THE UNIVERSITY OF CHICAGO

ESSAYS ON THE ECONOMETRICS OF RANDOMIZED CONTROLLED TRIALS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

KENNETH C. GRIFFIN DEPARTMENT OF ECONOMICS

BY
YUEHAO BAI

CHICAGO, ILLINOIS
JUNE 2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am deeply grateful for the continued guidance and support from my advisors Azeem Shaikh, Stephane Bonhomme, Alex Torgovitsky, and Leo Bursztyn. Azeem has always been encouraging me and putting up with my incessant display of concern and stress during my past six years. I am indebted to him for most things I know about econometrics, and especially thankful for two things he taught me: how to identify interesting research questions, and that I shouldn't be afraid to ask simple questions about anything. I am always humbled by the knowledge and passion of Stephane and without his help I might have abandoned Chapter 1 of my dissertation at a very early stage. Alex and Leo have always been generous with their time and they helped me overcome several obstacles in my dissertation I thought to be insurmountable. Together my advisors and the whole department have shaped my view of economics and I am proud to be a Chicago-produced economist and econometrician.

I would also like to thank several other professors at Chicago for their advice and support. In particular, I would like to thank Marinho Bertanha, Max Farrell, Guillaume Pouliot, and Max Tabord-Meehan. I also want to thank Wooyong Lee and Joshua Shea for their comments on my dissertation.

Throughout grad school I have received tremendous support from my peer, and I'm fortunate to count them among my close friends. Among them I would especially like to thank Fang Fu, Wenji Xu, Kai Hao Yang, and Cong Zhang.

Finally, I would like to thank my parents and my wife Xiaofan for their unwavering support. They have always had the greatest faith in me through the ups and downs.

ABSTRACT

This dissertation studies the econometrics of the design and analysis of randomized controlled trials (RCTs). Chapter 1, coauthored with Joseph Romano and Azeem Shaikh, studies inference for the average treatment effect (ATE) in RCTs where treatment status is determined according matched-pair designs. We assume that units are paired according to observed baseline covariates instead of some function of the covariates, and Chapter 2 extends these results to settings where units are paired according to (random) functions of the covariates. 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.

Chapter 2 studies the optimality of matched-pair designs in RCTs. Matched-pair designs are examples of stratified randomization, in which the researcher partitions a set of units into strata based on their observed covariates and assign a fraction of units in each stratum to

treatment. A matched-pair design is such a procedure with two units per stratum. Despite the prevalence of stratified randomization in RCTs, implementations differ vastly. We provide an econometric framework in which, among all stratified randomization procedures, the optimal one in terms of the mean-squared error of the difference-in-means estimator is a matched-pair design that orders units according to a scalar function of their covariates and matches adjacent units. Our framework captures a leading motivation for stratifying in the sense that it shows that the proposed matched-pair design additionally minimizes the magnitude of the ex-post bias, i.e., the bias of the estimator conditional on realized treatment status. We then consider empirical counterparts to the optimal stratification using data from pilot experiments and provide two different procedures depending on whether the sample size of the pilot is large or small. For each procedure, we develop methods for testing the null hypothesis that the average treatment effect equals a prespecified value. Each test we provide is asymptotically exact in the sense that the limiting rejection probability under the null equals the nominal level. We run an experiment on the Amazon Mechanical Turk using one of the proposed procedures, replicating one of the treatment arms in DellaVigna and Pope (2018), and find the standard error decreases by 29%, so that only half of the sample size is required to attain the same standard error.

Chapter 3 studies the more fundamental question of why randomization should be used in controlled trials when the objective is to estimate the ATE precisely. In particular, I study the minimax optimality of certain randomization schemes and assignment schemes in estimating "reasonable" parameters including the average treatment effect, when treatment effects are heterogeneous. By a randomization scheme, I mean the distribution over a group of permutations of a given treatment assignment vector. By an assignment scheme, I mean the joint distribution over assignment vectors, linear estimators, and permutations of assignment vectors. I show that for any given assignment vector and any estimator, the complete randomization scheme is minimax optimal for any objective function satisfying quasi-convexity, where the worst-case is over a permutation-invariant class of distributions

of the data. Objective functions satisfying quasi-convexity include the expectation operator, the quantile function, and the survival function. Under further conditions on the distribution of the data, I characterize the minimax optimal assignment scheme, where the worst-case is again over a permutation-invariant class of distributions of the data. Finally, I provide insights on how randomization might improve estimation, even when permutation invariance does not hold.

CHAPTER 1

INFERENCE IN EXPERIMENTS WITH MATCHED PAIRS

1.1 Introduction

This chapter 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 1.4.1–1.4.3 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 1.3.1 and 1.3.2 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 1.3.11, 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 1.2, 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 1.3. In Section 1.4, we develop some results that ensure that units in each pair are

suitably "close" in terms of their baseline covariates. Finally, in Section 1.5, we examine the finite-sample behavior of these tests via a small simulation study. Proofs of all results are provided in the Appendix.

1.2 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

$$Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i) \ . \tag{1.1}$$

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 (1.1), 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

$$\Delta(Q) = E[Y_i(1) - Y_i(0)] \ . \tag{1.2}$$

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

$$H_0 : \Delta(Q) = \Delta_0 \text{ versus } H_1 : \Delta(Q) \neq \Delta_0 \tag{1.3}$$

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

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

Assumption 1.2.1. The distribution Q is such that

(a) $0 < E[\text{Var}[Y_i(d)|X_i]]$ 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 1.2.1(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 A.9.3 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 1.2.1(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 1.2.2. 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 1.2.3. The pairs used in determining treatment status satisfy

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

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 1.2.4. The pairs used in determining treatment status satisfy

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

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

In Section 1.4 below, we provide results to facilitate constructing pairs satisfying Assumptions 1.2.3–1.2.4 under weak assumptions on Q. We emphasize, however, that Assumption 1.2.4, in contrast to Assumptions 1.2.1–1.2.3, will not be required for many of our results. Furthermore, given pairs satisfying Assumption 1.2.3, it will frequently be possible to "re-

order" them so that Assumption 1.2.4 is satisfied. See Theorem 1.4.3 below for further details.

Remark 1.2.1. Note that Assumption 1.2.2 implies that

$$(Y^{(n)}(1), Y^{(n)}(0)) \perp\!\!\!\perp D^{(n)} \big| X^{(n)} \ . \tag{1.4}$$

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

Remark 1.2.2. 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 1.2.3. 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. ■

1.3 Main Results

1.3.1 Two-Sample t-Test

In this section, we consider using the two-sample t-test to test (1.3) at level $\alpha \in (0,1)$. In order to define this test, for $d \in \{0,1\}$, define

$$\hat{\mu}_n(d) = \frac{1}{n} \sum_{1 \le i \le 2n : D_i = d} Y_i \tag{1.5}$$

$$\hat{\sigma}_n^2(d) = \frac{1}{n} \sum_{1 \le i \le 2n : D_i = d} (Y_i - \hat{\mu}_n(d))^2 \tag{1.6}$$

and let

$$\hat{\Delta}_n = \hat{\mu}_n(1) - \hat{\mu}_n(0) \ . \tag{1.7}$$

The two-sample t-test is given by

$$\phi_n^{t-\text{test}}(Z^{(n)}) = I\{|T_n^{t-\text{test}}(Z^{(n)})| > z_{1-\frac{\alpha}{2}}\} \ , \tag{1.8}$$

where

$$T_n^{t-\text{test}}(Z^{(n)}) = \frac{\sqrt{n}(\hat{\Delta}_n - \Delta_0)}{\sqrt{\hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0)}} \tag{1.9}$$

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 (1.9) and, as a consequence, the two-sample t-test defined in (1.8). 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 1.3.1. *Suppose Q satisfies Assumption 1.2.1 and the treatment assignment mechanism satisfies Assumptions 1.2.2–1.2.3. Then,*

$$\frac{\sqrt{n}(\hat{\Delta}_n - \Delta(Q))}{\sqrt{\hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0)}} \xrightarrow{d} N(0, \varsigma^2_{t-test}) \ , \tag{1.10}$$

where

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

Thus, for the problem of testing (1.3) *at level $\alpha \in (0,1)$, $\phi_n^{t-test}(Z^{(n)})$ defined in* (1.8) *satisfies*

$$\lim_{n\to\infty} E[\phi_n^{t-test}(Z^{(n)})] = P\{\varsigma_{t-test}\ |G| > z_{1-\frac{\alpha}{2}}\} \leq \alpha \ , \tag{1.11}$$

where $G \sim N(0,1)$, whenever Q additionally satisfies the null hypothesis, i.e., $\Delta(Q) = \Delta_0$. Furthermore, the inequality in (1.11) *is strict unless*

$$E[Y_i(1) + Y_i(0)] = E[Y_i(1) + Y_i(0)|X_i] \tag{1.12}$$

with probability one under Q.

Remark 1.3.1. Theorem 1.3.1 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 (1.12). 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 1.5, 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 1.3.2. 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 A.9.5 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 A.9.4 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. ■

1.3.2 "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

$$\phi_n^{\text{paired}}(Z^{(n)}) = I\{|T_n^{\text{paired}}(Z^{(n)})| > z_{1-\frac{\alpha}{2}}\} \ , \tag{1.13}$$

where

$$T_n^{\text{paired}}(Z^{(n)}) = \frac{\sqrt{n}(\hat{\Delta}_n - \Delta_0)}{\sqrt{\frac{1}{n}\sum_{1\leq j\leq n}(Y_{\pi(2j)} - Y_{\pi(2j-1)})^2 - \hat{\Delta}_n^2}} \tag{1.14}$$

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 (1.14), and, as a consequence, the "matched pairs" t-test defined in (1.13). 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 1.3.2. *Suppose Q satisfies Assumption 1.2.1 and the treatment assignment mechanism satisfies Assumptions 1.2.2–1.2.3. Then,*

$$\frac{\sqrt{n}(\hat{\Delta}_n - \Delta(Q))}{\sqrt{\frac{1}{n}\sum_{1\leq j\leq n}(Y_{\pi(2j)} - Y_{\pi(2j-1)})^2 - \hat{\Delta}_n^2}} \xrightarrow{d} N(0, \varsigma^2_{\text{paired}}) \ , \tag{1.15}$$

where

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

Thus, for the problem of testing (1.3) *at level $\alpha \in (0,1)$, $\phi_n^{paired}(Z^{(n)})$ defined in* (1.13) *satisfies*

$$\lim_{n\to\infty} E[\phi_n^{paired}(Z^{(n)})] = P\{\varsigma_{paired}\,|G| > z_{1-\frac{\alpha}{2}}\} \leq \alpha \ , \tag{1.16}$$

where $G \sim N(0,1)$, whenever Q additionally satisfies the null hypothesis, i.e., $\Delta(Q) = \Delta_0$. Furthermore, the inequality in (1.16) *is strict unless*

$$E[Y_i(1) - Y_i(0)] = E[Y_i(1) - Y_i(0)|X_i] \tag{1.17}$$

with probability one under Q.

Remark 1.3.3. While Theorem 1.3.2 is qualitatively similar to Theorem 1.3.1, it is worth emphasizing the difference between (1.12) and (1.17). 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 1.3.4. The test statistic in (1.14) 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 (1.3). In particular, the result in (1.16) continues to hold for the test formed by replacing the $\hat{\Delta}_n$ in the denominator on the right-hand side of (1.14) with Δ_0. ■

Remark 1.3.5. The literature has also at times advocated estimation of $\Delta(Q)$ via estimation by ordinary least squares of the coefficient on D_i in

$$Y_i = \beta D_i + \sum_{1 \leq j \leq n} \lambda_j I\{i \in \{\pi(2j), \pi(2j-1)\}\} + \epsilon_i \ . \tag{1.18}$$

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 (1.18) 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. ■

1.3.3 "Adjusted" t-Test

The proofs of Theorems 1.3.1 and 1.3.2 in the Appendix rely upon Lemma A.9.4, which establishes that

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

where

$$\nu^2 = \text{Var}[Y_i(1)] + \text{Var}[Y_i(0)]$$

$$-\frac{1}{2}E\left[((E[Y_i(1)|X_i]-E[Y_i(1)])+(E[Y_i(0)|X_i]-E[Y_i(0)]))^2\right] \ . \quad (1.19)$$

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 (1.19). As discussed further in Remark 1.3.7 below, there exist multiple consistent estimators of (1.19), but a convenient one for our purposes is given by

$$\hat{\nu}_n^2 = \hat{\tau}_n^2 - \frac{1}{2}(\hat{\lambda}_n^2 + \hat{\Delta}_n^2) \ , \quad (1.20)$$

where

$$\hat{\tau}_n^2 = \frac{1}{n}\sum_{1\le j\le n}(Y_{\pi(2j)}-Y_{\pi(2j-1)})^2 \quad (1.21)$$

$$\hat{\lambda}_n^2 = \frac{2}{n}\sum_{1\le j\le \lfloor\frac{n}{2}\rfloor}\Big((Y_{\pi(4j-3)}-Y_{\pi(4j-2)})(Y_{\pi(4j-1)}-Y_{\pi(4j)}) \times(D_{\pi(4j-3)}-D_{\pi(4j-2)})(D_{\pi(4j-1)}-D_{\pi(4j)})\Big) \ . \quad (1.22)$$

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

$$\phi_n^{t-\text{test,adj}}(Z^{(n)}) = I\{|T_n^{t-\text{test,adj}}(Z^{(n)})| > z_{1-\frac{\alpha}{2}}\} \quad (1.23)$$

with

$$T_n^{t-\text{test,adj}}(Z^{(n)}) = \frac{\sqrt{n}(\hat{\Delta}_n-\Delta_0)}{\hat{\nu}_n} \ , \quad (1.24)$$

satisfies the desired property.

Theorem 1.3.3. *Suppose Q satisfies Assumption 1.2.1 and the treatment assignment mech-*

anism satisfies Assumptions 1.2.2–1.2.4. Then,

$$\frac{\sqrt{n}(\hat{\Delta}_n - \Delta(Q))}{\hat{\nu}_n} \xrightarrow{d} N(0,1) \ . \tag{1.25}$$

Thus, for the problem of testing (1.3) *at level* $\alpha \in (0,1)$, $\phi_n^{t-test,adj}(Z^{(n)})$ *defined in* (1.23) *satisfies*

$$\lim_{n\to\infty} E[\phi_n^{t-test,adj}(Z^{(n)})] = \alpha \ , \tag{1.26}$$

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

Remark 1.3.6. While our discussion has focused on two-sided null hypotheses as described in (1.3), the convergence in distribution results described in (1.10), (1.15) and (1.25) have straightforward implications for other tests, such related tests of one-sided null hypotheses.

■

Remark 1.3.7. As mentioned previously, other consistent estimators of (1.19) exist. For instance, one may consider the estimator given by

$$\tilde{\nu}_n^2 = \hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0) - \frac{1}{2}\left(\tilde{\lambda}_n^2 - (\hat{\mu}_n(1) + \hat{\mu}_n(0))^2\right) \ , \tag{1.27}$$

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 1.3.3, it is possible to show that Theorem 1.3.3 remains true when $\hat{\nu}_n^2$ defined in (1.20) is replaced by $\tilde{\nu}_n^2$ defined in (1.27).

■

1.3.4 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 (1.3) with $\Delta_0 = 0$; for the problem of testing (1.3) more generally, the construction below may be applied with Y_i replaced with $Y_i - D_i\Delta_0$. See Remark 1.3.10 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.,

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

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

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

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

$$\phi_n^{\text{rand}}(Z^{(n)}) = I\{T_n(Z^{(n)}) > \hat{R}_n^{-1}(1-\alpha)\} \ , \tag{1.28}$$

where

$$\hat{R}_n(t) = \frac{1}{|\mathbf{G}_n(\pi)|} \sum_{g \in \mathbf{G}_n(\pi)} I\{T_n(gZ^{(n)}) \leq t\} \ . \tag{1.29}$$

Here, $\hat{R}_n^{-1}(1-\alpha)$ is understood to be $\inf\{t \in \mathbf{R} : \hat{R}_n(t) \geq 1-\alpha\}$. 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 1.3.8. 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 (1.29). 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 1.3.4 below remains true with such an

approximation; Theorem 1.3.5 below also remains true with such an approximation provided that $B \to \infty$ as $n \to \infty$. ■

Finite-Sample Results

Before developing the large-sample properties of the randomization test given by (1.28), 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 (1.28) has rejection probability no greater than the nominal level for the following more restrictive null hypothesis:

$$\tilde{H}_0 : Y_i(1)|X_i \stackrel{d}{=} Y_i(0)|X_i \ . \tag{1.30}$$

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 1.3.4. *Suppose the treatment assignment mechanism satisfies Assumption 1.2.2. For the problem of testing* (1.30) *at level* $\alpha \in (0,1)$, $\phi_n^{\text{rand}}(Z^{(n)})$ *defined in* (1.28) *with any* $T_n(Z^{(n)})$ *satisfies*

$$E[\phi_n^{\text{rand}}(Z^{(n)})] \leq \alpha \tag{1.31}$$

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

Remark 1.3.9. By modifying the test defined in (1.28) 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 (1.31). See Lehmann and Romano (2005, Chapter 15) for further details. ■

Large-Sample Properties

In this section, we establish the large-sample validity of the randomization test given by (1.28) with a suitable choice of test statistic for testing (1.3). In particular, we show that the limiting rejection probability of the proposed test equals the nominal level under the null hypothesis.

Theorem 1.3.5. *Suppose Q satisfies Assumption 1.2.1 and the treatment assignment mechanism satisfies Assumptions 1.2.2–1.2.4. Let $T_n(Z^{(n)}) = |T_n^{t-test,adj}(Z^{(n)})|$, where $T_n^{t-test,adj}(Z^{(n)})$ is defined in* (1.24)*. For such a choice of $T_n(Z^{(n)})$,*

$$\sup_{t \in \mathbf{R}} \left| \hat{R}_n(t) - (\Phi(t) - \Phi(-t)) \right| \xrightarrow{P} 0 \ , \tag{1.32}$$

where $\Phi(\cdot)$ is the standard normal c.d.f. Thus, for the problem of testing (1.3) *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*

$$\lim_{n \to \infty} E[\phi_n^{\text{rand}}(Z^{(n)})] = \alpha \ , \tag{1.33}$$

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

Remark 1.3.10. For completeness, we briefly describe the way in which Theorem 1.3.5 extends to testing (1.3) 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 1.3.5 with $\tilde{Z}^{(n)}$. For such a test, we have, under the assumptions of Theorem 1.3.5, 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 1.3.11. The conclusion in Theorem 1.3.5 depends heavily on the choice of test statistic in the definition of (1.28). In order to illustrate this phenomenon, consider the test defined by (1.28) with $T_n(Z^{(n)}) = |\sqrt{n}\hat{\Delta}_n|$. Using Lemmas A.9.4 and A.9.8 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 1.3.2. In particular, it has limiting rejection probability under the null hypothesis no greater than α and strictly less than α unless (1.17) 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 (1.19). 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 (1.28) with $T_n(Z^{(n)}) = \frac{|\sqrt{n}\hat{\Delta}_n|}{\tilde{\nu}_n}$, where $\tilde{\nu}_n^2$ is defined in (1.27). Even though $\tilde{\nu}_n^2$ is consistent for (1.19), as discussed in Remark 1.3.7, it is possible to show using arguments similar to those used in establishing Theorem 1.3.3 that this test also behaves similarly under the null hypothesis to the "matched pairs" t-test described in Section 1.3.2. ■

1.4 Algorithms for Pairing

In this section, we describe different algorithms for pairing units so that Assumptions 1.2.3–1.2.4 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 1.2.3–1.2.4 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 1.4.1. *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 1.2.3–1.2.4.

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

$$\frac{1}{n} \sum_{1 \leq j \leq n} |X_{\check{\pi}(2j-1)} - X_{\check{\pi}(2j)}|^2 \leq \frac{1}{n} \sum_{1 \leq j \leq n} |X_{\check{\pi}(2j-1)} - X_{\check{\pi}(2j)}| \; . \tag{1.34}$$

In order to satisfy Assumption 1.2.3, it is therefore natural to choose π so as to minimize the right-hand side of (1.34). 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 (1.34) for π minimizing the right-hand side of (1.34), which implies, in particular, that pairing units in this way satisfies Assumption 1.2.3.

Theorem 1.4.2. *Suppose $\text{supp}(X_i) \subseteq [0,1]^k$. Let π be any permutation of $2n$ elements minimizing the right-hand side of* (1.34)*. Then, for each integer $m > 1$, we have that*

$$\frac{1}{n} \sum_{1 \leq j \leq n} |X_{\pi(2j)} - X_{\pi(2j-1)}| \leq \frac{\sqrt{k}}{m} + \frac{m^{k-1} 2\sqrt{k}}{n} \; . \tag{1.35}$$

In particular, if $m \asymp n^{\frac{1}{k}}$, then π satisfies Assumption 1.2.3.

Given a pairing satisfying Assumption 1.2.3, we now turn our attention to ensuring that the pairing further satisfies Assumption 1.2.4. To this end, choose $\bar{\pi}$ so as to minimize

$$\frac{2}{n} \sum_{1 \leq j \leq \lfloor \frac{n}{2} \rfloor} |\bar{X}_{\check{\pi}(2j)} - \bar{X}_{\check{\pi}(2j-1)}| \; , \tag{1.36}$$

where

$$\bar{X}_j = \frac{X_{\pi(2j)} + X_{\pi(2j-1)}}{2} \; . \tag{1.37}$$

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 1.2.4 in addition to Assumption 1.2.3.

Theorem 1.4.3. *Suppose* $supp(X_i) \subseteq [0,1]^k$. *Let* π *be a permutation of* $2n$ *elements such that Assumption 1.2.3 is satisfied and* $\bar{\pi}$ *be any permutation of* n *elements minimizing* (1.36). *Define a permutation* $\tilde{\pi}$ *of* $2n$ *elements so that*

$$\tilde{\pi}(2j) = \pi(2\bar{\pi}(j)) \quad \textit{and} \quad \tilde{\pi}(2j-1) = \pi(2\bar{\pi}(j)-1) \tag{1.38}$$

for $1 \leq j \leq n$. *Then,* $\tilde{\pi}$ *satisfies Assumptions 1.2.3–1.2.4.*

1.5 Simulations

In this section, we examine the finite-sample behavior of several different tests of (1.3) 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 (1.12), Models 1–2 and 7 satisfy (1.17), and Models 1-2 and 7 with $\sigma_1 = 1$ satisfy (1.30) under the null hypothesis.

Treatment status is determined according to Assumption 1.2.2, 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 1.4.1. Note that this ensures that both Assumptions 1.2.3 and 1.2.4 are satisfied. If $\dim(X_i) > 1$, then the pairs are calculated by finding π that minimizes the right-hand side of (1.34) using the `R` package `nbpMatching`. Theorem 1.4.2 ensures that these pairs satisfy Assumption 1.2.3. In order to further ensure that the pairs satisfy Assumption 1.2.4, we re-order the pairs by finding $\bar{\pi}$ that minimizes (1.36) using the same `R` package and applying Theorem 1.4.3.

The results of our simulations are presented in Tables 1.1–1.3 below. Columns are labeled in the following way:

> ***t*-test**: The two-sample t-test studied in Theorem 1.3.1.
>
> **naïve**: The randomization test defined in (1.28) with $T_n(Z^{(n)}) = |\sqrt{n}\hat{\Delta}_n|$ and discussed in Remark 1.3.11. We henceforth refer to this test as the naïve randomization test.
>
> **MP**-t: The "matched pairs" t-test studied in Theorem 1.3.2.
>
> t-**adj**: The "adjusted" t-test studied in Theorem 1.3.3.
>
> **R-adj**: The randomization test studied in Theorem 1.3.5. 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 1.3.8 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 1.3.1, 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 (1.12) 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 1.3.11, 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 (1.17) and are therefore expected to exhibit this phenomenon. Models 1–2 and 7 with $\sigma_1 = 1$ (corresponding to Tables 1.1 and 1.3) in fact satisfy (1.30) under the null hypothesis, so the rejection probability is exactly equal to the nominal level up to simulation error, in agreement with Theorem 1.3.4. 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 1.3.2, 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 (1.17) 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 1.3.3, 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 1.3.5, 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 1.1 and 1.3), which, as mentioned previously, satisfy (1.30) under the null hypothesis, the rejection probability is again exactly equal to the nominal level up to simulation error, in agreement with Theorem 1.3.4. 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 1.1: 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 1.2: 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 1.3: Rej. prob. for Models 1–9 with $\gamma = 1$ for Models 1–6, $\gamma' = (4, 1)$ for Models 7–9, $\sigma_1 = 1$, $\rho = 0$.

CHAPTER 2
OPTIMALITY OF MATCHED-PAIR DESIGNS IN RANDOMIZED CONTROLLED TRIALS

2.1 Introduction

This chapter studies the optimality of matched-pair designs in randomized controlled trials (RCTs). Matched-pair designs are examples of stratified randomization, in which the researcher partitions a set of units into strata based on their observed covariates and assigns a fraction of units in each stratum to treatment. A matched-pair design is a stratified randomization procedure with two units in each stratum. Stratified randomization is prevalent in economics and more broadly the sciences. A simple search with the keyword "stratified" in the AEA RCT Registry reveals about 500 RCTs. The procedures in these papers, however, differ vastly in terms of variables being stratified on, how strata are formed, and numbers of strata. Among these procedures, matched-pair designs have recently gained popularity. 56% of researchers interviewed in Bruhn and McKenzie (2009) have used matched-pair designs at some point in their research. Moreover, more than 40 ongoing experiments in the AEA RCT Registry use matched-pair designs. See Section 2.1.1 for a list of papers. Despite the popularity of matched-pair designs, there is little theory justifying their use in RCTs. We provide an econometric framework in which a certain form of matched-pair design emerges as optimal among all stratified randomization procedures. As will be explained below, an attractive feature of our framework is that it captures a leading motivation for stratifying in the sense that it shows that the proposed matched-pair design minimizes the second moment of the ex-post bias, i.e., the bias of the estimator conditional on realized treatment status. We then provide empirical counterparts to the optimal procedure and illustrate one of the proposed procedures by conducting an actual experiment on the Amazon Mechanical Turk (MTurk). In particular, we replicate one of the treatment arms from the experiment in DellaVigna and Pope (2018) and show that the standard error decreases by 29% compared

to original results, which means that only half of the sample size is required to attain the same level of precision as in the original paper.

We begin by studying settings where treated fractions are identical across strata. In such settings, it is natural to estimate the average treatment effect (ATE) by the difference in means of the treated and control groups. The properties of the difference-in-means estimator, however, vary substantially with stratifications. In the main text, we further restrict treated fractions to be $\frac{1}{2}$ within each stratum, but in the appendix, we provide extensions to settings where treated fractions are identical across strata but not equal to $\frac{1}{2}$ and where they are in addition allowed to vary across a fixed number of subpopulations. Our first result shows the mean-squared error (MSE) of the difference-in-means estimator conditional on the covariates is remarkably minimized by a matched-pair design, where units are ordered by their values of a scalar function of the covariates and paired adjacently. The scalar function is defined by the sum of the expectations of potential outcomes if treated and not treated conditional on the covariates.

