Bai, Yuehao; Romano, Joseph P.; Shaikh, Azeem M.

Working Paper

Inference in experiments with matched pairs

cemmap working paper, No. CWP19/19

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Bai, Yuehao; Romano, Joseph P.; Shaikh, Azeem M. (2019) : Inference in experiments with matched pairs, cemmap working paper, No. CWP19/19, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2019.1919

This Version is available at:
https://hdl.handle.net/10419/211112

cemmap
centre for microdata methods and practice

Inference in Experiments with Matched Pairs

Yuehao Bai
Joseph P. Romano
Azeem M. Shaikh

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP19/19

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Inference in Experiments with Matched Pairs*

Yuehao Bai
Department of Economics
University of Chicago
ybai@uchicago.edu

Joseph P. Romano
Departments of Economics and Statistics
Stanford University
romano@stanford.edu

Azeem M. Shaikh
Department of Economics
University of Chicago
amshaikh@uchicago.edu

April 23, 2019

Abstract

This paper studies inference for the average treatment effect in randomized controlled trials where treatment status is determined according to a "matched pairs" design. By a "matched pairs" design, we mean that units are sampled i.i.d. from the population of interest, paired according to observed, baseline covariates and finally, within each pair, one unit is selected at random for treatment. This type of design is used routinely throughout the sciences, but results about its implications for inference about the average treatment effect are not available. The main requirement underlying our analysis is that pairs are formed so that units within pairs are suitably "close" in terms of the baseline covariates, and we develop novel results to ensure that pairs are formed in a way that satisfies this condition. Under this assumption, we show that, for the problem of testing the null hypothesis that the average treatment effect equals a pre-specified value in such settings, the commonly used two-sample t-test and "matched pairs" t-test are conservative in the sense that these tests have limiting rejection probability under the null hypothesis no greater than and typically strictly less than the nominal level. We show, however, that a simple adjustment to the standard errors of these tests leads to a test that is asymptotically exact in the sense that its limiting rejection probability under the null hypothesis equals the nominal level. We also study the behavior of randomization tests that arise naturally in these types of settings. When implemented appropriately, we show that this approach also leads to a test that is asymptotically exact in the sense described previously, but additionally has finite-sample rejection probability no greater than the nominal level for certain distributions satisfying the null hypothesis. A simulation study confirms the practical relevance of our theoretical results.

KEYWORDS: Experiment, matched pairs, matched pairs t-test, permutation test, randomized controlled trial, treatment assignment, two-sample t-test

JEL classification codes: C12, C14

*We thank John Duchi for references to the "blossom" algorithm and Panos Toulis for helpful comments. The research of the third author is supported by NSF Grant SES-1530661.

1 Introduction

This paper studies inference for the average treatment effect in randomized controlled trials where treatment status is determined according to a "matched pairs" design. By a "matched pairs" design, we mean that units are sampled i.i.d. from the population of interest, paired according to observed, baseline covariates and finally, within each pair, one unit is selected at random for treatment. This method is used routinely in all parts of the sciences. Indeed, commands to facilitate its implementation are included in popular software packages, such as `sampsi` in Stata. References to a variety of specific examples can be found, for instance, in the following surveys of various field experiments: Riach and Rich (2002), List and Rasul (2011), White (2013), Crépon et al. (2015), Bertrand and Duflo (2017), and Heard et al. (2017). See also Bruhn and McKenzie (2009), who, based on a survey of selected development economists, report that 56% of researchers have used such a design at some point. Despite the widespread use of "matched pairs" designs, results about its implications for inference about the average treatment effect are not available. The main requirement underlying our analysis is that pairs are formed so that units within pairs are suitably "close" in terms of the baseline covariates. We develop novel results to ensure that pairs are formed in a way that satisfies this condition. See, in particular, Theorems 4.1–4.3 below. Under this assumption, we derive a variety of results pertaining to the problem of testing the null hypothesis that the average treatment effect equals a pre-specified value in such settings.

We first study the behavior of the two-sample t-test and "matched pairs" t-test, which are both used routinely in the analysis of this type of data. Several specific references are provided in Sections 3.1 and 3.2 below. Our first pair of results establish that these commonly used tests are conservative in the sense that these tests have limiting rejection probability under the null hypothesis no greater than and typically strictly less than the nominal level. For each of these tests, we additionally provide a characterization of when the limiting rejection probability under the null hypothesis is in fact strictly less than the nominal level. In a simulation study, we find that the rejection probability of these tests may in fact be dramatically less than the nominal level, and, as a result, they may have very poor power when compared to other tests. Intuitively, the conservative feature of these tests is a consequence of the dependence in treatment status across units and between treatment status and baseline covariates resulting from the "matched pairs" design. We show, however, that a simple adjustment to the usual standard error of these tests leads to a test that is asymptotically exact in the sense that its limiting rejection probability under the null hypothesis equals the nominal level.

Next, we study the behavior of some randomization tests that arise naturally in these types of settings. More specifically, we study randomization tests based on the idea of permuting only treatment status for units within pairs. When implemented with a suitable choice of test statistic, we show that this approach also leads to a test that is asymptotically exact in the sense described previously. We emphasize, however, that this result relies heavily upon the choice of test statistic. Indeed, as explained further in Remark 3.11, when implemented with other choices of test statistics, randomization tests may behave in large samples like the "matched pairs" t-test described above. On the other hand, regardless of the specific way in which they are implemented, these tests have the attractive feature that they have finite-sample rejection probability

no greater than the nominal level for certain distributions satisfying the null hypothesis. We highlight these properties in a simulation study.

The remainder of the paper is organized as follows. In Section 2, we describe our setup and notation. In particular, there we describe the precise sense in which we require that units in each pair are "close" in terms of their baseline covariates. Our main results concerning the two-sample t-test, the "matched pairs" t-test, and randomization tests are contained in Section 3. In Section 4, we develop some results that ensure that units in each pair are suitably "close" in terms of their baseline covariates. Finally, in Section 5, we examine the finite-sample behavior of these tests via a small simulation study. Proofs of all results are provided in the Appendix.

2 Setup and Notation

Let Y_i denote the (observed) outcome of interest for the ith unit, D_i denote the treatment status of the ith unit, and X_i denote observed, baseline covariates for the ith unit. Further denote by $Y_i(1)$ the potential outcome of the ith unit if treated and by $Y_i(0)$ the potential outcome of the ith unit if not treated. As usual, the (observed) outcome and potential outcomes are related to treatment status by the relationship

$$Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i) \ . \tag{1}$$

For a random variable indexed by i, A_i, it will be useful to denote by $A^{(n)}$ the random vector $(A_1, \ldots, A_{2n})$. Denote by P_n the distribution of the observed data $Z^{(n)}$, where $Z_i = (Y_i, D_i, X_i)$, and by Q_n the distribution of $W^{(n)}$, where $W_i = (Y_i(1), Y_i(0), X_i)$. Note that P_n is jointly determined by (1), Q_n, and the mechanism for determining treatment assignment. We assume throughout that $W^{(n)}$ consists of $2n$ i.i.d. observations, i.e., $Q_n = Q^{2n}$, where Q is the marginal distribution of W_i. We therefore state our assumptions below in terms of assumptions on Q and the mechanism for determining treatment assignment. Indeed, we will not make reference to P_n in the sequel and all operations are understood to be under Q and the mechanism for determining treatment assignment.

Our object of interest is the average effect of the treatment on the outcome of interest, which may be expressed in terms of this notation as

$$\Delta(Q) = E[Y_i(1) - Y_i(0)] \ . \tag{2}$$

For a pre-specified choice of Δ_0, the testing problem of interest is

$$H_0 : \Delta(Q) = \Delta_0 \text{ versus } H_1 : \Delta(Q) \neq \Delta_0 \tag{3}$$

at level $\alpha \in (0, 1)$.

We now describe our assumptions on Q. We restrict Q to satisfy the following mild requirement:

Assumption 2.1. The distribution Q is such that

(a) $0 < E[\mathrm{Var}[Y_i(d)|X_i]]$ for $d \in \{0,1\}$.

(b) $E[Y_i^2(d)] < \infty$ for $d \in \{0,1\}$.

(c) $E[Y_i(d)|X_i = x]$ and $E[Y_i^2(d)|X_i = x]$ are Lipschitz for $d \in \{0,1\}$.

Assumptions 2.1(a)–(b) are mild restrictions imposed, respectively, to rule out degenerate situations and to permit the application of suitable laws of large numbers and central limit theorems. See, in particular, Lemma 6.3 in the Appendix for a novel law of large numbers for independent and non-identically distributed random variables that is useful in establishing our results. Assumption 2.1(c), on the other hand, is a smoothness requirement that ensures that units that are "close" in terms of their baseline covariates are suitably comparable.

Next, we describe our assumptions on the mechanism determining treatment assignment. In order to describe these assumptions more formally, we require some further notation to define the relevant pairs of units. The n pairs may be represented by the sets

$$\{\pi(2j-1), \pi(2j)\} \text{ for } j = 1, \ldots, n \ ,$$

where $\pi = \pi_n(X^{(n)})$ is a permutation of $2n$ elements. Because of its possible dependence on $X^{(n)}$, π encompasses a broad variety of different ways of pairing the $2n$ units according to the observed, baseline covariates $X^{(n)}$. Given such a π, we assume that treatment status is assigned as described in the following assumption:

Assumption 2.2. Conditional on $X^{(n)}$, $(D_{\pi(2j-1)}, D_{\pi(2j)}), j = 1, \ldots, n$ are i.i.d. and each uniformly distributed over the values in $\{(0,1),(1,0)\}$.

Our analysis will require some discipline on the way in which the pairs are formed. In particular, we will require that the units in each pair are "close" in terms of their baseline covariates in the sense described by the following assumption:

Assumption 2.3. The pairs used in determining treatment status satisfy

$$\frac{1}{n} \sum_{1 \le j \le n} |X_{\pi(2j)} - X_{\pi(2j-1)}|^r \xrightarrow{P} 0$$

for $r = 1$ and $r = 2$.

It will at times be convenient to require further that units in consecutive pairs are also "close" in terms of their baseline covariates. One may view this requirement, which is formalized in the following assumption, as "pairing the pairs" so that they are "close" in terms of their baseline covariates.

Assumption 2.4. The pairs used in determining treatment status satisfy

$$\frac{1}{n} \sum_{1 \le j \le \lfloor \frac{n}{2} \rfloor} |X_{\pi(4j-k)} - X_{\pi(4j-\ell)}|^2 \xrightarrow{P} 0$$

for any $k \in \{2,3\}$ and $\ell \in \{0,1\}$.

In Section 4 below, we provide results to facilitate constructing pairs satisfying Assumptions 2.3–2.4 under weak assumptions on Q. We emphasize, however, that Assumption 2.4, in contrast to Assumptions 2.1–2.3, will not be required for many of our results. Furthermore, given pairs satisfying Assumption 2.3, it will frequently be possible to "re-order" them so that Assumption 2.4 is satisfied. See Theorem 4.3 below for further details.

Remark 2.1. Note that Assumption 2.2 implies that

$$(Y^{(n)}(1), Y^{(n)}(0)) \perp\!\!\!\perp D^{(n)} \big| X^{(n)} \ . \tag{4}$$

In this sense, treatment status is determined exogenously conditional on $X^{(n)}$. ■

Remark 2.2. At the expense of some additional notation, it is straightforward to allow π to depend further on a uniform random variable U that is independent of $(Y^{(n)}(1), Y^{(n)}(0), X^{(n)})$, but we do not pursue this generalization here. ■

Remark 2.3. The treatment assignment scheme described in this section is an example of what is termed in some parts of the literature as a covariate-adaptive randomization scheme, in which treatment status is assigned so as to "balance" units assigned to treatment and the units assigned to control in terms of their baseline covariates. For a review of these types of treatment assignment schemes focused on their use in clinical trials, see Rosenberger and Lachin (2015). In some such schemes, units are sampled i.i.d. from the population of interest, stratified into a finite number of strata according to observed, baseline covariates, and finally, within each stratum, treatment status is assigned so as to achieve "balance" within each stratum. For instance, within each stratum, a researcher may assign (uniformly) at random half of the units to treatment and the remainder to control. Bugni et al. (2018, 2019) develop a variety of results pertaining to these ways of assigning treatment status, but their analysis relies heavily upon the requirement that the units are stratified using the baseline covariates into only a finite number of strata. As a result, their framework cannot accomodate "matched pairs" designs, where the number of strata is equal to the number of pairs and therefore proportional to the sample size. ■

3 Main Results

3.1 Two-Sample t-Test

In this section, we consider using the two-sample t-test to test (3) at level $\alpha \in (0,1)$. In order to define this test, for $d \in \{0,1\}$, define

$$\hat{\mu}_n(d) = \frac{1}{n} \sum_{1 \le i \le 2n : D_i = d} Y_i \tag{5}$$

$$\hat{\sigma}_n^2(d) = \frac{1}{n} \sum_{1 \le i \le 2n : D_i = d} (Y_i - \hat{\mu}_n(d))^2 \tag{6}$$

and let

$$\hat{\Delta}_n = \hat{\mu}_n(1) - \hat{\mu}_n(0) \ . \tag{7}$$

The two-sample t-test is given by

$$\phi_n^{t-\text{test}}(Z^{(n)}) = I\{|T_n^{t-\text{test}}(Z^{(n)})| > z_{1-\frac{\alpha}{2}}\} \ , \tag{8}$$

where

$$T_n^{t-\text{test}}(Z^{(n)}) = \frac{\sqrt{n}(\hat{\Delta}_n - \Delta_0)}{\sqrt{\hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0)}} \tag{9}$$

and $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution. While its properties are far from clear in our setting, this classical test is used routinely in the analysis of such data. See, for example, Riach and Rich (2002), Gelman and Hill (2006, page 174), Duflo et al. (2007), Bertrand and Duflo (2017) and the references therein. See also Imai et al. (2009) for the use of an analgous test in a setting with cluster-level treatment assignment.

The following theorem establishes the asymptotic behavior of the two-sample t-statistic defined in (9) and, as a consequence, the two-sample t-test defined in (8). In particular, the theorem shows that the limiting rejection probability of the two-sample t-test under the null hypothesis is generally strictly less than the nominal level.

Theorem 3.1. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumptions 2.2–2.3. Then,*

$$\frac{\sqrt{n}(\hat{\Delta}_n - \Delta(Q))}{\sqrt{\hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0)}} \xrightarrow{d} N(0, \varsigma_{t-test}^2) \ , \tag{10}$$

where

$$\varsigma_{t-test}^2 = 1 - \frac{1}{2}\frac{E\left[((E[Y_i(1)|X_i] - E[Y_i(1)]) + (E[Y_i(0)|X_i] - E[Y_i(0)]))^2\right]}{Var[Y_i(1)] + Var[Y_i(0)]} \ .$$

Thus, for the problem of testing (3) at level $\alpha \in (0,1)$, $\phi_n^{t-test}(Z^{(n)})$ defined in (8) satisfies

$$\lim_{n\to\infty} E[\phi_n^{t-test}(Z^{(n)})] = P\{\varsigma_{t-test}\,|G| > z_{1-\frac{\alpha}{2}}\} \leq \alpha \ , \tag{11}$$

where $G \sim N(0,1)$, whenever Q additionally satisfies the null hypothesis, i.e., $\Delta(Q) = \Delta_0$. Furthermore, the inequality in (11) is strict unless

$$E[Y_i(1) + Y_i(0)] = E[Y_i(1) + Y_i(0)|X_i] \tag{12}$$

with probability one under Q.

Remark 3.1. Theorem 3.1 shows that the limiting rejection probability of the two-sample t-test under the null hypothesis is strictly less than the nominal level unless the baseline covariates are irrelevant for potential outcomes in the sense described by (12). We note that the conservativeness of the two-sample t-test is mentioned in Athey and Imbens (2017), but without any formal results. The magnitude of the difference between the limiting rejection probability and the nominal level, however, will depend further on

Q through the value of $\varsigma^2_{t-\text{test}}$. In our simulation study in Section 5, we find that the rejection probability can be severely less than the nominal level and that this difference translates into significant power losses when compared with tests studied below that are (asymptotically) exact in the sense that they have limiting rejection probability under the null hypothesis equal to the nominal level. ■

Remark 3.2. In our definition of the two-sample t-test above, we have used the unpooled estimator of the variance rather than the pooled estimator of the variance. Using Lemma 6.5 in the Appendix, it is straightforward to show that the unpooled estimator of the variance tends in probability to

$$\frac{\text{Var}[Y_i(1)] + \text{Var}[Y_i(0)]}{2} + \frac{(E[Y_i(1)] - E[Y_i(1)])^2}{4} .$$

From this and Lemma 6.4 in the Appendix, it is possible to deduce that with this choice of an estimator of the variance the test may even have limiting rejection probability under the null hypothesis that strictly exceeds the nominal level. ■

3.2 "Matched Pairs" t-Test

Instead of the two-sample t-test studied in the preceding section, it is often recommended to use a "matched pairs" t-test when analyzing such data, which treats the differences of the outcomes within a pair as the observations. This test is also sometimes referred to as the "paired difference-of-means" test. For some examples of its use, see Athey and Imbens (2017), Hsu and Lachenbruch (2007), and Armitage et al. (2008). Formally, this test is given by

$$\phi^{\text{paired}}_n(Z^{(n)}) = I\{|T^{\text{paired}}_n(Z^{(n)})| > z_{1-\frac{\alpha}{2}}\} , \tag{13}$$

where

$$T^{\text{paired}}_n(Z^{(n)}) = \frac{\sqrt{n}(\hat{\Delta}_n - \Delta_0)}{\sqrt{\frac{1}{n}\sum_{1\le j\le n}(Y_{\pi(2j)} - Y_{\pi(2j-1)})^2 - \hat{\Delta}^2_n}} \tag{14}$$

and, as before, $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution. Again, despite its widespread use, the properties of this test are not transparent in our setting.

