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*

September 30, 2024

Abstract

We propose a powerful, nonparametric test of conditional independence $X \perp\!\!\!\perp Y \mid Z$ assuming only that X is stochastically increasing in Z. In particular, unlike recent work on conditional independence testing, our test does not require knowledge of the conditional distribution $X \mid Z$ beyond a shape constraint. Our method is a constrained form of permutation testing, affording the analyst a great deal of flexibility in designing a powerful test statistic.

1 Introduction

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

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

based on $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n) \stackrel{\text{iid}}{\sim} P$, where P is an unknown distribution on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Throughout, we write $P_{X|Z}$ and $P_{Y|Z}$ to denote these conditional distributions for the unknown joint distribution P.

In the case where the distribution of Z is nonatomic, 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; Kim et al., 2022). Existing work typically circumvent this hardness result by assuming one of the following:

(a) a parametric model, such as joint Gaussianity of (X, Y, Z),

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

(c) smoothness of the conditional distribution $P_{X|Z}$ (Shah and Peters, 2020).

1.1 Our contribution

In this work, we propose an alternative: we assume a shape constraint—specifically, a form of monotonicity—for the conditional distribution of $X \mid Z$. This approach falls outside of all of the categories (a)–(c) above. That is to say, we do not require any form of smoothness, known (or estimated) conditional distributions, nor parametric assumptions. The specific constraint we focus on in this paper is that X is stochastically increasing in Z.

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

$$P\{X \geq x|z\} \leq P\{X \geq x|z'\}.$$

We will often consider the case where the control variable Z is univariate, $\mathcal{Z} \subseteq \mathbb{R}$, under the usual total order $\leq$. For example, the variable Z may be a primary control variable that is a known risk factor associated with some adverse outcome X. Alternatively, if X is binary, $\mathcal{X} = \{0, 1\}$, and $\tilde{Z}$ represents some arbitrary covariates, then the propensity score, defined as $Z := \mathbb{P}\{X = 1 \mid \tilde{Z}\}$, automatically satisfies Assumption 1. Our framework also allows for multivariate $Z \in \mathbb{R}^d$. In this case, the most common partial order is the coordinate-wise order, where Assumption 1 arises naturally in reliability theory (Nevius, Proschan and Sethuraman, 1974; Barlow and Proschan, 1975). For example, if X indicates the presence of some medical condition, each covariate Z_j may represent some comorbidity or other risk factor associated with an outcome X.

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

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

Remark 1. *Rejecting the restricted null H_0^{ICI} tells us that either the conditional independence assumption H_0^{CI} is violated or Assumption 1 is violated (or it is a false rejection). Therefore, to reject the null hypothesis of conditional independence, we should apply our test in settings where Assumption 1 is well-motivated.*

1.2 Background: testing independence

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

Let $(X_i, Y_i, Z_i)_{i=1}^n$ be independent copies of (X, Y, Z), and let $\mathbf{X} = (X_i)_{i=1}^n$, $\mathbf{Y} = (Y_i)_{i=1}^n$ and $\mathbf{Z} = (Z_i)_{i=1}^n$. 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 $T : \mathcal{X}^n \to \mathbb{R}$. For any permutation $\sigma \in \mathcal{S}_n$, let $T_\sigma = T(\mathbf{X}^\sigma)$ denote the value of the statistic when the entries of $\mathbf{X}$ are permuted according to σ. Finally, define a p-value

$$p = \mathbb{P}\{T_\sigma \geq T \mid \mathbf{X}, \mathbf{Y}\} \qquad \text{where} \qquad \sigma \sim \text{Unif}(\mathcal{S}_n).$$

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. In our work, we set up 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 *much* larger null class H_0^{ICI}.

2 Methodology

In this section we give a general procedure 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 $X_i \leq X_j$ 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, allowing the analyst to specify, based on $\mathbf{Y}$ and $\mathbf{Z}$, pairs (i,j) such that $Z_i \preceq Z_j$ and then to use, for example, large differences $X_i - X_j$ as evidence against the null.

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

1. A sequence of L ordered pairs

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

of indices in $[n]$, 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 $\ell \in [L]$.

2. A sequence of L 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 $\Delta, \Delta' \geq 0$ such that $x, x', x + \Delta, x' - \Delta' \in \mathcal{X}$. Define

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

Next, the analyst computes the observed test statistic $T = T(\mathbf{X})$, and compares it to versions of T where pairs (i_ℓ, j_ℓ) are randomly swapped. That is, 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

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

Finally, define a p-value

$$p := \mathbb{P}_{\mathbf{s}}\big\{T_{\mathbf{s}} \geq T \;\big|\; \mathbf{X}, \mathbf{Y}, \mathbf{Z}\big\} \qquad \text{where} \qquad s_1, \ldots, s_L \overset{\text{iid}}{\sim} \text{Rademacher}. \tag{5}$$

Our method is quite flexible, in that the analyst may match any pairs (i_ℓ, j_ℓ) subject to monotonicity (2) and 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 $\mathbf{Y}, \mathbf{Z}$. However, the quality of the matches (i_ℓ, j_ℓ) and functions (ψ_ℓ) impact the power of our test; we discuss effective strategies for choosing a statistic T in Sections 3 and 4.

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}, conditional Type I error is controlled, i.e. $\mathbb{P}\{p \le \alpha \mid \mathbf{Y}, \mathbf{Z}\} \le \alpha$ for all $\alpha \in [0, 1]$.*

This result alone does not show that our method evades the impossibility result of Shah and Peters (2020), since it says nothing about the power of our test. However, we will see empirically that, for well-designed test statistics, the method has nontrivial power in simulations and real data experiments. Moreover, in Section 5, we perform power calculations showing our method have asymptotic power one under certain alternatives.

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. Define the anti-symmetrized kernel

$$\bar{\psi}_\ell(x, x') = \frac{1}{2}\big(\psi_\ell(x, x') - \psi_\ell(x', x)\big).$$

Our anti-monotonicity assumption (3) on ψ_ℓ means $\bar{\psi}_\ell$ is nondecreasing in its first coordinate, and nonincreasing in its second coordinate. Let $\bar{T}$ and $\bar{T}_{\mathbf{s}}$ be defined as for T and $T_{\mathbf{s}}$, but with the $\bar{\psi}_\ell$'s instead of with the ψ_ℓ's. Observe that the p-value p is unchanged if we use statistic $\bar{T}$ in place of T, since $\bar{\psi}_\ell(x, x') - \bar{\psi}_\ell(x', x) = \psi_\ell(x, x') - \psi_\ell(x', x)$.

Fix $\ell \in [L]$. Under H_0^{ICI} and given $\mathbf{Y}$ and $\mathbf{Z}$, X_{i_ℓ} and X_{j_ℓ} are independent with $X_{i_\ell} \le_{\text{st}} X_{j_\ell}$. We claim that, since s_ℓ is independent of the pair (X_{i_ℓ}, X_{j_ℓ}),

$$\bar{\psi}_\ell(X_{i_\ell}^{\mathbf{s}}, X_{j_\ell}^{\mathbf{s}}) \ge_{\text{st}} \bar{\psi}_\ell(X_{i_\ell}, X_{j_\ell}) \;\big|\; \mathbf{Y}, \mathbf{Z}. \tag{6}$$

To see this, note that for any t,

$$\mathbb{P}\left\{\bar{\psi}_\ell(X_{i_\ell}^{\mathbf{s}}, X_{j_\ell}^{\mathbf{s}}) \le t \,\middle|\, \mathbf{Y}, \mathbf{Z}\right\} = \frac{1}{2}\mathbb{P}\left\{\bar{\psi}_\ell(X_{i_\ell}, X_{j_\ell}) \le t \,\middle|\, \mathbf{Y}, \mathbf{Z}\right\} + \frac{1}{2}\mathbb{P}\left\{\bar{\psi}_\ell(X_{j_\ell}, X_{i_\ell}) \le t \,\middle|\, \mathbf{Y}, \mathbf{Z}\right\}.$$

Define the mixture conditional distribution

$$\bar{P}_{X|Z}^{(\ell)} = \frac{1}{2}P_{X|Z_{i_\ell}} + \frac{1}{2}P_{X|Z_{j_\ell}}.$$

By stochastic monotonicity (1), we can take X'_{i_ℓ}, X'_{j_ℓ} to be iid draws (given $\mathbf{Y}, \mathbf{Z}$) from the conditional distribution $\bar{P}_{X|Z}^{(\ell)}$ such that $X'_{i_\ell} \le X_{j_\ell}$ and $X_{i_\ell} \le X'_{j_\ell}$ almost surely. Thus

$$\begin{aligned}
\mathbb{P}\big\{\bar{\psi}_\ell(X_{j_\ell}, X_{i_\ell}) \le t \;\big|\; \mathbf{Y}, \mathbf{Z}\big\} &\le \mathbb{P}\big\{\bar{\psi}_\ell(X'_{i_\ell}, X'_{j_\ell}) \le t \;\big|\; \mathbf{Y}, \mathbf{Z}\big\} \\
&= \mathbb{P}\big\{\bar{\psi}_\ell(X'_{j_\ell}, X'_{i_\ell}) \le t \;\big|\; \mathbf{Y}, \mathbf{Z}\big\} \\
&\le \mathbb{P}\big\{\bar{\psi}_\ell(X_{i_\ell}, X_{j_\ell}) \le t \;\big|\; \mathbf{Y}, \mathbf{Z}\big\},
\end{aligned}$$

proving (6). Next, given $\mathbf{Y}$ and $\mathbf{Z}$, the terms are independent across $\ell \in [L]$, implying that

$$\bar{T}_{\mathbf{s}} \geq_{\text{st}} \bar{T} \;\big|\; \mathbf{Y}, \mathbf{Z}. \tag{7}$$

Finally, define $p(t) = \mathbb{P}\left\{\bar{T}_{\mathbf{s}} \geq t \,\middle|\, \mathbf{X}, \mathbf{Y}, \mathbf{Z}\right\}$, so $p = p(\bar{T})$. By (7) and monotonicity of $p(\cdot)$,

$$\mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} \leq \mathbb{P}\{p(\bar{T}_{\mathbf{s}}) \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} = \mathbb{E}\big[\mathbb{P}\{p(\bar{T}_{\mathbf{s}}) \leq \alpha \mid \mathbf{X}, \mathbf{Y}, \mathbf{Z}\} \;\big|\; \mathbf{Y}, \mathbf{Z}\big] \leq \alpha \qquad \square$$

The p-value constructed in (5) requires computing $T_{\mathbf{s}} = T(X^{\mathbf{s}})$ for all 2^L values of $\mathbf{s} \in \{-1, 1\}^L$, which may be computationally prohibitive for moderate or large L. In practice, we sample $\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(M)} \overset{\text{iid}}{\sim} \text{Unif}(\{-1, +1\}^L)$ and compute

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

The extra '1+' term in the numerator and denominator is necessary to make the Monte Carlo version of our test conservative (Davison and Hinkley, 1997; Phipson and Smyth, 2010).

Theorem 2. *Under H_0^{ICI}, for any $M \in \mathbb{N}$, $\hat{p}_M$ controls conditional Type I error, i.e. $\mathbb{P}\{\hat{p}_M \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} \leq \alpha$ for all α.*

Proof of Theorem 2. Let $B = \sum_{m=1}^{M} \mathbb{1}\{T_{\mathbf{s}^{(m)}} \geq T\}$, so that $\hat{p}_M \leq \alpha$ if and only if $B \in \{0, \ldots, \lfloor (1+M)\alpha - 1 \rfloor\}$. By (7), under H_0^{ICI}, B stochastically dominates the discrete uniform distribution on $\{0, \ldots, M\}$, conditionally on $\mathbf{Y}, \mathbf{Z}$. Thus

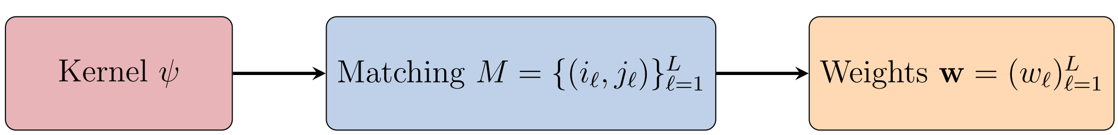
$$\mathbb{P}\{\hat{p}_M \leq \alpha | \mathbf{Y}, \mathbf{Z}\} = \mathbb{P}\left\{B \leq \lfloor (1+M)\alpha - 1 \rfloor | \mathbf{Y}, \mathbf{Z}\right\} \leq \frac{\lfloor (1+M)\alpha - 1 \rfloor + 1}{1 + M} \leq \alpha. \qquad \square$$

3 Designing the test

In this section, we construct a principled, powerful implementation of our test. Throughout, we assume that the test statistic $T(\mathbf{x})$ has the form

$$T(\mathbf{x}) = \sum_{\ell=1}^{L} w_\ell \, \psi(x_{i_\ell}, x_{j_\ell}). \tag{8}$$

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. With this simplification, designing a test statistic requires specifying the kernel ψ, deciding which pairs (i_ℓ, j_ℓ) get matched, and finally, how much weight w_ℓ to assign to each pair.



Our test controls the Type I error for any choice of ψ, M and $\mathbf{w}$, subject to the conditions 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. Throughout this section we adopt the convention that the kernel ψ is relatively simple, so we can see how M and $\mathbf{w}$ interact in greater depth and detail.

3.1 Specifying the kernel ψ

As a simple example, we can represent linear statistics $T(\mathbf{x}) = \mathbf{u}^\top \mathbf{x}$ in the form (8) above. To achieve this, we set $\psi(x, x') = x - x'$ and $w_\ell = u_{i_\ell} - u_{j_\ell}$. Then $T(\mathbf{x})$ satisfies the anti-monotonicity condition (3) as long as $u_{i_\ell} \geq u_{j_\ell}$ for any matched pair (i_ℓ, j_ℓ). If we believe that a linear model for $X \mid Y, Z$ is a good approximation to the true conditional distribution, then our test statistic (i.e., the weights $\mathbf{u}$) can be chosen to extract the coefficient of Y in the least squares regression. We remark that choosing such a test statistic is by no means implying an assumption that the linear model is correct—it may be the case that $T(\mathbf{u}) = \mathbf{u}^\top \mathbf{x}$ has good power for distinguishing the null from the alternative even if a linear model does not hold.

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 the linear statistic T can be very sensitive to extreme values. We can ameliorate this sensitivity by using $\psi(x, x') = \mathrm{sign}(x - x')$ or $\psi(x, x') = (-K) \vee (x - x') \wedge K$ for some appropriate $K > 0$.

3.2 An asymptotic oracle test

In this section, we build intuition by sketching the asymptotics of our test. Let us consider any statistic T of the form (8). Throughout this section, the kernel ψ is a fixed, anti-symmetric 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 CLT approximation gives, for large L, that

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

and where $\bar{\Phi}$ denotes the standard Gaussian survivor function. The above approximation holds for fixed $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ and relies only on the CLT approximation for a weighted sum of L independent Rademacher variables $\mathbf{s}$.

Next, we consider the distribution of $\bar{\Phi}(\hat{T})$ given $\mathbf{Y}$ and $\mathbf{Z}$. Under slightly stronger assumptions, we roughly have

$$\hat{T} \mid \mathbf{Y}, \mathbf{Z} \,\dot{\sim}\, \mathcal{N}\big(\hat{T}^*, 1\big) \qquad \text{where} \qquad \hat{T}^* := \frac{\sum_{\ell=1}^{L} w_\ell \, \mathbb{E}\left[\psi(X_{i_\ell}, X_{j_\ell}) | \mathbf{Y}, \mathbf{Z}\right]}{\sqrt{\sum_{\ell=1}^{L} w_\ell^2 \, \mathrm{Var}\left(\psi(X_{i_\ell}, X_{j_\ell}) | \mathbf{Y}, \mathbf{Z}\right)}}.$$

Up to these heuristics, we have shown that, given $\mathbf{Y}$ and $\mathbf{Z}$, the quantity $\hat{T}^*$ governs the power of our test. We can thus approximately optimize the power of our test over the choices of weights $\mathbf{w}$ and matching $M = \{(i_\ell, j_\ell)\}$.

3.2.1 The ideal weights (given a matching)

Given a matching $M = \{(i_\ell, j_\ell)\}$, we can directly optimize $\hat{T}^*$ over the weights $\mathbf{w} \geq 0$, yielding

$$w_\ell^* := \frac{\max\left\{\mathbb{E}\left[\psi(X_{i_\ell}, X_{j_\ell}) | \mathbf{Y}, \mathbf{Z}\right], 0\right\}}{\mathrm{Var}\left(\psi(X_{i_\ell}, X_{j_\ell}) | \mathbf{Y}, \mathbf{Z}\right)}. \tag{9}$$

3.2.2 The ideal matching

Once we fix the choice of weights to (9), the quantity $\hat{T}^*$ simplifies to

$$\hat{T}^* = \sqrt{\sum_{\ell=1}^{L} W^*_{i_\ell j_\ell}} \qquad \text{where} \qquad W^*_{ij} = \frac{\max\left\{\mathbb{E}\left[\psi(X_i, X_j)|\mathbf{Y}, \mathbf{Z}\right], 0\right\}^2}{\operatorname{Var}\left(\psi(X_i, X_j)|\mathbf{Y}, \mathbf{Z}\right)}. \tag{10}$$

We thus maximize $\hat{T}^*$ by solving the following maximum-weight matching problem:

$$M^* \in \underset{\text{Matchings } M \text{ on } [n]}{\operatorname{argmax}} \sum_{(i,j)\in M} W^*_{ij} \mathbf{1}\{Z_i \preceq Z_j\}. \tag{11}$$

3.3 Regression-based, maximum-weight matching

One strategy for designing a test is to imitate the asymptotic oracle. In order to estimate the conditional moments of ψ, we assume that we have access to an auxiliary set of $\tilde{n}$ observations $\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}, \tilde{\mathbf{Z}}$. Using this auxiliary data we estimate the $n \times n$ matrices (E^*_{ij}) and (V^*_{ij}), where

$$\begin{aligned} E^*_{ij} &:= \mathbb{E}\left[\psi(X_i, X_j) \mid \mathbf{Y}, \mathbf{Z}\right], \\ V^*_{ij} &:= \operatorname{Var}\left[\psi(X_i, X_j) \mid \mathbf{Y}, \mathbf{Z}\right]. \end{aligned} \tag{12}$$

Let $\hat{E}$ and $\hat{V}$ denote the estimated matrices. We solve for the optimal matching:

$$\hat{M} \in \underset{\text{Matchings } M \text{ on } [n]}{\operatorname{argmax}} \sum_{(i,j)\in M} \frac{\max\{\hat{E}_{ij}, 0\}^2}{\hat{V}_{ij}} \mathbf{1}\{Z_i \preceq Z_j\}.$$

Finally, we enumerate the matches $\hat{M} = \{(i_1, j_1), \ldots, (i_L, j_L)\}$ and define weights

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

For example, for linear statistics $\psi(x, x') = x - x'$, our regression-based procedure requires us to estimate the first two moments of $X_i \mid Y_i, Z_i$ for each i. Alternatively, if $\psi(x, x') = \operatorname{sign}(x - x')$, then we want to estimate $\mathbb{P}\left\{X_i > X_j \mid \mathbf{Y}, \mathbf{Z}\right\}$. The flexibility of our test allows us to model these quantities however we like, e.g. with our preferred state-of-the-art machine learning method.

The maximum-weight matching $\hat{M}$ can be computed in polynomial time (Edmonds, 1965; Duan and Pettie, 2014). Specifically, if $m = \#\{(i,j) : Z_i \preceq Z_j\}$, can be computed in time $O(mn + n^2 \log n)$ using an algorithm of Gabow (1985). For our experiments in Sections 6 and 7, we use the Python package `networkx` (Hagberg, Swart and S Chult, 2008), which uses the Blossom algorithm (Edmonds, 1965) and runs in time $O(n^3)$.

4 Heuristic matching strategies

We now discuss simple alternative approaches for the one-dimensional setting ($\mathcal{Z} \subseteq \mathbb{R}$) that do not require accurate estimation of the first two moments of $\psi(X_i, X_j)$. Our heuristic methods apply to the linear case $\psi(x, x') = x - x'$. As we will see in our experiments, these heuristic approaches may be preferable if we do not have a quality working model of the conditional distribution $X \mid Y, Z$ or if maximum-weight matching is computationally prohibitive.

4.1 Simple weighting scheme

To motivate this approach, suppose that matched pairs (i_ℓ, j_ℓ) are constrained such that $Z_{i_\ell} \approx Z_{j_\ell}$ for all ℓ. Furthermore, for simplicity, we assume that the conditional mean function $\mu^*(y,z) := \mathbb{E}[X|Y=y, Z=z]$ is *linear* in its first argument

$$\mu^*(y,z) = \beta^* y + \mu_Z^*(z), \tag{13}$$

and that the conditional variance σ^{*2} is constant. In this case, the ideal weights w_ℓ^* in (9) are approximately $\frac{\beta^*}{\sigma^{*2}}(Y_{i_\ell} - Y_{j_\ell})$. Crucially, our test is invariant to rescaling the weights—that is, we do not need to know $\frac{\beta^*}{\sigma^{*2}}$—so we instead set

$$\hat{w}_\ell := Y_{i_\ell} - Y_{j_\ell}.$$

In this case, the constraints (2) and (3) mean that the pair of observations (Y_{i_ℓ}, Z_{i_ℓ}) and (Y_{j_ℓ}, Z_{j_ℓ}) must be discordant.

4.2 Neighbor matching

Recall that we require (1) $Z_{i_\ell} \approx Z_{j_\ell}$, (2) $Z_{i_\ell} \leq Z_{j_\ell}$ and (3) $Y_{i_\ell} \geq Y_{j_\ell}$. A naïve matching strategy, then, is to sort (Y_i, Z_i) according to the Z-values in ascending order $Z_{(1)} \leq \cdots \leq Z_{(n)}$, matching $(2j, 2j+1)$ if $Y_{(2j)} \geq Y_{(2j+1)}$ for $j \in \{1, \ldots, \lfloor n/2 \rfloor\}$, and otherwise leaving $2j$ and $2j+1$ unmatched.

An obvious limitation of this naïve matching strategy is that many pairs can fail to have $Y_{(2j)} \geq Y_{(2j+1)}$ just by chance. For instance, even if $Z_{(2j)}$ and $Z_{(2j+1)}$ are extremely close, we might expect $Y_{(2j)}$ and $Y_{(2j+1)}$ to be roughly iid (given $\mathbf{Z}$), so $\mathrm{sign}(Y_{(2j)} - Y_{(2j+1)})$ is roughly a Rademacher variable. In particular, we are throwing out many observations that might be matched with other nearby observations.

4.3 Cross-bin matching

In order to increase to total number of matches L, we propose *cross-bin matching*. Unlike neighbor matching, where immediate neighbors in Z are matched, we bin the observations according to Z and match observations in adjacent bins. Figure 1 illustrates the approach.

