upper bound on the power of our test, as long as $\mu(Y, Z)$ is almost surely away from the extremities, i.e., 0 or 1.

  Corollary 8. *Suppose $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ satisfy (19) where $\mu(Y, Z) \in (\epsilon, 1 - \epsilon)$ almost surely for some $\epsilon > 0$. Then*

  $$\inf_{Q_{X,Y,Z} \in H_0^{\mathrm{ICI}}} \mathrm{d_{TV}}\left(P^n_{X,Y,Z}, Q^n_{X,Y,Z}\right) \leq \left(\frac{1}{\epsilon(1-\epsilon)}\right)^{1/2} \mathrm{ISS}_n,$$

  and consequently, for any test ϕ that controls Type I error at level α,

  $$\mathbb{E}_P\left[\phi(\mathbf{X}, \mathbf{Y}, \mathbf{Z})\right] \leq \alpha + \left(\frac{1}{\epsilon(1-\epsilon)}\right)^{1/2} \mathrm{ISS}_n.$$

To summarize, an appropriate *signal-to-noise ratio* determines the extent to which a test can outperform the trivial test that rejects the null H_0^{ICI} with probability α without using data. Consequently, in both examples, no valid test can achieve non-trivial power when ISS_n is negligible relative to the noise in $P_{X|Y,Z}$.

Note that this characterization relies on the oracle quantity ISS_n, while the conditional power of the `PairSwap-ICI` test is determined by the empirical version, $\widehat{\mathrm{ISS}}_n$. We would expect that, under mild conditions, $\mathrm{ISS}_n \approx \widehat{\mathrm{ISS}}_n$ (i.e., a concentration property should hold, as long as n is large)—we can therefore interpret ISS_n as characterizing the power, both in terms of upper and lower bounds, at least for certain cases.

4.3 Near-optimal power guarantees without knowledge of μ

The asymptotic upper and lower bounds on conditional power from Theorems 4 and 5 require either oracle knowledge of μ or an independent estimate of μ whose Euclidean norm error is small by comparison with $\widehat{\mathrm{ISS}}_n$. If this side information is not available, then a natural question arises: can near-optimal power guarantees be achieved with some of the simpler and more practical matching schemes from Section 3.3?

The aim of this subsection is to show that this is indeed possible, under some additional assumptions on the model class (16). Specifically, consider the class of partially linear Gaussian models given by

$$(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \text{ satisfy (16) with } P_\zeta = \mathcal{N}(0, \sigma^2),\ \mu(\mathbf{Y}, \mathbf{Z}) = \mu_0(\mathbf{Z}) + \beta_n \mathbf{Y} \tag{20}$$

for some $\mu_0 \in \mathcal{C}_{\mathrm{ISO}}$ (applied componentwise) and some $\beta_n \in [0, \infty)$. We also assume that Y is a bounded and mean-zero random variable, with $|Y| \leq 1$. Note that μ_0 is a fixed function, and so the dependence on n of the distribution of (X, Y, Z) is solely through β_n. The following lemma relates ISS_n to a more interpretable quantity, namely the expected conditional variance of Y given Z.

Lemma 9. *Under the model class (20),*

$$\sqrt{n}\beta_n \cdot \left(\mathbb{E}\left[\mathrm{Var}\left(Y \mid Z\right)\right]\right)^{1/2} \left(1 + \mathrm{o}_P(1)\right) \leq \mathrm{ISS}_n \leq \sqrt{n}\beta_n.$$

We can draw several conclusions from this result. By Lemma 9 and Corollary 7, no valid test can achieve asymptotically non-trivial power against an alternative in the model class (20) when $\beta_n = \mathrm{o}(1/\sqrt{n})$. From this point on, then, we will consider the regime $\beta_n \gtrsim 1/\sqrt{n}$. Another consequence of Lemma 9 is that the asymptotic lower bound on the conditional power of oracle matching in (18) further simplifies to give

$$\mathbb{P}\left\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\right\} \geq \Phi\left(\sqrt{n}\beta_n \left\{\frac{\mathbb{E}\left[\mathrm{Var}\left(Y \mid Z\right)\right]}{2\sigma^2}\right\}^{1/2} - \bar{\Phi}^{-1}(\alpha)\right) - \mathrm{o}_P(1). \tag{21}$$

Next, we aim to show that practical matching schemes, such as neighbour matching and cross-bin matching, can achieve conditional power at least as good as the lower bound above.

Theorem 10. *Let $\beta_n \gtrsim 1/\sqrt{n}$, and assume that $\mu_0(Z)$ is a sub-Gaussian random variable. Then the conditional power of neighbour matching (Algorithm 1), implemented with kernel $\psi(x, x') = x - x'$ and weights $w_\ell = \max\{Y_{i_\ell} - Y_{j_\ell}, 0\}$, satisfies*

$$\left|\mathbb{P}\left\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\right\} - \Phi\left(\sqrt{n}\beta_n \left\{\frac{\mathbb{E}\left[\mathrm{Var}\left(Y \mid Z\right)\right]}{4\sigma^2}\right\}^{1/2} - \bar{\Phi}^{-1}(\alpha)\right)\right| = \mathrm{o}_P(1).$$

Notably, the conditional power of neighbour matching matches the lower bound from (21) up to a factor of 2. This factor arises due to the inherent inefficiency of neighbour matching: as discussed in Section 3.3, we have seen that neighbour matching discards roughly half the data, which explains the extra factor of 2.

To avoid this loss of a factor of 2, we now consider the cross-bin matching strategy, which will typically assign nearly all data points to a matched pair. Consequently, the power of cross-bin matching can then meet the lower bound in (21).

Theorem 11. *Consider the setting as in Theorem 10. Then, the conditional power of cross-bin matching (Algorithm 2) with $K \propto \sqrt{n}$ many bins, implemented with kernel $\psi(x, x') = x - x'$, weights $w_\ell = \max\{Y_{i_\ell} - Y_{j_\ell}, 0\}$, satisfies*

$$\mathbb{P}\left\{p \le \alpha \mid \mathbf{Y}, \mathbf{Z}\right\} \ge \Phi\left(\sqrt{n}\beta_n\left\{\frac{\mathbb{E}\left[\mathrm{Var}\left(Y \mid Z\right)\right]}{2\sigma^2}\right\}^{1/2} - \bar{\Phi}^{-1}(\alpha)\right) - \mathrm{o}_P(1)$$

under suitable smoothness assumptions (stated formally in Theorem C.14).

In particular, this result matches the lower bound in (21), without an additional factor of 2 as for neighbour matching. (While the above result provides only an asymptotic lower bound on the conditional power of cross-bin matching, under additional assumptions we can characterize the power more exactly—see Appendix C.4.)

5 Simulations

In this section, we evaluate the performance of our method on simulated data, and compare the matching strategies from Section 3.3. For simplicity, we focus on the univariate case $\mathcal{Z} = \mathbb{R}$. We will test two versions of the `PairSwap-ICI` method:

- Neighbour matching (Algorithm 1), with the linear kernel $\psi(x, x') = x - x'$ and weights $w_\ell = \max\{Y_{i_\ell} - Y_{j_\ell}, 0\}$ as discussed in Section 3.3;
- Cross-bin matching (Algorithm 2) with $K = 2\left\lfloor n^{1/2} \right\rfloor$ bins, again with the linear kernel $\psi(x, x') = x - x'$ and weights $w_\ell = \max\{Y_{i_\ell} - Y_{j_\ell}, 0\}$. This particular choice of K is motivated by Theorem 11, which suggests choosing $K \propto \sqrt{n}$ bins in order to achieve competitive power with neighbour matching.

5.1 Conservativeness under the null H_0^{ICI}

Theorem 1 establishes valid, finite-sample Type I error control for our method. The purpose of this section is to evaluate how conservative the Type I error is under various null distributions. Because our inference relies on the fact that matched pairs (X_{i_ℓ}, X_{j_ℓ}) are stochastically ordered under the null, intuitively the conservativeness of our test depends on the strength of monotonicity in the conditional distribution.

To see how the dependence between X and Z affects the rejection probability, we sample X from an additive noise model

$$X \mid Y, Z \sim \mathcal{N}(\mu(\gamma Z), 1),$$

where Y, Z are independent standard normal random variables. As long as μ is nondecreasing and $\gamma \ge 0$, this joint distribution belongs to the null H_0^{ICI}. The scalar γ controls the strength of the monotonicity of $X \mid Z$. In particular, as $\gamma \downarrow 0$ we expect the Type I error $\mathbb{P}\{p \le \alpha\}$ to approach α.

In our simulations, we consider two functions μ, the identity $\mu(z) = z$ and the Gaussian CDF $\mu(z) = \Phi(z)$. Figure 2 shows the Type I error as a function of γ for two levels of α. We observe similar results for each α, where the test typically becomes more conservative as γ increases, as expected. Under the null, our test is more conservative for cross-bin matching than for neighbour matching, since the Z values are further apart in cross-bin matching.

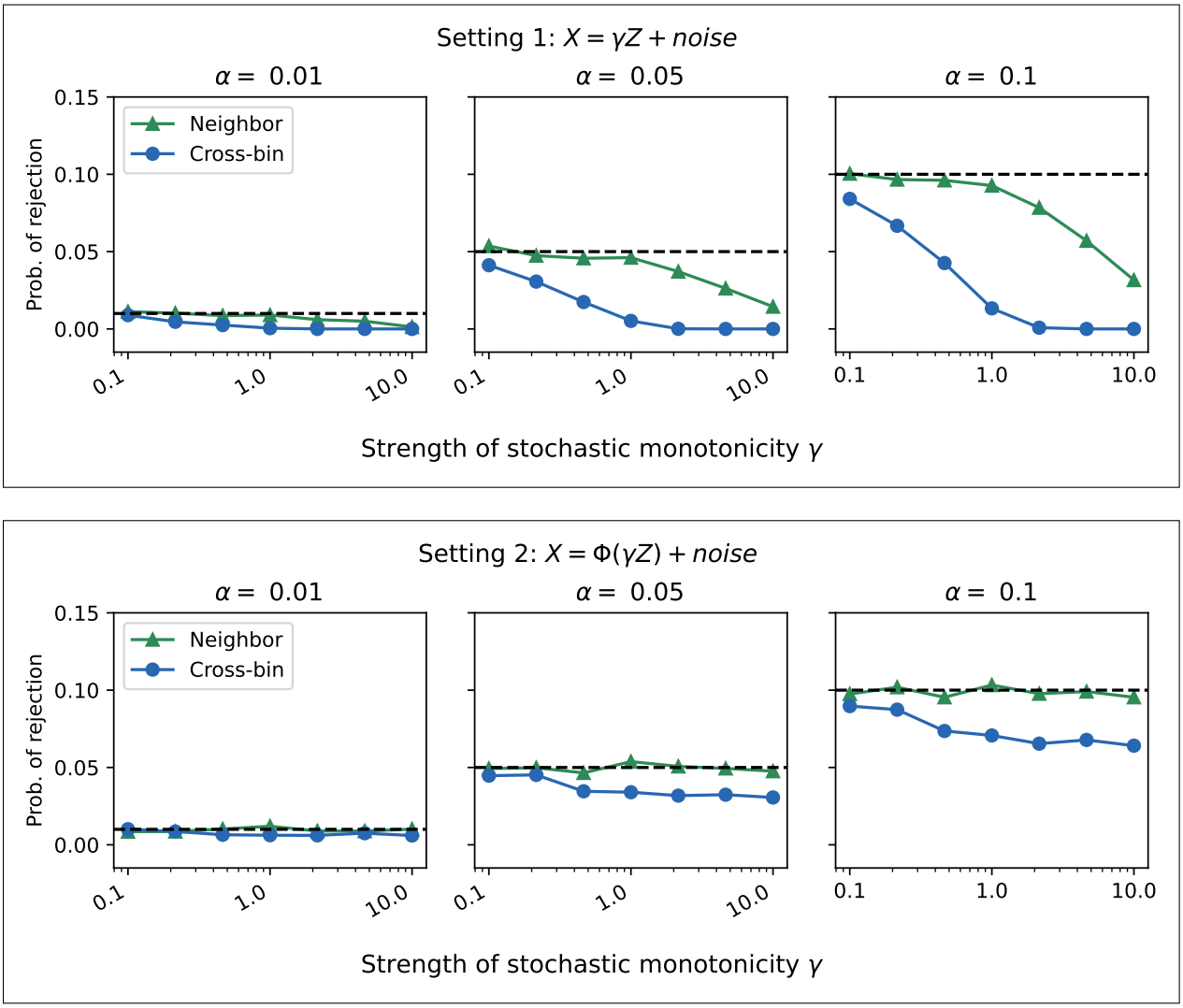


Figure 2: Simulation results illustrating Type I error control under the null H_0^{ICI} for two forms of the conditional mean $\mathbb{E}[X \mid Z]$. Each subplot shows the rejection probability of `PairSwap-ICI` test on data set of size 1000, averaged over 10^4 simulation trials, as a function of the strength of stochastic monotonicity γ.

5.2 Power under alternatives

In Section 4.3 we showed theoretically that our heuristic methods—neighbour matching and cross-bin matching—achieve high power (in fact, power that tends to 1) in the partial linear model (20) provided the signal β_n exceeds the detection threshold $n^{-1/2}$. In this section, we

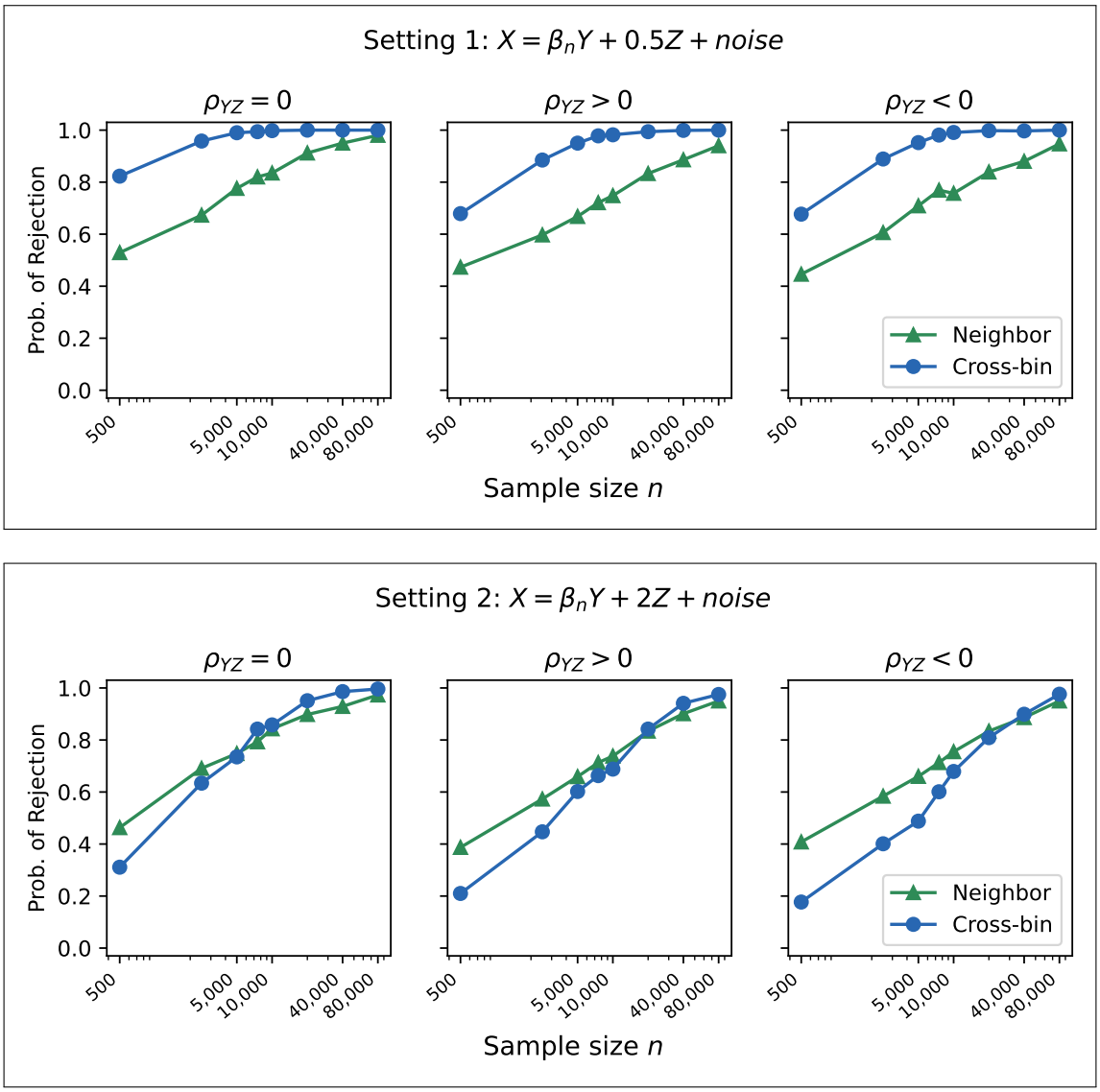


Figure 3: Simulation results demonstrating power for two alternatives at level $\alpha = 0.1$. Each subplot shows the rejection probability, averaged over 10^3 simulation trials, as a function of the sample size n. Columns correspond to different relationships between Y and Z. In each setting, X follows a Gaussian linear model with mean $\beta_n Y + \gamma Z$, where $\beta_n = n^{-1/3}$ and $\gamma = 0.5$ (above) or $\gamma = 2$ (below).

now examine this setting empirically. We sample data from the Gaussian linear model

$$X \mid Y, Z \sim \mathcal{N}(\beta_n Y + \gamma Z, 1),$$

with $\beta_n = n^{-1/3}$. The pair (Y, Z) is drawn from a bivariate Gaussian

$$\mathcal{N}\left(0, \begin{bmatrix} 1 & \rho_{YZ} \\ \rho_{YZ} & 1 \end{bmatrix}\right).$$

Figure 3 shows the power as a function of the sample size n for various choices of γ and ρ_{YZ}. In Setting 1, we set $\gamma = 0.5$, and cross-bin matching uniformly dominates neighbour

matching because it allows us to make many more matches of similar quality. On the other hand, in Setting 2 we set $\gamma = 2$, so the strong dependence of X on Z means the quality of a match (i_ℓ, j_ℓ) degrades much more quickly as the gap $Z_{j_\ell} - Z_{i_\ell}$ increases—that is, for cross-bin matching, where Z_{i_ℓ} and Z_{j_ℓ} may be farther apart than for neighbour matching, this gap may lead to conservativeness that results in a loss of power. However, with sufficiently large sample size, cross-bin matching performs at least as well as the neighbour matching. This is because the bin width decreases as n increases, so the quality of the cross-bin matches rivals that of the neighbour matches (with many more matches). The dependence ρ_{YZ} between Y and Z does not have a major impact on the power of these two methods.

6 Experiment on real data: risk factors for diabetes

In this section, we evaluate the performance of our proposed testing procedure on a real data set in three different experimental setting. The `PairSwap-ICI` method is implemented with the heuristic neighbour matching or with cross-bin matching (with $K = 50$ bins), and our statistic T takes the form in (11) with linear kernel $\psi(x, x') = x - x'$ and weights $w_\ell = \max\{Y_{i_\ell} - Y_{j_\ell}, 0\}$.

We use a data set[5] on the incidence of diabetes among the Pima population near Phoenix, Arizona, originally collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. The data set contains 768 observations, and it includes information on whether each of the patient has been diagnosed with diabetes according to World Health Organization standards. Additional variables provide data on the number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skinfold thickness, 2-hour serum insulin levels, body mass index (BMI), diabetes pedigree function and age.

It is well-known that the likelihood of developing diabetes increases with age (e.g., the CDC[6] lists advanced age as one of the risk factors for type 1 and type 2 diabetes). Therefore, if we choose X to represent the incidence of diabetes and Z as the age of the patient, then we would expect X to exhibit stochastic monotonicity with respect to Z (i.e., we expect that Assumption 1 holds, at least approximately). This is supported by the increasing trend we observe in Figure 4. Most of the other variables, such as `BloodPressure`, `BMI`, `Glucose`, and `Pregnancies`, are also considered potential risk factors for diabetes, as visualized in Figure 5—but does this association remain after we control for age? In this experiment, we aim to determine whether these variables remain significant risk factors for diabetes, even after controlling for age.

Experiment 1: marginal independence testing: In our first experiment, we consider six variables: `Pregnancies`, `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, and `BMI`, and aim to assess whether each of them is an individual risk factor for diabetes incidence. Specifically, we test the hypothesis $H_0 : X \perp\!\!\!\perp Y$, where Y represents one of the six variables listed above, while X is `Diabetes` (and since we are testing marginal rather than conditional independence, we do not attempt to control for Z, i.e., `Age`). For this purpose, we will be

[5]The data for this experiment were obtained from `https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database`. Additional data descriptions can be found in Smith et al. (1988).

[6]For more details, refer to the list of diabetes risk factors from U.S. Centers for Disease Control and Prevention.

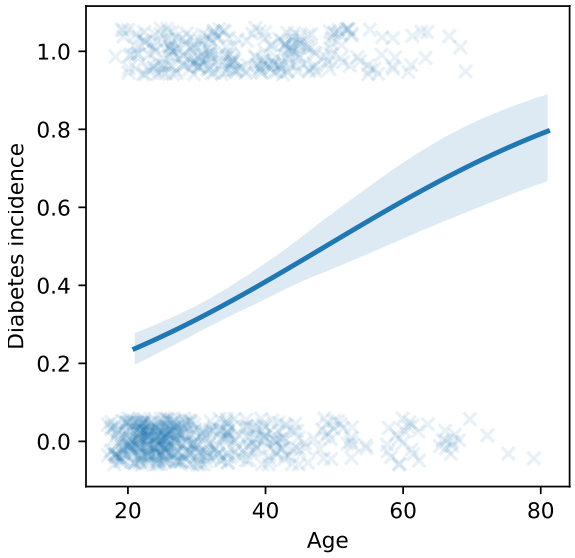


Figure 4: A scatter-plot (jittered for better visibility) of `Age` and `Diabetes Incidence` along with the fitted logistic regression model to demonstrate the stochastic monotonicity between them.

using the permutation test for independence with $T(\mathbf{X}, \mathbf{Y}) = \mathbf{X}^T\mathbf{Y}$, as outlined in Section 1.2.

Experiment 2: conditional independence testing, after controlling for age: Next, for the same set of six choices for Y, we test the hypothesis $H_0^{\mathrm{CI}} : X \perp\!\!\!\perp Y \mid Z$, where Z denotes `Age` (and X is `Diabetes` as before). This allows us to identify risk factors for diabetes after controlling for age. As noted earlier, we expect the distribution of $X \mid Z = z$ to be stochastically monotone in z, which supports the application of the `PairSwap-ICI` testing procedure developed in this paper for this purpose.

Experiment 3: conditional independence testing, with synthetic control $\tilde{X}$: Finally, we consider a semi-synthetic experiment where X is replaced by synthetic observations $\tilde{X}$, generated from an estimated model for $P_{X|Z}$ that satisfies stochastic monotonicity. We then test the hypothesis $\tilde{\mathcal{H}}_0 : \tilde{X} \perp\!\!\!\perp Y \mid Z$ for the same choices of Y from Experiment 1. Since $\tilde{X}$ is generated solely based on Z, the null hypothesis of conditional independence holds trivially in this synthetic setting. Tthe validity of our procedure should therefore ensure that the p-values generated by `PairSwap-ICI` are (super)uniformly distributed.

Now we give details on how the synthetic feature $\tilde{X}$ is generated. Since X is binary, it suffices to fit an isotonic regression to estimate the conditional mean $\mathbb{E}[X \mid Z]$ and then sample $\tilde{X}$ from the Bernoulli distribution with this fitted conditional mean. Following the theory of Henzi, Ziegel and Gneiting (2021, Theorem 1), this is the best approximation for $P_{X|Z}$ under *continuous ranked probability score (CRPS)*, while respecting the monotonocity constraint.