The following theorem describes the asymptotic behavior of the "matched pairs" t-statistic defined in (14), and, as a consequence, the "matched pairs" t-test defined in (13). The theorem shows, in particular, that the behavior of the "matched pairs" t-test is qualitatively similar to that of the two-sample t-test studied in the preceding section.

Theorem 3.2. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumptions 2.2–2.3. Then,*

$$\frac{\sqrt{n}(\hat{\Delta}_n - \Delta(Q))}{\sqrt{\frac{1}{n}\sum_{1\le j\le n}(Y_{\pi(2j)} - Y_{\pi(2j-1)})^2 - \hat{\Delta}^2_n}} \xrightarrow{d} N(0, \varsigma^2_{\text{paired}}) , \tag{15}$$

where

$$\varsigma^2_{paired} = 1 - \frac{1}{2}\frac{E\left[\left((E[Y_i(1)|X_i] - E[Y_i(1)]) - (E[Y_i(0)|X_i] - E[Y_i(0)])\right)^2\right]}{E[\,Var[Y_i(1)|X_i]] + E[\,Var[Y_i(0)|X_i]] + E\left[\left((E[Y_i(1)|X_i] - E[Y_i(1)]) - (E[Y_i(0)|X_i] - E[Y_i(0)])\right)^2\right]} \ .$$

Thus, for the problem of testing (3) *at level* $\alpha \in (0,1)$, $\phi_n^{paired}(Z^{(n)})$ *defined in* (13) *satisfies*

$$\lim_{n\to\infty} E[\phi_n^{paired}(Z^{(n)})] = P\{\varsigma_{paired}\,|G| > z_{1-\frac{\alpha}{2}}\} \leq \alpha \ , \tag{16}$$

where $G \sim N(0,1)$, *whenever* Q *additionally satisfies the null hypothesis, i.e.,* $\Delta(Q) = \Delta_0$. *Furthermore, the inequality in* (16) *is strict unless*

$$E[Y_i(1) - Y_i(0)] = E[Y_i(1) - Y_i(0)|X_i] \tag{17}$$

with probability one under Q.

Remark 3.3. While Theorem 3.2 is qualitatively similar to Theorem 3.1, it is worth emphasizing the difference between (12) and (17). Both conditions determine a sense in which the baseline covariates are irrrelevant for potential outcomes, but the latter condition holds, in particular, whenever the treatment effect $Y_i(1) - Y_i(0)$ is constant. ■

Remark 3.4. The test statistic in (14) is particularly convenient for the purposes of constructing a confidence interval for $\Delta(Q)$, but we note that it is possible to studentize differently if one is only interested in testing (3). In particular, the result in (16) continues to hold for the test formed by replacing the $\hat{\Delta}_n$ in the denominator on the right-hand side of (14) with Δ_0. ■

Remark 3.5. The literature has also at times advocated estimation of $\Delta(Q)$ via estimation by ordinary least squares of the coefficient on D_i in

$$Y_i = \beta D_i + \sum_{1\leq j\leq n} \lambda_j I\{i \in \{\pi(2j), \pi(2j-1)\}\} + \epsilon_i \ . \tag{18}$$

See, for example, Duflo et al. (2007) and Glennerster and Takavarasha (2013, page 363) as well as Crépon et al. (2015), who estimate $\Delta(Q)$ in the same way, but in a setting with cluster-level treatment assignment. In our setting, it is straightforward to see that the ordinary least squares estimator of β in (18) equals $\hat{\Delta}_n$. It is also possible to show that the usual heteroskedasticity-consistent estimator variance equals

$$\frac{1}{n}\sum_{1\leq j\leq n} (Y_{\pi(2j)} - Y_{\pi(2j-1)})^2 - \hat{\Delta}_n^2 \ .$$

Hence, the resulting test is identical to the "matched pairs" t-test studied in this section. ■

3.3 "Adjusted" t-Test

The proofs of Theorems 3.1 and 3.2 in the Appendix rely upon Lemma 6.4, which establishes that

$$\sqrt{n}(\hat{\Delta}_n - \Delta(Q)) \xrightarrow{d} N(0, \nu^2) \ ,$$

where

$$\nu^2 = \text{Var}[Y_i(1)] + \text{Var}[Y_i(0)] - \frac{1}{2}E\left[((E[Y_i(1)|X_i] - E[Y_i(1)]) + (E[Y_i(0)|X_i] - E[Y_i(0)]))^2\right] \ . \tag{19}$$

Using this observation, it is possible to provide an adjustment to these tests that leads to a test that is exact in the sense that its limiting rejection probability under the null hypothesis equals the nominal level by providing a consistent estimator of (19). As discussed further in Remark 3.7 below, there exist multiple consistent estimators of (19), but a convenient one for our purposes is given by

$$\hat{\nu}_n^2 = \hat{\tau}_n^2 - \frac{1}{2}(\hat{\lambda}_n^2 + \hat{\Delta}_n^2) \ , \tag{20}$$

where

$$\hat{\tau}_n^2 = \frac{1}{n}\sum_{1\le j\le n}(Y_{\pi(2j)} - Y_{\pi(2j-1)})^2 \tag{21}$$

$$\hat{\lambda}_n^2 = \frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2} \rfloor}\Big((Y_{\pi(4j-3)} - Y_{\pi(4j-2)})(Y_{\pi(4j-1)} - Y_{\pi(4j)}) \times (D_{\pi(4j-3)} - D_{\pi(4j-2)})(D_{\pi(4j-1)} - D_{\pi(4j)})\Big) \ . \tag{22}$$

The following theorem shows that the "adjusted" t-test, given by

$$\phi_n^{t-\text{test,adj}}(Z^{(n)}) = I\{|T_n^{t-\text{test,adj}}(Z^{(n)})| > z_{1-\frac{\alpha}{2}}\} \tag{23}$$

with

$$T_n^{t-\text{test,adj}}(Z^{(n)}) = \frac{\sqrt{n}(\hat{\Delta}_n - \Delta_0)}{\hat{\nu}_n} \ , \tag{24}$$

satisfies the desired property.

Theorem 3.3. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumptions 2.2–2.4. Then,*

$$\frac{\sqrt{n}(\hat{\Delta}_n - \Delta(Q))}{\hat{\nu}_n} \xrightarrow{d} N(0,1) \ . \tag{25}$$

Thus, for the problem of testing (3) at level $\alpha \in (0,1)$, $\phi_n^{t-test,adj}(Z^{(n)})$ defined in (23) satisfies

$$\lim_{n\to\infty} E[\phi_n^{t-test,adj}(Z^{(n)})] = \alpha \ , \tag{26}$$

whenever Q additionally satisfies the null hypothesis, i.e., $\Delta(Q) = \Delta_0$.

Remark 3.6. While our discussion has focused on two-sided null hypotheses as described in (3), the con-

vergence in distribution results described in (10), (15) and (25) have straightforward implications for other tests, such related tests of one-sided null hypotheses. ■

Remark 3.7. As mentioned previously, other consistent estimators of (19) exist. For instance, one may consider the estimator given by

$$\tilde{\nu}_n^2 = \hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0) - \frac{1}{2}\left(\tilde{\lambda}_n^2 - (\hat{\mu}_n(1) + \hat{\mu}_n(0))^2\right) , \tag{27}$$

where

$$\tilde{\lambda}_n^2 = \frac{2}{n} \sum_{1 \le j \le \lfloor \frac{n}{2} \rfloor} (Y_{\pi(4j-3)} + Y_{\pi(4j-2)})(Y_{\pi(4j-1)} + Y_{\pi(4j)}) .$$

Using arguments similar to those used in establishing Theorem 3.3, it is possible to show that Theorem 3.3 remains true when $\hat{\nu}_n^2$ defined in (20) is replaced by $\tilde{\nu}_n^2$ defined in (27). ■

3.4 Randomization Tests

In this section, we study the properties of randomization tests based on the idea of permuting treatment status for units within pairs. For ease of exposition, it is convenient to describe the test for the problem of testing (3) with $\Delta_0 = 0$; for the problem of testing (3) more generally, the construction below may be applied with Y_i replaced with $Y_i - D_i\Delta_0$. See Remark 3.10 below for further details.

In order to describe the test formally, it is useful to introduce some further notation. To this end, denote by $\mathbf{G}_n$ the group of all permutations of $2n$ elements and by $\mathbf{G}_n(\pi)$ the subgroup that only permutes elements within the the pairs defined by π, i.e.,

$$\mathbf{G}_n(\pi) = \{g \in \mathbf{G}_n : \{\pi(2j-1), \pi(2j)\} = \{g(\pi(2j-1)), g(\pi(2j)) \text{ for } 1 \le j \le n\}\} .$$

Define the action of $g \in \mathbf{G}_n(\pi)$ on $Z^{(n)}$ as follows:

$$gZ^{(n)} = \{(Y_i, D_{g(i)}, X_i) : 1 \le i \le 2n\} ,$$

i.e., $g \in \mathbf{G}_n(\pi)$ acts on $Z^{(n)}$ by permuting treatment assignment. For a given choice of test statistic $T_n(Z^{(n)})$, the randomization test is given by

$$\phi_n^{\text{rand}}(Z^{(n)}) = I\{T_n(Z^{(n)}) > \hat{R}_n^{-1}(1-\alpha)\} , \tag{28}$$

where

$$\hat{R}_n(t) = \frac{1}{|\mathbf{G}_n(\pi)|} \sum_{g \in \mathbf{G}_n(\pi)} I\{T_n(gZ^{(n)}) \le t\} . \tag{29}$$

Here, $\hat{R}_n^{-1}(1-\alpha)$ is understood to be $\inf\{t \in \mathbf{R} : \hat{R}_n(t) \ge 1-\alpha\}$. We also emphasize that difference choices of $T_n(Z^{(n)})$ lead to different randomization tests and some of our results below will rely upon a particular choice of $T_n(Z^{(n)})$.

Remark 3.8. In some situations, $|\mathbf{G}_n(\pi)| = 2^n$ may be too large to permit computation of $\hat{c}_n^{\text{rand}}(1-\alpha)$ defined in (29). In such cases, a stochastic approximation to the test may be used by replacing $\mathbf{G}_n(\pi)$ with $\hat{\mathbf{G}}_n = \{g_1, \ldots, g_B\}$, where g_1 is the identity permutation and let $g_2, \ldots, g_B$ are i.i.d. Unif($\mathbf{G}_n(\pi)$). Theorem 3.4 below remains true with such an approximation; Theorem 3.5 below also remains true with such an approximation provided that $B \to \infty$ as $n \to \infty$. ■

3.4.1 Finite-Sample Results

Before developing the large-sample properties of the randomization test given by (28), we present some finite-sample properties of the test. We show, in particular, that for any choice of test statistic the randomization test defined in (28) has rejection probability no greater than the nominal level for the following more restrictive null hypothesis:

$$\tilde{H}_0 : Y_i(1)|X_i \stackrel{d}{=} Y_i(0)|X_i \ . \tag{30}$$

While the proof of the result follows closely classical arguments that underlie the finite-sample validity of randomization tests more generally, it is presented in the Appendix for completeness. Similar results can also be found in Heckman et al. (2011) and Lee and Shaikh (2014).

Theorem 3.4. *Suppose the treatment assignment mechanism satisfies Assumption 2.2. For the problem of testing (30) at level $\alpha \in (0,1)$, $\phi_n^{\text{rand}}(Z^{(n)})$ defined in (28) with any $T_n(Z^{(n)})$ satisfies*

$$E[\phi_n^{\text{rand}}(Z^{(n)})] \leq \alpha \tag{31}$$

whenever Q additionally satisfies the null hypothesis, i.e., $Y_i(1)|X_i \stackrel{d}{=} Y_i(0)|X_i$.

Remark 3.9. By modifying the test defined in (28) so that it rejects with positive probability when $T_n(Z^{(n)}) = \hat{c}_n^{\text{rand}}(1-\alpha)$, it is possible to ensure that the test has rejection probability exactly equal to α whenever Q satisfies the null hypothesis, rather than simply less than or equal to α, as described in (31). See Lehmann and Romano (2005, Chapter 15) for further details. ■

3.4.2 Large-Sample Properties

In this section, we establish the large-sample validity of the randomization test given by (28) with a suitable choice of test statistic for testing (3). In particular, we show that the limiting rejection probability of the proposed test equals the nominal level under the null hypothesis.

Theorem 3.5. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumptions 2.2–2.4. Let $T_n(Z^{(n)}) = |T_n^{t-test,adj}(Z^{(n)})|$, where $T_n^{t-test,adj}(Z^{(n)})$ is defined in (24). For such a choice of $T_n(Z^{(n)})$,*

$$\sup_{t \in \mathbf{R}} \left| \hat{R}_n(t) - (\Phi(t) - \Phi(-t)) \right| \stackrel{P}{\to} 0 \ , \tag{32}$$

where $\Phi(\cdot)$ is the standard normal c.d.f. Thus, for the problem of testing (3) with $\Delta_0 = 0$ at level $\alpha \in (0,1)$,

$\phi_n^{\text{rand}}(Z^{(n)})$ *with such a choice of* $T_n(Z^{(n)})$ *satisfies*

$$\lim_{n\to\infty} E[\phi_n^{\text{rand}}(Z^{(n)})] = \alpha \ , \tag{33}$$

whenever Q *additionally satisfies the null hypothesis, i.e.,* $\Delta(Q) = 0$.

Remark 3.10. For completeness, we briefly describe the way in which Theorem 3.5 extends to testing (3) with $\Delta_0 \neq 0$ in further detail. To this end, let $\tilde{Z}_i = (Y_i - D_i\Delta_0, D_i, X_i)$ and define the action of $g \in \mathbf{G}_n(\pi)$ on $\tilde{Z}^{(n)}$ as follows:

$$g\tilde{Z}^{(n)} = \{(Y_i - D_i\Delta_0, D_{g(i)}, X_i) : 1 \leq i \leq 2n\} \ .$$

Consider the test, $\phi_n^{\text{rand}}(\tilde{Z}^{(n)})$, obtained by replacing $Z^{(n)}$ in the test described in Theorem 3.5 with $\tilde{Z}^{(n)}$. For such a test, we have, under the assumptions of Theorem 3.5, that

$$\lim_{n\to\infty} E[\phi_n^{\text{rand}}(\tilde{Z}^{(n)})] = \alpha$$

whenever Q additionally satisfies the null hypothesis, i.e., $\Delta(Q) = \Delta_0$. ■

Remark 3.11. The conclusion in Theorem 3.5 depends heavily on the choice of test statistic in the definition of (28). In order to illustrate this phenomenon, consider the test defined by (28) with $T_n(Z^{(n)}) = |\sqrt{n}\hat{\Delta}_n|$. Using Lemmas 6.4 and 6.8 in the Appendix, it is possible to show that this test behaves similarly under the null hypothesis to the "matched pairs" t-test described in Section 3.2. In particular, it has limiting rejection probability under the null hypothesis no greater than α and strictly less than α unless (17) holds. A growing literature suggests that it should be possible to achieve limiting rejection probability under the null hypothesis equal to α by studentizing the test statistic using a consistent estimator of (19). See, for example, Janssen (1997), Chung and Romano (2013), DiCiccio and Romano (2017) and Bugni et al. (2018). The problem considered here, however, illustrates that this need not be sufficient. To see this, consider the test defined by (28) with $T_n(Z^{(n)}) = \frac{|\sqrt{n}\hat{\Delta}_n|}{\tilde{\nu}_n}$, where $\tilde{\nu}_n^2$ is defined in (27). Even though $\tilde{\nu}_n^2$ is consistent for (19), as discussed in Remark 3.7, it is possible to show using arguments similar to those used in establishing Theorem 3.3 that this test also behaves similarly under the null hypothesis to the "matched pairs" t-test described in Section 3.2. ■

4 Algorithms for Pairing

In this section, we describe different algorithms for pairing units so that Assumptions 2.3–2.4 are satisfied. For the case where $\dim(X_i) = 1$, a particularly simple algorithm leads to pairs that satisfy these assumptions. In particular, we show that in order to satisfy Assumptions 2.3–2.4 it suffices to pair units simply by first ordering the units from smallest to largest according to X_i and then defining pairs according to adjacent units.

Theorem 4.1. *Suppose* $dim(X_i) = 1$ *and* $E[X_i^2] < \infty$. *Let* π *be any permutation of* $2n$ *elements such that that*

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

Then, π satisfies Assumptions 2.3–2.4.