Specifically, we sort the observations by Z values in increasing order and partition them into K bins such that approximately n/K observations fall in each bin. Formally, $A_1, \ldots, A_K$ form a partition of the $[n]$, where

$$\begin{aligned} A_1 = \{1, \ldots, \lfloor n/K \rfloor\}, \ldots, A_k = \{(k-1)\lfloor n/K \rfloor, \ldots, k\lfloor n/K \rfloor\}, \ldots \\ , A_K = \{(K-1)\lfloor n/K \rfloor, \ldots, n\}. \end{aligned}$$

Now given the binning, for any $k \in [K]$, we define the set of positive and negative samples as

$$\begin{aligned} J_k^+ = \{i \in A_k,\ Y_i \geq \mathrm{Median}(\{Y_i : i \in A_k\})\}, \\ J_k^- = \{i \in A_k,\ Y_i < \mathrm{Median}(\{Y_i : i \in A_k\})\} \end{aligned}$$

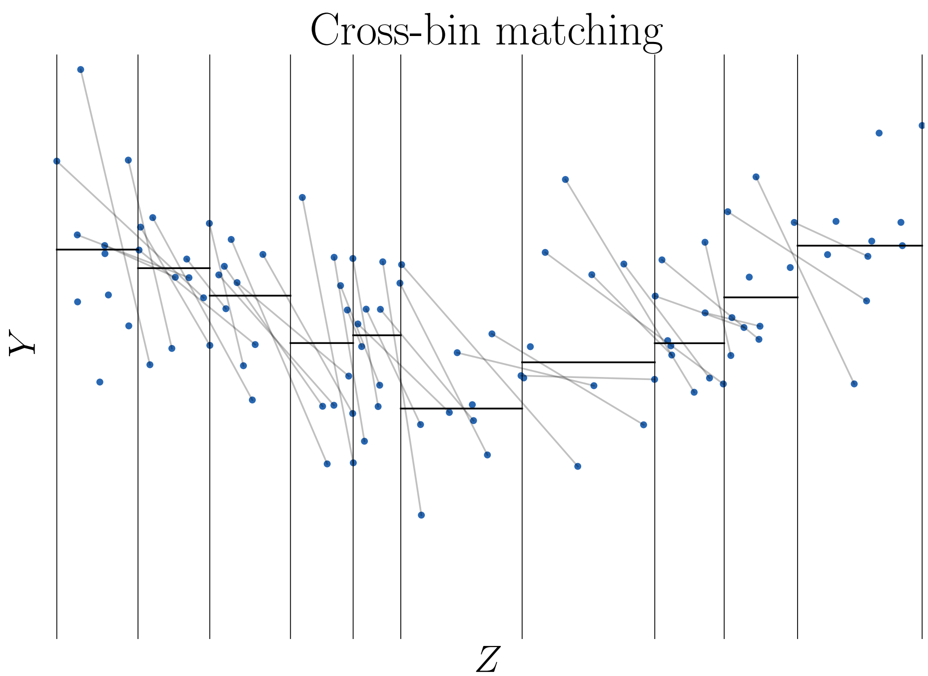


Figure 1: Demonstration of the cross-bin matching scheme described in Section 4.3.

respectively. Now, we order the observations in both the positive and negative set by increasing and decreasing order in Y. Formally, $i_{k,1}, i_{k,2}, \ldots$ are indices in J_k^+ such that $Y_{i_{k,1}} \geq Y_{i_{k,2}} \geq \ldots$ and similarly, $j_{k,1}, j_{k,2}, \ldots$ are indices in J_k^- such that $Y_{j_{k,1}} \leq Y_{j_{k,2}} \leq \ldots$ Finally, we take a greedy approach to maximize $\sum_{(i,j)\in M}(Y_i - Y_j)^2$ and for $k \in [K-1]$, we diagonally match the positive samples from left bin with the negative samples from each bin i.e. for any given k we match $(i_{k,1}, j_{k+1,1}), (i_{k,2}, j_{k+1,2}), \ldots$ until either one of the positive or negative samples are matched completely or for some ℓ, $Y_{i_{k,\ell}} < Y_{j_{k+1,\ell}}$. Observe, by the choice of binning, for matched samples it always holds that

$$Z_{i_{k,\ell}} \leq Z_{j_{k+1,\ell}} \quad \text{and} \quad Y_{i_{k,\ell}} \geq Y_{j_{k+1,\ell}}$$

i.e. we satisfy the anti-monotonicity constraint if we choose our weights for matched samples to be simply their respective Y value.

5 Power analysis

In this section, we study the power of our testing procedure under the following general model. We assume that we are given triples $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = (X_i, Y_i, Z_i)_{i\in[n]} \overset{\text{iid}}{\sim} P$, drawn according to the model

$$\mathbf{X} = \mu(\mathbf{Y}, \mathbf{Z}) + \boldsymbol{\zeta}, \tag{14}$$

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 σ^2. Throughout this section, we further assume that the functions (ψ_ℓ) employed in our testing procedure are linear in both coordinates. In order to state the power guarantees under this signal plus noise model, we first introduce some notation.

By a *test function*, we mean a measurable function $\phi : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to [0, 1]$, and say that ϕ is a *valid* test or *controls the Type I error* over $\mathcal{H}_0^{\mathrm{ICI}}$ if

$$\sup_{P \in \mathcal{H}_0^{\mathrm{ICI}}} \mathbb{E}_P \left[\phi(\mathbf{X}, \mathbf{Y}, \mathbf{Z})\right] \leq \alpha. \tag{15}$$

We denote by $\mathrm{Mon}(\mathcal{Z})$ the set of all measurable functions that are monotonic on the partially ordered set $(\mathcal{Z}, \preceq)$. Further, we define

$$\mathrm{ISS}_n = \inf_{g \in \mathrm{Mon}(\mathcal{Z})} \mathbb{E}_{P_{Y,Z}^n} \left[\|\mu(\mathbf{Y}, \mathbf{Z}) - g(\mathbf{Z})\|_2\right], \quad \widehat{\mathrm{ISS}}_n = \inf_{g \in \mathrm{Mon}(\mathcal{Z})} \|\mu(\mathbf{Y}, \mathbf{Z}) - g(\mathbf{Z})\|_2, \tag{16}$$

referring to them as the oracle and empirical *isotonic signal strength* respectively. We denote the corresponding oracle and empirical L_2 projections as μ_{ISO} and $\widehat{\mu}_{\mathrm{ISO}}$.

We will shortly demonstrate that the ability of our test procedure to distinguish the alternatives from the class of null models $\mathcal{H}_0^{\mathrm{ICI}}$ is governed by the empirical isotonic signal strength. In fact, in Section 5.2, we show that this connection is not exclusive to our testing procedure: the oracle ISS_n serves as a fundamental measure for characterizing the hardness of any valid test procedure. With this hardness result in mind, ISS_n can be interpreted as the *distance* of the model P from null class $\mathcal{H}_0^{\mathrm{ICI}}$. If ISS_n is too small, then the model P becomes essentially indistinguishable from $\mathcal{H}_0^{\mathrm{ICI}}$, implying that we can barely outperform the trivial testing procedure that ignores the data and rejects the null hypothesis randomly with probability α. Conversely, once $\widehat{\mathrm{ISS}}_n$ exceeds a threshold that we determine, the power of the test can approach 1, especially with appropriately designed matching schemes.

The organization of the rest of the section is as follows: in Section 5.1, we state asymptotic upper and lower bounds on the power of our test, optimized over all matching schemes, and conditioned on $(\mathbf{Y}, \mathbf{Z})$ under the general class of alternatives given in (14). While these power guarantees correspond to the max-weight matching, which relies on the oracle knowledge of μ, we show in Section 5.1.1 that even with an consistent estimate $\hat{\mu}$, we can derive very similar power guarantees. Next, in Section 5.2, we demonstrate how under suitable model assumptions, the oracle ISS_n characterizes the hardness of this test, and this connection is not limited to our proposed test procedure. Finally, in Section 5.3, we specialize our power guarantees to the special case of partially linear Gaussian models, and show that even without knowledge of the oracle μ, we can achieve near-optimal power guarantees with some of the natural choices of matching schemes proposed in Section 4.

5.1 ISS dictates the power of our testing procedure

To analyze the best-case performance of our testing procedure, we study the conditional power $\mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\}$ of max-weight matching from Section 3.2, while the power analysis for general matching schemes is deferred to Theorem A.10 in the appendix. We assume that the distributions $P_{Y,Z}$ and P_ζ in (14) do not depend on sample size n, while the regression function $\mu(\cdot,\cdot)$ does, but we suppress the dependence of n in our notation for simplicity. A more detailed finite-sample power analysis is presented in the Appendix A.2, where the exact dependence of the power on ISS_n and other related terms is discussed.

Theorem 3. *Suppose that* $(\|\mu(\mathbf{Y},\mathbf{Z})\|_\infty \vee \|\mu(\mathbf{Y},\mathbf{Z})\|_\infty^4)/\widehat{\mathrm{ISS}}_n = o_P(1)$. *Then, under the model* (14)*, the conditional power of max-weight matching satisfies*

$$\Phi\left(\frac{\widehat{\mathrm{ISS}}_n}{\sqrt{2}\sigma} - \bar{\Phi}^{-1}(\alpha)\right) - o_P(1) \le \mathbb{P}\{p \le \alpha \mid \mathbf{Y},\mathbf{Z}\} \le \Phi\left(\frac{\widehat{\mathrm{ISS}}_n}{\sigma} - \bar{\Phi}^{-1}(\alpha)\right) + o_P(1). \tag{17}$$

In both the upper and lower bounds, the signal-to-noise ratio $\widehat{\mathrm{ISS}}_n/\sigma$ dictates the dominant term, and hence the power of our test procedure. These bounds match up to a factor of $\sqrt{2}$ in the signal-to-noise ratio. The reason for this asymmetry is explained in Section 5.3.

It is important to note that the power result requires that $(\|\mu(\mathbf{Y},\mathbf{Z})\|_\infty \vee \|\mu(\mathbf{Y},\mathbf{Z})\|_\infty^4)/\widehat{\mathrm{ISS}}_n$ is $o_P(1)$, which is rather necessary and unavoidable. The above term quantifies the concentration of the signal $\mu(Y,Z)$ across samples, and the assumption ensures that the signal is not accumulated within a small subset of sample points. Otherwise, the power of our test will only depend on swaps of a small subset of matched pairs, leading to low power in practice. The exact dependence of this term on the upper and lower bounds is stated in Theorem A.11.

5.1.1 Optimal power guarantees with an estimate of μ

We further note that the power guarantees for max-weight matching in Theorem 3 requires the oracle knowledge of μ, which is not accessible in practice. A natural solution for bypassing this issue is data splitting — i.e., we start with learning an estimate $\hat{\mu}$ from one random split of the data, and then implement the max-weight matching with this estimated $\hat{\mu}$ on the remaining split. It is important to note that learning $\hat{\mu}$ on a different split makes sure that the calculated weights are independent of $\mathbf{X}$, which is crucial for validity of our test. While the above outlined data splitting is more accurate, for the sake of simplicity, while stating the following result we assume that we have a prior data where we can learn $\hat{\mu}$. Finally, under suitable consistency assumptions (details are given below) on $\hat{\mu}$, we can recover the similar power guarantees as in Theorem 3.

Theorem 4. *Consider the setting from Theorem 3. Suppose we use max-weight matching with an estimate $\hat{\mu}$ learnt on a prior data, where $\hat{\mu}$ satisfies*

$$\|\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \mu(\mathbf{Y},\mathbf{Z})\|_2/\widehat{\mathrm{ISS}}_n, \quad \|\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \mu(\mathbf{Y},\mathbf{Z})\|_\infty^4/\widehat{\mathrm{ISS}}_n \ \text{ is } \ o_P(1).$$

Then, the conditional power satisfies (17)*.*

To summarize, with or without the oracle knowledge of μ, the optimal power of our test is governed by $\widehat{\mathrm{ISS}}_n$. As briefly mentioned in the introduction of this section, the connection of ISS with power is not limited to the particular testing procedure, we propose. In fact, under additional model assumptions, with any valid testing procedure, it is impossible to attain non-trivial power if the oracle ISS_n is too small. We state and prove this in the following subsection.

5.2 ISS characterizes hardness of testing the null $\mathcal{H}_0^{\text{st}}$

In this subsection, we aim to characterize the hardness of testing the null $\mathcal{H}_0^{\text{st}}$ using the notion of ISS. In particular, via the oracle ISS_n under certain model assumptions, we identify the alternative models that are indistinguishable from the class of null models using any valid testing procedure i.e. any test Φ that satisfies (15). Towards this goal, we start with a simple total-variation calculation that gives a naive upper bound on power function, uniformly valid for any test procedure.

Lemma 5. *Fix $\alpha \in (0,1)$. For any test ϕ that satisfies (15), and for any distribution $P_{X,Y,Z}$,*

$$\mathbb{E}_{P_{X,Y,Z}}[\phi(\mathbf{X},\mathbf{Y},\mathbf{Z})] \le \alpha + \inf_{Q_{X,Y,Z}\in\mathcal{H}_0^{\text{st}}} \mathrm{d}_{\text{TV}}\left(P^n_{X,Y,Z}, Q^n_{X,Y,Z}\right).$$

The offset total variation term can be read as the distance of $P_{X,Y,Z}$ from null class $\mathcal{H}_0^{\text{st}}$, 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 model restrictions on P, we can come up with interpretable upper bounds for the same. Few such examples are listed below.

Gaussian setting: In the first example, we consider a special class of Gaussian alternatives, which is given by (14) where we assume that $P_\zeta = \mathcal{N}(0,\sigma^2)$ for some $\sigma > 0$. In this special case, ISS gives a meaningful upper bound on the offset total-variation term, up to a constant.

Corollary 1. *Suppose, $P_\zeta = \mathcal{N}(0,\sigma^2)$ and $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ satisfy (14). Then, for any test ϕ that satisfies (15), we have*

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

Proof of Corollary 1. By Lemma 5, it is enough to argue that

$$\inf_{Q_{X,Y,Z}\in\mathcal{H}_0^{\text{st}}} \mathrm{d}_{\text{TV}}\left(P^n_{X,Y,Z}, Q^n_{X,Y,Z}\right) \le \frac{\text{ISS}_n}{2\sigma}.$$

Consider $\mu_{\text{ISO}} \in \text{Mon}(\mathcal{Z})$. Similar to (14), define a model $Q_{X,Y,Z}$ as

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

By definition, $Q_{X,Y,Z} \in \mathcal{H}_0^{\text{st}}$ and further,

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

This proves the result. $\square$

As an immediate consequence, we note that it is impossible to achieve non-trivial power with any valid testing procedure when P_ζ is gaussian, and the ISS for alternative model is negligible. Next, we consider an example where $P_{X|Y,Z}$ is binary and we prove a similar characterization of power using ISS.

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

$$X_i \sim \mathrm{Ber}(\mu(Y_i, Z_i)), \quad (Y_1, Z_1), \cdots, (Y_n, Z_n) \overset{\text{iid}}{\sim} P_{Y,Z}. \tag{18}$$

We will show that ISS_n leads to a very interpretable upper bound on the power of our test, as long as $\mu(Y, Z)$ is away from the extremities i.e., 0 or 1. Below we state and prove the result.

Corollary 2. *Suppose* $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ *satisfy* (18) *where* $\mu(Y, Z) \in (\epsilon, 1-\epsilon)$ *almost surely for some* ϵ*. Then, for any test* ϕ *that satisfies* (15)*, we have*

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

Proof of Corollary 2. Similar to the proof of Corollary 1, it is enough to argue that

$$\inf_{Q_{X,Y,Z} \in \mathcal{H}_0^{\text{st}}} \mathrm{d_{TV}}\left(P^n_{X,Y,Z}, Q^n_{X,Y,Z}\right) \leq \left(\frac{1}{\epsilon(1-\epsilon)}\right)^{1/2} \mathrm{ISS}_n.$$

Now, consider the model $Q_{X,Y,Z} \in \mathcal{H}_0^{\text{st}}$ given by

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

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

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

Then, the inner total variation term can be upper bounded by the corresponding Hellinger distance as follows.

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

Now, for squared Hellinger distance, $H^2(P_1 \times \cdots \times P_k, Q_1 \times \cdots Q_k)$ is bounded by $\sum_{i=1}^k H^2(P_i, Q_i)$ and then, using Lemma 6 (stated below), the squared Hellinger distance can be further upper bounded by

$$\sum_{i=1}^n H^2\left(\mathrm{Ber}(\mu(Y_i, Z_i)), \mathrm{Ber}(\mu_{\mathrm{ISO}}(Z_i))\right) \leq \sum_{i=1}^n \frac{(\mu(Y_i, Z_i) - \mu_{\mathrm{ISO}}(Z_i))^2}{2\mu(Y_i, Z_i)(1 - \mu(Y_i, Z_i))}$$

Finally, since $\mu(Y, Z) \in (\epsilon, 1-\epsilon)$ almost surely, the final bound from above can be further upper bounded by $\frac{1}{\epsilon(1-\epsilon)} \cdot \|\mu(\mathbf{Y}, \mathbf{Z}) - \mu_{\mathrm{ISO}}(\mathbf{Z})\|_2^2$. This concludes the proof. □

Lemma 6. *For any* $p, q \in [0, 1]$*, it holds that*

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

While for both examples, the oracle ISS_n is used to bound power, in (17) the conditional power of our test is controlled by the empirical variant $\widehat{\mathrm{ISS}}_n$. While there is always a gap, under 'nice' settings both the oracle and empirical ISS terms will lie very close to each other. This justifies the nomenclature of 'Isotonic signal strength', at least for the gaussian and bernoulli examples. Beyond these settings, we can possibly define some other variants of oracle ISS_n (e.g., as we see in the analysis in Appendix B, we can consider $\|\mu(\mathbf{Y}, \mathbf{Z}) - \widetilde{\mu}_{\mathrm{ISO}}(\mathbf{Z})\|_2$ where $\widetilde{\mu}_{\mathrm{ISO}}(\mathbf{Z})$ is the isotonic L_1 projection.) to achieve similar upper bounds on power.

5.3 Near-optimal power guarantees without oracle knowledge of μ, and without sample-splitting

In earlier sections, we have established maximal power guarantees of our test procedure via the max-weight matching scheme; these are valid even without the oracle knowledge of μ with the cost of some additional consistency assumptions. In this subsection, we show that we can achieve similar non-trivial power guarantees with more natural choices of matching schemes. To prove such results, we specialize to the following model class, and show that some of these natural choices from Section 4 lead to near-optimal power guarantees.

Consider the class of partially linear gaussian models given by

$$(X, Y, Z) \text{ satisfy (14) with } P_\zeta = \mathcal{N}(0, \sigma^2),\ \mu(Y, Z) = \mu_0(Z) + \beta\, Y \tag{19}$$

for some $\mu_0 \in \mathrm{Mon}(\mathcal{Z})$ and some $\beta \in \mathbb{R}_{\geq 0}$. Without loss of generality, we further assume $\mathbb{E}\ Y = 0$. Observe that $\mathcal{H}_0^{\mathrm{st}}$ is satisfied if and only if $\beta = 0$. Thus, after parametrizing this model class with a single parameter β, we can consider a much simpler testing problem: $\mathcal{H}_0 : \beta = 0$ against $\mathcal{H}_1 : \beta > 0$. Under this specific model class, $\mathbf{Y}^T(\mathbf{X} - \mu(\mathbf{Z}))$ is a sufficient statistic for this parametric family indexed by β with

$$\mathbf{Y}^T(\mathbf{X} - \mu_0(\mathbf{Z})) \mid \mathbf{Y}, \mathbf{Z} \sim \mathcal{N}\big(\beta\|\mathbf{Y}\|_2^2, \sigma^2\|\mathbf{Y}\|_2^2\big).$$

Hence, an oracle test for $\mathcal{H}_0 : \beta = 0$ can be built by thresholding this sufficient statistic at $\sigma \cdot \|\mathbf{Y}\|_2 \cdot \Phi^{-1}(1 - \alpha)$, where Φ denotes the standard normal distribution function, and the power of this oracle test is given by

$$\mathbb{P}\big\{\mathcal{N}(\beta\|\mathbf{Y}\|_2^2, \|\mathbf{Y}\|_2^2) > \sigma \cdot \|\mathbf{Y}\|_2 \cdot \bar{\Phi}^{-1}(\alpha)\big\} = \Phi\left((\beta/\sigma) \cdot \|\mathbf{Y}\|_2 - \bar{\Phi}^{-1}(\alpha)\right).$$

Observe by weak law of large numbers, $\big|\|\mathbf{Y}\|_2 - \big(\mathrm{Var}(Y)\big)^{1/2}\big| = \mathrm{o}_P(1)$, and hence for large sample sizes, the oracle power can be computed as $\Phi\left((\beta/\sigma) \cdot \big(\mathrm{Var}(Y)\big)^{1/2} - \bar{\Phi}^{-1}(\alpha)\right) \pm \mathrm{o}_P(1)$. We would shortly present that matching schemes like immediate-neighbor and cross-bin matching depicts a very similar behavior in power.

In order to state these results, we go back to the asymptotic framework from Section 5.1 where now, the dependence of n on $\mu(Y, Z)$ is only through β and the function μ_0 remains fixed (to make this explicit, we write β_n now onwards). Under this framework and the model (19), we first note the following bounds on oracle ISS_n.

Lemma 7. *Under the model class (19), it holds that*

$$\sqrt{n}\beta_n\ \left(\mathbb{E}\big[\mathrm{Var}(Y \mid Z)\big]\right)^{1/2} \lessapprox\ \mathrm{ISS}_n \leq \sqrt{n}\beta_n\big(\mathrm{Var}(Y)\big)^{1/2}$$

Proof.

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

$$\text{ISS} = \mathbb{E}_{P_{Y,Z}}\left[\|\mu(\mathbf{Y},\mathbf{Z}) - \mu_{\text{ISO}}(\mathbf{Z})\|_2\right] \geq \mathbb{E}_{P_{Y,Z}}\left[\|\mu(\mathbf{Y},\mathbf{Z}) - \mathbb{E}\big[\mu(\mathbf{Y},\mathbf{Z}) \mid \mathbf{Z}\big]\|_2\right]$$
$$= \mathbb{E}_{P_{Y,Z}}\left[\|\beta_n\,\big(\mathbf{Y} - \mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]\big)\|_2\right] \gtrapprox \sqrt{n}\beta_n\,\big(\mathbb{E}\big[\mathrm{Var}(Y \mid Z)\big]\big)^{1/2}.$$

□

A direct implication of this lemma is that the threshold $1/\sqrt{n}$ determines a phase-transition of the power analysis. More specifically, if $\beta_n \ll 1/\sqrt{n}$, then by Corollary 1, every valid testing procedure is powerless. Hence, to ensure that we at least have non-trivial power, we will assume β_n scales as $\omega(1/\sqrt{n})$. Further, after scaling with $\sqrt{n}\beta_n$, the difference in between the upper and lower bound is driven by $\mathrm{Var}\big(\mathbb{E}[Y \mid Z]\big)$ which implies that if the conditional mean function is smooth, then upper and lower bounds are in fact very close.