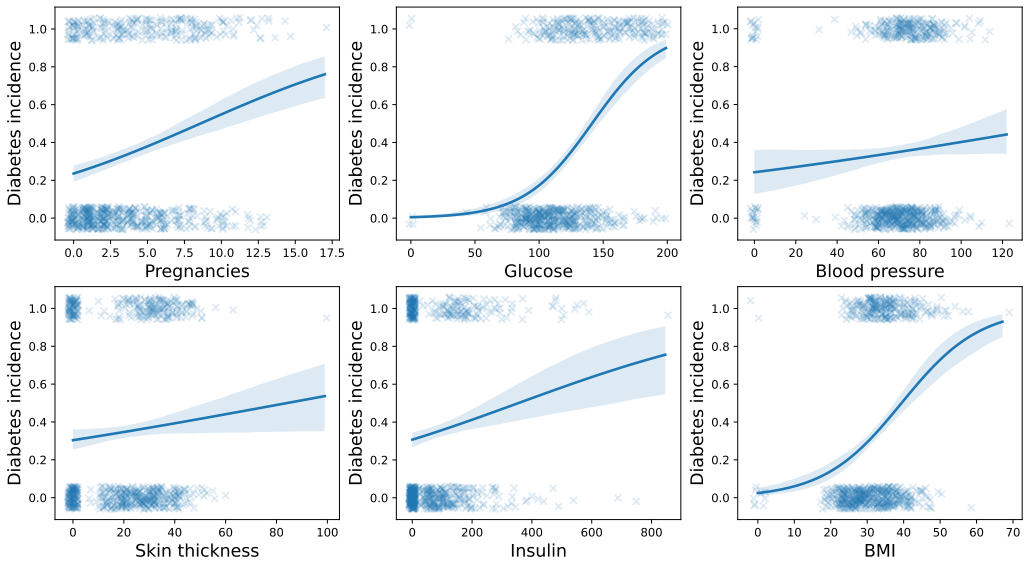


Figure 5: Scatter plots of X (jittered for better visibility) and other feature variables along with the fitted logistic regression models to demonstrate the dependence among these variables and `Diabetes Incidence`.

		Pregnancies	Glucose	Blood pressure	Skin thickness	Insulin	BMI
Permutation test (testing marginal indep.)		**0.001**(0.00)	**0.001**(0.00)	0.155(0.003)	0.117(0.002)	**0.023**(0.001)	**0.001**(0.00)
PairSwap-ICI	neighbour matching	0.433(0.005)	**0.002**(0.00)	0.496(0.005)	0.243(0.004)	0.159(0.003)	**0.033**(0.001)
	cross-bin matching	0.424(0.004)	**0.001**(0.00)	0.499(0.004)	0.123(0.003)	0.224(0.004)	**0.001**(0.00)
PairSwap-ICI (with synthetic control)	neighbour matching	0.511(0.005)	0.499(0.005)	0.501(0.005)	0.506(0.005)	0.497(0.005)	0.499(0.005)
	cross-bin matching	0.544(0.005)	0.550(0.005)	0.555(0.005)	0.558(0.005)	0.545(0.005)	0.555(0.005)

Table 1: p-values, averaged over 3000 random sub-samples along with the estimated standard errors (within brackets) for the different tests from different experiments, as outlined in Section 6. The p-values significant at the 0.05 level are marked in bold.

For each experiment, we generate 3,000 random sub-samples of the data, each consisting of half the size of the full data set. We then compute p-values using the permutation test for marginal independence and the `PairSwap-ICI` test for conditional independence. For experiments involving synthetic control, which require estimating $P_{X|Z}$, the sub-sampled data is further divided into training and test sets, with $\hat{P}_{X|Z}$ being computed in training set. For all the experiments, p-values are computed in the test set. Finally, we report the average p-values from the $3,000$ sub-samples, along with the corresponding choice of the Y variable in Table 1.

Results: Under the marginal independence test, four of the six variables are identified as having significant association with `Diabetes`—but, once we test conditional independence with the `PairSwap-ICI` test, only two of these associations are identified as significant. Specifically, `Glucose` and `BMI` both are identified as potential risk factors at the 0.05 level of significance by the marginal independence test, and also by the `PairSwap-ICI` test, even after controlling for `Age`. On the other hand, the variables `Pregnancies` and `Insulin` are significant only under the marginal test; this suggests that, after controlling for `Age`, the data does not provide sufficient evidence to support them as risk factors for `Diabetes`.

Finally, we also note that all the averaged p-value from Experiment 3 with synthetic control $\tilde{X}$ is concentrated around 0.5, for each of the choices of Y. Since $(\tilde{X}, Y, Z)$ satisfy H_0^{ICI} the p-values from Experiment 3 should be roughly uniform (or, if the test is conservative, superuniform), and thus this behaviour is expected as per the result we have established in Theorem 1.

7 Discussion

In this paper, we have developed a nonparametric test of conditional independence assuming only stochastic monotonicity of the conditional distribution $P_{X|Z}$. This nonparametric constraint is natural in many applications, and allows us to circumvent the impossibility of assumption-free conditional independence testing (Shah and Peters, 2020). We have introduced a variety of approaches to constructing a valid test statistic. Our test controls the Type I error in finite samples and has power against an array of alternatives. We close our discussion with some interesting connections to the literature, and potential avenues for future work.

- *Optimal power in general settings.* Theorems 4 and 5 bound the asymptotic power of our test above and below, where the upper and lower bounds differ by the appearance of the constant 2 in the lower bound—as we will see in Appendix C.4.3, this difference can be removed if we assume that the conditional distribution of $Y \mid Z$ is symmetric, but it remains an open question whether other tests (or, perhaps, the `PairSwap-ICI` but with a different kernel) may be able to avoid this assumption.

- *Avoiding data splitting.* The oracle matching test derived in Section 3 requires modeling the conditional mean and conditional variance of the kernel $\psi(X_i, X_j)$ as a function of Y_i, Y_j, Z_i, Z_j. We proposed to estimate these moments on a hold-out data set. Can we instead perform cross-fitting to improve power and retain finite-sample error control?

- *Alternative methods.* A notable benefit of our stochastic monotonicity assumption is that one can consistently estimate the conditional distribution $P_{X|Z}$ using isotonic distributional regression (Mösching and Dümbgen, 2020; Henzi, Ziegel and Gneiting, 2021). Hence, an alternative approach to testing the restricted null H_0^{ICI} is to first estimate this conditional distribution on one split of the data, and then run a conditional independence test which assumes knowledge of $P_{X|Z}$ (Berrett et al., 2020; Candès et al., 2018). Since we are plugging in the estimated conditional distribution, such tests will only be valid asymptotically. Is there any way to modify such tests to be valid in finite samples?

- *Connection with knockoffs and conditional randomization tests.* Creating synthetic copies of $\mathbf{X}$ via pairwise swaps has resemblance to other conditional independence testing procedures, such as knockoffs and the conditional randomization test (Candès et al., 2018), and the conditional permutation test (Berrett et al., 2020). One difference is that in our method, due to the stochastic ordering assumption, creating the swapped copies of $\mathbf{X}$ is potentially more conservative (i.e., the resulting p-value may be superuniform), since we are not working under the "sharp null".
- *Alternative shape constraints.* We view stochastic monotonicity as one form of positive dependence for the joint distribution (X, Z). Are there natural approaches to test conditional independence under other models of dependence, such as likelihood-ratio ordering or total positivity, or under other shape constraints, such as unimodality (Karlin, 1968; Shaked and Shanthikumar, 2007; Mösching and Dümbgen, 2024)?

References

Albert, M., Laurent, B., Marrel, A. and Meynaoui, A. (2022) Adaptive test of independence based on HSIC measures. *The Annals of Statistics*, **50**, 858–879.

Azadkia, M. and Chatterjee, S. (2021) A simple measure of conditional dependence. *The Annals of Statistics*, **49**, 3070–3102.

Barber, R. F., Candès, E. J. and Samworth, R. J. (2020) Robust inference with knockoffs. *The Annals of Statistics*, **48**, 1409–1431.

Berrett, T. B., Kontoyiannis, I. and Samworth, R. J. (2021) Optimal rates for independence testing via U-statistic permutation tests. *The Annals of Statistics*, **49**, 2457–2490.

Berrett, T. B. and Samworth, R. J. (2019) Nonparametric independence testing via mutual information. *Biometrika*, **106**, 547–566.

Berrett, T. B. and Samworth, R. J. (2021) USP: an independence test that improves on Pearson's chi-squared and the G-test. *Proceedings of the Royal Society A*, **477**, 20210549.

Berrett, T. B., Wang, Y., Barber, R. F. and Samworth, R. J. (2020) The conditional permutation test for independence while controlling for confounders. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **82**, 175–197.

Candès, E., Fan, Y., Janson, L. and Lv, J. (2018) Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **80**, 551–577.

Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap Methods and their Application*, vol. 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Duan, R. and Pettie, S. (2014) Linear-time approximation for maximum weight matching. *Journal of the ACM (JACM)*, **61**, 1–23.

Edmonds, J. (1965) Paths, trees, and flowers. *Canadian Journal of mathematics*, **17**, 449–467.

Gabow, H. N. (1985) A scaling algorithm for weighted matching on general graphs. In *26th Annual Symposium on Foundations of Computer Science (sfcs 1985)*, 90–100, IEEE.

Hagberg, A., Swart, P. and Schult, D. (2008) Exploring network structure, dynamics, and function using NetworkX. Tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Henzi, A., Ziegel, J. F. and Gneiting, T. (2021) Isotonic distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **83**, 963–993.

Kalisch, M. and Bühlmann, P. (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, **8**.

Karlin, S. (1968) *Total positivity. Vol. I.* Stanford University Press, Stanford, CA.

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

Kim, I., Neykov, M., Balakrishnan, S. and Wasserman, L. (2022) Local permutation tests for conditional independence. *Ann. Statist.*, **50**, 3388–3414.

Lei, J. (2020) Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, **26**, 767 – 798.

Lundborg, A. R., Kim, I., Shah, R. D. and Samworth, R. J. (2024+) The Projected Covariance Measure for assumption-lean variable significance testing. *The Annals of Statistics, to appear. arXiv preprint arXiv:2211.02039*.

Lundborg, A. R., Shah, R. D. and Peters, J. (2022) Conditional independence testing in Hilbert spaces with applications to functional data analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **84**, 1821–1850.

Mösching, A. and Dümbgen, L. (2020) Monotone least squares and isotonic quantiles. *Electron. J. Stat.*, **14**, 24–49.

Mösching, A. and Dümbgen, L. (2024) Estimation of a likelihood ratio ordered family of distributions. *Stat. Comput.*, **34**, Paper No. 58, 16.

Neykov, M., Balakrishnan, S. and Wasserman, L. (2021) Minimax optimal conditional independence testing. *Ann. Statist.*, **49**, 2151–2177.

Niu, Z., Chakraborty, A., Dukes, O. and Katsevich, E. (2024) Reconciling model-X and doubly robust approaches to conditional independence testing. *The Annals of Statistics*, **52**, 895–921.

O'Mahony, C., Jichi, F., Pavlou, M., Monserrat, L., Anastasakis, A., Rapezzi, C., Biagini, E., Gimeno, J. R., Limongelli, G., McKenna, W. J. et al. (2014) A novel clinical risk prediction model for sudden cardiac death in hypertrophic cardiomyopathy (HCM risk-SCD). *European Heart Journal*, **35**, 2010–2020.

Pfister, N., Bühlmann, P., Schölkopf, B. and Peters, J. (2018) Kernel-based tests for joint independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **80**, 5–31.

Phipson, B. and Smyth, G. K. (2010) Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.*, **9**, Art. 39, 14.

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

Shaked, M. and Shanthikumar, J. G. (2007) *Stochastic orders*. Springer Series in Statistics, Springer, New York.

Shevtsova, I. G. (2010) An improvement of convergence rate estimates in the Lyapunov theorem. *Doklady Mathematics*, **82**, 862–864.

Smith, J. W., Everhart, J. E., Dickson, W., Knowler, W. C. and Johannes, R. S. (1988) Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, 261, American Medical Informatics Association.

Villani, C. (2009) *Optimal transport: old and new*, vol. 338. Springer.

Yan, Z., Cai, M., Han, X., Chen, Q. and Lu, H. (2023) The interaction between age and risk factors for diabetes and prediabetes: a community-based cross-sectional study. *Diabetes, Metabolic Syndrome and Obesity*, 85–93.

A Proof of Theorem 2

The proof of this result follows the same structure as the proof of Theorem 1.

Step 1: some deterministic properties of the p-value. Define a function $\hat{p}_M : \mathbb{R}^n \times (\{\pm 1\}^L)^M \to [0,1]$ as

$$\hat{p}_M(\mathbf{x}; \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}) = \frac{1 + \sum_{m=1}^{M} \mathbb{1}\left\{T(\mathbf{x}^{\mathbf{s}^{(m)}}) \geq T(\mathbf{x})\right\}}{1+M}.$$

As in the proof of Theorem 1, this function is monotone nonincreasing in each x_{i_ℓ}, and monotone nondecreasing in each x_{j_ℓ}.

Step 2: compare to the sharp null. Define $\mathbf{X}_\sharp$ as in the proof of Theorem 1. Following identical arguments as in that proof, we can verify that, for any fixed $\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}$, it holds that

$$\hat{p}_M\big(\mathbf{X}_\sharp; \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}\big) \preceq_{\text{st}} \hat{p}_M\big(\mathbf{X}; \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}\big)$$

conditional on $\mathbf{Y}, \mathbf{Z}$. Since $\hat{p}_M = \hat{p}_M(\mathbf{X}; \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)})$ by construction, we therefore have

$$\mathbb{P}\left\{\hat{p}_M \leq \alpha \;\middle|\; \mathbf{Y}, \mathbf{Z}, \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}\right\} \leq \mathbb{P}\left\{\hat{p}_M\big(\mathbf{X}_\sharp; \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}\big) \leq \alpha \;\middle|\; \mathbf{Y}, \mathbf{Z}, \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}\right\}.$$

Marginalizing over the random draw of the swaps, $\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)} \overset{\text{iid}}{\sim} \text{Unif}(\{\pm 1\}^L)$, we therefore have

$$\mathbb{P}\left\{\hat{p}_M \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\right\} \leq \mathbb{P}\left\{\hat{p}_M\big(\mathbf{X}_\sharp; \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}\big) \leq \alpha \;\Big|\; \mathbf{Y}, \mathbf{Z}\right\}.$$

Step 3: validity under the sharp null. We now need to verify the validity of the Monte Carlo p-value, under the sharp null. Unlike the first two steps, for this step the arguments are somewhat different than in the proof of Theorem 1.

First, let $\mathbf{s}^{(0)}$ be an additional draw from $\text{Unif}(\{\pm 1\}^L)$, sampled independently from all other random variables. Then it holds that

$$(\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}) \overset{\text{d}}{=} \big(\mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\big),$$

where $\circ$ denotes the elementwise product, and so

$$\hat{p}_M\big(\mathbf{X}_\sharp; \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}\big) \overset{\text{d}}{=} \hat{p}_M\Big(\mathbf{X}_\sharp; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\Big)$$

conditional on $\mathbf{Y}, \mathbf{Z}$. Moreover, by construction of the sharp null data $\mathbf{X}_\sharp$,

$$\mathbf{X}_\sharp \overset{\text{d}}{=} (\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}$$

holds conditional on $\mathbf{Y}, \mathbf{Z}, \mathbf{s}^{(0)}, \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}$, and therefore

$$\hat{p}_M\Big(\mathbf{X}_\sharp; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\Big) \overset{\text{d}}{=} \hat{p}_M\Big((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\Big)$$

holds conditional on $\mathbf{Y}, \mathbf{Z}, \mathbf{s}^{(0)}, \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}$. Combining all these calculations so far, then, we have

$$\hat{p}_M\big(\mathbf{X}_\sharp; \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}\big) \overset{\text{d}}{=} \hat{p}_M\Big((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\Big), \tag{22}$$

conditional on $\mathbf{Y}, \mathbf{Z}$.

Next we calculate this last p-value: by definition,

$$\begin{aligned}\hat{p}_M\Big((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\Big) &= \frac{1 + \sum_{m=1}^M \mathbb{1}\left\{T\left((\mathbf{X}_\sharp)^{\mathbf{s}^{(m)}}\right) \geq T\left((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}\right)\right\}}{1+M} \\ &= \frac{\sum_{m=0}^M \mathbb{1}\left\{T\left((\mathbf{X}_\sharp)^{\mathbf{s}^{(m)}}\right) \geq T\left((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}\right)\right\}}{1+M},\end{aligned}$$

where the first step holds since, for each $m = 1, \ldots, M$,

$$\Big[(\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}\Big]^{\mathbf{s}^{(0)} \circ \mathbf{s}^{(m)}} = (\mathbf{X}_\sharp)^{\mathbf{s}^{(0)} \circ \mathbf{s}^{(0)} \circ \mathbf{s}^{(m)}} = (\mathbf{X}_\sharp)^{\mathbf{s}^{(m)}}$$

by definition of the swap operation. In other words, the p-value $\hat{p}_M\Big((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\Big)$ is simply comparing the value of the statistic $T((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}})$ against the list of $M+1$ values $T((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}), \ldots, T((\mathbf{X}_\sharp)^{\mathbf{s}^{(M)}})$. We therefore have

$$\mathbb{P}\left\{\hat{p}_M\Big((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\Big) \leq \alpha \;\Big|\; \mathbf{X}_\sharp, \mathbf{Y}, \mathbf{Z}\right\} \leq \alpha,$$

since, conditional on $\mathbf{X}_\sharp, \mathbf{Y}, \mathbf{Z}$, the sign vectors $\mathbf{s}^{(0)}, \ldots, \mathbf{s}^{(M)}$ are i.i.d., and therefore the rank of the statistic $T((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}})$ among the list $T((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}), \ldots, T((\mathbf{X}_\sharp)^{\mathbf{s}^{(M)}})$ is uniformly distributed. Marginalizing over $\mathbf{X}_\sharp$, therefore,

$$\mathbb{P}\left\{\hat{p}_M\Big((\mathbf{X}_\sharp)^{\mathbf{s}^{(0)}}; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\Big) \leq \alpha \;\Big|\; \mathbf{Y}, \mathbf{Z}\right\} \leq \alpha.$$

Finally, combining this with our earlier calculation (22), we have

$$\mathbb{P}\left\{\hat{p}_M\big(\mathbf{X}_\sharp; \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)}\big) \leq \alpha \;\big|\; \mathbf{Y}, \mathbf{Z}\right\} \leq \alpha,$$

which completes the proof.

B Proof of Lemma 3

For $\ell \in [L]$, let

$$u_\ell = \mathbb{E}\left[\psi(X_{i_\ell}, X_{j_\ell}) \mid \mathbf{Y}, \mathbf{Z}\right] \quad \text{and} \quad v_\ell = \big(\mathrm{Var}\left(\psi(X_{i_\ell}, X_{j_\ell}) \mid \mathbf{Y}, \mathbf{Z}\right)\big)^{1/2}.$$

Then by Cauchy–Schwarz, for any $(w_1, \ldots, w_L) \in [0, \infty)^L$,

$$\begin{aligned}\frac{\sum_{\ell=1}^L w_\ell \mathbb{E}\left[\psi(X_{i_\ell}, X_{j_\ell}) \mid \mathbf{Y}, \mathbf{Z}\right]}{\sqrt{\sum_{\ell=1}^L w_\ell^2 \mathrm{Var}\left(\psi(X_{i_\ell}, X_{j_\ell}) \mid \mathbf{Y}, \mathbf{Z}\right)}} &= \frac{\sum_{\ell=1}^L w_\ell u_\ell}{\sqrt{\sum_{\ell=1}^L w_\ell^2 v_\ell^2}} \leq \frac{\sum_{\ell=1}^L w_\ell \max\{u_\ell, 0\}}{\sqrt{\sum_{\ell=1}^L w_\ell^2 v_\ell^2}} \\ &= \frac{\sum_{\ell=1}^L (w_\ell v_\ell) \cdot \frac{\max\{u_\ell, 0\}}{v_\ell}}{\sqrt{\sum_{\ell=1}^L (w_\ell v_\ell)^2}} \leq \left(\sum_{\ell=1}^L \frac{\max\{u_\ell, 0\}^2}{v_\ell^2}\right)^{1/2},\end{aligned}$$

with equality if and only if $w_\ell \propto \max\{u_\ell, 0\}/v_\ell^2$ for $\ell \in [L]$.

C Proof of the results from Section 4

In this section, we prove the results presented in Section 4. Throughout this appendix, we assume that the statistic T admits the form in (11) with $\psi(x, x') = x - x'$, and that $\mathcal{Y} = \mathcal{Z} = \mathbb{R}$; the partial ordering $\preceq$ for Z will simply be the usual ordering $\leq$ on $\mathbb{R}$. The organization of this appendix is as follows.

- We begin in Appendix C.1 by proving finite sample and asymptotic upper and lower bounds on the conditional power of `PairSwap-ICI` test for any valid matching and weighting scheme, and any statistic T of the form (11) with a shared linear kernel $\psi(x, x') = x - x'$.
- Next, in Appendix C.2, we specialize these results to the oracle matching under two cases: one assuming access to oracle knowledge of μ (i.e., Theorem 4) and another with μ estimated from data (i.e., Theorem 5).
- Then, we shift our attention to the partially linear Gaussian models in (20). In Appendices C.3 and C.4 we prove the asymptotic behavior of conditional power for neighbour matching (Theorem 10) and for cross bin matching (Theorem 11), respectively.

- In Appendix C.5 we prove the corollaries and lemmas from Section 4.
- In Appendix C.6 we introduce an oracle matching, namely isotonic median matching, and discuss a key property of the same, which allows us to prove the results in Appendix C.2.
- Finally, in Appendix C.7 we prove the lemmas stated in Appendices C.1—C.4.

Notation. We write $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ as shorthand for independent triples $(X_i, Y_i, Z_i)_{i\in[n]}$. For $L \equiv L_n \in \mathbb{N}$, a matching $\{(i_1, j_1), (i_2, j_2), \ldots, (i_L, j_L)\} \in \mathcal{M}_n(\mathbf{Z})$ and a vector $\mathbf{V} = (V_1, \ldots, V_n) \in \mathbb{R}^n$, we define $\Delta\mathbf{V} = (\Delta_1\mathbf{V}, \ldots, \Delta_L\mathbf{V}) \in \mathbb{R}^L$ with entries $\Delta_\ell\mathbf{V} := V_{i_\ell} - V_{j_\ell}$. Given any vector $\mathbf{v}$, we write $\mathbf{v}^+$ to denote the vector with ith component $v_{i+} = \max\{v_i, 0\}$. We write $\mathbf{a} \circ \mathbf{b}$ for the Hadamard product of vectors $\mathbf{a}$, $\mathbf{b}$ of the same dimension, with ith component $a_i \cdot b_i$. For $k \geq 1$ and a distribution P_ζ on $\mathbb{R}$ with finite k-th moment, let

$$\rho_k = \left(\mathbb{E}_{\substack{\zeta,\zeta' \sim P_\zeta \\ \zeta \perp\!\!\!\perp \zeta'}} \left[|\zeta - \zeta'|^k \right] \right)^{1/k}. \tag{23}$$

In particular, $\rho_2 = \sqrt{2}\sigma$ where σ^2 is variance of P_ζ.