We then study the properties of empirical counterparts to this optimal stratification, in which we replace the unknown scalar function with estimates based on pilot data. Pilot experiments are frequently available in practice. Around 350 out of 3000 experiments in the AEA RCT Registry have pilot experiments. For more examples, see Karlan and Zinman (2008), Karlan and Appel (2016), Karlan and Wood (2017), DellaVigna and Pope (2018), and papers cited in Section 2.1.1. We first consider a plug-in procedure that estimates the scalar function using data from a pilot experiment and matches the units in the main experiment into pairs based on their values of the estimated function. Under a weak consistency requirement on the plug-in estimator, we show that as the sample sizes of both the pilot and the main experiments increase, the limiting variance of a suitable normalization of the difference-in-means estimator under the plug-in procedure is the same as that under the infeasible optimal procedure. Equivalently, under such a normalization, the limiting MSE of the estimator is the same as that under the optimal stratification. The consistency require-

ment is satisfied by a large class of nonparametric estimation methods including machine learning methods in high-dimensional settings, i.e., when the dimension of covariates is large. In this sense, when the sample size of the pilot is large, the plug-in procedure is optimal. Of course, this property no longer holds when the sample size of the pilot is small. Furthermore, we may be concerned that a poor estimate of the scalar function leads to a matched-pair design under which the MSE of the estimator is large. Therefore, we additionally consider a penalized procedure under which, according to simulation studies with small pilots, the MSE of the estimator is often smaller than those under plug-in and other commonly-used procedures. The procedure is named so because it can be viewed as penalizing the plug-in procedure by the standard error of the plug-in estimate. Another attractive feature of the penalized procedure is that it is optimal in integrated risk in a Bayesian framework with Gaussian priors and linear conditional expectations of potential outcomes.

For each procedure, we develop methods for testing the null hypothesis that the ATE equals a prespecified value. Each test we provide is asymptotically exact in the sense that the limiting rejection probability under the null equals the nominal level. Our results extend those in Chapter 1 to settings where units are matched according to (random) functions of their covariates instead of the covariates themselves. A special feature of inference under the plug-in procedure is that the same test is valid regardless of the sample size of the pilot. Inference methods under both the plug-in and the penalized procedures are computationally easy.

Our results on optimal stratification formalizes the motivation for using stratified randomization by showing that minimizing the conditional (on covariates) MSE is equivalent to minimizing the conditional second moment of the ex-post bias, i.e., the bias of the estimator conditional on both the covariates and realized treatment status. Furthermore, the two problems are both equivalent to minimizing the conditional variance of the ex-post bias. To illustrate the intuition behind this minimization problem, it is instructive to consider the special case when there is a single binary covariate. Consider an RCT with 100 units,

composed of 50 women and 50 men. The intuitive motivation for stratifying by gender is as follows: if all the units are in one stratum, then it could happen that 40 women are treated while only 10 men are so, so that a large part of the difference in treated and control units could be from the difference in gender instead of the treatment itself; on the other hand, if we stratify by gender, then we always end up treating 25 women and 25 men. The intuitive motivation is formalized by the comparison of the ex-post bias. Since the ex-post bias only depends on how many men and women treated instead of their identities, it varies across realized treatment status if all the units are in one stratum, but is identical if we stratify by gender. As a result, the conditional variance of the ex-post bias is positive if all the units are in one stratum but zero if we stratify by gender. When there are more covariates or when some of them are continuous, it is hard to see only by inspection which stratification minimizes the second moment or the variance of the ex-post bias, but the solution is given by the optimal stratification. Our results could also be viewed as formalizing the discussion about which covariates should be stratified on, e.g., the recommendation in Bruhn and McKenzie (2009) and Glennerster and Takavarasha (2013) for using covariates most correlated with the outcome.

While pilot experiments are common in RCTs, there are scenarios in which they are either not available or are performed on a different population from units in the main experiment. For those scenarios, we study a minimax problem that does not rely on pilot data, where we assume the data generating process is chosen by nature adversarially among a large class of distributions that could be characterized by bounded polyhedrons. In particular, we minimize the variance of the ex-post bias of the difference-in-means estimator conditional on the covariates under the worst possible distribution in this class by choosing across matched-pair designs. The framework accommodates many common shape restrictions on the conditional expectations of potential outcomes given the covariates, including Lipschitz continuity, monotonicity, and convexity. We then rewrite the minimax problem into a mixed integer linear program (MILP) which is computationally easy. Simulation evidence further

suggests although the minimax matched-pair design is in general not minimax-optimal among all stratifications except when the there is a single covariate, it is often close to being so.

The remainder of the chapter is organized as follows. In Section 2.2, we introduce the setup and notation. We study the optimal stratification in Section 2.3. In Section 2.4, we consider empirical counterparts to the optimal stratification, using data from pilot experiments. We consider the plug-in procedure with large pilots and the penalized procedure with small pilots. Section 2.5 includes asymptotic results and methods for inference for ATE. In Section 2.6, we illustrate the properties of different procedures in a small simulation study. Section 2.7 discusses results from the MTurk experiment using the penalized procedure. The experiment shows a 29% reduction in standard error compared to results in the original paper, which means that we need only half of the sample size to attain the same standard error. Section 2.8 briefly discusses the minimax procedure, the details of which are included in Appendix B.5. We conclude with recommendations for empirical practice in Section 2.9.

2.1.1 Related Literature

This chapter is most closely related to Barrios (2013) and Tabord-Meehan (2020). Barrios (2013) considers minimizing the variance of the difference-in-means estimator but assumes a homogeneous treatment effect and uses only information about untreated potential outcomes in his analysis. Despite having "optimal stratification" in the title, his paper only shows that a certain matched-pair design is optimal among all matched-pair designs, instead of all stratifications. We instead show that a certain matched-pair design is optimal among all stratifications, without assuming a homogeneous treatment effect. Moreover, we provide novel results relating the MSE to the ex-post bias. We also provide formal results on the large sample properties of empirical counterparts to the optimal procedure as well as formal results on inference. Tabord-Meehan (2020) considers optimality within a specific class of stratifications, which is a certain class of stratification trees. Since the number of strata is fixed in his asymptotic framework, his paper precludes matched-pair designs. We instead

provide analytical characterization of the optimal one among the set of all stratifications. Remark 2.5.5 elaborates the details of the comparison between the two papers, and in particular, notes that it is straightforward to combine the procedures in both papers. Under the combined procedure, the asymptotic variance of the fully saturated estimator is no greater than and typically strictly smaller than that when using the procedure in Tabord-Meehan (2020) alone.

Recent examples of stratified randomization in development economics include Aker et al. (2012, page 97), Alatas et al. (2012, page 1211), Ashraf et al. (2010, page 2393), Dupas and Robinson (2013, page 168), Callen et al. (2014, page 133), Banerjee et al. (2015, page 31), Duflo et al. (2015a, page 96), Duflo et al. (2015a, footnote 6), Chong et al. (2016, page 228), Berry et al. (2018, page 75), Bursztyn et al. (2018, page 1570), Callen et al. (2018, page 10), Dupas et al. (2018, page 264), Bursztyn et al. (2019, footnote 15), Casaburi and Macchiavello (2019, page 548), Chen and Yang (2019, page 2308), Dizon-Ross (2019, page 2738), Khan et al. (2019, page 254), and Muralidharan et al. (2019, page 1434). See Bruhn and McKenzie (2009) for more examples in economics and Rosenberger and Lachin (2015) and Lin et al. (2015) for examples in clinical trials. For examples of matched-pair designs, see Riach and Rich (2002), Ashraf et al. (2006), Panagopoulos and Green (2008), Angrist and Lavy (2009), Imai et al. (2009), Sondheimer and Green (2010), List and Rasul (2011), White (2013), Bhargava and Manoli (2015), Banerjee et al. (2015), Crépon et al. (2015), Bruhn et al. (2016), Glewwe et al. (2016), Groh and McKenzie (2016), Bertrand and Duflo (2017), Fryer (2017), Fryer et al. (2017), Heard et al. (2017), Fryer (2018), Chapter 1 of this dissertation, and the references therein. See Appendix B.6 for a list of ongoing experiments using matched-pair designs in the AEA RCT Registry. Matched-pair designs are also implemented in leading experimental design packages, including `sampsi_mcc` in `Stata`. Imbens (2011) and Athey and Imbens (2017) discuss the benefits of stratified randomization in a finite sample framework and a simple example with one binary covariate. These two papers, together with Chapter 10 in Imbens and Rubin (2015), recognize the merit of matched-pair designs in terms of

estimation but suggest they come with the cost that the asymptotic variance of the estimator is hard to estimate. Our inference procedure solves this problem and therefore eliminates this cost. Besides Chapter 1 of this dissertation, inference under matched-pair designs has also been studied in Fogarty (2018a) and Fogarty (2018b), who provide conservative estimators for the asymptotic variance, and de Chaisemartin and Ramirez-Cuellar (2019), under a sampling scheme different from that in Chapter 1 of this dissertation and a cluster setting.

For general references on RCTs, see Duflo et al. (2007), Bruhn and McKenzie (2009), Glennerster and Takavarasha (2013), Rosenberger and Lachin (2015), Peters et al. (2016), and the Handbook of Field Experiments, Duflo and Banerjee (2017). For earlier work on the optimal design of experiments under parametric models with block structures, see Cox and Reid (2000), Bailey (2004), and Pukelsheim (2006). A series of papers also examine optimal design in RCTs. Hahn et al. (2011) assume independent random sampling across units, whereas stratified randomization induces dependence within each stratum. Chambaz et al. (2015) adaptively assign treatment status for each new observation based on those of the previous units. Kallus (2018) studies optimal treatment assignment from a minimax perspective and optimizes over treatment assignments rather than stratifications. Freedman (2008) and Lin (2013) compare regression-adjusted estimators and the difference-in-means estimator, assuming all the units are in one stratum. Re-randomization, another commonly-used method to balance covariates, is studied in parametric models in Morgan et al. (2012), Morgan and Rubin (2015), Li et al. (2018), Schultzberg and Johansson (2019), and Johansson et al. (2019). Kasy (2016) considers a Bayesian problem in a parametric model, where both the prior and the distributions of potential outcomes are Gaussian with known parameters, and concludes that researchers should never randomize. On the contrary, Wu (1981), Li (1983), and Hooper (1989), and Chapter 3 of this dissertation show the optimality of certain randomization schemes in minimax frameworks. Carneiro et al. (2019) examine the trade-off between collecting more units and more covariates for each unit when designing an RCT under fixed budget. A growing literature, including Manski (2004), Kitagawa and Tetenov

(2018), and Mbakop and Tabord-Meehan (2018), considers empirical welfare maximization by assigning treatment status. Banerjee et al. (2019) study optimal experiments under a combination of Bayesian and minimax criteria in terms of welfare.

2.2 Setup and Notation

Let Y_i denote the observed outcome of interest for the ith unit, D_i denote the treatment status for the ith unit and $X_i = (X_{i,1}, \ldots, X_{i,p})' \in \mathbf{R}^p$ denote the observed, baseline covariates for the ith unit. Further denote by $Y_i(1)$ the potential outcome of the ith unit if treated and by $Y_i(0)$ if not treated. As usual, the observed outcome is related to the potential outcomes and treatment status by the relationship

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

In addition, we define $W_i = (Y_i, X_i', D_i)'$. For ease of exposition, we assume the sample size is even and denote it by $2n$. We assume that $((Y_i(1), Y_i(0), X_i) : 1 \leq i \leq 2n)$ is an i.i.d. sequence of random vectors with distribution Q. For any random vector indexed by i, A_i, define $A^{(n)} = (A_1, \ldots, A_{2n})'$. Our parameter of interest is the average treatment effect (ATE) under Q:

$$\theta(Q) = E_Q[Y_i(1) - Y_i(0)] \ . \tag{2.1}$$

For ease of exposition, we will at times suppress the dependence of various quantities on Q, e.g., use θ to refer to $\theta(Q)$. In stratified randomization, the first step is to partition the set of units into strata. Formally, we define a stratification $\lambda = \{\lambda_s : 1 \leq s \leq S\}$ as a partition of $\{1, \ldots, 2n\}$, i.e.,

(a) $\lambda_s \bigcap \lambda_{s'} = \varnothing$ for all s and s' such that $1 \leq s \neq s' \leq S$.

(b) $\bigcup_{1 \leq s \leq S} \lambda_s = \{1, \ldots, 2n\}$.

Let Λ_n denote the set of all stratifications of $2n$ units. Many results in the chapter will feature matched-pair designs. Recall that a permutation of $\{1, \ldots, 2n\}$ is a function that maps $\{1, \ldots, 2n\}$ onto itself. Let Π_n denote the group of all permutations of $\{1, \ldots, 2n\}$. A matched-pair design is a stratified randomization with

$$\lambda = \{\{\pi(2s-1), \pi(2s)\} : 1 \leq s \leq n\} \ ,$$

where $\pi \in \Pi_n$. Further define $\Lambda_n^{\text{pair}} \subseteq \Lambda_n$ as the set of all matched-pair designs for $2n$ units.

Define $n_s = |\lambda_s|$ and τ_s as the treated fraction in stratum λ_s. Under stratified randomization, given $X^{(n)}$, λ, and $(\tau_s : 1 \leq s \leq S)$, the treatment assignment scheme is as follows: independently for $1 \leq s \leq S$, uniformly at random choose $n_s\tau_s$ units in λ_s and assign $D_i = 1$ for them, and assign $D_i = 0$ for the other units. The treatment assignment scheme implies that

$$(Y^{(n)}(0), Y^{(n)}(1)) \perp\!\!\!\perp D^{(n)} | X^{(n)} \ . \tag{2.2}$$

It also implies that $n_s\tau_s$ is an integer for $1 \leq s \leq S$. Note that the distribution of $D^{(n)}$ depends on λ. In the remainder of the chapter, we assume the following about the treatment assignment scheme unless indicated otherwise:

Assumption 2.2.1. The treatment assignment scheme satisfies $\tau_s \equiv \frac{1}{2}$.

Assumption 2.2.1 implies that the size of each stratum has to be an even number. Most results below could be extended to settings where $\tau_s \equiv \tau \in (0, 1)$ or where they are in addition allowed to vary across subpopulations. See Appendix B.2 for more details.

We estimate the ATE by the difference in means between the treated and control groups. Formally, for $d \in \{0, 1\}$, define

$$\hat{\mu}_n(d) = \frac{\sum_{1 \leq i \leq 2n} Y_i I\{D_i = d\}}{\sum_{1 \leq i \leq 2n} I\{D_i = d\}} = \frac{1}{n} \sum_{1 \leq i \leq 2n : D_i = d} Y_i \ .$$

The difference-in-means estimator is defined as

$$\hat{\theta}_n = \hat{\mu}_n(1) - \hat{\mu}_n(0) \ . \tag{2.3}$$

The difference-in-means estimator is widely used because it is simple and transparent. Under Assumption 2.2.1, it coincides with the estimator from regressing the outcome on treatment status and strata fixed effects, and the estimator from the fully saturated regression, both of which are also widely used in the analysis of RCTs. See, for example, Duflo et al. (2007), Glennerster and Takavarasha (2013), and Crépon et al. (2015).

2.3 Optimal Stratification

For any stratification $\lambda \in \Lambda_n$, our objective function is the mean-squared error (MSE) of $\hat{\theta}_n$ for θ conditional on $X^{(n)}$ under λ:

$$\text{MSE}(\lambda|X^{(n)}) = E_\lambda[(\hat{\theta}_n - \theta)^2|X^{(n)}] \ . \tag{2.4}$$

Here, the subscript λ of E indicates that the expectation depends on λ, since the distribution of treatment status $D^{(n)}$ depends on λ. We consider minimizing the conditional MSE defined in (2.4) over the set of all stratifications:

$$\min_{\lambda \in \Lambda_n} \ \text{MSE}(\lambda|X^{(n)}) \ . \tag{2.5}$$

The solution will depend on features of the distribution which are generally unknown, and we will consider empirical counterparts to the solution, in which unknown quantities are replaced by estimates using data from pilot experiments, in Section 2.4. By Assumption 2.2.1, other aspects of the stratified randomization procedure, especially the treated fractions, are fixed. Therefore, the stratification that solves (2.5) corresponds to an optimal stratified randomization procedure among all those satisfying Assumption 2.2.1.

In order to describe an important result that leads to the solution to (2.5), we define the ex-ante bias of $\hat{\theta}_n$ for θ conditional on $X^{(n)}$ as

$$\text{Bias}^{\text{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}) = E_\lambda[\hat{\theta}_n|X^{(n)}] - \theta \ , \tag{2.6}$$

and the ex-post bias of $\hat{\theta}_n$ for θ conditional on $X^{(n)}$ and $D^{(n)}$ as

$$\text{Bias}^{\text{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)}) = E_\lambda[\hat{\theta}_n|X^{(n)}, D^{(n)}] - \theta \ . \tag{2.7}$$

Here, ex-ante bias refers to the bias conditional only on covariates, before treatment status is assigned; ex-post bias refers to the bias conditional on both the covariates and treatment status, i.e, after treatment status is assigned. By definition,

$$E_\lambda[\text{Bias}^{\text{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})|X^{(n)}] = \text{Bias}^{\text{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}) \ , \tag{2.8}$$

i.e., the expectation of the ex-post bias over the distribution of treatment status equals the ex-ante bias. Note that by (2.3),

$$\hat{\theta}_n = \frac{1}{n} \sum_{1 \leq i \leq 2n} (Y_i(1)D_i - Y_i(0)(1 - D_i)) \ .$$

Under Assumption 2.2.1,

$$\text{Bias}^{\text{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}) = \frac{1}{2n} \sum_{1 \leq i \leq 2n} (E[Y_i(1)|X_i] - E[Y_i(0)|X_i]) - \theta \ , \tag{2.9}$$

so that ex-ante bias is identical across $\lambda \in \Lambda_n$.

To solve (2.5), we decompose the conditional MSE as follows. First, note that

$$\text{MSE}(\lambda|X^{(n)}) = \text{Bias}^{\text{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)})^2 + \text{Var}_\lambda[\hat{\theta}_n|X^{(n)}] \ . \tag{2.10}$$

Here, Var_λ indicates that the distribution of treatment status depends on λ. By (2.9), the first term on the right-hand side is identical across all $\lambda \in \Lambda_n$. Hence, (2.5) is equivalent to minimizing the second term on the right-hand side of (2.10), which could be further decomposed into

$$\mathrm{Var}_\lambda[\hat{\theta}_n|X^{(n)}] = E_\lambda[\mathrm{Var}[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] + \mathrm{Var}_\lambda[E[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] \ . \tag{2.11}$$

By (2.2), conditional on $X^{(n)}$ and $D^{(n)}$, $(Y_i(0), Y_i(1))$'s are independent across i, so that for any $\lambda \in \Lambda_n$, the first term on the right-hand side of (2.11) equals

$$E_\lambda\left[\frac{1}{n^2}\sum_{1\le i\le 2n}(\mathrm{Var}[Y_i(1)|X_i]D_i + \mathrm{Var}[Y_i(0)|X_i](1-D_i))\middle| X^{(n)}\right] = \frac{1}{2n^2}\sum_{1\le i\le 2n}(\mathrm{Var}[Y_i(1)|X_i] + \mathrm{Var}[Y_i(0)|X_i]) \ , \tag{2.12}$$

which is also identical across all $\lambda \in \Lambda_n$. Here, we use (2.2), the facts that $D_i(1-D_i) = 0$ for $1 \le i \le 2n$, and that $E[D_i|X^{(n)}] = \frac{1}{2}$. Hence, (2.5) is further equivalent to minimizing the second term on the right-hand side of (2.11), which equals

$$\mathrm{Var}_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})|X^{(n)}] \ . \tag{2.13}$$

Furthermore, we have

$$\begin{aligned}
&\mathrm{Var}_\lambda[E[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] \\
&= E_\lambda[(E[\hat{\theta}_n|X^{(n)}, D^{(n)}] - E[\hat{\theta}_n|X^{(n)}])^2|X^{(n)}] \\
&= E_\lambda[(\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)}) - \mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}))^2|X^{(n)}] \\
&= E_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})^2|X^{(n)}] \\
&\quad - 2E_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})\,\mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)})|X^{(n)}] + \mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)})^2 \qquad (2.14) \\
&= E_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})^2|X^{(n)}]
\end{aligned}$$

$$- 2E_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})|X^{(n)}] \,\mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}) + \mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)})^2 \quad (2.15)$$

$$= E_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})^2|X^{(n)}] - 2\,\mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)})^2 + \mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)})^2$$

$$= E_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})^2|X^{(n)}] - \mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)})^2 \ , \quad (2.16)$$

where the first equality follows from definition, the second follows from (2.6) and (2.7), the third equality follows from expanding the square, the fourth equality follows since $\mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)})$ is constant conditional on $X^{(n)}$, and the fifth equality follows from (2.8) . By (2.9), $\mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)})$ is the same across λ, and therefore it follows from (2.10)–(2.16) that (2.5) is equivalent to minimizing the first term in (2.16), i.e., the second moment of the ex-post bias. We summarize the results in the following lemma:

Lemma 2.3.1. *Suppose the treatment assignment scheme satisfies Assumption 2.2.1. Then, the set of solutions to* (2.5) *is the same as the set of solutions to*

$$\min_{\lambda\in\Lambda_n} \ E_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})^2|X^{(n)}] \ , \quad (2.17)$$

and the set of solutions to

$$\min_{\lambda\in\Lambda_n} \ \mathrm{Var}_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})|X^{(n)}] \ . \quad (2.18)$$

Remark 2.3.1. We have shown that minimizing the conditional MSE is equivalent to (2.17), i.e., minimizing the second moment of the ex-post bias, and (2.18), i.e., minimizing the variance of the ex-post bias conditional on the covariates. This equivalence holds since the mean of the ex-post bias is the ex-ante bias, which is the same across stratifications by (2.9). (2.17) is more convenient for intuition, while (2.18) is easier to solve. ■

The following theorem contains our main result on optimal stratification, which shows that (2.5) is solved by a matched-pair design, where units are ordered by their values of a

scalar function of the covariates and paired adjacently. In particular, define the function

$$g(x) = E[Y_i(1) + Y_i(0)|X_i = x] \ . \tag{2.19}$$

For any measurable function $h : \mathbf{R}^p \to \mathbf{R}$, define $h_i = h(X_i)$. Let $\pi^g \in \Pi_n$ be such that $g_{\pi^g(1)} \leq \ldots \leq g_{\pi^g(2n)}$. Define the stratification

$$\lambda^g(X^{(n)}) = \{\{\pi^g(2s-1), \pi^g(2s)\} : 1 \leq s \leq n\} \ . \tag{2.20}$$

Theorem 2.3.1. *Suppose the treatment assignment scheme satisfies Assumption 2.2.1. Then, $\lambda^g(X^{(n)})$ defined in* (2.20) *solves* (2.5).

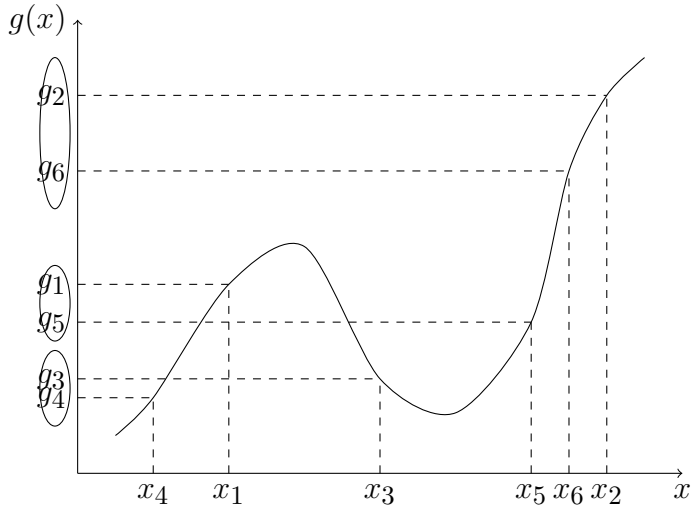


Figure 2.1: Illustration of the optimal stratification defined in (2.20). In the example, $p = 1$, i.e., X_i's are scalars. The optimal stratification is $\{\{3, 4\}, \{1, 5\}, \{2, 6\}\}$.

Remark 2.3.2. Figure 2.3 illustrates the optimal stratification in (2.20). The outline of the proof of Theorem 2.3.1 is as follows. Lemma B.3.1 shows that each stratification is a convex combination of matched-pair designs. Therefore, one of the solutions to (2.5) must be a "vertex" of these convex combinations, i.e., a matched-pair design. Using the second part of Lemma 2.3.1, we show that the conditional MSEs of $\hat{\theta}_n$ under matched-pair designs

differ only in terms of the sum of squared distances in g within pairs. The sum is minimized by the stratification defined in (2.20), according to a variant of the Hardy-Littlewood-Pólya rearrangement inequality for non-bipartite matching. ■

Remark 2.3.3. Note from (2.19) that g_i is a scalar regardless of the dimension p of X_i. Moreover, (2.20) depends not on the values but merely the ordering of $g_i, 1 \leq i \leq 2n$. For instance, if $p = 1$ and we are certain that $g(x)$ is monotonic in x, then it is optimal to order units by $X_i, 1 \leq i \leq n$ and pair the units adjacently, regardless of the values of $g_i, 1 \leq i \leq 2n$.

■

Remark 2.3.4. In a related paper, Barrios (2013) assumes a homogeneous treatment effect and uses only information about untreated potential outcomes in his analysis. Despite having "optimal stratification" in the title, his paper only shows that a certain matched-pair design is optimal among all matched-pair designs, instead of all stratifications. In contrast, Theorem 2.3.1 shows that a certain matched-pair design is optimal among all stratifications, without assuming a homogeneous treatment effect. Moreover, we provide novel results distinguishing the ex-ante or ex-post bias, as well as connecting the ex-post bias to the ex-ante MSE in (2.4). We also provide formal results on the large sample properties of empirical counterparts to the optimal procedure as well as formal results on inference. ■

Remark 2.3.5. Theorem B.2.1 in the appendix examines the scenario where $\tau_s \equiv \tau \in (0, 1)$. Assume $\tau = \frac{l}{k}$ where $l, k \in \mathbb{Z}$, $0 < l < k$, and they are relatively prime, and that the sample size is kn. Define

$$g^{\tau}(X_i) = \frac{E[Y_i(1)|X_i]}{\tau} + \frac{E[Y_i(0)|X_i]}{1-\tau} \ . \tag{2.21}$$

Let $\pi^{\tau,g^{\tau}}$ be a permutation of $\{1, \ldots, kn\}$ such that $g^{\tau}_{\pi^{\tau,g^{\tau}}(1)} \leq \ldots \leq g^{\tau}_{\pi^{\tau,g^{\tau}}(kn)}$. We show that (2.5) is solved by

$$\lambda^{\tau,g}(X^{(n)}) = \{\{\pi^{\tau,g^{\tau}}((s-1)k+1), \ldots \pi^{\tau,g^{\tau}}(sk)\} : 1 \leq s \leq n\} \ , \tag{2.22}$$

The scalar function g^{τ} adjusts for treatment probabilities by inverse probability weighting.

For a similar design, see Bold et al. (2018). ■

We illustrate Lemma 2.3.1, and in particular (2.17), in a small simulation study. In this example, $2n = 100$; $X_i = (X_{i,1}, X_{i,2})'$; $X_{i,1}$ and $X_{i,2}$ are both distributed as $N(0,1)$, independent from each other, and i.i.d. across $1 \leq i \leq 2n$; and $E[Y_i(d)|X_i] = X_i'\beta(d)$ for $\beta(0) = (0, 1.5)'$ and $\beta(1) = (0.5, 2)'$. As a result, $\theta = 0$. In Figure 2.2, we plot the densities of the distributions of $\text{Bias}^{\text{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})$ defined in (2.7) over 1000 draws of $X^{(n)}$ and $D^{(n)}$, for different treatment assignment schemes:

Oracle stratified randomization using the infeasible optimal procedure defined by (2.20).

by1 stratified randomization with two strata separated by the sample median of $X_{i,1}$.

by2 stratified randomization with two strata separated by the sample median of $X_{i,2}$.