Further Lemma 7 implies that under the model class (19) and mild assumptions stated in Theorem 3, the conditional power for max-weight matching satisfies

$$\Phi\left(\sqrt{n}\beta_n\left\{\frac{\mathbb{E}\big[\mathrm{Var}(Y \mid Z)\big]}{2\sigma^2}\right\}^{1/2} - \bar{\Phi}^{-1}(\alpha)\right) - o_P(1) \;\leq \mathbb{P}\{p \leq \alpha \mid \mathbf{Y},\mathbf{Z}\}$$
$$\leq \Phi\left(\sqrt{n}\beta_n\left\{\frac{\mathrm{Var}(Y)}{2\sigma^2}\right\}^{1/2} - \bar{\Phi}^{-1}(\alpha)\right) + o_P(1). \tag{20}$$

Below we show that other heuristic matching schemes from Section 4 match these power guarantees, taking advantage of the linearity in this model class. More precisely, if for our matched pairs $Z_{i_\ell} \approx Z_{j_\ell}$, then we have

$$\mu(Y_{i_\ell}, Z_{i_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell}) \approx \beta_n\big(Y_{i_\ell} - Y_{j_\ell}\big) \propto \big(Y_{i_\ell} - Y_{j_\ell}\big).$$

Since our test procedure depends on weights only through their 'in-pair differences', we can simply choose our weights to be the corresponding Y values without compromising much in power. In particular, the asymptotic power for neighbor matching matches the lower bound in (20) up to a constant. The result is formally stated below.

Theorem 8. *Let β_n be $\omega(1/\sqrt{n})$. Suppose Y is bounded, and $\mu_0(Z)$ is a sub-Gaussian random variable. Then, the power for immediate neighbor matching satisfies*

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

While this is a good starting point, immediate neighbor matching is inefficient since half of the sample points remain unmatched in this strategy. Cross-bin matching from Section 4 on the other hand improves in number of matched samples, and thus that is also reflected in power guarantees.

Theorem 9. *Consider the setting as in Theorem 8 and suppose, the number of bins i.e., K is chosen to satisfy $K = \omega_P(\sqrt{n})$. Then, under suitable smoothness assumptions (stated formally in Theorem A.13), the power for cross-bin matching satisfies*

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

Further, if $P_{Y|Z}$ is symmetric almost surely, then the power of cross-bin matching further satisfies

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

The first part of the result states that asymptotically, under no further model assumptions, the power of cross bin matching is no smaller than the lower bound from (20). However, with the additional assumption of symmetry of $P_{Y|Z}$, cross-bin matching can recover power as good as the upper bound in (20), when the conditional variance of Y given Z is constant across $\mathcal{Z}$-space or more particularly, when $\mathbb{E}[\mathrm{Var}(Y \mid Z)]$ is same as $\mathrm{Var}(Y)$.

Explaining the gap between the lower and upper bounds on power We see that under this specialized model, $\mathbb{E}\big[\mathrm{Var}(Y \mid Z)\big]/\sigma^2$ plays the role of signal-to-noise ratio in the power results stated above. Further, we rewrite the asymptotic power from Theorem 8 as

$$\Phi\left(\sqrt{n}\beta_n\left\{\frac{\mathbb{E}\big[\mathrm{Var}(Y \mid Z)\big]}{2 \cdot 2 \cdot \sigma^2}\right\}^{1/2} - \bar{\Phi}^{-1}(\alpha)\right),$$

and stress that the source of these two factors of 2 are different. Below, we explain the sources for each of the two in details.

- *Insufficient matched pairs:* For neighbor matching, a matched pair of consecutive observations contribute only if the anti-monotonicity holds. Since asymptotically, neighbours are almost distinguishable, the anti-monotonicity holds with probability $\approx 1/2$. As a result, half of the samples are discarded in the process of matching. This contributes to one of the factors of 2. Cross-bin matching, and max-weight matching however leaves only $\mathrm{o}(n)$ many samples unmatched, and that helps in recovering the gap.
- *Asymmetry in matching:* The other factor of 2 is rather unavoidable. We observe that for general matching schemes, in our general power calculations (Theorem A.10) the dominating term is governed by the term $\sum_\ell(\mu_{i_\ell} - \mu_{j_\ell})(w_{i_\ell} - w_{j_\ell})$ which is maximized when for matched pairs,

  $$w_{i_\ell} \approx \mu_{i_\ell}, w_{j_\ell} \approx \mu_{j_\ell}, \quad \text{and} \quad \mu_{i_\ell} \approx -\mu_{j_\ell}.$$

  Without further model assumptions, it is specially hard to guarantee $\mu_{i_\ell} \approx -\mu_{j_\ell}$, which contributes to the other factor of 2 in Theorem 8. This also explains the $\sqrt{2}$ in Theorem 3, and the factor of 2 in Theorem 9. Under the gaussian partial model, $\mu_{i_\ell} - \mu_{j_\ell} \approx Y_{i_\ell} - Y_{j_\ell}$ if the matched observations lie in vicinity in the $\mathcal{Z}$-space. Thus, the assumption of $P_{Y|Z}$ being symmetric recovers this factor of 2, as we note in Theorem 9.

6 Simulations

In this section, we evaluate the performance of our method on simulated data, and compare different matching strategies. For simplicity, we focus on the univariate case $\mathcal{Z} = \mathbb{R}$.

6.1 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 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 $\gamma \geq 0$, 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 $\mathbb{P}\{p \leq \alpha\}$ to approach α.

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

6.2 Power under alternatives

In Section 5.3 we show that our heuristic methods—neighbor matching and cross-bin matching—achieve asymptotic power one in the partial linear model (19) provided the signal β_n exceeds the detection threshold $n^{-1/2}$. In this section, we simulate 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 $\gamma = 1$, and cross-bin matching uniformly dominates neighbor matching because it allows us to make many more matches of similar quality. On the other hand, in Setting 2 we set $\gamma = 10$, 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. However, cross-bin matching still overtakes neighbor matching in sufficiently large samples. This is because the bin width decreases as n increases, so the quality of the cross-bin matches rivals that of the neighbor 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.

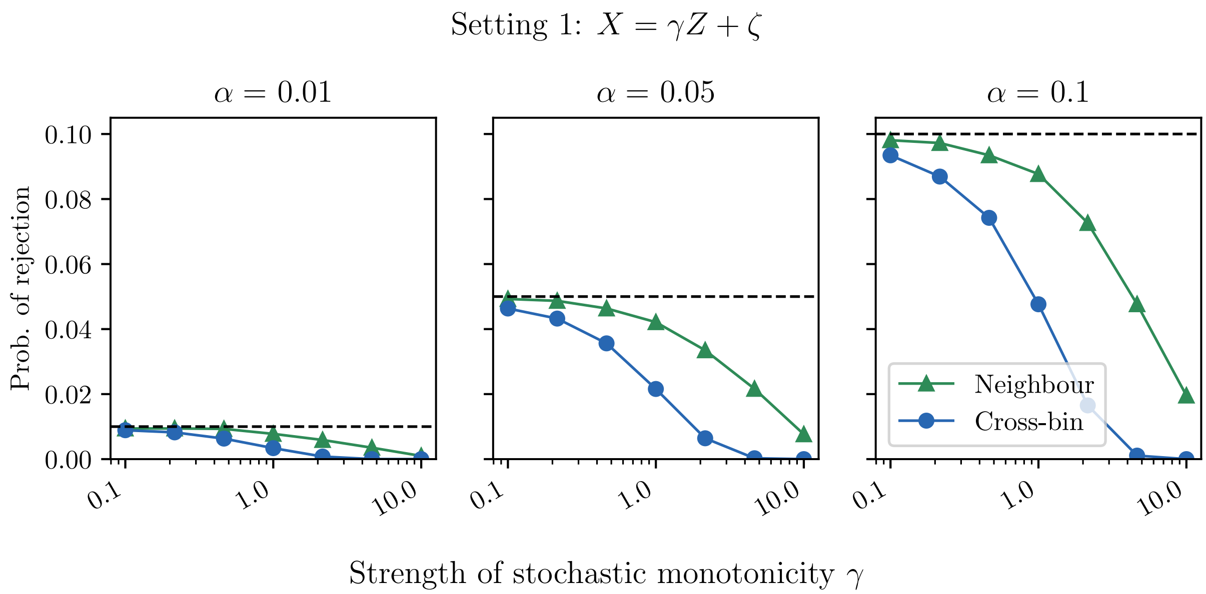
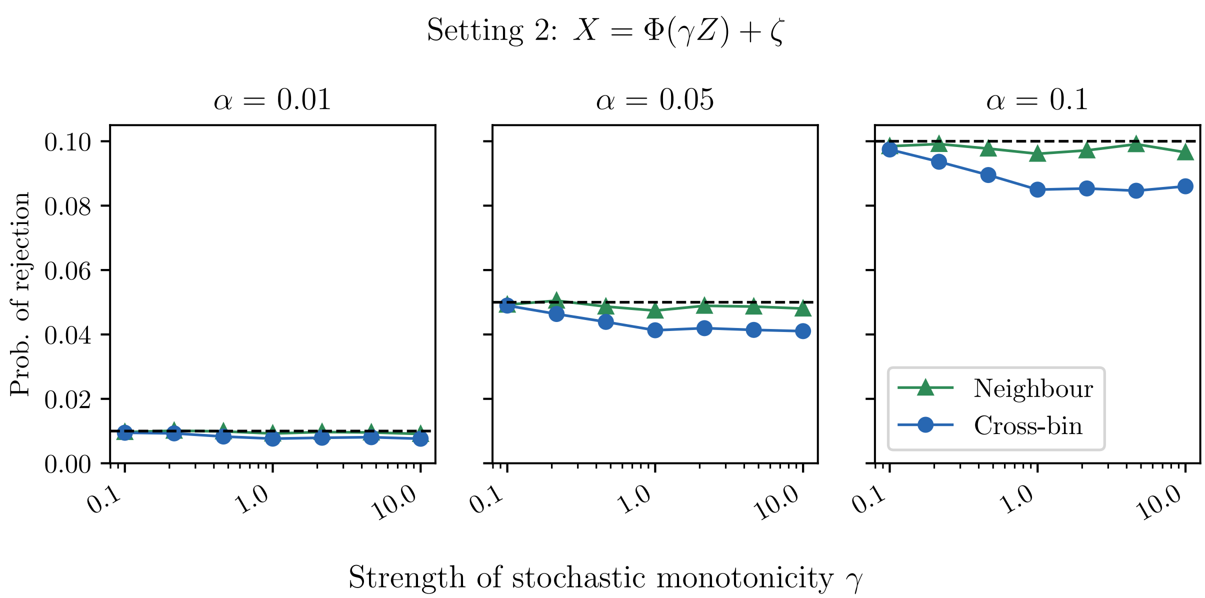


Figure 2: Simulation results illustrating Type I error control under the null H_0^{ICI} for two forms of the conditional mean $\mathbb{E}[X \mid Z]$. Each subplot shows the rejection probability, averaged over 10^6 simulation trials, as a function of the strength of stochastic monotonicity γ.



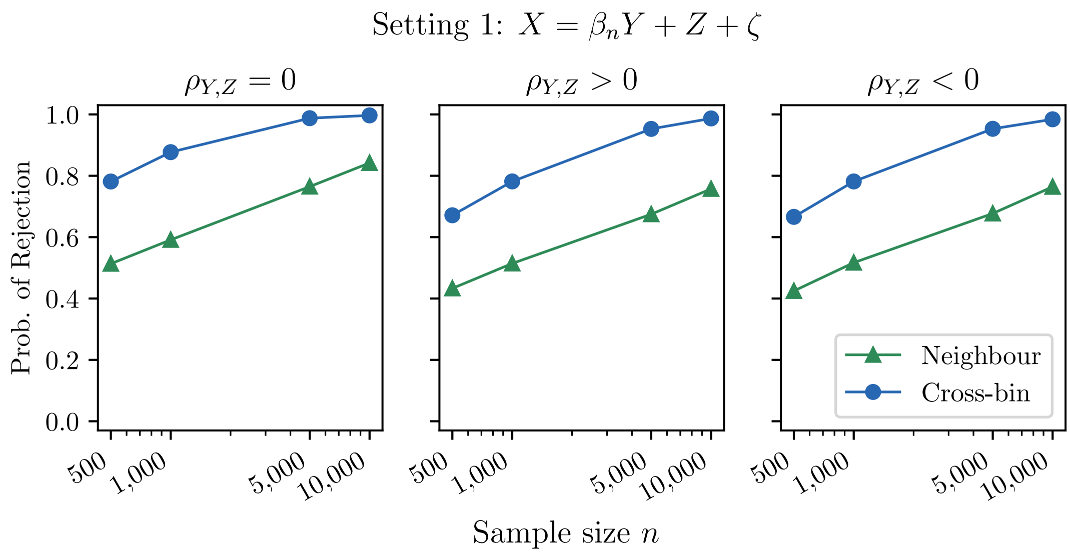
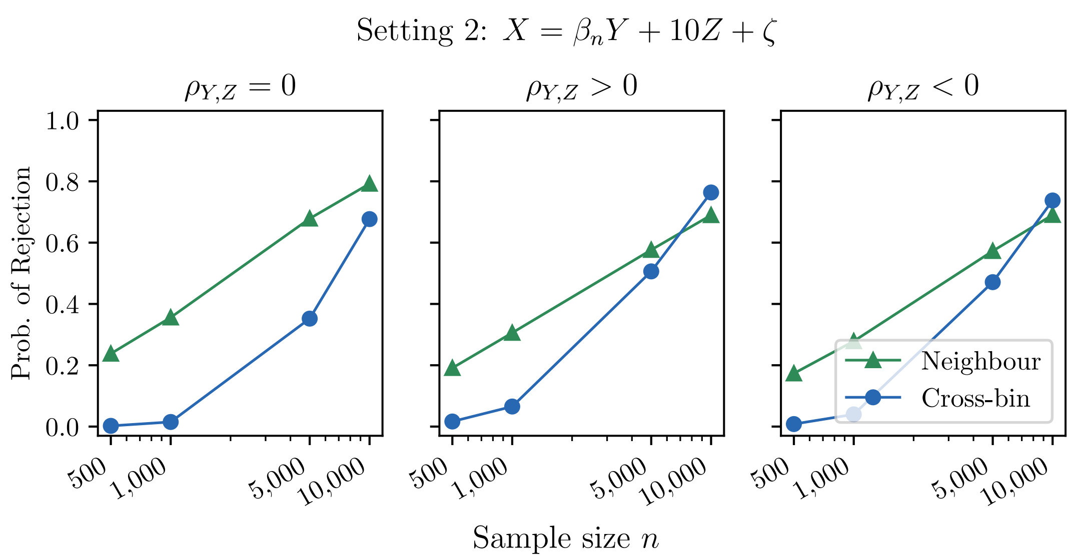


Figure 3: Simulation results demonstrating power for two alternatives at level $\alpha = 0.1$. Each subplot shows the rejection probability, averaged over 10^4 simulation trials, as a function of the sample size n. Columns correspond to different relationships between Y and Z. In each setting, X follows a Gaussian linear model with mean $\beta_n Y + \gamma Z$, where $\beta_n = n^{-1/3}$ and $\gamma = 1$ (above) or $\gamma = 10$ (below).



7 Experiment on real data: Risk factors for Diabetes

In this section, we evaluate the performance of our proposed testing procedure on a real dataset using various matching schemes, and compare it against other well-established conditional independence testing methods in the literature. We use a dataset[1] on the incidence of diabetes among the Pima Indian population near Phoenix, Arizona, originally collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. The dataset 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., CDC [2] 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. This relationship is also verified empirically in the left-most panel of Figure 5. Most of the other variables, such as `BloodPressure`, `BMI`, `Glucose`, and `Pregnancies`, are also considered potential risk factors for diabetes. In this experiment, we aim to determine whether these variables remain significant risk factors for diabetes, even after controlling for age.

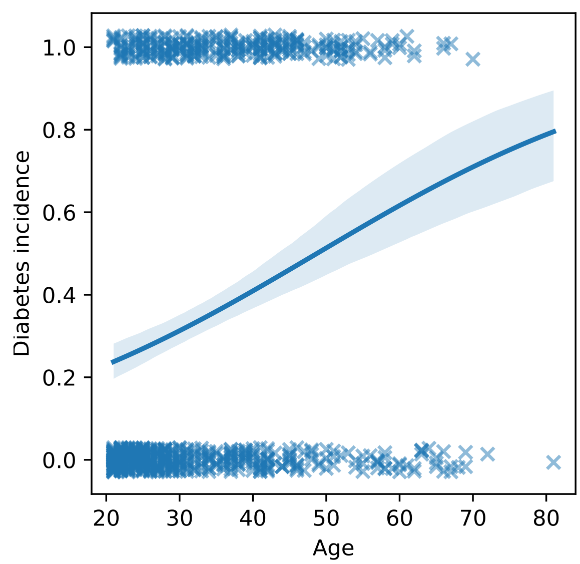


Figure 4: A scatter-plot of `Age` and `Diabetes Incidence` along with the fitted logistic regression model to demonstrate the stochastic monotonicity between them.

[1]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).

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

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. For this purpose, we will be 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`. 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 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 ghost 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, and we therefore expect our testing procedure to yield significantly larger p-values compared to the previous experiment. 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 $P_{X|Z}$ as per *continuous ranked probability score (CRPS)*, while respecting the monotonocity constraint.

		Pregnancies	Glucose	Blood pressure	Skin thickness	Insulin	BMI
Permutation test (testing marginal indep.)		0.001(0.001)	0.001(0.00)	0.151(0.154)	0.12(0.134)	0.024(0.049)	0.001(0.00)
PairSwap-ICI	neighbor matching	0.438(0.275)	0.002(0.01)	0.495(0.275)	0.234(0.218)	0.157(0.185)	0.031(0.065)
	cross-bin matching	0.415(0.254)	0.001(0.00)	0.49(0.256)	0.146(0.166)	0.123(0.157)	0.001(0.004)
PS-ICI (with synthetic control)	neighbor matching	0.507(0.286)	0.504(0.287)	0.504(0.287)	0.504(0.29)	0.507(0.287)	0.508(0.286)
	cross-bin matching	0.509(0.289)	0.522(0.288)	0.516(0.288)	0.509(0.289)	0.509(0.29)	0.521(0.287)

Table 1: p-values, averaged over 3000 bootstrap samples along with the estimated standard deviations (within brackets) for the different tests from different experiments, as outlined in Section 7.

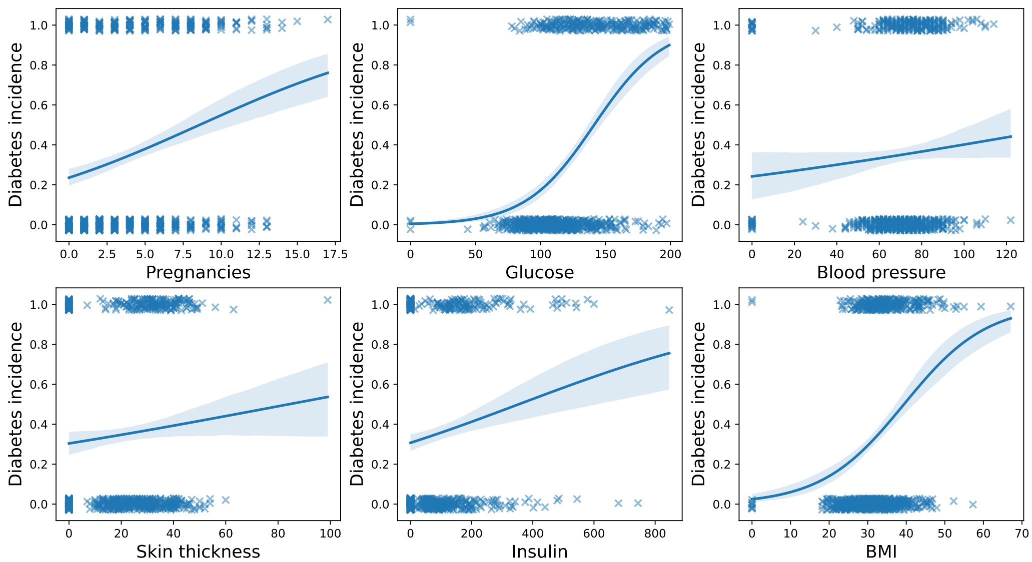


Figure 5: Scatter plots of X and other feature variables along with the fitted logistic regression models to demonstrate the dependence among these variables and `Diabetes Incidence`.

Results: The p-values from various experiments with different choices of the Y variable are presented in Table 1. We observe that, according to the test for marginal independence, most variables show significant dependence with the incidence of diabetes at the 0.1 level of significance.

Among the results from our conditional independence tests conditioned on `Age` in Table 1, the p-values for `Pregnancies` and `BloodPressure` are insignificant. This suggests that, after controlling for age, the data does not provide sufficient evidence to support them as risk factors for diabetes. However, we also observe that both variables show marginal dependence with diabetes incidence, as visually demonstrated in Figure 5 and confirmed by Experiment 1.

On the other hand, while `SkinThickness` is marginally independent of diabetes incidence at the 0.05 level of significance, indicating it is not a significant risk factor on its own, the conditional independence tests reveal that, after controlling for age, it shows significant dependence with diabetes incidence.

8 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 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 avenues for future work.

- *Optimal power in general settings.* Theorem 9 shows that the power for cross-bin matching can rival that of a parametric oracle with knowledge of the conditional mean μ, provided that the conditional distribution $P_{Y|Z}$ is symmetric. As we discussed, this condition appears to be essentially unavoidable. How can we achieve the oracle power against non-symmetric alternatives?
- *Avoiding data splitting.* The max-weight 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 and retain finite-sample error control?
- *Alternative methods.* A notable benefit of our stochastic monotonicity assumption is that we can consistently estimate the conditional distribution $P_{X|Z}$ using isotonic distributional regression (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?
- *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 (Shaked and Shanthikumar, 2007)?

References

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

Barlow, R. E. and Proschan, F. (1975) *Statistical Theory of Reliability and Life Testing: Probability Models*, vol. 1. Holt, Rinehart and Winston New York.

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 S Chult, 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.

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

Nevius, S. E., Proschan, F. and Sethuraman, J. (1974) Schur functions in statistics. III: A stochastic version of weak majorization, with applications. Tech. rep., Florida State University Statistics Report M307.

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

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.

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.

A Proofs of results from Section 5

In this section, we prove the power guarantees from Section 5. We begin in Appendix A.1 by stating the general power guarantees (both finite-sample and asymptotic) for any valid matching and weighting scheme, and then in Appendix A.2, we specialize these power calculations to max-weight matching, both with and without oracle knowledge of μ. Finally, in Appendix A.3, we prove the power guarantees for immediate neighbor matching and cross-bin matching under the special case of gaussian partial linear model from Section 5.3.