For the case where $\dim(X_i) > 1$, it is helpful to assume that $\text{supp}(X_i)$ lies in a known, bounded set, which, without loss of generality, we may assume to be $[0,1]^k$. Because $u^2 \leq u$ for all $0 \leq u \leq 1$, it follows that for any permutation $\check{\pi}$ of $2n$ elements

$$\frac{1}{n} \sum_{1 \leq j \leq n} |X_{\check{\pi}(2j-1)} - X_{\check{\pi}(2j)}|^2 \leq \frac{1}{n} \sum_{1 \leq j \leq n} |X_{\check{\pi}(2j-1)} - X_{\check{\pi}(2j)}| \,. \tag{34}$$

In order to satisfy Assumption 2.3, it is therefore natural to choose π so as to minimize the right-hand side of (34). Algorithms for solving this minimization problem in a polynomial number of operations exist. See, for example, the "blossom" algorithm described in Edmonds (1965) as well as the algorithm described in Derigs (1988) and implemented in the `R` package `nbpMatching`. The following theorem derives a finite-sample bound on the right-hand side of (34) for π minimizing the right-hand side of (34), which implies, in particular, that pairing units in this way satisfies Assumption 2.3.

Theorem 4.2. *Suppose $\text{supp}(X_i) \subseteq [0,1]^k$. Let π be any permutation of $2n$ elements minimizing the right-hand side of (34). Then, for each integer $m > 1$, we have that*

$$\frac{1}{n} \sum_{1 \leq j \leq n} |X_{\pi(2j)} - X_{\pi(2j-1)}| \leq \frac{\sqrt{k}}{m} + \frac{m^{k-1} 2\sqrt{k}}{n} \,. \tag{35}$$

In particular, if $m \asymp n^{\frac{1}{k}}$, then π satisfies Assumption 2.3.

Given a pairing satisfying Assumption 2.3, we now turn our attention to ensuring that the pairing further satisfies Assumption 2.4. To this end, choose $\bar{\pi}$ so as to minimize

$$\frac{2}{n} \sum_{1 \leq j \leq \lfloor \frac{n}{2} \rfloor} |\bar{X}_{\check{\pi}(2j)} - \bar{X}_{\check{\pi}(2j-1)}| \,, \tag{36}$$

where

$$\bar{X}_j = \frac{X_{\pi(2j)} + X_{\pi(2j-1)}}{2} \,. \tag{37}$$

We note that the aforementioned algorithms may also be used to solve this minimization problem in a polynomial number of operations. The following theorem establishes that by re-ordering the pairs according to $\bar{\pi}$, we can ensure that the pairing satisfies Assumption 2.4 in addition to Assumption 2.3.

Theorem 4.3. *Suppose $\text{supp}(X_i) \subseteq [0,1]^k$. Let π be a permutation of $2n$ elements such that Assumption 2.3 is satisfied and $\bar{\pi}$ be any permutation of n elements minimizing (36). Define a permutation $\tilde{\pi}$ of $2n$ elements so that*

$$\tilde{\pi}(2j) = \pi(2\bar{\pi}(j)) \quad \text{and} \quad \tilde{\pi}(2j-1) = \pi(2\bar{\pi}(j) - 1) \tag{38}$$

for $1 \leq j \leq n$. Then, $\tilde{\pi}$ satisfies Assumptions 2.3–2.4.

5 Simulations

In this section, we examine the finite-sample behavior of several different tests of (3) with $\Delta_0 = 0$ at nominal level $\alpha = .05$ with a simulation study. For $d \in \{0, 1\}$ and $1 \leq i \leq 2n$, potential outcomes are generated according to the equation:

$$Y_i(d) = \mu_d + m_d(X_i) + \sigma_d(X_i)\epsilon_{d,i} \ ,$$

where μ_d, $m_d(X_i)$, $\sigma_d(X_i)$ and $\epsilon_{d,i}$ are specified in each model as follows. In each of following specifications, $n = 100$, $(X_i, \epsilon_{0,i}, \epsilon_{1,i}), i = 1 \ldots 2n$ are i.i.d., $\mu_0 = 0$ and $\mu_1 = \Delta$, where $\Delta = 0$ to study the behavior of the tests under the null hypothesis and $\Delta = \frac{1}{4}$ to study the behavior of the tests under the alternative hypothesis.

Model 1: $X_i \sim \text{Unif}[0, 1]$; $m_1(X_i) = m_0(X_i) = \gamma(X_i - \frac{1}{2})$; $\epsilon_{d,i} \sim N(0, 1)$ for $d = 0, 1$; $\sigma_0(X_i) = \sigma_0 = 1$ and $\sigma_1(X_i) = \sigma_1$.

Model 2: As in Model 1, but $m_1(X_i) = m_0(X_i) = \sin(\gamma(X - \frac{1}{2}))$.

Model 3: As in Model 2, but with $m_1(X_i) = m_0(X_i) + X_i^2 - \frac{1}{3}$.

Model 4: As in Model 1, but $m_0(X_i) = 0$ and $m_1(X_i) = 10(X_i^2 - \frac{1}{3})$.

Model 5: As in Model 4, but $m_0(X_i) = -10(X_i^2 - \frac{1}{3})$.

Model 6: As in Model 4, but $\sigma_0(X_i) = X_i^2$ and $\sigma_1(X_i) = \sigma_1 X_i^2$.

Model 7: $X_i = (\Phi(V_{i1}), \Phi(V_{i2}))'$, where $\Phi(\cdot)$ is the standard normal c.d.f. and

$$V_i \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \ ;$$

$m_1(X_i) = m_0(X_i) = \gamma' X_i - 1$; $\epsilon_{d,i} \sim N(0, 1)$ for $d = 0, 1$; $\sigma_0(X_i) = \sigma_0 = 1$ and $\sigma_1(X_i) = \sigma_1$.

Model 8: As in Model 7, but $m_1(X_i) = m_0(X_i) + 10(\Phi^{-1}(X_{i1})\Phi^{-1}(X_{i2}) - \rho)$.

Model 9: As in Model 7, but $m_0(X_i) = 5(\Phi^{-1}(X_{i1})\Phi^{-1}(X_{i2}) - \rho)$ and $m_1(X_i) = -m_0(X_i)$.

For our subsequent discussion, it is useful to note that Models 5 and 9 satisfy (12), Models 1–2 and 7 satisfy (17), and Models 1-2 and 7 with $\sigma_1 = 1$ satisfy (30) under the null hypothesis.

Treatment status is determined according to Assumption 2.2, where the pairs are calculated as follows. If $\dim(X_i) = 1$, then pairs are calculated by sorting the X_i as described in Theorem 4.1. Note that this ensures that both Assumptions 2.3 and 2.4 are satisfied. If $\dim(X_i) > 1$, then the pairs are calculated by finding π that minimizes the right-hand side of (34) using the `R` package `nbpMatching`. Theorem 4.2 ensures that these pairs satisfy Assumption 2.3. In order to further ensure that the pairs satisfy Assumption 2.4, we re-order the pairs by finding $\bar{\pi}$ that minimizes (36) using the same `R` package and applying Theorem 4.3.

The results of our simulations are presented in Tables 1–3 below. Columns are labeled in the following way:

t-test: The two-sample t-test studied in Theorem 3.1.

naïve: The randomization test defined in (28) with $T_n(Z^{(n)}) = |\sqrt{n}\hat{\Delta}_n|$ and discussed in Remark 3.11. We henceforth refer to this test as the naïve randomization test.

MP-t: The "matched pairs" t-test studied in Theorem 3.2.

t-adj: The "adjusted" t-test studied in Theorem 3.3.

R-adj: The randomization test studied in Theorem 3.5. We henceforth refer to this test as the "adjusted" randomization test.

The tables vary according to the values of γ, σ_1 and ρ, which were not specified in the description of the different models above. Rejection probabilities are calculated using 10^4 replications and presented in percentage points. Because 2^n is large, we employ a stochastic approximation as described in Remark 3.8 with $B = 1000$ when computing each of the randomizaton tests. We organize our discussion of the results by test:

t-test: As expected in light of Theorem 3.1, the two-sample t-test has rejection probability under the null hypothesis no greater than the nominal level. In some cases, the rejection probability under the null hypothesis is far below the nominal level – see, for instance, Models 4 and 6–8. In other cases, the rejection probability is close to the nominal level – see, in particular, Models 5 and 9, which satisfy (12) and are therefore expected to exhibit this phenomenon. In almost all cases, the two-sample t-test is among the least powerful tests, but, as expected, this feature is especially acute when it has rejection probability under the null hypothesis severely below the nominal level.

naïve: As expected following the discussion in Remark 3.11, the naïve randomization test has rejection probability under the null hypothesis no greater than the nominal level. In some cases, the rejection probability under the null hypothesis is far below the nominal level – see, for instance, Models 4–6 and 8–9. In other cases, the rejection probability is close to the nominal level – see, in particular, Models 1–2 and 7, which satisfy (17) and are therefore expected to exhibit this phenomenon. Models 1–2 and 7 with $\sigma_1 = 1$ (corresponding to Tables 1 and 3) in fact satisfy (30) under the null hypothesis, so the rejection probability is exactly equal to the nominal level up to simulation error, in agreement with Theorem 3.4. If its rejection probability is close to the nominal level, then it is also among the most powerful tests, but it otherwise fares poorly in terms of power, especially when compared to the "adjusted" randomization test.

MP-t: As expected in light of Theorem 3.2, the "matched pairs" t-test has rejection probability under the null hypothesis no greater than the nominal level. In some cases, the rejection probability under the null hypothesis is far below the nominal level – see, for instance, Models 4–6 and 8–9. In other cases, the rejection probability is close to the nominal level – see, in particular, Models 1–2 and 7, which satisfy (17) and are therefore expected to exhibit this phenomenon. In almost all cases, the "matched pairs" t-test is among the least powerful tests, but, as expected, this feature is especially acute when it has rejection probability under the null hypothesis severely below the nominal level.

t-adj: As expected in light of Theorem 3.3, the "adjusted" t-test has rejection probability under the null hypothesis close to the nominal level in all cases. In all cases, it is the most powerful test.

R-adj: As expected in light of Theorem 3.5, the "adjusted" randomization test has rejection probability under the null hypothesis close to the nominal level in almost all cases. The exception is Model 8, for which the test exhibits some under-rejection under the null hypothesis. For Models 1–2 and 7 with $\sigma_1 = 1$ (corresponding to Tables 1 and 3), which, as mentioned previously, satisfy (30) under the null hypothesis, the rejection probability is again exactly equal to the nominal level up to simulation error, in agreement with Theorem 3.4. In all cases, it is nearly as powerful as our most powerful test, the "adjusted" t-test.

We conclude with some recommendations for empirical practice based on our theoretical results as well as the simulation study above. We do not recommend the two-sample t-test, the "matched pairs" t-test or the naïve randomization test, which are often considerably less powerful than both the "adjusted" t-test and the "adjusted" randomization test. In our simulations the "adjusted" t-test is always the most powerful among the tests we consider, though sometimes by a small margin in comparison to the "adjusted" randomization test. We also note that the modest gain in power of the "adjusted" t-test is accompanied by the generally higher rejection probability under the null hypothesis of the "adjusted" t-test as well. As mentioned previously, the "adjusted" randomization test retains the attractive feature that the finite-sample rejection probability under the null hypothesis is no greater than the nominal size for certain distributions satisfying the null hypothesis. To the extent that this feature is deemed important, the "adjusted" randomization test may be preferred to the "adjusted" t-test despite having slightly lower power.

	Under H_0 — $\Delta = 0$					Under H_1 — $\Delta = 1/4$				
Model	t-**test**	**näive**	**MP-**t	t-**adj**	**R-adj**	t-**test**	**näive**	**MP-**t	t-**adj**	**R-adj**
1	4.25	5.02	5.31	5.29	4.97	40.16	41.87	43.20	43.17	41.44
2	4.32	4.93	5.43	5.42	4.93	39.23	41.37	42.52	42.29	40.78
3	3.51	4.73	5.04	5.15	4.73	35.90	40.09	41.56	42.05	40.67
4	1.28	1.13	1.29	4.89	4.27	5.43	5.12	5.51	15.97	14.45
5	5.69	0.79	0.90	5.68	4.98	9.65	1.94	2.18	9.61	8.60
6	0.87	0.65	0.75	5.33	4.83	4.80	4.03	4.70	19.41	17.36
7	3.29	4.94	5.30	5.44	5.28	35.82	41.56	43.07	43.17	42.16
8	1.00	0.93	1.03	4.56	4.26	0.94	0.93	0.96	4.75	4.37
9	5.30	0.65	0.71	4.28	3.87	7.18	1.52	1.65	6.17	5.83

Table 1: Rej. prob. for Models 1–9 with $\gamma = 1$ for Models 1–6, $\gamma' = (1, 1)$ for Models 7–9, $\sigma_1 = 1$, $\rho = 0.2$.

	Under H_0 — $\Delta = 0$					Under H_1 — $\Delta = 1/4$				
Model	t-**test**	**näive**	**MP-**t	t-**adj**	**R-adj**	t-**test**	**näive**	**MP-**t	t-**adj**	**R-adj**
1	4.75	5.11	5.37	5.46	5.06	29.46	30.26	31.51	31.49	30.24
2	4.23	4.59	5.03	5.20	4.70	29.33	29.99	31.39	30.89	29.52
3	4.16	4.84	5.27	5.39	5.09	26.60	28.78	30.07	30.30	29.27
4	1.65	1.53	1.65	5.24	4.74	5.80	5.31	5.91	14.95	13.72
5	5.27	0.68	0.81	5.21	4.67	9.59	2.19	2.53	9.54	8.45
6	0.83	0.81	0.91	5.50	4.86	4.89	4.23	4.66	18.25	16.43
7	0.39	5.21	5.66	5.85	5.54	7.38	30.04	31.01	31.20	30.56
8	1.50	1.58	1.66	5.71	5.27	0.69	0.70	0.77	4.80	4.36
9	5.73	1.34	1.42	5.24	4.87	8.28	2.13	2.22	7.33	6.93

Table 2: Rej. prob. for Models 1–9 with $\gamma = 1$ for Models 1–6, $\gamma' = (1, 4)$ for Models 7–9, $\sigma_1 = 2$, $\rho = 0.7$.

	Under H_0 — $\Delta = 0$					Under H_1 — $\Delta = 1/4$				
Model	t-**test**	**näive**	**MP-**t	t-**adj**	**R-adj**	t-**test**	**näive**	**MP-**t	t-**adj**	**R-adj**
1	4.51	5.19	5.62	5.66	5.24	39.09	40.88	42.09	41.92	40.56
2	4.09	4.68	5.03	5.08	4.58	39.95	41.59	42.84	42.43	41.20
3	3.67	4.91	5.26	5.55	5.26	35.10	39.48	40.89	41.48	40.15
4	1.07	0.98	1.13	4.83	4.28	5.43	5.00	5.47	16.52	14.95
5	5.21	0.69	0.79	5.21	4.61	9.98	2.17	2.35	9.93	8.89
6	0.67	0.65	0.69	5.17	4.44	5.11	4.50	4.89	19.03	17.23
7	0.28	4.91	5.19	5.50	5.23	11.20	41.61	43.01	43.18	42.06
8	0.70	0.67	0.81	4.41	4.03	0.95	0.96	1.11	5.26	4.75
9	5.37	0.71	0.79	4.30	4.00	6.93	0.95	1.02	5.52	5.10

Table 3: Rej. prob. for Models 1–9 with $\gamma = 1$ for Models 1–6, $\gamma' = (4, 1)$ for Models 7–9, $\sigma_1 = 1$, $\rho = 0$.

6 Appendix

Please note that in what follows we will use the notation $a \lesssim b$ to denote $a \leq cb$ for some constant c.

6.1 Proof of Theorem 3.1

The theorem follows immediately upon noting that (10) follows from Lemmas 6.4–6.5 below. ■

6.2 Proof of Theorem 3.2

The theorem follows immediately upon noting that (15) follows from Lemmas 6.4–6.5 and 6.6 below. ■

6.3 Proof of Theorem 3.3

From Lemma 6.4, we see that it suffices to show that $\hat{\nu}_n^2$ defined in (20) tends in probability to (54). Since

$$\begin{aligned} & E[\mathrm{Var}[Y_i(1)|X_i]] + E[\mathrm{Var}[Y_i(0)|X_i]] + \frac{1}{2}E\left[((E[Y_i(1)|X_i] - E[Y_i(1)]) - (E[Y_i(0)|X_i] - E[Y_i(0)]))^2\right] \\ = \quad & E[\mathrm{Var}[Y_i(1)|X_i]] + E[\mathrm{Var}[Y_i(0)|X_i]] + \frac{1}{2}\left(E\left[(E[Y_i(1)|X_i] - E[Y_i(0)|X_i])^2\right] - (E[Y_i(1)] - E[Y_i(0)])^2\right) . \end{aligned}$$

the desired conclusion follows immediately from Lemmas 6.5–6.7. ■

6.4 Proof of Theorem 3.4

Let Q satisfying (30) be given. For such a Q, we first argue that

$$gZ^{(n)}|X^{(n)} \stackrel{d}{=} Z^{(n)}|X^{(n)} . \tag{39}$$

Since $\pi = \pi_n(X^{(n)})$, we have from Assumption 2.2 that

$$gD^{(n)}|X^{(n)} \stackrel{d}{=} D^{(n)}|X^{(n)} . \tag{40}$$

Furthermore,

$$Y^{(n)} \perp\!\!\!\perp D^{(n)}|X^{(n)} . \tag{41}$$

To see this, note for any set A and any d and d' in the support of $D^{(n)}|X^{(n)}$ that

$$\begin{aligned} P\{Y^{(n)} \in A|D^{(n)} = (d_1, \ldots d_{2n}), X^{(n)}\} &= P\{(Y_1(d_1), \ldots Y_{2n}(d_{2n})) \in A|D^{(n)} = (d_1, \ldots d_{2n}), X^{(n)}\} \\ &= P\{(Y_1(d_1), \ldots Y_{2n}(d_{2n})) \in A|X^{(n)}\} \\ &= P\{(Y_1(d'_1), \ldots Y_{2n}(d'_{2n})) \in A|X^{(n)}\} \\ &= P\{(Y_1(d'_1), \ldots Y_{2n}(d'_{2n})) \in A|D^{(n)} = (d'_1, \ldots d'_{2n}), X^{(n)}\} \\ &= P\{Y^{(n)} \in A|D^{(n)} = (d'_1, \ldots d'_{2n}), X^{(n)}\} , \end{aligned}$$

where the first and fifth equalities follow from (1), the second and fourth equalities follow from (4), the third follows from the fact that Q satisfies (30). It now follows from (40) and (41) that (39) holds.