SRS Simple Random Sampling, i.e., $(D_i, 1 \leq i \leq 2n)$ are i.i.d. Bernoulli($\frac{1}{2}$).

Note that the distribution of $\text{Bias}^{\text{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})$ under **Oracle** is much more concentrated than those under other treatment assignment schemes.

2.4 Empirical Counterparts

The optimal procedure in (2.20) depends on the function g defined in (2.19), which needs to estimated in practice. Fortunately, pilot experiments are common in RCTs, and we could use data from pilot experiments to estimate g. In this section, we consider empirical counterparts to the optimal procedure defined by (2.20), when there is a pilot experiment. We describe the procedures in this section and comment on their asymptotic properties, formally introducing asymptotic results in Section 2.5. For any random vector A, we denote by $\tilde{A}_j$ the corresponding random vector of the jth unit in the pilot experiment. Suppose $\tilde{W}^{(m)} = ((\tilde{Y}_j, \tilde{X}_j', \tilde{D}_j)' : 1 \leq j \leq m)$ comes from the pilot experiment. We assume that

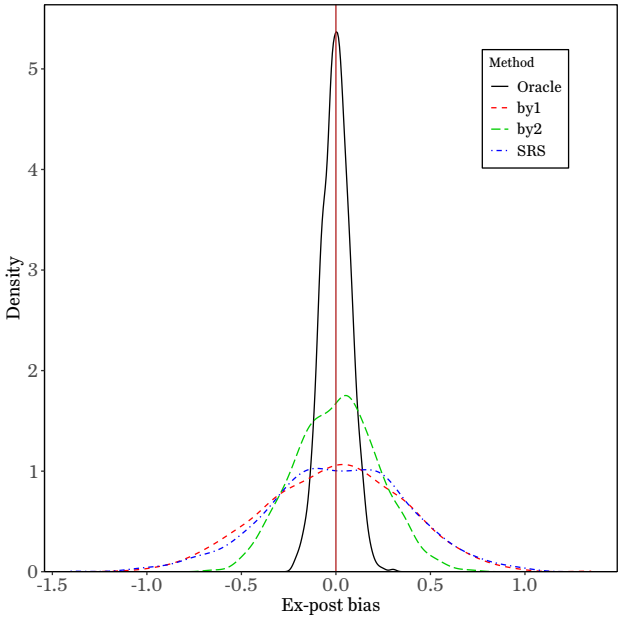


Figure 2.2: Densities of the distributions of the $\text{Bias}^{\text{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})$ over 1000 draws of $X^{(n)}$ and $D^{(n)}$ under all treatment assignment schemes.

$((\tilde{Y}_j(1), \tilde{Y}_j(0), \tilde{X}_j) : 1 \leq j \leq m)$ is an i.i.d. sequence of random vectors with distribution Q, i.e., the units in the pilot are drawn from the same population as the units in the main experiment.

We first consider a plug-in procedure. Suppose $\hat{g}_m$ is an estimator of g defined in (2.19). Concretely, $\hat{g}_m$ is a random function from $\mathbf{R}^p$ to $\mathbf{R}$ that depends on $\tilde{W}^{(m)}$. We will abstract away from how $\hat{g}_m$ is obtained but directly impose conditions on $\hat{g}_m$ itself. Recall Π_n is the set of all permutations of $\{1, \ldots, 2n\}$ and let $\pi^{\hat{g}_m} \in \Pi_n$ be such that $\hat{g}_{m,\pi^{\hat{g}_m}(1)} \leq \ldots \leq \hat{g}_{m,\pi^{\hat{g}_m}(2n)}$. We define the following plug-in stratification for the main experiment:

$$\lambda^{\hat{g}_m}(X^{(n)}) = \{\{\pi^{\hat{g}_m}(2s-1), \pi^{\hat{g}_m}(2s)\} : 1 \leq s \leq n\} \ . \tag{2.23}$$

As Theorem 2.5.1 shows, the plug-in procedure enjoys the property that as the sample size of the pilot increases, the asymptotic variance of $\hat{\theta}_n$ in (2.3) is that same as that under the optimal procedure defined by (2.20). The key condition for the property is that $\hat{g}_m$ is consistent for g in a certain cense. See Assumption 2.5.3 below for more details. The

assumption is satisfied by a large class of nonparametric estimation methods, including machine learning methods in high-dimensional settings, i.e., when the dimension of the covariates is large.

When the sample size of the pilot is small, the plug-in procedure generally does not have the efficiency property as in settings with large pilot. Indeed, we may be concerned that the plug-in estimator $\hat{g}_m$ is a poor approximation for g in (2.19), and as a result, that under the plug-in stratification defined in (2.23), the conditional MSE and the asymptotic variance of $\hat{\theta}_n$ is large. Therefore, we consider a penalized procedure under which, according to simulation studies in Section 2.6, the conditional MSE of $\hat{\theta}_n$ is often smaller than that under the stratification defined in (2.23). The procedure is named so because it can be viewed as penalizing the plug-in procedure by the standard error of the plug-in estimate. To describe the procedure, for $d \in \{0, 1\}$, define the least-square estimators based on the treated or control units as

$$\hat{\beta}_m(d) = \left(\sum_{1 \le j \le m : \tilde{D}_j = d} \tilde{X}_j \tilde{X}_j' \right)^{-1} \sum_{1 \le j \le m : \tilde{D}_j = d} \tilde{X}_j \tilde{Y}_j \ , \tag{2.24}$$

and the variance estimators assuming homoskedasticity as

$$\hat{\Sigma}_m(d) = \hat{\nu}_m^2(d) \left(\sum_{1 \le j \le m : \tilde{D}_j = d} \tilde{X}_j \tilde{X}_j' \right)^{-1} \ , \tag{2.25}$$

where

$$\hat{\nu}_m^2(d) = \frac{\sum_{1 \le j \le m} (\tilde{Y}_j - \tilde{X}_j' \hat{\beta}_m(d))^2 I\{\tilde{D}_j = d\}}{\sum_{1 \le j \le m} I\{\tilde{D}_j = d\}} \ .$$

Further define

$$\hat{\beta}_m = \hat{\beta}_m(1) + \hat{\beta}_m(0) \tag{2.26}$$

$$\hat{\Sigma}_m = \hat{\Sigma}_m(1) + \hat{\Sigma}_m(0) \ . \tag{2.27}$$

Next, we define R_m as the result of the following Cholesky decomposition:

$$R_m' R_m = \hat{\beta}_m \hat{\beta}_m' + \hat{\Sigma}_m \ , \tag{2.28}$$

and the following transformation of the covariates:

$$Z_i = R_m X_i \ . \tag{2.29}$$

The penalized stratification matches units to minimize the sum of distances in terms of Z_i within pairs. Compared with $\hat{g}_m(X_i)$, the main difference is that Z_i is a vector of the same dimension p of X_i, instead of a scalar. Let π^{pen} denote the solution to the following problem:

$$\min_{\pi \in \Pi_n} \frac{1}{n} \sum_{1 \le s \le n} \| Z_{\pi(2s-1)} - Z_{\pi(2s)} \| \ . \tag{2.30}$$

When the dimension p of X_i is not too large, the problem could be solved quickly by the package `nbpMatching` in `R`. Finally, define the penalized stratification as

$$\lambda^{\text{pen}}(X^{(n)}) = \{ \{ \pi^{\text{pen}}(2s-1), \pi^{\text{pen}}(2s) \} : 1 \le s \le n \} \ . \tag{2.31}$$

(2.31) can be viewed as penalizing the plug-in procedure in (2.23) by the variance of the plug-in estimator.

We now briefly explain the intuition behind (2.30). For simplicity, suppose $E[Y_i(d)|X_i] = X_i' \beta(d)$ for $d \in \{0, 1\}$. In addition, define $\beta = \beta(1) + \beta(0)$. (2.30) penalizes the the plug-in stratification by the standard error of the plug-in estimate. Indeed, the objective in (2.30) equals

$$\frac{1}{n} \sum_{1 \le s \le n} \hat{d}^{\frac{1}{2}}(X_{\pi(2s-1)}, X_{\pi(2s)}) \ ,$$

where for any $x_1, x_2 \in \mathbf{R}^p$,

$$\hat{d}(x_1, x_2) = (x_1'\hat{\beta}_m - x_2'\hat{\beta}_m)^2 + (x_1 - x_2)'\hat{\Sigma}_m(x_1 - x_2) \ . \tag{2.32}$$

If $\hat{\Sigma}_m = 0$, then (2.30) is solved by $\pi^{\hat{g}_m}$ in the plug-in stratification in (2.23) with $\hat{g}_m = X_i'\hat{\beta}_m$. If on the other hand $\hat{\Sigma}_m$ is large, which means that $\hat{\beta}_m$ is a very noisy estimate for β, then the second term in (2.32) dominates, and $\hat{g}_m$ contributes little to the solution to (2.30).

Remark 2.4.1. We now provide a further justification for (2.31) by discussing its optimality in a Bayesian framework. To begin with, note that the problem in (2.30) could also be defined with the squared norm $\|Z_{\pi(2s-1)} - Z_{\pi(2s)}\|^2$, and the two definitions are asymptotically equivalent. For more details, see Section 1.4. This asymptotically equivalent formulation is in fact optimal in the sense that it minimizes the integrated risk in a Bayesian framework with a diffuse normal prior, where the conditional expectations of potential outcomes are linear. With some abuse of notation, denote the conditional MSE in (2.4) by $\mathrm{MSE}(\lambda|g, X^{(n)})$, where we make explicit the dependence on g. Suppose we have a prior distribution of g, denoted by $F(dg)$, which is normal. Let $Q_X^n(dx^{(n)})$ denote the distribution of $X^{(n)}$ and $Q_{\tilde{W}}^m(d\tilde{w}^{(m)})$ denote the distribution of $\tilde{W}^{(m)}$. Consider the solution to following problem of minimizing the integrated risk across all measurable functions of the form $u : (\tilde{w}^{(m)}, x^{(n)}) \mapsto \lambda \in \Lambda_n$:

$$\min_u \iiint \mathrm{MSE}(u(\tilde{w}^{(m)}, x^{(n)})|g, x^{(n)}) Q_X^n(dx^{(n)}) Q_{\tilde{W}}^m(d\tilde{w}^{(m)}) F(dg) \ . \tag{2.33}$$

In Appendix B.4, we first show that the problem in (2.33) under any prior F is solved by a matched-pair design. To the best of our knowledge, this is the first result showing that matched-pair designs are optimal in general Bayesian frameworks. Next, we specialize the model by assuming $E[Y_i(d)|X_i] = X_i'\beta(d)$, define $\beta = \beta(1) + \beta(0)$, and show that F could be equivalently expressed as a distribution on β, which we further assume to be normal. One may be tempted to conjecture that the solution to (2.33) is to näively match units on the the value of $X_i'\bar{\beta}$, where $\bar{\beta}$ is posterior mean of β, i.e., $\hat{\beta}_m$ in (2.26) shrunk towards

the prior mean. We show, however, that the solution to (2.33) depends not only on the posterior mean of β, but also on the posterior variance of it. The posterior variance serves as a penalty to matching naively on the posterior mean of β: the larger the variance, the more it penalizes matching on the posterior mean. In the end, we show that when F diverges to the diffuse prior, the posterior mean converges to the OLS estimate, and the posterior variance converges to the variance estimate from OLS. As a result, the solution to (2.33) converges to the procedure defined by (2.30) with the squared norm $\|Z_{\pi(2s-1)} - Z_{\pi(2s)}\|^2$.

■

2.5 Asymptotic Results and Inference

Under matched-pair designs, it is challenging to derive asymptotic properties of the difference-in-means estimator and conduct inference for ATE, because of the heavy dependence of treatment status across units. Even if g in (2.19) is known, commonly-used inference procedures under matched-pair designs, including the two-sample t-test and the "matched pairs" t-test, are conservative in the sense that the limiting rejection probability under the null is equal to the nominal level. The issue is further complicated since g needs to be estimated, so that the stratifications in (2.23) and (2.31) depend on data from the pilot experiment. Extending results from Chapter 1 of this dissertation, we develop novel results of independent interest on the limiting behavior of the difference-in-means estimator under procedures involving a large number of strata, when the stratifications depend on data from the pilot experiment. These results enable us to establish the desired property of our proposed inference procedures. To begin with, we make the following mild moment restriction on the distributions of potential outcomes:

Assumption 2.5.1. $E[Y_i^2(d)] < \infty$ for $d \in \{0, 1\}$.

2.5.1 Asymptotic Results for Plug-in with Large Pilot

In this subsection, we study the properties of $\hat{\theta}_n$ defined in (2.3) under settings where the sample sizes of both the pilot and the main experiments increase. We henceforth refer to such a setting as an experiment with a large pilot. We first impose the following assumption on g defined in (2.19).

Assumption 2.5.2. The function g satisfies

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

(b) $\mathrm{Var}[Y_i(d)|g(X_i) = z]$ is Lipschitz in z.

(c) $E[g^2(X_i)] < \infty$.

Assumption 2.5.2(a)–(c) are conditions imposed on the target function g instead of the plug-in estimator $\hat{g}_m$. Assumption 2.5.2(a) is a mild restriction to rule out degenerate situations and to permit the application of suitable laws of large numbers and central limit theorems. Assumption 2.5.2(c) is another mild moment restriction to ensure the pairs are "close" in the limit. New sufficient conditions for Assumption 2.5.2(b) are provided in Appendix B.3.1. The results therein about the conditional expectation of a random variable given a manifold are new and may be of independent interest.

We additionally impose the following restriction on the estimator $\hat{g}_m$. In what follows, we use Q_X to denote the marginal distribution of X_i under Q.

Assumption 2.5.3. The sequence of estimators $\{\hat{g}_m\}$ satisfies

$$\int_{\mathbf{R}^p} |\hat{g}_m(x) - g(x)|^2 Q_X(dx) \xrightarrow{P} 0$$

as $m \to \infty$.

Assumption 2.5.3 is commonly referred to as the L^2-consistency of the $\hat{g}_m$ for g. When p is fixed and suitable smoothness conditions hold, L^2-consistency is satisfied by series and sieves

estimators (Newey, 1997; Chen, 2007) and kernel estimators (Li and Racine, 2007). In high-dimensional settings, when p increases with n at suitable rates, it is satisfied by the LASSO estimator (Bühlmann and Van De Geer, 2011; Belloni et al., 2012, 2014; Chatterjee, 2013; Bellec et al., 2018), regression trees and random forests (Györfi et al., 2006; Biau, 2012; Denil et al., 2014; Scornet et al., 2015; Wager and Walther, 2015), neural nets (White, 1990; Chen and White, 1999; Chen, 2007; Farrell et al., 2018), and support vector machines (Steinwart and Christmann, 2008). The results therein are either exactly as stated in Assumption 2.5.3 or one of the following:

(a) $\sup_{x \in \mathbf{R}^p} |\hat{g}_m(x) - g(x)| \xrightarrow{P} 0$ as $m \to \infty$.

(b) $E[|\hat{g}_m(x) - g(x)|^2] \to 0$ as $m \to \infty$.

It is straightforward to see (a) implies Assumption 2.5.3. (b) also implies Assumption 2.5.3 by Markov's inequality.

The next theorem reveals that under L^2-consistency of the estimator $\hat{g}_m$, the asymptotic variance of $\hat{\theta}_n$ under the plug-in procedure is the same with that under the infeasible optimal procedure defined by (2.20).

Theorem 2.5.1. *Suppose the treatment assignment scheme satisfies Assumption 2.2.1, Q satisfies Assumption 2.5.1, g satisfies Assumption 2.5.2. Then, under $\lambda^g(X^{(n)})$, as $n \to \infty$,*

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

where

$$\varsigma_g^2 = \mathrm{Var}[Y_i(1)] + \mathrm{Var}[Y_i(0)] - \frac{1}{2} E[(g(X_i) - E[Y_i(1) + Y_i(0)])^2] \ . \tag{2.34}$$

In addition, suppose $\hat{g}_m$ satisfies Assumption 2.5.3. Then, under $\lambda^{\hat{g}_m}(X^{(n)})$ defined in (2.23), as $m, n \to \infty$,

$$\sqrt{n}(\hat{\theta}_n - \theta(Q)) \xrightarrow{d} N(0, \varsigma_g^2) \ .$$

2.5.2 *Inference under Plug-In Procedure*

Next, we consider inference for the ATE. For any prespecified $\theta_0 \in \mathbf{R}$, we are interested in testing

$$H_0 : \theta(Q) = \theta_0 \text{ versus } H_1 : \theta(Q) \neq \theta_0 \tag{2.35}$$

at level $\alpha \in (0, 1)$. In order to do so, for $d \in \{0, 1\}$, define

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

Define

$$\hat{\rho}_n = \frac{2}{n} \sum_{1 \leq j \leq \lfloor \frac{n}{2} \rfloor} (Y_{\pi^{\hat{g}_m}(4j-3)} + Y_{\pi^{\hat{g}_m}(4j-2)})(Y_{\pi^{\hat{g}_m}(4j-1)} + Y_{\pi^{\hat{g}_m}(4j)}) \tag{2.36}$$

and define $\hat{\varsigma}_n^{\hat{g}_m}$ such that

$$(\hat{\varsigma}_n^{\hat{g}_m})^2 = \hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0) - \frac{1}{2}\hat{\rho}_n + \frac{1}{2}(\hat{\mu}_n(1) + \hat{\mu}_n(0))^2 \ . \tag{2.37}$$

The test is

$$\phi_n^{\hat{g}_m}(W^{(n)}) = I\{|T_n^{\hat{g}_m}(W^{(n)})| > \Phi^{-1}(1 - \frac{\alpha}{2})\} \ , \tag{2.38}$$

where

$$T_n^{\hat{g}_m}(W^{(n)}) = \frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\hat{\varsigma}_n^{\hat{g}_m}} \ , \tag{2.39}$$

and $\Phi^{-1}(1-\frac{\alpha}{2})$ denotes the $(1-\frac{\alpha}{2})$-th quantile of the standard normal distribution. Although the right-hand side of (2.37) is possibly negative, its limit in probability must be positive under assumptions imposed below. For ease of exposition, we assume all quantities like (2.37) are positive for the rest of the chapter.

We start by studying the limiting behavior of the test defined in (2.38) with a large pilot. The following theorem shows that the test defined in (2.38) is asymptotically exact in the sense that when the sample sizes of both the pilot and the main experiments increase, the

limiting rejection probability is equal to the nominal level.

Theorem 2.5.2. *Suppose the treatment assignment scheme satisfies Assumption 2.2.1, Q satisfies Assumption 2.5.1, g satisfies Assumption 2.5.2, and $\hat{g}_m$ satisfies Assumption 2.5.3. Then, under $\lambda^{\hat{g}_m}(X^{(n)})$ defined in* (2.23)*, as $m, n \to \infty$,*

$$(\hat{\varsigma}_n^{\hat{g}_m})^2 \xrightarrow{P} \varsigma_g^2 \ .$$

Thus, for the problem of testing (2.35) *at level $\alpha \in (0,1)$, $\phi_n^{\hat{g}_m}(W^{(n)})$ defined in* (2.38) *satisfies*

$$\lim_{m,n\to\infty} E[\phi_n^{\hat{g}_m}(W^{(n)})] = \alpha \ ,$$

when Q additionally satisfies the null hypothesis, i.e., $\theta(Q) = \theta_0$.

Remark 2.5.1. The studentization by (2.37) is crucial for the asymptotic exactness of (2.38). Commonly-used tests including the two-sample t-test (Riach and Rich, 2002; Gelman and Hill, 2006; Duflo et al., 2007) and the "matched pairs" t-test (Moses, 2006; Hsu and Lachenbruch, 2007; Armitage et al., 2008; Athey and Imbens, 2017) are asymptotically conservative in the sense that the limiting rejection probabilities under the null are no greater than and typically strictly less than the nominal level. See Chapter 1 of this dissertation for more details. ■

Remark 2.5.2. In order for $\hat{g}_m$ to satisfy Assumption 2.5.3, the following selection on observables condition is usually required on the pilot experiment:

$$(\tilde{Y}^{(m)}(1), \tilde{Y}^{(m)}(0)) \perp\!\!\!\perp \tilde{D}^{(m)} | \tilde{X}^{(m)} \ ,$$

The condition is satisfied by a large class of treatment assignment schemes, including simple random sampling, covariate-adaptive randomization, re-randomization, etc. For more details, see Bugni et al. (2018) and Chapter 1 of this dissertation. ■

Next, we consider settings where the sample size of the main experiment increases while that of the pilot experiment is allowed to be fixed. We henceforth refer to such a setting as an experiment with a small pilot. We show that test defined in (2.38) is again asymptotically exact in the sense that the limiting rejection probability under the null is equal to the nominal level when the sample size of the main experiment increases, regardless of the sample size of the pilot. The restrictions that we put on $\hat{g}_m$, however, are more likely to be satisfied when $\hat{g}_m$ is constructed using simple methods such as least squares. We impose the following restriction in addition to Assumption 2.5.1:

Assumption 2.5.4. The estimator $\hat{g}_m$ satisfies

$$Q\{\hat{g}_m \in \mathbf{H}\} = 1 \ ,$$

where $\mathbf{H}$ is the set of all measurable functions $h : \mathbf{R}^p \to \mathbf{R}$ such that

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

(b) $E[Y_i^r(d)|h(X_i) = z]$ is Lipschitz in z for $r = 1, 2$ and $d = 0, 1$.

(c) $E[h^2(X_i)] < \infty$.

Assumption 2.5.4 is imposed on the distributions of potential outcomes conditional on $\hat{g}_m$, where $\hat{g}_m$ is viewed as a fixed function given data from the pilot experiment. In fact, with small pilots, Assumption 2.5.4 contains the same set of conditions as those in Assumption 2.5.2, the only difference being that they are imposed on $\hat{g}_m$ instead of g. In the definition of $\mathbf{H}$, (a) is a mild restriction to rule out degenerate situations and to permit the application of suitable laws of large numbers and central limit theorems, and (c) is another mild moment restriction to ensure the pairs are "close" in the limit. New sufficient conditions for (b) are provided in Appendix B.3.1. Note, in particular, that (b) is more likely to be satisfied when $\hat{g}_m$ is constructed using simple estimation methods such as least squares.

The following theorem shows that the test defined in (2.38) is asymptotically exact in the sense that as the sample size of the main experiment increases, the limiting rejection probability under the null is equal to the nominal level. Note, in particular, that the sample size of the pilot is allowed to be fixed.

Theorem 2.5.3. *Suppose the treatment assignment scheme satisfies Assumption 2.2.1, Q satisfies Assumption 2.5.1, and $\hat{g}_m$ satisfies Assumption 2.5.4. Suppose Q additionally satisfies the null hypothesis, i.e., $\theta(Q) = \theta_0$. Then, under $\lambda^{\hat{g}_m}(X^{(n)})$ defined in* (2.23)*, for the problem of testing* (2.35) *at level $\alpha \in (0,1)$, $\phi_n^{\hat{g}_m}(W^{(n)})$ defined in* (2.38) *satisfies*

$$\lim_{n\to\infty} E[\phi_n^{\hat{g}_m}(W^{(n)})] = \alpha \ .$$

Remark 2.5.3. Note that we use the same test $\phi_n^{\hat{g}_m}$ with large (Theorem 2.5.2) and small (Theorem 2.5.3) pilots, and it is asymptotically exact either way. When m increases at a rate such that Assumption 2.5.3 is satisfied, the asymptotic variance of $\hat{\theta}_n$ as $m, n \to \infty$ is ς_g^2, which equals the asymptotic variance under the infeasible optimal procedure defined by (2.20). Yet when m is fixed, the asymptotic variance of $\hat{\theta}_n$ as $n \to \infty$ is generally larger than ς_g^2. Moreover, as previously commented, the assumptions in the two settings are non-nested. Assumption 2.5.4 is more likely to be satisfied when the plug-in estimator $\hat{g}_m$ is constructed using simple estimation methods, but does not require $\hat{g}_m$ to be consistent for g in any sense. On the other hand, Assumptions 2.5.2 and Assumption 2.5.3 could potentially allow for more complicated estimation methods but require $\hat{g}_m$ to be L^2-consistent for g. ■

Remark 2.5.4. In fact, the asymptotic exactness of $\phi_n^{\hat{g}_m}(W^{(n)})$ holds conditional on data from the pilot experiment, i.e.,

$$\lim_{n\to\infty} E[\phi_n^{\hat{g}_m}(W^{(n)})|\tilde{W}^{(m)}] = \alpha \tag{2.40}$$

with probability one for $\tilde{W}^{(m)}$. See the proof of Theorem 2.5.3 in the appendix for more

details. Furthermore, it follows from the proof that the test is also asymptotically exact under any procedure defined by (2.23) with $\hat{g}_m$ replaced by a fixed function $h \in \mathbf{H}$, for $\mathbf{H}$ defined in Assumption 2.5.4. In addition, it is possible to show that the asymptotic variance of $\hat{\theta}_n$ under the plug-in procedure or any procedure just mentioned defined by a fixed function is no greater than and typically strictly less than that under procedures with $\lambda = \{\{1, \ldots, 2n\}\}$, i.e., when all the units are in one stratum, or that under simple random sampling, i.e., when treatment status is determined by i.i.d. coin flips. See Lemma B.3.4 for more details. ■

Remark 2.5.5. Sometimes political or logistical considerations or estimation of subpopulation treatment effects require researchers to prespecify different treated fractions across subpopulations. In those settings, as discussed in Appendix B.2, $\hat{\theta}_n$ is no longer consistent for θ in (2.1). Instead, it is natural to use the estimator from the fully saturated regression with all interaction terms of treatment status and strata indicators, i.e., $\hat{\theta}_n^{\text{sat}}$ defined in (B.16). Appendix B.2 discusses straightforward extensions of the optimality result in Theorem 2.3.1 and empirical counterparts including that in (2.23). These results are closely related to Tabord-Meehan (2020), who considers stratification trees which lead to a small number of large strata. In particular, Remark B.2.1 discusses a way to combine his procedure and procedures in this chapter, under which the asymptotic variance of $\hat{\theta}_n^{\text{sat}}$ is no greater than and typically strictly less than that under his procedure alone. ■

Remark 2.5.6. Under SRS, as long as Assumption 2.5.1 holds, $\sqrt{n}(\hat{\theta}_n - \theta(Q)) \xrightarrow{d} N(0, \varsigma^2_{\text{srs}})$ as $n \to \infty$, where

$$\varsigma^2_{\text{srs}} = \text{Var}[Y_i(1)] + \text{Var}[Y_i(0)] \ . \tag{2.41}$$

By (B.4), for any fixed $\tilde{W}^{(m)}$, the asymptotic variance of $\hat{\theta}_n$ under $\lambda^{\hat{g}_m}(X^{(n)})$ defined in (2.23) is no greater than ς^2_{srs}, while being strictly smaller unless $E[Y_i(1)+Y_i(0)|\hat{g}_m(X_i), \tilde{W}^{(m)}]$ $= E[Y_i(1)+Y_i(0)]$. The condition is equivalent to the two-sample t-test being asymptotically exact. For more details, see Chapter 1 of this dissertation. ■

2.5.3 Inference under Penalized Procedure

We now consider inference under the penalized procedure defined by (2.31) with a small pilot. This subsection follows closely the exposition in Section 1.4. Since in general Z defined in (2.29) is not a scalar, the correction term in (2.36) could no longer be defined as before since it relies on $\pi^{\hat{g}_m}$, where $\hat{g}_m$ is a scalar. Instead, we need to match the pairs to ensure that the two pairs matched are close in terms of Z. Define

$$\bar{Z}_s = \frac{Z_{\pi^{\text{pen}}(2s-1)} + Z_{\pi^{\text{pen}}(2s)}}{2} ,$$

and $\bar{\pi}$ as the solution of the following problem:

$$\min_{\pi \in \Pi_n} \frac{1}{n} \sum_{1 \le j \le \lfloor \frac{n}{2} \rfloor} \|\bar{Z}_{\pi(2j-1)} - \bar{Z}_{\pi(2j)}\| .$$

Let $\tilde{\pi}^{\text{pen}} \in \Pi_n$ be such that for $1 \le s \le n$,

$$\tilde{\pi}^{\text{pen}}(2s-1) = \pi^{\text{pen}}(2\bar{\pi}(s)-1) \text{ and } \tilde{\pi}^{\text{pen}}(2s) = \pi^{\text{pen}}(2\bar{\pi}(s)) .$$

In other words, $\tilde{\pi}^{\text{pen}}$ matches the pairs defined by π^{pen} based on the midpoints of pairs. Since $\tilde{\pi}^{\text{pen}}$ rearranges π^{pen} in (2.31) while preserving the units in each stratum, it follows that for $\lambda^{\text{pen}}(X^{(n)})$ defined in (2.31), we have

$$\lambda^{\text{pen}}(X^{(n)}) = \{\{\tilde{\pi}^{\text{pen}}(2s-1), \tilde{\pi}^{\text{pen}}(2s)\} : 1 \le s \le n\} .$$

We then define the test similarly to (2.38), with $\pi^{\hat{g}_m}$ replaced by $\tilde{\pi}^{\text{pen}}$. In particular, define

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

and let $\hat{\varsigma}_n^{\text{pen}}$ be such that

$$(\hat{\varsigma}_n^{\text{pen}})^2 = \hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0) - \frac{1}{2}\hat{\rho}_n^{\text{pen}} + \frac{1}{2}(\hat{\mu}_n(1) + \hat{\mu}_n(0))^2 \ .$$

The test is

$$\phi_n^{\text{pen}}(W^{(n)}) = I\{|T_n^{\text{pen}}(W^{(n)})| > \Phi^{-1}(1 - \frac{\alpha}{2})\} \ , \tag{2.42}$$

where

$$T_n^{\text{pen}}(W^{(n)}) = \frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\hat{\varsigma}_n^{\text{pen}}} \ , \tag{2.43}$$

and $\Phi^{-1}(1 - \frac{\alpha}{2})$ denotes the $(1 - \frac{\alpha}{2})$-th quantile of the standard normal distribution.