Notation We write $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ as shorthand for independent triples $(X_i, Y_i, Z_i)_{i\in[n]}$. For a matching $\{(i_{1,n}, j_{1,n}), (i_{2,n}, j_{2,n}), \ldots, (i_{L_n,n}, j_{L_n,n})\}$ and a vector $\mathbf{V} = (V_1, \ldots, V_n) \in \mathbb{R}^n$, we define $\Delta\mathbf{V} = (\Delta_1\mathbf{V}, \ldots, \Delta_{L_n}\mathbf{V}) \in \mathbb{R}^{L_n}$ with entries $\Delta_\ell\mathbf{V} := V_{i_{\ell,n}} - V_{j_{\ell,n}}$. 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

$$\rho_k = \left(\mathbb{E}_{\substack{\zeta,\zeta' \sim P_\zeta \\ \zeta \perp\!\!\!\perp \zeta'}}\left[|\zeta - \zeta'|^k\right]\right)^{1/k}; \tag{21}$$

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

A.1 Power guarantees for any valid matching and weighting scheme

In this section, we consider any valid matching and weighting scheme as outlined in Section 2, and state finite-sample lower and upper bounds on the conditional power $\mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\}$. Further, under the asymptotic setting from Section 5.1, we state asymptotic high-probability upper and lower bounds for the same quantity. The proof of this result is deferred to Appendix C. Before going to the result, we first define several quantities in terms of the weight vector $\mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$, $\delta > 0$ and $\{\rho_k : k \leq 6\}$, defined in (21): let

$$\begin{aligned}
\epsilon_{1,\delta,U} &= \frac{\|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2^2}{\rho_2^2\|\Delta\mathbf{w}\|_2^2} + \frac{\rho_4^2}{\rho_2^2\sqrt{\delta}}\frac{\|\Delta\mathbf{w}\|_\infty}{\|\Delta\mathbf{w}\|_2} + \frac{2}{\rho_2\sqrt{\delta}}\frac{\|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\Delta\mathbf{w}\|_2}, \\
\epsilon_{1,\delta,L} &= \frac{\|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2^2}{\rho_2^2\|\Delta\mathbf{w}\|_2^2} - \frac{\rho_4^2}{\rho_2^2\sqrt{\delta}}\frac{\|\Delta\mathbf{w}\|_\infty}{\|\Delta\mathbf{w}\|_2} - \frac{2}{\rho_2\sqrt{\delta}}\frac{\|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\Delta\mathbf{w}\|_2}, \\
\epsilon_{2,\delta} &= \frac{1}{(1+\epsilon_{1,\delta,L})^{3/2}}\Bigg\{\left(\frac{\|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2}{\rho_2\|\Delta\mathbf{w}\|_2}\right)^{2/3} \cdot \left(\frac{\|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty}{\rho_2\|\Delta\mathbf{w}\|_2}\right)^{1/3} \\
&\qquad\qquad + \left(\frac{\rho_3^3}{\rho_2^3}\frac{\|\Delta\mathbf{w}\|_\infty}{\|\Delta\mathbf{w}\|_2} + \frac{\rho_6^3}{\sqrt{\delta}\rho_2^3} \cdot \frac{\|\Delta\mathbf{w}\|_\infty^2}{\|\Delta\mathbf{w}\|_2^2}\right)^{1/3}\Bigg\}, \\
\epsilon_{3,\delta} &= \frac{0.56\rho_3^3}{\rho_2^3} \cdot \frac{\|\Delta\mathbf{w}\|_\infty}{\|\Delta\mathbf{w}\|_2} + 5\delta.
\end{aligned} \tag{22}$$

Theorem A.10. *Suppose that* $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ *satisfies* (14), *and that* $\psi_\ell = w_{i_\ell}X_{i_\ell} + w_{j_\ell}X_{j_\ell}$ *for* $\ell \in [L]$, *where* $\mathbf{w}$ *is our chosen weight vector. Then*

(i) For any $\delta > 0$,

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

where $\epsilon_{1,\delta,U}, \epsilon_{1,\delta,L}, \epsilon_{2,\delta}$, *and* $\epsilon_{3,\delta}$ *are as in* (22).

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

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

Assumption A2. $\|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2/\Delta\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z}) = o_P(1)$.

Then

$$\Phi\left(\frac{\Delta\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\Delta\mathbf{w}\|_2}-\bar{\Phi}^{-1}(\alpha)\right)-\mathrm{o}_P(1)\le\mathbb{P}\left\{p\le\alpha|\mathbf{Y},\mathbf{Z}\right\}$$
$$\le\Phi\left(\frac{\Delta\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\Delta\mathbf{w}\|_2}-\bar{\Phi}^{-1}(\alpha)\right)+\mathrm{o}_P(1).$$

A.2 Proving power results for max-weight matching

We now refine our power guarantees for max-weight matching, beginning with the ideal scenario where oracle knowledge of μ is available. We then extend these results to a more practical setting, where μ is estimated from one half of the data, and the test procedure is applied to the remaining half.

A.2.1 Power of max-weight matching with the oracle knowledge of μ

In Theorem A.10, we observe that the term $\frac{\Delta\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\mathbf{w}\|_2}$ dominates both the upper and lower bounds on conditional power. To maximize this term—and consequently the power—max-weight matching selects the weight vector $\Delta\mu^+(\mathbf{Y},\mathbf{Z})$, with entries $\max\{0,\ \mu(Yi_\ell,Z_{i_\ell})-\mu(Y_{j_\ell},Z_{j_\ell})\}$ (the detailed arguments are presented in Section 4). Below we present the exact lower and upper bounds to conditional power and then extend these results to the asymptotic setting in Theorem 3. We call the residual vector after projecting $\mu(\mathbf{Y},\mathbf{Z})$ onto $\mathrm{Mon}(\mathcal{Z})$,

$$\mathrm{Res}_n(\mathbf{Y},\mathbf{Z})=\mu(\mathbf{Y},\mathbf{Z})-\widehat{\mu}_{\mathrm{ISO}}(\mathbf{Z}) \tag{23}$$

and then we define the following quantities in terms of $\mathrm{Res}_n(\mathbf{Y},\mathbf{Z})$, $\delta>0$ and $\{\rho_k:k\le 6\}$, defined in (21): let

$$\begin{aligned}
\epsilon_{1,\delta,U}&=\frac{8\|\mathrm{Res}_n(\mathbf{Y},\mathbf{Z})\|_4^4}{\rho_2^2\widehat{\mathrm{ISS}}_n^2}+\frac{2\|\mathrm{Res}_n(\mathbf{Y},\mathbf{Z})\|_\infty}{\widehat{\mathrm{ISS}}_n}\cdot\left(\frac{\rho_4^2}{\rho_2^2\sqrt{\delta}}+\frac{4}{\rho_2\sqrt{\delta}}\|\mathrm{Res}_n(\mathbf{Y},\mathbf{Z})\|_\infty\right),\\
\epsilon_{1,\delta,L}&=\frac{8\|\mathrm{Res}_n(\mathbf{Y},\mathbf{Z})\|_4^4}{\rho_2^2\widehat{\mathrm{ISS}}_n^2}-\frac{2\|\mathrm{Res}_n(\mathbf{Y},\mathbf{Z})\|_\infty}{\widehat{\mathrm{ISS}}_n}\cdot\left(\frac{\rho_4^2}{\rho_2^2\sqrt{\delta}}+\frac{4}{\rho_2\sqrt{\delta}}\|\mathrm{Res}_n(\mathbf{Y},\mathbf{Z})\|_\infty\right),\\
\epsilon_{2,\delta}&=\frac{1}{(1+\epsilon_{1,\delta,L})^{3/2}}\left[\left(\frac{2\sqrt{2}\|\mathrm{Res}_n(\mathbf{Y},\mathbf{Z})\|_4^2}{\rho_2\widehat{\mathrm{ISS}}_n}\right)^{2/3}\cdot\left(\frac{4\|\mathrm{Res}_n(\mathbf{Y},\mathbf{Z})\|_\infty^2}{\rho_2\widehat{\mathrm{ISS}}_n}\right)^{1/3}+\right.\\
&\qquad\left.\left(\frac{2\rho_3^3}{\rho_2^3}\frac{\|\mathrm{Res}_n(\mathbf{Y},\mathbf{Z})\|_\infty}{\widehat{\mathrm{ISS}}_n}+\frac{4\rho_6^3}{\sqrt{\delta}\rho_2^3}\cdot\left(\frac{\|\mathrm{Res}_n(\mathbf{Y},\mathbf{Z})\|_\infty}{\widehat{\mathrm{ISS}}_n}\right)^2\right)^{1/3}\right]^3,\\
\text{and}\quad \epsilon_{3,\delta}&=\frac{1.12\rho_3^3}{\rho_2^3}\cdot\frac{\|\mathrm{Res}_n(\mathbf{Y},\mathbf{Z})\|_\infty}{\widehat{\mathrm{ISS}}_n}+5\delta,
\end{aligned}\tag{24}$$

Theorem A.11. *Suppose* $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ *is as defined in* (14). *Then, the conditional power of the max-weight matching satisfies the following.*

1. *For any $\delta > 0$,*

$$\Phi\left(\frac{\widehat{\mathrm{ISS}}_n}{\rho_2} - \bar{\Phi}^{-1}\left(\alpha - \epsilon_{2,\delta}\right) \cdot \left(1+\epsilon_{1,\delta,U}\right)^{1/2}\right) - \epsilon_{3,\delta} \le \mathbb{P}\left\{p \le \alpha | \mathbf{Y}, \mathbf{Z}\right\}$$
$$\le \Phi\left(\frac{\sqrt{2}\widehat{\mathrm{ISS}}_n}{\rho_2} - \bar{\Phi}^{-1}\left(\alpha - \epsilon_{2,\delta}\right) \cdot \left(1+\epsilon_{1,\delta,L}\right)^{1/2}\right) + \epsilon_{3,\delta}.$$

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

2. *Further, if $\left(\|\mu(\mathbf{Y},\mathbf{Z})\|_\infty \vee \|\mu(\mathbf{Y},\mathbf{Z})\|_\infty^4\right)/\widehat{\mathrm{ISS}}_n$ is $\mathrm{o}_P(1)$, then the conditional power of max-weight matching satisfies the following.*

$$\Phi\left(\frac{\widehat{\mathrm{ISS}}_n}{\rho_2} - \bar{\Phi}^{-1}(\alpha)\right) - \mathrm{o}_P(1) \ \le \mathbb{P}\{p \le \alpha \mid \mathbf{Y}, \mathbf{Z}\} \le \ \Phi\left(\frac{\sqrt{2}\widehat{\mathrm{ISS}}_n}{\rho_2} - \bar{\Phi}^{-1}(\alpha)\right) + \mathrm{o}_P(1).$$

Proof. Firstly, for max-weight matching the weight vector is chosen to be $\Delta\mu^+(\mathbf{Y},\mathbf{Z})$. Thus, an application of Theorem A.10 gives us an immediate upper and lower bound on conditional power. Replacing $\Delta\mathbf{w} = \Delta\mu^+(\mathbf{Y},\mathbf{Z})$, the dominating term simplifies as

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

and the terms $\epsilon_{1,\delta,U}, \epsilon_{1,\delta,L}, \epsilon_{2,\delta}$ from (22) simplify to the following.

$$\begin{aligned}
\epsilon_{1,\delta,U} &= \frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_4^4}{\rho_2^2\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^2} + \frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2}\cdot\left(\frac{\rho_4^2}{\rho_2^2\sqrt{\delta}} + \frac{2}{\rho_2\sqrt{\delta}}\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty\right), \\
\epsilon_{1,\delta,L} &= \frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_4^4}{\rho_2^2\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^2} - \frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2}\cdot\left(\frac{\rho_4^2}{\rho_2^2\sqrt{\delta}} + \frac{2}{\rho_2\sqrt{\delta}}\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty\right), \text{ and} \\
\epsilon_{2,\delta} &= \frac{1}{(1+\epsilon_{1,\delta,L})^{3/2}}\left[\left(\frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_4^2}{\rho_2\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2}\right)^{2/3}\cdot\left(\frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty^2}{\rho_2\|\Delta\mu(^+\mathbf{Y},\mathbf{Z})\|_2}\right)^{1/3} + \right. \\
&\qquad \left.\left(\frac{\rho_3^3}{\rho_2^3}\frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2} + \frac{\rho_6^3}{\sqrt{\delta}\rho_2^3}\cdot\left(\frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2}\right)^2\right)^{1/3}\right]^3, \\
\epsilon_{3,\delta} &= \frac{0.56\rho_3^3}{\rho_2^3}\cdot\frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2} + 5\delta.
\end{aligned} \tag{25}$$

By Lemma B.15, there exists a matching (more details in Appendix B) such that

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

Since max-weight matching maximizes $\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2$ over all valid matching schemes, the above lower bound on $\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2$ is valid for max-weight matching too. Above mentioned lower bound on L_2-norm and upper bounds on norms of $\Delta\mu^+(\mathbf{Y},\mathbf{Z})$ from Lemma A.12 (stated below) implies that the $\epsilon_{1,\delta,U}, \epsilon_{1,\delta,L}, \epsilon_{2,\delta}, \epsilon_{3,\delta}$ terms from (25) are further upper bounded by the corresponding terms in (24). This proves the first part of the result.

In order to prove the second part, it suffices to show that Assumption (A1), (A2) follow from

$$\left(\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty \vee \|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty^4\right)/\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 \;\text{ is }\; o_P(1), \tag{26}$$

which, in turn, is implied by $\left(\|\mu(\mathbf{Y},\mathbf{Z})\|_\infty \vee \|\mu(\mathbf{Y},\mathbf{Z})\|_\infty^4\right)/\widehat{\mathrm{ISS}}_n = o_P(1)$ by Lemma B.15, and noting that $\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty \leq 2\|\mu(\mathbf{Y},\mathbf{Z})\|$. Towards that goal, we start with observing

$$\begin{aligned}\frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2}{\Delta\mu^+(\mathbf{Y},\mathbf{Z})^T\Delta\mu(\mathbf{Y},\mathbf{Z})} = \frac{\|\left(\Delta\mu^+(\mathbf{Y},\mathbf{Z})\right)^2\|}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^2} &= \frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_4^2}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^2}\\ &\leq \frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^2} \;\text{ is }\; o_P(1) \;\text{ by }\; (26).\end{aligned}$$

This proves that assumption (A2) holds. Similarly, we observe

$$\begin{aligned}&\frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{1/3}\cdot\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2^{2/3}}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2}\\ &\quad= \frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty^{2/3}\cdot\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_4^{4/3}}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2} \leq \frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty^{4/3}\cdot\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^{2/3}}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2}\\ &\quad= \left(\frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty^4}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2}\right)^{1/3} \;\text{ is }\; o_P(1) \;\text{ by }\; (26)\end{aligned}$$

and thus (A1) holds too. This completes the proof for the second part of our result.

□

Lemma A.12. *For any valid matching, the following upper bounds on norms of* $\Delta\mu^+(\mathbf{Y},\mathbf{Z})$ *hold valid.*

1. $\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 \leq \sqrt{2}\cdot\|\mathrm{Res}_n(\mathbf{Y},\mathbf{Z})\|_2 = \sqrt{2}\widehat{\mathrm{ISS}}_n$.
2. $\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty \leq 2\cdot\|\mathrm{Res}_n(\mathbf{Y},\mathbf{Z})\|_\infty$.
3. $\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_4 \leq 2^{3/4}\cdot\|\mathrm{Res}_n(\mathbf{Y},\mathbf{Z})\|_4$,

while $\mathrm{Res}_n(\mathbf{Y},\mathbf{Z})$ *is as defined in* (23)*.*

A.2.2 Power of max-weight matching without oracle knowledge of μ

Proof of Theorem 4. By Lemma B.15 $\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 \geq \widehat{\mathrm{ISS}}_n$ and also, it holds that

$$\begin{aligned}&\|\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2 \leq \sqrt{2}\cdot\|\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \mu(\mathbf{Y},\mathbf{Z})\|_2,\\ &\qquad\qquad \|\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty \leq 2\cdot\|\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \mu(\mathbf{Y},\mathbf{Z})\|_\infty.\end{aligned}$$

Hence, it is enough to prove the power guarantees under the following assumptions.

$$\begin{aligned}&\|\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2/\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2,\\ &\qquad\qquad \|\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^4/\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 \;\text{ are }\; o_P(1),\end{aligned} \tag{27}$$

Since now we use an estimated $\hat{\mu}$, our weight vector satisfies $\Delta\mathbf{w} = \Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})$. Firstly, we note that

$$\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 \leq \|\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2, \tag{28}$$

which simply follows by noting that $|a^+ - b^+| \leq |a - b|$ for any $a, b \in \mathbb{R}$. Further, the dominating term in our power guarantees from Theorem A.10, after scaling with ρ_2 can be upper bounded as follows.

$$\begin{aligned}
&\frac{\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})\|_2} = \frac{\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})^T\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z})}{\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})\|_2} + \frac{\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})^T\left(\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \Delta\mu(\mathbf{Y},\mathbf{Z})\right)}{\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})\|_2} \\
&= \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})\|_2 + \frac{\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})\|_2 \cdot \|\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2}{\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})\|_2} \\
&\leq \|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 + \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 + \|\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2 \\
&\leq \|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 \cdot \left(1 + \frac{2\|\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2}\right) \qquad \text{(by (28))}
\end{aligned}$$

which is $\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2(1 + o_P(1))$ by (27). Now, if (A1), (A2) holds then by arguments similar to that in the proof of Theorem A.11, we note that

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

which will complete the proof.

Proving that (A2) holds We start with observing

$$\begin{aligned}
\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2 &\leq \|\Delta\mu^+(\mathbf{Y},\mathbf{Z})^2\|_2 + \|\Delta\mu(\mathbf{Y},\mathbf{Z}) \circ \left(\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\right)\|_2 \\
&\leq \|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_4^2 + \|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty \cdot \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 \\
&\leq \|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty + \|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty \cdot \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2
\end{aligned}$$

and hence, by (28)

$$\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^2\left(\frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2} + \frac{\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2} \cdot \frac{\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2}\right)$$

is a upper bound on $\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2$. Similarly, we also note that

$$\begin{aligned}
\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})^T\Delta\mu(\mathbf{Y},\mathbf{Z}) &= \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})\|_2^2 + \left(\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \Delta\mu(\mathbf{Y},\mathbf{Z})\right)^T\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) \\
&\geq \frac{1}{2}\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^2 - \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^2 - \\
&\quad \|\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2\left(\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 + \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2\right),
\end{aligned}$$

which via (28) is further lower bounded by

$$\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^2\left(\frac{1}{2} - \frac{\|\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \Delta\mu\mathbf{Y},\mathbf{Z})\|_2}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2} - 2\frac{\|\Delta\hat{\mu}(\mathbf{Y},\mathbf{Z}) - \Delta\mu\mathbf{Y},\mathbf{Z})\|_2^2}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^2}\right).$$

Together these two inequalities imply that (A2) holds.

Proving that (A1) holds Next, we will show that (A1) holds, and we observe that

$$\frac{\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z})\|_2} \leq \frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty + \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2 - \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2}$$
$$\leq \frac{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty/\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2 + (\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2/\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2)}{1 - (\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2/\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2)},$$

where the last expression is $o_P(1)$ by (27), (28), and noting that $\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty/\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2$ is $o_P(1)$ by the original assumption from the Theorem 3 i.e., $(\|\mu(\mathbf{Y},\mathbf{Z})\|_\infty \vee \|\mu(\mathbf{Y},\mathbf{Z})\|_\infty^4)/\widehat{\mathrm{ISS}}_n$ is $o_P(1)$. Further, $\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{1/3}$ is upper bounded by

$$\left(\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty^2 + \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty\right)^{1/3}$$
$$\leq \|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty^{2/3} + \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{1/3}\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{1/3}$$

since $\|a^+\|_\infty \leq \|a\|_\infty$ for any vector a. Also, by a similar calculation $\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2^{2/3}$ is upper bounded by

$$\left(\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})^2\|_2 + \|\left(\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\right) \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2\right)^{2/3}$$
$$\leq \left(\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_4^2 + \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty\right)^{2/3}$$
$$\leq 2^{1/3}\cdot\left(\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^{2/3}\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{2/3} + \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^{2/3}\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{2/3}\right).$$

Combining both, $\left(\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{1/3} \cdot \|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2^{2/3}\right)/\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2$ is upper bounded by

$$2^{1/3}\cdot\left(\frac{\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{4/3}}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^{1/3}} + \frac{\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty^{1/3}\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^{1/3}}\right)\cdot$$
$$\left(1 + \frac{\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^{2/3}}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^{2/3}}\right)$$

The last expression is $o_P(1)$ as long as $\frac{\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{1/3}\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^{1/3}}$ is $o_P(1)$. Towards that, we note that

$$\frac{\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty^{1/3}\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^{1/3}} = \frac{\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty^{1/3}\cdot\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^{1/12}\cdot\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2^{3/12}}$$
$$= \left(\frac{\|\Delta\hat{\mu}^+(\mathbf{Y},\mathbf{Z}) - \Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_\infty^4}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2}\right)^{1/12}\left(\frac{\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^4}{\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2}\right)^{1/4}.$$

As we have noted in the proof of Theorem 3, $(\|\mu(\mathbf{Y},\mathbf{Z})\|_\infty \vee \|\mu(\mathbf{Y},\mathbf{Z})\|_\infty^4)/\widehat{\mathrm{ISS}}_n$ being $o_P(1)$ implies $\|\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^4/\|\Delta\mu^+(\mathbf{Y},\mathbf{Z})\|_2$ is also $o_P(1)$. Finally by (27), (28), the last expression is $o_P(1)$ which completes the proof. □

A.3 Proving power results from Section 5.3

Finally in this section, we specialize our power calculations for gaussian partial linear model from Section 5.3 and prove the power guarantees for immediate neighbour matching and cross-bin matching.

A.3.1 Proving power guarantees for immediate neighbor matching

Proof of Theorem 8. We assume that n is even, $Z_1 \leq \cdots \leq Z_n$ are ordered w.r.t the partial order $\preccurlyeq$, and Y_i is the observed Y value corresponding to Z_i. Since neighbor matching uses $\mathbf{Y}$ as the weight vector, by Theorem A.10 $\frac{\Delta\mathbf{Y}^{+T}\Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\Delta\mathbf{Y}^+\|_2}$ dominates the power calculations. Furthermore, under the gaussian partial linear model (19), it holds that

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

where the first term is negative due to anti-monotonicity of our weights, and its magnitude can be controlled suitably. The rest of the proof proceeds in three key steps, outlined as follows:

1. We would prove a control on magnitude of $\Delta\mathbf{Y}^{+T}\Delta\mu_0(\mathbf{Z})$, more precisely that
$$\left|\Delta\mathbf{Y}^{+T}\Delta\mu_0(\mathbf{Z})\right| = \mathrm{o}_P(\log n).$$

2. Next, we will show that $\frac{1}{n/2}\|\Delta\mathbf{Y}^+\|_2^2$ concentrates as follows.
$$\left|\frac{1}{n/2}\|\Delta\mathbf{Y}^+\|_2^2 - \mathbb{E}\left[\mathrm{Var}(Y \mid Z)\right]\right| = \mathrm{o}_P(1). \tag{29}$$

3. Finally, we will prove that (A1), (A2) holds for neighbor matching.

Observe, that an immediate consequence of (29) is that $\|\Delta\mathbf{Y}^+\|_2$ is $\omega_P(\sqrt{n})$ which will come useful in the following arguments. Firstly, by the first two key steps, the dominating term concentrates as follows.