C.1 A general result on power of `PairSwap-ICI` test

Here, we consider any valid matching and weighting scheme, and state finite-sample lower and upper bounds on the conditional power $\mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\}$ of `PairSwap-ICI` test. Further, under the asymptotic regime of Section 4.1, we state asymptotic high-probability upper and lower bounds for the same quantity. We first define several quantities in terms of the weight vector $\mathbf{w} = (w_1, \ldots, w_L) \in \mathbb{R}^L$, $\delta > 0$ and $\{\rho_k : k \leq 6\}$, defined in (23): let

$$\begin{aligned}
\epsilon_{1,\delta,U} &= \frac{\|\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_2^2}{\rho_2^2 \|\mathbf{w}\|_2^2} + \frac{\rho_4^2}{\rho_2^2 \delta^{1/2}} \frac{\|\mathbf{w}\|_\infty}{\|\mathbf{w}\|_2} + \frac{2}{\rho_2 \delta^{1/2}} \frac{\|\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_\infty}{\|\mathbf{w}\|_2}, \\
\epsilon_{1,\delta,L} &= \frac{\|\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_2^2}{\rho_2^2 \|\mathbf{w}\|_2^2} - \frac{\rho_4^2}{\rho_2^2 \delta^{1/2}} \frac{\|\mathbf{w}\|_\infty}{\|\mathbf{w}\|_2} - \frac{2}{\rho_2 \delta^{1/2}} \frac{\|\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_\infty}{\|\mathbf{w}\|_2}, \\
\epsilon_{2,\delta} &= \frac{0.56}{\big((1 + \epsilon_{1,\delta,L}) \vee 0\big)^{3/2}} \left\{ \left(\frac{\|\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_2}{\rho_2 \|\mathbf{w}\|_2} \right)^{2/3} \cdot \left(\frac{\|\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_\infty}{\rho_2 \|\mathbf{w}\|_2} \right)^{1/3} \right. \\
&\qquad \left. + \left(\frac{\rho_3^3}{\rho_2^3} \frac{\|\mathbf{w}\|_\infty}{\|\mathbf{w}\|_2} + \frac{\rho_6^3}{\delta^{1/2} \rho_2^3} \cdot \frac{\|\mathbf{w}\|_\infty^2}{\|\mathbf{w}\|_2^2} \right)^{1/3} \right\}^3, \\
\epsilon_{3,\delta} &= \frac{0.56 \rho_3^3}{\rho_2^3} \cdot \frac{\|\mathbf{w}\|_\infty}{\|\mathbf{w}\|_2} + 5\delta.
\end{aligned} \tag{24}$$

Theorem C.12. *Suppose that $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ satisfies (16), and that $\mathbf{w} = \mathbf{w}(\mathbf{Y}, \mathbf{Z}) \in \mathbb{R}^n$ is our chosen weight vector, which is assumed to satisfy $\|\mathbf{w}\|_2 > 0$. Write $\Omega_0 := \{\|\mathbf{w} \circ \Delta\mathbf{X}\|_2 > 0\}$. Then*

(i) For any $\delta > 0$ and $\alpha \in (0, 1/2 - \epsilon_{2,\delta}]$,

$$\Phi\left(\frac{\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\mathbf{w}\|_2} - \bar{\Phi}^{-1}\big((\alpha-\epsilon_{2,\delta})\vee 0\big)(1+\epsilon_{1,\delta,U})^{1/2}\right) - \epsilon_{3,\delta} - \mathbb{P}\left\{\Omega_0^c \mid \mathbf{Y},\mathbf{Z}\right\}$$
$$\leq \mathbb{P}\left\{p \leq \alpha \mid \mathbf{Y},\mathbf{Z}\right\} \leq \Phi\left(\frac{\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha+\epsilon_{2,\delta})\cdot\big((1+\epsilon_{1,\delta,L})\vee 0\big)^{1/2}\right) + \epsilon_{3,\delta},$$

where $\epsilon_{1,\delta,U}, \epsilon_{1,\delta,L}, \epsilon_{2,\delta}$, and $\epsilon_{3,\delta}$ are as in (24).

(ii) Further, suppose that the weights and matching scheme satisfy

Assumption A1. $\|\mathbf{w}\|_\infty \vee \big(\|\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2^{2/3}\cdot\|\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{1/3}\big) = o_P(\|\mathbf{w}\|_2)$.

Assumption A2. $\|\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2 = o_P\big(\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z})\big)$.

Then for $\alpha \in (0, 1/2)$,

$$\Phi\left(\frac{\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\sqrt{2}\sigma\|\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha)\right) - o_P(1) \leq \mathbb{P}\left\{p \leq \alpha \mid \mathbf{Y},\mathbf{Z}\right\}$$
$$\leq \Phi\left(\frac{\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\sqrt{2}\sigma\|\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha)\right) + o_P(1).$$

Proof. We note that for fixed $n \in \mathbb{N}$, and conditional on $(\mathbf{X},\mathbf{Y},\mathbf{Z})$, the quantity $\mathbf{s}^T(\mathbf{w}\circ\Delta\mathbf{X})$ is a sum of independent random variables $\{s_\ell\cdot w_\ell\cdot\Delta_\ell\mathbf{X}\}_{\ell\in[L]}$, and has mean 0 and variance $\|\mathbf{w}\circ\Delta\mathbf{X}\|_2^2$. Hence, writing $U := \frac{\mathbf{1}^T(\mathbf{w}\circ\Delta\mathbf{X})}{\|\mathbf{w}\circ\Delta\mathbf{X}\|_2}\mathbb{1}_{\Omega_0}$, we have by the Berry–Esseen theorem (Shevtsova, 2010, Theorem 1) that on Ω_0,

$$|p - \bar{\Phi}(U)| \leq \sup_{x\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\mathbf{s}^T(\mathbf{w}\circ\Delta\mathbf{X})}{\|\mathbf{w}\circ\Delta\mathbf{X}\|_2} \geq x \;\middle|\; \mathbf{X},\mathbf{Y},\mathbf{Z}\right\} - \bar{\Phi}(x)\right| \leq 0.56\cdot\frac{\|\mathbf{w}\circ\Delta\mathbf{X}\|_3^3}{\|\mathbf{w}\circ\Delta\mathbf{X}\|_2^3}.$$

Hence, since $p = 1$ on Ω_0^c, we therefore have

$$\mathbb{P}\left[\left\{U \geq \bar{\Phi}^{-1}\left(\left(\alpha - 0.56\frac{\|\mathbf{w}\circ\Delta\mathbf{X}\|_3^3}{\|\mathbf{w}\circ\Delta\mathbf{X}\|_2^3}\right)\vee 0\right)\right\}\bigcap\Omega_0 \;\middle|\; \mathbf{Y},\mathbf{Z}\right] \leq \mathbb{P}\left\{p \leq \alpha \mid \mathbf{Y},\mathbf{Z}\right\}$$
$$\leq \mathbb{P}\left[\left\{U \geq \bar{\Phi}^{-1}\left(\left(\alpha + 0.56\frac{\|\mathbf{w}\circ\Delta\mathbf{X}\|_3^3}{\|\mathbf{w}\circ\Delta\mathbf{X}\|_2^3}\right)\wedge 1\right)\right\}\bigcap\Omega_0 \;\middle|\; \mathbf{Y},\mathbf{Z}\right].$$

Now $\mathbf{1}^T(\mathbf{w}\circ\Delta\mathbf{X}) = \mathbf{w}^T\Delta\mathbf{X} = \mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z}) + \mathbf{w}^T\Delta\boldsymbol{\zeta}$ and $\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z})$ is a measurable function of $(\mathbf{Y},\mathbf{Z})$. On the other hand, conditional on $(\mathbf{Y},\mathbf{Z})$, the quantity $\mathbf{w}^T\Delta\boldsymbol{\zeta}$ is a sum of independent random variables $\{w_\ell\cdot\Delta_\ell\boldsymbol{\zeta}\}_{\ell\in[L]}$, and has mean 0 and variance $\rho_2^2\|\mathbf{w}\|_2^2$. Another application of the Berry–Esseen theorem then yields that

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\mathbf{w}^T\Delta\boldsymbol{\zeta}}{\rho_2\|\mathbf{w}\|_2} \geq x \middle| \mathbf{Y},\mathbf{Z}\right\} - \bar{\Phi}(x)\right| \leq \frac{0.56\rho_3^3}{\rho_2^3}\cdot\frac{\|\mathbf{w}\|_3^3}{\|\mathbf{w}\|_2^3} \leq \frac{0.56\rho_3^3}{\rho_2^3}\cdot\frac{\|\mathbf{w}\|_\infty}{\|\mathbf{w}\|_2}. \tag{25}$$

Thus, recalling the events Ω_2 and Ω_3 from Lemma C.19 we have by that lemma that

$$\begin{aligned}
&\mathbb{P}\left[\left\{U \geq \bar{\Phi}^{-1}\left(\left(\alpha + 0.56\frac{\|\mathbf{w}\circ\Delta\mathbf{X}\|_3^3}{\|\mathbf{w}\circ\Delta\mathbf{X}\|_2^3}\right)\wedge 1\right)\right\}\bigcap\Omega_0 \;\middle|\; \mathbf{Y},\mathbf{Z}\right] \\
&\leq \mathbb{P}\left[\left\{\frac{\mathbf{w}^T\Delta\boldsymbol{\zeta}}{\rho_2\|\mathbf{w}\|_2} \geq -\frac{\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\mathbf{w}\|_2} + \frac{\|\mathbf{w}\circ\Delta\mathbf{X}\|_2}{\rho_2\|\mathbf{w}\|_2}\bar{\Phi}^{-1}(\alpha+\epsilon_{2,\delta})\right\}\bigcap\Omega_0 \;\middle|\; \mathbf{Y},\mathbf{Z}\right] + \mathbb{P}\left\{\Omega_3^c\right\} \\
&\leq \mathbb{P}\left[\left\{\frac{\mathbf{w}^T\Delta\boldsymbol{\zeta}}{\rho_2\|\mathbf{w}\|_2} \geq -\frac{\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\mathbf{w}\|_2} + \bar{\Phi}^{-1}(\alpha+\epsilon_{2,\delta})\big((1+\epsilon_{1,\delta,L})\vee 0\big)^{1/2}\right\}\bigcap\Omega_0 \;\middle|\; \mathbf{Y},\mathbf{Z}\right] \\
&\qquad + \mathbb{P}\left\{\Omega_2^c\right\} + \mathbb{P}\left\{\Omega_3^c\right\} \\
&\leq \Phi\left(\frac{\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha+\epsilon_{2,\delta})\cdot\big((1+\epsilon_{1,\delta,L})\vee 0\big)^{1/2}\right) + \frac{0.56\rho_3^3}{\rho_2^3}\cdot\frac{\|\mathbf{w}\|_\infty}{\|\mathbf{w}\|_2} + 5\delta \\
&\leq \Phi\left(\frac{\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha+\epsilon_{2,\delta})\cdot\big((1+\epsilon_{1,\delta,L})\vee 0\big)^{1/2}\right) + \epsilon_{3,\delta}.
\end{aligned}$$

Similarly, recalling the events Ω_1, Ω_2 and Ω_3 from Lemma C.19 for the lower bound, and writing $\Omega_0^* = \Omega_0 \cap \Omega_1 \cap \Omega_3$,

$$\begin{aligned}
&\mathbb{P}\left[\left\{U \geq \bar{\Phi}^{-1}\left(\left(\alpha - 0.56\frac{\|\mathbf{w}\circ\Delta\mathbf{X}\|_3^3}{\|\mathbf{w}\circ\Delta\mathbf{X}\|_2^3}\right)\vee 0\right)\right\}\bigcap\Omega_0 \;\middle|\; \mathbf{Y},\mathbf{Z}\right] \\
&\geq \mathbb{P}\left[\left\{\frac{\mathbf{w}^T\Delta\boldsymbol{\zeta}}{\rho_2\|\mathbf{w}\|_2} \geq -\frac{\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\mathbf{w}\|_2} + \frac{\|\mathbf{w}\circ\Delta\mathbf{X}\|_2}{\rho_2\|\mathbf{w}\|_2}\bar{\Phi}^{-1}\big((\alpha-\epsilon_{2,\delta})\vee 0\big)\right\}\bigcap\Omega_0^* \;\middle|\; \mathbf{Y},\mathbf{Z}\right] \\
&\geq \mathbb{P}\left[\left\{\frac{\mathbf{w}^T\Delta\boldsymbol{\zeta}}{\rho_2\|\mathbf{w}\|_2} \geq -\frac{\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\mathbf{w}\|_2} + \bar{\Phi}^{-1}\big((\alpha-\epsilon_{2,\delta})\vee 0\big)(1+\epsilon_{1,\delta,U})^{1/2}\right\}\bigcap\Omega_0^* \;\middle|\; \mathbf{Y},\mathbf{Z}\right] \\
&\geq \Phi\left(\frac{\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\mathbf{w}\|_2} - \bar{\Phi}^{-1}\big((\alpha-\epsilon_{2,\delta})\vee 0\big)(1+\epsilon_{1,\delta,U})^{1/2}\right) - \frac{0.56\rho_3^3\|\mathbf{w}\|_\infty}{\rho_2^3\|\mathbf{w}\|_2} - \mathbb{P}\left\{\Omega_0^{*c} \mid \mathbf{Y},\mathbf{Z}\right\} \\
&\geq \Phi\left(\frac{\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\mathbf{w}\|_2} - \bar{\Phi}^{-1}\big((\alpha-\epsilon_{2,\delta})\vee 0\big)(1+\epsilon_{1,\delta,U})^{1/2}\right) - \epsilon_{3,\delta} - \mathbb{P}\left\{\Omega_0^c \mid \mathbf{Y},\mathbf{Z}\right\}.
\end{aligned}$$

This completes the proof of the first part of the result.

For the second part, define

$$\epsilon_{1,\delta} := \frac{\rho_4^2}{\rho_2^2\delta^{1/2}}\cdot\frac{\|\mathbf{w}\|_\infty}{\|\mathbf{w}\|_2} + \frac{2}{\rho_2\delta^{1/2}}\cdot\frac{\|\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\mathbf{w}\|_2} = o_P(1),$$

by Assumption **A1**, and observe that $(1+\epsilon_{1,\delta,U})^{1/2} \leq \frac{\|\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2}{\rho_2\|\mathbf{w}\|_2} + (1+\epsilon_{1,\delta})^{1/2}$ and $(1+\epsilon_{1,\delta,L})^{1/2} \geq (1-\epsilon_{1,\delta})^{1/2}$. It follows by Assumption **A1** that $\epsilon_{2,\delta} = o_P(1)$. Moreover,

$$\begin{aligned}
\limsup_{n\to\infty}\mathbb{P}\left\{\Omega_0^c \mid \mathbf{Y},\mathbf{Z}\right\} &\leq \limsup_{n\to\infty}\mathbb{P}\left\{\Omega_2^c\cap\{1+\epsilon_{1,\delta,L}>0\} \mid \mathbf{Y},\mathbf{Z}\right\} + \limsup_{n\to\infty}\mathbb{P}\left\{1+\epsilon_{1,\delta,L}\leq 0\right\} \\
&= \limsup_{n\to\infty}\mathbb{P}\left\{\Omega_2^c \mid \mathbf{Y},\mathbf{Z}\right\} \leq 2\delta.
\end{aligned}$$

Now, $\epsilon_{3,\delta} - 5\delta = \mathrm{o}_P(1)$ by Assumption **A1**. Thus, by Assumption **A2** and part *(i)*,

$$\begin{aligned}
&\mathbb{P}\{p \le \alpha \mid \mathbf{Y}, \mathbf{Z}\} \\
&\ge \Phi\left(\frac{\mathbf{w}^T \Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\mathbf{w}\|_2} - \bar{\Phi}^{-1}\big(\alpha - \mathrm{o}_P(1)\big)\left(\frac{\|\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2}{\rho_2\|\mathbf{w}\|_2} + \big(1+\mathrm{o}_P(1)\big)\right)\right) - \mathrm{o}_P(1) - 7\delta \\
&= \Phi\left(\left(\frac{\mathbf{w}^T \Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha) + \mathrm{o}_P(1)\right)\big(1+\mathrm{o}_P(1)\big)\right) - \mathrm{o}_P(1) - 7\delta,
\end{aligned}$$

and since $\delta > 0$ was arbitrary, the desired asymptotic lower bound holds by noting that $\rho_2 = \sqrt{2}\sigma$. The asymptotic upper bound follows by a very similar (in fact, slightly more straightforward) argument. □

C.2 Proof of Theorems 4 and 5

We first consider the more practical setting where μ is estimated from data, and prove Theorem 5 using the general upper and lower bounds on the conditional power of the `PairSwap-ICI` test from Theorem C.12. Theorem 4 will then follow as a special case of Theorem 5.

Before turning to the proofs, we observe from Lemma 3 that under the model class (16) and with the linear kernel $\psi(x, x') = x - x'$, the oracle weights in (13) satisfy $w^*_\ell \propto \big(\mu(Y_{i_\ell}, Z_{i_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell})\big)_+$ for $\ell \in [L]$. Since the p-value of our test is independent of the scales of weights, we may take $\mathbf{w}^* = \Delta\mu^+(\mathbf{Y}, \mathbf{Z})$. When we do not have access to oracle knowledge of μ, we assume that we are able to learn estimates $\hat{\mu}$ and $\hat{\sigma}$ of μ and σ respectively from a prior dataset D_{prior}, independent of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$. With these estimates in place, the plug-in weights satisfy $\hat{w}_\ell \propto \big(\hat{\mu}(Y_{i_\ell}, Z_{i_\ell}) - \hat{\mu}(Y_{j_\ell}, Z_{j_\ell})\big)_+$, i.e., we may take $\hat{\mathbf{w}} = \Delta\hat{\mu}^+(\mathbf{Y}, \mathbf{Z})$.

Given $(\mathbf{Y}, \mathbf{Z}) \in \mathbb{R}^n \times \mathbb{R}^n$, we define

$$\hat{\mu}_{\text{ISO}}(\mathbf{Y}, \mathbf{Z}) = \operatorname*{argmin}_{g \in \mathcal{C}_{\text{ISO}}} \|\mu(\mathbf{Y}, \mathbf{Z}) - g(\mathbf{Z})\|_2, \tag{26}$$

the empirical isotonic Euclidean projection of μ onto $\mathcal{C}_{\text{ISO}}$.

C.2.1 Proof of Theorem 5

Our proof is split into three steps. First we establish a key property of the oracle matching $M^\star$ and use this to deduce relative error properties of $\hat{\mu}$. These in turn enable us to show that Assumptions **A1** and **A2** of Theorem C.12 are satisfied by $\hat{M}$ and weight vector $\mathbf{w} = \Delta\hat{\mu}^+(\mathbf{Y}, \mathbf{Z})$. Finally we apply part *(ii)* of Theorem C.12 to conclude the proof.

We write

$$\begin{aligned}
\mathrm{Err}_2(\hat{\mu}, \mu) &= \frac{\|\Delta\hat{\mu}(\mathbf{Y}, \mathbf{Z}) - \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_2}{\|\Delta\mu^+(\mathbf{Y}, \mathbf{Z})\|_2}, \\
\mathrm{Err}_\infty(\hat{\mu}, \mu) &= \frac{\|\Delta\hat{\mu}(\mathbf{Y}, \mathbf{Z}) - \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_\infty^4}{\|\Delta\mu^+(\mathbf{Y}, \mathbf{Z})\|_2}.
\end{aligned}$$

Step 1: a key property of oracle matching and its implications. Our notation $\Delta\mu(\mathbf{Y},\mathbf{Z})$ and $\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z})$ suppresses the dependence of the matching M for which it is computed. Since we will need both matchings $M^\star$ and $\hat{M}$ in this step, we make the dependence on the matching explicit by writing $\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z};M)$, where $M \in \mathcal{M}_n(\mathbf{Z})$.

By Lemma C.13, Theorem C.18 and the definition of $M^* \in \mathcal{M}_n(\mathbf{Z})$ from (14), we have

$$\widehat{\mathrm{ISS}}_n \le \|\Delta\mu^+(\mathbf{Y},\mathbf{Z};M^\star)\|_2 = \sup_{M\in\mathcal{M}_n(\mathbf{Z})} \|\Delta\mu^+(\mathbf{Y},\mathbf{Z};M)\|_2 \le \sqrt{2}\,\widehat{\mathrm{ISS}}_n. \tag{27}$$

Now, for any $M \in \mathcal{M}_n(\mathbf{Z})$,

$$\begin{aligned}\big|\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z};M)\|_2 - \|\Delta\mu^+(\mathbf{Y},\mathbf{Z};M)\|_2\big| &\le \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z};M) - \Delta\mu^+(\mathbf{Y},\mathbf{Z};M)\|_2 \\ &\le \|\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z};M) - \Delta\mu(\mathbf{Y},\mathbf{Z};M)\|_2 \\ &\le \sqrt{2}\|\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \mu(\mathbf{Y},\mathbf{Z})\|_2 = \mathrm{o}_P(\widehat{\mathrm{ISS}}_n), \end{aligned}\tag{28}$$

where the final step is from the hypothesis in the statement of the result. Further by (15),

$$\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z};\hat{M})\|_2 \ge \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z};M^\star)\|_2 \ge \|\Delta\mu^+(\mathbf{Y},\mathbf{Z};M^\star)\|_2 - \mathrm{o}_P(\widehat{\mathrm{ISS}}_n) \ge \widehat{\mathrm{ISS}}_n\big(1-\mathrm{o}_P(1)\big). \tag{29}$$

Similarly, by (28),

$$\|\Delta\mu^+(\mathbf{Y},\mathbf{Z};\hat{M})\|_2 \ge \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z};\hat{M})\|_2 - \mathrm{o}_P(\widehat{\mathrm{ISS}}_n) \ge \widehat{\mathrm{ISS}}_n\big(1-\mathrm{o}_P(1)\big). \tag{30}$$

We deduce that for plug-in matching,

$$\max\left\{\frac{\|\mu(\mathbf{Y},\mathbf{Z})\|_\infty \vee \|\mu(\mathbf{Y},\mathbf{Z})\|_\infty^4}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2},\ \mathrm{Err}_2(\hat{\mu},\mu),\ \mathrm{Err}_\infty(\hat{\mu},\mu)\right\} = \mathrm{o}_P(1). \tag{31}$$

Step 2: establishing Assumptions A1 and A2. Henceforth we work with the matching $\hat{M}$. We have by (27) that

$$\begin{aligned}&\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2 \\ &\qquad\le \|\Delta\mu^+(\mathbf{Y},\mathbf{Z}) \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2 + \big\|\big(\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\big)\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\big\|_2 \\ &\qquad\le \|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_4^2 + \|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty \cdot \|\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2 \\ &\qquad\le 2\|\mu(\mathbf{Y},\mathbf{Z})\|_\infty\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 \cdot \big(1 + \mathrm{Err}_2(\hat{\mu},\mu)\big) = \mathrm{o}_P\big(\widehat{\mathrm{ISS}}_n^2\big).\end{aligned}$$

Moreover, by the Cauchy–Schwarz inequality, (27) and (28),

$$\begin{aligned}\big|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})^T\Delta\mu(\mathbf{Y},\mathbf{Z}) - \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})\|_2^2\big| &= \big|\big(\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \Delta\mu(\mathbf{Y},\mathbf{Z})\big)^T\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})\big| \\ &\le \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})\|_2 \cdot \|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 \cdot \mathrm{Err}_2(\hat{\mu},\mu) \\ &= \mathrm{o}_P\big(\widehat{\mathrm{ISS}}_n^2\big).\end{aligned}\tag{32}$$

Hence, by (31), Assumption **A2** is satisfied by $\hat{M}$. Moreover, (31) also yields that

$$\begin{aligned}\frac{\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})\|_2} &\le \frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty + \|\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 - \|\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2} \\ &\le \frac{2\big(\|\mu(\mathbf{Y},\mathbf{Z})\|_\infty/\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2\big) + \mathrm{Err}_2(\hat{\mu},\mu)}{1 - \mathrm{Err}_2(\hat{\mu},\mu)} = \mathrm{o}_P(1).\end{aligned}$$

Now, we focus on the other term that appears in Assumption **A1**. By (31) and (27),

$$\begin{aligned}
&\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2^{2/3}\\
&\le \Big(\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\circ\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2+\|\big(\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})-\Delta\mu^+(\mathbf{Y},\mathbf{Z})\big)\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2\Big)^{2/3}\\
&\le \Big(\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_4^2+\mathrm{Err}_2(\hat{\mu},\mu)\cdot\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2\cdot\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty\Big)^{2/3}\\
&\le \|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{2/3}\cdot\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^{2/3}\cdot\big(1+\mathrm{o}_P(1)\big)\\
&\le \|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{2/3}\cdot 2^{1/3}\widehat{\mathrm{ISS}}_n^{2/3}\cdot\big(1+\mathrm{o}_P(1)\big).
\end{aligned}$$

Moreover, again by (31),

$$\begin{aligned}
&\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{1/3}\\
&\qquad\le \|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty^{2/3}+\|\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z})-\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{1/3}\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{1/3}\\
&\qquad\le \frac{\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2^{1/3}}{\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{2/3}}\left(\frac{2^{4/3}\|\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{4/3}}{\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2^{1/3}}+\mathrm{Err}_\infty^{1/12}(\hat{\mu},\mu)\frac{2\|\mu(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^{1/4}}\right)\\
&\qquad= \mathrm{o}_P\left(\frac{\widehat{\mathrm{ISS}}_n^{1/3}}{\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{2/3}}\right).
\end{aligned}$$

Hence, by (29) we see that Assumption **A1** is satisfied by plug-in matching.