Next, observe that

$$E\left[\sum_{g \in \mathbf{G}_n(\pi)} \phi_n^{\mathrm{rand}}(gZ^{(n)})\right] = E\left[E\left[\sum_{g \in \mathbf{G}_n(\pi)} \phi_n^{\mathrm{rand}}(gZ^{(n)})\middle| X^{(n)}\right]\right]$$

$$
\begin{aligned}
&= E\left[\sum_{g\in\mathbf{G}_n(\pi)} E\left[\phi_n^{\mathrm{rand}}(Z^{(n)})\middle|X^{(n)}\right]\right] \\
&= E\left[2^n E\left[\phi_n^{\mathrm{rand}}(Z^{(n)})\middle|X^{(n)}\right]\right] \\
&= 2^n E\left[\phi_n^{\mathrm{rand}}(Z^{(n)})\right] , \qquad (42)
\end{aligned}
$$

where the first and final equalities follow from the law of iterated expectations, the second follows from (39), and the third exploits the fact that $|\mathbf{G}_n(\pi)| = 2^n$. Using the fact that $\mathbf{G}_n(\pi)$ is a group, we have with probability one that

$$
\sum_{g\in\mathbf{G}_n(\pi)} \phi_n^{\mathrm{rand}}(gZ^{(n)}) \le 2^n\alpha .
$$

Hence,

$$
E\left[\sum_{g\in\mathbf{G}_n(\pi)} \phi_n^{\mathrm{rand}}(gZ^{(n)})\right] \le 2^n\alpha . \qquad (43)
$$

Combining (42) and (43), we see that (31) holds, as desired. ■

6.5 Proof of Theorem 3.5

Note that

$$
\hat{\Delta}_n = \frac{1}{n}\sum_{1\le j\le n}(Y_{\pi(2j)} - Y_{\pi(2j-1)})(D_{\pi(2j)} - D_{\pi(2j-1)}) .
$$

This observation, together with the definition of $\hat{\nu}_n$ in (20), implies that

$$
\hat{R}_n(t) = P\left\{\frac{\left|\frac{1}{\sqrt{n}}\sum_{1\le j\le n}\epsilon_j(Y_{\pi(2j)} - Y_{\pi(2j-1)})(D_{\pi(2j)} - D_{\pi(2j-1)})\right|}{\check{\nu}_n(\epsilon_1,\ldots,\epsilon_n)} \le t\middle| W^{(n)}\right\} ,
$$

where, independently of $W^{(n)}$, $\epsilon_j, j = 1,\ldots,n$ are i.i.d. Rademacher random variables and $\check{\nu}_n^2$ is defined as in (79). Note further that

$$
\hat{R}_n(t) = \check{R}_n(t) - \check{R}_n(-t) ,
$$

where

$$
\check{R}_n(t) = P\left\{\frac{\frac{1}{\sqrt{n}}\sum_{1\le j\le n}\epsilon_j(Y_{\pi(2j)} - Y_{\pi(2j-1)})(D_{\pi(2j)} - D_{\pi(2j-1)})}{\check{\nu}_n(\epsilon_1,\ldots,\epsilon_n)} \le t\middle| W^{(n)}\right\} .
$$

The desired conclusion now follows immediately from Lemmas 6.8–6.9 together with Theorem 5.2 of Chung and Romano (2013). ■

6.6 Proof of Theorem 4.1

For $1 \le i \le 2n$, let $U_i = |X_i|$ and write $U_{(1)} \le \cdots \le U_{(2n)}$. Note that

$$
\begin{aligned}
\frac{1}{n}\sum_{1\le j\le n}|X_{\pi(2j)} - X_{\pi(2j-1)}| &= \frac{1}{n}\sum_{1\le j\le n}(X_{\pi(2j)} - X_{\pi(2j-1)}) \\
&\le \frac{1}{n}(X_{\pi(2n)} - X_{\pi(1)}) \\
&\le \frac{1}{n}2U_{(2n)} \\
&\xrightarrow{P} 0 ,
\end{aligned}
$$

where the equality exploits the fact that $X_{\pi(2j-1)} \le X_{\pi(2j)}$, the two inequalities follow by inspection, and the convergence in probability to zero follows from Lemma 6.1. Similarly,

$$
\frac{1}{n}\sum_{1\le j\le n}|X_{\pi(2j)} - X_{\pi(2j-1)}|^2 \le |X_{\pi(2n)} - X_{\pi(1)}|\left(\frac{1}{n}\sum_{1\le j\le n}|X_{\pi(2j)} - X_{\pi(2j-1)}|\right)
$$

$$\begin{aligned} &\leq \left(\frac{U_{(2n)}}{\sqrt{n}}\right)^2 \\ &\xrightarrow{P} 0 , \end{aligned}$$

where the first inequality follows by inspection, the second follows by arguing as before, and the convergence in probability to zero again follows from Lemma 6.1. Finally, for any $k \in \{2,3\}$ and $\ell \in \{0,1\}$, we have that

$$\begin{aligned} \frac{2}{n} \sum_{1 \leq j \leq \frac{n}{2}} |X_{\pi(4j-k)} - X_{\pi(4j-\ell)}|^2 &\leq \frac{2}{n} \sum_{1 \leq j \leq \frac{n}{2}} |X_{\pi(4j-3)} - X_{\pi(4j)}|^2 \\ &\leq |X_{\pi(2n)} - X_{\pi(1)}| \left(\frac{2}{n} \sum_{1 \leq j \leq \frac{n}{2}} |X_{\pi(4j-3)} - X_{\pi(4j)}| \right) \\ &\leq \left(\frac{U_{(2n)}}{\sqrt{n}}\right)^2 \\ &\xrightarrow{P} 0 , \end{aligned}$$

where the first and second inequalities follow by inspection, the third follows by arguing as before, and the convergence in probability to zero again follows from Lemma 6.1. It thus follows that Assumptions 2.3–2.4 hold.

6.7 Proof of Theorem 4.2

We describe an algorithm that leads to a pairing that does not minimize the right-hand side of (34) exactly, but which leads to the desired bound, from which the result follows.

In order to describe the algorithm, it is useful to introduce some further notation. For an integer $m > 1$, divide $[0,1]^k$ into m^k hypercubes with sides of length m^{-1}. We index these cubes by k-tuples of the form $(i_1, \ldots, i_k)$ with $1 \leq i_j \leq m$ for all $1 \leq j \leq k$. Specifically, the k-tuple $(i_1, \ldots, i_k)$ corresponds to the (closed) cube with vertices

$$\left\{ \frac{1}{m}(i_1 - 1 + \delta_1, \ldots, i_k - 1 + \delta_k) : \delta_j \in \{0,1\} \text{ for all } 1 \leq j \leq k \right\} .$$

We further order these cubes in a "contiguous" way. We do so by defining an algorithm f_k that takes as an input a k-dimensional hypercube of the form $(i_1, \ldots, i_k)$ with $i_j \in \{1, m\}$ for all $1 \leq j \leq k$ and returns a "path" starting from $(i_1, \ldots, i_k)$ and ending at $(i'_1, \ldots, i'_k)$ with $i'_j \in \{1, m\}$ for all $1 \leq j \leq k$ that traverses all m^k of the possible k-dimensional hypercubes. We define f_1 so that

$$f_1((i_1)) = \begin{cases} (1) \mapsto (2) \mapsto \cdots \mapsto (m-1) \mapsto (m) & \text{if} \quad (i_1) = (1) \\ (m) \mapsto (m-1) \mapsto \cdots \mapsto (2) \mapsto (1) & \text{if} \quad (i_1) = (m) . \end{cases} \tag{44}$$

Given f_{k-1}, we define $f_k((i_1^0, \ldots, i_k^0))$ as follows. If $i_k^0 = 1$, then $f_k((i_1^0, \ldots, i_k^0))$ equals

$$\begin{aligned} & (i_1^0, \ldots, i_{k-1}^0, 1) \mapsto \cdots \mapsto (i_1^1, \ldots, i_{k-1}^1, 1) \\ \mapsto \quad & (i_1^1, \ldots, i_{k-1}^1, 2) \mapsto \cdots \mapsto (i_1^2, \ldots, i_{k-1}^2, 2) \\ & \vdots \\ \mapsto \quad & (i_1^{j-1}, \ldots, i_{k-1}^{j-1}, j) \mapsto \cdots \mapsto (i_1^j, \ldots, i_{k-1}^j, j) \\ & \vdots \\ \mapsto \quad & (i_1^{m-1}, \ldots, i_{k-1}^{m-1}, m) \mapsto \cdots \mapsto (i_1^m, \ldots, i_{k-1}^m, m) , \end{aligned}$$

where in the preceding display it is understood that the "path" for a fixed "row," i.e.,

$$(i_1^{j-1}, \ldots, i_{k-1}^{j-1}, j) \mapsto \cdots \mapsto (i_1^j, \ldots, i_{k-1}^j, j) , \tag{45}$$

is given by applying f_{k-1} first to obtain a "path" starting from $(i_1^{j-1}, \ldots, i_{k-1}^{j-1})$ and ending at $(i_1^j, \ldots, i_{k-1}^j)$ and then "ap-

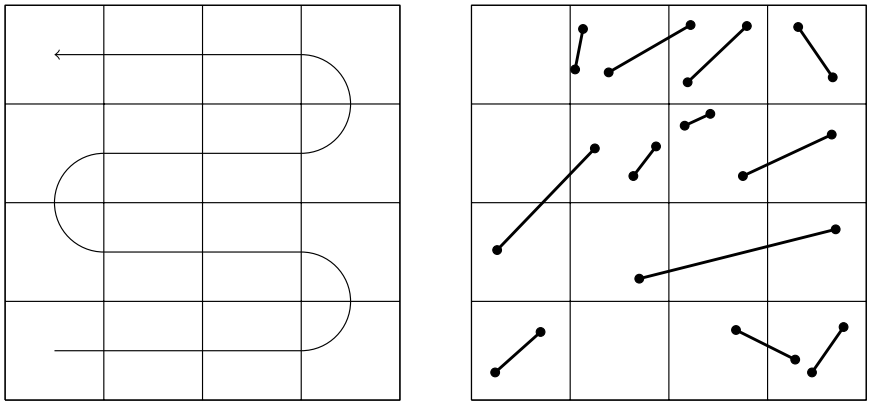
pending" j to obtain a "path" of the form (45). If, on the other hand, $i_k^0 = m$, then $f_k((i_1^0, \ldots, i_k^0))$ equals

$$\begin{aligned}
& (i_1^0, \ldots, i_{k-1}^0, m) \mapsto \cdots \mapsto (i_1^1, \ldots, i_{k-1}^1, m) \\
\mapsto \quad & (i_1^1, \ldots, i_{k-1}^1, m-1) \mapsto \cdots \mapsto (i_1^2, \ldots, i_{k-1}^2, m-1) \\
& \qquad \vdots \\
\mapsto \quad & (i_1^{j-1}, \ldots, i_{k-1}^{j-1}, m-j+1) \mapsto \cdots \mapsto (i_1^j, \ldots, i_{k-1}^j, m-j+1) \\
& \qquad \vdots \\
\mapsto \quad & (i_1^{m-1}, \ldots, i_{k-1}^{m-1}, 1) \mapsto \cdots \mapsto (i_1^m, \ldots, i_{k-1}^m, 1) \ ,
\end{aligned}$$

where, as before, in the preceding display it is understood that the "path" for a fixed "row," i.e.,

$$(i_1^{j-1}, \ldots, i_{k-1}^{j-1}, m-j+1) \mapsto \cdots \mapsto (i_1^j, \ldots, i_{k-1}^j, m-j+1) \ , \tag{46}$$

is given by applying f_{k-1} first to obtain a "path" starting from $(i_1^{j-1}, \ldots, i_{k-1}^{j-1})$ and ending at $(i_1^j, \ldots, i_{k-1}^j)$ and then "appending" $m-j+1$ to obtain a "path" of the form (45).



(a) (b)

Figure 1: (a) Illustration of the "path" obtained by applying f_k with $k = 2$ and $m = 4$; (b) Illustration of a pairing obtained by applying Algorithm 6.1 with $k = 2$, $n = 12$ and $m = 4$. Note that the endpoints of the line segments correspond to units and the pairs correspond to units connected by a line segments.

With f_k so defined, we may obtain a "path" starting with $(1, \ldots, 1)$. Figure 1(a) above illustrates the "path" obtained in this way for the case of $k = 2$ and $m = 4$. Using this "path," we are now prepared to describe our algorithm for pairing units below. We emphasize that the algorithm depends on the choice of m. For clarity, we also note that when we say in our description of the algorithm that a unit i belongs to a hypercube, we mean that X_i belongs to the hypercube. To avoid any ambiguity, whenever a unit belongs to more than one hypercube, we assign it the hypercube that appears earliest along the "path."

Algorithm 6.1.

> Begin with the first nonempty hypercube along the "path." If there are an even number of units in that hypercube, pair them together in any fashion; if there are an odd number of units in that hypercube, pair as many as possible together. Now proceed to the "next" nonempty hypercube along the "path." If in the previous hypercube there was an unpaired unit, pair one of the units in the present hypercube with the remaining unit from the previous hypercube. If, after doing so, there are an even number of unpaired units in the hypercube,

pair them in any fashion; if, after doing so, there are an odd number of unpaired units in the hypercube, pair as many as possible together. Proceed to the next nonempty hypercube along the "path." Continue in this fashion until there are no more nonempty hypercubes.

Figure 1(b) above illustrates a pairing obtained by applying Algorithm 6.1 with $k = 2$, $n = 12$ and $m = 4$.

We now argue that Algorithm 6.1 leads to a pairing satisfying the desired bound. To this end, first note that the maximum distance between any two points in the a k-dimensional hypercube with sides of length $\frac{1}{m}$ is $\frac{\sqrt{k}}{m}$. Note further that the maximum distance between two points in two such cubes that are contiguous (as understood according to ordering described in Section 4) is $\frac{2\sqrt{k}}{m}$. Using these facts, the bound in (35) now easily follows. Indeed, simply note that the sum that appears on the left-hand side of (35) may contain at most n terms corresponding to pairs of points within hypercubes and at most m^k terms corresponding to pairs of points in contiguous hypercubes. The desired conclusion now follows immediately. ■

6.8 Proof of Theorem 4.3

We prove the result for $k = 3$ and $\ell = 0$; the other values of k and ℓ can be handled similarly.

By arguing as in the proof of Theorem 4.2 and using (34), we see that

$$\frac{2}{n} \sum_{1 \le j \le \lfloor \frac{n}{2} \rfloor} |\bar{X}_{\tilde{\pi}(2j)} - \bar{X}_{\tilde{\pi}(2j-1)}|^2 \xrightarrow{P} 0 \ . \tag{47}$$

Note that

$$\begin{aligned}
& \frac{1}{n} \sum_{1 \le j \le \frac{n}{2}} |X_{\tilde{\pi}(4j-3)} - X_{\tilde{\pi}(4j)}|^2 \\
= \ & \frac{1}{n} \sum_{1 \le j \le \frac{n}{2}} |X_{\tilde{\pi}(4j-3)} - \bar{X}_{\tilde{\pi}(2j-1)} + \bar{X}_{\tilde{\pi}(2j-1)} - \bar{X}_{\tilde{\pi}(2j)} + \bar{X}_{\tilde{\pi}(2j)} - X_{\tilde{\pi}(4j)}|^2 \\
\lesssim \ & \frac{1}{n} \sum_{1 \le j \le \frac{n}{2}} |X_{\tilde{\pi}(4j-3)} - \bar{X}_{\tilde{\pi}(2j-1)}|^2 + |\bar{X}_{\tilde{\pi}(2j-1)} - \bar{X}_{\tilde{\pi}(2j)}|^2 + |\bar{X}_{\tilde{\pi}(2j)} - X_{\tilde{\pi}(4j)}|^2 \\
\lesssim \ & \frac{1}{n} \sum_{1 \le j \le \frac{n}{2}} |X_{\tilde{\pi}(4j-3)} - X_{\tilde{\pi}(4j-2)}|^2 + |\bar{X}_{\tilde{\pi}(2j-1)} - \bar{X}_{\tilde{\pi}(2j)}|^2 + |X_{\tilde{\pi}(4j-1)} - X_{\tilde{\pi}(4j)}|^2 \\
\lesssim \ & \frac{1}{n} \sum_{1 \le j \le n} |X_{\pi(2j)} - X_{\pi(2j-1)}|^2 + \frac{1}{n} \sum_{1 \le j \le \frac{n}{2}} |\bar{X}_{\tilde{\pi}(2j-1)} - \bar{X}_{\tilde{\pi}(2j)}|^2 \\
\xrightarrow{P} \ & 0 \ ,
\end{aligned}$$

where the first equality follows by inspection, the first inequality follows using the fact that $|a + b|^2 \le 2(|a|^2 + |b|^2)$ for any real vectors a and b, the second inequality follows from (37) and (38), the second equality follows again from (38), and the convergence to zero in probability follows from the assumption that π satisfies Assumption 2.3 and (47). ■

6.9 Auxiliary Results

Lemma 6.1. *Let $U_i, i = 1, \ldots, n$ an i.i.d. sequence of random vectors such that $E[|U_i|^r] < \infty$. Then,*

$$n^{-\frac{1}{r}} \max_{1 \le i \le n} |U_i| \xrightarrow{P} 0$$

as $n \to \infty$.