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

Step 1 and 2 together establishes the expression for asymptotic power for neighbor matching. Finally, by step 3 and an application of Theorem A.10 concludes the proof for neighbor matching. Now, we go through the proof of the key steps one by one.

Step 1: control on $\left|\Delta\mathbf{Y}^{+T}\Delta\mu_0(\mathbf{Z})\right|$ Firstly, $\left|\Delta\mathbf{Y}^{+T}\Delta\mu_0(\mathbf{Z})\right|$ can be upper bounded as follows.

$$\begin{aligned}\left|\Delta\mathbf{Y}^{+T}\Delta\mu_0(\mathbf{Z})\right| &= \sum_{\ell=1}^{n/2} |Y_{2\ell-1} - Y_{2\ell}|\cdot(\mu_0(Z_{2\ell}) - \mu_o(Z_{2\ell-1})) \\ &\leq \|\Delta\mathbf{Y}^+\|_\infty \sum_{i=2}^{n} \mu_0(Z_i) - \mu(Z_{i-1}) = \|\Delta\mathbf{Y}^+\|_\infty\cdot(\mu_0(Z_n) - \mu_0(Z_1)).\end{aligned} \tag{30}$$

In above calculations, the first equality uses the fact that $\mu_0(Z_{2\ell}) \geq \mu_0(Z_{2\ell-1})$ which holds since $Z_1, \ldots, Z_n$ are ordered. Now, $\|\Delta \mathbf{Y}^+\|_\infty \cdot (\mu_0(Z_n) - \mu_0(Z_1))$ is $o_P(\log n)$ by boundedness of Y, and sub-gaussianity of $\mu_0(Z)$.

Step 2: Concentration of $\frac{1}{n/2}\|\Delta \mathbf{Y}^+\|_2^2$ Now, we analyze the large sample behaviour of

$$\|\Delta \mathbf{Y}^+\|_2^2 = \sum_{i=1}^{n/2} (Y_{2i-1} - Y_{2i})_+^2 .$$

Conditioned on $\{Z_1, Z_2, \ldots, Z_n\}$, $\|\Delta \mathbf{Y}^+\|_2^2$ is a sum of $n/2$ many independent and uniformly bounded terms. Hence, by the law of large numbers,

$$\left| \frac{1}{n/2}\|\Delta \mathbf{Y}^+\|_2^2 - \mathbb{E}\left[\frac{1}{n/2}\|\Delta \mathbf{Y}^+\|_2^2 \mid Z_1, Z_2, \ldots, Z_n\right] \right| = o_P(1).$$

Further, Lemma D.17 implies that

$$\left| \mathbb{E}\left[\frac{1}{n/2}\|\Delta \mathbf{Y}^+\|_2^2 \;\middle|\; Z_1, \ldots, Z_n\right] - \frac{1}{n}\sum_{i=1}^{n} \sigma_{Z_i}^2 \right| = o_P(1),$$

where we call $\sigma_z^2 = \mathrm{Var}(Y \mid Z = z)$. Finally, another application of law of large numbers imply

$$\left| \frac{1}{n}\sum_{i=1}^{n} \sigma_{Z_i}^2 - \mathbb{E}\left[\mathrm{Var}(Y \mid Z)\right] \right| = o_P(1),$$

and thus (29) holds.

Step 3: Proving (A1), (A2) holds for neighbor matching Firstly,

$$\|\Delta \mathbf{Y}^+ \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_\infty \leq \|\Delta \mathbf{Y}^+ \circ \Delta\mu_0(\mathbf{Z})\|_\infty + \beta\|\Delta \mathbf{Y}^+\|_\infty^2 \leq 4\left(\|\mathbf{Y}\|_\infty\|\mu_0(\mathbf{Z})\|_\infty + \beta\|\mathbf{Y}\|_\infty^2\right),$$

and it also holds that

$$\begin{aligned}\|\Delta \mathbf{Y}^+ \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_2 &= \|\Delta \mathbf{Y}^+ \circ \left(\Delta\mu_0(\mathbf{Z}) + \beta\Delta \mathbf{Y}\right)\|_2 \\ &\leq \|\Delta \mathbf{Y}^+ \circ \Delta\mu_0(\mathbf{Z})\|_2 + \left\|\beta\Delta {\mathbf{Y}^+}^2\right\|_2 \leq 2 \cdot \|\mu_0(\mathbf{Z})\|_\infty\|\Delta \mathbf{Y}^+\|_2 + \beta\|\Delta \mathbf{Y}^+\|_4^2. \quad (31)\end{aligned}$$

To verify (A1), we first note that $\|\Delta \mathbf{Y}^+\|_\infty / \|\Delta \mathbf{Y}^+\|_2$ is $o_P(1)$ since Y is bounded and $\|\Delta \mathbf{Y}^+\|_2 = \omega_P(\sqrt{n})$. Furthermore, $\|\Delta \mathbf{Y}^+ \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_\infty^{1/3} \cdot \|\Delta \mathbf{Y}^+ \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_2^{2/3}$ is bounded by

$$\left(4^{1/3} \cdot \|\mathbf{Y}\|_\infty^{1/3}\|\mu_0(\mathbf{Z})\|_\infty^{1/3} + 4^{1/3}\beta^{1/3}\|\mathbf{Y}\|_\infty^{2/3}\right)\left(2^{2/3} \cdot \|\mu_0(\mathbf{Z})\|_\infty^{2/3}\|\Delta \mathbf{Y}^+\|_2^{2/3} + 2^{1/3}\beta^{2/3}\|\Delta \mathbf{Y}^+\|_4^{4/3}\right).$$

Hence, $\|\Delta \mathbf{Y}^+ \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_\infty^{1/3} \cdot \|\Delta \mathbf{Y}^+ \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_2^{2/3} / \|\Delta \mathbf{Y}^+\|_2$ is upper bounded by

$$\begin{aligned}&2^{4/3} \cdot \frac{\|\mathbf{Y}\|_\infty^{1/3}\|\mu_0(\mathbf{Z})\|_\infty}{\|\Delta \mathbf{Y}^+\|_2^{1/3}} + 2\beta^{2/3} \cdot \frac{\|\mathbf{Y}\|_\infty^{1/3}\|\mu_0(\mathbf{Z})\|_\infty^{1/3}\|\Delta \mathbf{Y}^+\|_\infty^{2/3}}{\|\Delta \mathbf{Y}^+\|_2^{1/3}} \\ &\qquad + (16\beta)^{1/3} \cdot \frac{\|\mathbf{Y}\|_\infty^{2/3}\|\mu_0(\mathbf{Z})\|_\infty^{2/3}}{\|\Delta \mathbf{Y}^+\|_2^{1/3}} + 2\beta \cdot \frac{\|\mathbf{Y}\|_\infty^{2/3}\|\Delta \mathbf{Y}^+\|_\infty^{2/3}}{\|\Delta \mathbf{Y}^+\|_2^{1/3}},\end{aligned}$$

where each of these four terms are $o_P(1)$, implying (A1) holds. Now, combining (31) and (30), we also have

$$\frac{\|\Delta\mathbf{Y}^+ \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2}{\Delta\mathbf{Y}^{+T}\Delta\mu(\mathbf{Y},\mathbf{Z})} \leq \frac{2\cdot\|\mu_0(\mathbf{Z})\|_\infty \cdot \|\Delta\mathbf{Y}^+\|_2 + \beta\|\Delta\mathbf{Y}^+\|_4^2}{-2\cdot\|\Delta\mathbf{Y}^+\|_\infty\|\mu_0(\mathbf{Z})\|_\infty + \beta\|\Delta\mathbf{Y}^+\|_2^2}$$
$$\leq \frac{2\cdot\|\mu_0(\mathbf{Z})\|_\infty/\|\Delta\mathbf{Y}^+\|_2 + \beta\|\Delta\mathbf{Y}^+\|_\infty/\|\Delta\mathbf{Y}^+\|_2}{-\big(2\cdot\|\Delta\mathbf{Y}^+\|_\infty\|\mu_0(\mathbf{Z})\|_\infty\big)/\|\Delta\mathbf{Y}^+\|_2^2 + \beta}$$

which is $o_P(1)$ since $\beta_n = \omega(1/\sqrt{n})$. Hence (A2) holds too, which completes the proof.

□

A.3.2 Proving power guarantees for cross-bin matching

In order to state the power guarantees for cross bin matching, we start with defining the following quantities for any distribution P. We denote q-th quantile of the distribution by $Q_P(q) = \inf\{x : \mathbb{P}_{X\sim P}(X \leq x) \geq q\}$ and we call

$$\mathrm{Dev}(P) = \mathbb{E}_{q\sim\mathrm{Unif}[0.0.5)}(Q_P(1-q) - Q_P(q))^2 \tag{32}$$

the deviation of the distribution P. Unlike in Section 5.3 Theorem 9, in this section we lift the symmetry assumption on $P_{Y|Z}$ and state and prove the general power guarantees using the above-mentioned deviation term. We first state this general result below, and prove Theorem 9 using this general result. Later in this section, we also prove the general result.

Theorem A.13. *Consider the model class (19) and let β_n be $\omega(1/\sqrt{n})$. Suppose, for cross bin matching the number of bins i.e., K satisfy $K = \omega_P(\sqrt{n})$. Further, suppose Y is bounded and $\mu_0(Z)$ is a sub-Gaussian random variable and the conditional distribution $P_{Y|Z}$ is almost surely non-atomic and satisfy the following.*

- *There exists $L_\infty \geq 0$ such that for any $z_1, z_2 \in \mathcal{Z}$ it holds that*

$$\|P_{Y|Z}(\cdot \mid z_1) - P_{Y|Z}(\cdot \mid z_2)\|_\infty \leq L_P\,|P_Z(z_1) - P_Z(z_2)|^2,$$

  where for any two distributions P_1, P_2 we define

$$\|P_1 - P_2\|_\infty = \sup_{x\in\mathbb{R}} |P_1(x) - P_2(x)|.$$

- *There exists $L_W \geq 0$ such that for any $z_1, z_2 \in \mathcal{Z}$ it holds that*

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

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

Then, the power for cross-bin matching satisfies

$$\Big|\mathbb{P}\{p \leq \alpha \mid \mathbf{Y},\mathbf{Z}\} - \Phi\Big(\sqrt{n/2}\cdot(\beta_n/\sigma)\cdot\big(\mathbb{E}\{\mathrm{Dev}(P_{Y|Z})\}\big)^{1/2} - \bar{\Phi}^{-1}(\alpha)\Big)\Big| = o_P(1).$$

Proving Theorem 9 Before proving the general power guarantees for cross-bin matching, firstly we show that Theorem A.13 implies Theorem 9.

Proof. Since for any $q < 0.5$ $Q_P(1-q) \geq Q_P(0.5) \geq Q_P(q)$, it holds that

$$\begin{aligned}\text{Dev}(P) &= \mathbb{E}_{q\sim\text{Unif}[0.0.5)}[Q_P(1-q) - Q_P(0.5) + Q_P(0.5) - Q_P(q)]^2 \\ &\geq \mathbb{E}_{q\sim\text{Unif}[0.0.5)}[Q_P(1-q) - Q_P(0.5)]^2 + \mathbb{E}_{q\sim\text{Unif}[0.0.5]}[Q_P(0.5) - Q_P(q)]^2 \\ &= 2\mathbb{E}_{q\sim\text{Unif}[0,1]}[Q_P(q) - Q_P(0.5)]^2 = 2\mathbb{E}_{X\sim P}[X - \text{Median}(P)]^2,\end{aligned}$$

which is further lower bounded by $2\text{Var}_{X\sim P}(X)$. Hence, the first part of the result follows from power guarantees in Theorem A.13.

Furthermore, if P is symmetric, $Q_P(q) = Q_P(1-q)$ and therefore

$$(Q_P(1-q) - Q_P(0.5))(Q_P(0.5) - Q_P(q)) = (Q_P(1-q) - Q_P(0.5))^2.$$

Hence, $\text{Dev}(P)$ now can be exactly computed as

$$\begin{aligned}\text{Dev}(P) = \mathbb{E}_{q\sim\text{Unif}[0.0.5)}[Q_P(1-q) - Q_P(0.5)]^2 &+ \mathbb{E}_{q\sim\text{Unif}[0.0.5]}[Q_P(0.5) - Q_P(q)]^2 \\ &+ 2\cdot\mathbb{E}_{q\sim\text{Unif}[0.0.5)}[Q_P(1-q) - Q_P(0.5)]^2,\end{aligned}$$

By symmetry, all of the threes terms are same and hence $\text{Dev}(P)$ equals $4\mathbb{E}_{q\sim\text{Unif}[0.0.5]}[Q_P(0.5) - Q_P(q)]^2 = 4\mathbb{E}_{X\sim P}[X - \text{Median}(P)]^2$. Further, P being symmetric, median and mean are the same i.e., $Q_P(0.5)$ is mean of P. Hence, $\text{Dev}(P) = 4\text{Var}_{X\sim P}(X)$. This proves the second part of the result. ☐

Proving Theorem A.13

Proof of Theorem A.13. The proof for cross bin matching imitates the key steps from the proof for Immediate neighbor matching from Appendix A.3.1. Here as well, $\frac{\Delta\mathbf{Y}^{+T}\Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\Delta\mathbf{Y}^+\|_2}$ dominates the power guarantees since we use the Y values for computing weights. Recall that, under the gaussian partial linear model (19),

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

As in Appendix A.3.1, we will analyze each of the two terms separately. The rest of the proof proceeds in three key steps, outlined as follows:

1. We would prove a control on size of the cross bin matching i.e., L_n and upper bound on the magnitude of $\Delta\mathbf{Y}^{+T}\Delta\mu_0(\mathbf{Z})$, more precisely that

$$\left|\Delta\mathbf{Y}^{+T}\Delta\mu_0(\mathbf{Z})\right| = o_P(\sqrt{n}), \quad \text{and} \quad \left|\frac{L_n}{n/2} - 1\right| = o_P(1/\sqrt{n}). \tag{33}$$

2. Next, we will show that $\frac{1}{n/2}\|\Delta\mathbf{Y}^+\|_2^2$ concentrates as follows.

$$\left|\frac{1}{n/2}\|\Delta\mathbf{Y}^+\|_2^2 - \mathbb{E}\left[\text{Dev}(P_{Y|Z})\right]\right| = o_P(1). \tag{34}$$

3. Finally, we will prove that (A1), (A2) holds for cross bin matching.

Observe, that an immediate consequence of (34) is that $\|\Delta\mathbf{Y}^+\|_2$ is $\omega_P(\sqrt{n})$ which will come useful in the following arguments. Firstly, by the first two key steps, the dominating term concentrates as follows.

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

Step 1 and 2 together establishes the expression for asymptotic power for neighbor matching. Finally, by step 3 and an application of Theorem A.10 concludes the proof for cross bin matching. Now, we go through the proof of the key steps one by one.

Step 1: control on L_n and $\left|\Delta\mathbf{Y}^{+T}\Delta\mu_0(\mathbf{Z})\right|$ We recall from the definition of cross-bin matching in Section 4.3 that we have K bins in $\mathcal{Z}$ of approximately uniform mass given by

$$B_1=\{z\le Z_{\lfloor n/K\rfloor}\},\ldots,B_k=\{Z_{(k-1)\lfloor n/K\rfloor}\le z\le Z_{k\lfloor n/K\rfloor}\},\ldots,B_K=\{z\ge Z_{(K-1)\lfloor n/K\rfloor}\},$$

and for $k\in[K-1]$, it matches samples between J_k^+ and J_{k+1}^-. Let us denote $L_{k,n}$ as the number of matched samples between J_k^+ and J_{k+1}^-. In total, $L_n=\sum_{k=1}^{K-1}L_{k,n}$ be the total number of matched samples. Let us denote

$$m_k=\mathrm{Median}\{Y_i:\ i\in A_k\}\ \text{ for any } k,$$

i.e., the median of Y observations lying in the k-th bin. With this definition, we note that

$$L_{k,n}=\left|\{i\in J_{k+1}^-:\ Y_i\le m_k\right|=\hat{P}_{Y|Z\in B_{k+1}}(m_k),$$

where for any $k\in[K-2]$, $\hat{P}_{Y|Z\in B_k}$ denotes the empirical distribution $\frac{1}{\lfloor n/K\rfloor}\sum_{i:Z_i\in B_k}\delta_{Y_i}$. Finally, by application of union bounds and Dvoretzky–Kiefer–Wolfowitz inequality, it holds that for fixed $\epsilon>0$,

$$\mathbb{P}\left\{\max_{k\in[K-1]}\left\|\hat{P}_{Y|Z\in B_k}-P_{Y|Z\in B_k}\right\|_\infty\ge\sqrt{\frac{\log\sqrt{K\epsilon}}{n/K}}\right\}\le\frac{C}{\epsilon}$$

for some universal constant C, i.e. $\max_{k\in[K-1]}\left\|\hat{P}_{Y|Z\in B_k}-P_{Y|Z\in B_k}\right\|_\infty=\mathrm{O}_P\left(\sqrt{\frac{\log\sqrt{K}}{n/K}}\right)$. Now, it follows that $L_n\ge\sum_{k=1}^{K-2}L_{k,n}=\sum_{k=1}^{K-2}\hat{P}_{Y|Z\in B_{k+1}}(m_k)$ can be lower bounded by

$$\begin{aligned}&\sum_{k=1}^{K-2}P_{Y|Z\in B_{k+1}}(m_k)-(K-2)\cdot\max_{k\in[K-1]}\left\|\hat{P}_{Y|Z\in B_k}-P_{Y|Z\in B_k}\right\|_\infty\\ &\ge\sum_{k=1}^{K-2}P_{Y|Z\in B_k}(m_k)-(K-2)\left(\max_{k\in[K-1]}\left\|\hat{P}_{Y|Z\in B_k}-P_{Y|Z\in B_k}\right\|_\infty+\max_{k\in[K-1]}\left\|P_{Y|Z\in B_{k+1}}-P_{Y|Z\in B_k}\right\|_\infty\right)\\ &\ge\sum_{k=1}^{K-2}\hat{P}_{Y|Z\in B_k}(m_k)-(K-2)\left(2\max_{k\in[K-1]}\left\|\hat{P}_{Y|Z\in B_k}-P_{Y|Z\in B_k}\right\|_\infty+\max_{k\in[K-1]}\left\|P_{Y|Z\in B_{k+1}}-P_{Y|Z\in B_k}\right\|_\infty\right).\end{aligned}$$

Firstly, observe that by definition $\sum_{k=1}^{K-2} \hat{P}_{Y|Z\in B_k}(m_k) = \frac{(K-2)}{2}\lfloor n/K \rfloor \geq \frac{(K-2)}{2}(n/K-1)$ and also,

$$(K-2)\max_{k\in[K-1]}\left\|\hat{P}_{Y|Z\in B_k} - P_{Y|Z\in B_k}\right\|_\infty = \mathrm{O}_P\left(K\sqrt{\frac{K\log\sqrt{K}}{n}}\right) = \mathrm{O}_P(n^{1/4}\log n),$$

and further it also holds that

$$(K-2)\max_{k\in[K-1]}\left\|P_{Y|Z\in B_{k+1}} - P_{Y|Z\in B_k}\right\|_\infty \leq (K-2)\max_{k\in[K-1]}\sup_{\substack{z_1\in B_k\\ z_2\in B_{k+1}}}\left\|P_{Y|Z}(\cdot\mid z_1) - P_{Y|Z}(\cdot\mid z_2)\right\|_\infty,$$

which is further upper bounded by $(K-2)L_\infty \max_{k\in[K-1]}\sup_{\substack{z_1\in B_k\\ z_2\in B_{k+1}}} |P_Z(z_1)-P_Z(z_2)|^2$ and thus bounded by

$$(K-2)L_\infty \max_{k\in[K-1]}\sup_{\substack{z_1\in B_k\\ z_2\in B_{k+1}}}\left(\left|\hat{P}_Z(z_1)-\hat{P}_Z(z_2)\right| + \|\hat{P}_Z - P_Z\|_\infty\right)^2.$$

Since for $z_1 \in B_k$ and $z_2 \in B_{k+1}$ it holds that $\left|\hat{P}_Z(z_1)-\hat{P}_Z(z_2)\right| \leq 2/K$ and thus, $(K-2)\max_{k\in[K-1]}\left\|P_{Y|Z\in B_{k+1}} - P_{Y|Z\in B_k}\right\|_\infty$ is $\mathrm{O}_P(L_\infty/\sqrt{n})$. Thus, combining all these parts, it holds that $L_n \geq n/2 - \mathrm{O}_P(\sqrt{n})$. Since $L_n \leq n/2$ it holds that $\left|\frac{L_n}{n/2}-1\right| \leq o_P(1/\sqrt{n})$, which proves the control on L_n.

Now, for cross bin matching $\left|\Delta\mathbf{Y}^{+T}\Delta\mu_0(\mathbf{Z})\right|$ can be upper bounded as follows.

$$\left|\Delta\mathbf{Y}^{+T}\Delta\mu_0(\mathbf{Z})\right| = \sum_{\ell=1}^{L_n}|Y_{i_{\ell,n}} - Y_{j_{\ell,n}}|\cdot\left|\mu_0(Z_{j_{\ell,n}}) - \mu_o(Z_{i_{\ell,n}})\right| \leq n\|\Delta\mathbf{Y}^+\|_\infty\|\Delta\mu_0(\mathbf{Z})\|_\infty. \quad (35)$$

Now note that Y is bounded and so $|\Delta\mathbf{Y}^+\|_\infty = O_P(1)$. Furthermore, we note that by construction of matching,

$$\max_{(i_\ell,j_\ell)\in\mathcal{M}_n}\left|\hat{P}_Z(Z_{i_\ell}) - \hat{P}_Z(Z_{j_\ell})\right| \leq \frac{1}{K} = \omega_P(1/\sqrt{n}),$$

where $\hat{P}_Z$ denotes the empirical distribution $\frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$. By Dvoretzky–Kiefer–Wolfowitz inequality $\max_{i\in[n]}\left|\hat{P}_Z(Z_i) - P_Z(Z_i)\right| = \mathrm{o}_P(1/\sqrt{n})$ and hence it holds that

$$\begin{aligned}\|\Delta\mu_0(\mathbf{Z})\|_\infty &= \max_{(i_\ell,j_\ell)\in\mathcal{M}_n}|\mu_0(Z_{i_\ell}) - \mu_0(Z_{j_\ell})| \leq L_{\mu_0}\max_{(i_\ell,j_\ell)\in\mathcal{M}_n}|P_Z(Z_{i_\ell}) - P_Z(Z_{j_\ell})| \\ &\leq L_{\mu_0}\left(\max_{(i_\ell,j_\ell)\in\mathcal{M}_n}\left|\hat{P}_Z(Z_{i_\ell}) - \hat{P}_Z(Z_{j_\ell})\right| + \max_{i\in[n]}\left|\hat{P}_Z(Z_i) - P_Z(Z_i)\right|\right),\end{aligned}$$

which is $\omega_P\left(\frac{1}{\sqrt{n}}\right)$. Thus, $\left|\Delta\mathbf{Y}^{+T}\Delta\mu_0(\mathbf{Z})\right|$ is $\omega_P(\sqrt{n})$.