Step 3: applying Theorem C.12. By (28), (29) and (32),

$$\left|\frac{\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})\|_2}-\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2\right|=\mathrm{o}_P(\widehat{\mathrm{ISS}}_n).$$

Hence, by (27) and (30),

$$\widehat{\mathrm{ISS}}_n\big(1-\mathrm{o}_P(1)\big)\le\frac{\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})\|_2}\le\sqrt{2}\cdot\widehat{\mathrm{ISS}}_n\big(1+\mathrm{o}_P(1)\big).$$

Finally the result follows from part *(ii)* of Theorem C.12. □

Lemma C.13. *For any $M\in\mathcal{M}_n(\mathbf{Z})$, we have $\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2\le\sqrt{2}\,\widehat{\mathrm{ISS}}_n$.*

Proof. By definition, $\widehat{\mathrm{ISS}}_n=\|\mu(\mathbf{Y},\mathbf{Z})-\hat{\mu}_{\mathrm{ISO}}(\mathbf{Z})\|_2$ where $\hat{\mu}_{\mathrm{ISO}}$ is as defined in (26). Recall that $\big(\Delta\mu^+(\mathbf{Y},\mathbf{Z})\big)_\ell=\big(\mu(Y_{i_\ell},Z_{i_\ell})-\mu(Y_{j_\ell},Z_{j_\ell})\big)_+$ for $\ell\in[L]$. We claim that

$$\Big(\mu\big(Y_{i_\ell},Z_{i_\ell}\big)-\mu\big(Y_{j_\ell},Z_{j_\ell}\big)\Big)_+\le\big|\mu\big(Y_{i_\ell},Z_{i_\ell}\big)-\widehat{\mu}_{\mathrm{ISO}}\big(Z_{i_\ell}\big)\big|+\big|\widehat{\mu}_{\mathrm{ISO}}\big(Z_{j_\ell}\big)-\mu\big(Y_{j_\ell},Z_{j_\ell}\big)\big|\qquad(33)$$

for every $\ell\in[L]$. To see this, we may assume that $\mu\big(Y_{i_\ell},Z_{i_\ell}\big)>\mu\big(Y_{j_\ell},Z_{j_\ell}\big)$ since otherwise the left-hand side is zero. Now, if $\hat{\mu}_{\mathrm{ISO}}\big(Z_{i_\ell}\big)\ge\mu\big(Y_{j_\ell},Z_{j_\ell}\big)$, then

$$\begin{aligned}
\mu\big(Y_{i_\ell},Z_{i_\ell}\big)-\mu\big(Y_{j_\ell},Z_{j_\ell}\big)&\le\big|\mu\big(Y_{i_\ell},Z_{i_\ell}\big)-\widehat{\mu}_{\mathrm{ISO}}\big(Z_{i_\ell}\big)\big|+\widehat{\mu}_{\mathrm{ISO}}\big(Z_{i_\ell}\big)-\mu\big(Y_{j_\ell},Z_{j_\ell}\big)\\
&\le\big|\mu\big(Y_{i_\ell},Z_{i_\ell}\big)-\widehat{\mu}_{\mathrm{ISO}}\big(Z_{i_\ell}\big)\big|+\widehat{\mu}_{\mathrm{ISO}}\big(Z_{j_\ell}\big)-\mu\big(Y_{j_\ell},Z_{j_\ell}\big).
\end{aligned}$$

Otherwise, if $\hat{\mu}_{\mathrm{ISO}}(Z_{i_\ell}) < \mu(Y_{j_\ell}, Z_{j_\ell})$, then

$$\begin{aligned}\mu(Y_{i_\ell}, Z_{i_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell}) &< \mu(Y_{i_\ell}, Z_{i_\ell}) - \widehat{\mu}_{\mathrm{ISO}}(Z_{i_\ell}) \\ &\leq \mu(Y_{i_\ell}, Z_{i_\ell}) - \widehat{\mu}_{\mathrm{ISO}}(Z_{i_\ell}) + \big|\widehat{\mu}_{\mathrm{ISO}}(Z_{j_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell})\big|.\end{aligned}$$

This proves the claim (33), and it follows that

$$\begin{aligned}\|\Delta\mu^+(\mathbf{Y}, \mathbf{Z})\|_2^2 &\leq 2\sum_{\ell=1}^{L}\big\{\mu(Y_{i_\ell}, Z_{i_\ell}) - \widehat{\mu}_{\mathrm{ISO}}(Z_{i_\ell})\big\}^2 + 2\sum_{\ell=1}^{L}\big\{\mu(Y_{j_\ell}, Z_{j_\ell}) - \widehat{\mu}_{\mathrm{ISO}}(Z_{j_\ell})\big\}^2 \\ &\leq 2\sum_{i=1}^{n}\big\{\mu(Y_i, Z_i) - \widehat{\mu}_{\mathrm{ISO}}(Z_i)\big\}^2 = 2\|\mathrm{Res}_n(\mathbf{Y}, \mathbf{Z})\|_2^2,\end{aligned}$$

which proves the result. □

C.2.2 Proof of Theorem 4

Given access to oracle knowledge of μ, we observe that $\hat{\mu} = \mu$ satisfies the assumptions of Theorem 5. Hence, Theorem 4 follows as an immediate corollary of Theorem 5. □

C.3 Proof of Theorem 10

Let us define π as any permutation of $[n]$ for which

$$Z_{\pi(1)} \leq Z_{\pi(2)} \leq \ldots \leq Z_{\pi(n)},$$

and note that the collection of matched pairs for neighbour matching is given by $M := \big\{\big(\pi(2\ell-1), \pi(2\ell)\big)\big\}_{\ell \in \lfloor n/2 \rfloor}$. By part *(ii)* of Theorem C.12, with $\mathbf{w} = \Delta\mathbf{Y}^+$ the dominant term in the upper and lower bound on conditional power reduces to

$$\frac{\Delta\mathbf{Y}^{+T}\Delta\mu(\mathbf{Y}, \mathbf{Z})}{\rho_2\|\Delta\mathbf{Y}^+\|_2}.$$

Since $\rho_2 = \sqrt{2}\sigma$, by the same result it suffices to prove a concentration of the dominant term as

$$\frac{\Delta\mathbf{Y}^{+T}\Delta\mu(\mathbf{Y}, \mathbf{Z})}{\|\Delta\mathbf{Y}^+\|_2} = \sqrt{n/2}\cdot\beta_n\cdot\big\{\mathbb{E}\left[\mathrm{Var}\left(Y \mid Z\right)\right]\big\}^{1/2} + \mathrm{o}_P(1),$$

and that the Assumptions A1 and A2 are satisfied.

Step 1: concentration of the dominant term. Under the partially linear Gaussian models (20),

$$\Delta\mathbf{Y}^{+T}\Delta\mu(\mathbf{Y}, \mathbf{Z}) = \Delta\mathbf{Y}^{+T}\Delta\mu_0(\mathbf{Z}) + \beta_n\|\Delta\mathbf{Y}^+\|_2^2,$$

where the first term is negative since $\mu_0 \in \mathcal{C}_{\mathrm{ISO}}$. In fact, we can say more. Firstly,

$$\big|\Delta\mathbf{Y}^{+T}\Delta\mu_0(\mathbf{Z})\big| \leq \|\Delta\mathbf{Y}^+ \circ \Delta\mu_0(\mathbf{Z})\|_1 \leq \|\Delta\mathbf{Y}^+\|_\infty\|\Delta\mu_0(\mathbf{Z})\|_1.$$

Since $\mu_0 \in \mathcal{C}_{\text{ISO}}$, we further have that

$$\|\Delta\mu_0(\mathbf{Z})\|_1 = \sum_{\ell=1}^{\lfloor n/2 \rfloor} \big(\mu_0(Z_{\pi(2\ell)}) - \mu_0(Z_{\pi(2\ell-1)})\big) \leq \sum_{i=2}^{n} \big(\mu_0(Z_{\pi(i)}) - \mu_0(Z_{\pi(i-1)})\big)$$
$$= \mu_0(Z_{\pi(n)}) - \mu_0(Z_{\pi(1)}) \leq 2\|\mu_0(\mathbf{Z})\|_\infty. \quad (34)$$

Moreover by Lemma C.20, we have that

$$\Big|\frac{1}{n/2}\|\Delta\mathbf{Y}^+\|_2^2 - \mathbb{E}\left[\mathrm{Var}\left(Y \mid Z\right)\right]\Big| = \mathrm{o}_P(1),$$

which also implies that $\|\Delta\mathbf{Y}^+\|_2 \geq \sqrt{n/2} \cdot \big\{\mathbb{E}\left[\mathrm{Var}\left(Y \mid Z\right)\right]\big\}^{1/2} - \mathrm{o}_P(\sqrt{n})$. Since $Y \in [-1, 1]$ and $\mu_0(\mathbf{Z})$ is sub-Gaussian, it follows that

$$\begin{aligned}
&\Big|\frac{\Delta\mathbf{Y}^{+T}\Delta\mu(\mathbf{Y},\mathbf{Z})}{\|\Delta\mathbf{Y}^+\|_2} - \sqrt{n/2} \cdot \beta_n \cdot \big\{\mathbb{E}\left[\mathrm{Var}\left(Y \mid Z\right)\right]\big\}^{1/2}\Big| \\
&\leq \Big|\frac{\Delta\mathbf{Y}^{+T}\Delta\mu(\mathbf{Y},\mathbf{Z})}{\|\Delta\mathbf{Y}^+\|_2} - \beta_n\|\Delta\mathbf{Y}^+\|_2\Big| + \Big|\beta_n\|\Delta\mathbf{Y}^+\|_2 - \sqrt{n/2} \cdot \beta_n \cdot \big\{\mathbb{E}\left[\mathrm{Var}\left(Y \mid Z\right)\right]\big\}^{1/2}\Big| \\
&\leq 4\,\frac{\|\mu_0(\mathbf{Z})\|_\infty}{\|\Delta\mathbf{Y}^+\|_2} + \mathrm{o}_P(\beta_n\sqrt{n}) = \mathrm{o}_P(1),
\end{aligned}$$

where the last equality follows by recalling that $\mu_0(Z)$ is sub-Gaussian and $\beta_n \gtrsim n^{-1/2}$.

Step 2: establishing Assumptions A1 and A2. Since $Y \in [-1, 1]$,

$$\|\Delta\mathbf{Y}^+ \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty \leq \|\Delta\mathbf{Y}^+ \circ \Delta\mu_0(\mathbf{Z})\|_\infty + \beta_n\|\Delta\mathbf{Y}^+\|_\infty^2 \leq 4\big(\|\mu_0(\mathbf{Z})\|_\infty + \beta_n\big).$$

Moreover, by an argument, similar to that in (34), we have that

$$\begin{aligned}
\|\Delta\mathbf{Y}^+ \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2 &\leq \|\Delta\mathbf{Y}^+ \circ \Delta\mu_0(\mathbf{Z})\|_2 + \beta_n\|\Delta\mathbf{Y}^+\|_4^2 \\
&\leq \|\Delta\mathbf{Y}^+\|_\infty\|\Delta\mu_0(\mathbf{Z})\|_\infty\|\Delta\mu_0(\mathbf{Z})\|_1 + \beta_n\|\Delta\mathbf{Y}^+\|_\infty\|\Delta\mathbf{Y}^+\|_2 \\
&\leq \|\Delta\mathbf{Y}^+\|_\infty\Big(\|\Delta\mu_0(\mathbf{Z})\|_\infty \sum_{i=2}^{n}\big(\mu_0(Z_{\pi(i)}) - \mu_0(Z_{\pi(i-1)})\big) + \beta_n\|\Delta\mathbf{Y}^+\|_2\Big) \\
&\leq 2\big(4\|\mu_0(\mathbf{Z})\|_\infty^2 + \beta_n\|\Delta\mathbf{Y}^+\|_2\big).
\end{aligned}$$

Since $\mu_0(Z)$ is sub-Gaussian and that $\|\Delta\mathbf{Y}^+\|_2 \geq \sqrt{n/2} \cdot \big\{\mathbb{E}\left[\mathrm{Var}\left(Y \mid Z\right)\right]\big\}^{1/2} - \mathrm{o}_P(\sqrt{n})$, Assumption **A1** is satisfied by the neighbour matching. Similarly, by (34) and recalling that $\beta_n \gtrsim 1/\sqrt{n}$, it follows that

$$\frac{\|\Delta\mathbf{Y}^+ \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2}{\Delta\mathbf{Y}^{+T}\Delta\mu(\mathbf{Y},\mathbf{Z})} \leq \frac{2\big(4\|\mu_0(\mathbf{Z})\|_\infty^2 + \beta_n\|\Delta\mathbf{Y}^+\|_2\big)}{-4\|\mu_0(\mathbf{Z})\|_\infty + \beta_n\|\Delta\mathbf{Y}^+\|_2^2} = \mathrm{o}_P(1).$$

Hence Assumption **A2** holds too, which completes the proof. □

C.4 Power of cross bin matching

In this section, we study power of the `PairSwap-ICI` test for cross-bin matching. We will start with stating and proving a concentration result for the conditional power of cross-bin matching in Theorem C.14. This subsequently leads to the proof of Theorem 11 in Appendix C.4.2. Before diving into the main results, we establish some necessary notations.

Definition 1. *For any distribution P, we define the deviation of P as*

$$\mathrm{Dev}(P) := \mathbb{E}_{q\sim \mathrm{Unif}[0,1]}\left[(F^{-1}(1-q) - F^{-1}(q))^2\right], \tag{35}$$

where F^{-1} is the generalized inverse CDF of P.

C.4.1 A general version of Theorem 11: asymptotic conditional power of cross bin matching

Theorem C.14. *Under the model class (20) suppose $\beta_n \gtrsim 1/\sqrt{n}$. Additionally, also assume that*

(i) $\mu_0(Z)$ is a sub-Gaussian random variable, and the function μ_0 satisfies the Lipschitz property

(36)

(ii) there exists a constant $L_W \geq 0$ such that for any $z_1, z_2 \in \mathbb{R}$,

$$\mathrm{d}_{\mathrm{W}_1}\left(P_{Y|Z}(\cdot \mid z_1), P_{Y|Z}(\cdot \mid z_2)\right) \leq L_W \left|P_Z(z_1) - P_Z(z_2)\right|, \tag{37}$$

where $\mathrm{d}_{\mathrm{W}_1}(\cdot,\cdot)$ is the 1-Wasserstein distance.

Then, the conditional power of cross-bin matching (Algorithm 2), implemented with kernel $\psi(x,x') = x - x'$ and weights $w_\ell = \max\{(Y_{i_\ell} - Y_{j_\ell}, 0)\}$, satisfies

$$\left|\mathbb{P}\left\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\right\} - \Phi\left(\sqrt{n/2}\cdot(\beta_n/\sigma)\cdot\left(\mathbb{E}\left[\mathrm{Dev}(P_{Y|Z})\right]\right)^{1/2} - \bar{\Phi}^{-1}(\alpha)\right)\right| = \mathrm{o}_P(1).$$

Proof. The proof follows the same key steps as from the Appendix C.3. By part *(ii)* of Theorem C.12 and with replacing $\mathbf{w} = \Delta\mathbf{Y}^+$, the dominant term in the upper and lower bound on conditional power reduces to

$$\frac{\Delta\mathbf{Y}^{+T}\Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\Delta\mathbf{Y}^+\|_2},$$

Since $\rho_2 = \sqrt{2}\sigma$, it suffices to prove a concentration of the dominant term as

$$\frac{\Delta\mathbf{Y}^{+T}\Delta\mu(\mathbf{Y},\mathbf{Z})}{\|\Delta\mathbf{Y}^+\|_2} = \sqrt{n/2}\cdot\beta_n\cdot\left\{\mathbb{E}\left[\mathrm{Dev}(P_{Y|Z})\right]\right\}^{1/2} + \mathrm{o}_P(1),$$

and that the Assumptions **A1** and **A2** are satisfied.

Step 1: concentration of the dominant term. Under the partially linear Gaussian model (20),

$$\Delta\mathbf{Y}^{+T}\Delta\mu(\mathbf{Y},\mathbf{Z}) = \Delta\mathbf{Y}^{+T}\Delta\mu_0(\mathbf{Z}) + \beta\|\Delta\mathbf{Y}^+\|_2^2,$$

where for the first term, it holds that

$$\left|\Delta\mathbf{Y}^{+T}\Delta\mu_0(\mathbf{Z})\right| = \sum_{\ell=1}^{L_n} |Y_{i_{\ell,n}} - Y_{j_{\ell,n}}| \left|\mu_0(Z_{j_{\ell,n}}) - \mu_o(Z_{i_{\ell,n}})\right| \leq n\|\Delta\mathbf{Y}^+\|_\infty \|\Delta\mu_0(\mathbf{Z})\|_\infty. \quad (38)$$

Next, we analyze the large sample behaviour of $\|\Delta\mathbf{Y}^+\|_2^2$. Conditioned on $\mathbf{Z}$, $\|\Delta\mathbf{Y}^+\|_2^2$ is a sum of independent and uniformly bounded terms. Hence, by the law of large numbers,

$$\left|\frac{1}{n/2}\|\Delta\mathbf{Y}^+\|_2^2 - \mathbb{E}\left[\frac{1}{n/2}\|\Delta\mathbf{Y}^+\|_2^2 \;\middle|\; \mathbf{Z}\right]\right| = \mathrm{o}_P(1).$$

Finally, by Lemma C.21 we have that

$$\left|\frac{1}{n/2}\|\Delta\mathbf{Y}^+\|_2^2 - \mathbb{E}\left[\mathrm{Dev}(P_{Y|Z})\right]\right| = \mathrm{o}_P(1). \quad (39)$$

Since μ_0 is L_μ- Lipschitz, we note that $\|\Delta\mu_0(\mathbf{Z})\|_\infty = \mathrm{O}_P(1/\sqrt{n})$. Combining behaviour of both terms, we have that

$$\begin{aligned}\frac{\Delta\mathbf{Y}^{+T}\Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\Delta\mathbf{Y}^+\|_2} &= \frac{-\mathrm{O}_P(\sqrt{n}) + \beta\|\Delta\mathbf{Y}^+\|_2^2}{\rho_2\|\Delta\mathbf{Y}^+\|_2} = (\beta/\rho_2)\cdot\|\Delta\mathbf{Y}^+\|_2 - \mathrm{o}_P(1)\\ &= (\beta/\rho_2)\left(\sqrt{(n/2)}\cdot\sqrt{\mathbb{E}\left[\mathrm{Dev}(P_{Y|Z})\right]} + \mathrm{o}_P(\sqrt{n})\right) - \mathrm{o}_P(1).\end{aligned}$$

Since $\rho_2 = \sqrt{2}\sigma$, the result follows by Theorem C.12 as long as we can show that Assumption **A1** and **A2** are satisfied.

Step 2: establishing Assumptions A1 and A2. Since $Y \in [-1,1]$,

$$\|\Delta\mathbf{Y}^+ \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty \leq \|\Delta\mathbf{Y}^+ \circ \Delta\mu_0(\mathbf{Z})\|_\infty + \beta_n\|\Delta\mathbf{Y}^+\|_\infty^2 \leq 4\big(\|\mu_0(\mathbf{Z})\|_\infty + \beta_n\big).$$

Similarly, we also have that

$$\begin{aligned}\|\Delta\mathbf{Y}^+ \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2 &\leq \|\Delta\mathbf{Y}^+ \circ \Delta\mu_0(\mathbf{Z})\|_2 + \beta_n\|\Delta\mathbf{Y}^+\|_4^2\\ &\leq \|\Delta\mathbf{Y}^+\|_2\big(\|\Delta\mu_0(\mathbf{Z})\|_\infty + \beta_n\|\Delta\mathbf{Y}^+\|_\infty\big)\\ &\leq 2\|\Delta\mathbf{Y}^+\|_2\big(\|\mu_0(\mathbf{Z})\|_\infty + \beta_n\big).\end{aligned}$$

Since $\mu_0(Z)$ is sub-Gaussian and that $\|\Delta\mathbf{Y}^+\|_2 \geq \sqrt{n/2}\cdot\left\{\mathbb{E}\left[\mathrm{Dev}(P_{Y|Z})\right]\right\}^{1/2} - \mathrm{o}_P(\sqrt{n})$, Assumption **A1** is satisfied by the cross-bin matching. Moreover, by (38) and recalling that $\beta_n \gtrsim 1/\sqrt{n}$, it follows that

$$\frac{\|\Delta\mathbf{Y}^+ \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2}{\Delta\mathbf{Y}^{+T}\Delta\mu(\mathbf{Y},\mathbf{Z})} \leq \frac{2\|\Delta\mathbf{Y}^+\|_2\big(\|\mu_0(\mathbf{Z})\|_\infty + \beta_n\big)}{-n\|\Delta\mathbf{Y}^+\|_\infty\|\Delta\mu_0(\mathbf{Z})\|_\infty + \beta_n\|\Delta\mathbf{Y}^+\|_2^2} = \mathrm{o}_P(1).$$

Hence Assumption **A2** holds too, which completes the proof. □

C.4.2 Proof of Theorem 11

Observe that for any $q \neq 0.5$,

$$(Q_P(1-q) - Q_P(0.5) \cdot (Q_P(0.5) - Q_P(q)) \geq 0.$$

Therefore, for any distribution P, we have that

$$\begin{aligned}
\mathrm{Dev}(P) &= \mathbb{E}_{q\sim\mathrm{Unif}[0,1]}\left[\left[Q_P(1-q) - Q_P(0.5) + Q_P(0.5) - Q_P(q)\right]^2\right] \\
&\geq \mathbb{E}_{q\sim\mathrm{Unif}[0,1]}\left[\left[Q_P(1-q) - Q_P(0.5)\right]^2 + \left[Q_P(0.5) - Q_P(q)\right]^2\right] \\
&= 2\,\mathbb{E}_{q\sim\mathrm{Unif}[0,1]}\left[(Q_P(q) - Q_P(0.5))^2\right] \\
&= 2\,\mathbb{E}_{X\sim P}\left[(X - \mathrm{Median}(P))^2\right] \geq 2\,\mathrm{Var}_{X\sim P}\left(X\right).
\end{aligned}$$

Hence, it follows that

$$\mathbb{E}\left[\mathrm{Dev}(P_{Y|Z})\right] \geq \sqrt{2}\,\mathbb{E}\left[\mathrm{Var}\left(Y \mid Z\right)\right]$$

The result now follows from Theorem C.14. □

C.4.3 Matching upper and lower bounds on conditional power

In (18), the asymptotic upper and lower bounds on conditional power for oracle matching match up to a factor of $\sqrt{2}$. In this section, we try to answer the natural question: can we close out this gap? We note that in Appendix C.2.1, this gap arises from (27), i.e.,

$$\widehat{\mathrm{ISS}}_n \leq \|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 \leq \sqrt{2}\,\widehat{\mathrm{ISS}}_n,$$

where $\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2$ is computed for the oracle matching.