Under the penalized procedure, we impose the following assumption on Q:

Assumption 2.5.5. (a) $0 < E[\text{Var}[Y_i(d)|R_m X_i]]$ for $d \in \{0, 1\}$.

(b) $E[Y_i^r(d)|R_m X_i = z]$ is Lipschitz in z for $r \in \{1, 2\}$ and $d \in \{0, 1\}$.

(c) The support of $R_m X_i$ is compact.

Assumption 2.5.5(a)–(b) are the counterparts to Assumption 2.1(a) and (c) of Chapter 1 of this dissertation. Assumption 2.5.5(c) is also imposed in Section 1.4. The following theorem establishes the asymptotic exactness of the test defined in (2.42), in the sense that the limiting rejection probability under the null equals the nominal level. Note, in particular, that the sample size of the pilot is allowed to be fixed.

Theorem 2.5.4. *Suppose the treatment assignment scheme satisfies Assumption 2.2.1 and Q satisfies Assumptions 2.5.1 and 2.5.5. Suppose Q additionally satisfies the null hypothesis, i.e., $\theta(Q) = \theta_0$. Then, under $\lambda^{\text{pen}}(X^{(n)})$ defined in* (2.31)*, for the problem of testing* (2.35) *at level $\alpha \in (0, 1)$, $\phi_n^{\text{pen}}(W^{(n)})$ defined in* (2.38) *satisfies*

$$\lim_{n\to\infty} E[\phi_n^{\text{pen}}(W^{(n)})] = \alpha \ .$$

Remark 2.5.7. In some setups, it may be possible to improve the estimator $\hat{g}_m$ by imposing shape restrictions on g. See, for instance, Chernozhukov et al. (2015) and Chetverikov et al. (2018). ■

2.5.4 Inference with Pooled Data

So far we have disregarded data from the pilot experiment in the test defined in (2.38) except when computing $\hat{g}_m$. We end this section by describing a test that combines data from the pilot and the main experiments. Define

$$\tilde{\theta}_m = \tilde{\mu}_m(1) - \tilde{\mu}_m(0) \ ,$$

where

$$\tilde{\mu}_m(d) = \frac{\sum_{1 \le j \le m} \tilde{Y}_j I\{\tilde{D}_j = d\}}{\sum_{1 \le j \le m} I\{\tilde{D}_j = d\}}$$

for $d \in \{0, 1\}$. We define the new estimator for $\theta(Q)$ as

$$\hat{\theta}_n^{\text{combined}} = \frac{m}{m + 2n} \tilde{\theta}_m + \frac{2n}{2n + m} \hat{\theta}_n \ .$$

We define the test as

$$\phi_n^{\text{combined}}(W^{(n)}, \tilde{W}^{(m)}) = I\{|T_n^{\text{combined}}(W^{(n)}, \tilde{W}^{(m)})| > \Phi^{-1}(1 - \frac{\alpha}{2})\} \ , \tag{2.44}$$

where

$$T_n^{\text{combined}}(W^{(n)}, \tilde{W}^{(m)}) = \frac{\sqrt{m + 2n}(\hat{\theta}_n^{\text{combined}} - \theta_0)}{\sqrt{\frac{m}{m+2n} \tilde{\varsigma}^2_{\text{pilot},m} + \frac{2n}{m+2n} 2(\hat{\varsigma}_n^{\hat{g}_m})^2}} \ , \tag{2.45}$$

and $\Phi^{-1}(1 - \frac{\alpha}{2})$ denotes the $(1 - \frac{\alpha}{2})$-th quantile of the standard normal distribution.

The following theorem shows that the test defined in (2.44) is asymptotically exact as the sample sizes of both the pilot and the main experiments increase. The main additional requirement is that as $m \to \infty$, $\sqrt{m}(\tilde{\theta}_m - \theta(Q))$ converges in distribution to a normal

distribution whose variance is consistently estimable. The assumption is satisfied by many treatment assignment schemes, including simple random sampling and covariate-adaptive randomization. See Bugni et al. (2018) and Bugni et al. (2019) for more details.

Theorem 2.5.5. *Suppose the treatment assignment scheme satisfies Assumption 2.2.1, Q satisfies Assumptions 2.5.1, g satisfies Assumption 2.5.2, and $\hat{g}_m$ satisfies Assumption 2.5.3. Suppose in addition that as $m \to \infty$, $\sqrt{m}(\tilde{\theta}_m - \theta(Q)) \xrightarrow{d} N(0, \varsigma^2_{\text{pilot}})$, $\tilde{\varsigma}^2_{\text{pilot},m} \xrightarrow{P} \varsigma^2_{\text{pilot}}$, and that as $m, n \to \infty$,*

$$\frac{m}{m+2n} \to \nu \in [0,1] \ .$$

Then, under $\lambda^{\hat{g}_m}(X^{(n)})$ defined in (2.23), as $m, n \to \infty$,

$$\frac{\sqrt{m+2n}(\hat{\theta}^{\text{combined}}_n - \theta(Q))}{\sqrt{\frac{m}{m+2n}\tilde{\varsigma}^2_{\text{pilot},m} + \frac{2n}{m+2n}2(\hat{\varsigma}^{\hat{g}_m}_n)^2}} \xrightarrow{d} N(0,1) \ .$$

Thus, for the problem of testing (2.35) at level $\alpha \in (0,1)$, $\phi^{\text{combined}}_n(W^{(n)}, \tilde{W}^{(m)})$ in (2.44) satisfies

$$\lim_{m,n\to\infty} E[\phi^{\text{combined}}_n(W^{(n)}, \tilde{W}^{(m)})] = \alpha \ ,$$

whenever Q additionally satisfies the null hypothesis, i.e. $\theta(Q) = \theta_0$.

Remark 2.5.8. Although Theorem 2.5.5 is stated under $\lambda^{\hat{g}_m}(X^{(n)})$ in (2.23), it is straightforward to establish similar results when $\lambda^{\hat{g}_m}(X^{(n)})$ in the main experiment is replaced by other stratifications, e.g., (2.31). ■

2.6 Simulation

In this section, we examine the properties of the procedures discussed in Section 2.4 in a small simulation study. For $d \in \{0,1\}$ and $1 \leq i \leq 2n$, potential outcomes are generated according to the equation:

$$Y_i(d) = \mu(d) + m_d(X_i) + \sigma_d(X_i)\epsilon_i(d) \ ,$$

where $\mu(d)$, $m_d(X_i)$, $\sigma_d(X_i)$, and $\epsilon_i(d)$ are specified in each model as follows. In each of the following specifications, $2n = 200$; $((X_i, \epsilon_i(0), \epsilon_i(1)) : 1 \leq i \leq 2n)$ are i.i.d.; $X_i, \epsilon_i(0), \epsilon_i(1)$ are independent; and $\mu(0) = 0$. For each model, we generate data from a very small pilot experiment of sample size $m = 20$, in which half of the units are treated.

Model 1 $p = 2$; $X_{i,1} \sim \text{Beta}(2,2), X_{i,2} \sim \text{Beta}(2,2)$; $m_d(X_i) = X_i'\beta(d)$ and $\epsilon_i(d) \sim N(0,1)$ for $d \in \{0,1\}$; $\beta(1) = \beta(0) = (1,1)'$; $\sigma_0(X_i) = \sigma_1(X_i) = 0.1$.

Model 2 as in Model 1, but $\beta(1) = \beta(0) = (3, 0.1)'$.

Model 3 as in Model 1, but $\sigma_0(X_i) = \sigma_1(X_i) = 1$ and $\epsilon_i(d) \sim \text{Unif}\left[-\frac{1}{2}, \frac{1}{2}\right]$ for $d \in \{0,1\}$.

Model 4 as in Model 2, but $\sigma_0(X_i) = \sigma_1(X_i) = 1$ and $\epsilon_i(d) \sim \text{Unif}\left[-\frac{1}{2}, \frac{1}{2}\right]$ for $d \in \{0,1\}$.

Model 5 as in Model 1, but $m_1(X_i) = m_0(X_i) = X_{i,1}^2$, $\sigma_0(X_i) = \sigma_1(X_i) = 0.1$, and $\epsilon_i(d) \sim N(0,1)$ for $d \in \{0,1\}$.

Model 6 as in Model 5, but $m_1(X_i) = m_0(X_i) = X_{i,1}^2 + X_{i,2}^2$.

Model 1 is a symmetric model with small variances in error terms. Model 2 differs from Model 1 in that $X_{i,1}$ is the predominant component in potential outcomes. Models 3 and 4 are similar to Models 1 and 2, the only difference being that the error terms have larger variances. Models 5 and 6 are non-linear and are designed to study properties of the plug-in and the penalized procedures under misspecification. In Model 5, only $X_{i,1}$ affects the potential outcomes, while $X_{i,1}$ and $X_{i,2}$ are symmetric in Model 6.

We consider the following procedures:

Oracle matched-pair design with the infeasible optimal stratification in (2.20).

Plug-in matched-pair design with the plug-in stratification in (2.23) with $\hat{g}_m(x) = x'\hat{\beta}_m$ for $\hat{\beta}_m$ in (2.26).

Pen matched-pair design with the penalized stratification in (2.31).

MPeuc matched-pair design minimizing the sum of Euclidean distances within pairs.

by1 stratified randomization with two strata separated by the sample median of $X_{i,1}$.

by2 stratified randomization with two strata separated by the sample median of $X_{i,2}$.

MP1 matched-pair design using $X_{i,1}$ only, i.e., stratification in (2.23) with $\hat{g}_m(x) = x_1$.

MP2 matched-pair design using $X_{i,2}$ only, i.e., stratification in (2.23) with $\hat{g}_m(x) = x_2$.

Stratifications in **Pen** and **MPeuc** are computed using the package `nbpMatching` in `R`.

We first present results on the conditional MSE of $\hat{\theta}_n$ defined in (2.4). In these results, we set $\mu(1) = \mu(0) = 0$, so that $\theta(Q) = 0$ as well. By Lemma 2.3.1 and in particular (2.18), the conditional MSEs of $\hat{\theta}_n$ under stratifications differ only in terms of the variance of the ex-post bias conditional on the covariates. Therefore, for a given stratification λ, a set of covariates $X^{(n)}$, and the function g defined in (2.19), we define a constant multiple of the objective in (2.18) as the loss:

$$L(\lambda|g, X^{(n)}) = 4n^2 \operatorname{Var}_\lambda[E[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] \ . \tag{2.46}$$

Table 2.1 displays the summary statistics of the values of the loss defined in (2.46) for different stratifications across 1000 draws of $X^{(n)}$. We label the columns according to the procedures. In each model, we calculate ratios of values of the loss for each procedure against those for **Oracle**, and present the quartiles and means of the ratios across the 1000 draws of $X^{(n)}$.

Unsurprisingly, **Oracle** always has the smallest values of the loss. Ad-hoc procedures including **by1**, **by2**, **MP1**, **MP2** perform miserably most of the time. Although **MP1** performs well under Models 2, 4, and 5, it is because there $X_{i,1}$ is a predominant element of potential outcomes. In particular, Model 5 is an example where g defined in (2.19) is a

Model		**Oracle**	**Plug-in**	**Pen**	**MPeuc**	**by1**	**by2**	**MP1**	**MP2**
1	25%	1.00	2.50	3.69	22.51	2344.62	2353.34	885.77	903.36
	50%	1.00	8.46	5.76	35.86	3852.52	3848.06	1455.54	1435.83
	75%	1.00	28.03	9.93	55.50	5853.40	5866.36	2238.42	2183.49
	Mean	1.00	25.07	8.22	40.76	4281.19	4293.87	1653.90	1641.51
2	25%	1.00	2.08	4.33	67.39	3238.83	10723.31	6.89	5192.29
	50%	1.00	5.34	5.96	86.24	4211.93	14112.38	8.48	6954.55
	75%	1.00	15.21	9.53	108.13	5239.90	17414.65	10.57	8640.57
	Mean	1.00	12.85	8.14	89.26	4305.93	14377.14	8.90	7169.01
3	25%	1.00	16.28	8.57	22.52	2329.58	2340.74	894.50	902.57
	50%	1.00	68.97	14.04	35.55	3835.03	3850.74	1455.64	1466.20
	75%	1.00	230.33	25.64	54.02	5734.63	5783.08	2230.34	2226.22
	Mean	1.00	205.52	21.42	40.65	4288.67	4299.91	1650.97	1662.17
4	25%	1.00	8.86	10.27	67.58	3266.09	10924.10	6.91	5440.39
	50%	1.00	43.88	15.49	87.50	4125.96	13824.46	8.57	6847.59
	75%	1.00	131.81	26.43	109.16	5197.76	17364.76	10.65	8744.17
	Mean	1.00	104.72	22.07	89.41	4291.05	14343.10	8.97	7168.34
5	25%	1.00	27.39	71.83	415.34	19128.24	57595.61	1.00	27631.81
	50%	1.00	116.62	103.72	501.70	22248.95	66572.16	1.00	32579.67
	75%	1.00	333.13	176.04	599.89	26430.16	77215.74	1.00	38871.75
	Mean	1.00	318.20	150.85	520.67	23158.28	68653.31	1.00	34162.98
6	25%	1.00	244.36	115.27	214.18	27727.82	11878.77	13124.19	1424.60
	50%	1.00	342.09	150.88	265.06	32936.14	14190.12	15817.15	1726.21
	75%	1.00	517.14	197.98	328.09	39810.35	17243.41	18864.44	2118.38
	Mean	1.00	424.81	168.61	276.24	34031.92	14659.61	16327.06	1798.22

Table 2.1: Summary statistics for ratios of the values of the loss in (2.46) under all stratifications against those under the infeasible optimal stratifications (**Oracle**), over 1000 draws of $X^{(n)}$, in Models 1–6.

monotonic function of the first covariate, so that **MP1** solves (2.5) and has the same values of loss with **Oracle**. We separately discuss the remaining three procedures, **Plug-in**, **Pen**, and **MPeuc**:

Plug-in: In most models, **Plug-in** outperforms ad-hoc procedures including **by1**, **by2**, **MP1**, **MP2**, which is somewhat surprising since the sample size of pilot is only $m = 20$. In Models 1–2, where the variances of $\epsilon_i(d)$'s are small, **Plug-in** also improves upon **MPeuc**, and the improvement is pronounced in Model 2. But when the variances of $\epsilon_i(d)$'s are large, it performs worse than **Pen** and **MPeuc**, as could be seen from Models 3–6.

Pen: In Models 1–4, **Pen** is the best among all procedures. In all models, it performs better than **Plug-in** and **MPeuc**, remarkably so than **Plug-in** in Models 3–6. The improvement upon **MPeuc** is most pronounced in Models 2 and 4, where $X_{i,2}$ contributes little to potential outcomes. These are examples in which **MPeuc** assigns equal weights to two covariates while regression-based methods could detect that one of them dominates. Even when potential outcomes are non-linear (Models 5–6), the values of its loss are smaller than those under **MPeuc**.

MPeuc: In all models, it is not as poor as the ad-hoc procedures including **by1**, **by2**, **MP1**, **MP2**, but is obviously worse than **Pen**. In Models 2 and 4, where only $X_{i,1}$ matters, it is obviously worse than **Pen** and **Plug-in**, because the pilot informs us that $X_{i,1}$ is much more important than $X_{i,2}$, which is not taken into account by Euclidean matching.

Next, for $\theta_0 = 0$, we consider the problem of testing (2.35) at level $\alpha = 0.05$. For Models 1–6, we compute the rejection probabilities of suitable tests under stratifications mentioned previously, when $\mu(0) = 0$ and $\theta = \mu(1) = 0, 0.0.1, 0.02, 0.04$. In particular, we use the following tests under each stratification:

Oracle: test in (2.38) with $\hat{g}_m = g$ for g defined in (2.19).

Plug-in: test in (2.38) with $\hat{g}_m(x) = x'\hat{\beta}_m$ for $\hat{\beta}_m$ defined in (2.26).

Pen: test in (2.42).

MPeuc: test in (2.42) with Z replaced by X.

by1: test in (2.38) with $\hat{g}_m(x) = I\{x_1 > \text{med}(X_{i,1} : 1 \le i \le 2n)\}$.

by2: test in (2.38) with $\hat{g}_m(x) = I\{x_2 > \text{med}(X_{i,2} : 1 \le i \le 2n)\}$.

MP1: test in (2.38) with $\hat{g}_m(x) = x_1$.

MP2: test in (2.38) with $\hat{g}_m(x) = x_2$.

Table 2.2 displays the rejection probabilities for Models 1–6 under all stratifications using tests described above. Note that loss properties in Table 2.1 translate into power properties in Table 2.2. Indeed, while all tests under all stratifications have correct sizes, the test in (2.42) under the penalized stratification in (2.31) has higher power than most other tests under other stratifications, except that under **Oracle**. In Models 1–2, the corresponding tests under **Plug-in** and **Pen** have higher power than that under **MPeuc**, while being comparable in other models, except in Model 6, where potential outcomes are highly non-linear. The comparison is most pronounced in Model 2, where g in (2.19) depends mostly on x_1, because **Plug-in** and **Pen** incorporate information from the pilot while **MPeuc** doesn't. The test under **Pen** performs better than that under **Plug-in** in Models 1–5. Finally, note that tests under matched-pair designs, including **Plug-in**, **Pen**, and **MPeuc** usually perform much better than tests under stratifications with a small number of large strata, including **by1** and **by2**.

2.7 Empirical Application

To illustrate our procedures in practice, we replicate part of the experiment in DellaVigna and Pope (2018) on Amazon Mechanical Turk (MTurk) and the TurkPrime Prime Panels, using the penalized procedure defined by (2.31). MTurk is an online crowdsourcing platform widely used to conduct economic and behavioral experiments. For more information about running experiments on Amazon MTurk, see Horton et al. (2011), Mason and Suri (2012), Paolacci and Chandler (2014), Kuziemko et al. (2015), and Litman et al. (2017). Prime Panels is another online platform with over 30 million participants and their reliable demographics.

DellaVigna and Pope (2018) run a large-scale experiment to compare the effectiveness of multiple incentives for efforts in one setting, as well as compare experimental results with expert forecasts. The 18 treatments include various monetary and behavioral incentives. We

Model		**Oracle**	**Plug-in**	**Pen**	**MPeuc**	**by1**	**by2**	**MP1**	**MP2**
1	$\theta = 0$	5.63	5.15	5.61	5.48	5.02	5.27	5.44	5.45
	$\theta = 0.01$	11.21	10.63	11.2	11	6.34	6.41	6.15	6.24
	$\theta = 0.02$	30.26	28.32	29.76	27.31	8.02	8.19	9.83	9.6
	$\theta = 0.04$	79.44	76.86	79.98	75.4	17.71	18.12	20.87	23.19
2	$\theta = 0$	5.43	5.05	5.12	5.24	5.37	5.47	5.32	5.88
	$\theta = 0.01$	11.72	10.84	11.06	9.68	5.54	5.57	10.96	5.53
	$\theta = 0.02$	28.52	27.45	27.88	20.5	7.35	5.6	27.14	5.81
	$\theta = 0.04$	79.82	76.23	78.6	62.6	11.98	6.79	77.77	7.19
3	$\theta = 0$	5.08	5.61	5.32	5.34	5.51	5.7	5.37	5.26
	$\theta = 0.01$	5.69	6.11	6.33	5.58	5.93	5.46	5.51	5.57
	$\theta = 0.02$	8.22	7.49	8.18	8.43	6.92	6.92	7.27	7.67
	$\theta = 0.04$	17.52	16.66	16.94	16.84	11.82	12.31	12.67	12.84
4	$\theta = 0$	5.69	5.55	5.7	5.31	5.43	5.16	5.2	5.14
	$\theta = 0.01$	6.31	6.2	6.69	5.98	5.72	5.49	6.32	5.72
	$\theta = 0.02$	8.1	7.98	8.13	7.87	6.97	5.91	8.05	5.88
	$\theta = 0.04$	16.73	16.77	17.02	16.75	9.69	7.28	16.81	7.28
5	$\theta = 0$	5.33	5.26	5.66	5.5	5.47	5.38	5.6	5.16
	$\theta = 0.01$	11.44	10.93	11.57	11.5	7.78	6.56	11.64	6.5
	$\theta = 0.02$	30.34	28.2	30.02	28.44	14.36	9.28	30.02	9.23
	$\theta = 0.04$	80.81	77.12	79.89	77.46	40.39	20.83	80.52	21.93
6	$\theta = 0$	5.15	5.47	3.51	4.94	5.57	5.78	5.78	5.72
	$\theta = 0.01$	6.77	6.84	4.44	6.46	5.72	5.7	5.62	6.52
	$\theta = 0.02$	12.41	11.49	8.72	11.22	6.79	7.91	6.69	10.55
	$\theta = 0.04$	31.94	29.34	24.37	29.18	10.45	16.31	10.94	25.43

Table 2.2: Rejection probabilities for Models 1–6 under all stratifications using tests in Section 4.

focus on one of the treatments, which is a monetary incentive. In the experiment, subjects are asked to alternately press the "a" and "b" buttons on their keyboard as quickly as possible in 10 minutes. One alternate press counts as 1 point. All subjects are paid some base rate upon finishing the experiment. In the treatment we replicate, subjects in the treated group are paid an extra \$0.01 for every 100 points they score, while subjects in the control group receive no extra payment. In DellaVigna and Pope (2018), the base payment is \$1, but we use about \$1.25 in the pilot and \$2 in the main experiment to minimize attrition. In our notation, the outcome Y is the points scored, the treatment D indicates whether the subject

receives extra payment ($D = 1$) or not ($D = 0$). The covariates X include a constant term, age, gender, ethnicity, education, and income. We re-index gender and ethnicity as binary variables and regard the rest as continuous.

The sample size in the original experiment in DellaVigna and Pope (2018) is 1098. In the original experiment, all the units are in one stratum and the treated fraction is approximately $\frac{1}{2}$. There is a pilot experiment in the preregistration stage but the results used in neither designing the main experiment nor analysis in their paper. In our replication, we perform the pilot experiment on Prime Panels and the main experiment on MTurk. The sample size of the pilot experiment is $m = 44$, and that of the main experiment is $2n = 176$. We could not replicate the original experiment with 1098 units because of the budget constraint.

After collecting data from the pilot experiment, we calculate the penalized stratification defined in (2.31), and conduct inference on the ATE in two ways: disregarding data from the pilot experiment as in (2.42), and combining data from the pilot and main experiments as in (2.44). We compare the results with the original ones in DellaVigna and Pope (2018). For a meaningful comparison, we also present the scaled-up version of the original standard errors in DellaVigna and Pope (2018) to match the sample size in our replication. Table 2.3 lists the sample sizes and difference-in-means estimates, standard errors, and t-statistics. Since there is only one stratum in DellaVigna and Pope (2018), the two-sample t-test is asymptotically exact in their setup. The columns correspond to the following:

Pen penalized stratification in (2.31) and the test statistic in (2.43).

Combined penalized stratification in (2.31) and the test statistic in (2.45).

Original (scaled) results in DellaVigna and Pope (2018), with sample size scaled down to $2n + m$ and standard error scaled up accordingly.

Original results in DellaVigna and Pope (2018) and the two-sample t-statistic.

We see that the standard error under **Combined** is 29% smaller than that under **Original (scaled)**. Equivalently, to attain the same standard error, **Combined** requires only

about half the sample size of that under the stratification in DellaVigna and Pope (2018).

	Pen	**Combined**	**Original (scaled)**	**Original**
sample size	176	220	220	1098
$\hat{\theta}_n$	644	624	-	499
s.e.	108.16	**92.05**	**129.95**	58.70
t-statistic	5.95	6.78	-	8.50

Table 2.3: Summary statistics from DellaVigna and Pope (2018) and our replication.

2.8 Minimax Procedure

Finally, we discuss alternative procedures without reliable pilot data. In some experiments pilot data is not available, or even if there is a pilot experiment, the units might not be drawn from the same population as the main experimental units. On the other hand, the procedure in Theorem 2.3.1 is optimal in population, which translates into optimality with large pilots in Theorem 2.5.1, while the penalized procedure in (2.31) is based on optimality in integrated risk in a Bayesian framework, assuming linearity and normality. It is then natural to ask about finite sample optimality without linearity and normality. To answer the question, we introduce a minimax problem. We briefly highlight the results and leave all details to Appendix B.5. By Lemma 2.3.1 and in particular (2.18), the conditional MSEs of $\hat{\theta}_n$ under stratifications differ only in terms of the variance of the ex-post bias conditional on the covariates, and hence we define a constant multiple of it as the loss in (2.46). Moreover, we have

$$L(\lambda|g, X^{(n)}) = 4n^2 \operatorname{Var}_\lambda[E[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] = \sum_{1\le s\le S} \frac{1}{n_s - 1} \sum_{i,j\in\lambda_s, i<j} (g_i - g_j)^2 \ . \tag{2.47}$$

Consider the following minimax problem to find the stratification λ that has the best worst-case performance in terms of the loss in (2.47), where the worst-case is among a class of

functions $\mathcal{G}$.

$$\min_{\lambda \in \Lambda} \max_{h \in \mathcal{G}} L(\lambda | h, X^{(n)}) \ . \tag{2.48}$$

Our framework requires $\mathcal{G}$ to have a bounded polyhedron structure, in the sense made precise by Assumption B.5.1. The assumption is satisfied by a large class of shape restrictions on $\mathcal{G}$, including Lipschitz continuity, monotonicity, and convexity.