Step 2: Concentration of $\frac{1}{n/2}\|\Delta\mathbf{Y}^+\|_2^2$ Conditioned on $\{Z_1, Z_2, \ldots, Z_n\}$, the sum $\|\Delta\mathbf{Y}^+\|_2^2$ is a sum of L_n many independent and uniformly bounded terms. Hence, by the law of large numbers it holds that

$$\left|\frac{1}{L_n}\|\Delta\mathbf{Y}^+\|_2^2 - \mathbb{E}\left[\frac{1}{L_n}\|\Delta\mathbf{Y}^+\|_2^2 \,\middle|\, Z_1, Z_2, \ldots, Z_n\right]\right| = \mathrm{o}_P(1).$$

By (33), it further holds that

$$\left|\frac{1}{n/2}\|\Delta\mathbf{Y}^+\|_2^2 - \mathbb{E}\left[\frac{1}{L_n}\|\Delta\mathbf{Y}^+\|_2^2 \,\middle|\, Z_1, Z_2, \ldots, Z_n\right]\right| = \mathrm{o}_P(1).$$

Finally, Lemma D.18 implies that

$$\left|\mathbb{E}\left[\frac{1}{L_n}\|\Delta\mathbf{Y}^+\|_2^2 \,\middle|\, Z_1, Z_2, \ldots, Z_n\right] - \frac{1}{K}\sum_{k=1}^{K-1}\mathrm{Dev}(P_{Y|Z_{k\lfloor n/K\rfloor}})\right| = \mathrm{o}_P(1).$$

While denoting $\hat{P}_Z = \frac{1}{n}\sum_{i=1}^{n}\delta_{Z_i}$ and $\tilde{P}_Z = \frac{1}{K}\sum_{k=1}^{K}\delta_{Z_{k\lfloor n/K\rfloor}}$, we also note that

$$\|P_n - \tilde{P}_n\|_\infty = \frac{1}{K} = \omega(1/\sqrt{n}).$$

Since $P_n \xrightarrow{d} P$ and therefore $\tilde{P}_n \xrightarrow{d} P$ as $n \to \infty$, by continuous mapping theorem it holds that

$$\left|\frac{1}{K}\sum_{k=1}^{K-1}\mathrm{Dev}(P_{Y|Z_{k\lfloor n/K\rfloor}}) - \mathbb{E}\left[\mathrm{Dev}(P_{Y|Z})\right]\right| = \mathrm{o}_P(1),$$

which proves (34).

Step 3: Proving (A1), (A2) holds for cross bin matching Finally, we need to show that (A1),(A2) holds. We start with noting that

$$\|\Delta\mathbf{Y}^+ \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty \le \|\Delta\mathbf{Y}^+ \circ \Delta\mu_0(\mathbf{Z})\|_\infty + \beta\|\Delta\mathbf{Y}^+\|_\infty^2 \le 2\|\mathbf{Y}\|_\infty\|\Delta\mu_0(\mathbf{Z})\|_\infty + 4\beta\|\mathbf{Y}\|_\infty^2$$

and further, it holds that $\|\Delta\mathbf{Y}^+ \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2$ is upper bounded by

$$\begin{aligned}\|\Delta\mathbf{Y}^+ \circ \big(\Delta\mu_0(\mathbf{Z}) + \beta\Delta\mathbf{Y}\big)\|_2 &\le \|\Delta\mathbf{Y}^+ \circ \Delta\mu_0(\mathbf{Z})\|_2 + \left\|\beta{\Delta\mathbf{Y}^+}^2\right\|_2 \\ &\le \|\Delta\mu_0(\mathbf{Z})\|_\infty\|\Delta\mathbf{Y}^+\|_2 + \beta\|\Delta\mathbf{Y}^+\|_4^2.\end{aligned}$$

To verify that (A1) holds, we first note that $\|\Delta\mathbf{Y}^+\|_\infty/\|\Delta\mathbf{Y}^+\|_2$ is $\mathrm{o}_P(1)$ since Y is bounded and $\|\Delta\mathbf{Y}^+\|_2 = \omega_P(\sqrt{n})$. Therefore, $\|\Delta\mathbf{Y}^+\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty^{1/3}\cdot\|\Delta\mathbf{Y}^+\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2^{2/3}$ is bounded by

$$\begin{aligned}&\big(2^{1/3}\cdot\|\mathbf{Y}\|_\infty^{1/3}\|\Delta\mu_0(\mathbf{Z})\|_\infty^{1/3} + 4^{1/3}\beta^{1/3}\|\mathbf{Y}\|_\infty^{2/3}\big) \\ &\qquad\times\left(2^{1/3}\|\Delta\mu_0(\mathbf{Z})\|_\infty^{2/3}\|\Delta\mathbf{Y}^+\|_2^{2/3} + 2^{1/3}\beta^{2/3}\|\Delta\mathbf{Y}^+\|_4^{4/3}\right).\end{aligned}$$

Hence, after scaling with $\|\Delta \mathbf{Y}^+\|_2$ it is upper bounded by

$$2^{2/3} \cdot \frac{\|\mathbf{Y}\|_\infty^{1/3} \|\Delta\mu_0(\mathbf{Z})\|_\infty}{\|\Delta \mathbf{Y}^+\|_2^{1/3}} + 2^{2/3}\beta^{2/3} \cdot \frac{\|\mathbf{Y}\|_\infty^{1/3} \|\Delta\mu_0(\mathbf{Z})\|_\infty^{1/3} \|\Delta \mathbf{Y}^+\|_\infty^{2/3}}{\|\Delta \mathbf{Y}^+\|_2^{1/3}} \\ + 2\beta^{1/3} \cdot \frac{\|\mathbf{Y}\|_\infty^{2/3} \|\Delta\mu_0(\mathbf{Z})\|_\infty^{2/3}}{\|\Delta \mathbf{Y}^+\|_2^{1/3}} + 2\beta \cdot \frac{\|\mathbf{Y}\|_\infty^{2/3} \|\Delta \mathbf{Y}^+\|_\infty^{2/3}}{\|\Delta \mathbf{Y}^+\|_2^{1/3}},$$

where each of these four terms are $o_P(1)$, implying (A1) holds. Further, combining the inequalities from above with (35) it holds that

$$\begin{aligned} \frac{\|\Delta \mathbf{Y}^+ \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2}{\Delta \mathbf{Y}^{+T} \Delta\mu(\mathbf{Y},\mathbf{Z})} &\le \frac{\|\Delta\mu_0(\mathbf{Z})\|_\infty \|\Delta \mathbf{Y}^+\|_2 + \beta \|\Delta \mathbf{Y}^+\|_4^2}{-n\|\Delta \mathbf{Y}^+\|_\infty \|\Delta\mu_0(\mathbf{Z})\|_\infty + \beta\|\Delta \mathbf{Y}^+\|_2^2} \\ &\le \frac{2 \cdot \|\mu_0(\mathbf{Z})\|_\infty / \|\Delta \mathbf{Y}^+\|_2 + \beta \|\Delta \mathbf{Y}^+\|_\infty / \|\Delta \mathbf{Y}^+\|_2}{-\big(n \cdot \|\Delta \mathbf{Y}^+\|_\infty \|\Delta\mu_0(\mathbf{Z})\|_\infty\big) / \|\Delta \mathbf{Y}^+\|_2^2 + \beta} \quad \text{is} \quad o_P(1) \end{aligned}$$

since $\beta_n = \omega(1/\sqrt{n})$. This proves that (A2) holds, which completes the proof. □

B An oracle matching—Isotonic median matching

Our theoretical guarantees on power depend on $\widehat{\mathrm{ISS}}_n$, which represents the L_2 cost of projecting onto the $\mathrm{Mon}_n(\mathcal{Z})$ space. While Lemma A.12 establishes that, for any valid matching, the "signal strength" is upper bounded by $\widehat{\mathrm{ISS}}_n$ up to a constant, a key question remains: can we recover the full signal strength with an appropriate matching? Indeed, we can and towards this goal, we first introduce isotonic median matching (IMM), and shortly, we will demonstrate that for IMM, $\|\Delta\mu^+\|_2$ is lower bounded by another related quantity $\widetilde{\mathrm{ISS}}_n = \|\mu(\mathbf{Y},\mathbf{Z}) - \widetilde{\mu}_{\mathrm{ISO}}(\mathbf{Z})\|_2$ where the function $\widetilde{\mu}_{\mathrm{ISO}}$ is defined as

$$\widetilde{\mu}_{\mathrm{ISO}}(Y_i, Z_i) = \max_{1\le j\le i} \min_{i \le k \le n} \mathrm{Med}\left\{\mu(Y_\ell, Z_\ell)\right\}_{j \le \ell \le k}. \tag{36}$$

Here we assume that $Z_1 \le Z_2 \le \cdots \le Z_n$ are ordered with respect to the partial order $\preccurlyeq$. Also, given $a_1, a_2, \ldots, a_m \in \mathbb{R}$, we let $\mathrm{Med}(a_1, a_2, \ldots, a_m)$ denote the (uniquely defined) median of this collection, given by

$$\mathrm{Med}(a_1, a_2, \ldots, a_m) = \begin{cases} a_{(k+1)} & \text{if } m = 2k+1, \\ \frac{a_{(k)} + a_{(k+1)}}{2} & \text{if } m = 2k \text{ for some } k, \end{cases}$$

where $a_{(i)}$ denotes the i-th order statistic of $a_1, a_2, \ldots, a_m$ and where any ties are broken randomly. One can check that $\widetilde{\mu}_{\mathrm{ISO}}$ belongs to the class $\mathrm{Mon}_n(\mathcal{Z})$ and thus, $\widetilde{\mathrm{ISS}}$ can then be considered as an interpretation of *isotonic signal strength* in this context. Here, $\widetilde{\mu}_{\mathrm{ISO}}$ can be thought of an approximate version of empirical isotonic L_1 projection of $\mu(\cdot)$ onto the $\mathcal{Z}$-space

$$\widetilde{\mu}_{\mathrm{ISO}} \in \arg \inf_{g \in \mathrm{Mon}_n(\mathcal{Z})} \|\mu(\mathbf{Y},\mathbf{Z}) - g(\mathbf{Z})\|_1,$$

Similar to the max-min characterization of isotonic L_2 projection, we start with a similar max-min characterization of isotonic L_1 projection. Since the optimization problem is not strictly convex, there are many solutions to the L_1 projection, and we stick with the definition in (36) to get a clean theory.

Firstly from the definition (36), it follows that for some $m \in [n]$, there exists $1 = n_1 < n_2 < n_3 < \cdots < n_m = n+1$ such that for any $i \in [n]$, if $n_j \le i < n_{j+1}$,

$$\widetilde{\mu}_{\text{ISO}}(Y_i, Z_i) = \text{Med}\left\{\mu(Y_\ell, Z_\ell)\right\}_{n_j \le \ell \le n_{j+1}-1},$$

i.e. $\widetilde{\mu}_{\text{ISO}}$ is a monotonic piece-wise constant function with m many blocks, and within each block $\tilde{\mu}^{\text{ISO}}$ is exactly the median of the $\mu(\cdot)$ values falling inside that block. Now, we are ready to formally define the IMM matching procedure using this block-representation of $\widetilde{\mu}_{\text{ISO}}$.

IMM: an oracle matching scheme

- Fix any block $j \in [m]$. Within this j-th block $\widetilde{\mu}_{\text{ISO}}$ takes a constant value, say m_j. We call any sample point $i \in [n_j, n_{j+1})$ inside this block "negative" if $\mu(Y_i, Z_i) < m_j$, and we will call it "positive" if $\mu(Y_i, Z_i) \ge m_j$. Let us denote the positive indices inside this block as $i_1^+ < i_2^+ < \cdots < i_{n_{1j}}^+$, and similarly we denote the negative 'indices' inside the block as $i_1^- < i_2^- < \cdots < i_{n_{2j}}^-$.
- Define k_j to be $n_{2j} \wedge n_{1j}$. Within this jth block, $\{(i_1^+, i_1^-), \cdots, (i_{k_j}^+, i_{k_j}^-)\}$ be the matched pairs. One can interpret k_j as the number of matches in the j-th block.

We follow this procedure for each of the m blocks separately, and combine the matched indices to get the overall matching. Along with the matching, we consider the corresponding μ values to be the weights.

Now by definition, the IMM matching procedure does not depend on $\mathbf{X}$. Now if we can prove that the matched pairs satisfy anti-monotonicity in μ values, IMM provides a valid matching and weighting scheme. Hence, we need to show that for any t, $i_t^+ < i_t^-$, where (i_t^+, i_t^-) are the t-th paired indices within say, the j-th block. To prove this, we start with the following useful property of $\widetilde{\mu}_{\text{ISO}}$.

Lemma B.14. *Fix any $j \in [m-1]$. For any i such that $n_j \le i \le n_{j+1} - 1$,*

$$\text{Med}\left\{\mu(Y_\ell, Z_\ell)\right\}_{n_j \le \ell \le i} \ge \text{Med}\left\{\mu(Y_\ell, Z_\ell)\right\}_{n_j \le \ell \le n_{j+1}-1}.$$

Proof. Suppose for some i, the above doesn't hold i.e.

$$\text{Med}\left\{\mu(Y_\ell, Z_\ell)\right\}_{n_j \le \ell \le i} < \text{Med}\left\{\mu(Y_\ell, Z_\ell)\right\}_{n_j \le \ell \le n_{j+1}-1}.$$

Hence, the following holds.

$$\begin{aligned}
\widetilde{\mu}_{\text{ISO}}(Y_{n_j}, Z_{n_j}) &= \max_{1 \le a \le n_j} \min_{n_j \le b \le n} \text{Med}\left\{\mu(Y_\ell, Z_\ell)\right\}_{a \le \ell \le b} \\
&= \min_{n_j \le b \le n} \text{Med}\left\{\mu(Y_\ell, Z_\ell)\right\}_{n_j \le \ell \le b} \le \text{Med}\left\{\mu(Y_\ell, Z_\ell)\right\}_{n_j \le \ell \le i} \\
&< \text{Med}\left\{\mu(Y_\ell, Z_\ell)\right\}_{n_{(j)} \le \ell \le n_{(j+1)}-1} \le \max_{1 \le a \le n_{(j+1)}-1} \text{Med}\left\{\mu(Y_\ell, Z_\ell)\right\}_{a \le \ell \le n_{(j+1)}-1} \\
&\le \max_{1 \le a \le n_{(j+1)}-1} \min_{n_{(j+1)}-1 \le b \le n} \text{Med}\left\{\mu(Y_\ell, Z_\ell)\right\}_{a \le \ell \le b} = \widetilde{\mu}_{\text{ISO}}(Y_{n_{j+1}-1}, Z_{n_{j+1}-1}).
\end{aligned}$$

But, $\widetilde{\mu}_{\text{ISO}}(Y_{n_j}, Z_{n_j}) = \widetilde{\mu}_{\text{ISO}}(Y_{n_{j+1}-1}, Z_{n_{j+1}-1})$ as they belong to the same block. Hence, this is a contradiction. $\square$

The above lemma further implies the anti-monotonicity of our weights. To see why, suppose inside some block it holds that for some t, $i_t^- < i_t^+$ where (i_t^+, i_t^-) are the t-th paired indices inside that block. Hence, within j-th block, to the left of i_t^-, including i_t^- there are t many negative samples while there are strictly less that t many positive samples. Equivalently, this means, to the left of i_t^-, there are t many samples with $\mu(Y_t, Z_t) < \widetilde{\mu}_{\text{ISO}}(Y_t, Z_t)$, and there are less than t many samples with $\mu(Y_t, Z_t) \geq \widetilde{\mu}_{\text{ISO}}(Y_t, Z_t)$. Hence, this implies

$$\text{Med}\left\{\mu(Y_\ell, Z_\ell)\right\}_{n_j \leq \ell \leq t} < \text{Med}\left\{\mu(Y_\ell, Z_\ell)\right\}_{n_j \leq \ell \leq n_{j+1}-1},$$

which is a direct contradiction to Lemma B.14. Observe, this additionally also implies $k_j = n_{2j} \wedge n_{1j} = n_{2j}$ for any $j \in [m]$. Now that validity of IMM matching is established, we are ready to state and prove the key-property of IMM.

Lemma B.15. *For IMM matching,*

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

Proof. The second inequality follows directly from noting $\widehat{\mu}_{\text{ISO}}$ minimizes L_2 cost over all functions in $\text{Mon}_n(\mathcal{Z})$, and hence $\widehat{\text{ISS}}_n \leq \widetilde{\text{ISS}}_n$. For the fisrt inequality, we start with noting

$$\|\Delta\mu^+(\mathbf{Y}, \mathbf{Z})\|_2^2 = \sum_{\ell=1}^{L} (\mu(Y_{i_\ell}, Z_{i_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell}))^2.$$

Now, by construction, the matched pairs (i_ℓ, j_ℓ) belong to the same block where the blocks are defined by $\widetilde{\mu}_{\text{ISO}}$. Hence, $\widetilde{\mu}_{\text{ISO}}(Z_{i_\ell}) = \widetilde{\mu}_{\text{ISO}}(Z_{j_\ell})$ and thus,

$$\begin{aligned}(\mu(Y_{i_\ell}, Z_{i_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell}))^2 &= (\mu(Y_{i_\ell}, Z_{i_\ell}) - \widetilde{\mu}_{\text{ISO}}(Z_{i_\ell}) + \widetilde{\mu}_{\text{ISO}}(Z_{j_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell}))^2 \\ &\geq (\mu(Y_{i_\ell}, Z_{i_\ell}) - \widetilde{\mu}_{\text{ISO}}(Z_{i_\ell}))^2 + (\widetilde{\mu}_{\text{ISO}}(Z_{j_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell}))^2.\end{aligned}$$

where the last step follows again because $(\mu(Y_{i_\ell}, Z_{i_\ell}) - \widetilde{\mu}_{\text{ISO}}(Z_{i_\ell})) \cdot (\widetilde{\mu}_{\text{ISO}}(Z_{j_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell})) \geq 0$ because by construction, i_ℓ is positive sample and j_ℓ is a negative sample within the same block. $\square$

C Proof of the master theorem

Proof of Theorem A.10. We note that for fixed $n \in \mathbb{N}$, and conditional on $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, the quantity $\mathbf{s}^T(\Delta\mathbf{w} \circ \Delta\mathbf{X})$ is a weighted sum of independent and identically distributed random variables $\{s_\ell \cdot \Delta_\ell\mathbf{w} \cdot \Delta_\ell\mathbf{X}\}_{\ell \in [L_n]}$ with mean 0 and variance $\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_2^2$. Hence, writing $T := \frac{\mathbf{1}^T(\Delta\mathbf{w} \circ \Delta\mathbf{X})}{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_2^2}$, we have by the Berry–Esseen theorem that

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

Hence, the conditional power $\mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\}$ is bounded above and below as follows.

$$\mathbb{P}\left\{\bar{\Phi}(T) \leq \alpha - 0.56\frac{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_3^3}{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_2^3} \mid \mathbf{Y}, \mathbf{Z}\right\} \leq \mathbb{P}\{p \leq \alpha | \mathbf{Y}, \mathbf{Z}\}$$
$$\leq \mathbb{P}\left\{\bar{\Phi}(T) \leq \alpha + 0.56\frac{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_3^3}{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_2^3} \mid \mathbf{Y}, \mathbf{Z}\right\},$$

or equivalently, one can write

$$\mathbb{P}\left\{T \geq \bar{\Phi}^{-1}\left(\alpha - 0.56\frac{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_3^3}{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_2^3}\right) \mid \mathbf{Y}, \mathbf{Z}\right\} \leq \mathbb{P}\{p \leq \alpha | \mathbf{Y}, \mathbf{Z}\} \leq$$
$$\mathbb{P}\left\{T \geq \bar{\Phi}^{-1}\left(\alpha + 0.56\frac{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_3^3}{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_2^3}\right) \mid \mathbf{Y}, \mathbf{Z}\right\}.$$

Now observe, $\mathbf{1}^T\Delta\mathbf{w} \circ \Delta\mathbf{X}$ can be equivalently written as $\Delta\mathbf{w}^T\Delta\mathbf{X} = \Delta\mathbf{w}^T\Delta\mu(\mathbf{Y}, \mathbf{Z}) + \Delta\mathbf{w}^T\Delta\boldsymbol{\zeta}$. Now, conditioned on $(\mathbf{Y}, \mathbf{Z})$, the first term i.e. $\Delta\mathbf{w}^T\Delta\mu(\mathbf{Y}, \mathbf{Z})$ is constant since $\mathbf{w}$ is a deterministic function of $\mathbf{Y}, \mathbf{Z}$. For the second term, observe that for a fixed n, conditioned on $\mathbf{Y}, \mathbf{Z}$, $\Delta\mathbf{w}^T\Delta\boldsymbol{\zeta}$ is again a weighted sum of i.i.d. random variables $\{\Delta_l\mathbf{w} \cdot \Delta_l\boldsymbol{\zeta}\}_{\ell \in [L_n]}$ with mean 0 and variance $\rho_2^2\|\Delta\mathbf{w}\|_2^2$. Another implication of the Berry–Esseen theorem implies that

$$\sup_{x \in \mathbb{R}} \left|\mathbb{P}\left\{\frac{\Delta\mathbf{w}^T\Delta\boldsymbol{\zeta}}{\rho_2\|\Delta\mathbf{w}\|_2} \geq x \middle| \mathbf{Y}, \mathbf{Z}\right\} - \bar{\Phi}(x)\right| \leq \frac{0.56\rho_3^3}{\rho_2^3} \cdot \frac{\|\Delta\mathbf{w}\|_3^3}{\|\Delta\mathbf{w}\|_2^3} \leq \frac{0.56\rho_3^3}{\rho_2^3} \cdot \frac{\|\Delta\mathbf{w}\|_\infty}{\|\Delta\mathbf{w}\|_2}, \tag{37}$$

Finally, note that an upper bound on conditional power $\mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\}$ can be computed as

$$\mathbb{P}\left\{T \geq \bar{\Phi}^{-1}\left(\alpha + 0.56\frac{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_3^3}{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_2^3}\right) \mid \mathbf{Y}, \mathbf{Z}\right\}$$
$$= \mathbb{P}\left\{\frac{\Delta\mathbf{w}^T\Delta\mathbf{X}}{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_2} \geq \bar{\Phi}^{-1}\left(\alpha + 0.56\frac{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_3^3}{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_2^3}\right) \mid \mathbf{Y}, \mathbf{Z}\right\},$$

which can be further broken into the following three probability terms using the high probability bounds of L_2, L_3 norms of $\Delta\mathbf{w} \circ \Delta\mathbf{X}$ as follows.

$$\mathbb{P}\left\{\frac{\Delta\mathbf{w}^T\Delta\mathbf{X}}{\rho_2\|\Delta\mathbf{w}\|_2} \geq \bar{\Phi}^{-1}(\alpha + \epsilon_{2,\delta}) \cdot \sqrt{1 + \epsilon_{1,\delta,L}} \mid \mathbf{Y}, \mathbf{Z}\right\}$$
$$+ \mathbb{P}\left\{0.56\frac{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_3^3}{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_2^3} > \epsilon_{2,\delta} \mid \mathbf{Y}, \mathbf{Z}\right\} + \mathbb{P}\left\{\frac{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_2}{\sqrt{\rho}_2\|\Delta\mathbf{w}\|_2} < \sqrt{1 + \epsilon_{1,\delta,L}} \mid \mathbf{Y}, \mathbf{Z}\right\}$$

Now, for the given choices of $\epsilon_{1,\delta,U}, \epsilon_{1,\delta,L}$, and $\epsilon_{2,\delta}$, the last two probability terms are correspondingly bounded by 3δ and 2δ by Lemma C.16. Finally, using equation (37), we can bound the first term by

$$\Phi\left(\frac{\Delta\mathbf{w}^T\Delta\mu(\mathbf{Y}, \mathbf{Z})}{\rho_2\|\Delta\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha + \epsilon_{2,\delta}) \cdot \sqrt{1 + \epsilon_{1,\delta,L}}\right) + \frac{0.56\rho_3^3}{\rho_2^3} \cdot \frac{\|\Delta\mathbf{w}\|_\infty}{\|\Delta\mathbf{w}\|_2}.$$

This completes the proof for the upper bound. The proof for lower bound follows similarly which concludes the first part of the result.