Without additional assumptions on the model, it is unclear whether for oracle matching one can ensure $\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 \approx \sqrt{2}\,\widehat{\mathrm{ISS}}_n$. However, the proof of Lemma C.13 suggests that it possible if oracle matching satisfies

$$Z_{i_\ell} \approx Z_{j_\ell} \;\text{ so that }\; \widehat{\mu}_{\mathrm{ISO}}(Z_{i_\ell}) \approx \widehat{\mu}_{\mathrm{ISO}}(Z_{i_\ell}), \;\text{ and }\; \mu(Y_{i_\ell}, Z_{i_\ell}) \approx -\mu(Y_{j_\ell}, Z_{j_\ell}).$$

However, we need to assume symmetry of $\mu(Y,Z)$ conditional on Z to have $\mu(Y_{i_\ell}, Z_{i_\ell}) \approx -\mu(Y_{j_\ell}, Z_{j_\ell})$. We investigate this in greater details for the special case of partial linear Gaussian model (20). With this additional structure, and the restriction $Z_{i_\ell} \approx Z_{j_\ell}$, we only need to ensure $Y_{i_\ell} \approx -Y_{j_\ell}$.

We show that under symmetry of the conditional distribution of $P_{Y|Z}$, even cross-bin matching can attain the upper bound on asymptotic conditional power of oracle matching in (18) and thus, oracle matching will attain it too.

Corollary C.15. *Under the setting of Theorem C.14, if $P_{Y|Z}$ is symmetric almost surely, then the conditional power of cross-bin matching satisfies*

$$\left|\mathbb{P}\left\{p \leq \alpha \mid \mathbf{Y},\mathbf{Z}\right\} - \Phi\left(\sqrt{n}\beta_n\left\{\frac{\mathbb{E}\left[\mathrm{Var}\left(Y \mid Z\right)\right]}{\sigma^2}\right\}^{1/2} - \bar{\Phi}^{-1}(\alpha)\right)\right| = \mathrm{o}_P(1).$$

Proof. If $P_{Y|Z=z}$ is symmetric, then

$$Q_{P_{Y|Z=z}}(1-q) - Q_{P_{Y|Z=z}}(q) = 2\left(Q_{P_{Y|Z=z}}(1-q) - Q_{P_{Y|Z=z}}(0.5)\right).$$

Consequently,

$$\mathrm{Dev}(P_{Y|Z}) = 4\,\mathbb{E}_{q\sim\mathrm{Unif}[0,0.5]}\left[\left(Q_P(0.5) - Q_P(q)\right)^2\right] = 4\,\mathbb{E}_{X\sim P_{Y|Z}}\left[\left(X - \mathrm{Median}(P_{Y|Z})\right)^2\right]$$

almost surely. Moreover, the median and mean coincide and hence, $\mathbb{E}\left[\mathrm{Dev}(P_{Y|Z})\right] = 4\,\mathrm{Var}\left(Y \mid Z\right)$. The result now follows by Theorem C.14. □

Since by Lemma 9, under the model class (20), $\sqrt{n}\beta_n\left(\mathbb{E}\left[\mathrm{Var}\left(Y \mid Z\right)\right]\right)^{1/2}(1+\mathrm{o}_P(1)) \le \mathrm{ISS}_n$, the conditional power for cross-bin matching further satisfies

$$\left|\mathbb{P}\left\{p \le \alpha \mid \mathbf{Y}, \mathbf{Z}\right\} - \Phi\left(\frac{\widehat{\mathrm{ISS}}_n}{\sigma} - \bar{\Phi}^{-1}(\alpha)\right)\right| = \mathrm{o}_P(1),$$

as required.

C.5 Proof of propositions, corollaries and lemmas from Section 4

C.5.1 Proof of Proposition 6

Fix a distribution $Q_{X,Y,Z} \in H_0^{\mathrm{ICI}}$. For any test function

$$\phi : (\mathcal{X}\times\mathbb{R}\times\mathbb{R})^n \to [0,1] \text{ such that } \sup_{P\in H_0^{\mathrm{ICI}}} \mathbb{E}_P\left[\phi(\mathbf{X},\mathbf{Y},\mathbf{Z})\right] \le \alpha,$$

by definition of total-variation distance it holds that

$$\begin{aligned}\mathbb{E}_{P_{X,Y,Z}}\left[\phi(\mathbf{X},\mathbf{Y},\mathbf{Z})\right] &\le \mathbb{E}_{Q_{X,Y,Z}}\left[\phi(\mathbf{X},\mathbf{Y},\mathbf{Z})\right] + \mathrm{d_{TV}}\left(P^n_{X,Y,Z}, Q^n_{X,Y,Z}\right)\\ &\le \alpha + \mathrm{d_{TV}}\left(P^n_{X,Y,Z}, Q^n_{X,Y,Z}\right).\end{aligned}$$

The result now follows since the the last inequality holds for any $Q_{X,Y,Z} \in H_0^{\mathrm{ICI}}$.

C.5.2 Proof of Corollary 7

By Proposition 6, it is enough to argue that

$$\inf_{Q_{X,Y,Z}\in H_0^{\mathrm{ICI}}} \mathrm{d_{TV}}\left(P^n_{X,Y,Z}, Q^n_{X,Y,Z}\right) \le \frac{\mathrm{ISS}_n}{2\sigma}.$$

Consider any $\mu' \in \mathcal{C}_{\mathrm{ISO}}$. Similar to (16), define a model $Q_{X,Y,Z}$ as

$$\mathbf{X} = \mu'(\mathbf{Z}) + \boldsymbol{\zeta}, \quad (Y_1,Z_1),\cdots,(Y_n,Z_n) \overset{\mathrm{iid}}{\sim} P_{Y,Z}, \quad \zeta_1,\cdots,\zeta_n \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,\sigma^2).$$

By definition, $Q_{X,Y,Z} \in H_0^{\mathrm{ICI}}$ and

$$\begin{aligned}\mathrm{d_{TV}}(P_{X,Y,Z}, Q_{X,Y,Z}) &\le \mathbb{E}_{P_{Y,Z}}\left[\mathrm{d_{TV}}\left(\mathcal{N}\left(\mu(\mathbf{Y},\mathbf{Z}),\sigma^2 I_n\right), \mathcal{N}\left(\mu'(\mathbf{Z}),\sigma^2 I_n\right) \mid \mathbf{Y},\mathbf{Z}\right)\right]\\ &= \frac{\mathbb{E}_{P_{Y,Z}}\left[\|\mu(\mathbf{Y},\mathbf{Z}) - \mu'(\mathbf{Z})\|_2\right]}{2\sigma}\end{aligned}$$

Since this is true for any $\mu' \in \mathcal{C}_{\text{ISO}}$, then,

$$\mathrm{d_{TV}}(P_{X,Y,Z}, Q_{X,Y,Z}) \leq \inf_{\mu' \in \mathcal{C}_{\text{ISO}}} \frac{\mathbb{E}_{P_{Y,Z}}\left[\|\mu(\mathbf{Y},\mathbf{Z}) - \mu'(\mathbf{Z})\|_2\right]}{2\sigma} = \frac{\text{ISS}_n}{2\sigma}.$$

This proves the result. □

C.5.3 Proof of Corollary 8

Similar to the proof of Corollary 7, it is enough to argue that

$$\inf_{Q_{X,Y,Z} \in H_0^{\text{ICI}}} \mathrm{d_{TV}}\left(P^n_{X,Y,Z}, Q^n_{X,Y,Z}\right) \leq \left(\frac{1}{\epsilon(1-\epsilon)}\right)^{1/2} \text{ISS}_n.$$

Consider any $\mu' \in \mathcal{C}_{\text{ISO}}$, and define the model $Q_{X,Y,Z} \in H_0^{\text{ICI}}$ given by

$$X_i \sim \text{Ber}(\mu'(Z_i)), \quad (Y_1, Z_1), \cdots, (Y_n, Z_n) \overset{\text{iid}}{\sim} P_{Y,Z}.$$

We note that $\mathrm{d_{TV}}\left(P^n_{X,Y,Z}, Q^n_{X,Y,Z}\right)$ can be computed as

$$\mathbb{E}_{\mathbf{Y},\mathbf{Z}}\left[\mathrm{d_{TV}}\left(\text{Ber}(\mu(Y_1,Z_1)) \times \cdots \times \text{Ber}(\mu(Y_n,Z_n)), \text{Ber}(\mu'(Z_1)) \times \cdots \times \text{Ber}(\mu'(Z_n))\right)\right],$$

where the inner total variation term can be upper bounded using the Hellinger distance as

$$\begin{aligned}&\mathrm{d_{TV}}\left(\text{Ber}(\mu(Y_1,Z_1)) \times \cdots \times \text{Ber}(\mu(Y_n,Z_n)), \text{Ber}(\mu'(Z_1)) \times \cdots \times \text{Ber}(\mu'(Z_n))\right)\\ &\quad \leq \sqrt{2} \cdot H\left(\text{Ber}(\mu(Y_1,Z_1)) \times \cdots \times \text{Ber}(\mu(Y_n,Z_n)), \text{Ber}(\mu'(Z_1)) \times \cdots \times \text{Ber}(\mu'(Z_n))\right).\end{aligned}$$

Further, $H^2(P_1 \times \cdots \times P_k, Q_1 \times \cdots Q_k) \leq \sum_{i=1}^k H^2(P_i, Q_i)$ and thus by Lemma D.25, the inner total variation can be further upper bounded by

$$\left(\sum_{i=1}^n \frac{(\mu(Y_i,Z_i) - \mu'(Z_i))^2}{2\mu(Y_i,Z_i)(1-\mu(Y_i,Z_i))}\right)^{1/2}.$$

Finally, since $\mu(Y,Z) \in (\epsilon, 1-\epsilon)$ almost surely, the aforementioned term is upper bounded by $\frac{1}{\epsilon(1-\epsilon)} \cdot \|\mu(\mathbf{Y},\mathbf{Z}) - \mu'(\mathbf{Z})\|_2^2$. Since this is true for any $\mu' \in \mathcal{C}_{\text{ISO}}$, we have

$$\begin{aligned}&\mathrm{d_{TV}}\left(\text{Ber}(\mu(Y_1,Z_1)) \times \cdots \times \text{Ber}(\mu(Y_n,Z_n)), \text{Ber}(\mu'(Z_1)) \times \cdots \times \text{Ber}(\mu'(Z_n))\right)\\ &\qquad \leq \left(\inf_{\mu' \in \mathcal{C}_{\text{ISO}}} \frac{1}{\epsilon(1-\epsilon)} \cdot \|\mu(\mathbf{Y},\mathbf{Z}) - \mu'(\mathbf{Z})\|_2^2\right)^{1/2} = \left(\frac{1}{\epsilon(1-\epsilon)} \cdot \text{ISS}_n^2\right)^{1/2},\end{aligned}$$

which concludes the proof. □

C.5.4 Proof of Lemma 9

Firstly, under the model class (20) with $\mathbb{E}[Y] = 0$, we have that

$$\begin{aligned}\text{ISS}_n &\overset{(1)}{\leq} \mathbb{E}_{P_{Y,Z}}\left[\|\mu(\mathbf{Y},\mathbf{Z}) - \mu_0(\mathbf{Z})\|_2\right]\\ &= \mathbb{E}_{P_{Y,Z}}\left[\|\beta_n \mathbf{Y}\|_2\right]\\ &\overset{(2)}{\leq} \beta_n \left(\mathbb{E}_{P_{Y,Z}}\left[\sum_i Y_i^2\right]\right)^{1/2} = \sqrt{n}\beta_n \big(\text{Var}\,(Y)\big)^{1/2},\end{aligned}$$

where the inequality (1) holds since $\mu_0 \in \mathcal{C}_{\text{ISO}}$ and (2) holds by Jensen's inequality. Next, we note that under the model class (20),

$$\begin{aligned}
\text{ISS}_n &= \inf_{g \in \mathcal{C}_{\text{ISO}}} \mathbb{E}_{P_{Y,Z}} \left[\|\mu(\mathbf{Y}, \mathbf{Z}) - g(\mathbf{Z})\|_2 \right] \\
&\geq \inf_{g : \mathbb{R} \to \mathbb{R}} \mathbb{E}_{P_{Y,Z}} \left[\|\mu(\mathbf{Y}, \mathbf{Z}) - g(\mathbf{Z})\|_2 \right] \\
&= \mathbb{E}_{P_{Y,Z}} \left[\left\| \mu(\mathbf{Y}, \mathbf{Z}) - \mathbb{E}\left[\mu(\mathbf{Y}, \mathbf{Z}) \mid \mathbf{Z} \right] \right\|_2 \right] \\
&= \mathbb{E}_{P_{Y,Z}} \left[\left\| \beta_n \left(\mathbf{Y} - \mathbb{E}\left[\mathbf{Y} \mid \mathbf{Z} \right] \right) \right\|_2 \right].
\end{aligned}$$

Finally, since Y is a bounded random variable, we have

$$\frac{1}{n} \|\mathbf{Y} - \mathbb{E}\left[\mathbf{Y} \mid \mathbf{Z}\right]\|_2^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - E[Y \mid Z_i])^2 = \mathbb{E}\left[\text{Var}\left(Y \mid Z\right)\right] \left(1 + \text{o}_P(1)\right).$$

Thus, we have $\text{ISS}_n \geq \sqrt{n} \beta_n \left(\mathbb{E}\left[\text{Var}\left(Y \mid Z\right)\right] \right)^{1/2} \left(1 + \text{o}_P(1)\right)$. □

C.6 An oracle matching—Isotonic median matching

A key step in Appendix C.2.1 is demonstrating (27) for a matching in $\mathcal{M}_n(\mathbf{Z})$. The purpose of this section is to show that in the case where $\mathbb{R}$ is a totally ordered set, this is achievable via a method that we refer to as isotonic median matching (IMM). Since $\mathbb{R}$ is totally ordered, we write $\leq$ instead of $\leq$, and assume that $Z_1 \leq \cdots \leq Z_n$. Given $a_1, \ldots, a_m \in \mathbb{R}$, we define their median by

$$\text{Med}(a_1, \ldots, a_m) = \begin{cases} a_{(k+1)} & \text{if } m = 2k+1 \text{ for some } k \in \mathbb{N}_0, \\ \frac{a_{(k)} + a_{(k+1)}}{2} & \text{if } m = 2k \text{ for some } k \in \mathbb{N}, \end{cases}$$

where $a_{(1)} \leq \cdots \leq a_{(m)}$ denote the order statistics of $a_1, \ldots, a_m$. Now given $(\mathbf{Y}, \mathbf{Z})$ and any function $\gamma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we consider the function $\widetilde{\gamma}_{\text{ISO}} : \mathbb{R} \to \mathbb{R}$ where

$$\widetilde{\gamma}_{\text{ISO}}(z) := \max_{j \in [n] : Z_j \leq z} \min_{z_r \in \mathbb{R} : z_r \geq z} \text{Med}\Big(\big\{ \gamma(Y_\ell, Z_\ell) : Z_j \leq Z_\ell \leq z_r \big\} \Big). \tag{40}$$

We claim that $\widetilde{\gamma}_{\text{ISO}} \in \mathcal{C}_{\text{ISO}}$. To see this, observe that for $z \leq z'$,

$$\begin{aligned}
\widetilde{\gamma}_{\text{ISO}}(z) &= \max_{j \in [n] : Z_j \leq z} \min_{z_r \in \mathbb{R} : z_r \geq z} \text{Med}\Big(\big\{ \gamma(Y_\ell, Z_\ell) : Z_j \leq Z_\ell \leq z_r \big\} \Big) \\
&\leq \max_{j \in [n] : Z_j \leq z} \min_{z_r \in \mathbb{R} : z_r \geq z'} \text{Med}\Big(\big\{ \gamma(Y_\ell, Z_\ell) : Z_j \leq Z_\ell \leq z_r \big\} \Big) \\
&\leq \max_{j \in [n] : Z_j \leq z'} \min_{z_r \in \mathbb{R} : z_r \geq z'} \text{Med}\Big(\big\{ \gamma(Y_\ell, Z_\ell) : Z_j \leq Z_\ell \leq z_r \big\} \Big) = \widetilde{\gamma}_{\text{ISO}}(z')
\end{aligned}$$

which proves the claim. Further, we note that $\widetilde{\gamma}_{\text{ISO}}$ is a piecewise constant function in $\mathcal{C}_{\text{ISO}}$, i.e., there exist $m \in [n+1]$, integers $1 = n_1 < n_2 < n_3 < \cdots < n_m = n+1$ and real numbers $r_1 < \ldots < r_{m-1}$ such that

$$\widetilde{\gamma}_{\text{ISO}}(z) = \sum_{i=1}^{m-1} r_i \, \mathbb{1}\left\{ Z_{n_i} \leq z < Z_{n_{i+1}} \right\}, \tag{41}$$

with the convention that $Z_{n+1} = \infty$. In fact, we can say more:

Lemma C.16. *In the representation (41), we have* $r_i = \mathrm{Med}\big(\{\gamma(Y_\ell, Z_\ell)\}_{n_i \le \ell \le n_{i+1}-1}\big)$ *for* $i \in [m-1]$. *Further, for any* $i \in [m-1]$ *and any integer* $k \in \{n_i, n_i+1, \ldots, n_{i+1}-1\}$,

$$\mathrm{Med}\big(\{\gamma(Y_\ell, Z_\ell)\}_{n_i \le \ell \le k}\big) \ge r_i. \tag{42}$$

Proof. Fix $i \in [m-1]$ and let $z_0 \in \mathbb{R}$ be such that $Z_{n_i-1} < z_0 < Z_{n_i}$. For any $j \in [n]$ with $Z_j \le z_0$, we have that

$$\min_{z_r \in \mathbb{R}: z_r \ge z_0} \mathrm{Med}\big(\{\gamma(Y_\ell, Z_\ell) : Z_j \le Z_\ell \le z_r\}\big) = \min_{z_r \in \mathbb{R}: z_r \ge Z_{n_i}} \mathrm{Med}\big(\{\gamma(Y_\ell, Z_\ell) : Z_j \le Z_\ell \le z_r\}\big).$$

Since $\widetilde{\gamma}_{\mathrm{ISO}}(z_0) = r_{i-1} < r_i = \widetilde{\gamma}_{\mathrm{ISO}}(Z_{n_i})$, it follows that

$$\begin{aligned}\widetilde{\gamma}_{\mathrm{ISO}}(Z_{n_i}) &= \max\Big\{\min_{z_r \in \mathbb{R}: z_r \ge Z_{n_i}} \mathrm{Med}\Big(\{\gamma(Y_\ell, Z_\ell) : Z_{n_i} \le Z_\ell \le z_r\}\Big), \widetilde{\gamma}_{\mathrm{ISO}}(z_0)\Big\} \\ &= \min_{k \in [n]: k \ge n_i} \mathrm{Med}\Big(\{\gamma(Y_\ell, Z_\ell) : n_i \le \ell \le k\}\Big). \end{aligned} \tag{43}$$

Let $\hat{k} \in [n]$ denote the largest index k at which the minimum in (43) is attained. Since $\widetilde{\gamma}_{\mathrm{ISO}}(Z_{n_i}) = \widetilde{\gamma}_{\mathrm{ISO}}(Z_{n_i+1}) = \cdots = \widetilde{\gamma}_{\mathrm{ISO}}(Z_{n_{i+1}-1})$, we have $\hat{k} \ge n_{i+1} - 1$. Now, it suffices to show that

$$\mathrm{Med}\Big(\{\gamma(Y_\ell, Z_\ell) : n_i \le \ell \le \hat{k}\}\Big) = \mathrm{Med}\Big(\{\gamma(Y_\ell, Z_\ell) : n_i \le \ell \le n_{i+1} - 1\}\Big). \tag{44}$$

If $i = m-1$, then $\hat{k} = n_m - 1 = n$, so we may assume that $i \in [m-2]$. Suppose for a contradiction that (44) is not true, so $\hat{k} \ge n_{i+1}$. Now by (43), we have that

$$\begin{aligned}\widetilde{\gamma}_{\mathrm{ISO}}(Z_{n_{i+1}}) &= \min_{k \in [n]: k \ge n_{i+1}} \mathrm{Med}\Big(\{\gamma(Y_\ell, Z_\ell) : n_{i+1} \le \ell \le k\}\Big) \\ &\le \mathrm{Med}\Big(\{\gamma(Y_\ell, Z_\ell) : n_{i+1} \le \ell \le \hat{k}\}\Big)\end{aligned}$$

But since we have assumed that (44) is not true, we have

$$\mathrm{Med}\Big(\{\gamma(Y_\ell, Z_\ell) : n_i \le \ell \le \hat{k}\}\Big) < \mathrm{Med}\Big(\{\gamma(Y_\ell, Z_\ell) : n_i \le \ell \le n_{i+1} - 1\}\Big),$$

and thus by Lemma D.26, we further have

$$\widetilde{\gamma}_{\mathrm{ISO}}(Z_{n_{i+1}}) \le \mathrm{Med}\Big(\{\gamma(Y_\ell, Z_\ell) : n_i \le \ell \le \hat{k}\}\Big).$$

However by (41), we also have

$$\widetilde{\gamma}_{\mathrm{ISO}}(Z_{n_{i+1}}) > \widetilde{\gamma}_{\mathrm{ISO}}(Z_{n_i}) = \mathrm{Med}\Big(\{\gamma(Y_\ell, Z_\ell) : n_i \le \ell \le \hat{k}\}\Big),$$

which is a contradiction. This establishes (44) and completes the proof of the first part of the result.

For the final part, fix any $i \in [m-1]$ and $k_0 \in \{n_i, n_i+1, \ldots, n_{i+1}-1\}$. Then by (43),

$$\begin{aligned} r_i = \widetilde{\gamma}_{\mathrm{ISO}}(Z_{n_i}) &= \min_{k \in [n]: k \ge n_i} \mathrm{Med}\Big(\{\gamma(Y_\ell, Z_\ell) : n_i \le \ell \le k\}\Big) \\ &\le \mathrm{Med}\Big(\{\gamma(Y_\ell, Z_\ell) : n_i \le \ell \le k_0\}\Big), \end{aligned}$$

as required. □

We are now in a position to define the isotonic median matching. For $i \in [m-1]$, define

$$P_i := \big\{t \in \{n_i, n_i+1, \dots, n_{i+1}-1\} : \gamma(Y_t, Z_t) > r_i\big\},$$
$$N_i := \big\{t \in \{n_i, n_i+1, \dots, n_{i+1}-1\} : \gamma(Y_t, Z_t) < r_i\big\}.$$

Let us order the indices in P_i as $t^+_{i,1} < \cdots < t^+_{i,n_i^+}$ and the indices in N_i as $t^-_{i,1} < \cdots < t^-_{i,n_i^-}$. Defining $L_i := n_i^+ \wedge n_i^-$, we consider the collection of ordered pairs

$$\mathcal{C}_i := \big\{(t^+_{i,1}, t^-_{i,1}), \cdots, (t^+_{i,L_i}, t^-_{i,L_i})\big\}.$$

Finally, $\widetilde{M}(\gamma) := \bigcup_{i=1}^{m-1} \mathcal{C}_i$ defines the isotonic median matching.

Lemma C.17. *Suppose that the elements of $\{\gamma(Y_\ell, Z_\ell) : \ell \in [n]\}$ are all distinct. Then for any $i \in [m-1]$, we have $n_i^+ = n_i^-$ and $\gamma(Y_\ell, Z_\ell) = \widetilde{\gamma}_{\mathrm{ISO}}(Z_\ell)$ for all $\ell \in \{n_i, n_i+1, \dots, n_{i+1}-1\} \setminus \big(P_i \cup N_i\big)$. Moreover, $\widetilde{M}(\gamma) \in \mathcal{M}_n(\mathbf{Z})$.*

Proof. Fix $i \in [m-1]$. When $n_{i+1} - n_i$ is even, we have by the first part of Lemma C.16 that $n_i^+ = n_i^- = (n_{i+1}-n_i)/2$ and $\{n_i, n_i+1, \dots, n_{i+1}-1\} \setminus \big(P_i \cup N_i\big) = \emptyset$. On the other hand, when $n_{i+1}-n_i$ is odd, we have $n_i^+ = n_i^- = (n_{i+1}-n_i-1)/2$ and $\{n_i, n_i+1, \dots, n_{i+1}-1\} \setminus \big(P_i \cup N_i\big) =: \{\ell_0\}$ is a singleton set. Moreover, $\gamma(Y_{\ell_0}, Z_{\ell_0}) = \widetilde{\gamma}_{\mathrm{ISO}}(Z_{\ell_0})$ which proves the first two claims.