Proof: Let $\epsilon > 0$ be given. Note that

$$P\left\{ n^{-\frac{1}{r}} \max_{1 \le i \le n} |U_i| > \epsilon \right\} = P\left\{ \bigcup_{1 \le i \le n} \{|U_i|^r > \epsilon^r n\} \right\}$$

$$
\begin{aligned}
&\leq \sum_{1\leq i\leq n} P\{|U_i|^r > \epsilon^r n\} \\
&\leq \frac{1}{n\epsilon^r} \sum_{1\leq i\leq n} E[|U_i|^r I\{|U_i|^r > \epsilon^r n\}] \\
&= \frac{1}{\epsilon^r} E[|U_i|^r I\{|U_i|^r > \epsilon^r n\}] \\
&\to 0
\end{aligned}
$$

as $n \to \infty$, where the first equality follows by inspection, the first inequality follows from Bonferonni's inequality, the second inequality follows from Markov's inequality, the final equality follows from the i.i.d. assumption, and the convergence to zero follows from the assumption that $E[|U_i|^r] < \infty$. ■

Lemma 6.2. *For $n \geq 1$, let U_n and V_n be real-valued random variables and $\mathcal{F}_n$ a σ-field. Suppose*

$$P\{U_n \leq u|\mathcal{F}_n\} \to \Phi(u/\tau_1) \text{ a.s. }, \tag{48}$$

where $\Phi(\cdot)$ is the standard normal c.d.f. Further assume V_n is $\mathcal{F}_n$-measurable and

$$V_n \xrightarrow{d} N(0, \tau_2^2) \ .$$

Then,

$$U_n + V_n \xrightarrow{d} N(0, \tau_1^2 + \tau_2^2) \ .$$

PROOF: Note that the convergence (48) holds with probability one for all u in a countable dense set, and hence the conditional distributions converge weakly to $N(0, \tau_1^2)$ with probability one. Use characteristic functions and calculate

$$E \exp[it(U_n + V_n)] = E\{\exp(itV_n) E[\exp(itU_n)|\mathcal{F}_n]\} \ .$$

But, $E[\exp it(V_n)] \to \exp(-\frac{t^2}{2}\tau_2^2)$. Also, on the set where we have weak convergence, we have convergence of characteristic functions, so that

$$E[\exp(itU_n)|\mathcal{F}_n] \to \exp\left(-\frac{t^2}{2}\tau_1^2\right) \text{ a.s.}$$

The result follows from dominated convergence. ■

Lemma 6.3. *Let $(U_{n,1}, \ldots, U_{n,n}) \sim G_n^* = \bigotimes_{1\leq i\leq n} G_{n,i}$ with $\mu(G_{n,i}) = 0$ for all $1 \leq i \leq n$. Define*

$$\bar{G}_n = \frac{1}{n} \sum_{1\leq i\leq n} G_{n,i} \ .$$

If

$$\lim_{\lambda\to\infty} \limsup_{n\to\infty} E_{\bar{G}_n}[|U|I\{|U| > \lambda\}] = 0 \ , \tag{49}$$

then $\bar{U}_n \xrightarrow{G_n^*} 0$.

PROOF: Define

$$Z_{n,i} = U_{n,i} I\{|U_{n,i}| \leq n\} \ .$$

Let $m_{n,i} = E[Z_{n,i}]$ and $\bar{m}_n = E[\bar{Z}_n]$. For any $\epsilon > 0$, we have that

$$P\{|\bar{U}_n - \bar{m}_n| > \epsilon\} \leq P\{|\bar{Z}_n - \bar{m}_n| > \epsilon\} + P\{\bar{U}_n \neq \bar{Z}_n\} \ .$$

Furthermore,

$$P\{\bar{U}_n \neq \bar{Z}_n\} \leq P\left\{\bigcup_{1\leq i\leq n} \{U_{n,i} \neq Z_{n,i}\}\right\} \leq \sum_{1\leq i\leq n} P\{U_{n,i} \neq Z_{n,i}\} = \sum_{1\leq i\leq n} P\{|U_{n,i}| > n\} \ .$$

By Chebychev's inequality, we have that

$$P\{|\bar{Z}_n - \bar{m}_n| > \epsilon\} \leq \frac{\mathrm{Var}[\bar{Z}_n]}{\epsilon^2} = \frac{1}{n} \sum_{1\leq i\leq n} \frac{\mathrm{Var}[Z_{n,i}]}{n\epsilon^2} \leq \frac{1}{n} \sum_{1\leq i\leq n} \frac{E[Z_{n,i}^2]}{n\epsilon^2} \ .$$

Hence,

$$P\{|\bar{U}_n - \bar{m}_n| > \epsilon\} \leq \frac{1}{n} \sum_{1 \leq i \leq n} \frac{E[Z_{n,i}^2]}{n\epsilon^2} + \frac{1}{n} \sum_{1 \leq i \leq n} nP\{|U_{n,i}| > n\} \ .$$

For $t > 0$, let

$$\begin{aligned} \tau_{n,i}(t) &= tP\{|U_{n,i}| > t\} = t(1 - G_{n,i}(t) + G_{n,i}(-t)) \\ \kappa_{n,i}(t) &= \frac{1}{t} E[Z_{n,i}^2] = \frac{1}{t} \int_{-t}^{t} x^2 dG_{n,i}(t) \ . \end{aligned}$$

In this notation, we have that

$$P\{|\bar{U}_n - \bar{m}_n| > \epsilon\} \leq \frac{1}{n} \sum_{1 \leq i \leq n} \frac{\kappa_{n,i}(n)}{\epsilon^2} + \frac{1}{n} \sum_{1 \leq i \leq n} \tau_{n,i}(n) \ . \tag{50}$$

Since

$$tP\{|U_{n,i}| > t\} \leq E[|U_{n,i}| I\{|U_{n,i}| > t\}] \ , \tag{51}$$

we see that

$$\frac{1}{n} \sum_{1 \leq i \leq n} \tau_{n,i}(t) \leq \frac{1}{n} \sum_{1 \leq i \leq n} E[|U_{n,i}| I\{|U_{n,i}| > t\}] = E_{\bar{G}_n}[|U| I\{|U| > t\}] \ .$$

Hence,

$$\frac{1}{n} \sum_{1 \leq i \leq n} \tau_{n,i}(n) \to 0 \ .$$

Using integration by parts, it is possible to show that

$$\kappa_{n,i}(t) = -\tau_{n,i}(t) + \frac{2}{t} \int_0^t \tau_{n,i}(x) dx \ .$$

In order to show that the left-hand side of (50) tends to zero, it therefore suffices to argue that

$$\frac{1}{n} \sum_{1 \leq i \leq n} \frac{1}{n} \int_0^n \tau_{n,i}(x) dx \to 0 \ . \tag{52}$$

To this end, note that (51) implies that

$$\begin{aligned} \frac{1}{n} \sum_{1 \leq i \leq n} \frac{1}{n} \int_0^n \tau_{n,i}(x) dx &\leq \frac{1}{n} \sum_{1 \leq i \leq n} \frac{1}{n} \int_0^n E[|U_{n,i}| I\{|U_{n,i}| > x\}] dx \\ &= \frac{1}{n} \int_0^n E_{\bar{G}_n}[|U| I\{|U| > x\}] dx \ . \end{aligned}$$

Let $\delta > 0$ be given and choose n_0 and λ_0 so that

$$E_{\bar{G}_n}[|U| I\{|U| > x\}] < \frac{\delta}{2}$$

whenever $n > n_0$ and $x > \lambda_0$. For $x \leq \lambda_0$ and $n > n_0$, we have that

$$E_{\bar{G}_n}[|U| I\{|U| > x\}] \leq E_{\bar{G}_n}[|U|] = E_{\bar{G}_n}[|U| I\{|U| \leq \lambda_0\}] + E_{\bar{G}_n}[|U| I\{|U| > \lambda_0\}] \leq \lambda_0 + \frac{\delta}{2}$$

It follows that

$$\frac{1}{n} \int_0^n E_{\bar{G}_n}[|U| I\{|U| > x\}] dx \leq \frac{\lambda_0(\lambda_0 + \frac{\delta}{2})}{n} + \frac{\delta}{2}$$

for $n > n_0$ and $n > \lambda_0$, which is less than δ for all n sufficiently large. Since the choice of $\delta > 0$ was arbitrary, (52) follows. To complete the proof, note that

$$|\bar{m}_n| \leq \frac{1}{n} \sum_{1 \leq i \leq n} E[|U_{n,i}| I\{U_{n,i}| > n\}] = E_{\bar{G}_n}[|U| I\{|U| > n\}] \ ,$$

which tends to zero by assumption. ■

Lemma 6.4. *If Assumptions 2.1–2.3 hold, then*

$$\sqrt{n}(\hat{\Delta}_n - \Delta(Q)) \xrightarrow{d} N(0, \nu^2) \ , \tag{53}$$

where

$$\begin{aligned}\nu^2 &= E[Var[Y_i(1)|X_i]] + E[Var[Y_i(0)|X_i]] \\ &\quad + \frac{1}{2}E\left[((E[Y_i(1)|X_i] - E[Y_i(1)]) - (E[Y_i(0)|X_i] - E[Y_i(0)]))^2\right] \\ &= Var[Y_i(1)] + Var[Y_i(0)] - \frac{1}{2}E\left[((E[Y_i(1)|X_i] - E[Y_i(1)]) + (E[Y_i(0)|X_i] - E[Y_i(0)]))^2\right]\end{aligned} \tag{54}$$

as $n \to \infty$.

PROOF: Note that

$$\begin{aligned}\frac{1}{n}\sum_{1\le i\le 2n: D_i=1} Y_i &= \frac{1}{n}\sum_{1\le i\le 2n} Y_i(1)D_i \\ \frac{1}{n}\sum_{1\le i\le 2n: D_i=0} Y_i &= \frac{1}{n}\sum_{1\le i\le 2n} Y_i(0)(1-D_i)\ .\end{aligned}$$

Hence, we may write

$$\sqrt{n}(\hat{\Delta}_n - \Delta(Q)) = A_n - B_n + C_n - D_n\ ,$$

where

$$\begin{aligned}A_n &= \frac{1}{\sqrt{n}}\sum_{1\le i\le 2n}\left(Y_i(1)D_i - E[Y_i(1)D_i|X^{(n)}, D^{(n)}]\right) \\ B_n &= \frac{1}{\sqrt{n}}\sum_{1\le i\le 2n}\left(Y_i(0)(1-D_i) - E[Y_i(0)(1-D_i)|X^{(n)}, D^{(n)}]\right) \\ C_n &= \frac{1}{\sqrt{n}}\sum_{1\le i\le 2n}\left(E[Y_i(1)D_i|X^{(n)}, D^{(n)}] - D_i E[Y_i(1)]\right) \\ D_n &= \frac{1}{\sqrt{n}}\sum_{1\le i\le 2n}\left(E[Y_i(0)(1-D_i)|X^{(n)}, D^{(n)}] - (1-D_i)E[Y_i(0)]\right)\ .\end{aligned}$$

Note that, conditional on $X^{(n)}$ and $D^{(n)}$, A_n and B_n are independent and C_n and D_n are constant.

We first analyze the limiting behavior of A_n. Conditional on $X^{(n)}$ and $D^{(n)}$, the terms in this sum are independent, but not identically distributed. We proceed by verifying that the condition in Linderberg's Central Limit Theorem holds in probability conditional on $X^{(n)}$ and $D^{(n)}$. To that end, define

$$s_n^2 = s_n^2(X^{(n)}, D^{(n)}) = \sum_{1\le i\le 2n} \text{Var}[Y_i(1)D_i|X^{(n)}, D^{(n)}]$$

and note that

$$\begin{aligned}s_n^2 &= \sum_{1\le i\le 2n: D_i=1} \text{Var}[Y_i(1)|X^{(n)}, D^{(n)}] \\ &= \sum_{1\le i\le 2n: D_i=1} \text{Var}[Y_i(1)|X^{(n)}] \\ &= \sum_{1\le i\le 2n: D_i=1} \text{Var}[Y_i(1)|X_i]\ ,\end{aligned}$$

where the first equality follows from Assumption 2.2 and the second follows from the fact that $Q_n = Q^n$. It follows that

$$\frac{s_n^2}{n} = \frac{1}{2n}\sum_{1\le i\le 2n} \text{Var}[Y_i(1)|X_i] + \left(\frac{1}{2n}\sum_{1\le i\le 2n: D_i=1} \text{Var}[Y_i(1)|X_i] - \frac{1}{2n}\sum_{1\le i\le 2n: D_i=0} \text{Var}[Y_i(1)|X_i]\right)\ .$$

Using Assumption 2.1(b), we have that

$$\frac{1}{2n}\sum_{1\le i\le 2n} \text{Var}[Y_i(1)|X_i] \xrightarrow{p} E[\text{Var}[Y_i(1)|X_i]]\ .$$

Note further that

$$\begin{aligned}
&\left|\frac{1}{2n}\sum_{1\leq i\leq 2n:D_i=1}\mathrm{Var}[Y_i(1)|X_i]-\frac{1}{2n}\sum_{1\leq i\leq 2n:D_i=0}\mathrm{Var}[Y_i(1)|X_i]\right| \\
\leq\ & \frac{1}{2n}\sum_{1\leq j\leq n}\left|\mathrm{Var}[Y_{\pi(2j)}(1)|X_{\pi(2j)}]-\mathrm{Var}[Y_{\pi(2j-1)}(1)|X_{\pi(2j-1)}]\right| \\
\lesssim\ & \frac{1}{n}\sum_{1\leq j\leq n}|X_{\pi(2j)}-X_{\pi(2j-1)}|\xrightarrow{P}0\ ,
\end{aligned}$$

where the first inequality follows by inspection, the second inequality exploits Assumption 2.1(c) and the convergence to zero follows from Assumption 2.3. Hence,

$$\frac{s_n^2}{n}\xrightarrow{P}E[\mathrm{Var}[Y_i(1)|X_i]]>0\ , \tag{55}$$

where the final inequality exploits Assumption 2.1(a). Next, we argue for any $\epsilon>0$ that

$$\frac{1}{s_n^2}\sum_{1\leq i\leq 2n}E[|Y_i(1)D_i-E[Y_i(1)D_i|X^{(n)},D^{(n)}]|^2I\{|Y_i(1)D_i-E[Y_i(1)D_i|X^{(n)},D^{(n)}]|>\epsilon s_n\}|X^{(n)},D^{(n)}]\xrightarrow{P}0\ .$$

To this end, first note that for any $m>0$ we have that

$$P\{\epsilon s_n>m\}\to 1\ . \tag{56}$$

Note further that Assumption 2.2 implies that

$$E[Y_i(1)D_i|X^{(n)},D^{(n)}]=D_iE[Y_i(1)|X_i]\ , \tag{57}$$

so the lefthand-side of the preceding display may be written as

$$\begin{aligned}
&\frac{1}{s_n^2}\sum_{1\leq i\leq 2n:D_i=1}E[|Y_i(1)-E[Y_i(1)|X_i]|^2I\{|Y_i(1)-E[Y_i(1)|X_i]|>\epsilon s_n\}|X^{(n)},D^{(n)}] \\
\leq\ & \left(\frac{s_n^2}{n}\right)^{-1}\frac{1}{n}\sum_{1\leq i\leq 2n}E[|Y_i(1)-E[Y_i(1)|X_i]|^2I\{|Y_i(1)-E[Y_i(1)|X_i]|>\epsilon s_n\}|X^{(n)},D^{(n)}] \\
\leq\ & \left(\frac{s_n^2}{n}\right)^{-1}\frac{1}{n}\sum_{1\leq i\leq 2n}E[|Y_i(1)-E[Y_i(1)|X_i]|^2I\{|Y_i(1)-E[Y_i(1)|X_i]|>m\}|X^{(n)},D^{(n)}]+o_P(1) \\
=\ & \left(\frac{s_n^2}{n}\right)^{-1}\frac{1}{n}\sum_{1\leq i\leq 2n}E[|Y_i(1)-E[Y_i(1)|X_i]|^2I\{|Y_i(1)-E[Y_i(1)|X_i]|>m\}|X_i]+o_P(1) \\
\xrightarrow{P}\ & (E[\mathrm{Var}[Y_i(1)|X_i]])^{-1}E[|Y_i(1)-E[Y_i(1)|X_i]|^2I\{|Y_i(1)-E[Y_i(1)|X_i]|>m\}]\ ,
\end{aligned}$$

where the first inequality follows by inspection, the second inequality exploits (55)–(56), the equality follows from Assumption 2.2 and the fact that $Q_n=Q^n$, and the convergence in probability follows from (55) and the fact that Assumption 2.1(b) implies

$$E[|Y_i(1)-E[Y_i(1)|X_i]|^2]=E[\mathrm{Var}[Y_i(1)|X_i]]\leq E[Y_i^2(1)]<\infty\ . \tag{58}$$