For the second part, we start with noting

$$\frac{\|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_\infty}{\rho_2\|\Delta\mathbf{w}\|_2} \leq \frac{\|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_\infty^{1/3} \cdot \|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_2^{2/3}}{\rho_2\|\Delta\mathbf{w}\|_2} \quad \text{is} \quad \mathrm{o}_P(1)$$

by assumption (A1). Further, observe that

$$(1+\epsilon_{1,\delta,U})^{1/2} \leq \frac{\|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_2}{\rho_2\|\Delta\mathbf{w}\|_2} + (1+\epsilon_{1,\delta})^{1/2}, \text{ where}$$
$$\epsilon_{1,\delta} = \frac{\rho_4^2}{\rho_2^2\sqrt{\delta}}\frac{\|\Delta\mathbf{w}\|_\infty}{\|\Delta\mathbf{w}\|_2} + \frac{2}{\rho_2\sqrt{\delta}}\frac{\|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_\infty}{\|\Delta\mathbf{w}\|_2} \quad \text{is} \quad \mathrm{o}_P(1) \quad \text{by assumption (A1)}.$$

Since, $\epsilon_{1,\delta}$ is $\mathrm{o}_P(1)$, it also holds that $(1+\epsilon_{1,\delta,L})^{1/2} \geq (1-\epsilon_{1,\delta})^{1/2}$ Now, moving on to $\epsilon_{2,\delta}$, the first term in the numerator i.e.

$$\frac{\|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_\infty^{1/3} \cdot \|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_2^{2/3}}{\rho_2\|\Delta\mathbf{w}\|_2} \quad \text{is} \quad \mathrm{o}_P(1) \quad \text{by assumption (A1)},$$

while the second term is also $\mathrm{o}_P(1)$ similarly by assumption (A1). Combining both terms, $\epsilon_{2,\delta} = \mathrm{o}_P(1)$ since $(1+\epsilon_{1,\delta,L})^{1/2} \geq (1-\mathrm{o}_P(1))$ as we have establishes earlier. Finally, this means that for large n, with high probability $\alpha - \epsilon_{2,\delta}$ is smaller than $\alpha/2$ and thus, observe that $\frac{\Delta\mathbf{w}^T\Delta\mu(\mathbf{Y},\mathbf{Z})}{\rho_2\|\Delta\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha-\epsilon_{2,\delta}) \cdot (1+\epsilon_{1,\delta,U})^{1/2}$ can be bounded from below by

$$\frac{\Delta\mathbf{w}^T\Delta\mu(\mathbf{Y}, \mathbf{Z})}{\rho_2\|\Delta\mathbf{w}\|_2} \cdot \left(1 - \bar{\Phi}^{-1}(\alpha/2) \cdot \frac{\|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_2}{\Delta\mathbf{w}^T\Delta\mu(\mathbf{Y}, \mathbf{Z})}\right) - \bar{\Phi}^{-1}(\alpha - \mathrm{o}_P(1)) \cdot (1+\mathrm{o}_P(1))^{1/2}.$$

Now, $\big(\|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y}, \mathbf{Z})\|_2\big)/\Delta\mathbf{w}^T\Delta\mu(\mathbf{Y}, \mathbf{Z})$ is $\mathrm{o}_P(1)$ by assumption (A2). Also, $\epsilon_{3,\delta}$ is $\mathrm{o}_P(1)$ by assumption (A1) and hence, the conditional power $\mathbb{P}\{p \leq \alpha | \mathbf{Y}, \mathbf{Z}\}$ is almost surely lower bounded by

$$\begin{aligned}\Phi\left(\frac{\Delta\mathbf{w}^T\Delta\mu(\mathbf{Y}, \mathbf{Z})}{\rho_2\|\Delta\mathbf{w}\|_2} \cdot (1-\mathrm{o}_P(1)) - \bar{\Phi}^{-1}(\alpha - \mathrm{o}_P(1)) \cdot (1+\mathrm{o}_P(1))^{1/2}\right) - \mathrm{o}_P(1)& \\ \geq \Phi\left(\frac{\Delta\mathbf{w}^T\Delta\mu(\mathbf{Y}, \mathbf{Z})}{\rho_2\|\Delta\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha)\right) - \mathrm{o}_P(1).&\end{aligned}$$

By a similar argument, the result for upper bound follows too. □

Lemma C.16. *For any $\delta \in [0, 1/2)$, the following holds.*

$$\begin{aligned}\mathbb{P}\Big\{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_2^2 \leq \rho_2^2\|\Delta\mathbf{w}\|_2^2 \cdot (1+\epsilon_{1,\delta,U}) \mid \mathbf{Y}, \mathbf{Z}\Big\} &\geq 1 - 2\delta,\\ \mathbb{P}\Big\{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_2^2 \geq \rho_2^2\|\Delta\mathbf{w}\|_2^2 \cdot (1+\epsilon_{1,\delta,L}) \mid \mathbf{Y}, \mathbf{Z}\Big\} &\geq 1 - 2\delta,\\ \mathbb{P}\Big\{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_3^3 \leq \|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_2^3 \cdot \epsilon_{2,\delta} \mid \mathbf{Y}, \mathbf{Z}\Big\} &\geq 1 - 3\delta,\end{aligned}$$

where $\epsilon_{1,\delta,U}, \epsilon_{1,\delta,L}, \epsilon_{2,\delta}$ are as defined in (22).

D Additional lemmas and their proofs

D.1 Proving concentration for the key terms in neighbor matching and cross-bin matching

Theorem D.17. *For immediate neighbor matching under any distribution $P_{Y,Z}$ with Y taking values in a bounded support, it holds that*

$$\left|\mathbb{E}\left[\frac{1}{n/2}\|\Delta\mathbf{Y}^{+}\|_2^2 \;\middle|\; Z_1,\ldots,Z_n\right] - \frac{1}{n}\sum_{i=1}^{n}\sigma_{Z_i}^2\right| \\ \leq \frac{1}{n/2}\sum_{i=1}^{n/2} 2\mathrm{d}_{\mathrm{W}_1}\left(P_{Y|Z}(\cdot \mid Z_{2i-1}), P_{Y|Z}(\cdot \mid Z_{2i})\right) = \mathrm{o}_P(1),$$

where $\mathrm{d}_{\mathrm{W}_1}(\cdot,\cdot)$ is the 1-Wasserstein distance and $\sigma_z^2 = \mathrm{Var}(Y \mid Z = z)$.

Proof. Recall that, for neighbor matching it holds that

$$\|\Delta\mathbf{Y}^{+}\|_2^2 = \sum_{i=1}^{n/2}\left(Y_{2i-1} - Y_{2i}\right)_+^2.$$

Now suppose, n is even and $Z_1 \leq Z_2 \leq \cdots \leq Z_n$ are ordered w.r.t the partial order $\preccurlyeq$. Without loss of generality, let us also assume that $|Y| \leq 1$ almost surely. Let $F_{Y|Z}^{-1}(\cdot \mid z)$ denote the inverse CDF of the conditional distribution $P_{Y|Z}(\cdot \mid z)$. Consider $\{(U_{2i-1}, U_{2i})\}_{i\in[n/2]} \overset{\text{iid}}{\sim}$ Uniform$[0,1]$. Without loss of generality, for any $i \in [n/2]$ we can assume

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

Now, we define the coupling

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

Conditioned on $\{Z_1, Z_2, \ldots, Z_n\}$, Y'_{2i-1}'s and Y'_{2i}'s share the same conditional distribution with that of Y_{2i}'s and Y_{2i-1}'s. Moreover, by construction

$$\mathbb{E}\left[|Y_{2i-1} - Y'_{2i-1}| \mid Z_1,\ldots,Z_n\right] = \mathbb{E}\left[|Y_{2i} - Y'_{2i}| \mid Z_1,\ldots,Z_n\right] = \mathrm{d}_{\mathrm{W}_1}\left(P_{Y|Z}(\cdot \mid Z_{2i-1}), P_{Y|Z}(\cdot \mid Z_{2i})\right).$$

Since $t \to t_+^2$ is 2-lipschitz for $t \in [-1,1]$, it holds that

$$\left|\left(Y_{2i-1} - Y_{2i}\right)_+^2 - \left(Y'_{2i-1} - Y_{2i}\right)_+^2\right| \leq 2|Y_{2i-1} - Y'_{2i-1}|$$

and similarly, $\left|(Y_{2i-1} - Y_{2i})_+^2 - (Y_{2i-1} - Y'_{2i})_+^2\right| \leq 2|Y_{2i} - Y'_{2i}|$. Combining the above two inequalities, it follows that for any $i \in [n/2]$,

$$\left|\left(Y_{2i-1} - Y_{2i}\right)_+^2 - \frac{1}{2}\left[\left(Y'_{2i-1} - Y_{2i}\right)_+^2 + \left(Y_{2i-1} - Y'_{2i}\right)_+^2\right]\right| \leq |Y_{2i-1} - Y'_{2i-1}| + |Y_{2i} - Y'_{2i}|.$$

Since Y_{2i-1}, Y'_{2i} are i.i.d given $\{Z_1, \ldots, Z_n\}$, we further have $\mathbb{E}\left[(Y_{2i-1} - Y'_{2i})_+^2 \mid Z_1, \ldots, Z_n\right] = \sigma^2_{Z_{2i-1}}$ and similarly, $\mathbb{E}\left[\left(Y_{2i} - Y'_{2i-1}\right)_+^2 \mid Z_1, \ldots, Z_n\right] = \sigma^2_{Z_{2i}}$. Therefore, it holds that

$$\left|\mathbb{E}\left[(Y_{2i-1} - Y_{2i})_+^2 \mid Z_1, \ldots, Z_n\right] - \frac{\left(\sigma^2_{Z_{2i-1}} + \sigma^2_{Z_{2i}}\right)}{2}\right| \\ \leq \mathbb{E}\left[|Y_{2i-1} - Y'_{2i}| + |Y'_{2i-1} - Y_{2i}|\right] = 2\mathrm{d}_{\mathrm{W}_1}\left(P_{Y|Z}(\cdot \mid Z_{2i-1}), P_{Y|Z}(\cdot \mid Z_2)\right).$$

Finally, recall that $\|\Delta\mathbf{Y}^+\|_2^2 = \sum_{i=1}^{n/2}(Y_{2i-1} - Y_{2i})_+^2$. Thus, it holds that

$$\left|\mathbb{E}\left[\frac{1}{n/2}\|\Delta\mathbf{Y}^+\|_2^2 \mid Z_1, \ldots, Z_n\right] - \frac{1}{n}\sum_{i=1}^{n}\sigma^2_{Z_i}\right| \leq \frac{1}{n/2}\sum_{i=1}^{n/2} 2\mathrm{d}_{\mathrm{W}_1}\left(P_{Y|Z}(\cdot \mid Z_{2i-1}), P_{Y|Z}(\cdot \mid Z_{2i})\right).$$

Finally, it is enough to argue that $\frac{1}{n/2}\sum_{i=1}^{n/2} 2\mathrm{d}_{\mathrm{W}_1}\left(P_{Y|Z}(\cdot \mid Z_{2i-1}), P_{Y|Z}(\cdot \mid Z_{2i})\right) = \mathrm{o}_P(1)$. Consider some $N \in \mathbb{N}$ and define $\tilde{Y} = (1/N) \cdot \lfloor NY \rfloor$ which essentially amounts to rounding off the Y value to the closest point in the grid $(\mathbb{N})/N$ to the left of Y. Fix any $i \in [n/2]$. By triangle inequality, we observe that

$$\mathrm{d}_{\mathrm{W}_1}\big(P_{Y|Z}(\cdot \mid Z_{2i-1}), P_{Y|Z}(\cdot \mid Z_{2i})\big) \leq \mathrm{d}_{\mathrm{W}_1}\big(P_{\tilde{Y}|Z}(\cdot \mid Z_{2i-1}), P_{Y|Z}(\cdot \mid Z_{2i-1})\big) + \\ \mathrm{d}_{\mathrm{W}_1}\big(P_{\tilde{Y}|Z}(\cdot \mid Z_{2i-1}), P_{\tilde{Y}|Z}(\cdot \mid Z_{2i})\big) + \mathrm{d}_{\mathrm{W}_1}\big(P_{\tilde{Y}|Z}(\cdot \mid Z_{2i}), P_{Y|Z}(\cdot \mid Z_{2i})\big),$$

where the first and the third term is upper bounded by $1/N$ by construction. Hence,

$$\mathrm{d}_{\mathrm{W}_1}\big(P_{Y|Z}(\cdot \mid Z_{2i-1}), P_{Y|Z}(\cdot \mid Z_{2i})\big) \leq 2/N + \mathrm{d}_{\mathrm{W}_1}\big(P_{\tilde{Y}|Z}(\cdot \mid Z_{2i-1}), P_{\tilde{Y}|Z}(\cdot \mid Z_{2i})\big).$$

Since $\tilde{Y}$ is a discrete random variable, we can further bound the 1-wasserstein metric by the total-variation distance as follows.

$$\mathrm{d}_{\mathrm{W}_1}\big(P_{\tilde{Y}|Z}(\cdot \mid Z_{2i-1}), P_{\tilde{Y}|Z}(\cdot \mid Z_{2i})\big) \leq 2\cdot(1+1/N) \sum_{k=-(N+1)}^{N} \left|P_{\tilde{Y}|Z}\big(k/N \mid Z_{2i-1}\big) - P_{\tilde{Y}|Z}\big(k/N \mid Z_{2i}\big)\right|$$

Above we have a sum of finitely many elements. Hence, it is enough to show that

$$\frac{1}{n/2}\sum_{i=1}^{n/2}\left|P_{\tilde{Y}|Z}\big(k/N \mid Z_{2i-1}\big) - P_{\tilde{Y}|Z}\big(k/N \mid Z_{2i}\big)\right| \leq \frac{1}{n/2}\sum_{i=1}^{n-1}\left|P_{\tilde{Y}|Z}\big(k/N \mid Z_i\big) - P_{\tilde{Y}|Z}\big(k/N \mid Z_{i+1}\big)\right|$$

is $\mathrm{o}_P(1)$ for any integer k in between $-(N+1)$ and N. To prove this, we start with noting that, by bounded difference inequality it holds that

$$\frac{1}{n/2}\sum_{i=1}^{n-1}\left|P_{\tilde{Y}|Z}\big(k/N \mid Z_i\big) - P_{\tilde{Y}|Z}\big(k/N \mid Z_{i+1}\big)\right| \to 2\mathbb{E}\left|P_{\tilde{Y}|Z}\big(k/N \mid Z\big) - P_{\tilde{Y}|Z}\big(k/N \mid Z_{RN}\big)\right|,$$

where Z_{RN} is the immediate neighbour, to the right of Z. Fix an $\epsilon > 0$. Finally by Lusin's theorem, there exists a compactly supported measurable function g_k such that

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

We note that, for any $\delta > 0$

$$\begin{aligned}\mathbb{P}\left\{\left|P_{\tilde{Y}|Z}\big(k/N \mid Z\big) - P_{\tilde{Y}|Z}\big(k/N \mid Z_{RN}\big)\right| \geq \delta\right\} &\leq \mathbb{P}\left\{|g_k(Z) - g_k(Z_{RN})| \geq \delta\right\} \\ &+ \mathbb{P}\left\{P_{\tilde{Y}|Z}\big(k/N \mid Z\big) \neq g_k(Z)\right\} + \mathbb{P}\left\{P_{\tilde{Y}|Z}\big(k/N \mid Z_{RN}\big) \neq g_k(Z_{RN})\right\}\end{aligned}$$

By construction, $\mathbb{P}\left\{P_{\tilde{Y}|Z}\big(k/N \mid Z\big) \neq g_k(Z)\right\} \leq \epsilon$ and by a calculation similar to Azadkia and Chatterjee (2021, Lemma 11.5), there exists a constant C_0 such that

$$\mathbb{P}\left\{P_{\tilde{Y}|Z}\big(k/N \mid Z_{RN}\big) \neq g_k(Z_{RN})\right\} \leq C_0 \mathbb{P}\left\{P_{\tilde{Y}|Z}\big(k/N \mid Z\big) \neq g_k(Z)\right\} \leq C_0\epsilon$$

Finally, by continuity of g_k and Azadkia and Chatterjee (2021, Lemma 11.3) we note that

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

Combining all of these, it holds that

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

Since ϵ, δ are arbitrary, we finally have for any k in between $-(N+1)$ and N,

$$\left|P_{\tilde{Y}|Z}\big(k/N \mid Z\big) - P_{\tilde{Y}|Z}\big(k/N \mid Z_{RN}\big)\right| \xrightarrow{P} 0 \quad \text{as } n \to \infty.$$

Furthermore, since both are probability terms and therefore bounded by 1, it follows that

$$\mathbb{E}\left|P_{\tilde{Y}|Z}\big(k/N \mid Z\big) - P_{\tilde{Y}|Z}\big(k/N \mid Z_{RN}\big)\right| \to 0 \quad \text{as } n \to \infty.$$

This completes the proof. $\square$

Theorem D.18. *For cross bin matching under any distribution $P_{Y,Z}$, it holds that*

$$\left|\frac{1}{n/2}\|\Delta\mathbf{Y}^+\|_2^2 - \frac{1}{K}\sum_{k=1}^{K-1} \mathrm{Dev}(P_{Y|Z_{k\lfloor n/K\rfloor}})\right| \overset{??}{\leq} \frac{1}{K-1}\sum_{i=1}^{K-1} 2\mathrm{d}_{\mathrm{W}_1}\left(P_{Y|Z\in B_k}, P_{Y|Z}(\cdot \mid Z_{k\lfloor n/K\rfloor})\right),$$

where $\mathrm{d}_{\mathrm{W}_1}(\cdot,\cdot)$ is the 1-Wasserstein distance, and $\sigma_z^2 = \mathrm{Var}(Y \mid Z = z)$.