Since $Z_1 \le \cdots \le Z_n$, in order to prove that $\widetilde{M}(\gamma)$ is a valid matching, it suffices to show that $t^+_{i,\ell} < t^-_{i,\ell}$ for any $i \in [m-1]$ and $\ell \in [L_i]$. To see this, noting that $t^+_{i,\ell} \neq t^-_{i,\ell}$, suppose for a contradiction that $t^+_{i,\ell_0} > t^-_{i,\ell_0}$ for some $i \in [m-1]$ and some minimal $\ell_0 \in [L_i]$. Since the elements of $\{\gamma(Y_\ell, Z_\ell) : \ell \in [n]\}$ are all distinct, we have $|\{t \in P_i : t \le t^-_{i,\ell_0}\}| = \ell_0 - 1$ and $|\{t \in N_i : t \le t^-_{i,\ell_0}\}| = \ell_0$, so

$$\mathrm{Med}\big(\{\gamma(Y_\ell, Z_\ell)\}_{n_i \le \ell \le t^-_{i,\ell_0}}\big) < r_i,$$

which contradicts (42). □

The following key property of IMM ensures that when the `PairSwap-ICI` test is run with the oracle matching and oracle choice of weights from (13), it has valid Type I error control by Theorem 1 and moreover satisfies the power guarantees of Corollary **??**.

Theorem C.18. *Given $(\mathbf{Y}, \mathbf{Z}) \in \mathbb{R}^n \times \mathbb{R}^n$, there exists $\gamma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ for which, with the corresponding isotonic median matching $\widetilde{M}(\gamma) \in \mathcal{M}_n(\mathbf{Z})$, we have*

$$\|\Delta\mu^+(\mathbf{Y}, \mathbf{Z})\|_2 \ge \widehat{\mathrm{ISS}}_n.$$

Proof. Fix $\epsilon > 0$. We can find $\gamma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that the coordinates of $\gamma(\mathbf{Y}, \mathbf{Z})$ are all distinct and $\|\mu(\mathbf{Y}, \mathbf{Z}) - \gamma(\mathbf{Y}, \mathbf{Z})\|_2 < \epsilon$. Consider the IMM $\widetilde{M}(\gamma) := \{(i_\ell, j_\ell)\}_{\ell \in \tilde{L}}$, where $\widetilde{M}(\gamma) \in \mathcal{M}_n(\mathbf{Z})$ by Lemma C.17. Furthermore by Lemma C.17 and (41), we have for any

$\ell \in [\tilde{L}]$ that $\gamma(Y_{i_\ell}, Z_{i_\ell}) > \widetilde{\gamma}_{\mathrm{ISO}}(Z_{i_\ell}) = \widetilde{\gamma}_{\mathrm{ISO}}(Z_{j_\ell}) > \gamma(Y_{j_\ell}, Z_{j_\ell})$. Hence,

$$\begin{aligned}
\|\Delta\gamma^+(\mathbf{Y},\mathbf{Z})\|_2^2 &= \sum_{\ell=1}^{\tilde{L}} \big(\gamma(Y_{i_\ell}, Z_{i_\ell}) - \gamma(Y_{j_\ell}, Z_{j_\ell})\big)^2 \\
&= \sum_{\ell=1}^{\tilde{L}} \big(\gamma(Y_{i_\ell}, Z_{i_\ell}) - \widetilde{\gamma}_{\mathrm{ISO}}(Z_{i_\ell}) + \widetilde{\gamma}_{\mathrm{ISO}}(Z_{j_\ell}) - \gamma(Y_{j_\ell}, Z_{j_\ell})\big)^2 \\
&\geq \sum_{\ell=1}^{\tilde{L}} \big\{ \big(\gamma(Y_{i_\ell}, Z_{i_\ell}) - \widetilde{\gamma}_{\mathrm{ISO}}(Z_{i_\ell})\big)^2 + \big(\widetilde{\gamma}_{\mathrm{ISO}}(Z_{j_\ell}) - \gamma(Y_{j_\ell}, Z_{j_\ell})\big)^2 \big\} \\
&= \sum_{i=1}^{n} \big(\gamma(Y_i, Z_i) - \widetilde{\gamma}_{\mathrm{ISO}}(Z_i)\big)^2 = \|\gamma(\mathbf{Y},\mathbf{Z}) - \widetilde{\gamma}_{\mathrm{ISO}}(\mathbf{Z})\|_2^2,
\end{aligned}$$

where the penultimate equality holds because $\gamma(Y_i, Z_i) = \widetilde{\gamma}_{\mathrm{ISO}}(Z_i)$ for $i \in [n] \backslash \{i_1, j_1, \ldots, i_{\tilde{L}}, j_{\tilde{L}}\}$, by Lemma C.17. Next, we observe that

$$\begin{aligned}
\|\Delta\gamma^+(\mathbf{Y},\mathbf{Z})\|_2 - \|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 &\leq \|\Delta\gamma^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 \\
&\leq \|\Delta\gamma(\mathbf{Y},\mathbf{Z}) - \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2 \\
&\leq \sqrt{2}\|\gamma(\mathbf{Y},\mathbf{Z}) - \mu(\mathbf{Y},\mathbf{Z})\|_2 \leq \sqrt{2}\epsilon,
\end{aligned}$$

and that by (17),

$$\widehat{\mathrm{ISS}}_n \leq \|\mu(\mathbf{Y},\mathbf{Z}) - \widetilde{\gamma}_{\mathrm{ISO}}(\mathbf{Z})\|_2 \leq \|\gamma(\mathbf{Y},\mathbf{Z}) - \widetilde{\gamma}_{\mathrm{ISO}}(\mathbf{Z})\|_2 + \epsilon.$$

Therefore,

$$\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 + \sqrt{2}\epsilon \geq \|\Delta\gamma^+(\mathbf{Y},\mathbf{Z})\|_2 \geq \|\gamma(\mathbf{Y},\mathbf{Z}) - \widetilde{\gamma}_{\mathrm{ISO}}(\mathbf{Z})\|_2 \geq \widehat{\mathrm{ISS}}_n - \epsilon.$$

Since $\epsilon > 0$ was arbitrary, the result follows. □

C.7 Proof of lemmas from Appendices C.1—C.4

Lemma C.19. *In the setting of Theorem C.12, let*

$$\begin{aligned}
\Omega_1 &:= \big\{\|\mathbf{w} \circ \Delta\mathbf{X}\|_2^2 \leq \rho_2^2 \|\mathbf{w}\|_2^2 \cdot (1 + \epsilon_{1,\delta,U})\big\}, \\
\Omega_2 &:= \big\{\|\mathbf{w} \circ \Delta\mathbf{X}\|_2^2 \geq \rho_2^2 \|\mathbf{w}\|_2^2 \cdot \max\{1 + \epsilon_{1,\delta,L}, 0\}\big\}, \\
\Omega_3 &:= \big\{0.56\|\mathbf{w} \circ \Delta\mathbf{X}\|_3^3 \leq \|\mathbf{w} \circ \Delta\mathbf{X}\|_2^3 \cdot \epsilon_{2,\delta}\big\},
\end{aligned}$$

where $\epsilon_{1,\delta,U}, \epsilon_{1,\delta,L}, \epsilon_{2,\delta}$ are as defined in (24). Then for any $\delta \in [0,1]$, we have $\mathbb{P}\{\Omega_1 \mid \mathbf{Y},\mathbf{Z}\} \geq 1 - 2\delta$, $\mathbb{P}\{\Omega_2\}\mathbf{Y},\mathbf{Z} \mid \geq\} 1 - 2\delta$ and $\mathbb{P}\{\Omega_3 \mid \mathbf{Y},\mathbf{Z}\} \geq 1 - 2\delta$.

Proof. We have

$$\|\mathbf{w} \circ \Delta\mathbf{X}\|_2^2 = \|\mathbf{w} \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2^2 + \|\mathbf{w} \circ \Delta\boldsymbol{\zeta}\|_2^2 + 2 \cdot \big(\mathbf{w} \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\big)^T (\mathbf{w} \circ \Delta\boldsymbol{\zeta}).$$

Now, conditional on $(\mathbf{Y}, \mathbf{Z})$, both $\|\mathbf{w} \circ \Delta\boldsymbol{\zeta}\|_2^2$, and $(\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z}))^T(\mathbf{w} \circ \Delta\boldsymbol{\zeta})$ are weighted sums of independent and identically distributed random variables. Hence,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{w} \circ \Delta\boldsymbol{\zeta}\|_2^2 \;\middle|\; \mathbf{Y}, \mathbf{Z}\right] &= \rho_2^2 \sum_{\ell=1}^{L} w_\ell^2 = \rho_2^2 \|\mathbf{w}\|_2^2, \\
\mathbb{E}\left[\big(\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\big)^T (\mathbf{w} \circ \Delta\boldsymbol{\zeta}) \;\middle|\; \mathbf{Y}, \mathbf{Z}\right] &= 0 \\
\mathrm{Var}\left(\|\mathbf{w} \circ \Delta\boldsymbol{\zeta}\|_2^2 \;\middle|\; \mathbf{Y}, \mathbf{Z}\right) &= \sum_{\ell=1}^{L} w_\ell^4 \mathrm{Var}\left((\Delta_\ell\boldsymbol{\zeta})^2\right) \\
&\leq \sum_{\ell=1}^{L} w_\ell^4 \mathbb{E}\left[(\Delta_\ell\boldsymbol{\zeta})^4\right] = \rho_4^4 \|\mathbf{w}\|_4^4, \\
\mathrm{Var}\left(\big(\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\big)^T (\mathbf{w} \circ \Delta\boldsymbol{\zeta}) \;\middle|\; \mathbf{Y}, \mathbf{Z}\right) &= \sum_{\ell=1}^{L} w_\ell^4 \big(\Delta_\ell\mu(\mathbf{Y}, \mathbf{Z})\big)^2 \mathbb{E}\left[(\Delta_\ell\boldsymbol{\zeta})^2\right] \\
&= \rho_2^2 \|\mathbf{w}^2 \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_2^2.
\end{aligned}$$

By Chebychev's inequality, for any $\delta \in [0, \infty)$,

$$\begin{aligned}
\mathbb{P}\left\{\left|\|\mathbf{w} \circ \Delta\boldsymbol{\zeta}\|_2^2 - \rho_2^2\|\mathbf{w}\|_2^2\right| \geq \frac{\rho_4^2}{\delta^{1/2}} \|\mathbf{w}\|_4^2 \;\middle|\; \mathbf{Y}, \mathbf{Z}\right\} &\leq \delta, \\
\mathbb{P}\left\{\left|(\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z}))^T (\mathbf{w} \circ \Delta\boldsymbol{\zeta})\right| \geq \frac{\rho_2}{\delta^{1/2}} \|\mathbf{w}^2 \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_2 \;\middle|\; \mathbf{Y}, \mathbf{Z}\right\} &\leq \delta.
\end{aligned}$$

Hence,

$$\begin{aligned}
1 - 2\delta &\leq \mathbb{P}\Bigg\{\|\mathbf{w} \circ \Delta\mathbf{X}\|_2^2 \leq \|\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_2^2 + \rho_2^2\|\mathbf{w}\|_2^2 + \frac{\rho_4^2}{\delta^{1/2}}\|\mathbf{w}\|_4^2 \\
&\qquad\qquad + \frac{2\rho_2}{\delta^{1/2}}\|\mathbf{w}^2 \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_2 \;\middle|\; \mathbf{Y}, \mathbf{Z}\Bigg\} \\
&\leq \mathbb{P}\left\{\Omega_1 \mid \mathbf{Y}, \mathbf{Z}\right\},
\end{aligned}$$

where, in the final step, we have used the facts that $\|\mathbf{w}\|_4^2 \leq \|\mathbf{w}\|_\infty \cdot \|\mathbf{w}\|_2$ and $\|\mathbf{w}^2 \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_2 \leq \|\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_\infty \cdot \|\mathbf{w}\|_2$.

The lower bound on $\mathbb{P}\left\{\Omega_2 \mid \mathbf{Y}, \mathbf{Z}\right\}$ follows by a very similar argument.

For the final bound, observe that $\|\mathbf{w} \circ \Delta\boldsymbol{\zeta}\|_3^3$ is also a weighted sum of independent and identically distributed random variables, and has (conditional) mean and variance given by

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{w} \circ \Delta\boldsymbol{\zeta}\|_3^3 \;\middle|\; \mathbf{Y}, \mathbf{Z}\right] &= \sum_{\ell=1}^{L} |w_\ell|^3 \rho_3^3 = \rho_3^3 \|\mathbf{w}\|_3^3, \\
\mathrm{Var}\left(\|\mathbf{w} \circ \Delta\boldsymbol{\zeta}\|_3^3 \;\middle|\; \mathbf{Y}, \mathbf{Z}\right) &= \sum_{\ell=1}^{L} w_\ell^6 \mathrm{Var}\left((|\Delta_\ell\boldsymbol{\zeta}|)^3\right) \leq \sum_{\ell=1}^{L} w_\ell^6 \mathbb{E}\left[(\Delta_\ell\boldsymbol{\zeta})^6\right] = \|\Delta\mathbf{w}\|_6^6 \rho_6^6.
\end{aligned}$$

Thus, by the triangle inequality and Chebychev's inequality,

$$\begin{aligned}
1-\delta &\le \mathbb{P}\left\{ \|\mathbf{w}\circ\Delta\mathbf{X}\|_3 \le \|\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_3 + \left(\rho_3^3\|\mathbf{w}\|_3^3 + \frac{\rho_6^3}{\delta^{1/2}}\cdot\|\mathbf{w}\|_6^3\right)^{1/3} \;\middle|\; \mathbf{Y},\mathbf{Z}\right\} \\
&\le \mathbb{P}\Bigg\{ \|\mathbf{w}\circ\Delta\mathbf{X}\|_3 \le \rho_2\|\mathbf{w}\|_2 \Bigg\{ \left(\frac{\|\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2}{\rho_2\|\mathbf{w}\|_2}\right)^{2/3}\left(\frac{\|\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty}{\rho_2\|\mathbf{w}\|_2}\right)^{1/3} \\
&\qquad + \left(\frac{\rho_3^3}{\rho_2^3}\frac{\|\mathbf{w}\|_\infty}{\|\mathbf{w}\|_2} + \frac{\rho_6^3}{\rho_2^3\delta^{1/2}}\cdot\frac{\|\mathbf{w}\|_\infty^2}{\|\mathbf{w}\|_2^2}\right)^{1/3} \Bigg\} \;\Bigg|\; \mathbf{Y},\mathbf{Z}\Bigg\} \\
&\le \mathbb{P}\left\{\Omega_2\cap\Omega_3 \mid \mathbf{Y},\mathbf{Z}\right\} + \mathbb{P}\left\{\Omega_2^c \mid \mathbf{Y},\mathbf{Z}\right\} \le \mathbb{P}\left\{\Omega_3 \mid \mathbf{Y},\mathbf{Z}\right\} + 2\delta,
\end{aligned}$$

as required. □

Lemma C.20. *Suppose we have i.i.d. samples $(Y_1,Z_1),\ldots,(Y_n,Z_n)$ from some distribution $P_{Y,Z}\in[-1,1]\times\mathbb{R}$, and π is any permutation of $[n]$ such that $Z_{\pi(1)}\le Z_{\pi(2)}\le\ldots\le Z_{\pi(n)}$ are ordered. Then,*

$$\left|\frac{1}{n/2}\sum_{i=1}^{\lfloor n/2\rfloor}(Y_{\pi(2i-1)}-Y_{\pi(2i)})_+^2 - \mathbb{E}\left[\mathrm{Var}\left(Y \mid Z\right)\right]\right| = \mathrm{o}_P(1).$$

Proof. Consider any joint distribution $P_{Y,Z}$ on $[-1,1]\times\mathbb{R}$. We write P_Z and $P_{Y|Z}(\cdot\mid z)$ for the distributions of Z and Y given $Z=z$, and we write F_Z^{-1} and $F_{Y|Z}^{-1}(\cdot\mid z)$ to denote the generalized inverse for P_Z and $P_{Y|Z}(\cdot\mid z)$ respectively. Without loss of generality we can also assume that π is the identity permutation, i.e., $Z_1\le Z_2\le\ldots\le Z_n$ are ordered and Y_i is the Y value corresponding to Z_i for any $i\in[n]$.

Step 1: concentration, conditional on Z. Conditioned on $\mathbf{Z}=(Z_1,Z_2,\ldots,Z_n)$, $\{(Y_{2i-1}-Y_{2i})\}_{i\in\lfloor n/2\rfloor}$ is a collection of independent and uniformly bounded random variables. Thus, by weak law of large numbers,

$$\left|\frac{1}{n/2}\sum_{i=1}^{\lfloor n/2\rfloor}(Y_{2i-1}-Y_{2i})_+^2 - \frac{1}{n/2}\sum_{i=1}^{\lfloor n/2\rfloor}\mathbb{E}\left[(Y_{2i-1}-Y_{2i})_+^2 \mid \mathbf{Z}\right]\right| = \mathrm{o}_P(1).$$

Hence, it suffices to show that

$$\left|\frac{1}{n/2}\sum_{i=1}^{\lfloor n/2\rfloor}\mathbb{E}\left[(Y_{2i-1}-Y_{2i})_+^2 \mid \mathbf{Z}\right] - \mathbb{E}\left[\mathrm{Var}\left(Y \mid Z\right)\right]\right| = \mathrm{o}_P(1),$$

Step 2: constructing a coupling. Consider $\{(U_{2i-1},U_{2i})\}_{i\in\lfloor n/2\rfloor}\overset{\text{iid}}{\sim}\text{Uniform}[0,1]$. Without loss of generality, we may take

$$Y_{2i-1}=F_{Y|Z}^{-1}\big(U_{2i-1}\mid Z_{2i-1}\big),\quad Y_{2i}=F_{Y|Z}^{-1}\big(U_{2i}\mid Z_{2i}\big).$$

We also define for $i\in\lfloor n/2\rfloor$

$$Y'_{2i-1}=F_{Y|Z}^{-1}\big(U_{2i-1}\mid Z_{2i}\big),\quad Y'_{2i}=F_{Y|Z}^{-1}\big(U_{2i}\mid Z_{2i-1}\big).$$

Conditioned on $\mathbf{Z}$, $Y'_{2i-1} \stackrel{d}{=} Y_{2i}$ and $Y'_{2i} \stackrel{d}{=} Y_{2i-1}$. Since $t \to t_+^2$ is 4-Lipschitz on $[-2,2]$, it follows that

$$\begin{aligned}\Big|2(Y_{2i-1}-Y_{2i})_+^2 - \big[(Y'_{2i-1}-Y_{2i})_+^2 + (Y_{2i-1}-Y'_{2i})_+^2\big]\Big| \\ \le \big|(Y_{2i-1}-Y_{2i})_+^2 - (Y'_{2i-1}-Y_{2i})_+^2\big| + \big|(Y_{2i-1}-Y_{2i})_+^2 - (Y_{2i-1}-Y'_{2i})_+^2\big| \\ \le 4\big(|Y_{2i-1}-Y'_{2i-1}| + |Y_{2i}-Y'_{2i}|\big).\end{aligned}$$

Therefore, we have

$$\begin{aligned}\left|\frac{1}{n/2}\sum_{i=1}^{\lfloor n/2\rfloor}\mathbb{E}\left[(Y_{2i-1}-Y_{2i})_+^2 \;\middle|\; \mathbf{Z}\right] - \frac{1}{n}\sum_{i=1}^{\lfloor n/2\rfloor}\mathbb{E}\left[(Y'_{2i-1}-Y_{2i})_+^2 + (Y_{2i-1}-Y'_{2i})_+^2 \;\middle|\; \mathbf{Z}\right]\right| \\ \le \frac{4}{n}\sum_{i=1}^{n}\mathbb{E}\left[|Y_{2i-1}-Y'_{2i-1}| + |Y_{2i}-Y'_{2i}| \;\middle|\; \mathbf{Z}\right].\end{aligned}$$

Moreover, by construction

$$\mathbb{E}\left[|Y_{2i-1}-Y'_{2i-1}| \;\middle|\; \mathbf{Z}\right] = \mathbb{E}\left[|Y_{2i}-Y'_{2i}| \mid \mathbf{Z}\right] = \mathrm{d}_{\mathrm{W}_1}\big(P_{Y|Z}(\cdot \mid Z_{2i-1}), P_{Y|Z}(\cdot \mid Z_{2i})\big),$$

where $\mathrm{d}_{\mathrm{W}_1}(\cdot,\cdot)$ denotes the 1-Wasserstein distance. We also write $\sigma_z^2 := \mathrm{Var}\left(Y \mid Z = z\right)$ and note that

$$\mathbb{E}\left[(Y_{2i-1}-Y'_{2i})_+^2 \;\middle|\; \mathbf{Z}\right] = \sigma^2_{Z_{2i-1}} \quad \text{and} \quad \mathbb{E}\left[(Y_{2i}-Y'_{2i-1})_+^2 \;\middle|\; \mathbf{Z}\right] = \sigma^2_{Z_{2i}}.$$

Hence, it follows that

$$\begin{aligned}\left|\frac{1}{n/2}\sum_{i=1}^{\lfloor n/2\rfloor}\mathbb{E}\left[(Y_{2i-1}-Y_{2i})_+^2 \;\middle|\; \mathbf{Z}\right] - \frac{1}{n}\sum_{i=1}^{\lfloor n/2\rfloor}\big(\sigma^2_{Z_{2i-1}} + \sigma^2_{Z_{2i}}\big)\right| \\ \le \frac{8}{n}\sum_{i=1}^{\lfloor n/2\rfloor}\mathrm{d}_{\mathrm{W}_1}\big(P_{Y|Z}(\cdot \mid Z_{2i-1}), P_{Y|Z}(\cdot \mid Z_{2i})\big).\end{aligned}$$

Step 3: another concentration step. Since $Y \in [-1,1]$,

$$\left|\frac{1}{n}\sum_{i=1}^{\lfloor n/2\rfloor}\big(\sigma^2_{Z_{2i-1}} + \sigma^2_{Z_{2i}}\big) - \frac{1}{n}\sum_{i=1}^{n}\sigma^2_{Z_i}\right| \le \frac{4}{n}.$$

Furthermore, $\sigma^2 Z_1, \ldots, \sigma^2_{Z_n}$ are independent and uniformly bounded random variables. Thus, by weak law of large numbers,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\sigma^2_{Z_i} - \mathbb{E}\left[\mathrm{Var}\left(Y \mid Z\right)\right]\right| = \mathrm{o}_P(1).$$

Hence, so far we have established that

$$\left|\frac{1}{n/2}\sum_{i=1}^{\lfloor n/2\rfloor}\mathbb{E}\left[(Y_{2i-1}-Y_{2i})_+^2 \;\middle|\; \mathbf{Z}\right] - \mathbb{E}\left[\mathrm{Var}\left(Y \mid Z\right)\right]\right| \le \frac{8}{n}\sum_{i=1}^{\lfloor n/2\rfloor}\mathrm{d}_{\mathrm{W}_1}\big(P_{Y|Z}(\cdot \mid Z_{2i-1}), P_{Y|Z}(\cdot \mid Z_{2i})\big) + \mathrm{o}_P(1).$$

Step 4: bounding the Wasserstein distance by a TV distance. Fix $N \in \mathbb{N}$, and define the random variable $\tilde{Y} := (1/N)\lfloor NY \rfloor$, which essentially amounts to rounding off Y to the closest point to the left of Y on the grid $(\mathbb{N})/N$.