Note further that (58) implies that

$$\lim_{m\to\infty}E[|Y_i(1)-E[Y_i(1)|X_i]|^2I\{|Y_i(1)-E[Y_i(1)|X_i]|>m\}]=0\ .$$

The condition in Lindeberg's Central Limit Theorem therefore holds in probability. It follows by a subsequencing argument similar to that used in the proof of Lemma 6.5 below that

$$\sup_{t\in\mathbf{R}}\left|P\{A_n\leq t|X^{(n)},D^{(n)}\}-\Phi(t/\sqrt{E[\mathrm{Var}[Y_i(1)|X_i]]})\right|\xrightarrow{P}0\ .$$

A similar argument establishes that

$$\sup_{t\in\mathbf{R}}\left|P\{B_n\leq t|X^{(n)},D^{(n)}\}-\Phi(t/\sqrt{E[\mathrm{Var}[Y_i(0)|X_i]]})\right|\xrightarrow{P}0\ .$$

Since A_n and B_n are independent conditional on $X^{(n)}$ and $D^{(n)}$, it follows by another subsequencing argument that

$$\sup_{t\in\mathbf{R}}\left|P\{A_n - B_n \le t|X^{(n)}, D^{(n)}\} - \Phi(t/\sqrt{E[\mathrm{Var}[Y_i(0)|X_i]] + E[\mathrm{Var}[Y_i(0)|X_i]]})\right| \xrightarrow{P} 0 \ . \tag{59}$$

To analyze C_n, first note that (57) implies that

$$C_n = \frac{1}{\sqrt{n}} \sum_{1\le i\le 2n} D_i \left(E[Y_i(1)|X_i] - E[Y_i(1)]\right) \ , \tag{60}$$

so

$$E[C_n|X^{(n)}] = \frac{1}{2\sqrt{n}} \sum_{1\le i\le 2n} \left(E[Y_i(1)|X_i] - E[Y_i(1)]\right) \ . \tag{61}$$

Furthermore,

$$\begin{aligned}
\mathrm{Var}[C_n|X^{(n)}] &= \mathrm{Var}[C_n - E[C_n|X^{(n)}]|X^{(n)}] \\
&= \mathrm{Var}\left[\frac{1}{\sqrt{n}} \sum_{1\le i\le 2n} \left(D_i - \frac{1}{2}\right)\left(E[Y_i(1)|X_i] - E[Y_i(1)]\right)\middle| X^{(n)}\right] \\
&= \mathrm{Var}\left[\frac{1}{\sqrt{n}} \sum_{1\le i\le 2n} \left(D_i - \frac{1}{2}\right) E[Y_i(1)|X_i]\middle| X^{(n)}\right] \\
&= \frac{1}{4n} \sum_{1\le j\le n} \left(E[Y_{\pi(2j)}(1)|X_{\pi(2j)}] - E[Y_{\pi(2j-1)}(1)|X_{\pi(2j-1)}]\right)^2 \\
&\lesssim \frac{1}{n} \sum_{1\le j\le n} \left(X_{\pi(2j)} - X_{\pi(2j-1)}\right)^2 \xrightarrow{P} 0 \ ,
\end{aligned}$$

where the first equality exploits properties of conditional variances, the second follows from (60)–(61), the third exploits the fact that $\sum_{1\le i\le 2n} D_i = n$, the fourth exploits the distribution of $D^{(n)}|X^{(n)}$, the inequality follows from Assumption 2.1(c), and the convergence in probability follows from Assumption 2.3. For any $\epsilon > 0$, it thus follows from Chebychev's inequality that

$$P\{|C_n - E[C_n|X^{(n)}]| > \epsilon|X^{(n)}\} \le \frac{\mathrm{Var}[C_n|X^{(n)}]}{\epsilon^2} \xrightarrow{P} 0 \ .$$

Since probabilities are bounded, we have further that

$$P\{|C_n - E[C_n|X^{(n)}]| > \epsilon\} \xrightarrow{P} 0 \ .$$

Hence,

$$C_n = \frac{1}{2\sqrt{n}} \sum_{1\le i\le 2n} \left(E[Y_i(1)|X_i] - E[Y_i(1)]\right) + o_P(1) \ . \tag{62}$$

A similar argument establishes that

$$D_n = \frac{1}{2\sqrt{n}} \sum_{1\le i\le 2n} \left(E[Y_i(0)|X_i] - E[Y_i(0)]\right) + o_P(1) \ . \tag{63}$$

Hence,

$$\begin{aligned}
C_n - D_n &= \frac{1}{2\sqrt{n}} \sum_{1\le i\le 2n} \left(\left(E[Y_i(1)|X_i] - E[Y_i(1)]\right) - \left(E[Y_i(0)|X_i] - E[Y_i(0)]\right)\right) + o_P(1) \\
&= \frac{\sqrt{2}}{2}\frac{1}{\sqrt{2n}} \sum_{1\le i\le 2n} \left(\left(E[Y_i(1)|X_i] - E[Y_i(1)]\right) - \left(E[Y_i(0)|X_i] - E[Y_i(0)]\right)\right) + o_P(1) \\
&\xrightarrow{d} N\left(0, \frac{1}{2}E\left[\left(\left(E[Y_i(1)|X_i] - E[Y_i(1)]\right) - \left(E[Y_i(0)|X_i] - E[Y_i(0)]\right)\right)^2\right]\right) \ ,
\end{aligned}$$

where the first equality follows from (62)–(63), the second equality follows by inspection, and the convergence in distribution follows from Slutsky's theorem and the Central Limit Theorem.

The desired conclusion (53) now follows by a subsequencing argument. To see this, suppose by way of contradiction that

(53) fails. This implies that there exists $\delta > 0$ and a subsequence n_k along which

$$\sup_{t \in \mathbf{R}} |P\{\sqrt{n_k}(\hat{\Delta}_{n_k} - \Delta(Q)) \leq t\} - \Phi(t/\nu)| \to \delta \ . \tag{64}$$

By considering a further subsequence if necessary, which, by an abuse of notation, we continue to denote by n_k, it follows from (59) that

$$A_{n_k} - B_{n_k} \xrightarrow{d} N(0, E[\mathrm{Var}[Y_i(0)|X_i]] + E[\mathrm{Var}[Y_i(0)|X_i]]) \text{ w.p.1 (conditional on } X^{(n_k)} \text{ and } D^{(n_k)}) \ .$$

Since $C_{n_k} - D_{n_k}$ is constant conditional on $X^{(n_k)}$ and $D^{(n_k)}$, Lemma 6.2 establishes that

$$\sqrt{n_k}(\hat{\Delta}_{n_k} - \Delta) = A_{n_k} - B_{n_k} + C_{n_k} - D_{n_k} \xrightarrow{d} N(0, \nu^2) \ ,$$

which, by Polya's Theorem, implies a contradiction to (64).

Finally, in order to complete the proof, note that

$$\begin{aligned}
& E[\mathrm{Var}[Y_i(1)|X_i]] + E[\mathrm{Var}[Y_i(0)|X_i]] + \frac{1}{2}E\left[((E[Y_i(1)|X_i] - E[Y_i(1)]) - (E[Y_i(0)|X_i] - E[Y_i(0)]))^2\right] \\
= \ & \mathrm{Var}[Y_i(1)] + \mathrm{Var}[Y_i(0)] - \mathrm{Var}[E[Y_i(1)|X_i]] - \mathrm{Var}[E[Y_i(0)|X_i]] \\
& \qquad + \frac{1}{2}E\left[((E[Y_i(1)|X_i] - E[Y_i(1)]) - (E[Y_i(0)|X_i] - E[Y_i(0)]))^2\right] \\
= \ & \mathrm{Var}[Y_i(1)] + \mathrm{Var}[Y_i(0)] - \frac{1}{2}\mathrm{Var}[E[Y_i(1)|X_i]] - \frac{1}{2}\mathrm{Var}[E[Y_i(0)|X_i]] \\
& \qquad - E[(E[Y_i(1)|X_i] - E[Y_i(1)])(E[Y_i(0)|X_i] - E[Y_i(0)])] \\
= \ & \mathrm{Var}[Y_i(1)] + \mathrm{Var}[Y_i(0)] - \frac{1}{2}E\left[((E[Y_i(1)|X_i] - E[Y_i(1)]) + (E[Y_i(0)|X_i] - E[Y_i(0)]))^2\right] \ ,
\end{aligned}$$

which establishes that the two expressions for ν^2 in the statement of the theorem are in fact equivalent. ■

Lemma 6.5. *If Assumptions 2.1–2.3 hold, then* $\hat{\mu}_n(d) \xrightarrow{P} E[Y_i(d)]$ *and* $\hat{\sigma}_n^2(d) \xrightarrow{P} Var[Y_i(d)]$*, where* $\hat{\mu}_n(d)$ *and* $\hat{\sigma}_n^2(d)$ *are defined in (5) and (6), respectively.*

Proof: Note that

$$\begin{aligned}
\hat{\mu}_n(d) &= \frac{1}{n}\sum_{1 \leq i \leq 2n} Y_i(d) I\{D_i = d\} \\
\hat{\sigma}_n^2(d) &= \frac{1}{n}\sum_{1 \leq i \leq 2n} (Y_i - \hat{\mu}_n(d))^2 I\{D_i = d\} \\
&= \frac{1}{n}\sum_{1 \leq i \leq 2n} Y_i^2(d) I\{D_i = d\} - \hat{\mu}_n^2(d) \ .
\end{aligned}$$

It therefore suffices to show that

$$\frac{1}{n}\sum_{1 \leq i \leq 2n} Y_i^r(d) I\{D_i = d\} \xrightarrow{P} E[Y_i^r(d)]$$

for $r \in \{1, 2\}$. We prove this result only for $r = 1$ and $d = 1$; the other cases can be proven similarly. To this end, write

$$\begin{aligned}
\frac{1}{n}\sum_{1 \leq i \leq 2n} Y_i(1) I\{D_i = 1\} &= \frac{1}{n}\sum_{1 \leq i \leq 2n} Y_i(1) D_i \\
&= \frac{1}{n}\sum_{1 \leq i \leq 2n} \left(Y_i(1) D_i - E[Y_i(1) D_i | X^{(n)}, D^{(n)}]\right) + \frac{1}{n}\sum_{1 \leq i \leq 2n} E[Y_i(1) D_i | X^{(n)}, D^{(n)}] \ .
\end{aligned}$$

Next, note that

$$\begin{aligned}
& \frac{1}{n}\sum_{1 \leq i \leq 2n} E[Y_i(1) D_i | X^{(n)}, D^{(n)}] \\
= \ & \frac{1}{n}\sum_{1 \leq i \leq 2n} D_i E[Y_i(1) | X_i] \\
= \ & \frac{1}{n}\sum_{1 \leq i \leq 2n : D_i = 1} E[Y_i(1) | X_i]
\end{aligned}$$

$$= \frac{1}{2n}\sum_{1\le i\le 2n} E[Y_i(1)|X_i] + \left(\frac{1}{2n}\sum_{1\le i\le 2n:D_i=1} E[Y_i(1)|X_i] - \frac{1}{2n}\sum_{1\le i\le 2n:D_i=0} E[Y_i(1)|X_i]\right) ,$$

where the first equality exploits (57) and the second and third equalities follow by inspection. Note further that

$$\begin{aligned}
&\left|\frac{1}{2n}\sum_{1\le i\le 2n:D_i=1} E[Y_i(1)|X_i] - \frac{1}{2n}\sum_{1\le i\le 2n:D_i=0} E[Y_i(1)|X_i]\right| \\
\le\ & \frac{1}{2n}\sum_{1\le j\le n} |E[Y_{\pi(2j)}(1)|X_{\pi(2j)}] - E[Y_{\pi(2j-1)}(1)|X_{\pi(2j-1)}]| \\
\lesssim\ & \frac{1}{n}\sum_{1\le j\le n} |X_{\pi(2j)} - X_{\pi(2j-1)}| \xrightarrow{P} 0 ,
\end{aligned}$$

where the first inequality follows by inspection, the second exploits Assumption 2.1(c) and the convergence in probability follows from Assumption 2.3. Since Assumption 2.1(b) implies that $E[|E[Y_i(1)|X_i]|] \le E[|Y_i(1)|] < \infty$, it follows that

$$\frac{1}{n}\sum_{1\le i\le 2n} E[Y_i(1)D_i|X^{(n)}, D^{(n)}] \xrightarrow{P} E[E[Y_i(1)|X_i]] = E[Y_i(1)] .$$

To complete the argument, we argue that

$$\frac{1}{n}\sum_{1\le i\le 2n}\Big(Y_i(1)D_i - E[Y_i(1)D_i|X^{(n)}, D^{(n)}]\Big) \xrightarrow{P} 0 . \tag{65}$$

For this purpose, we proceed by verifying that (49) in Lemma 6.3 holds in probability conditional on $X^{(n)}$ and $D^{(n)}$. To that end, note for any $m > 0$ that

$$\begin{aligned}
&\frac{1}{2n}\sum_{1\le i\le 2n} E[|Y_i(1)D_i - E[Y_i(1)D_i|X^{(n)}, D^{(n)}]|I\{|Y_i(1)D_i - E[Y_i(1)D_i|X^{(n)}, D^{(n)}]| > m\}|X^{(n)}, D^{(n)}] \\
=\ & \frac{1}{2n}\sum_{1\le i\le 2n} E[|Y_i(1)D_i - D_iE[Y_i(1)|X_i]|I\{|Y_i(1)D_i - D_iE[Y_i(1)|X_i]| > m\}|X^{(n)}, D^{(n)}] \\
\le\ & \frac{1}{2n}\sum_{1\le i\le 2n} E[|Y_i(1) - E[Y_i(1)|X_i]|I\{|Y_i(1) - E[Y_i(1)|X_i]| > m\}|X^{(n)}, D^{(n)}] \\
=\ & \frac{1}{2n}\sum_{1\le i\le 2n} E[|Y_i(1) - E[Y_i(1)|X_i]|I\{|Y_i(1) - E[Y_i(1)|X_i]| > m\}|X_i] \\
\xrightarrow{P}\ & E[|Y_i(1) - E[Y_i(1)|X_i]|I\{|Y_i(1) - E[Y_i(1)|X_i]| > m\}] ,
\end{aligned} \tag{66}$$

where the first and fourth equalities follow from (57), the inequality follows by inspection, and the convergence in probability follows from (58). The desired conclusion (65) now follows by a subsequencing argument. To see this, suppose by way of contradiction that (65) fails. This implies that there exists $\epsilon > 0$, $\delta > 0$ and a subsequence n_k along which

$$P\left\{\left|\frac{1}{n_k}\sum_{1\le i\le 2n_k}\Big(Y_i(1)D_i - E[Y_i(1)D_i|X^{(n_k)}, D^{(n_k)}]\Big)\right| > \epsilon\right\} \to \delta . \tag{67}$$

By considering a further subsequence if necessary, which, by an abuse of notation, we continue to denote by n_k, it follows from (57), (58) and (66) that

$$\begin{aligned}
\lim_{m\to\infty}\limsup_{k\to\infty}\frac{1}{2n}\sum_{1\le i\le 2n}\Big(&E[|Y_i(1)D_i - E[Y_i(1)D_i|X^{(n)}, D^{(n)}]| \\
&\times I\{|Y_i(1)D_i - E[Y_i(1)D_i|X^{(n)}, D^{(n)}]| > m\}|X^{(n)}, D^{(n)}]\Big) = 0
\end{aligned}$$

w.p.1 (conditional on $X^{(n_k)}$ and $D^{(n_k)}$). Lemma 6.3 implies, however, that

$$\frac{1}{n_k}\sum_{1\le i\le 2n_k}\Big(Y_i(1)D_i - E[Y_i(1)D_i|X^{(n_k)}, D^{(n_k)}]\Big) \to 0 \text{ w.p.1 (conditional on } X^{(n_k)} \text{ and } D^{(n_k)}) ,$$

which implies a contradiction to (67). ■

Lemma 6.6. *If Assumptions 2.1–2.3 hold, then*

$$\hat{\tau}_n^2 \xrightarrow{d} E[\,Var[Y_i(1)|X_i]] + E[\,Var[Y_i(0)|X_i]] + E\left[(E[Y_i(1)|X_i] - E[Y_i(0)|X_i])^2\right] \ ,$$

where $\hat{\tau}_n^2$ is defined in (21).