Our first result shows that when $p = 1$, under a Lipschitz model, (2.48) is solved by matching on X directly. It reflects the intuition to match on the covariate itself when little information is available on how the covariate affects potential outcomes. For more details, see Theorem B.5.1. Unfortunately, such a result no longer holds when $p > 1$. Indeed, Example B.5.7 shows that matched-pair designs may not even be minimax-optimal. We show, however, that under Assumption B.5.1 it is possible to reformulate (2.48) into a mixed-integer linear program. The reformulation is based on the special structure in (2.47), which enables us to rewrite (2.48) into a problem in graph theory, related to but more complicated than what is known in the literature as the clique partitioning problem. The program is computationally intensive, and therefore we consider a relaxation which replaces $\lambda \in \Lambda$ in the minimization in (2.48) with $\lambda \in \Lambda^{\text{pair}}$. The resulting program, related to what is known in the literature as the minimum-weight perfect matching problem, is computationally much easier and could be computed using modern solvers such as `Gurobi`. In Appendix B.5, we compute the solutions in a simulation study. Simulation evidence suggests that although the minimax matched-pair design is in general not minimax-optimal among all stratifications, it is often close to optimal in a sense we make precise in the appendix.

2.9 Conclusion and Recommendations for Empirical Practice

This chapter provides a framework under which a certain matched-pair design is optimal among all stratified randomization procedures. To the best of our knowledge, this is the first formal justification in the literature on the use of matched-pair designs based on optimality

results. We show it is optimal to match units according to the sum of expectations of potential outcomes if treated and untreated conditional on the covariates. We then provide empirical counterparts to the optimal stratification and study their properties. In particular, we provide different procedures under large and small pilots, as well as inference procedures under each of them. From the theoretical point of view, stratifying impacts the estimation efficiency of RCTs in terms of the ex-ante MSE, i.e., before treatment status is assigned, and the ex-post bias, i.e., after treatment status is assigned. Lemma 2.3.1 shows that ex-post bias translates into ex-ante MSE, and hence impacts the estimation of treatment effects in an RCT. From a practical point of view, matched-pair designs weakly improve estimation and typically strictly do so, as long as the function used in matching satisfies the regularity conditions laid out in Assumption 2.5.4. Therefore, we recommend researchers to consider using matched-pair designs, or corresponding procedures in Appendix B.2, when treated fractions are identical across strata but not $\frac{1}{2}$ and when they are in addition allowed to vary across subpopulations.

Both our theoretical and simulation results suggest that the efficiency for estimation of ATE could be improved, often notably, by incorporating information from pilot data. Therefore, we recommend researchers to perform pilot studies, on the same population as the main experiment. Based on Theorem 2.5.2, we recommend researchers to use machine learning methods when the pilot is large. When the pilot is small, although the plug-in procedure improves over not stratifying at all, we recommend them to be used with simple plug-in estimators, such as least squares. Moreover, researchers should consider using the penalized procedure as it shows sizable improvement over the plug-in procedure in simulation studies.

CHAPTER 3

RANDOMIZATION UNDER PERMUTATION INVARIANCE

3.1 Introduction

This chapter studies the minimax optimality of certain randomization schemes and assignment schemes in estimating "reasonable" parameters including the average treatment effects, when treatment effects are heterogeneous. By a randomization scheme, I mean the distribution over a group of permutations of a given treatment assignment vector. By an assignment scheme, I mean the joint distribution over assignment vectors, linear estimators, and permutations of assignment vectors. Randomization is prevalent in the sciences, especially in randomized controlled trials. See Rosenberger and Lachin (2015) for a review in clinical trials and Bruhn and McKenzie (2009) for a review in development economics. Common randomization schemes include simple random sampling, stratified block randomization (Bertrand et al., 2007; Olken, 2007; Duflo et al., 2015b; Dupas et al., 2018), and matched-pair designs (List and Rasul, 2011; Groh and McKenzie, 2016; Bertrand and Duflo, 2017). But the benefits of randomization are frequently taken as granted. In a recent paper, Kasy (2016) studies a Bayesian framework with a given prior on the distribution of potential outcomes and concludes that the optimal Bayesian decision never involves randomization. He therefore recommends experimental designers to never randomize.

Wu (1981), on the other hand, provides three arguments for using randomization. First, it helps researchers conduct inference. See Bugni et al. (2018), Bugni et al. (2019), and Chapter 1 of this dissertation for several inference procedures that are asymptotically exact. Second, it guarantees impartiality since experimental objects could not manipulate themselves towards treatment groups. Third, it guards the researcher against model inadequacies. By model inadequacy, I mean the researcher has only partial knowledge of how covariates affect potential outcomes. This chapter studies the third argument, following a series of papers by Wu (1981), Li (1983), and Hooper (1989).

I show that for any given assignment vector and any estimator, the complete randomization scheme is minimax optimal for any objective function satisfying quasi-convexity, where the worst-case is over a permutation-invariant class of distributions of the data. Objective functions satisfying quasi-convexity include the expectation operator, the quantile function, and the survival function. Under further conditions on the distribution of the data, I characterize the minimax optimal assignment scheme, where the worst-case is again over a permutation-invariant class of distributions of the data. Importantly, these results are derived under heterogeneous treatment effects.

There is a large literature on minimax estimation, including Donoho (1994) and Armstrong and Kolesár (2018). But these papers either don't involve randomization or take randomization schemes as given. This chapter derives the optimal assignment scheme, which comprises an assignment vector, a linear estimator, and a distribution over permutations of the assignment vector. The results, however, holds under strong assumptions. Chapter 3 of this dissertation characterizes optimal stratifications in the oracle, as well as provides an algorithm to calculate minimax matched-pair designs. The results, however, are derived taking as given the difference-in-means estimator.

This chapter is organized as follows. Section 3.2 proves the minimax optimality of the complete randomization scheme for any given assignment vector and any given estimator. Under strong assumptions, Section 3.3 derives the optimal assignment scheme. Section 3.4 concludes. Since the optimality results in Section 3.3 are complicated, I provide an example in Appendix C.1 under homogeneous treatment effects and show the optimal assignment scheme is intuitive. Appendix C.2 contains all the proofs.

3.2 Minimaxity under Permutation Invariance

Let $i = 1, \ldots, n$ denote the underlying units. Denote by $\tilde{Y}_i(1)$ the potential outcome of the ith unit if treated and by $\tilde{Y}_i(0)$ if not treated. Let Z_i denotes the observed covariates of the

ith unit. For $a \in \{0, 1\}$, the potential outcome $\tilde{Y}_i(a)$ is determined by the model

$$\tilde{Y}_i(a) = \mu_a + m_a(Z_i) + \epsilon_i(a) \ , \tag{3.1}$$

where $m_a(Z_i) = \mu_a(Z_i) - \mu_a$ for $\mu_a(Z_i) = E[\tilde{Y}_i(a)|Z_i]$. Further define

$$\begin{aligned}
\tilde{Y}_i &= (\tilde{Y}_i(0), \tilde{Y}_i(1))' \\
\mu &= (\mu_0, \mu_1)' \\
\epsilon_i &= (\epsilon_i(0), \epsilon(1))' \\
m(Z_i) &= (m_0(Z_i), m_1(Z_i))' \ .
\end{aligned}$$

Define $\tilde{\mathbf{Y}} = (Y_1, \ldots, Y_n)'$ and similarly for $\mathbf{Z}$, $\boldsymbol{\epsilon}$, and $\mathbf{m}$. Define $\xi = \mathcal{L}(\boldsymbol{\epsilon}|\mathbf{Z})$, the distribution of $\boldsymbol{\epsilon}$ conditional on $\mathbf{Z}$.

Let H denote a group of permutations. An element $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ maps each integer from 1 to n to an integer in the same set. I identify each $\pi \in H$ as a $n \times n$ permutation matrix such that the $(i, \pi^{-1}(i))$-th elements are ones for $1 \leq i \leq n$ and the rest are zeros. Pre-multiplying the vector $(1, \ldots, n)'$ by the matrix π generates a vector with i on the $\pi(i)$-th position. For instance, if $\pi(5) = 1, \pi(3) = 2, \pi(2) = 3, \pi(4) = 4, \pi(1) = 5$, then π is described by

$$\begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix} \xrightarrow{\pi} \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix} = \begin{pmatrix} 5 \\ 3 \\ 2 \\ 4 \\ 1 \end{pmatrix} \ .$$

See Artin (2013) for a detailed introduction of permutations and group theory. H doesn't have to include all permutations on $\{1, \ldots, n\}$, as will be seen in Example 3.2.2.

For any vector X, $\pi X = (X_{\pi^{-1}(1)}, \ldots, X_{\pi^{-1}(n)})$, so permuting a vector is equivalent to applying the inverse of permutation to its indices. For a matrix $\mathbf{X} = (X_1, \ldots, X_n)'$, define $\pi\mathbf{X} = (X_{\pi^{-1}(1)}, \ldots, X_{\pi^{-1}(n)})'$. Furthermore, define $\pi\xi = \mathcal{L}(\pi\boldsymbol{\epsilon}|\mathbf{Z})$ if $\xi = \mathcal{L}(\boldsymbol{\epsilon}|\mathbf{Z})$.

Define $s = (\mu, \mathbf{m}, \xi)$ as the state and $\pi s = (\mu, \pi\mathbf{m}, \pi\xi)$. I make the following assumption on S, the set of possible states.

Assumption 3.2.1. S is H-invariant, i.e. $s \in S \Rightarrow \pi s \in S$ for all $\pi \in H$, where H is a group of permutations.

The following example demonstrates Assumption 3.2.1.

Example 3.2.1 (Complete Invariance)**.** In model (3.1), denote by M and $\mathcal{E}$ the set of possible values of $\mathbf{m}$ and ξ. Suppose the researcher doesn't want to impose any assumption on $\mathbf{m}$ and $\boldsymbol{\epsilon}$ beyond permutation invariance and therefore H is the set of all permutations on $\{1, \ldots, n\}$. Then, for any permutation $\pi \in H$,

$$\begin{aligned} \mathbf{m} \in M \Rightarrow \pi\mathbf{m} \in M \\ \xi \in \mathcal{E} \Rightarrow \pi\xi \in \mathcal{E} \ . \end{aligned} \tag{3.2}$$

Complete invariance is natural when the researcher doesn't have ex-ante information on either how Z_i affects $Y_i(a)$ for $a \in \{0, 1\}$ or how the errors are distributed. ■

Example 3.2.2 (Block Model)**.** Suppose $Z_i = (Z_{1i}, Z_{2i})$ where $Z_{1i} \in \{1, \ldots, B\}$. Consider model (3.1) with

$$m_a(Z_i) = \gamma_{Z_{1i}}(a) + \tilde{m}_a(Z_i) \tag{3.3}$$

for $\gamma_b(a) = E[Y_i(a)|Z_{1i} = b]$ and $\tilde{m}_a(Z_i) = E[Y_i(a)|Z_i] - \gamma_{Z_{1i}}(a)$. Suppose H consists of all permutations that respects block structure:

$$H = \{\pi : Z_{1\pi(i)} = Z_{1i} \text{ for } i = 1, \ldots, n\} \ .$$

The model reflects the belief that in determining the potential outcomes, different values of

Z_{1i} implies different means of potential outcomes, but the researcher either doesn't observe Z_{2i} or doesn't know how Z_{2i} affects the potential outcome. Note that in this case $\pi\mathbf{m} = (\gamma, \pi\tilde{\mathbf{m}})$ for $\tilde{\mathbf{m}} = (\tilde{m}(Z_1), \ldots, \tilde{m}(Z_n))'$ and $\tilde{m}(Z_i) = (\tilde{m}_0(Z_i), \tilde{m}_1(Z_i))'$. ■

We distinguish between the labels of underlying units and observed units. Let $l = 1, \ldots, n$ denote the observed units. Note that the same value of l and i need not correspond to the same unit because i indexes underlying units while l indexes observed units. Define $A_l = (0, 1)'$ if the lth observed unit is treated and $A_l = (1, 0)'$ otherwise. Define the matrix $\mathbf{A} = \sum_{l=1}^{n} e_{ll} \otimes A_l'$ where e_{ll} is $n \times n$ matrix with all zeros except the (l, l)-th element being 1. For example, if

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} ,$$

then observed units $l = 1, 2$ are always treated and $l = 3, 4$ are always control, no matter what underlying units they correspond to.

The observed quantities are

$$Y_l = A_l' \tilde{Y}_l \ , \tag{3.4}$$

and its matrix version is

$$\mathbf{Y} = \mathbf{A} \operatorname{vec}(\tilde{\mathbf{Y}}') = \mathbf{A}(\mathbf{1}_n \otimes \mu + \operatorname{vec}(\mathbf{m}' + \boldsymbol{\epsilon}')) \ , \tag{3.5}$$

where $\mathbf{1}_n$ is the $n \times 1$ vectors of 1's. For $\pi \in H$, $(\pi^{-1}\mathbf{A}) \operatorname{vec}(\tilde{\mathbf{Y}}')$ is observationally equivalent to $\mathbf{A} \operatorname{vec}(\tilde{\mathbf{Y}}'\pi')$ since each underlying unit $i = 1, \ldots, n$ has the same observed outcome $A_{\pi(i)}' \tilde{Y}_i$ in the two cases. Since the group of permutation is closed under inversion, instead of applying it to $\mathbf{A}$, I apply a permutation $\pi \in H$ to $\tilde{\mathbf{Y}}$ while keeping $\mathbf{A}$ fixed. In the above example with four units, $l = 1, 2$ are always treated, but their correspond underlying units are $i = \pi^{-1}(1), \pi^{-1}(2)$, which changes with π. Accordingly, define the observed quantities

under permutation using the following notation:

$$Y_l^\pi = A_l' \tilde{Y}_{\pi^{-1}(l)} \tag{3.6}$$

and its matrix version

$$\mathbf{Y}^\pi = \mathbf{A}(\mathbf{1}_n \otimes \mu + \text{vec}[(\mathbf{m}' + \boldsymbol{\epsilon}')\pi']) \ . \tag{3.7}$$

Remark 3.2.1. The seemingly unconventional description of the model in (3.6) serves a practical purpose: I keep the matrix $\mathbf{A}$ fixed across permutations and only permute $\mathbf{m}$ and $\boldsymbol{\epsilon}$ and hence avoid simultaneously relabelling of units and assignments. In Example 3.2.1, H is the group all permutations of $\{1, \ldots, n\}$, and fixing $\mathbf{A}$ means that the total numbers of treated and control units stay the same. In Example 3.2.2, H is the group all permutations of $\{1, \ldots, n\}$ which respects the block structure, and fixing $\mathbf{A}$ means that the numbers of treated and control units within each block stay the same. ■

I now formally define randomization schemes. Recall that $\mathbf{H}$ is a uniform distribution on a group H if $\pi\mathbf{H} \stackrel{d}{=} \mathbf{H} \stackrel{d}{=} \mathbf{H}\pi$ for all $\pi \in H$. I define a randomization scheme $\mathbf{G}$ as a distribution on H. In addition, I define the complete randomization scheme as the uniform distribution $\mathbf{H}$ on H.

Let Φ denote the set of all randomization schemes. I use an estimator δ which depends only on $\mathbf{Y}^\pi$, the observed data. I impose an additional assumption on the loss function L, which says the researcher is only interested in μ instead of $\mathbf{m}$ or ξ.

Assumption 3.2.2. $L(s, d) = L(\pi s, d)$ for all $\pi \in H$, $s \in S$ and action d.

Let δ be the estimator used. An example that satisfies Assumption 3.2.2 is the Average Treatment Effect (ATE), $\delta(\mathbf{Y}^\pi) = \bar{Y}_1^\pi - \bar{Y}_0^\pi$ with the squared loss $L(s, d) = (\bar{Y}_1^\pi - \bar{Y}_0^\pi - (\mu_1 - \mu_0))^2$, where

$$\bar{Y}_1^\pi = \frac{\sum\limits_{l:A_l=(0,1)'} Y_l^\pi}{\sum\limits_{1 \le l \le n} I\{A_l = (0,1)\}} \ ,$$

the mean of the treated group, and similarly for $\bar{Y}_0^\pi$.

Define the risk $r(\pi, \mathbf{A}, \delta; s) = E_s[L(s, \delta(\mathbf{Y}^\pi))|\mathbf{Z}]$, where the expectation is based on the state s and conditional on the covariates $\mathbf{Z}$. In addition, define the randomized risk $r(\mathbf{G}, \mathbf{A}, \delta; s) : (H, 2^H, \mathbf{G}) \to (\mathbb{R}, \mathcal{R}, \text{Leb})$ as a random variable with realizations $r(\pi, \mathbf{A}, \delta; s) \in \mathbb{R}$ where $\pi \sim \mathbf{G}$. Note the distribution of the risk $r(\mathbf{G}, \mathbf{A}, \delta; s)$ is a measure on $\mathbb{R}$. The performance of $\mathbf{G}$ is summarized by a functional f which maps each measure on $\mathbb{R}$ to a real number. f could the expectation operator but I only impose quasiconvexity on f.

Assumption 3.2.3. A functional f on the collection of probability measures on $\mathbb{R}$ is quasiconvex, i.e., $f(\lambda\mu_1 + (1-\lambda)\mu_2) \leq \max\{f(\mu_1), f(\mu_2)\}$ for all probability measures μ_1, μ_2 on $\mathbb{R}$ and for all $\lambda \in [0, 1]$.

Remark 3.2.2. Other examples which satisfy Assumption 3.2.3 include the quantile function and the survival function, i.e., $1 - F$ where F is the distribution function of the measure. With the survival function the distributions of the risk are compared in terms of First Order Stochastic Dominance (FOSD). ■

The following theorem shows the minimax optimality of the complete randomization scheme, for any given assignments $\mathbf{A}$ and any estimator δ.

Theorem 3.2.1. *Suppose Assumptions 3.2.1–3.2.3 hold. Then for any* $\mathbf{A}$ *and* δ, $\mathbf{H}$ *solves*

$$\min_{\mathbf{G} \in \Phi} \max_{s \in S} f(r(\mathbf{G}, \mathbf{A}, \delta; s)) \ . \tag{3.8}$$

Remark 3.2.3. Theorem 3.2.1 says that given any assignment vector $\mathbf{A}$ and any estimator δ, the complete randomization scheme achieves the minimax risk. The intuition is as follows. To begin with, I need only show the conclusion with the worst-case over the orbit of each fixed state s_0, $\{\pi s_0 : \pi \in H\}$. Given model (3.7), πs_0 generates the same distribution of $\mathbf{Y}$ given $\mathbf{Z}$ as if $\mathbf{G}$ is post-multiplied by π. The problem then reduces to finding the maximum risk of $\mathbf{G}\pi$ over π. However, $\mathbf{GH} \stackrel{d}{=} \mathbf{H}$, so the average randomized risk over π of $\mathbf{G}\pi$ is

equal in distribution to the randomized risk of **H**. Since f is quasiconvex and the average is weakly smaller than the maximum, the conclusion holds. ■

Remark 3.2.4. The main takeaway from Theorem 3.2.1 is as follows. If the researcher believes the distributions of unit-specific effects m and ϵ are invariant to permutations but has no idea of the exact values, he should randomize to guard himself against the worst case. If **G** is a degenerate distribution on the identity matrix I_n so that he never randomizes, and complete invariance in Example 3.2.1 holds, then there exists a very adverse outcome that hurts him very much. But if he applies the complete randomization scheme **H**, then because of the quasi-convexity of E, he guarantees the average risk is weakly smaller than the worst case. ■

Remark 3.2.5. Note that Theorem 3.2.1 holds under any **A** and δ. It is possible though, that there exists no optimal **A** and δ. Theorem 3.3.1, however, provides a positive result in many setups. ■

3.3 Optimal Assignment Scheme

Remark 3.2.5 says that optimal **A** and δ need not exist. Hooper (1989) provides a recipe for finding the optimal ones under homogeneous treatment effect. In this section, I derive a new result under heterogeneous treatment effects.

I continue to work with the model defined by (3.1) and (3.7), but with a particular loss function and using linear estimators. The parameter of interest is

$$\tau'\mu \text{ for } \tau = (\tau_0, \tau_1)', \tag{3.9}$$

the ATE is an example of which with $\tau = (-1, 1)'$. The estimator is $\delta(\mathbf{Y}) = \beta'\mathbf{Y}$ for some $\beta \in \mathbb{R}^n$ and

$$L(s, d) = \|d - \tau'\mu\|^2 . \tag{3.10}$$

The loss function in (3.10) clearly satisfies Assumption 3.2.2 and hence I need only impose Assumption 3.2.1. Define

$$\mathcal{A}_n = \left\{ \sum_{l=1}^{n} e_{ll} \otimes A_l' : A_l \in \{(1,0)', (0,1)'\} \text{ for } l = 1, \ldots, n \right\} .$$

I allow $\mathbf{A}$ and β to be random.

I formally define an assignment scheme $\mathcal{L}(\mathbf{A}, \beta, \pi)$ as a joint distribution of $(\mathbf{A}, \beta, \pi) \in \mathcal{A}_n \times \mathbb{R}^n \times H$. To simplify notations, I define $U_i = m(Z_i) + \epsilon_i$ and $\mathbf{U} = (U_1, \ldots, U_n)'$. Also define $P_{\mathbf{U}} = \mathcal{L}(\mathbf{U}|\mathbf{Z})$ and $\pi P_{\mathbf{U}} = \mathcal{L}(\pi\mathbf{U}|\mathbf{Z})$ if $P_{\mathbf{U}} = \mathcal{L}(\mathbf{U}|\mathbf{Z})$. Denote by $\mathcal{P}_{\mathbf{U}}$ the set of all possible $P_{\mathbf{U}}$'s. Assumption 3.2.1 could be reformulated as follows.

Assumption 3.3.1. $\mathcal{P}_{\mathbf{U}}$ is H-invariant, i.e., $P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}} \Rightarrow \pi P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}$ for all $P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}$ and all $\pi \in H$, where H is a group of permutations.

Define the following set

$$C = \left\{ c \in \mathbb{R}^{2n} : \sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} c' E[(\mathbf{H} \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')' (\mathbf{H}' \otimes I_2) | \mathbf{Z}] c < \infty \right\} \tag{3.11}$$

and its projection matrix Q. As a projection matrix, Q is symmetric and idempotent.

For a family of positive semi-definite (p.s.d.) matrices $S_n \in \mathbb{R}^{2n \times 2n}$, introduce the partial order $\leq$ such that $V_1 \leq V_2$ if $V_2 - V_1$ is p.s.d. Define

$$V_{P_{\mathbf{U}}} = E[Q \operatorname{vec}(\mathbf{H} \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')' (\mathbf{H}' \otimes I_2) Q | \mathbf{Z}] . \tag{3.12}$$

The following assumption is nontrivial but holds in Example C.1.1.

Assumption 3.3.2. There exists a V such that

$$V_{P_{\mathbf{U}}} \leq V \text{ for all } \mathcal{P}_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}} \text{ and } \sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} \operatorname{tr} V_{P_{\mathbf{U}}} = \operatorname{tr} V . \tag{3.13}$$

Theorem 3.3.1. *Suppose Assumptions 3.3.1 and 3.3.2 hold for $V < \infty$ and $(\mathbf{A}_0, \beta_0)$ solves*

$$\begin{aligned} \min_{\mathbf{A},\beta} \ & \beta' \mathbf{A} V \mathbf{A}' \beta \\ \textit{subject to} \ \ & Q\mathbf{A}'\beta = \mathbf{A}'\beta, \\ & \beta' \mathbf{A} \mathbf{1}_n \otimes (1,0)' = \tau_0 \\ & \beta' \mathbf{A} \mathbf{1}_n \otimes (0,1)' = \tau_1 \end{aligned}$$

for $\mathbf{1}_n$ the $n \times 1$ vector of 1's. Then, $(\mathbf{A}_0, \beta_0, \mathbf{H})$ solves

$$\min_{\mathcal{L}(\mathbf{A},\beta,\pi)} \sup_{(\mu, P_{\mathbf{U}}) \in \mathbb{R}^2 \times \mathcal{P}_{\mathbf{U}}} E[\|\beta' \mathbf{Y}^{\pi} - \tau' \mu\|^2 | \mathbf{Z}] \ . \tag{3.14}$$

Remark 3.3.1. The constraints of the optimization problem in Theorem 3.3.1 are easily interpretable in many examples. Consider the homogeneous treatment effect example Example C.1.1, where the parameter of interest is the treatment effect and a block structure holds. Then, the first constraint restricts the coefficients β_l's to sum to zero in each block, while the next two constraints guarantee that the sum of β_l's of the treated group to sum up to 1, while those of the control group sum up to -1. If in addition block sizes are equal, then the difference-in-means estimator with half of each block assigned to treatment is minimax-optimal. ■

Remark 3.3.2. Although I allow $\mathbf{A}$ and β to be random, the optimal assignment scheme in Theorem 3.3.1 has degenerate $\mathbf{A}$ and β, so it is optimal to randomize a given assignment vector $\mathbf{A}$ using $\mathbf{H}$ and then apply a given estimator. In Example 3.2.1, it means that the total numbers of treated and control units are fixed. In Example 3.2.2, it means that the total numbers of treated and control units within each block are fixed and we should only permute within blocks. ■

3.4 Conclusion

This chapter shows the minimax optimality of the complete randomization scheme under permutation invariance. I also characterize under heterogeneous treatment effects the optimal assignment scheme which is the complete randomization of a deterministic assignment vector together with a fixed linear estimator.

It is natural to ask what would happen if invariance assumptions do not hold. For instance, the data might not display the exact block structure as in Example 3.2.2, although there are some discrete variables. Section 2.8 in Chapter 2 of this dissertation studies the optimal stratified randomization both with and without an estimate of conditional expectation functions. When the researcher does not have access to pilot data so could not estimate conditional expectation functions, I recommend that they solve a minimax problem in a similar spirit to the one studied here, where the worst-case is over a bounded polyhedron. I also provide fast algorithms to search over a restricted set of randomization schemes, the set of all matched-pair designs.

APPENDIX A

APPENDIX FOR CHAPTER 1

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

A.1 Proof of Theorem 1.3.1

The theorem follows immediately upon noting that (1.10) follows from Lemmas A.9.4–A.9.5 below. ■

A.2 Proof of Theorem 1.3.2

The theorem follows immediately upon noting that (1.15) follows from Lemmas A.9.4–A.9.5 and A.9.6 below. ■

A.3 Proof of Theorem 1.3.3

From Lemma A.9.4, we see that it suffices to show that $\hat{\nu}_n^2$ defined in (1.20) tends in probability to (A.16). Since

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

the desired conclusion follows immediately from Lemmas A.9.5–A.9.7. ■

A.4 Proof of Theorem 1.3.4

Let Q satisfying (1.30) be given. For such a Q, we first argue that

$$gZ^{(n)}|X^{(n)} \stackrel{d}{=} Z^{(n)}|X^{(n)} \ . \tag{A.1}$$

Since $\pi = \pi_n(X^{(n)})$, we have from Assumption 1.2.2 that

$$gD^{(n)}|X^{(n)} \stackrel{d}{=} D^{(n)}|X^{(n)} \ . \tag{A.2}$$

Furthermore,

$$Y^{(n)} \perp\!\!\!\perp D^{(n)}|X^{(n)} \ . \tag{A.3}$$

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 (1.1), the second and fourth equalities follow from (1.4), the third follows from the fact that Q satisfies (1.30). It now follows from (A.2) and (A.3) that (A.1) holds.