For any $i \in \lfloor n/2 \rfloor$,

$$\begin{aligned}&\left|\mathrm{d}_{\mathrm{W}_1}\big(P_{Y|Z}(\cdot \mid Z_{2i-1})P_{Y|Z}(\cdot \mid Z_{2i})\big) - \mathrm{d}_{\mathrm{W}_1}\big(P_{\tilde{Y}|Z}(\cdot \mid Z_{2i-1}), P_{\tilde{Y}|Z}(\cdot \mid Z_{2i})\big)\right| \\ &\quad \le \mathrm{d}_{\mathrm{W}_1}\big(P_{Y|Z}(\cdot \mid Z_{2i-1}), P_{\tilde{Y}|Z}(\cdot \mid Z_{2i-1})\big) + \mathrm{d}_{\mathrm{W}_1}\big(P_{Y|Z}(\cdot \mid Z_{2i}), P_{\tilde{Y}|Z}(\cdot \mid Z_{2i})\big) \le 2/N,\end{aligned}$$

where the last inequality follows by noting that $|Y - \tilde{Y}| \le 1/N$. Furthermore, $\tilde{Y} \in [-1, 1]$ and thus, by Villani (2009, Theorem 6.15)

$$\begin{aligned}\mathrm{d}_{\mathrm{W}_1}\big(P_{\tilde{Y}|Z}(\cdot \mid Z_{2i-1}), P_{\tilde{Y}|Z}(\cdot \mid Z_{2i})\big) &\le 2\,\mathrm{d}_{\mathrm{TV}}\big(P_{\tilde{Y}|Z}(\cdot \mid Z_{2i-1}), P_{\tilde{Y}|Z}(\cdot \mid Z_{2i})\big) \\ &\le 2\sum_{k=-N}^{N}\left|P_{\tilde{Y}|Z}(k/N \mid Z_{2i-1}) - P_{\tilde{Y}|Z}(k/N \mid Z_{2i})\right|.\end{aligned}$$

Writing $f_k(Z_1, \ldots, Z_n) := \frac{2}{n-1}\sum_{i=1}^{n-1}\left|P_{\tilde{Y}|Z}(k/N \mid Z_i) - P_{\tilde{Y}|Z}(k/N \mid Z_{i+1})\right|$ for any $k \in [-N, N]$, we have that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{\lfloor n/2 \rfloor}\mathrm{d}_{\mathrm{W}_1}\big(P_{Y|Z}(\cdot \mid Z_{2i-1}), P_{Y|Z}(\cdot \mid Z_{2i})\big) \\ &\le \frac{1}{N} + \frac{2}{n}\sum_{i=1}^{\lfloor n/2 \rfloor}\sum_{k=-N}^{N}\left|P_{\tilde{Y}|Z}(k/N \mid Z_{2i-1}) - P_{\tilde{Y}|Z}(k/N \mid Z_{2i})\right| \\ &\le \frac{1}{N} + \frac{n-1}{n}\sum_{k=-N}^{N} f_k(Z_1, \ldots, Z_n).\end{aligned}$$

Step 5: controlling the total-variation term. Fix $K \in [-N, N]$. Observe that if we resample a single Z_i (and write $(Z_1', \ldots, Z_n')$ to denote this new sample), the perturbation in f_k is given by

$$\left|f_k(Z_1, \ldots, Z_n) - f_k(Z_1', \ldots, Z_n')\right| \le \frac{2}{n} \cdot 2 = \frac{4}{n}.$$

This follows by noting that the resampled Z_i' can alter at most two of the summands.

Now, let $\tilde{Z}_1, \tilde{Z}_2, \ldots$ be an infinite sequence of i.i.d. copies of Z. For each $n \ge 2$ and each $1 \le i \le n$, let $\tilde{Z}_{n,i}$ be the Euclidean-nearest neighbour of $\tilde{Z}_i$ among $\{\tilde{Z}_j : 1 \le j \le n, j \ne i \text{ and } \tilde{Z}_j \ge \tilde{Z}_i\}$. Now, observe that for any $i \in [n]$,

$$\mathbb{E}\left[\left|P_{\tilde{Y}|Z}(k/N \mid Z_i) - P_{\tilde{Y}|Z}(k/N \mid Z_{i+1})\right|\right] = \mathbb{E}_{P_Z}\left[\left|P_{\tilde{Y}|Z}(k/N \mid \tilde{Z}_1) - P_{\tilde{Y}|Z}(k/N \mid \tilde{Z}_{n,1})\right|\right].$$

Since $Z_1, \ldots, Z_n$ are independent, by McDiarmid's inequality, for any $\delta > 0$

$$\left|f_k(Z_1, \ldots, Z_n) - 2\mathbb{E}_{P_Z}\left[\left|P_{\tilde{Y}|Z}(k/N \mid \tilde{Z}_1) - P_{\tilde{Y}|Z}(k/N \mid \tilde{Z}_{n,1})\right|\right]\right| = \mathrm{o}_P(1). \tag{45}$$

Now, we claim that

$$\left|P_{\tilde{Y}|Z}(k/N \mid \tilde{Z}_1) - P_{\tilde{Y}|Z}(k/N \mid \tilde{Z}_{n,1})\right| = \mathrm{o}_P(1). \tag{46}$$

To prove this claim, fix an $\epsilon > 0$ and a $\delta > 0$. By Lusin's theorem, there exists a compactly supported continuous function g_k such that

$$\mathbb{P}_{\tilde{Z}\sim P_Z}\left\{P_{\tilde{Y}|Z}(k/N \mid \tilde{Z}) \neq g_k(\tilde{Z})\right\} < \epsilon.$$

Now, $\mathbb{P}\left\{\left|P_{\tilde{Y}|Z}(k/N \mid \tilde{Z}_1) - P_{\tilde{Y}|Z}(k/N \mid \tilde{Z}_{n,1})\right| \geq \delta\right\}$ is upper bounded by

$$\mathbb{P}\left\{|g_k(\tilde{Z}_1) - g_k(\tilde{Z}_{n,1})| \geq \delta\right\} + \mathbb{P}\left\{P_{\tilde{Y}|Z}(k/N \mid \tilde{Z}_1) \neq g_k(\tilde{Z}_1)\right\} + \mathbb{P}\left\{P_{\tilde{Y}|Z}(k/N \mid \tilde{Z}_{n,1}) \neq g_k(\tilde{Z}_{n,1})\right\}.$$

By construction, $\mathbb{P}\left\{P_{\tilde{Y}|Z}\big(k/N \mid \tilde{Z}_1\big) \neq g_k(\tilde{Z}_1)\right\} \leq \epsilon$ and by Lemma D.27, there exists an universal constant C_0 such that

$$\mathbb{P}\left\{P_{\tilde{Y}|Z}(k/N \mid \tilde{Z}_{n,1}) \neq g_k(\tilde{Z}_{n,1})\right\} \leq C_0\,\mathbb{P}\left\{P_{\tilde{Y}|Z}(k/N \mid \tilde{Z}_1) \neq g_k(\tilde{Z}_1)\right\} \leq C_0\epsilon.$$

Moreover, by continuity of g_k and by Lemma D.27,

$$\lim_{n\to\infty} \mathbb{P}\left\{|g_k(Z) - g_k(Z_{RN})| \geq \delta\right\} = 0.$$

Thus, we have that

$$\lim_{n\to\infty} \mathbb{P}\left\{P_{\tilde{Y}|Z}\big(k/N \mid Z\big) - P_{\tilde{Y}|Z}\big(k/N \mid Z_{RN}\big)| \geq \delta\right\} \leq \epsilon(1 + C_0).$$

Since ϵ, δ was arbitrary, (46) holds.

Since the sequence of random variables in (46) is uniformly bounded, it further implies that

$$\mathbb{E}_{P_Z}\left[\left|P_{\tilde{Y}|Z}(k/N \mid \tilde{Z}_1) - P_{\tilde{Y}|Z}(k/N \mid \tilde{Z}_{n,1})\right|\right] = \mathrm{o}_P(1).$$

Since the conclusion above holds for any $k \in [-N, N]$, we finally have that

$$\frac{1}{n}\sum_{i=1}^{\lfloor n/2\rfloor} \mathrm{d}_{\mathrm{W}_1}\big(P_{Y|Z}(\cdot \mid Z_{2i-1}), P_{Y|Z}(\cdot \mid Z_{2i})\big) \leq \frac{1}{N} + \frac{n-1}{n}\sum_{k=-N}^{N} f_k(Z_1, \ldots, Z_n) \leq \frac{1}{N} + \mathrm{o}_P(1).$$

But, N was arbitrary to start with. Thus, it follows that

$$\frac{1}{n}\sum_{i=1}^{\lfloor n/2\rfloor} \mathrm{d}_{\mathrm{W}_1}\big(P_{Y|Z}(\cdot \mid Z_{2i-1}), P_{Y|Z}(\cdot \mid Z_{2i})\big) = \mathrm{o}_P(1),$$

which concludes the proof. □

Lemma C.21. *For cross bin matching under any distribution $P_{Y,Z}$, it holds that*

$$\left|\frac{1}{n/2}\|\Delta\mathbf{Y}^+\|_2^2 - \mathbb{E}\left[\mathrm{Dev}(P_{Y|Z})\right]\right| \leq C\big(L_W \vee 1\big)\left(\frac{1}{\sqrt{K}} + \frac{1}{\sqrt{n/K}} + \sqrt{\frac{\log(4/\delta)}{n}}\right)$$

where C is a universal constant.

Proof. Let $P_{(Y,Z)}$ be a joint distribution on $(Y, Z) \in \mathbb{R} \times \mathbb{R}$. Write P_Z and $P_{Y|Z}$ for the marginal and conditional distributions. Define also F_Z^{-1} and $F_{Y|Z}^{-1}(\cdot \mid z)$ as the generalized inverse CDF for P_Z and $P_{Y|Z}(\cdot \mid z)$.

We define some notation. Let $A_k \subseteq [n]$ denote the "kth bin"—that is, if we sort Z values from smallest to largest, A_k corresponds to the indices appearing in positions $(k-1)m+1, \dots, m$ in the sorted list. If we write $Y_{k,(1)} \geq \cdots \geq Y_{k,(m)}$ as the sorted values of Y in the kth bin $A_k = \{(k-1)m+1, \dots, km\}$, then the cross-bin matching returns

$$\widehat{\mathrm{Dev}}_{\text{cross-bin}} = \sum_{k=1}^{K-1} \sum_{i=1}^{\lfloor m/2 \rfloor} (Y_{k,(i)} - Y_{k+1,(m+1-i)})_+^2.$$

Step 1: rewriting with a Lipschitz function. Define a function $f : [-1,1]^m \times [-1,1]^m \to \mathbb{R}$ as

$$f(y, y') = \sum_{i=1}^{\lfloor m/2 \rfloor} (y_{(i)} - y'_{(m+1-i)})_+^2$$

where $y_{(1)} \geq \cdots \geq y_{(m)}$ and $y'_{(1)} \geq \cdots \geq y'_{(m)}$ denote the sorted values of y and of y'. Note that, by construction,

$$\widehat{\mathrm{Dev}}_{\text{cross-bin}} = \sum_{k=1}^{K-1} f(Y_{A_k}, Y_{A_{k+1}}).$$

This function satisfies several key properties:

Lemma C.22. *The function f satisfies the Lipschitz property*

$$|f(y, y') - f(\tilde{y}, \tilde{y}')| \leq 4\|y - \tilde{y}\|_1 + 4\|y' - \tilde{y}'\|_1.$$

Lemma C.23. *Let Q be any distribution on $\mathbb{R}$, and let $A = (A_1, \dots, A_m) \sim Q^m$ and $B = (B_1, \dots, B_m) \sim Q^m$. Then*

$$\left| \frac{1}{m/2} \mathbb{E}\left[f(A, B)\right] - \mathrm{Dev}(Q) \right| \leq \frac{C}{\sqrt{m}}$$

for a universal constant C.

Step 2: concentration. First, observe that if we resample a single Y_i value (and write $(Y'_1, \dots, Y'_n)$ to denote this new sample), the perturbation in $\widehat{\mathrm{Dev}}_{\text{cross-bin}}$ is given by

$$\left| \sum_{k=1}^{K-1} f(Y_{A_k}, Y_{A_{k+1}}) - \sum_{k=1}^{K-1} f(Y'_{A_k}, Y'_{A_{k+1}}) \right| \leq \sum_{k=1}^{K-1} \left(4\|Y_{A_k} - Y'_{A_k}\|_1 + 4\|Y_{A_{k+1}} - Y'_{A_{k+1}}\|_1 \right) \leq 16$$

where the last step holds since, for each k, $\|Y_{A_k} - Y'_{A_k}\|_1 \leq 2 \cdot \mathbb{1}_{i \in A_k}$ and similarly $\|Y_{A_{k+1}} - Y'_{A_{k+1}}\|_1 \leq 2 \cdot \mathbb{1}_{i \in A_{k+1}}$. Therefore by McDiarmid's inequality, since the Y_i's are independent conditional on $Z_1, \dots, Z_n$, with probability $\geq 1 - \delta/2$ it holds that

$$\left| \widehat{\mathrm{Dev}}_{\text{cross-bin}} - \mathbb{E}\left[\widehat{\mathrm{Dev}}_{\text{cross-bin}} \;\middle|\; Z_1, \dots, Z_n \right] \right| \leq 8\sqrt{2n \log(4/\delta)}.$$

Step 3: an equivalent way to sample the data. Now let $V_1, \dots, V_n, U_1, \dots, U_n \overset{\text{iid}}{\sim} \text{Unif}[0,1]$. First we define $Z_i = F_Z^{-1}(V_i)$, so that we have $Z_1, \dots, Z_n \overset{\text{iid}}{\sim} P_Z$. Then we define

$$Y_i = F_{Y|Z}^{-1}(U_i \mid Z_i)$$

for each i—this is equivalent to sampling $(Y_i, Z_i) \overset{\text{iid}}{\sim} P_{(Y,Z)}$.

Step 4: replace the Y values with oracle samples. Now let $0 < u_1^* \le \cdots \le u_{K-1}^* < 1$ be any values and let $z_k^* = F_Z^{-1}(u_k^*)$. For each $i \in A_k$, define

$$\tilde{Y}_i = F_{Y|Z}^{-1}(U_i \mid z_k^*)$$

and for each $i \in A_{k+1}$, define

$$\tilde{Y}_i' = F_{Y|Z}^{-1}(U_i \mid z_k^*).$$

Now define

$$\widetilde{\text{Dev}}_{\text{cross-bin}} = \sum_{k=1}^{K-1} f(\tilde{Y}_{A_k}, \tilde{Y}'_{A_{k+1}}),$$

so that

$$\left|\widehat{\text{Dev}}_{\text{cross-bin}} - \widetilde{\text{Dev}}_{\text{cross-bin}}\right| \le \sum_{k=1}^{K-1} \left| f(Y_{A_k}, Y_{A_{k+1}}) - f(\tilde{Y}_{A_k}, \tilde{Y}'_{A_{k+1}}) \right|.$$

By Lemma C.22, for each k,

$$\left| f(Y_{A_k}, Y_{A_{k+1}}) - f(\tilde{Y}_{A_k}, \tilde{Y}'_{A_{k+1}}) \right| \le 4\|Y_{A_k} - \tilde{Y}_{A_k}\|_1 + 4\|Y_{A_{k+1}} - \tilde{Y}'_{A_{k+1}}\|_1.$$

Taking expected values, then,

$$\mathbb{E}\left[\left| f(Y_{A_k}, Y_{A_{k+1}}) - f(\tilde{Y}_{A_k}, \tilde{Y}'_{A_{k+1}}) \right| \;\middle|\; Z_1, \dots, Z_n \right] \le \mathbb{E}\left[4\|Y_{A_k} - \tilde{Y}_{A_k}\|_1 + 4\|Y_{A_{k+1}} - \tilde{Y}'_{A_{k+1}}\|_1 \;\middle|\; Z_1, \dots, Z_n \right].$$

Now we calculate an upper bound. For any $k = 1, \dots, K-1$ and any $i \in A_k$,

$$\mathbb{E}\left[|Y_i - \tilde{Y}_i| \;\middle|\; Z_1, \dots, Z_n \right] = \mathrm{d_W}(P_{Y|Z}(\cdot \mid Z_i), P_{Y|Z}(\cdot \mid z_k^*)),$$

by definition of the 1-Wasserstein distance, and by construction of Y_i and $\tilde{Y}_i$. Similarly for $i \in A_{k+1}$,

$$\mathbb{E}\left[|Y_i - \tilde{Y}_i'| \;\middle|\; Z_1, \dots, Z_n \right] = \mathrm{d_W}(P_{Y|Z}(\cdot \mid Z_i), P_{Y|Z}(\cdot \mid z_k^*)).$$

And, since we have made a smoothness assumption on the conditional distributions, since $Z_i = F_Z^{-1}(V_i)$ and $z_k^* = F_Z^{-1}(u_k^*)$, for each i and each k we have

$$\mathrm{d_W}(P_{Y|Z}(\cdot \mid Z_i), P_{Y|Z}(\cdot \mid z_k^*)) \le L_W |V_i - u_k^*|.$$

Therefore,

$$\mathbb{E}\left[\left| f(Y_{A_k}, Y_{A_{k+1}}) - f(\tilde{Y}_{A_k}, \tilde{Y}'_{A_{k+1}}) \right| \;\middle|\; Z_1, \dots, Z_n \right] \le 4L_W \sum_{i \in A_k \cup A_{k+1}} |V_i - u_k^*|.$$

We therefore have

$$\left|\mathbb{E}\left[\widehat{\mathrm{Dev}}_{\text{cross-bin}} \;\middle|\; Z_1,\dots,Z_n\right] - \mathbb{E}\left[\widetilde{\mathrm{Dev}}_{\text{cross-bin}} \;\middle|\; Z_1,\dots,Z_n\right]\right| \\ \leq \mathbb{E}\left[\left|\widetilde{\mathrm{Dev}}_{\text{cross-bin}} - \widehat{\mathrm{Dev}}_{\text{cross-bin}}\right| \;\middle|\; Z_1,\dots,Z_n\right] \leq 4L_W \sum_{k=1}^{K-1} \sum_{i\in A_k\cup A_{k+1}} |V_i - u_k^*|.$$

Next, by Lemma C.23, since $\tilde{Y}_{A_k}$ and $\tilde{Y}'_{A_{k+1}}$ each represent m i.i.d. draws from $P_{Y|Z}(\cdot \mid z_k^*)$ for each k, we have

$$\left|\frac{1}{m/2}\mathbb{E}\left[f(\tilde{Y}_{A_k}, \tilde{Y}'_{A_{k+1}}) \;\middle|\; Z_1,\dots,Z_n\right] - \mathrm{Dev}(P_{Y|Z}(\cdot \mid z_k^*)\right| \leq \frac{C}{\sqrt{m}}.$$

Therefore,

$$\left|\mathbb{E}\left[\widetilde{\mathrm{Dev}}_{\text{cross-bin}} \;\middle|\; Z_1,\dots,Z_n\right] - \frac{m}{2}\sum_{k=1}^{K-1} \mathrm{Dev}(P_{Y|Z}(\cdot \mid z_k^*)\right| \leq \frac{m}{2}\cdot(K-1)\cdot\frac{C}{\sqrt{m}}.$$

Combined with the calculations above, then,

$$\left|\mathbb{E}\left[\widehat{\mathrm{Dev}}_{\text{cross-bin}} \;\middle|\; Z_1,\dots,Z_n\right] - \frac{m}{2}\sum_{k=1}^{K-1} \mathrm{Dev}(P_{Y|Z}(\cdot \mid z_k^*)\right| \\ \leq 4L_W \sum_{k=1}^{K-1} \sum_{i\in A_k\cup A_{k+1}} |V_i - u_k^*| + \frac{m}{2}\cdot(K-1)\cdot\frac{C}{\sqrt{m}}.$$

Step 5: another Wasserstein distance. By definition of the bins A_k (i.e., these bins reflect the ordering of the V_i's), we can rewrite

$$\sum_{k=1}^{K-1}\sum_{i\in A_k} |V_i - u_k^*| = \sum_{k=1}^{K-1}\sum_{i=(k-1)m+1}^{km} |V_{(i)} - u_k^*|.$$

Since $u_1^* \leq \cdots \leq u_{K-1}^*$, by definition of the 1-Wasserstein distance, we have

$$\frac{1}{m(K-1)}\sum_{k=1}^{K-1}\sum_{i=(k-1)m+1}^{km} |V_{(i)} - u_k^*| = \mathrm{d_W}\big(\widehat{P}_V^{(1)}, \widehat{P}_{u^*}\big),$$

where $\widehat{P}_V^{(1)}$ is the empirical distribution of $V_{(1)},\dots,V_{((K-1)m)}$, and $\widehat{P}_{u^*}$ is the empirical distribution of $u_1^*,\dots,u_{K-1}^*$. By the triangle inequality,

$$\mathrm{d_W}\big(\widehat{P}_V^{(1)}, \widehat{P}_{u^*}\big) \leq \mathrm{d_W}\big(\widehat{P}_V, \mathrm{Unif}[0,1]\big) + \mathrm{d_W}\big(\widehat{P}_{u^*}, \mathrm{Unif}[0,1]\big) + \mathrm{d_W}\big(\widehat{P}_V^{(1)}, \widehat{P}_V\big) \\ \leq \mathrm{d_W}\big(\widehat{P}_V, \mathrm{Unif}[0,1]\big) + \mathrm{d_W}\big(\widehat{P}_{u^*}, \mathrm{Unif}[0,1]\big) + \frac{2m}{n},$$

where $\widehat{P}_V$ is the empirical distribution of $V_1, \ldots, V_n$ (note that the last inequality holds since $V_{(1)}, \ldots, V_{((K-1)m)}$ is a subset of the list $V_1, \ldots, V_n$, containing all but $n - (K-1)m \leq 2m$ many entries). Therefore,

$$\sum_{k=1}^{K-1} \sum_{i \in A_k} |V_i - u_k^*| \leq m(K-1) \left(\mathrm{d_W}\big(\widehat{P}_V, \mathrm{Unif}[0,1]\big) + \mathrm{d_W}\big(\widehat{P}_{u^*}, \mathrm{Unif}[0,1]\big) \right) + 2m.$$

By an identical argument the same bound holds for $\sum_{k=1}^{K-1} \sum_{i \in A_{k+1}} |V_i - u_k^*|$. Therefore, combining with the results of the previous step,

$$\left| \mathbb{E}\left[\widehat{\mathrm{Dev}}_{\text{cross-bin}} \;\middle|\; Z_1, \ldots, Z_n \right] - \frac{m}{2} \sum_{k=1}^{K-1} \mathrm{Dev}(P_{Y|Z}(\cdot \mid z_k^*)) \right|$$
$$\leq 4L_W m(K-1) \left(\mathrm{d_W}\big(\widehat{P}_V, \mathrm{Unif}[0,1]\big) + \mathrm{d_W}\big(\widehat{P}_{u^*}, \mathrm{Unif}[0,1]\big) \right) + 8L_W m + \frac{m}{2} \cdot (K-1) \cdot \frac{C}{\sqrt{m}}.$$

Step 6: averaging over the reference values. The above calculations hold for any fixed $0 < u_1^* \leq \ldots u_{K-1}^* < 1$. Now we will make these random: let $U_1^*, \ldots, U_{K-1}^* \overset{\text{iid}}{\sim} \mathrm{Unif}[0,1]$ and let $U_{(1)}^* \leq \cdots \leq U_{(K-1)}^*$ be the order statistics. And, let $Z_k^* = F_Z^{-1}(U_k^*)$, with $Z_{(k)}^* = F_Z^{-1}(U_{(k)}^*)$ being the order statistics. Since $Z_k^* \sim P_Z$ for each k, we have

$$\mathbb{E}\left[\sum_{k=1}^{K-1} \mathrm{Dev}(P_{Y|Z}(\cdot \mid Z_{(k)}^*)) \right] = \mathbb{E}\left[\sum_{k=1}^{K-1} \mathrm{Dev}(P_{Y|Z}(\cdot \mid Z_k^*)) \right] = (K-1)\mathbb{E}\left[\mathrm{Dev}(P_{Y|Z}) \right].$$