Proof: Note that

$$\hat{\tau}_n^2 = \frac{1}{n}\sum_{1\le j\le n}(Y_{\pi(2j)} - Y_{\pi(2j-1)})^2 = \frac{1}{n}\sum_{1\le i\le 2n} Y_i^2 - \frac{2}{n}\sum_{1\le j\le n} Y_{\pi(2j)}Y_{\pi(2j-1)} \ .$$

Since

$$\frac{1}{n}\sum_{1\le i\le 2n} Y_i^2 = \hat{\sigma}_n^2(1) - \hat{\mu}_n^2(1) + \hat{\sigma}_n^2(0) - \hat{\mu}_n^2(0) \ ,$$

it follows from Lemma 6.5 that

$$\frac{1}{n}\sum_{1\le i\le 2n} Y_i^2 \xrightarrow{P} E[Y_i^2(1)] + E[Y_i^2(0)] \ .$$

Next, we argue that

$$\frac{2}{n}\sum_{1\le j\le n} Y_{\pi(2j)}Y_{\pi(2j-1)} \xrightarrow{P} 2E[\mu_1(X_i)\mu_0(X_i)] \ ,$$

where we use the notation $\mu_d(X_i)$ to denote $E[Y_i(d)|X_i]$. To this end, first note that

$$E\left[Y_{\pi(2j)}Y_{\pi(2j-1)}\middle|X^{(n)}\right] = \frac{1}{2}\mu_1(X_{\pi(2j)})\mu_0(X_{\pi(2j-1)}) + \frac{1}{2}\mu_0(X_{\pi(2j)})\mu_1(X_{\pi(2j-1)}) \ , \tag{68}$$

so

$$\begin{aligned}
& E\left[\frac{2}{n}\sum_{1\le j\le n} Y_{\pi(2j)}Y_{\pi(2j-1)}\middle|X^{(n)}\right] \\
= \ & \frac{2}{n}\sum_{1\le j\le n} E\left[Y_{\pi(2j)}Y_{\pi(2j-1)}\middle|X^{(n)}\right] \\
= \ & \frac{1}{n}\sum_{1\le j\le n} \mu_1(X_{\pi(2j)})\mu_0(X_{\pi(2j-1)}) + \mu_0(X_{\pi(2j)})\mu_1(X_{\pi(2j-1)}) \\
= \ & \frac{1}{n}\sum_{1\le j\le n} \Big(\mu_1(X_{\pi(2j)})(\mu_0(X_{\pi(2j-1)}) - \mu_0(X_{\pi(2j)})) + \mu_1(X_{\pi(2j)})\mu_0(X_{\pi(2j)}) \\
& \qquad + \mu_1(X_{\pi(2j-1)})(\mu_0(X_{\pi(2j)}) - \mu_0(X_{\pi(2j-1)})) + \mu_1(X_{\pi(2j-1)})\mu_0(X_{\pi(2j-1)})\Big) \\
= \ & \frac{1}{n}\sum_{1\le i\le 2n} \mu_1(X_i)\mu_0(X_i) + \frac{1}{n}\sum_{1\le j\le n} (\mu_1(X_{\pi(2j-1)}) - \mu_1(X_{\pi(2j)}))(\mu_0(X_{\pi(2j)}) - \mu_0(X_{\pi(2j-1)})) \ ,
\end{aligned}$$

where the second equality follows from (68) and the other equalities follow by inspection. Assumption 2.3 implies that

$$\left|\frac{1}{n}\sum_{1\le j\le n} (\mu_1(X_{\pi(2j-1)}) - \mu_1(X_{\pi(2j)}))(\mu_0(X_{\pi(2j)}) - \mu_0(X_{\pi(2j-1)}))\right| \lesssim \frac{1}{n}\sum_{1\le j\le n} |X_{\pi(2j-1)}) - X_{\pi(2j)})|^2 \xrightarrow{P} 0 \ .$$

Furthermore, since

$$E[|\mu_1(X_i)\mu_0(X_i)|] \lesssim E[\mu_1^2(X_i)] + E[\mu_1^2(X_i)] \le E[Y_i^2(1)] + E[Y_i^2(0)] < \infty \ ,$$

we have that

$$E\left[\frac{2}{n}\sum_{1\le j\le n} Y_{\pi(2j)}Y_{\pi(2j-1)}\middle|X^{(n)}\right] \xrightarrow{P} 2E[\mu_1(X_i)\mu_0(X_i)] \ .$$

To complete the argument, we show that

$$\frac{1}{n}\sum_{1\le j\le n}\left(Y_{\pi(2j)}Y_{\pi(2j-1)} - E\left[Y_{\pi(2j)}Y_{\pi(2j-1)}\middle|X^{(n)}\right]\right) \xrightarrow{P} 0 \ . \tag{69}$$

For this purpose, we proceed by verifying that (49) in Lemma 6.3 holds in probability conditional on $X^{(n)}$. In what follows, we make repeated use of the following facts for any real numbers a and b and $\lambda > 0$:

$$|a+b|I\{|a+b| > \lambda\} \quad \le \quad 2|a|I\left\{|a| > \frac{\lambda}{2}\right\} + 2|b|I\left\{|b| > \frac{\lambda}{2}\right\} \tag{70}$$

$$|ab|I\{|ab|>\lambda\} \quad \leq \quad a^2 I\{|a|>\sqrt{\lambda}\} + b^2 I\{|b|>\sqrt{\lambda}\}\ . \tag{71}$$

Note that the second of these facts follows from the first together with the inequality $2|ab| \leq a^2+b^2$. Next, note that

$$\begin{aligned}
& \frac{1}{n}\sum_{1\leq j\leq n} E\left[\left|Y_{\pi(2j)}Y_{\pi(2j-1)} - E\left[Y_{\pi(2j)}Y_{\pi(2j-1)}\middle|X^{(n)}\right]\right| I\left\{\left|Y_{\pi(2j)}Y_{\pi(2j-1)} - E\left[Y_{\pi(2j)}Y_{\pi(2j-1)}\middle|X^{(n)}\right]\right|>\lambda\right\}\middle|X^{(n)}\right] \\
\lesssim\ & \frac{1}{n}\sum_{1\leq j\leq n} E\left[\left|Y_{\pi(2j)}Y_{\pi(2j-1)}\right| I\left\{\left|Y_{\pi(2j)}Y_{\pi(2j-1)}\right|>\frac{\lambda}{2}\right\}\middle|X^{(n)}\right] \\
& \quad + \frac{1}{n}\sum_{1\leq j\leq n} E\left[\left|E\left[Y_{\pi(2j)}Y_{\pi(2j-1)}\middle|X^{(n)}\right]\right| I\left\{\left|E\left[Y_{\pi(2j)}Y_{\pi(2j-1)}\middle|X^{(n)}\right]\right|>\frac{\lambda}{2}\right\}\middle|X^{(n)}\right] \\
\leq\ & \frac{1}{n}\sum_{1\leq j\leq n} E\left[Y^2_{\pi(2j)} I\left\{|Y_{\pi(2j)}|>\sqrt{\frac{\lambda}{2}}\right\}\middle|X^{(n)}\right] + \frac{1}{n}\sum_{1\leq j\leq n} E\left[Y^2_{\pi(2j-1)} I\left\{|Y_{\pi(2j-1)}|>\sqrt{\frac{\lambda}{2}}\right\}\middle|X^{(n)}\right] \\
& \quad + \frac{1}{n}\sum_{1\leq j\leq n} \left|E\left[Y_{\pi(2j)}Y_{\pi(2j-1)}\middle|X^{(n)}\right]\right| I\left\{\left|E\left[Y_{\pi(2j)}Y_{\pi(2j-1)}\middle|X^{(n)}\right]\right|>\frac{\lambda}{2}\right\} \\
\lesssim\ & \frac{1}{n}\sum_{1\leq j\leq n} E\left[Y^2_{\pi(2j)} I\left\{|Y_{\pi(2j)}|>\sqrt{\frac{\lambda}{2}}\right\}\middle|X^{(n)}\right] + \frac{1}{n}\sum_{1\leq j\leq n} E\left[Y^2_{\pi(2j-1)} I\left\{|Y_{\pi(2j-1)}|>\sqrt{\frac{\lambda}{2}}\right\}\middle|X^{(n)}\right] \\
& \quad + \frac{1}{n}\sum_{1\leq j\leq n} \left|\mu_1(X_{\pi(2j)})\mu_0(X_{\pi(2j-1)})\right| I\left\{\left|\mu_1(X_{\pi(2j)})\mu_0(X_{\pi(2j-1)})\right|>\frac{\lambda}{2}\right\} \\
& \quad + \frac{1}{n}\sum_{1\leq j\leq n} \left|\mu_0(X_{\pi(2j)})\mu_1(X_{\pi(2j-1)})\right| I\left\{\left|\mu_0(X_{\pi(2j)})\mu_1(X_{\pi(2j-1)})\right|>\frac{\lambda}{2}\right\} \\
\lesssim\ & \frac{1}{n}\sum_{1\leq j\leq n} E\left[Y^2_{\pi(2j)} I\left\{|Y_{\pi(2j)}|>\sqrt{\frac{\lambda}{2}}\right\}\middle|X^{(n)}\right] + \frac{1}{n}\sum_{1\leq j\leq n} E\left[Y^2_{\pi(2j-1)} I\left\{|Y_{\pi(2j-1)}|>\sqrt{\frac{\lambda}{2}}\right\}\middle|X^{(n)}\right] \\
& \quad + \frac{1}{n}\sum_{1\leq j\leq n} \mu_1^2(X_{\pi(2j)}) I\left\{|\mu_1(X_{\pi(2j)})|>\sqrt{\frac{\lambda}{2}}\right\} + \frac{1}{n}\sum_{1\leq j\leq n} \mu_0^2(X_{\pi(2j-1)}) I\left\{|\mu_0(X_{\pi(2j-1)})|>\sqrt{\frac{\lambda}{2}}\right\} \\
& \quad + \frac{1}{n}\sum_{1\leq j\leq n} \mu_0^2(X_{\pi(2j)}) I\left\{|\mu_0(X_{\pi(2j)})|>\sqrt{\frac{\lambda}{2}}\right\} + \frac{1}{n}\sum_{1\leq j\leq n} \mu_1^2(X_{\pi(2j-1)}) I\left\{|\mu_1(X_{\pi(2j-1)})|>\sqrt{\frac{\lambda}{2}}\right\} \\
\lesssim\ & \frac{1}{n}\sum_{1\leq j\leq n} E\left[Y^2_{\pi(2j)}(1) I\left\{|Y_{\pi(2j)}(1)|>\sqrt{\frac{\lambda}{2}}\right\}\middle|X^{(n)}\right] + \frac{1}{n}\sum_{1\leq j\leq n} E\left[Y^2_{\pi(2j)}(0) I\left\{|Y_{\pi(2j)}(0)|>\sqrt{\frac{\lambda}{2}}\right\}\middle|X^{(n)}\right] \\
& \quad + \frac{1}{n}\sum_{1\leq j\leq n} E\left[Y^2_{\pi(2j-1)}(1) I\left\{|Y_{\pi(2j-1)}(1)|>\sqrt{\frac{\lambda}{2}}\right\}\middle|X^{(n)}\right] + \frac{1}{n}\sum_{1\leq j\leq n} E\left[Y^2_{\pi(2j-1)}(0) I\left\{|Y_{\pi(2j-1)}(0)|>\sqrt{\frac{\lambda}{2}}\right\}\middle|X^{(n)}\right] \\
& \quad + \frac{1}{n}\sum_{1\leq j\leq n} \mu_1^2(X_{\pi(2j)}) I\left\{|\mu_1(X_{\pi(2j)})|>\sqrt{\frac{\lambda}{2}}\right\} + \frac{1}{n}\sum_{1\leq j\leq n} \mu_0^2(X_{\pi(2j-1)}) I\left\{|\mu_0(X_{\pi(2j-1)})|>\sqrt{\frac{\lambda}{2}}\right\} \\
& \quad + \frac{1}{n}\sum_{1\leq j\leq n} \mu_0^2(X_{\pi(2j)}) I\left\{|\mu_0(X_{\pi(2j)})|>\sqrt{\frac{\lambda}{2}}\right\} + \frac{1}{n}\sum_{1\leq j\leq n} \mu_1^2(X_{\pi(2j-1)}) I\left\{|\mu_1(X_{\pi(2j-1)})|>\sqrt{\frac{\lambda}{2}}\right\} \\
\lesssim\ & \frac{1}{n}\sum_{1\leq i\leq 2n} E\left[Y_i^2(1) I\left\{|Y_i(1)|>\sqrt{\frac{\lambda}{2}}\right\}\middle|X_i\right] + \frac{1}{n}\sum_{1\leq i\leq 2n} E\left[Y_i^2(0) I\left\{|Y_i(0)|>\sqrt{\frac{\lambda}{2}}\right\}\middle|X_i\right] \\
& \quad + \frac{1}{n}\sum_{1\leq i\leq 2n} \mu_0^2(X_i) I\left\{|\mu_0(X_i)|>\sqrt{\frac{\lambda}{2}}\right\} + \frac{1}{n}\sum_{1\leq i\leq 2n} \mu_1^2(X_i) I\left\{|\mu_1(X_i)|>\sqrt{\frac{\lambda}{2}}\right\} \\
\xrightarrow{P}\ & E\left[Y_i^2(1) I\left\{|Y_i(1)|>\sqrt{\frac{\lambda}{2}}\right\}\right] + E\left[Y_i^2(0) I\left\{|Y_i(0)|>\sqrt{\frac{\lambda}{2}}\right\}\right] \\
& \quad + E\left[\mu_1^2(X_i) I\left\{|\mu_1(X_i)|>\sqrt{\frac{\lambda}{2}}\right\}\right] + E\left[\mu_0^2(X_i) I\left\{|\mu_0(X_i)|>\sqrt{\frac{\lambda}{2}}\right\}\right]
\end{aligned}$$

where the third inequality exploits (68). Since $E[Y_i^2(d)] < \infty$ and $E[\mu_d^2(X_i)] \leq E[Y_i^2(d)]$, we have that

$$\lim_{\lambda\to\infty} E\left[\mu_d^2(X_i) I\left\{|\mu_d(X_i)|>\sqrt{\frac{\lambda}{2}}\right\}\right] = 0$$

$$\lim_{\lambda\to\infty} E\left[Y_i^2(1) I\left\{|Y_i(1)| > \sqrt{\frac{\lambda}{2}}\right\}\right] = 0 \ .$$

It now follows from a subsequencing argument as in the proof of Lemma 6.5 that (69) holds. Hence,

$$\begin{aligned}
\hat{\tau}_n^2 &\xrightarrow{P} E[Y_i^2(1)] + E[Y_i^2(0)] - 2E[\mu_1(X_i)\mu_0(X_i)] \\
&= E[\mathrm{Var}[Y_i(1)|X_i]] + E[\mathrm{Var}[Y_i(0)|X_i]] + E\left[(\mu_1(X_i) - \mu_0(X_i))^2\right] \\
&= E[\mathrm{Var}[Y_i(1)|X_i]] + E[\mathrm{Var}[Y_i(0)|X_i]] + E\left[(E[Y_i(1)|X_i] - E[Y_i(0)|X_i])^2\right] \ ,
\end{aligned}$$

as desired. ■

Lemma 6.7. *If Assumptions 2.1–2.4 hold, then*

$$\hat{\lambda}_n^2 \xrightarrow{P} E[(E[Y_i(1)|X_i] - E[Y_i(0)|X_i])^2] \ , \tag{72}$$

where $\hat{\lambda}_n$ is defined in (22).