Proof. We start with noting that for cross-bin matching, it holds that

$$\|\Delta\mathbf{Y}^+\|_2^2 = \sum_{k=1}^{K-1}\sum_{\ell\geq 1}(Y_{i_{k,\ell}} - Y_{j_{k+1,\ell}})_+^2,$$

where we assume that there are K bins $B_1, \ldots, B_K$ of approximately uniform mass. Let us call $\hat{P}_Z$ the empirical CDF i.e., $\frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$, and then the k-th bin is defined as

$$B_k = \left\{z : \ \lfloor n/K\rfloor \cdot \hat{P}_Z(z) \in ((k-1)/n, k/n]\right\}$$

Further, the cross-bin matching scheme always ensures that $Z_{i_{k,\ell}} \in B_k,\ Z_{j_{k+1,\ell}} \in B_{k+1}$ for all ℓ, k. Let $F_{Y|Z}(\cdot \mid z)$, and $F_{Y|Z\in B_k}(\cdot)$ denote the CDF of the conditional distribution $P_{Y|Z}(\cdot \mid z)$, and $P_{Y|Z\in B_k}(\cdot)$ respectively. For any $i \in [n]$, if $Z_i \in B_k$ then we define

$$\tilde{Y}_i^R = F_{Y|Z}^{-1}\left(F_{Y|Z\in B_k}(Y_{i_{\ell,k}}) \mid Z_{k\lfloor n/K\rfloor}\right), \quad \text{and} \quad \tilde{Y}_i^L = F_{Y|Z}^{-1}\left(F_{Y|Z\in B_k}(Y_{i_{\ell,k}}) \mid Z_{k-1\lfloor n/K\rfloor}\right)$$

Without loss of generality, we assume that $|Y|$ is almost surely bounded by 1. Since, $t \to t_+^2$ is 2-lipschitz for $t \in [-1, 1]$ it holds that

$$\begin{aligned}\left|(Y_{i_{k,\ell}} - Y_{j_{k+1,\ell}})_+^2 - (\tilde{Y}^R_{i_{k,\ell}} - \tilde{Y}^L_{j_{k+1,\ell}})_+^2\right| &\leq 2\left(\left|Y_{i_{\ell,k}} - \tilde{Y}^R_{i_{\ell,k}}\right| + \left|Y_{j_{\ell,k+1}} - \tilde{Y}^L_{j_{\ell,k+1}}\right|\right) \\ &\leq 2\left(\left|Y_{i_{\ell,k}} - \tilde{Y}^L_{i_{\ell,k}}\right| + \left|Y_{i_{\ell,k}} - \tilde{Y}^R_{i_{\ell,k}}\right| + \left|Y_{j_{\ell,k+1}} - \tilde{Y}^L_{j_{\ell,k+1}}\right| + \left|Y_{j_{\ell,k+1}} - \tilde{Y}^R_{j_{\ell,k+1}}\right|\right).\end{aligned}$$

Combining the above inequalities for every matched pairs, it follows that

$$\left|\frac{1}{n/2}\|\Delta\mathbf{Y}^+\|_2^2 - \frac{1}{n/2}\sum_{k=1}^{K-1}\sum_{\ell\geq 1}(\tilde{Y}^R_{i_{k,\ell}} - \tilde{Y}^L_{j_{k+1,\ell}})_+^2\right| \leq \frac{2}{n}\sum_{i=1}^{n}\left(\left|Y_i - \tilde{Y}_i^L\right| + \left|Y_i - \tilde{Y}_i^R\right|\right).$$

Since Y is bounded, $|Y_i - \tilde{Y}_i^L|$ is bounded by 2 and thus, by Hoeffding's inequality we note that

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \tilde{Y}_i^L\right| - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \tilde{Y}_i^L\right|\right] \geq \delta \mid Z_1, \ldots, Z_n\right\} \leq 2\exp\left\{-n\delta^2/2\right\}.$$

Furthermore, conditioned on $Z_1, \ldots, Z_n$ we observe that $\mathbb{E}\left|Y_i - \tilde{Y}_i^L\right|$ can be computed as $\mathrm{d}_{\mathrm{W}_1}\left(P_{Y|Z\in B_k}, P_{Y|Z}(\cdot \mid Z_{k-1\lfloor n/K\rfloor})\right)$ if $Z_i \in B_k$ and thus,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \tilde{Y}_i^L\right|\right] = \frac{1}{K}\sum_{k=1}^{K}\frac{|\{i : Z_i \in B_k\}|}{n/K}\mathrm{d}_{\mathrm{W}_1}\left(P_{Y|Z\in B_k}, P_{Y|Z}(\cdot \mid Z_{k-1\lfloor n/K\rfloor})\right)$$

Finally, we note that for $k < K$, $|\{i : Z_i \in B_k\}| = \lfloor n/K\rfloor \leq n/K$, and for $k = K$ $|\{i : Z_i \in B_k\}| \leq 2n/K$, and thus, it follows that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \tilde{Y}_i^L\right|\right] \leq 2\max_{k\in[K]}\mathrm{d}_{\mathrm{W}_1}\left(P_{Y|Z\in B_k}, P_{Y|Z}(\cdot \mid Z_{k-1\lfloor n/K\rfloor})\right).$$

By a similar argument,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \tilde{Y}_i^R\right|\right] \leq 2\max_{k\in[K]}\mathrm{d}_{\mathrm{W}_1}\left(P_{Y|Z\in B_k}, P_{Y|Z}(\cdot \mid Z_{k\lfloor n/K\rfloor})\right).$$

Hence, it holds that

$$\begin{aligned}\left|\frac{1}{n/2}\|\Delta\mathbf{Y}^+\|_2^2 - \frac{1}{n/2}\sum_{k=1}^{K-1}\sum_{\ell\geq 1}(\tilde{Y}^R_{i_{k,\ell}} - \tilde{Y}^L_{j_{k+1,\ell}})_+^2\right| \leq 4\Bigg(&\max_{k\in[K]}\mathrm{d}_{\mathrm{W}_1}\left(P_{Y|Z\in B_k}, P_{Y|Z}(\cdot \mid Z_{k-1\lfloor n/K\rfloor})\right) \\ &+ \max_{k\in[K]}\mathrm{d}_{\mathrm{W}_1}\left(P_{Y|Z\in B_k}, P_{Y|Z}(\cdot \mid Z_{k\lfloor n/K\rfloor})\right)\Bigg).\end{aligned}$$

Next we observe that

$$\left|\frac{1}{n/2}\sum_{k=1}^{K-1}\sum_{\ell\geq 1}(\tilde{Y}^R_{i_{k,\ell}}-\tilde{Y}^L_{j_{k+1,\ell}})^2_+-\frac{1}{n/2}\sum_{k=1}^{K-1}\sum_{\ell\geq 1}(\tilde{Y}^R_{i_{k,\ell}}-\tilde{Y}^{R,M}_{i_{k,\ell}})^2_+\right|\leq\frac{2}{n}\sum_{i=1}^{n}\left(\left|Y_i-\tilde{Y}^L_i\right|+\left|Y_i-\tilde{Y}^R_i\right|\right).$$

□

Lemma D.19. *Let $X_{(1)}\leq X_{(2)}\leq\ldots\leq X_{(n)}$ be the the order statistics for $\{X_i\}_{0\leq i\leq n}\overset{\text{iid}}{\sim}P$, for some distribution P with bounded support and CDF F. Let us define*

$$X^M_{(i)}=F^{-1}\big(1-F(X_{(n+1-i)}\big),$$

which is the i-th order statistic, reflected at the median of F, and furthermore let us denote the mirrored observation for any X by X^M. Then, it holds that

$$\frac{1}{n}\sum_{i=1}^{n}|X_i-X^M_i|\to 0 \text{ almost surely, as } n\to\infty.$$

D.2 Proof of Lemmas

Proof of Lemma C.16

Proof. Under our working model (14), $\Delta\mathbf{w}\circ\Delta\mathbf{X}$ admits the simple form $\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z})+\mathbf{w}\circ\Delta\boldsymbol{\zeta}$, and thus,

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

Further, conditioned on both $\mathbf{Y}$ and $\mathbf{Z}$, both $\|\Delta\mathbf{w}\circ\Delta\boldsymbol{\zeta}\|_2^2$, and $(\Delta\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z}))^T(\Delta\mathbf{w}\circ\Delta\boldsymbol{\zeta})$ are weighted sum of i.i.d. random variables. Hence, the means and variances of these quantities satisfy the following.

$$\mathbb{E}\{\|\Delta\mathbf{w}\circ\Delta\boldsymbol{\zeta}\|_2^2\}=\rho_2^2\sum_{\ell=1}^{L_n}(\Delta_\ell\mathbf{w})^2=\rho_2^2\|\Delta\mathbf{w}\|_2^2,\quad\mathbb{E}\{(\Delta\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z}))^T(\Delta\mathbf{w}\circ\Delta\boldsymbol{\zeta})\}=0$$

$$\mathrm{Var}\{\|\Delta\mathbf{w}\circ\Delta\boldsymbol{\zeta}\|_2^2\}=\sum_{\ell=1}^{L_n}(\Delta_\ell\mathbf{w})^4\mathrm{Var}((\Delta_\ell\boldsymbol{\zeta})^2)\leq\sum_{\ell=1}^{L_n}(\Delta_\ell\mathbf{w})^4\mathbb{E}((\Delta_\ell\boldsymbol{\zeta})^4)=\|\Delta\mathbf{w}\|_4^4\rho_4^4,$$

$$\mathrm{Var}\{(\Delta\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z}))^T(\Delta\mathbf{w}\circ\Delta\boldsymbol{\zeta})\}=\sum_{\ell=1}^{L_n}(\Delta_\ell\mathbf{w})^4(\Delta_\ell\mu(\mathbf{Y},\mathbf{Z}))^2\mathbb{E}((\Delta_\ell\boldsymbol{\zeta})^2)=\rho_2^2\|(\Delta\mathbf{w})^2\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2^2.$$

By Chebychev's inequality, for any $\delta\in\mathbb{R}_{\geq 0}$, it holds that

$$\mathbb{P}\left\{\left|\|\Delta\mathbf{w}\circ\Delta\boldsymbol{\zeta}\|_2^2-\rho_2^2\|\Delta\mathbf{w}\|_2^2\right|\geq\sqrt{\frac{\|\Delta\mathbf{w}\|_4^4\rho_4^4}{\delta}}\mid\mathbf{Y},\mathbf{Z}\right\}\leq\delta,$$

$$\mathbb{P}\left\{\left|(\Delta\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z}))^T(\Delta\mathbf{w}\circ\Delta\boldsymbol{\zeta})\right|\geq\sqrt{\frac{\rho_2^2\|(\Delta\mathbf{w})^2\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2^2}{\delta}}\mid\mathbf{Y},\mathbf{Z}\right\}\leq\delta.$$

Hence, conditioned on both $\mathbf{Y}$ and $\mathbf{Z}$, each of the following two statements holds with probability at least $1-2\delta$.

$$\|\Delta\mathbf{w}\circ\Delta\mathbf{X}\|_2^2 \le \|\Delta\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2^2 + \rho_2^2\|\Delta\mathbf{w}\|_2^2 + \frac{\rho_4^2}{\sqrt{\delta}}\|\Delta\mathbf{w}\|_4^2 + \frac{2\rho_2}{\sqrt{\delta}}\|(\Delta\mathbf{w})^2\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2,$$

$$\|\Delta\mathbf{w}\circ\Delta\mathbf{X}\|_2^2 \ge \|\Delta\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2^2 + \rho_2^2\|\Delta\mathbf{w}\|_2^2 - \frac{\rho_4^2}{\sqrt{\delta}}\|\Delta\mathbf{w}\|_4^2 - \frac{2\rho_2}{\sqrt{\delta}}\|(\Delta\mathbf{w})^2\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2.$$

Observe, $\|\Delta\mathbf{w}\|_4^2 \le \|\Delta\mathbf{w}\|_\infty\cdot\|\Delta\mathbf{w}\|_2$, and similarly $\|(\Delta\mathbf{w})^2\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2 \le \|\Delta\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty\cdot\|\Delta\mathbf{w}\|_2$. Thus, we further have

$$\mathbb{P}\left\{\|\Delta\mathbf{w}\circ\Delta\mathbf{X}\|_2^2 \le \rho_2^2\|\Delta\mathbf{w}\|_2^2\cdot\left(1+\frac{\|\Delta\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_2^2}{\rho_2^2\|\Delta\mathbf{w}\|_2^2} + \frac{\rho_4^2}{\rho_2^2\sqrt{\delta}}\frac{\|\Delta\mathbf{w}\|_\infty}{\|\Delta\mathbf{w}\|_2} + \frac{2}{\rho_2\sqrt{\delta}}\frac{\|\Delta\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty}{\|\Delta\mathbf{w}\|_2}\right) \mid \mathbf{Y},\mathbf{Z}\right\} \ge 1-2\delta.$$

Hence, by definition of $\epsilon_{1,\delta,U}$,

$$\mathbb{P}\left\{\|\Delta\mathbf{w}\circ\Delta\mathbf{X}\|_2^2 \le \rho_2^2\|\Delta\mathbf{w}\|_2^2\cdot(1+\epsilon_{1,\delta,U}) \mid \mathbf{Y},\mathbf{Z}\right\} \le 1-2\delta.$$

Imitating calculations for the upper bounds, the same also holds for the lower bpund implying

$$\mathbb{P}\left\{\|\Delta\mathbf{w}\circ\Delta\mathbf{X}\|_2^2 \ge \rho_2^2\|\Delta\mathbf{w}\|_2^2\cdot(1+\epsilon_{1,\delta,L}) \mid \mathbf{Y},\mathbf{Z}\right\} \le 1-2\delta,$$

where $\epsilon_{1,\delta,L}$ is as defined in (22).

Similarly, $\|\Delta\mathbf{w}\circ\Delta\boldsymbol{\zeta}\|_3^3$ too is a weighted sum of i.i,.d. random variables, and has mean and variance given by

$$\mathbb{E}\{\|\Delta\mathbf{w}\circ\Delta\boldsymbol{\zeta}\|_3^3\} = \sum_{\ell=1}^{L}(|\Delta_\ell\mathbf{w}|)^3\rho_3^3 = \rho_3^3\|\Delta\mathbf{w}\|_3^3,$$

$$\mathrm{Var}\{\|\Delta\mathbf{w}\circ\Delta\boldsymbol{\zeta}\|_3^3\} = \sum_{\ell=1}^{L}(\Delta_\ell\mathbf{w})^6\mathrm{Var}((|\Delta_\ell\boldsymbol{\zeta}|)^3) \le \sum_{\ell=1}^{L}(\Delta_\ell\mathbf{w})^6\mathbb{E}((\Delta_\ell\boldsymbol{\zeta})^6) = \|\Delta\mathbf{w}\|_6^6\rho_6^6.$$

Thus, by Chebychev's inequality

$$\mathbb{P}\left\{\left|\|\Delta\mathbf{w}\circ\Delta\boldsymbol{\zeta}\|_3^3 - \rho_3^3\|\Delta\mathbf{w}\|_3^3\right| \ge \sqrt{\frac{\|\Delta\mathbf{w}\|_6^6\rho_6^6}{\delta}} \mid \mathbf{Y},\mathbf{Z}\right\} \le \delta \text{ or equivalently,}$$

$$\mathbb{P}\left\{\|\Delta\mathbf{w}\circ\Delta\boldsymbol{\zeta}\|_3^3 \le \rho_3^3\|\Delta\mathbf{w}\|_3^3 + \frac{\rho_6^3}{\sqrt{\delta}}\cdot|\Delta\mathbf{w}\|_6^3 \mid \mathbf{Y},\mathbf{Z}\right\} \le \delta$$

and hence, by a simple application of trianagle inequality we have

$$\mathbb{P}\left\{\|\Delta\mathbf{w}\circ\Delta\mathbf{X}\|_3 \le \|\Delta\mathbf{w}\circ\Delta\mu(\mathbf{Y},\mathbf{Z})\|_3 + \left(\rho_3^3\|\Delta\mathbf{w}\|_3^3 + \frac{\rho_6^3}{\sqrt{\delta}}\cdot\|\Delta\mathbf{w}\|_6^3\right)^{1/3} \mid \mathbf{Y},\mathbf{Z}\right\} \ge 1-\delta.$$

Finally observe, $\|\Delta\mathbf{w}\|_3^3 \leq \|\Delta\mathbf{w}\|_\infty \cdot \|\Delta\mathbf{w}\|_2^2$, and $\|\Delta\mathbf{w}\|_6^3 \leq \|\Delta\mathbf{w}\|_\infty^2 \cdot \|\Delta\mathbf{w}\|_2$ and similarly, $\|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_3^3 \leq \|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2^2 \cdot \|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty$. As a result, we have

$$\mathbb{P}\Bigg\{\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_3 \leq \rho_2\|\Delta\mathbf{w}\|_2 \cdot \Bigg(\left(\frac{\|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_2}{\rho_2\|\Delta\mathbf{w}\|_2}\right)^{2/3} \cdot \left(\frac{\|\Delta\mathbf{w} \circ \Delta\mu(\mathbf{Y},\mathbf{Z})\|_\infty}{\rho_2\|\Delta\mathbf{w}\|_2}\right)^{1/3} + \left(\frac{\rho_3^3}{\rho_2^3}\frac{\|\Delta\mathbf{w}\|_\infty}{\|\Delta\mathbf{w}\|_2} + \frac{\rho_6^3}{\rho_2^3\sqrt{\delta}} \cdot \left(\frac{\|\Delta\mathbf{w}\|_\infty}{\|\Delta\mathbf{w}\|_2}\right)^2\right)^{1/3}\Bigg) \mid \mathbf{Y},\mathbf{Z}\Bigg\} \geq 1-\delta$$

Combining the above result, with the high probability lower bound on $\|\Delta\mathbf{w} \circ \Delta\mathbf{X}\|_2$, the final result holds too.

□

Proof of Lemma 6

Proof. We start with observing that the Hellinger distance between two Bernoulli distributions can be computed as

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

Since the cross terms in the denominators are positive, the distance can be further upper bounded by

$$\frac{1}{2}\left[\frac{(p-q)^2}{p+q} + \frac{(p-q)^2}{(1-p)+(1-q)}\right] \leq \frac{1}{2}\left[\frac{(p-q)^2}{p} + \frac{(p-q)^2}{(1-p)}\right] \leq \frac{(p-q)^2}{2p(1-p)}.$$

Since, $H^2\left(\mathrm{Ber}(p),\mathrm{Ber}(q)\right)$ is symmetric, it is also bounded by $\frac{(p-q)^2}{2q(1-q)}$ by a similar argument. This proves the lemma. □

Proof of Lemma A.12

Proof. We start with noting that $\Delta\mu^+$ is a vector of size L with entries $\max\{0, \mu(Y_{i_\ell}, Z_{i_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell})\}$ where $i_\ell < j_\ell$. Now, for any $a,b,c \in \mathbb{R}$ note that

$$|a-b| \leq \begin{cases} |a-c| \vee |b-c| & \text{if } (a-c)(b-c) \geq 0 \\ |a-c|+|b-c| & \text{otherwise} \end{cases} \tag{38}$$

Note, $\Delta\mu_\ell^+$ is positive only when $\mu(Y_{i_\ell}, Z_{i_\ell}) > \mu(Y_{j_\ell}, Z_{j_\ell})$. So, we will focus on this case only. Now, if $\mu(Y_{i_\ell}, Z_{i_\ell}) \geq \widehat{\mu}_{\mathrm{ISO}}(Z_{i_\ell}) \geq \mu(Y_{j_\ell}, Z_{j_\ell})$, then further

$$\left|\widehat{\mu}_{\mathrm{ISO}}(Z_{i_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell})\right| \leq \left|\widehat{\mu}_{\mathrm{ISO}}(Z_{j_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell})\right|$$

since $Z_{i_\ell} \preccurlyeq Z_{j_\ell}$ and $\widehat{\mu}_{\text{ISO}} \in \text{Mon}_n(\mathcal{Z})$. Hence, it implies from (38) that

$$|\mu(Y_{i_\ell}, Z_{i_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell})| \leq \left|\mu(Y_{i_\ell}, Z_{i_\ell}) - \widehat{\mu}_{\text{ISO}}(Z_{i_\ell})\right| + \left|\widehat{\mu}_{\text{ISO}}(Z_{j_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell})\right|.$$

Otherwise, using the first case of (38), we have

$$|\mu(Y_{i_\ell}, Z_{i_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell})| \leq \left|\mu(Y_{i_\ell}, Z_{i_\ell}) - \widehat{\mu}_{\text{ISO}}(Z_{i_\ell})\right| \vee \left|\widehat{\mu}_{\text{ISO}}(Z_{j_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell})\right|.$$

Combining both cases, $|\mu(Y_{i_\ell}, Z_{i_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell})|$ is always bounded by $\left|\mu(Y_{i_\ell}, Z_{i_\ell}) - \widehat{\mu}_{\text{ISO}}(Z_{i_\ell})\right| + \left|\widehat{\mu}_{\text{ISO}}(Z_{j_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell})\right|$. Now, all the results follow from here using the inequalities

$$(|a| + |b|)^2 \leq 2(|a|^2 + |b|^2) \text{ and } (|a| + |b|)^4 \leq 8(|a|^4 + |b|^4) \text{ for any } a, b.$$

□

Proof of Lemma D.19

Proof. We start with claiming that $\max_{i=1,\ldots,n} \left|F(X_{(i)}) - 1 + F(X_{(n+1-i)})\right| \to 0$ almost surely, as $n \to \infty$. This follows by simply applying triangle inequality, and noting that

$$\max_{i=1,\ldots,n} \left|F(X_{(i)}) - 1 + F(X_{(n+1-i)})\right| \leq \max_{i=1,\ldots,n} \left\{\left|F(X_{(i)}) - \widehat{F}_n(X_{(i)})\right| + \left|F(X_{(n+1-i)}) - \widehat{F}_n(X_{(n+1-i)})\right| + \left|\widehat{F}_n(X_{(i)}) - 1 + \widehat{F}_n(X_{(n+1-i)})\right|\right\},$$

where $\widehat{F}_n = \frac{1}{n}\delta_{X_i}$ be the empirical CDF. We further note that the first two terms almost surely tend to 0 by Glivenko-Cantelli theorem and finally, for the last term

$$\left|\widehat{F}_n(X_{(i)}) - 1 + \widehat{F}_n(X_{(n+1-i)})\right| = \left|\frac{i}{n} - 1 + \frac{n+1-i}{n}\right| = 1/n.$$

Combining everything, the claim holds. By definition, $F(X_i^M) = 1 - F(X_{(n+1-i)}$ and thus,

$$\max_{i=1,\ldots,n} \left|F(X_i) - F(X_i^M)\right| \to 0 \text{ almost surely.}$$

Finally, consider the sum $S_n(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^n |X_i - X_i^M|$ and furthermore, observe that the sum S_n satisfies the bounded difference assumption i.e.,

$$\sup_{x_1,\ldots,x_n,x_i'} |S_n(x_1, \ldots, x_i, \ldots, x_n) - S_n(x_1, \ldots, x_i', \ldots, x_n)| \leq 4B/n.$$

Thus by the bounded difference inequality, we have

$$S_n \xrightarrow{P} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n |X_i - X_i^M|\right] = \mathbb{E}|X - X^M|$$

Since X and X^M are both bounded, in order to argue that $\mathbb{E}|X - X^M| \to 0$, it is enough to show that $X - X^M \xrightarrow{P} 0$. Towards that, we note that by Lusin's theorem there exists a compactly supported measurable function g such that

$$\mathbb{P}_{U \overset{d}{=} F(X)} \left\{F^{-1}(U) \neq g(U)\right\} < \epsilon$$

for any given $\epsilon > 0$. We denote $F(X), F(X^M)$ by U, U^M respectively, and then note that for any $\delta > 0$

$$\begin{aligned}
\mathbb{P}\left\{\left|X - X^M\right| \geq \delta\right\} &= \mathbb{P}\left\{\left|F^{-1}(U) - F^{-1}(U^M)\right| \geq \delta\right\} \\
&\leq \mathbb{P}\left\{\left|g(U) - g(U^M)\right| \geq \delta\right\} + \mathbb{P}\left\{F^{-1}(U) \neq g(U)\right\} + \mathbb{P}\left\{F^{-1}(U^M) \neq g(U^M)\right\}
\end{aligned}$$

By construction, $\mathbb{P}\{F^{-1}(U) \neq g(U)\} < \epsilon$ and also, we observe that by symmetry

$$\mathbb{P}\left\{F^{-1}(U^M) \neq g(U^M)\right\} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left\{F^{-1}(U_i^M) \neq g(U_i^M)\right\}$$

Since the sets $\{U_1, \ldots, U_n\}$ and $\{U_1^M, \ldots, U_n^M\}$ are identical, it further holds that, by symmetry,

$$\mathbb{P}\left\{F^{-1}(U^M) \neq g(U^M)\right\} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left\{F^{-1}(U_i) \neq g(U_i)\right\} = \mathbb{P}\left\{F^{-1}(U) \neq g(U)\right\} < \epsilon$$

Finally, we recall that $|U - U^M| \to 0$ almost surely, as $n \to \infty$. Thus, by continuity of g and Azadkia and Chatterjee (2021, Lemma 11.3) we note that

$$\lim_{n\to\infty} \mathbb{P}\left\{\left|g(U) - g(U^M)\right| \geq \delta\right\} = 0.$$

Combining all of these calculations, it holds that $\lim_{n\to\infty} \mathbb{P}\left\{\left|X - X^M\right| \geq \delta\right\} \leq \epsilon(1 + C_0)$. Since ϵ, δ are arbitrary, it holds that $X - X^M \xrightarrow{P} 0$, and this completes the proof. $\square$