Next, observe that

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

$$
\begin{aligned}
&= E\left[\sum_{g\in\mathbf{G}_n(\pi)} E\left[\phi_n^{\mathrm{rand}}(Z^{(n)})\middle| X^{(n)}\right]\right] \\
&= E\left[2^n E\left[\phi_n^{\mathrm{rand}}(Z^{(n)})\middle| X^{(n)}\right]\right] \\
&= 2^n E\left[\phi_n^{\mathrm{rand}}(Z^{(n)})\right] \ , \qquad \text{(A.4)}
\end{aligned}
$$

where the first and final equalities follow from the law of iterated expectations, the second follows from (A.1), 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,

$$
E\left[\sum_{g\in\mathbf{G}_n(\pi)} \phi_n^{\mathrm{rand}}(gZ^{(n)})\right] \le 2^n\alpha \ . \qquad \text{(A.5)}
$$

Combining (A.4) and (A.5), we see that (1.31) holds, as desired. ■

A.5 Proof of Theorem 1.3.5

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 (1.20), 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 (A.41). Note further that

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

where

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

The desired conclusion now follows immediately from Lemmas A.9.8–A.9.9 together with Theorem 5.2 of Chung and Romano (2013). ■

A.6 Proof of Theorem 1.4.1

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

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

where the equality exploits the fact that $X_{\pi(2j-1)} \leq X_{\pi(2j)}$, the two inequalities follow by inspection, and the convergence in probability to zero follows from Lemma A.9.1. Similarly,

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

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 A.9.1. Finally, for any $k \in \{2,3\}$ and $\ell \in \{0,1\}$, we have that

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

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 A.9.1. It thus follows that Assumptions 1.2.3–1.2.4 hold.

A.7 Proof of Theorem 1.4.2

We describe an algorithm that leads to a pairing that does not minimize the right-hand side of (1.34) 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 $1 \le i_j \le m$ for all $1 \le j \le k$. Specifically, the k-tuple $(i_1, \ldots, i_k)$ corresponds to the (closed) cube with vertices

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

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

$$f_1((i_1)) = \begin{cases} (1) \mapsto (2) \mapsto \cdots \mapsto (m-1) \mapsto (m) & \text{if} \quad (i_1) = (1) \\ (m) \mapsto (m-1) \mapsto \cdots \mapsto (2) \mapsto (1) & \text{if} \quad (i_1) = (m) \ . \end{cases} \tag{A.6}$$

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

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

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

$$(i_1^{j-1}, \ldots, i_{k-1}^{j-1}, j) \mapsto \cdots \mapsto (i_1^j, \ldots, i_{k-1}^j, j) \ , \tag{A.7}$$

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" j to obtain a "path" of the form (A.7). If, on the other hand, $i_k^0 = m$, then $f_k((i_1^0, \ldots, i_k^0))$ equals

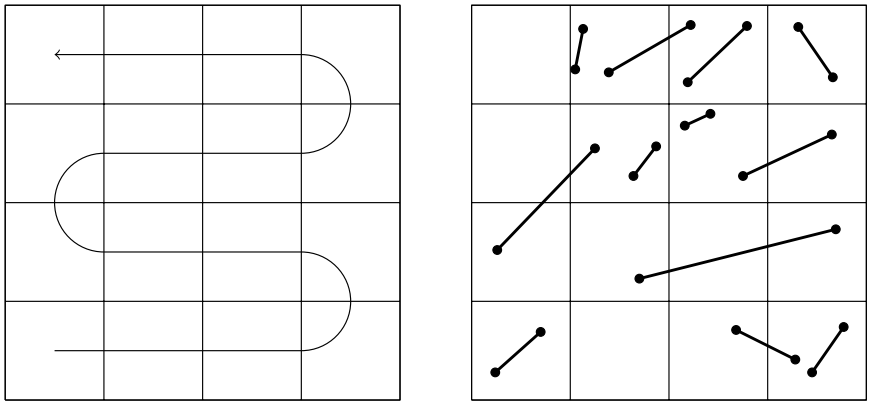
$$\begin{aligned} & (i_1^0, \ldots, i_{k-1}^0, m) \mapsto \cdots \mapsto (i_1^1, \ldots, i_{k-1}^1, m) \\ \mapsto \ & (i_1^1, \ldots, i_{k-1}^1, m-1) \mapsto \cdots \mapsto (i_1^2, \ldots, i_{k-1}^2, m-1) \\ & \qquad \vdots \end{aligned}$$

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

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

$$(i_1^{j-1}, \ldots, i_{k-1}^{j-1}, m-j+1) \mapsto \cdots \mapsto (i_1^{j}, \ldots, i_{k-1}^{j}, m-j+1) \ , \tag{A.8}$$

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 (A.7).



(a) (b)

Figure A.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 A.7.1 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 A.7.1. 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 A.7.1 with $k = 2$, $n = 12$ and $m = 4$.

We now argue that Algorithm A.7.1 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 1.4) is $\frac{2\sqrt{k}}{m}$. Using these facts, the bound in (1.35) now easily follows. Indeed, simply note that the sum that appears on the left-hand side of (1.35) 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. ■

A.8 Proof of Theorem 1.4.3

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 1.4.2 and using (1.34), we see that

$$\frac{2}{n}\sum_{1\le j\le\lfloor\frac{n}{2}\rfloor}|\bar{X}_{\check{\pi}(2j)}-\bar{X}_{\check{\pi}(2j-1)}|^2\xrightarrow{P}0\ . \tag{A.9}$$

Note that

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

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 (1.37) and (1.38), the second equality follows again from (1.38), and the convergence to zero in probability follows from the assumption that π satisfies Assumption 1.2.3 and (A.9). ■

A.9 Auxiliary Results

Lemma A.9.1. *Let $U_i, i=1,\dots,n$ 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

$$\begin{aligned} P\left\{n^{-\frac{1}{r}} \max_{1\le i\le n} |U_i| > \epsilon\right\} &= P\left\{\bigcup_{1\le i\le n} \{|U_i|^r > \epsilon^r n\}\right\} \\ &\le \sum_{1\le i\le n} P\{|U_i|^r > \epsilon^r n\} \\ &\le \frac{1}{n\epsilon^r} \sum_{1\le i\le n} E[|U_i|^r I\{|U_i|^r > \epsilon^r n\}] \\ &= \frac{1}{\epsilon^r} E[|U_i|^r I\{|U_i|^r > \epsilon^r n\}] \\ &\to 0 \end{aligned}$$

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 A.9.2. *For $n \ge 1$, let U_n and V_n be real-valued random variables and $\mathcal{F}_n$ a σ-field. Suppose*

$$P\{U_n \le u|\mathcal{F}_n\} \to \Phi(u/\tau_1) \text{ a.s. }, \tag{A.10}$$

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 (A.10) 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 A.9.3. *Let* $(U_{n,1},\ldots,U_{n,n}) \sim G_n^* = \bigotimes_{1\le i\le n} G_{n,i}$ *with* $\mu(G_{n,i}) = 0$ *for all* $1 \le i \le n$. *Define*

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

If

$$\lim_{\lambda\to\infty}\limsup_{n\to\infty} E_{\bar{G}_n}[|U|I\{|U|>\lambda\}] = 0 \ , \tag{A.11}$$

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

Proof: Define

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

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

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

Furthermore,

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

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

$$P\{|\bar{U}_n - \bar{m}_n| > \epsilon\} \leq \frac{1}{n}\sum_{1\leq i\leq n}\frac{\kappa_{n,i}(n)}{\epsilon^2} + \frac{1}{n}\sum_{1\leq i\leq n}\tau_{n,i}(n) \ . \tag{A.12}$$

Since

$$tP\{|U_{n,i}| > t\} \leq E[|U_{n,i}|I\{|U_{n,i}| > t\}] \ , \tag{A.13}$$

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 (A.12) tends to zero, it therefore suffices to argue that

$$\frac{1}{n}\sum_{1\le i\le n}\frac{1}{n}\int_0^n \tau_{n,i}(x)dx \to 0 \ . \tag{A.14}$$

To this end, note that (A.13) implies that

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

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 $x \le \lambda_0$ and $n > n_0$, we have that

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

It follows that

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

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, (A.14) follows. To complete the proof, note that

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

which tends to zero by assumption. ■

Lemma A.9.4. *If Assumptions 1.2.1–1.2.3 hold, then*

$$\sqrt{n}(\hat{\Delta}_n - \Delta(Q)) \xrightarrow{d} N(0,\nu^2) \ , \tag{A.15}$$

where

$$\begin{aligned}\nu^2 &= E[\,Var[Y_i(1)|X_i]] + E[\,Var[Y_i(0)|X_i]] \\ &\quad + \frac{1}{2}E\left[((E[Y_i(1)|X_i] - E[Y_i(1)]) - (E[Y_i(0)|X_i] - E[Y_i(0)]))^2\right] \qquad \text{(A.16)} \\ &= Var[Y_i(1)] + Var[Y_i(0)] - \frac{1}{2}E\left[((E[Y_i(1)|X_i] - E[Y_i(1)]) + (E[Y_i(0)|X_i] - E[Y_i(0)]))^2\right]\end{aligned}$$

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

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

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

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

and note that

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

where the first equality follows from Assumption 1.2.2 and the second follows from the fact that $Q_n = Q^n$. It follows that

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

Using Assumption 1.2.1(b), we have that

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

Note further that

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

where the first inequality follows by inspection, the second inequality exploits Assumption 1.2.1(c) and the convergence to zero follows from Assumption 1.2.3. Hence,

$$\frac{s_n^2}{n} \xrightarrow{P} E[\text{Var}[Y_i(1)|X_i]] > 0 \ , \tag{A.17}$$

where the final inequality exploits Assumption 1.2.1(a). Next, we argue for any $\epsilon > 0$ that

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

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

$$P\{\epsilon s_n > m\} \to 1 \ . \tag{A.18}$$

Note further that Assumption 1.2.2 implies that

$$E[Y_i(1)D_i|X^{(n)}, D^{(n)}] = D_i E[Y_i(1)|X_i] \ , \tag{A.19}$$

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

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

where the first inequality follows by inspection, the second inequality exploits (A.17)–(A.18), the equality follows from Assumption 1.2.2 and the fact that $Q_n = Q^n$, and the convergence in probability follows from (A.17) and the fact that Assumption 1.2.1(b) implies

$$
E[|Y_i(1) - E[Y_i(1)|X_i]|^2] = E[\mathrm{Var}[Y_i(1)|X_i]] \le E[Y_i^2(1)] < \infty \ . \tag{A.20}
$$

Note further that (A.20) implies that

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

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 A.9.5 below that

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

A similar argument establishes that

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

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

$$\sup_{t\in\mathbf{R}}\left|P\{A_n - B_n \le t|X^{(n)}, D^{(n)}\} - \Phi(t/\sqrt{E[\mathrm{Var}[Y_i(0)|X_i]] + E[\mathrm{Var}[Y_i(0)|X_i]]})\right| \xrightarrow{P} 0 \ . \tag{A.21}$$

To analyze C_n, first note that (A.19) implies that

$$C_n = \frac{1}{\sqrt{n}} \sum_{1\le i\le 2n} D_i \left(E[Y_i(1)|X_i] - E[Y_i(1)]\right) \ , \tag{A.22}$$

so

$$E[C_n|X^{(n)}] = \frac{1}{2\sqrt{n}} \sum_{1\le i\le 2n} \left(E[Y_i(1)|X_i] - E[Y_i(1)]\right) \ . \tag{A.23}$$

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 (A.22)–(A.23), 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 1.2.1(c), and the convergence in probability follows from Assumption 1.2.3. For any $\epsilon > 0$, it thus follows from Chebychev's inequality that

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

Since probabilities are bounded, we have further that

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

Hence,

$$C_n = \frac{1}{2\sqrt{n}} \sum_{1 \leq i \leq 2n} (E[Y_i(1)|X_i] - E[Y_i(1)]) + o_P(1) \ . \tag{A.24}$$

A similar argument establishes that

$$D_n = \frac{1}{2\sqrt{n}} \sum_{1 \leq i \leq 2n} (E[Y_i(0)|X_i] - E[Y_i(0)]) + o_P(1) \ . \tag{A.25}$$

Hence,

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

where the first equality follows from (A.24)–(A.25), the second equality follows by inspection, and the convergence in distribution follows from Slutsky's theorem and the Central Limit Theorem.

The desired conclusion (A.15) now follows by a subsequencing argument. To see this, suppose by way of contradiction that (A.15) fails. This implies that there exists $\delta > 0$ and a subsequence n_k along which

$$\sup_{t \in \mathbf{R}} |P\{\sqrt{n_k}(\hat{\Delta}_{n_k} - \Delta(Q)) \leq t\} - \Phi(t/\nu)| \to \delta \ . \tag{A.26}$$

By considering a further subsequence if necessary, which, by an abuse of notation, we continue to denote by n_k, it follows from (A.21) that

$$A_{n_k} - B_{n_k} \xrightarrow{d} N(0, E[\mathrm{Var}[Y_i(0)|X_i]] + E[\mathrm{Var}[Y_i(0)|X_i]])$$

w.p.1 conditional on $X^{(n_k)}$ and $D^{(n_k)}$. Since $C_{n_k} - D_{n_k}$ is constant conditional on $X^{(n_k)}$ and $D^{(n_k)}$, Lemma A.9.2 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 (A.26).

Finally, in order to complete the proof, note that

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

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

Lemma A.9.5. *If Assumptions 1.2.1–1.2.3 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* (1.5) *and* (1.6)*, 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

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

Next, note that

$$\begin{aligned}
& \frac{1}{n} \sum_{1 \leq i \leq 2n} E[Y_i(1) D_i | X^{(n)}, D^{(n)}] \\
= \; & \frac{1}{n} \sum_{1 \leq i \leq 2n} D_i E[Y_i(1) | X_i] \\
= \; & \frac{1}{n} \sum_{1 \leq i \leq 2n : D_i = 1} E[Y_i(1) | X_i] \\
= \; & \frac{1}{2n} \sum_{1 \leq i \leq 2n} E[Y_i(1) | X_i] \\
& \quad + \left(\frac{1}{2n} \sum_{1 \leq i \leq 2n : D_i = 1} E[Y_i(1) | X_i] - \frac{1}{2n} \sum_{1 \leq i \leq 2n : D_i = 0} E[Y_i(1) | X_i] \right) ,
\end{aligned}$$

where the first equality exploits (A.19) and the second and third equalities follow by inspection. Note further that

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

where the first inequality follows by inspection, the second exploits Assumption 1.2.1(c) and the convergence in probability follows from Assumption 1.2.3. Since Assumption 1.2.1(b) implies that $E[|E[Y_i(1)|X_i]|] \leq E[|Y_i(1)|] < \infty$, it follows that

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

To complete the argument, we argue that

$$\frac{1}{n}\sum_{1\leq i\leq 2n}\Big(Y_i(1)D_i - E[Y_i(1)D_i|X^{(n)},D^{(n)}]\Big)\xrightarrow{P}0\ . \tag{A.27}$$

For this purpose, we proceed by verifying that (A.11) in Lemma A.9.3 holds in probability conditional on $X^{(n)}$ and $D^{(n)}$. To that end, note for any $m>0$ that

$$\begin{aligned}
&\frac{1}{2n}\sum_{1\leq i\leq 2n} E[|Y_i(1)D_i - E[Y_i(1)D_i|X^{(n)},D^{(n)}]|\\
&\quad I\{|Y_i(1)D_i - E[Y_i(1)D_i|X^{(n)},D^{(n)}]|>m\}|X^{(n)},D^{(n)}]\\
&=\frac{1}{2n}\sum_{1\leq i\leq 2n} E[|Y_i(1)D_i - D_iE[Y_i(1)|X_i]|I\{|Y_i(1)D_i - D_iE[Y_i(1)|X_i]|>m\}|X^{(n)},D^{(n)}]\\
&\leq\frac{1}{2n}\sum_{1\leq i\leq 2n} E[|Y_i(1) - E[Y_i(1)|X_i]|I\{|Y_i(1) - E[Y_i(1)|X_i]|>m\}|X^{(n)},D^{(n)}]\\
&=\frac{1}{2n}\sum_{1\leq i\leq 2n} E[|Y_i(1) - E[Y_i(1)|X_i]|I\{|Y_i(1) - E[Y_i(1)|X_i]|>m\}|X_i]\\
&\xrightarrow{P} E[|Y_i(1) - E[Y_i(1)|X_i]|I\{|Y_i(1) - E[Y_i(1)|X_i]|>m\}]\ ,
\end{aligned}\tag{A.28}$$

where the first and fourth equalities follow from (A.19), the inequality follows by inspection, and the convergence in probability follows from (A.20). The desired conclusion (A.27) now follows by a subsequencing argument. To see this, suppose by way of contradiction that (A.27) fails. This implies that there exists $\epsilon>0$, $\delta>0$ and a subsequence n_k along which

$$P\left\{\left|\frac{1}{n_k}\sum_{1\leq i\leq 2n_k}\Big(Y_i(1)D_i - E[Y_i(1)D_i|X^{(n_k)},D^{(n_k)}]\Big)\right|>\epsilon\right\}\to\delta\ . \tag{A.29}$$

By considering a further subsequence if necessary, which, by an abuse of notation, we continue to denote by n_k, it follows from (A.19), (A.20) and (A.28) that

$$\lim_{m\to\infty}\limsup_{k\to\infty}\frac{1}{2n}\sum_{1\leq i\leq 2n}\Big(E[|Y_i(1)D_i - E[Y_i(1)D_i|X^{(n)},D^{(n)}]|$$

$$\times\, I\{|Y_i(1)D_i - E[Y_i(1)D_i|X^{(n)}, D^{(n)}]| > m\}|X^{(n)}, D^{(n)}]\Big) = 0$$

w.p.1 (conditional on $X^{(n_k)}$ and $D^{(n_k)}$). Lemma A.9.3 implies, however, that

$$\frac{1}{n_k}\sum_{1\le i\le 2n_k}\Big(Y_i(1)D_i - E[Y_i(1)D_i|X^{(n_k)}, D^{(n_k)}]\Big) \to 0$$

w.p.1 conditional on $X^{(n_k)}$ and $D^{(n_k)}$, which implies a contradiction to (A.29). ■

Lemma A.9.6. *If Assumptions 1.2.1–1.2.3 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* (1.21).

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 A.9.5 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

$$E\left[Y_{\pi(2j)}Y_{\pi(2j-1)}\middle|X^{(n)}\right] = \frac{1}{2}\mu_1(X_{\pi(2j)})\mu_0(X_{\pi(2j-1)}) + \frac{1}{2}\mu_0(X_{\pi(2j)})\mu_1(X_{\pi(2j-1)}) \ , \quad \text{(A.30)}$$

so

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

where the second equality follows from (A.30) and the other equalities follow by inspection. Assumption 1.2.3 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

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

To complete the argument, we show that

$$\frac{1}{n}\sum_{1\leq j\leq n}\left(Y_{\pi(2j)}Y_{\pi(2j-1)} - E\left[Y_{\pi(2j)}Y_{\pi(2j-1)}\middle| X^{(n)}\right]\right) \xrightarrow{P} 0 \ . \tag{A.31}$$

For this purpose, we proceed by verifying that (A.11) in Lemma A.9.3 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 $\lambda > 0$:

$$\begin{aligned} |a+b|I\{|a+b|>\lambda\} &\leq 2|a|I\left\{|a|>\frac{\lambda}{2}\right\} + 2|b|I\left\{|b|>\frac{\lambda}{2}\right\} &\text{(A.32)}\\ |ab|I\{|ab|>\lambda\} &\leq a^2 I\{|a|>\sqrt{\lambda}\} + b^2 I\{|b|>\sqrt{\lambda}\} \ . &\text{(A.33)}\end{aligned}$$

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

$$\begin{aligned}
&\frac{1}{n}\sum_{1\leq j\leq n} E\left[\left|Y_{\pi(2j)}Y_{\pi(2j-1)} - E\left[Y_{\pi(2j)}Y_{\pi(2j-1)}\middle| X^{(n)}\right]\right| \right.\\
&\qquad \left. I\left\{\left|Y_{\pi(2j)}Y_{\pi(2j-1)} - E\left[Y_{\pi(2j)}Y_{\pi(2j-1)}\middle| X^{(n)}\right]\right| > \lambda\right\}\middle| X^{(n)}\right]\\
&\lesssim \frac{1}{n}\sum_{1\leq j\leq n} E\left[\left|Y_{\pi(2j)}Y_{\pi(2j-1)}\right| I\left\{\left|Y_{\pi(2j)}Y_{\pi(2j-1)}\right| > \frac{\lambda}{2}\right\}\middle| X^{(n)}\right]\\
&\qquad + \frac{1}{n}\sum_{1\leq j\leq n} E\left[\left|E\left[Y_{\pi(2j)}Y_{\pi(2j-1)}\middle| X^{(n)}\right]\right| I\left\{\left|E\left[Y_{\pi(2j)}Y_{\pi(2j-1)}\middle| X^{(n)}\right]\right| > \frac{\lambda}{2}\right\}\middle| X^{(n)}\right]\\
&\leq \frac{1}{n}\sum_{1\leq j\leq n} E\left[Y^2_{\pi(2j)} I\left\{\left|Y_{\pi(2j)}\right| > \sqrt{\frac{\lambda}{2}}\right\}\middle| X^{(n)}\right]\\
&\qquad + \frac{1}{n}\sum_{1\leq j\leq n} E\left[Y^2_{\pi(2j-1)} I\left\{\left|Y_{\pi(2j-1)}\right| > \sqrt{\frac{\lambda}{2}}\right\}\middle| X^{(n)}\right]
\end{aligned}$$

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

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

where the third inequality exploits (A.30). Since $E[Y_i^2(d)] < \infty$ and $E[\mu_d^2(X_i)] \le E[Y_i^2(d)]$, we have that

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

It now follows from a subsequencing argument as in the proof of Lemma A.9.5 that (A.31) 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 A.9.7. *If Assumptions 1.2.1–1.2.4 hold, then*

$$\hat{\lambda}_n^2 \xrightarrow{P} E[(E[Y_i(1)|X_i] - E[Y_i(0)|X_i])^2] \ , \tag{A.34}$$

where $\hat{\lambda}_n$ *is defined in* (1.22).

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

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

Hence, in order to show that

$$E[\hat{\lambda}_n^2|X^{(n)}] \xrightarrow{P} E[(\mu_1(X_i) - \mu_0(X_i))^2] \ , \tag{A.35}$$

it suffices to show that

$$\frac{1}{2n}\sum_{1\le j\le \frac{n}{2}}\sum_{k\in\{2,3\},\ell\in\{0,1\}} \mu_1(X_{\pi(4j-k)})\mu_1(X_{\pi(4j-\ell)}) \xrightarrow{P} E[\mu_1^2(X_i)] \tag{A.36}$$

$$\frac{1}{2n}\sum_{1\le j\le \frac{n}{2}}\sum_{k\in\{2,3\},\ell\in\{0,1\}}\mu_0(X_{\pi(4j-k)})\mu_0(X_{\pi(4j-\ell)})\xrightarrow{P}E[\mu_0^2(X_i)] \tag{A.37}$$

$$\frac{1}{2n}\sum_{1\le j\le \frac{n}{2}}\sum_{k\in\{2,3\},\ell\in\{0,1\}}\Big(\mu_0(X_{\pi(4j-k)})\mu_1(X_{\pi(4j-\ell)})+\mu_1(X_{\pi(4j-k)})\mu_0(X_{\pi(4j-\ell)})\Big)$$
$$\xrightarrow{P}2E[\mu_1(X_i)\mu_0(X_i)]\ . \tag{A.38}$$

We first prove (A.36). To see this, note that

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

so

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

It follows that

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

But, Assumption 1.2.1 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 1.2.4. 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 (A.36) holds. Similar arguments may be used to establish (A.37)-(A.38), from which (A.35) 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 (A.11) in Lemma A.9.3 holds in probability conditionally on $X^{(n)}$, which may be accomplished by repeated application of (A.32) and (A.33), as in the proof of Lemma A.9.6. ■

Lemma A.9.8. *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 1.2.1–1.2.3 hold, then

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

where

$$\tau^2 = E[\, Var[Y_i(1)|X_i]] + E[\, Var[Y_i(0)|X_i]] + E\left[(E[Y_i(1)|X_i] - E[Y_i(0)|X_i])^2 \right] \ . \quad \text{(A.39)}$$

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

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

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

$$\frac{\max_{1 \leq j \leq n} C_{n,j}^2}{\sum_{1 \leq j \leq n} C_{n,j}^2} \xrightarrow{P} 0 . \tag{A.40}$$

From Lemma A.9.6, we see that $\frac{1}{n} \sum_{1 \leq j \leq n} C_{n,j}^2 = \hat{\tau}_n^2 \xrightarrow{P} \tau^2 > 0$, where the inequality exploits Assumption 1.2.1(a). Furthermore,

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

where the first inequality follows by exploiting the fact that $|a-b|^2 \leq 2(a^2+b^2)$ 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 A.9.1 and Assumption 1.2.1(b). Hence, (A.40) holds, from which the desired conclusion now follows easily by appealing to the aforementioned lemma and Polya's theorem. ■

Lemma A.9.9. *Let*

$$\check{\nu}_n^2(\epsilon_1, \ldots, \epsilon_n) = \hat{\tau}_n^2 - \frac{1}{2}(\check{\lambda}_n^2(\epsilon_1, \ldots, \epsilon_n) + \check{\Delta}_n^2(\epsilon_1, \ldots, \epsilon_n)) , \tag{A.41}$$

where $\hat{\tau}_n^2$ *is defined in* (1.21),

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

and, independently of $W^{(n)}$, $\epsilon_j, j = 1, \ldots, n$ *are i.i.d. Rademacher random variables. If Assumptions 1.2.1–1.2.3 hold, then*

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

where τ^2 *is defined in* (A.39).

Proof: From Lemma A.9.6, we see that $\hat{\tau}_n^2 \xrightarrow{P} \tau^2$. From Lemma A.9.8, 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

$$\frac{2}{n} \sum_{1 \leq j \leq \lfloor \frac{n}{2} \rfloor} \epsilon_{2j-1}\epsilon_{2j} Y_{\pi(4j-k)} Y_{\pi(4j-\ell)} D_{\pi(4j-k')} D_{\pi(4j-\ell')} \ . \tag{A.42}$$

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 (A.11) in Lemma A.9.3 holds in probability conditional on $W^{(n)}$. To this end, note that

$$\frac{2}{n} \sum_{1 \leq j \leq \lfloor \frac{n}{2} \rfloor} E\Big[|\epsilon_{2j-1}\epsilon_{2j} Y_{\pi(4j-k)} Y_{\pi(4j-\ell)} D_{\pi(4j-k')} D_{\pi(4j-\ell')}|$$

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

where the first inequality follows from the fact that $|\epsilon_j|=1$ for all $1\leq j\leq n$ and $|D_i|\leq 1$ for all $1\leq i\leq 2n$, 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 (A.33) used in the proof of Lemma A.9.6, the fourth inequality follows by inspection, the fifty inequality uses the fact that $Y_i^2\leq Y_i^2(1)+Y_i^2(0)$, and the convergence in probability follows from Assumption 1.2.1(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 A.9.5 that (A.42) tends to zero in probability. The desired result thus follows. ■

APPENDIX B

APPENDIX FOR CHAPTER 2

B.1 Proof of Main Results

For the rest of the appendix we introduce the following definition for the convex combination of matched-pair designs.