Proof: Let $\mu_d(X_i)$ denote $E[Y_i(d)|X_i]$ and note that

$$\begin{aligned}
& E\left[(Y_{\pi(4j-3)} - Y_{\pi(4j-2)})(Y_{\pi(4j-1)} - Y_{\pi(4j)})(D_{\pi(4j-3)} - D_{\pi(4j-2)})(D_{\pi(4j-1)} - D_{\pi(4j)})\Big| X^{(n)}\right] \\
= \ & \frac{1}{4}(\mu_1(X_{\pi(4j-3)}) - \mu_0(X_{\pi(4j-2)}))(\mu_1(X_{\pi(4j-1)}) - \mu_0(X_{\pi(4j)})) \\
& \quad -\frac{1}{4}(\mu_0(X_{\pi(4j-3)}) - \mu_1(X_{\pi(4j-2)}))(\mu_1(X_{\pi(4j-1)}) - \mu_0(X_{\pi(4j)})) \\
& \quad -\frac{1}{4}(\mu_1(X_{\pi(4j-3)}) - \mu_0(X_{\pi(4j-2)}))(\mu_0(X_{\pi(4j-1)}) - \mu_1(X_{\pi(4j)})) \\
& \quad +\frac{1}{4}(\mu_0(X_{\pi(4j-3)}) - \mu_1(X_{\pi(4j-2)}))(\mu_0(X_{\pi(4j-1)}) - \mu_1(X_{\pi(4j)})) \\
= \ & \frac{1}{4}\sum_{k\in\{2,3\},\ell\in\{0,1\}} \mu_1(X_{\pi(4j-k)})\mu_1(X_{\pi(4j-\ell)}) + \frac{1}{4}\sum_{k\in\{2,3\},\ell\in\{0,1\}} \mu_0(X_{\pi(4j-k)})\mu_0(X_{\pi(4j-\ell)}) \\
& \quad -\frac{1}{4}\sum_{k\in\{2,3\},\ell\in\{0,1\}} \left(\mu_0(X_{\pi(4j-k)})\mu_1(X_{\pi(4j-\ell)}) + \mu_1(X_{\pi(4j-k)})\mu_0(X_{\pi(4j-\ell)})\right) \ .
\end{aligned}$$

Hence, in order to show that

$$E[\hat{\lambda}_n^2|X^{(n)}] \xrightarrow{P} E[(\mu_1(X_i) - \mu_0(X_i))^2] \ , \tag{73}$$

it suffices to show that

$$\frac{1}{2n}\sum_{1\le j\le \frac{n}{2}}\sum_{k\in\{2,3\},\ell\in\{0,1\}} \mu_1(X_{\pi(4j-k)})\mu_1(X_{\pi(4j-\ell)}) \xrightarrow{P} E[\mu_1^2(X_i)] \tag{74}$$

$$\frac{1}{2n}\sum_{1\le j\le \frac{n}{2}}\sum_{k\in\{2,3\},\ell\in\{0,1\}} \mu_0(X_{\pi(4j-k)})\mu_0(X_{\pi(4j-\ell)}) \xrightarrow{P} E[\mu_0^2(X_i)] \tag{75}$$

$$\begin{aligned}
&\frac{1}{2n}\sum_{1\le j\le \frac{n}{2}}\sum_{k\in\{2,3\},\ell\in\{0,1\}} \left(\mu_0(X_{\pi(4j-k)})\mu_1(X_{\pi(4j-\ell)}) + \mu_1(X_{\pi(4j-k)})\mu_0(X_{\pi(4j-\ell)})\right) \\
&\qquad \xrightarrow{P} 2E[\mu_1(X_i)\mu_0(X_i)] \ .
\end{aligned} \tag{76}$$

We first prove (74). To see this, note that

$$\begin{aligned}
\mu_1(X_{\pi(4j-k)})\mu_1(X_{\pi(4j-\ell)}) &= \mu_1^2(X_{\pi(4j-k)}) + \mu_1(X_{\pi(4j-k)})(\mu_1(X_{\pi(4j-\ell)}) - \mu_1(X_{\pi(4j-k)})) \\
\mu_1(X_{\pi(4j-k)})\mu_1(X_{\pi(4j-\ell)}) &= \mu_1^2(X_{\pi(4j-\ell)}) - \mu_1(X_{\pi(4j-\ell)})(\mu_1(X_{\pi(4j-\ell)}) - \mu_1(X_{\pi(4j-k)})) \ ,
\end{aligned}$$

so

$$\mu_1(X_{\pi(4j-k)})\mu_1(X_{\pi(4j-\ell)}) = \frac{1}{2}\mu_1^2(X_{\pi(4j-k)}) + \frac{1}{2}\mu_1^2(X_{\pi(4j-\ell)}) - \frac{1}{2}(\mu_1(X_{\pi(4j-\ell)}) - \mu_1(X_{\pi(4j-k)}))^2 \ .$$

It follows that

$$\frac{1}{2n}\sum_{1\le j\le \frac{n}{2}}\sum_{k\in\{2,3\},\ell\in\{0,1\}} \mu_1(X_{\pi(4j-k)})\mu_1(X_{\pi(4j-\ell)})$$

$$= \frac{1}{2n}\sum_{1\le i\le 2n}\mu_1^2(X_i) - \frac{1}{4n}\sum_{k\in\{2,3\},\ell\in\{0,1\}}\sum_{1\le j\le\frac{n}{2}}(\mu_1(X_{\pi(4j-\ell)})-\mu_1(X_{\pi(4j-k)}))^2 \ .$$

But, Assumption 2.1 implies that

$$\frac{1}{4n}\sum_{k\in\{2,3\},\ell\in\{0,1\}}\sum_{1\le j\le\frac{n}{2}}(\mu_1(X_{\pi(4j-\ell)})-\mu_1(X_{\pi(4j-k)}))^2 \lesssim \frac{1}{n}\sum_{1\le j\le\frac{n}{2}}|X_{\pi(4j-k)}-X_{\pi(4j-\ell)}|^2 \xrightarrow{P} 0 \ ,$$

where the convergence in probability to zero follows from Assumption 2.4. Since $E[\mu_1^2(X_i)] \le E[Y_i^2(1)]$, we have that

$$\frac{1}{2n}\sum_{1\le i\le 2n}\mu_1^2(X_i) \xrightarrow{P} E[\mu_1^2(X_i)] \ .$$

It thus follows that (74) holds. Similar arguments may be used to establish (75)-(76), from which (73) follows.

To complete the proof, it remains only to show that

$$\hat{\lambda}_n^2 - E[\hat{\lambda}_n^2|X^{(n)}] \xrightarrow{P} 0 \ .$$

This fact may be established by verifying that (49) in Lemma 6.3 holds in probability conditionally on $X^{(n)}$, which may be accomplished by repeated application of (70) and (71), as in the proof of Lemma 6.6. ■

Lemma 6.8. *Let*

$$\tilde{R}_n(t) = P\left\{\frac{1}{\sqrt{n}}\sum_{1\le j\le n}\epsilon_j(Y_{\pi(2j)}-Y_{\pi(2j-1)})(D_{\pi(2j)}-D_{\pi(2j-1)}) \le t \middle| W^{(n)}\right\} \ ,$$

where, independently of $W^{(n)}$, $\epsilon_j, j=1,\ldots,n$ are i.i.d. Rademacher random variables. If Assumptions 2.1–2.3 hold, then

$$\sup_{t\in\mathbf{R}}\left|\tilde{R}_n(t)-\Phi(t/\tau)\right| \xrightarrow{P} 0 \ ,$$

where

$$\tau^2 = E[Var[Y_i(1)|X_i]] + E[Var[Y_i(0)|X_i]] + E\left[(E[Y_i(1)|X_i]-E[Y_i(0)|X_i])^2\right] \ . \tag{77}$$

Proof: Using the fact that $\epsilon_j, j=1,\ldots,n$ and $\epsilon_j(D_{\pi(2j)}-D_{\pi(2j-1)}), j=1,\ldots,n$ have the same distribution conditional on $W^{(n)}$, we have that

$$\tilde{R}_n(t) = P\left\{\frac{1}{\sqrt{n}}\sum_{1\le j\le n}\epsilon_j(Y_{\pi(2j)}-Y_{\pi(2j-1)}) \le t \middle| W^{(n)}\right\} \ .$$

We now proceed by applying part (ii) of Lemma 11.3.3 in Lehmann and Romano (2005) with $C_{n,j} = (Y_{\pi(2j)}-Y_{\pi(2j-1)})$, which requires

$$\frac{\max_{1\le j\le n}C_{n,j}^2}{\sum_{1\le j\le n}C_{n,j}^2} \xrightarrow{P} 0 \ . \tag{78}$$

From Lemma 6.6, we see that $\frac{1}{n}\sum_{1\le j\le n}C_{n,j}^2 = \hat{\tau}_n^2 \xrightarrow{P} \tau^2 > 0$, where the inequality exploits Assumption 2.1(a). Furthermore,

$$\begin{aligned}\frac{\max_{1\le j\le n}C_{n,j}^2}{n} \xrightarrow{P} 0 \quad &\lesssim \quad \frac{\max_{1\le j\le n}(Y_{\pi(2j-1)}^2+Y_{\pi(2j)}^2)}{n} \\ &\lesssim \quad \frac{\max_{1\le i\le 2n}Y_i^2}{n} \\ &\lesssim \quad \frac{\max_{1\le i\le 2n}(Y_i(1)^2+Y_i(0)^2)}{n} \\ &\xrightarrow{P} \quad 0 \ ,\end{aligned}$$

where the first inequality follows by exploiting the fact that $|a-b|^2 \le 2(a^2+b^2)$ for any real numbers a and b, the second and third inequalities follow by inspection, and the convergence in probability to zero follows from Lemma 6.1 and Assumption 2.1(b). Hence, (78) holds, from which the desired conclusion now follows easily by appealing to the aforementioned lemma and Polya's theorem. ■

Lemma 6.9. *Let*

$$\check{\nu}_n^2(\epsilon_1,\ldots,\epsilon_n) = \hat{\tau}_n^2 - \frac{1}{2}(\check{\lambda}_n^2(\epsilon_1,\ldots,\epsilon_n)+\check{\Delta}_n^2(\epsilon_1,\ldots,\epsilon_n)) \ , \tag{79}$$

where $\hat{\tau}_n^2$ *is defined in* (21),

$$\begin{aligned}
&\check{\lambda}_n^2(\epsilon_1,\ldots,\epsilon_n) \\
&\quad = \frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} \epsilon_{2j-1}\epsilon_{2j}(Y_{\pi(4j-3)} - Y_{\pi(4j-2)})(Y_{\pi(4j-1)} - Y_{\pi(4j)})(D_{\pi(4j-3)} - D_{\pi(4j-2)})(D_{\pi(4j-1)} - D_{\pi(4j)}) \\
&\check{\Delta}_n(\epsilon_1,\ldots,\epsilon_n) \\
&\quad = \frac{1}{n}\sum_{1\le j\le n} \epsilon_j(Y_{\pi(2j)} - Y_{\pi(2j-1)})(D_{\pi(2j)} - D_{\pi(2j-1)}) \ ,
\end{aligned}$$

and, independently of $W^{(n)}$, $\epsilon_j, j = 1,\ldots,n$ *are i.i.d. Rademacher random variables. If Assumptions 2.1–2.3 hold, then*

$$\check{\nu}_n^2(\epsilon_1,\ldots,\epsilon_n) \xrightarrow{P} \tau^2 \ ,$$

where τ^2 *is defined in* (77).

Proof: From Lemma 6.6, we see that $\hat{\tau}_n^2 \xrightarrow{P} \tau^2$. From Lemma 6.8, we have further that $\check{\Delta}_n(\epsilon_1,\ldots,\epsilon_n) \xrightarrow{P} 0$. It therefore suffices to show that $\check{\lambda}_n^2(\epsilon_1,\ldots,\epsilon_n) \xrightarrow{P} 0$. In order to do so, note that $\check{\lambda}_n^2(\epsilon_1,\ldots,\epsilon_n)$ may be decomposed into sums of the form

$$\frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} \epsilon_{2j-1}\epsilon_{2j}Y_{\pi(4j-k)}Y_{\pi(4j-\ell)}D_{\pi(4j-k')}D_{\pi(4j-\ell')} \ . \tag{80}$$

where $(k,k') \in \{2,3\}^2$ and $(\ell,\ell') \in \{0,1\}^2$. Furthermore, conditional on $W^{(n)}$, the terms in any such sum are independent with mean zero. We may therefore argue that any such sum tends to zero in probability by verifying that (49) in Lemma 6.3 holds in probability conditional on $W^{(n)}$. To this end, note that

$$\begin{aligned}
&\frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} E\Big[|\epsilon_{2j-1}\epsilon_{2j}Y_{\pi(4j-k)}Y_{\pi(4j-\ell)}D_{\pi(4j-k')}D_{\pi(4j-\ell')}| \\
&\qquad \times I\left\{|\epsilon_{2j-1}\epsilon_{2j}Y_{\pi(4j-k)}Y_{\pi(4j-\ell)}D_{\pi(4j-k')}D_{\pi(4j-\ell')}| > \lambda\right\}\Big|W^{(n)}\Big] \\
\le\ & \frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} E\left[|Y_{\pi(4j-k)}Y_{\pi(4j-\ell)}|I\left\{|Y_{\pi(4j-k)}Y_{\pi(4j-\ell)}| > \lambda\right\}\Big|W^{(n)}\right] \\
\le\ & \frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} |Y_{\pi(4j-k)}Y_{\pi(4j-\ell)}|I\left\{|Y_{\pi(4j-k)}Y_{\pi(4j-\ell)}| > \lambda\right\} \\
\le\ & \frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} Y_{\pi(4j-k)}^2 I\left\{|Y_{\pi(4j-k)}| > \sqrt{\lambda}\right\} + \frac{2}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} Y_{\pi(4j-\ell)}^2 I\left\{|Y_{\pi(4j-\ell)}| > \sqrt{\lambda}\right\} \\
\lesssim\ & \frac{1}{n}\sum_{1\le i\le 2n} Y_i^2 I\left\{|Y_i| > \sqrt{\lambda}\right\} \\
\le\ & \frac{1}{n}\sum_{1\le i\le 2n} (Y_i^2(1) + Y_i^2(0)) I\left\{(Y_i^2(1) + Y_i^2(0))^{\frac{1}{2}} > \sqrt{\lambda}\right\} \\
\xrightarrow{P}\ & E\left[(Y_i^2(1) + Y_i^2(0)) I\left\{(Y_i^2(1) + Y_i^2(0))^{\frac{1}{2}} > \sqrt{\lambda}\right\}\right] \ ,
\end{aligned}$$

where the first inequality follows from the fact that $|\epsilon_j| = 1$ for all $1 \le j \le n$ and $|D_i| \le 1$ for all $1 \le i \le 2n$, the second inequality exploits the fact that $\pi = \pi_n(X^{(n)})$ and both $Y^{(n)}$ and $X^{(n)}$ are contained in $W^{(n)}$, the third inequality follows from (71) used in the proof of Lemma 6.6, the fourth inequality follows by inspection, the fifty inequality uses the fact that $Y_i^2 \le Y_i^2(1) + Y_i^2(0)$, and the convergence in probability follows from Assumption 2.1(b). Since $E[Y_i^2(d)] < \infty$, we have that

$$\lim_{\lambda\to\infty} E\left[(Y_i^2(1) + Y_i^2(0)) I\left\{(Y_i^2(1) + Y_i^2(0))^{\frac{1}{2}} > \sqrt{\lambda}\right\}\right] = 0 \ .$$

It now follows from a subsequencing argument as in the proof of Lemma 6.5 that (80) tends to zero in probability. The desired result thus follows. ■

References

Armitage, P., Berry, G. and Matthews, J. N. S. (2008). *Statistical Methods in Medical Research*. John Wiley & Sons.

Athey, S. and Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments*, vol. 1. Elsevier, 73–140.

Bertrand, M. and Duflo, E. (2017). Field experiments on discrimination. In *Handbook of Economic Field Experiments*, vol. 1. Elsevier, 309–393.

Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, **1** 200–232.

Bugni, F. A., Canay, I. A. and Shaikh, A. M. (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, **113** 1784–1796.

Bugni, F. A., Canay, I. A. and Shaikh, A. M. (2019). Inference under covariate-adaptive randomization with multiple treatments. Becker Friedman Institute for Economics Working Paper 2019-19, University of Chicago.

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

Crépon, B., Devoto, F., Duflo, E. and Parienté, W. (2015). Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in Morocco. *American Economic Journal: Applied Economics*, **7** 123–50.

Derigs, U. (1988). Solving non-bipartite matching problems via shortest path techniques. *Annals of Operations Research*, **13** 225–261.

DiCiccio, C. J. and Romano, J. P. (2017). Robust permutation tests for correlation and regression coefficients. *Journal of the American Statistical Association*, **112** 1211–1220.

Duflo, E., Glennerster, R. and Kremer, M. (2007). Using randomization in development economics research: A toolkit. In *Handbook of Development Economics*, vol. 4. Elsevier, 3895–3962.

Edmonds, J. (1965). Maximum matching and a polyhedron with 0, 1-vertices. *Journal of Research of the National Bureau of Standards B*, **69** 55–56.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Glennerster, R. and Takavarasha, K. (2013). *Running randomized evaluations: A practical guide*. Princeton University Press.

Heard, K., OToole, E., Naimpally, R. and Bressler, L. (2017). Real world challenges to randomization and their solutions.

Heckman, J. J., Pinto, R., Shaikh, A. M. and Yavitz, A. (2011). Inference with imperfect randomization: The case of the Perry preschool program. Tech. rep., National Bureau of Economic Research.

Hsu, H. and Lachenbruch, P. A. (2007). Paired t-test. Wiley Online Library, 1–3.

Imai, K., King, G., Nall, C. et al. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the mexican universal health insurance evaluation. *Statistical Science*, **24** 29–53.

Janssen, A. (1997). Studentized permutation tests for non-iid hypotheses and the generalized behrens-fisher problem. *Statistics & Probability Letters*, **36** 9–21.

Lee, S. and Shaikh, A. M. (2014). Multiple testing and heterogeneous treatment effects: Re-evaluating the effect of Progresa on school enrollment. *Journal of Applied Econometrics*, **29** 612–626.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. 3rd ed. Springer, New York.

List, J. A. and Rasul, I. (2011). Field experiments in labor economics. vol. 4. Elsevier, 103–228.

Riach, P. A. and Rich, J. (2002). Field experiments of discrimination in the market place. *The Economic Journal*, **112** 480–518.

Rosenberger, W. F. and Lachin, J. M. (2015). *Randomization in clinical trials: Theory and Practice*. John Wiley & Sons.

White, H. (2013). An introduction to the use of randomised control trials to evaluate development interventions. *Journal of Development Effectiveness*, **5** 30–49.