Combining with the previous step, and applying Jensen's inequality when taking expectation over the distribution of the Z_k^*'s (while conditioning on $Z_1, \ldots, Z_n$), we then have

$$\left| \mathbb{E}\left[\widehat{\mathrm{Dev}}_{\text{cross-bin}} \;\middle|\; Z_1, \ldots, Z_n \right] - \frac{m}{2} \cdot (K-1)\mathbb{E}\left[\mathrm{Dev}(P_{Y|Z}) \right] \right|$$
$$\leq 4L_W m(K-1) \mathrm{d_W}\big(\widehat{P}_V, \mathrm{Unif}[0,1]\big) + 8L_W m + \frac{m}{2} \cdot (K-1) \cdot \frac{C}{\sqrt{m}}$$
$$+ 4L_W n \mathbb{E}\left[\mathrm{d_W}\big(\widehat{P}_{U^*}, \mathrm{Unif}[0,1]\big) \right],$$

where $\widehat{P}_{U^*}$ is the empirical distribution of $U_1^*, \ldots, U_{K-1}^*$. Since $n - 2m \leq m(K-1) \leq n$, and $\mathbb{E}\left[\mathrm{Dev}(P_{Y|Z}) \right] \in [0,4]$, then,

$$\left| \mathbb{E}\left[\widehat{\mathrm{Dev}}_{\text{cross-bin}} \;\middle|\; Z_1, \ldots, Z_n \right] - \frac{n}{2}\mathbb{E}\left[\mathrm{Dev}(P_{Y|Z}) \right] \right|$$
$$\leq 4L_W m(K-1) \mathrm{d_W}\big(\widehat{P}_V, \mathrm{Unif}[0,1]\big) + (8L_W + 4)m + \frac{m}{2} \cdot (K-1) \cdot \frac{C}{\sqrt{m}}$$
$$+ 4L_W m(K-1) \mathbb{E}\left[\mathrm{d_W}\big(\widehat{P}_{U^*}, \mathrm{Unif}[0,1]\big) \right].$$

By bounds on the empirical 1-Wasserstein distance from Lei (2020, Corollary 5.2.), since $V_i \overset{\text{iid}}{\sim} \mathrm{Unif}[0,1]$ and $U_k^* \overset{\text{iid}}{\sim} \mathrm{Unif}[0,1]$, we have

$$\mathbb{E}\left[\mathrm{d_W}\big(\widehat{P}_V, \mathrm{Unif}[0,1]\big) \right] \leq C' n^{-1/2}, \quad \mathbb{E}\left[\mathrm{d_W}\big(\widehat{P}_{U^*}, \mathrm{Unif}[0,1]\big) \right] \leq C'(K-1)^{-1/2}$$

for a universal constant C', and so

$$\begin{aligned}\left|\mathbb{E}\left[\widehat{\mathrm{Dev}}_{\text{cross-bin}} \;\middle|\; Z_1,\dots,Z_n\right] - \frac{n}{2}\mathbb{E}\left[\mathrm{Dev}(P_{Y|Z})\right]\right| \\ \leq 4L_W m(K-1)\left(\mathrm{d_W}\big(\widehat{P}_V, \mathrm{Unif}[0,1]\big) - \mathbb{E}\left[\mathrm{d_W}\big(\widehat{P}_V, \mathrm{Unif}[0,1]\big)\right]\right) \\ + 4CL_W\sqrt{n} + (8L_W+4)m + \frac{m}{2}\cdot(K-1)\cdot\frac{C}{\sqrt{m}} + 4C'L_W m\sqrt{K-1}.\end{aligned}$$

Step 7: another concentration step. Since $V_1,\dots,V_n$ are i.i.d., and the quantity $\mathrm{d_W}\big(\widehat{P}_V, \mathrm{Unif}[0,1]\big)$ can change value by at most $\frac{2}{n}$ if we resample one value V_i, by McDiarmid's inequality, with probability $\geq 1-\delta/2$,

$$\left|\mathrm{d_W}\big(\widehat{P}_V, \mathrm{Unif}[0,1]\big) - \mathbb{E}\left[\mathrm{d_W}\big(\widehat{P}_V, \mathrm{Unif}[0,1]\big)\right]\right| \leq \sqrt{\frac{2\log(4/\delta)}{n}}.$$

Step 8: combining everything. From all the steps above, with probability $\geq 1-\delta$,

$$\begin{aligned}\left|\widehat{\mathrm{Dev}}_{\text{cross-bin}} - \frac{n}{2}\mathbb{E}\left[\mathrm{Dev}(P_{Y|Z})\right]\right| \\ \leq 8\sqrt{2n\log(4/\delta)} + 4L_W m(K-1)\cdot\sqrt{\frac{2\log(4/\delta)}{n}} \\ + 4CL_W\sqrt{n} + (8L_W+4)m + \frac{m}{2}\cdot(K-1)\cdot\frac{C}{\sqrt{m}} + 4C'L_W m\sqrt{K-1}.\end{aligned}$$

Letting C now denote a different universal constant, we have the final result □

D Auxiliary lemmas

Lemma D.24. *It holds that deterministically, for any $\alpha \in [0,1]$,*

$$\frac{1}{2^L}\sum_{\mathbf{s}\in\{\pm1\}^L} \mathbb{1}\left\{p(\mathbf{x}^{\mathbf{s}}) \leq \alpha\right\} \leq \alpha,$$

where $p(\cdot)$ is as defined in (7).

Proof. Consider a bijection $\sigma : [2^L] \to \{-1,1\}^L$ such that $T(\mathbf{x}^{\sigma(1)}) \geq \ldots \geq T(\mathbf{x}^{\sigma(2^L)})$, and let $r \in \{0,1,\dots,2^L\}$ be such that $\alpha \in \big[r/2^L, (r+1)/2^L\big)$. Then, since $\sum_{k=1}^{2^L} \mathbb{1}\left\{T(\mathbf{x}^{\sigma(k)}) \geq T(\mathbf{x}^{\sigma(j)})\right\} \in [2^L]$ for each $j \in [2^L]$, we have deterministically that

$$\begin{aligned}\frac{1}{2^L}\sum_{\mathbf{s}\in\{\pm1\}^L} \mathbb{1}\left\{p(\mathbf{x}^{\mathbf{s}}) \leq \alpha\right\} &= \frac{1}{2^L}\sum_{j=1}^{2^L} \mathbb{1}\left\{\frac{1}{2^L}\sum_{k=1}^{2^L} \mathbb{1}\left\{T(\mathbf{x}^{\sigma(k)}) \geq T(\mathbf{x}^{\sigma(j)})\right\} \leq \alpha\right\} \\ &= \frac{1}{2^L}\sum_{j=1}^{2^L} \mathbb{1}\left\{\sum_{k=1}^{2^L} \mathbb{1}\left\{T(\mathbf{x}^{\sigma(k)}) \geq T(\mathbf{x}^{\sigma(j)})\right\} \leq r\right\} \\ &\leq \frac{1}{2^L}\sum_{j=1}^{2^L} \mathbb{1}\left\{\sum_{k=1}^{2^L} \mathbb{1}\left\{k \leq j\right\} \leq r\right\} = \frac{1}{2^L}\sum_{j=1}^{2^L} \mathbb{1}\left\{j \leq r\right\} = \frac{r}{2^L} \leq \alpha.\end{aligned}$$

□

Lemma D.25. *For any $p, q \in (0,1)$, it holds that*

$$H^2\big(\mathrm{Ber}(p), \mathrm{Ber}(q)\big) \le \min\left\{ \frac{(p-q)^2}{2p(1-p)}, \frac{(p-q)^2}{2q(1-q)} \right\}.$$

Proof. We note that for any $p, q \in (0,1)$,

$$\begin{aligned} H^2\big(\mathrm{Ber}(p), \mathrm{Ber}(q)\big) &= \frac{1}{2}\left[(\sqrt{p} - \sqrt{q})^2 + (\sqrt{1-p} - \sqrt{1-q})^2 \right] \\ &= \frac{1}{2}\left[\frac{(p-q)^2}{(\sqrt{p}+\sqrt{q})^2} + \frac{(p-q)^2}{(\sqrt{1-p}+\sqrt{1-q})^2} \right] \\ &\le \frac{1}{2}\left[\frac{(p-q)^2}{p+q} + \frac{(p-q)^2}{(1-p)+(1-q)} \right] \\ &\le \frac{1}{2}\left[\frac{(p-q)^2}{p} + \frac{(p-q)^2}{(1-p)} \right] \le \frac{(p-q)^2}{2p(1-p)}. \end{aligned}$$

By symmetry, it is also bounded by $(p-q)^2/\big(2q(1-q)\big)$, which proves the lemma. □

Lemma D.26. *If S_1 and S_2 are finite subsets of $\mathbb{R}$ with $\mathrm{Med}(S_1 \cup S_2) < \mathrm{Med}(S_1)$, then $\mathrm{Med}(S_1 \cup S_2) \ge \mathrm{Med}(S_2)$.*

Proof. Let $S_1 = \{a_1, \dots, a_m\}$ with $a_1 \le \cdots \le a_m$, $S_2 = \{b_1, \dots, b_n\}$ with $b_1 \le \cdots \le b_n$, and $S_1 \cup S_2 = \{c_1, \dots, c_{m+n}\}$ with $c_1 \le \cdots \le c_{m+n}$.

We first claim that $\mathrm{Med}(S_2) < \mathrm{Med}(S_1)$. To see this, observe that

$$\begin{aligned} \big|\{k \in [m+n] : c_k \le \mathrm{Med}(S_1 \cup S_2)\}\big| &\le \big|\{k \in [m+n] : c_k < \mathrm{Med}(S_1)\}\big| \\ &= \big|\{i \in [m] : a_i < \mathrm{Med}(S_1)\}\big| + \big|\{j \in [n] : b_j < \mathrm{Med}(S_1)\}\big|. \end{aligned}$$

Recalling that $\big|\{j \in [n] : b_j < \mathrm{Med}(S_2)\}\big| \le n/2$, there are four cases to consider:

Case 1: m and n both are odd. In this case,

$\big|\{k \in [m+n] : c_k \le \mathrm{Med}(S_1 \cup S_2)\}\big| \ge (m+n)/2$, $\big|\{i \in [m] : a_i < \mathrm{Med}(S_1)\}\big| \le (m-1)/2$, and hence $\big|\{j \in [n] : b_j < \mathrm{Med}(S_1)\}\big| \ge (n+1)/2$, which proves the claim.

Case 2: m is odd, and n is even. In this case,

$\big|\{k \in [m+n] : c_k \le \mathrm{Med}(S_1 \cup S_2)\}\big| \ge (m+n+1)/2$, $\big|\{i \in [m] : a_i < \mathrm{Med}(S_1)\}\big| \le (m-1)/2$, and hence $\big|\{j \in [n] : b_j < \mathrm{Med}(S_1)\}\big| \ge (n+2)/2$, which proves the claim.

Case 3: m is even, and n is odd. In this case,

$\big|\{k \in [m+n] : c_k \le \mathrm{Med}(S_1 \cup S_2)\}\big| \ge (m+n+1)/2$, $\big|\{i \in [m] : a_i < \mathrm{Med}(S_1)\}\big| \le m/2$, and hence $\big|\{j \in [n] : b_j < \mathrm{Med}(S_1)\}\big| \ge (n+1)/2$, which proves the claim.

Case 4: m and n both are even. By definition,

$$\text{Med}(S_1) = \frac{a_{m/2} + a_{m/2+1}}{2}, \quad \text{Med}(S_2) = \frac{b_{n/2} + b_{n/2+1}}{2}$$
$$\text{and} \quad \text{Med}(S_1 \cup S_2) = \frac{c_{m/2+n/2} + c_{m/2+n/2+1}}{2}.$$

Now, if $a_{m/2} = a_{m/2+1}$, then

$$\big|\{k \in [m+n] : c_k \le \text{Med}(S_1 \cup S_2)\}\big| \ge (m+n)/2, \quad \big|\{i \in [m] : a_i < \text{Med}(S_1)\}\big| \le m/2 - 1,$$

and thus, $\big|\{j \in [n] : b_j < \text{Med}(S_1)\}\big| \ge n/2 + 1$, which proves the claim.

Otherwise, when $a_{m/2} < a_{m/2+1}$

$$\big|\{k \in [m+n] : c_k \le \text{Med}(S_1 \cup S_2)\}\big| \ge (m+n)/2, \quad \big|\{i \in [m] : a_i < \text{Med}(S_1)\}\big| = m/2,$$

and hence, $\big|\{j \in [n] : b_j < \text{Med}(S_1)\}\big| \ge n/2$ and $b_{n/2} < \text{Med}(S_1)$. Now again, there are four cases to consider.

(i) If $b_{n/2} \le a_{m/2} < a_{m/2+1} \le b_{n/2+1}$, then $\text{Med}(S_1 \cup S_2) = \frac{a_{m/2}+a_{m/2+1}}{2} = \text{Med}(S_1)$ which is a contradiction.

(ii) If $a_{m/2} < b_{n/2} \le a_{m/2+1} \le b_{n/2+1}$, then $\text{Med}(S_1 \cup S_2) = \frac{b_{n/2}+a_{m/2+1}}{2} > \frac{a_{m/2}+a_{m/2+1}}{2} = \text{Med}(S_1)$ which is a contradiction.

(iii) If $a_{m/2} < b_{n/2} \le b_{n/2+1} \le a_{m/2+1}$, then $\text{Med}(S_1) > \text{Med}(S_1 \cup S_2) = \frac{b_{n/2}+b_{n/2+1}}{2} = \text{Med}(S_2)$.

(iv) If $b_{n/2} \le a_{m/2} \le b_{n/2+1} < a_{m/2+1}$, then $\text{Med}(S_1) = \frac{a_{m/2}+a_{m/2+1}}{2} > \frac{b_{n/2}+b_{n/2+1}}{2} = \text{Med}(S_2)$.

This establishes the claim that $\text{Med}(S_2) < \text{Med}(S_1)$. If we had $\text{Med}(S_1 \cup S_2) < \text{Med}(S_2)$, then by the same argument as above, we could conclude that $\text{Med}(S_1) < \text{Med}(S_2)$ which is a contradiction. This proves the result. □

Proof of Lemma C.22. First, for each $i = 1, \dots, \lfloor m/2 \rfloor$, since $t \mapsto (t)_+^2$ is 4-Lipschitz on $t \in [-2, 2]$,

$$\big|(y_{(i)} - y'_{(m+1-i)})_+^2 - (\tilde{y}_{(i)} - \tilde{y}'_{(m+1-i)})_+^2\big| \le 4|y_{(i)} - \tilde{y}_{(i)}| + 4|y'_{(m+1-i)} - \tilde{y}'_{(m+1-i)}|.$$

Therefore,

$$|f(y, y') - f(\tilde{y}, \tilde{y}')| \le 4 \sum_{i=1}^{\lfloor m/2 \rfloor} |y_{(i)} - \tilde{y}_{(i)}| + 4 \sum_{i=1}^{\lfloor m/2 \rfloor} |y'_{(m+1-i)} - \tilde{y}'_{(m+1-i)}|.$$

Writing $y_{()} = (y_{(1)}, \dots, y_{(m)})$ as the sorted vector, and same for $y'_{()}$, $\tilde{y}_{()}$, and $\tilde{y}'_{()}$, then,

$$|f(y, y') - f(\tilde{y}, \tilde{y}')| \le 4\|y_{()} - \tilde{y}_{()}\|_1 + 4\|y'_{()} - \tilde{y}'_{()}\|_1.$$

Finally, it holds that

$$\|y_{()} - \tilde{y}_{()}\|_1 \le \|y - \tilde{y}\|_1$$

by an ℓ_1 version of the rearrangement inequality, and similarly for $y', \tilde{y}'$. This completes the proof. □

Proof of Lemma C.23. First let $A_{(1)} \geq \cdots \geq A_{(m)}$ and $B_{(1)} \geq \cdots \geq B_{(m)}$ be the order statistics, and write $A_{()} = (A_{(1)}, \ldots, A_{(m)})$ and $B_{()} = (B_{(1)}, \ldots, B_{(m)})$ for the sorted vectors. Then

$$f(A, B) = f(A_{()}, B_{()})$$

since $f(y, y')$ is invariant to permutations of y, and invariant to permutations of y'.

Next, let F_Q^{-1} be a generalized CDF for Q. Let $U_1, \ldots, U_m \overset{\text{iid}}{\sim} \text{Unif}[0,1]$, so that we can equivalently define $A_i = F_Q^{-1}(U_i)$ for $i = 1, \ldots, m$. Define $A'_i = F_Q^{-1}(1 - U_i)$, and let $A'_{()}$ denote the sorted vector as before. Then

$$|f(A, B) - f(A, A')| = \left|f(A_{()}, B_{()}) - f(A_{()}, A'_{()})\right| \leq 4\|A'_{()} - B_{()}\|_1$$

by Lemma C.22, and so applying Jensen's inequality,

$$\left|\mathbb{E}\left[f(A_{()}, B_{()})\right] - \mathbb{E}\left[f(A_{()}, A'_{()}\right]\right| \leq 4\mathbb{E}\left[\|A'_{()} - B_{()}\|_1\right].$$

Next, let $U_{(1)} \leq \cdots \leq U_{(m)}$ be the sorted values. Then we have $A_{(i)} = F_Q^{-1}(U_{(m+1-i)})$ and $A'_{(i)} = F_Q^{-1}(U_{(i)})$ for each i. Then by construction,

$$f(A, A') = \sum_{i=1}^{\lfloor m/2 \rfloor} (A_{(i)} - A'_{(m+1-i)})_+^2 = \sum_{i=1}^{\lfloor m/2 \rfloor} (F_Q^{-1}(U_{(m+1-i)}) - F_Q^{-1}(1 - U_{(m+1-i)}))_+^2.$$

Therefore,

$$\mathbb{E}\left[f(A, A')\right] = \mathbb{E}\left[\sum_{i=1}^{\lfloor m/2 \rfloor} (F_Q^{-1}(U_{(m+1-i)}) - F_Q^{-1}(1 - U_{(m+1-i)}))_+^2\right].$$

And by symmetry,

$$\mathbb{E}\left[f(A, A')\right] = \mathbb{E}\left[f(A', A)\right] = \mathbb{E}\left[\sum_{i=1}^{\lfloor m/2 \rfloor} (F_Q^{-1}(1 - U_{(i)}) - F_Q^{-1}(U_{(i)}))_+^2\right].$$

Therefore,

$$\mathbb{E}\left[f(A, A')\right] = \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{m} (F_Q^{-1}(1 - U_{(i)}) - F_Q^{-1}(U_{(i)}))^2 \cdot \mathbb{1}_{\mathcal{E}_i}\right]$$

where $\mathcal{E}_i$ is the event that $i < \frac{m}{2}$ and $U_{(i)} < \frac{1}{2}$, or, $i > \frac{m}{2}$ and $U_{(i)} > \frac{1}{2}$. Therefore, since each term in the sum is bounded by 1,

$$\left|\mathbb{E}\left[f(A, A')\right] - \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{m} (F_Q^{-1}(1 - U_{(i)}) - F_Q^{-1}(U_{(i)}))^2\right]\right| \leq \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{m} \mathbb{1}_{\mathcal{E}_i^c}\right].$$

We can rewrite this as

$$\left|\mathbb{E}\left[f(A, A')\right] - \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{m} (F_Q^{-1}(1 - U_i) - F_Q^{-1}(U_i))^2\right]\right| \leq \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{m} \mathbb{1}_{\mathcal{E}_i^c}\right].$$

By definition, for each i,

$$\mathbb{E}\left[(F_Q^{-1}(U_i) - F_Q^{-1}(1-U_i))^2\right] = \mathrm{Dev}(Q).$$

Therefore,

$$\left|\mathbb{E}\left[f(A,A')\right] - \frac{m}{2}\mathrm{Dev}(Q)\right| \leq \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{m} \mathbb{1}_{\mathcal{E}_i^c}\right] \leq C''\sqrt{m},$$

where the last step holds for a universal constant C'' by properties of the Binomial distribution, since

$$\sum_{i=1}^{m} \mathbb{1}_{\mathcal{E}_i^c} \leq 1 + \left|\sum_{i=1}^{m} \mathbb{1}_{U_i<1/2} - \frac{m}{2}\right|$$

and $\sum_{i=1}^{m} \mathbb{1}_{U_i<1/2} \sim \mathrm{Binomial}(m, \frac{1}{2})$.

Returning to the initial calculations, then, we have

$$\left|\mathbb{E}\left[f(A,B)\right] - \frac{m}{2}\mathrm{Dev}(Q)\right| \leq 4\mathbb{E}\left[\|A'_{()} - B_{()}\|_1\right] + C''\sqrt{m}.$$

Next we turn to bounding $\|A'_{()} - B_{()}\|_1$. Write $\widehat{P}_{A'}$ and $\widehat{P}_B$ as the empirical distributions of A' and of B, respectively. Then

$$\|A'_{()} - B_{()}\|_1 = m\mathrm{d_W}(\widehat{P}_{A'}, \widehat{P}_B) \leq m\mathrm{d_W}(\widehat{P}_{A'}, \mathrm{Unif}[0,1]) + m\mathrm{d_W}(\widehat{P}_B, \mathrm{Unif}[0,1]).$$

By bounds on the empirical 1-Wasserstein distance from Lei (2020, Corollary 5.2.) we have

$$\mathbb{E}\left[\mathrm{d_W}(\widehat{P}_{A'}, \mathrm{Unif}[0,1])\right] \leq C' m^{-1/2}$$

for a universal constant C', and same for $\widehat{P}_B$. Therefore,

$$\left|\mathbb{E}\left[f(A,B)\right] - \frac{m}{2}\mathrm{Dev}(Q)\right| \leq 8C'\sqrt{m} + C''\sqrt{m},$$

which completes the proof. □

Lemma D.27. *Let $Z_1, Z_2, \ldots$ be an infinite sequence of i.i.d. samples from P_Z. For each $n \geq 2$, let $Z_{n,i}$ be the Euclidean-nearest neighbour of Z_i among $\{Z_j : 1 \leq j \leq n, j \neq i$ and $Z_j \geq \tilde{Z}_i\}$. Then,*

(i) $Z_1 - Z_{n,1} \xrightarrow{P} 0$, *and*

(ii) there exists an universal constant C_0 such that for any measurable function $f : \mathbb{R} \to [0,\infty)$ and any n,

$$\mathbb{E}\left[f(Z_{n,1})\right] \leq C_0\,\mathbb{E}\left[f(Z_1)\right].$$

Proof. Let S be the support of P_Z, i.e., for any $z \in S$, any open interval containing z has strictly positive measure. Consequently, $\mathbb{P}_{P_Z}\{Z \in S\} = 1$. Take any $\epsilon > 0$, and consider the closed interval $[Z_1, Z_1 + \epsilon]$. Note that,

$$\mathbb{P}\left\{|Z_1 - Z_{n,1}| \geq \epsilon\right\} = \mathbb{E}\left[\mathbb{P}\left\{|Z_1 - Z_{n,1}| \geq \epsilon \mid Z_1\right\}\right] \leq \mathbb{E}\left[\left(1 - P_Z\left([Z_1, Z_1+\epsilon)\right)\right)^{n-1}\right]$$

Since $Z_1 \in S$ almost surely, $P_Z\big([Z_1, Z_1 + \epsilon)\big) > 0$ almost surely. Hence,

$$\lim_{n\to\infty} \mathbb{P}\left\{|Z_1 - Z_{n,1}| \geq \epsilon\right\} = 0,$$

which proves the first part of the result.

a calculation[7] similar to Azadkia and Chatterjee (2021, Lemma 11.5), □

[7]Note that, in Lemma 11.5 of Azadkia and Chatterjee (2021) they work with nearest Euclidean neighbour, while here we deal with nearest neighbour to the right. However, the argument follows in a very similar fashion.