Definition B.1.1. For $\lambda, \lambda' \in \Lambda_n^{\text{pair}}$ and $\delta \in [0,1]$, define $\delta\lambda \bigoplus (1-\delta)\lambda'$ as the randomization between λ and λ' such that λ is implemented with probability δ. Define the convex hull formed by all convex combinations of any matched-pair designs as

$$\text{co}(\Lambda_n^{\text{pair}}) = \left\{ \bigoplus_{1\le j\le J} \delta_j \lambda^j : \lambda^j \in \Lambda_n^{\text{pair}} \text{ and } \delta_j \ge 0 \text{ for } 1 \le j \le J, \sum_{1\le j\le J} \delta_j = 1, 1 \le J < \infty \right\}. \tag{B.1}$$

B.1.1 Proof of Theorem 2.3.1

Define $V(\lambda)$ as the objective in (2.18) multiplied by n^2. We have

$$\begin{aligned}
V(\lambda) &= n^2 \operatorname{Var}_\lambda[E[\hat{\theta}_n | X^{(n)}, D^{(n)}] | X^{(n)}] \\
&= \operatorname{Var}_\lambda \left[\sum_{1\le i\le 2n} [D_i E[Y_i(1)|X_i] - (1-D_i)E[Y_i(0)|X_i]] \middle| X^{(n)} \right] \\
&= \operatorname{Var}_\lambda \left[\sum_{1\le i\le 2n} D_i (E[Y_i(0)|X_i] + E[Y_i(1)|X_i]) \middle| X^{(n)} \right] \\
&= (g^{(n)})' \operatorname{Var}_\lambda[D^{(n)}] g^{(n)} \ .
\end{aligned}$$

Recall from Section 2.2 that Λ_n^{pair} is the set of all matched-pair designs. For any $\lambda = \{\{\pi(1), \pi(2)\}, \ldots, \{\pi(2n-1), \pi(2n)\}\} \in \Lambda_n^{\text{pair}}$,

$$V(\lambda) = \frac{1}{4} \sum_{1 \le s \le n} \left(g_{\pi(2s-1)} - g_{\pi(2s)}\right)^2 . \tag{B.2}$$

By Lemma B.3.2, we have $V(\lambda^g(X^{(n)})) \le V(\lambda)$.

Recall the definition of convex combinations of matched-pair designs from Definition B.1.1. To conclude the proof, note that by Lemma B.3.1, for any $\lambda \in \Lambda$ we have

$$\lambda = \bigoplus_{1 \le j \le J} \delta_j \lambda^j ,$$

where $\lambda^j \in \Lambda_n^{\text{pair}}$ and $\delta_j \ge 0$ for $1 \le j \le J$, $\sum_{1 \le j \le J} \delta_j = 1$, and $1 \le J < \infty$. Then,

$$\text{MSE}(\lambda | X^{(n)}) = \sum_{1 \le j \le J} \delta_j \, \text{MSE}(\lambda^j | X^{(n)}) \ge \min_{1 \le j \le J} \text{MSE}(\lambda^j | X^{(n)}) \ge \text{MSE}(\lambda^g(X^{(n)}) | X^{(n)}) ,$$

where the last inequality follows because $\lambda^g(X^{(n)})$ minimizes $\text{MSE}(\lambda | X^{(n)})$ over Λ_n^{pair}. The theorem therefore follows. ■

B.1.2 Proof of Theorem 2.5.3

First, note that the assumptions in Lemma B.3.4 hold because of Lemma B.3.7 and Assumption 2.5.4, and that $\hat{g}_m$ is a fixed function conditional on $\tilde{W}^{(m)}$. Hence, by Lemma B.3.4 with $\tau = \frac{1}{2}$, with probability one for $\tilde{W}^{(m)}$, as $n \to \infty$,

$$\sup_{t \in \mathbf{R}} \left| Q\{\sqrt{n}(\hat{\theta}_n - \theta(Q)) \le t | \tilde{W}^{(m)}\} - \Phi(z/\varsigma_{\hat{g}_m}) \right| \to 0 , \tag{B.3}$$

where

$$\varsigma^2_{\hat{g}_m} = \mathrm{Var}[Y_i(1)] + \mathrm{Var}[Y_i(0)] - \frac{1}{2}E[(E[Y_i(1) + Y_i(0)|\hat{g}_m(X_i), \tilde{W}^{(m)}] - E[Y_i(1) + Y_i(0)])^2] \ . \tag{B.4}$$

On the other hand, note that the assumptions in Lemma B.3.5 hold because of Lemma B.3.7 and Assumption 2.5.4, and that $\hat{g}_m$ is a fixed function conditional on $\tilde{W}^{(m)}$. Hence, by Lemma B.3.5 with $\tau = \frac{1}{2}$, with probability one for $\tilde{W}^{(m)}$, for all $\epsilon > 0$, as $n \to \infty$,

$$Q\{|(\hat{\varsigma}^{\hat{g}_m}_n)^2 - \varsigma^2_{\hat{g}_m}| > \epsilon | \tilde{W}^{(m)}\} \to 0 \ . \tag{B.5}$$

The conditional convergence in (2.40) follows immediately from (B.3) and (B.5). Since the conditional convergence holds with probability one for $\tilde{W}^{(m)}$, and $\phi^{\hat{g}_m}_n(W^{(n)}) \in [0, 1]$, the unconditional convergence follows from the dominated convergence theorem. ■

B.1.3 Proof of Theorem 2.5.1

The first assertion follows from Lemma B.3.4 with $h = g$ and $\tau = \frac{1}{2}$. We now show that under $\lambda^{\hat{g}_m}(X^{(n)})$ defined in (2.23), $\sqrt{n}(\hat{\theta}_n - \theta(Q)) \xrightarrow{d} N(0, \varsigma^2_g)$ for ς^2_g defined in (2.34) as $m, n \to \infty$. By repeating arguments in the proof of Lemma B.3.4, we write

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

where

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

$$D_n = \frac{1}{\sqrt{n}} \sum_{1 \le i \le 2n} (E[Y_i(0)|g^{(n)}, D^{(n)}] - E[Y_i(0)]) \ .$$

Note that unlike in Lemma B.3.4, the quantities above are conditioned on $g^{(n)}$ for g defined in (2.19), instead of $\hat{g}_m^{(n)}$. Note that by Assumptions 2.5.2(c), 2.5.3, and Lemma B.3.8,

$$\frac{1}{n} \sum_{1 \le s \le n} (g_{\pi \hat{g}m}(2s-1) - g_{\pi \hat{g}m}(2s))^2 \xrightarrow{P} 0 \ . \tag{B.6}$$

Since Assumption 2.5.2(a)–(b) and (B.6) hold, by repeating arguments in the proof of Lemma B.3.4 with $\tau = \frac{1}{2}$, it is straightforward to establish that as $m, n \to \infty$,

$$\sqrt{n}(\hat{\theta}_n - \theta(Q)) \xrightarrow{d} N(0, \varsigma_g^2) \ , \tag{B.7}$$

Note that (B.6) is enough to derive the asymptotic representation for C_n so that we need not impose Lipschitz conditions on $E[Y_i(d)|g(X_i)]$. ■

B.1.4 Proof of Theorem 2.5.2

In light of Theorem 2.5.1, we only need to show that $(\hat{\varsigma}_n^{\hat{g}m})^2 \xrightarrow{P} \varsigma_g^2$ as $m, n \to \infty$. Similar arguments as those used in Lemma B.3.5 go through if (B.6) holds and

$$\frac{1}{n} \sum_{1 \le j \le \lfloor \frac{n}{2} \rfloor} |g_{\pi \hat{g}m}(4j-k) - g_{\pi \hat{g}m}(4j-l)|^2 \xrightarrow{P} 0 \tag{B.8}$$

for $k \in \{2, 3\}$ and $l \in \{0, 1\}$. Since (B.8) follows from Assumptions 2.5.3 by Lemma B.3.8, the proof is concluded.

B.1.5 Proof of Theorem 2.5.5

To begin with, note that we need only establish that as $m, n \to \infty$,

$$\sqrt{m+2n}(\hat{\theta}_n^{\text{combined}} - \theta(Q)) \xrightarrow{d} N(0, \nu\varsigma_{\text{pilot}}^2 + (1-\nu)2\varsigma^2) \ , \tag{B.9}$$

and the rest follows from Slutsky's lemma. We prove (B.9) by contradiction. Suppose (B.9) does not hold. Then, there exists a subsequence still denoted by $\{m, n\}$ for notational simplicity, along which as $m, n \to \infty$,

$$\sup_{t \in \mathbf{R}} \left| \sqrt{m+2n}(\hat{\theta}_n^{\text{combined}} - \theta(Q)) - \Phi(z/\sqrt{\nu\varsigma_{\text{pilot}}^2 + (1-\nu)2\varsigma^2}) \right| \to c \ , \tag{B.10}$$

where $c > 0$, and

$$\frac{m}{m+2n} \to \nu \in [0,1] \ .$$

Now consider this subsequence. Since the two convergences in the Lemma B.3.8 hold in probability, there exists a further subsequence along which they hold with probability one. By repeating the proof of Theorem 2.5.2, we could see that along this subsequence, as $m, n \to \infty$, with probability one for $\tilde{W}^{(m)}$,

$$\sup_{t \in \mathbf{R}} \left| Q\{\sqrt{n}(\hat{\theta}_n - \theta(Q)) \le t | \tilde{W}^{(m)}\} - \Phi(z/\varsigma_g) \right| \to 0 \ . \tag{B.11}$$

Along the subsequence we construct, since $\frac{m}{m+2n} \to \nu$, by (B.11), Slutsky's lemma and Lemma B.3.3,

$$\sqrt{m+2n}(\hat{\theta}_n^{\text{combined}} - \theta(Q)) \xrightarrow{d} N(0, \nu\varsigma_{\text{pilot}}^2 + (1-\nu)2\varsigma^2) \ ,$$

which is a contradiction to (B.10). The theorem therefore holds. ■

B.1.6 Proof of Theorem 2.5.4

Follows from Theorem A.7.1 and by repeating arguments in the proof of Lemma B.3.4. ■

B.2 Supplementary Results

The next theorem shows that the infeasible optimal stratification has a similar structure to (2.20) when $\tau \neq \frac{1}{2}$.

Theorem B.2.1. *Suppose the sample size is kn for $k \in \mathbb{Z}$ and the treatment assignment scheme satisfies $\tau_s \equiv \tau = \frac{l}{k}$, where $l \in \mathbb{Z}$, $0 < l < k$, and k and l are relatively prime. Then, (2.5) is solved by $\lambda^{\tau,g}$ defined in (2.22), where $g^{\tau}_{\pi^{\tau},g^{\tau}}(1) \leq \ldots \leq g^{\tau}_{\pi^{\tau},g^{\tau}}(kn)$ for g^{τ} defined in (2.21).*

Proof of Theorem B.2.1. First, note that

$$\hat{\theta}_n = \frac{1}{kn} \sum_{1 \leq i \leq kn} \left(\frac{1}{\tau} Y_i(1) D_i - \frac{1}{1-\tau} Y_i(0)(1 - D_i) \right) .$$

Next,

$$\text{MSE}(\lambda | X^{(n)}) = (E_{\lambda}[\hat{\theta}_n | X^{(n)}] - \theta(Q))^2 + \text{Var}_{\lambda}[\hat{\theta}_n | X^{(n)}] .$$

By repeating arguments in the proof of Lemma 2.3.1,

$$E_{\lambda}[\hat{\theta}_n | X^{(n)}] - \theta(Q) = \frac{1}{kn} \sum_{1 \leq i \leq kn} (E[Y_i(1)|X_i] - E[Y_i(0)|X_i]) - \theta(Q) ,$$

identical across all $\lambda \in \Lambda_n$, so that we need only consider conditional variances of $\hat{\theta}$ given $X^{(n)}$ which could be decomposed as in (2.11). By repeating arguments in the proof of Lemma 2.3.1, for any $\lambda \in \Lambda_n$, the first term of the right-hand side of (2.11) equals

$$\frac{1}{k^2 n^2} \sum_{1 \leq i \leq kn} \left(\frac{\text{Var}[Y_i(1)|X_i]}{\tau} + \frac{\text{Var}[Y_i(0)|X_i]}{1-\tau} \right) ,$$

again identical across all $\lambda \in \Lambda_n$. Therefore, we need only consider

$$\mathrm{Var}_\lambda[E[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] \ .$$

By repeating arguments in the proof of Lemma B.3.1, a stratum of size kl where $l > 1$ is a convex combination of stratifications with strata only of size k. We could therefore focus on the case where each stratum is of size k. For any stratification of the form $\lambda = \{\{\pi((s-1)k+1, \ldots \pi(sk)\} : 1 \leq s \leq n\}$,

$$\mathrm{Var}_\lambda[E[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] \propto \sum_{1 \leq s \leq n} \sum_{(s-1)k+1 \leq j \leq sk} (g^\tau_{\pi(j)} - \bar{g}^\tau_s)^2 \ ,$$

where g^τ_i is defined in (2.21) and

$$\bar{g}^\tau_s = \frac{1}{k} \sum_{(s-1)k+1 \leq j \leq sk} g^\tau_{\pi(j)} \ .$$

To see this, first note that units are independent across strata, so that by repeating arguments in the proof of Lemma 2.3.1,

$$\mathrm{Var}_\lambda[E[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] \propto \sum_{1 \leq s \leq n} \mathrm{Var}_\lambda \left[\sum_{(s-1)k+1 \leq j \leq sk} g^\tau_{\pi(j)} D_{\pi(j)} \right] \ .$$

Next,

$$\begin{aligned}
&\mathrm{Var}_\lambda \left[\sum_{(s-1)k+1 \leq j \leq sk} g^\tau_{\pi(j)} D_{\pi(j)} \right] \\
&= \frac{1}{\binom{k}{l}} \sum_{(s-1)k+1 \leq j_1 < \ldots < j_l \leq sk} \left(\sum_{1 \leq \iota \leq l} g^\tau_{\pi(j_\iota)} - l\bar{g}^\tau_s \right)^2 \\
&= \frac{l}{k} \sum_{(s-1)k+1 \leq j \leq sk} (g^\tau_{\pi(j)} - \bar{g}^\tau_s)^2
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{\binom{k}{l}} \sum_{(s-1)k+1 \le j_1 < \ldots < j_l \le sk} \sum_{1 \le \iota_1 \ne \iota_2 \le l} (g_{\pi(j_{\iota_1})} - \bar{g}_s^\tau)(g_{\pi(j_{\iota_2})} - \bar{g}_s^\tau) \\
&= \frac{l}{k} \sum_{(s-1)k+1 \le j \le sk} (g_{\pi(j)}^\tau - \bar{g}_s^\tau)^2 \\
&\quad + \frac{\binom{k-2}{l-2}}{\binom{k}{l}} \left[\left(\sum_{(s-1)k+1 \le j \le sk} g_{\pi(j)}^\tau - k\bar{g}_s^\tau \right)^2 - \sum_{(s-1)k+1 \le j \le sk} (g_{\pi(j)}^\tau - \bar{g}_s^\tau)^2 \right] \\
&\propto \sum_{(s-1)k+1 \le j \le sk} (g_{\pi(j)}^\tau - \bar{g}_s^\tau)^2 \ ,
\end{aligned}
$$

where the first equality holds by definition, the second holds by expanding the square, the third holds by accounting for cross product terms, and the fourth holds because the first term inside the square bracket on the fourth line is 0. Therefore, the problem is reduced to optimal univariate clustering of kn units on the real line where each cluster is of size k, and the conclusion follows by arguing similarly to in the proof of Lemma B.3.2. ■

For a measurable function $h : \mathbf{R}^p \to \mathbf{R}$, let π^h be a permutation of $\{1, \ldots, kn\}$ such that $h_{\pi^{\tau,h}(1)} \le \ldots \le h_{\pi^{\tau,h}(kn)}$. Define

$$
\lambda^{\tau,h}(X^{(n)}) = \{\{\pi^{\tau,h}((s-1)k+1), \ldots \pi^{\tau,h}(sk)\} : 1 \le s \le n\} \ . \tag{B.12}
$$

Further define $\bar{h}_s^\tau = \frac{1}{k} \sum_{(s-1)k+1 \le j \le sk} h_{\pi^{\tau,h}(j)}$.

Assumption B.2.1. h satisfies

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

(b) $E[Y_i^r(d)|h(X_i) = z]$ is Lipschitz for $r = 1, 2$ and $d = 0, 1$.

(c) $\frac{1}{n} \sum_{1 \le s \le n} \sum_{(s-1)k+1 \le j \le sk} |h_{\pi^{\tau,h}(j)} - \bar{h}_s^\tau|^2 \xrightarrow{P} 0$.

The next theorem is the limiting counterpart to Theorems 2.3.1 and B.2.1. It shows that across all stratifications defined by (B.12) for h satisfying Assumption B.2.1, the asymptotic variance of $\hat{\theta}_n$ is minimized by choosing $h = g^\tau$ defined in (2.21).

Theorem B.2.2. *Suppose $h : \mathbf{R}^p \to \mathbf{R}$ be a measurable function that satisfies Assumption B.2.1. Then,*

$$\varsigma^2_{\tau,g^\tau} \leq \varsigma^2_{\tau,h} \ ,$$

for $\varsigma^2_{\tau,g^\tau}$ and $\varsigma^2_{\tau,h}$ defined in (B.18) *and g^τ defined in* (2.21)*. Moreover, the inequality is strict unless $E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right] = g^\tau(X_i)$ with probability one under Q.*

Proof of Theorem B.2.2. By the definition of $\varsigma^2_{\tau,h}$ in (B.18), minimizing $\varsigma^2_{\tau,h}$ with respect to h is equivalent to maximizing

$$E\left[\left(E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right] - \left(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\right)\right)^2\right] \ .$$

Next, note that

$$\begin{aligned}
&E\left[\left(g^\tau(X_i) - \left(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\right)\right)^2\right] \\
&= E\left[\left(g^\tau(X_i) - E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right] + E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right]\right.\right. \\
&\qquad \left.\left. - \left(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\right)\right)^2\right] \\
&= E\left[\left(g^\tau(X_i) - E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right]\right)^2\right] \qquad \text{(B.13)} \\
&\qquad + E\left[\left(E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right] - \left(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\right)\right)^2\right] \ , \qquad \text{(B.14)} \\
&\geq E\left[\left(E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right] - \left(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\right)\right)^2\right] \ . \qquad \text{(B.15)}
\end{aligned}$$

where the last inequality is strict except unless $E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right] = g^\tau(X_i)$ with probability one under Q. To show (B.13), note that

$$E\left[\left(g^\tau(X_i) - E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right]\right)\right.$$

$$
\begin{aligned}
&\left(E\left[\frac{Y_i(1)}{\tau}+\frac{Y_i(0)}{1-\tau}\middle|h(X_i)\right]-\left(\frac{E[Y_i(1)]}{\tau}+\frac{E[Y_i(0)]}{1-\tau}\right)\right)\Bigg] \\
&= E\left[E\left[g^\tau(X_i)-E\left[\frac{Y_i(1)}{\tau}+\frac{Y_i(0)}{1-\tau}\middle|h(X_i)\right]\middle|h(X_i)\right]\right. \\
&\quad \left.\left(E\left[\frac{Y_i(1)}{\tau}+\frac{Y_i(0)}{1-\tau}\middle|h(X_i)\right]-\left(\frac{E[Y_i(1)]}{\tau}+\frac{E[Y_i(0)]}{1-\tau}\right)\right)\right] \\
&= 0 \ ,
\end{aligned}
$$

where the second equality holds because

$$
E[g^\tau(X_i)|h(X_i)] = E\left[\frac{Y_i(1)}{\tau}+\frac{Y_i(0)}{1-\tau}\middle|h(X_i)\right]
$$

by the law of iterated expectation. The lemma is thus proved. ■

If τ_s's are allowed to differ across s, then $\hat{\theta}_n$ is generally inconsistent for θ. In such settings researchers often use the estimator from the fully saturated regression in Bugni et al. (2019). For $1 \leq s \leq S$ and $d \in \{0,1\}$, define

$$
\hat{\mu}_{n,s}(1) = \frac{1}{n_s \tau_s} \sum_{i \in \lambda_s : D_i = 1} Y_i
$$

and

$$
\hat{\mu}_{n,s}(0) = \frac{1}{n_s (1-\tau_s)} \sum_{i \in \lambda_s : D_i = 0} Y_i \ .
$$

The estimator is

$$
\hat{\theta}_n^{\text{sat}} = \sum_{1 \leq s \leq S} \frac{n_s}{n} (\hat{\mu}_{n,s}(1) - \hat{\mu}_{n,s}(0)) \ . \tag{B.16}
$$

Note that $\hat{\theta}_n^{\text{sat}}$ and $\hat{\theta}_n$ coincide whenever $\tau_s \equiv \tau \in (0,1)$. See Bugni et al. (2018), Tabord-Meehan (2020), and Bugni et al. (2019) for more details. By repeating arguments used in the proof of Theorem 2.3.1 and Theorem B.2.1, we could find the stratification that minimizes $\text{MSE}(\hat{\theta}_n^{\text{sat}}|X^{(n)})$, which is defined as in (2.4) with $\hat{\theta}_n$ replaced by θ_n^{sat}. The solution is as follows: we first calculate the stratification defined in (2.22) with τ, g, and $X^{(n)}$ defined

separately for each subpopulation, and then take the union of those stratifications. Moreover, the next theorem enables us to derive feasible procedures similar to (2.23) when treated fractions are allowed to vary across subpopulations. In particular, it reveals any plug-in estimator that satisfies the regularity conditions in Assumption B.2.1 leads to a procedure under which the asymptotic variance of $\hat{\theta}_n^{\text{sat}}$ is no greater than and typically strictly less than that under procedures with each subpopulation as a stratum.

Theorem B.2.3. *Suppose the sample size is n. Define a function $f : \mathbf{R}^p \to \{1, \ldots, R\}$ where $R \geq 1$ is an integer. Define $N_r = \{i : f(X_i) = r\}$, $X^{N_r} = (X_i : i \in N_r)$, $n_r = |N_r|$, and $p(r) = Q\{f(X_i) = r\}$. Define $\lambda^{\text{large}} = \bigcup_{1 \leq r \leq R} N_r$. For $1 \leq r \leq R$, let τ_r be the treated fraction in N_r. Define functions $h^r : \mathbf{R}^p \to \mathbf{R}$ for $1 \leq r \leq R$. Define $\lambda^{\text{small}} = \bigcup_{1 \leq r \leq R} \lambda^{\tau_r, h^r}(X^{N_r})$, where $\lambda^{\tau_r, h^r}(X^{N_r})$ is defined in* (B.12)*. Suppose Q satisfies Assumption 2.5.1. Then, under λ^{large}, for $\hat{\theta}_n^{\text{sat}}$ defined in* (B.16)*, as $n \to \infty$,*

$$\sqrt{n}(\hat{\theta}_n^{\text{sat}} - \theta(Q)) \xrightarrow{d} N(0, \varsigma^2_{\text{large}}) \ ,$$

where

$$\varsigma^2_{\text{large}} = E\left[\frac{\text{Var}[Y_i(1)]}{\tau_{f_i}} + \frac{\text{Var}[Y_i(0)]}{1 - \tau_{f_i}} - \tau_{f_i}(1 - \tau_{f_i}) \right.$$
$$\left. E\left[\left(E\left[\frac{Y_i(1)}{\tau_{f_i}} + \frac{Y_i(0)}{1 - \tau_{f_i}}\middle| f(X_i)\right] - \left(\frac{E[Y_i(1)]}{\tau_{f_i}} + \frac{E[Y_i(0)]}{1 - \tau_{f_i}}\right)\right)^2\right]\right] \ .$$

Suppose in addition that $h^r, 1 \leq r \leq R$ satisfy Assumption B.2.1, under Q restricted to $\{x \in \mathbf{R}^p : f(x) = r\}$. Then, under λ^{small}, for $\hat{\theta}_n^{\text{sat}}$ defined in (B.16)*, as $n \to \infty$,*

$$\sqrt{n}(\hat{\theta}_n^{\text{sat}} - \theta(Q)) \xrightarrow{d} N(0, \varsigma^2_{\text{small}}) \ ,$$

where

$$\varsigma^2_{\text{small}} = E\left[\frac{\text{Var}[Y_i(1)]}{\tau_{f_i}} + \frac{\text{Var}[Y_i(0)]}{1-\tau_{f_i}} - \tau_{f_i}(1-\tau_{f_i}) \right.$$
$$\left. E\left[\left(E\left[\frac{Y_i(1)}{\tau_{f_i}} + \frac{Y_i(0)}{1-\tau_{f_i}}\middle| h^{f_i}(X_i)\right] - \left(\frac{E[Y_i(1)]}{\tau_{f_i}} + \frac{E[Y_i(0)]}{1-\tau_{f_i}}\right)\right)^2\right]\right] .$$

In addition, $\varsigma^2_{\text{small}} \leq \varsigma^2_{\text{large}}$*, where the inequality is strict unless*

$$E\left[\frac{Y_i(1)}{\tau_{f_i}} + \frac{Y_i(0)}{1-\tau_{f_i}}\middle| h^{f_i}(X_i)\right] = E\left[\frac{Y_i(1)}{\tau_{f_i}} + \frac{Y_i(0)}{1-\tau_{f_i}}\middle| f(X_i)\right]$$

with probability one under Q*. Moreover, among all choices of* $(h^r : 1 \leq r \leq R)$*,* $\varsigma^2_{\text{small}}$ *is minimized by setting* $h^r = g^{\tau_r}$*, where* g^{τ_r} *is defined in* (2.21)*.*

Remark B.2.1. Tabord-Meehan (2020) considers stratification trees, which leads to a small number of large strata, with different treated fractions in each stratum. Using results from Theorem B.2.3, it is straightforward to combine his procedure with procedures in this chapter. The asymptotic variance of $\hat{\theta}^{\text{sat}}_n$ under the combined procedure is no greater than and typically strictly less than that under his procedure alone. The combined procedure is as follows: First, perform the procedure in Tabord-Meehan (2020), which produces a finite number of strata with a target treated fraction for each stratum. Second, we view each stratum as a subpopulation and calculate the stratification in (B.12) either with a fixed function h or some plug-in estimate, with τ equal the target treated fraction. Finally, we take the union of these stratifications. The desired property now follows from Theorem B.2.3. ■

Proof of Theorem B.2.3. The first convergence holds by Theorem 3.1 of Bugni et al. (2019). For the second convergence, note that

$$\begin{pmatrix} \sqrt{n_1}(\hat{\mu}_{n,1}(1) - \hat{\mu}_{n,1}(0)) \\ \vdots \\ \sqrt{n_R}(\hat{\mu}_{n,R}(1) - \hat{\mu}_{n,R}(0)) \end{pmatrix} \xrightarrow{d} N\left(0, \text{diag}(\varsigma^2_{\tau_r,h^r} : 1 \leq r \leq R)\right) .$$

Meanwhile, note that $\frac{n_r}{n} \xrightarrow{P} p(r)$ for $1 \leq r \leq R$. The convergence then follows by the Slutsky's lemma. The last two results could be shown similarly to Theorem B.2.2. ■

B.3 Auxiliary Lemmas

In the rest of the appendix, we use $a \lesssim b$ to denote that there exists $c \geq 0$ such that $a \leq cb$.

Lemma B.3.1. *If the treatment assignment scheme satisfies Assumption 2.2.1, then* $\Lambda_n \subseteq \text{co}(\Lambda_n^{\text{pair}})$.

Proof of Lemma B.3.1. We first prove that $\lambda_0 = \{\{X_1, \ldots, X_{2n}\}\}$ is a convex combination of matched-pair designs. Indeed,

$$\lambda_0 = \frac{1}{|\Lambda_n^{\text{pair}}|} \bigoplus_{\lambda \in \Lambda_n^{\text{pair}}} \lambda \ ,$$

where

$$|\Lambda_n^{\text{pair}}| = \frac{\binom{2n}{n} n!}{2^n} \ .$$

Next, consider $\lambda = \{\lambda_1, \ldots, \lambda_S\}$. Let $\Lambda_n^{\text{pair}}(\lambda_s)$ denote the set of all matched-pair designs of units in λ_s. Then,

$$\lambda = \frac{1}{\prod_{1 \leq s \leq S} |\Lambda_n^{\text{pair}}(\lambda_s)|} \bigoplus_{\xi^s \in \Lambda_n^{\text{pair}}(\lambda_s): 1 \leq s \leq S} \bigcup_{1 \leq s \leq S} \xi^s \ ,$$

and the conclusion follows. ■

Example B.3.1. Let $n = 4$ and define

$$\lambda^0 = \{\{1, 2, 3, 4\}\}$$

$$\lambda^1 = \{\{1, 2\}, \{3, 4\}\}$$

$$\lambda^2 = \{\{1, 3\}, \{2, 4\}\}$$

$$\lambda^3 = \{\{1,4\},\{2,3\}\} \ .$$

We have $\lambda^0 = \frac{1}{3}\lambda^1 \bigoplus \frac{1}{3}\lambda^2 \bigoplus \frac{1}{3}\lambda^3$. ■

Lemma B.3.2. *Suppose $m \geq 2$, and $x_1, \ldots, x_{2m}$ are real number such that $x_1 \leq \ldots \leq x_{2m}$. Then, for any $\pi \in \Pi_n$,*

$$\sum_{k=1}^{m} x_{\pi(2k-1)} x_{\pi(2k)} \leq \sum_{k=1}^{m} x_{2k-1} x_{2k} \ . \tag{B.17}$$

Proof of Lemma B.3.2. We need only consider the case where there exists $k_1 < k_2 < k_3 < k_4$ such that at least one of $\pi(k_1), \pi(k_2)$ is greater than at least one of $\pi(k_3), \pi(k_4)$ because the lemma trivially holds otherwise. Suppose without loss of generality that $\pi(k_2) < \pi(k_3) < \pi(k_4) < \pi(k_1)$, then it is easy to verify that

$$x_{\pi(k_1)} x_{\pi(k_2)} + x_{\pi(k_3)} x_{\pi(k_4)} \leq x_{\pi(k_2)} x_{\pi(k_3)} + x_{\pi(k_1)} x_{\pi(k_4)}$$

so that by interchanging two indices we decrease the sum weakly. A finite number of those interchanges maps π back to the identity operator, and hence (B.17) holds. ■

Lemma B.3.3. *Let X_n, Y_n, Z_n be random variables. Suppose $Y_n = g(Z_n) \xrightarrow{d} Y$ as $n \to \infty$, where $g : \mathbf{R} \to \mathbf{R}$ is measurable and $X_n \xrightarrow{d} X$ conditional on Z_n, with probability one for Z_n. Furthermore, suppose the distributions of both X and Y are continuous everywhere. Then*

$$(X_n, Y_n) \xrightarrow{d} (X, Y) \ ,$$

where $X \perp\!\!\!\perp Y$.

Proof of Lemma B.3.3. Since X and Y both have continuous distribution function, we need only show for any $x, y \in \mathbf{R}$,

$$P\{X_n \leq x, Y_n \leq y\} \to P\{X \leq x\} P\{Y \leq y\} \ .$$

To this end, note that

$$
\begin{aligned}
&P\{X_n \le x, Y_n \le y\} - P\{X \le x\}P\{Y \le y\} \\
&= E[E[I\{X_n \le x\}I\{Y_n \le y\}|Z_n]] - P\{X \le x\}P\{Y \le y\} \\
&= E[E[I\{X_n \le x\}|Z_n]I\{Y_n \le y\}] - P\{X \le x\}P\{Y \le y\} \\
&= E[(E[I\{X_n \le x\}|Z_n] - P\{X \le x\})I\{Y_n \le y\}] + E[P\{X \le x\}(I\{Y_n \le y\} - P\{Y \le y\})] \\
&= E[(P\{X_n \le x|Z_n\} - P\{X \le x\})I\{Y_n \le y\}] + (P\{Y_n \le y\} - P\{Y \le y\})P\{X \le x\}
\end{aligned}
$$

For the first term on the right-hand side, note that

$$P\{X_n \le x|Z_n\} - P\{X \le x\} \to 0$$

with probability one for Z_n, and hence the product inside the expectation converges to 0 with probability one as well, which in turn implies the expectation converges to 0 by the dominated convergence theorem since probabilities are bounded. The second term converges to 0 because of the definition of convergence in distribution and the fact that the distribution of Y has no discontinuity. ■

Lemma B.3.4. *Suppose the sample size is kn for $k \in \mathbb{Z}$ and the treatment assignment scheme satisfies $\tau_s \equiv \tau = \frac{l}{k}$, where $l \in \mathbb{Z}$, $0 < l < k$, and they are relatively prime. Suppose Q satisfies Assumption 2.5.1 and h satisfies Assumption B.2.1. Then, under $\lambda^{\tau,h}(X^{(n)})$ defined in* (B.12)*, as $n \to \infty$,*

$$\sqrt{kn}(\hat{\theta}_n - \theta(Q)) \xrightarrow{d} N(0, \varsigma^2_{\tau,h}) \ ,$$

where

$$\varsigma^2_{\tau,h} = \frac{\mathrm{Var}[Y_i(1)]}{\tau} + \frac{\mathrm{Var}[Y_i(0)]}{1-\tau}$$

$$- \tau(1-\tau)E\left[\left(E\left[\frac{Y_i(1)}{\tau} + \frac{Y_i(0)}{1-\tau}\middle| h(X_i)\right] - \left(\frac{E[Y_i(1)]}{\tau} + \frac{E[Y_i(0)]}{1-\tau}\right)\right)^2\right] \quad . \quad \text{(B.18)}$$

Proof of Lemma B.3.4. To begin with, note that

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

where

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

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

We first study the limiting behavior of A_n. Conditional on $h^{(n)}$ and $D^{(n)}$, the terms in the sum are independent but not identically distributed. Therefore, we proceed to verify that the Lindeberg condition holds in probability conditional on $h^{(n)}$ and $D^{(n)}$. To that end, define

$$s_n^2 = s_n^2(h^{(n)}, D^{(n)}) = \sum_{1\le i\le kn} \operatorname{Var}\left[\frac{Y_i(1)D_i}{\tau}\middle| h^{(n)}, D^{(n)}\right]$$

and note that

$$\begin{aligned}
s_n^2 &= \sum_{1\le i\le kn} \operatorname{Var}\left[\frac{Y_i(1)D_i}{\tau}\middle| h^{(n)}, D^{(n)}\right] \\
&= \frac{1}{\tau^2} \sum_{1\le i\le kn} D_i \operatorname{Var}[Y_i(1)|h^{(n)}]
\end{aligned}$$

$$= \frac{1}{\tau^2} \sum_{1 \le i \le kn : D_i = 1} \text{Var}[Y_i(1) | h(X_i)] \ ,$$

where the second equality follows from (2.2) and the third follows from the fact that units are i.i.d. It follows that

$$\tau \frac{s_n^2}{kn} = \frac{1}{kn} \sum_{1 \le i \le kn} \text{Var}[Y_i(1) | h(X_i)] + \left(\frac{1-\tau}{\tau kn} \sum_{1 \le i \le kn : D_i = 1} \text{Var}[Y_i(1) | h(X_i)] \right.$$
$$\left. - \frac{1}{kn} \sum_{1 \le i \le kn : D_i = 0} \text{Var}[Y_i(1) | h(X_i)] \right) \ . \quad \text{(B.19)}$$

By Assumption 2.5.1,

$$\frac{1}{kn} \sum_{1 \le i \le kn} \text{Var}[Y_i(1) | h(X_i)] \xrightarrow{P} E[\text{Var}[Y_i(1) | h(X_i)]] < E[Y_i(1)] < \infty \ . \quad \text{(B.20)}$$

Meanwhile,

$$\begin{aligned}
&\left| \frac{1-\tau}{\tau kn} \sum_{1 \le i \le kn : D_i = 1} \text{Var}[Y_i(1) | h(X_i)] - \frac{1}{kn} \sum_{1 \le i \le kn : D_i = 0} \text{Var}[Y_i(1) | h(X_i)] \right| \\
&\lesssim \left| \frac{1-\tau}{\tau kn} \sum_{1 \le i \le kn : D_i = 1} h_i - \frac{1}{kn} \sum_{1 \le i \le kn : D_i = 0} h_i \right| \\
&= \frac{1}{\tau kn} \left| \sum_{1 \le s \le n} \sum_{(s-1)k+1 \le j \le sk : D_{\pi^{\tau,h}(j)} = 1} (h_{\pi^{\tau,h}(j)} - \bar{h}_s^\tau) \right| \\
&\le \frac{1}{\tau kn} \sum_{1 \le s \le n} \sum_{(s-1)k+1 \le j \le sk : D_{\pi^{\tau,h}(j)} = 1} |h_{\pi^{\tau,h}(j)} - \bar{h}_s^\tau| \\
&\lesssim \frac{1}{n} \sum_{1 \le s \le n} \sum_{(s-1)k+1 \le j \le sk} |h_{\pi^{\tau,h}(j)} - \bar{h}_s^\tau| \\
&\le \left(\frac{1}{n} \sum_{1 \le s \le n} \sum_{(s-1)k+1 \le j \le sk} |h_{\pi^{\tau,h}(j)} - \bar{h}_s^\tau|^2 \right)^{1/2} \xrightarrow{P} 0 \ , \quad \text{(B.21)}
\end{aligned}$$

where the first inequality holds by Assumption B.2.1(b), the second holds by using Assumption B.2.1(c), the third holds by inspection, the last holds by the Cauchy-Schwarz inequality, and the equality holds by inspection. Combining (B.19), (B.20), and (B.21), we have

$$\frac{s_n^2}{kn} \xrightarrow{P} \frac{E[\mathrm{Var}[Y_i(1)|h(X_i)]]}{\tau} > 0 \ , \tag{B.22}$$

where the inequality holds by Assumption B.2.1(a).

We now argue that the Lindeberg condition holds in probability conditional on $h^{(n)}$ and $D^{(n)}$, i.e., for any $\epsilon > 0$,

$$E_n = \frac{1}{s_n^2 \tau^2} \sum_{1 \le i \le kn} E[|Y_i(1)D_i - E[Y_i(1)D_i|h^{(n)}, D^{(n)}]|^2 \\ I\{|Y_i(1)D_i - E[Y_i(1)D_i|h^{(n)}, D^{(n)}]| > \epsilon \tau s_n\}|h^{(n)}, D^{(n)}] \xrightarrow{P} 0 \ .$$

To this end, first note that for any $M > 0$,

$$P\{\epsilon \tau s_n > M\} \to 1 \tag{B.23}$$

because of (B.22). Next, note that

$$E[Y_i(1)D_i|h^{(n)}, D^{(n)}] = E[Y_i(1)|h(X_i)]D_i$$

because of (2.2). As a result, for any $M > 0$

$$\begin{aligned} & E_n \\ &= \frac{1}{s_n^2 \tau^2} \sum_{1 \le i \le kn : D_i = 1} E[|Y_i(1) - E[Y_i(1)|h^{(n)}, D^{(n)}]|^2 \\ &\qquad I\{|Y_i(1) - E[Y_i(1|h^{(n)}, D^{(n)}]| > \epsilon \tau s_n\}|h^{(n)}, D^{(n)}] \\ &\le \frac{1}{s_n^2 \tau^2} \sum_{1 \le i \le kn} E[|Y_i(1) - E[Y_i(1)|h^{(n)}, D^{(n)}]|^2 \end{aligned}$$

$$
\begin{aligned}
&\quad I\{|Y_i(1) - E[Y_i(1)|h^{(n)}, D^{(n)}]| > \epsilon\tau s_n\}|h^{(n)}, D^{(n)}] \\
&\leq \frac{1}{s_n^2\tau^2} \sum_{1\leq i\leq kn} E[|Y_i(1) - E[Y_i(1)|h(X_i)]|^2 I\{|Y_i(1) - E[Y_i(1)|h(X_i)] > M\}|h^{(n)}, D^{(n)}] \\
&\quad + o_p(1) \\
&= \frac{kn}{s_n^2\tau^2}\frac{1}{kn} \sum_{1\leq i\leq kn} E[|Y_i(1) - E[Y_i(1)|h(X_i)]|^2 I\{|Y_i(1) - E[Y_i(1)|h(X_i)]| > M\}|h^{(n)}, D^{(n)}] \\
&\quad + o_p(1) \qquad \text{(B.24)} \\
&\xrightarrow{P} (E[\mathrm{Var}[Y_i(1)|h(X_i)]])^{-1} E[|Y_i(1) - E[Y_i(1)|h(X_i)]|^2 I\{|Y_i(1) - E[Y_i(1)|h(X_i)]| > M\}] \ , \qquad \text{(B.25)}
\end{aligned}
$$

where the first inequality holds by inspection, the second holds because of (B.23) and the equality follows because (2.2) and $Q_n = Q^{kn}$, and the convergence in probability follows from (B.22) and the fact that Assumption B.2.1(a) implies

$$
\begin{aligned}
&E[|Y_i(1) - E[Y_i(1)|h(X_i)]|^2 I\{|Y_i(1) - E[Y_i(1)|h(X_i)]| > M\}] \\
&\qquad \leq E[|Y_i(1) - E[Y_i(1)|h(X_i)]|^2] = E[\mathrm{Var}[Y_i(1)|h(X_i)]] \leq E[Y_i^2(1)] < \infty \ .
\end{aligned}
$$

In addition, by the dominated convergence theorem,

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

To show that $E_n \xrightarrow{P} 0$, fix any subsequence, which we still call $\{n\}$ with some abuse of notation, and we argue that there is a further subsequence along which E_n converges to 0 almost surely. Indeed, for the subsequence $\{n\}$, for any fixed M, the preceding display is bounded by (B.24), which we define as $U_n(M)$. We know from above that $U_n(M) \xrightarrow{P} U(M)$, where $U(M)$ is defined as (B.25). Hence, there exists a further subsequence $\{n\}$ along which $U_n(M) \to U(M)$ almost surely. We then choose a sequence $\{M_n\}_{n\geq 1}$ such that $M_n \to \infty$. By the dominated convergence theorem, $\lim_{n\to\infty} U(M_n) = 0$. By a diagonalizing

argument, we could construct a further subsequence $\{n\}$ along which $U_n(M_n) \to 0$. Along this subsequence, since $E_n \leq U_n(M_n)$ for each n, the almost sure limit of E_n must be zero because it is non-negative.

We now argue that

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

Fix any subsequence. Since $E_n \xrightarrow{P} 0$, there exists a further subsequence along which $E_n \to 0$ with probability one for $h^{(n)}, D^{(n)}$. Because of the Lindeberg condition and (B.22), it follows that with probability one for $h^{(n)}, D^{(n)}$, $A_n \xrightarrow{d} N(0, E[\mathrm{Var}[Y_i(1)|h(X_i)]]/\tau)$ conditional on $h^{(n)}, D^{(n)}$. But then the left-hand side of the preceding display must converge almost surely to 0 by Pólya's theorem. Since for any subsequence there exists a further subsequence along which it converges to 0 almost surely, it must converge to 0 in probability.

A similar argument establishes that

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

Since A_n and B_n are independent conditional on $h^{(n)}$ and $D^{(n)}$, it follows by a similar subsequencing argument as above that

$$\sup_{t \in \mathbf{R}} \Big| P\{A_n - B_n \leq t | h^{(n)}, D^{(n)}\} \\ - \Phi\left(t / \sqrt{E[\mathrm{Var}[Y_i(1)|h(X_i)]/\tau + E[\mathrm{Var}[Y_i(0)|h(X_i)]]/(1-\tau)}\right) \Big| \xrightarrow{P} 0 \ . \quad \text{(B.26)}$$

To study C_n, note that by (2.2),

$$C_n = \frac{1}{\sqrt{kn}} \sum_{1 \leq i \leq kn} \left(E\left[\frac{Y_i(1)}{\tau} \middle| h(X_i) \right] D_i - E[Y_i(1)] \right) \ .$$

So we have

$$E[C_n|h^{(n)}] = \frac{1}{\sqrt{kn}} \sum_{1\le i\le kn} (E[Y_i(1)|h(X_i)] - E[Y_i(1)]) \ .$$

Furthermore, by Assumptions B.2.1(b)–(c),

$$\mathrm{Var}[C_n|h^{(n)}] \propto \frac{1}{kn} \sum_{1\le s\le n} (h_{\pi^{\tau,h}(i)} - \bar{h}^s_\tau)^2 \xrightarrow{P} 0 \ ,$$

where the first relation could be established by repeating the arguments used in the last step of establishing Theorem B.2.1. It therefore follows by Markov's inequality that for any $\epsilon > 0$,

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

and since probabilities are bounded and hence uniformly integrable,

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

and hence

$$C_n = \frac{1}{\sqrt{kn}} \sum_{1\le i\le kn} (E[Y_i(1)|h(X_i)] - E[Y_i(1)]) + o_p(1) \ .$$

A similar proof shows that

$$D_n = \frac{1}{\sqrt{kn}} \sum_{1\le i\le kn} (E[Y_i(0)|h(X_i)] - E[Y_i(0)]) + o_p(1) \ .$$

and therefore

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

We now show by contradiction that

$$\sup_{t \in \mathbf{R}} |P\{\sqrt{n}(\hat{\theta}_n - \theta(Q)) \leq t\} - \Phi(t/\varsigma_h)| \to 0 \ .$$

Suppose not, then there must exist a subsequence along which the left-hand side of the above display converges to some $\delta > 0$. Along this subsequence, we could find a further subsequence along which the left-hand side of (B.26) converges to 0 with probability one for $h^{(n)}$ and $D^{(n)}$, i.e.,

$$A_n - B_n \xrightarrow{d} N\left(0, \frac{E[\mathrm{Var}[Y_i(1)|h(X_i)]}{\tau} + \frac{E[\mathrm{Var}[Y_i(0)|h(X_i)]]}{1-\tau}\right)$$

with probability one for $h^{(n)}$ and $D^{(n)}$. Since $C_n - D_n$ is constant for each $h^{(n)}$ and $D^{(n)}$, Lemma B.3.3 establishes that

$$A_n - B_n + C_n - D_n \xrightarrow{d} N\Bigg(0, \frac{E[\mathrm{Var}[Y_i(1)|h(X_i)]}{\tau} + \frac{E[\mathrm{Var}[Y_i(0)|h(X_i)]]}{1-\tau} + E\left[(E[Y_i(1)|h(X_i)] - E[Y_i(1)] - (E[Y_i(0)|h(X_i)] - E[Y_i(0)]))^2\right]\Bigg) \ ,$$

which, by Pólya's Theorem, implies a contradiction.

Finally, note that

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

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

and the result follows. ■

Assumption B.3.1. h satisfies

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

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

Lemma B.3.5. *Define*

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

and

$$
(\hat{\varsigma}_n^h)^2 = \hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0) - \frac{1}{2}\hat{\rho}_n + \frac{1}{2}(\hat{\mu}_n(1) + \hat{\mu}_n(0))^2 .
$$

Suppose the treatment assignment scheme satisfies Assumption 2.2.1, Q satisfies Assumption 2.5.1, and h satisfies Assumptions B.2.1 and B.3.1. Then, under $\lambda^{\frac{1}{2},h}$ defined in (B.12)*,*

$$
(\hat{\varsigma}_n^h)^2 \xrightarrow{P} \varsigma^2_{\frac{1}{2},h} .
$$

Proof of Lemma B.3.5. To begin with, note that $\hat{\mu}_n(d) \xrightarrow{P} E[Y_i(d)]$ and $\hat{\sigma}_n^2(d) \xrightarrow{P} \mathrm{Var}[Y_i(d)]$ for $d \in \{0,1\}$, by Lemma A.9.5. Next, we show that

$$
E[\hat{\rho}_n | h^{(n)}] \xrightarrow{P} \rho^2 . \tag{B.27}
$$

For notational simplicity, we define $\mu_d(h_i) = E[Y_i(d)|h(X_i) = h_i]$ for $d \in \{0,1\}$. To see this, note that

$$
\begin{aligned}
&E[(Y_{\pi^h(4j-3)} + Y_{\pi^h(4j-2)})(Y_{\pi^h(4j-1)} + Y_{\pi^h(4j)})|h^{(n)}] \\
&= \frac{1}{4}(\mu_1(h_{\pi^h(4j-3)}) + \mu_0(h_{\pi^h(4j-2)}))(\mu_1(h_{\pi^h(4j-1)}) + \mu_0(h_{\pi^h(4j)})) \\
&\quad + \frac{1}{4}(\mu_1(h_{\pi^h(4j-3)}) + \mu_0(h_{\pi^h(4j-2)}))(\mu_1(h_{\pi^h(4j)}) + \mu_0(h_{\pi^h(4j-1)})) \\
&\quad + \frac{1}{4}(\mu_1(h_{\pi^h(4j-2)}) + \mu_0(h_{\pi^h(4j-3)}))(\mu_1(h_{\pi^h(4j-1)}) + \mu_0(h_{\pi^h(4j)})) \\
&\quad + \frac{1}{4}(\mu_1(h_{\pi^h(4j-2)}) + \mu_0(h_{\pi^h(4j-3)}))(\mu_1(h_{\pi^h(4j)}) + \mu_0(h_{\pi^h(4j-1)})) \\
&= \frac{1}{4}(g_h(h_{\pi^h(4j-3)}) + g_h(h_{\pi^h(4j-2)}))(g_h(h_{\pi^h(4j-1)}) + g_h(h_{\pi^h(4j)})) \\
&= \frac{1}{4}\sum_{k\in\{2,3\},l\in\{0,1\}} g_h^2(h_{\pi^h(4j-k)}) + g_h^2(h_{\pi^h(4j-l)}) - (g_h(h_{\pi^h(4j-k)}) - g_h(h_{\pi^h(4j-l)}))^2 \ .
\end{aligned}
$$

As a result,

$$
\begin{aligned}
&E[\hat{\rho}_n|h^{(n)}] \\
&= \sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} E[(Y_{\pi^h(4j-3)} + Y_{\pi^h(4j-2)})(Y_{\pi^h(4j-1)} + Y_{\pi^h(4j)})|h^{(n)}] \\
&= \frac{1}{2n}\sum_{1\le i\le 2n} g_h^2(h(X_i)) - \frac{1}{4n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor}\sum_{k\in\{2,3\},l\in\{0,1\}} (g_h(h_{\pi^h(4j-k)}) - g_h(h_{\pi^h(4j-l)}))^2 \ .
\end{aligned}
$$

(B.27) then follows from Assumption B.2.1(b), B.3.1, the fact that

$$
\begin{aligned}
E[g_h^2(h(X_i))] &\lesssim E[E[Y_i(1)|h(X_i)]^2] + E[E[Y_i(0)|h(X_i)]^2] \\
&\le E[E[Y_i^2(1)|h(X_i)]] + E[E[Y_i^2(0)|h(X_i)]] = E[Y_i^2(1) + Y_i^2(0)] < \infty
\end{aligned}
$$

because of Assumption 2.5.1, and an application of the WLLN.

It remains to show $\hat{\rho}_n - E[\hat{\rho}_n|h^{(n)}] \xrightarrow{P} 0$. We will prove

$$\frac{2}{n} \sum_{1 \le j \le \lfloor \frac{n}{2} \rfloor} (Y_{\pi^h(4j-2)} Y_{\pi^h(4j)} - E[Y_{\pi^h(4j-2)} Y_{\pi^h(4j)}|h^{(n)}]) \xrightarrow{P} 0 \ ,$$

and the others follow similarly. We will repeatedly use the following elementary inequalities for any $a, b \in \mathbf{R}$ and $\lambda > 0$:

$$\begin{aligned} |a+b| I\{|a+b| > \lambda\} &\le 2|a| I\{|a| > \lambda/2\} + 2|b| I\{|b| > \lambda/2\} \\ |ab| I\{|ab| > \lambda\} &\le |a|^2 I\{|a| > \sqrt{\lambda}\} + |b|^2 I\{|b| > \sqrt{\lambda}\} \ . \end{aligned}$$

To begin with,

$$E[Y_{\pi^h(4j-2)} Y_{\pi^h(4j)}|h^{(n)}] = \frac{1}{2}\mu_1(h_{\pi^h(4j-2)})\mu_0(h_{\pi^h(4j)}) + \frac{1}{2}\mu_1(h_{\pi^h(4j)})\mu_0(h_{\pi^h(4j-2)})$$

Next, note that

$$\begin{aligned} &\frac{2}{n} \sum_{1 \le j \le \lfloor \frac{n}{2} \rfloor} E[|Y_{\pi^h(4j-2)} Y_{\pi^h(4j)} - E[Y_{\pi^h(4j-2)} Y_{\pi^h(4j)}|h^{(n)}]| \\ &\quad I\{|Y_{\pi^h(4j-2)} Y_{\pi^h(4j)} - E[Y_{\pi^h(4j-2)} Y_{\pi^h(4j)}|h^{(n)}]| > \lambda\}|h^{(n)}] \\ &\le \frac{2}{n} \sum_{1 \le j \le \lfloor \frac{n}{2} \rfloor} E[|Y_{\pi^h(4j-2)} Y_{\pi^h(4j)}| I\{|Y_{\pi^h(4j-2)} Y_{\pi^h(4j)}| > \sqrt{\lambda/2}\}|h^{(n)}] \\ &\quad + E[|E[Y_{\pi^h(4j-2)} Y_{\pi^h(4j)}|h^{(n)}]| I\{|E[Y_{\pi^h(4j-2)} Y_{\pi^h(4j)}|h^{(n)}]| > \sqrt{\lambda/2}\}|h^{(n)}] \\ &\le \frac{2}{n} \sum_{1 \le j \le \lfloor \frac{n}{2} \rfloor} E[Y^2_{\pi^h(4j-2)} I\{|Y_{\pi^h(4j-2)}| > \sqrt{\lambda/2}\}|h^{(n)}] \\ &\quad + E[Y^2_{\pi^h(4j)} I\{|Y_{\pi^h(4j)}| > \sqrt{\lambda/2}\}|h^{(n)}] \\ &\quad + |\mu_1(h_{\pi^h(4j-2)})\mu_0(h_{\pi^h(4j)})| I\{\mu_1(h_{\pi^h(4j-2)})\mu_0(h_{\pi^h(4j)}) > \lambda/2\} \\ &\quad + |\mu_1(h_{\pi^h(4j)})\mu_0(h_{\pi^h(4j-2)})| I\{\mu_1(h_{\pi^h(4j)})\mu_0(h_{\pi^h(4j-2)}) > \lambda/2\} \end{aligned}$$

$$
\begin{aligned}
&\leq \frac{2}{n} \sum_{1 \leq j \leq \lfloor \frac{n}{2} \rfloor} E[Y^2_{\pi^h(4j-2)}(1) I\{|Y_{\pi^h(4j-2)}(1)| > \sqrt{\lambda/2}\} | h^{(n)}] \\
&\quad + E[Y^2_{\pi^h(4j-2)}(0) I\{|Y_{\pi^h(4j-2)}(0)| > \sqrt{\lambda/2}\} | h^{(n)}] \\
&\quad + E[Y^2_{\pi^h(4j)}(1) I\{|Y_{\pi^h(4j)}(1)| > \sqrt{\lambda/2}\} | h^{(n)}] \\
&\quad + E[Y^2_{\pi^h(4j)}(0) I\{|Y_{\pi^h(4j)}(0)| > \sqrt{\lambda/2}\} | h^{(n)}] \\
&\quad + \mu_1^2(h_{\pi^h(4j-2)}) I\{|\mu_1(h_{\pi^h(4j-2)})| > \sqrt{\lambda/2}\} + \mu_0^2(h_{\pi^h(4j)}) I\{|\mu_0(h_{\pi^h(4j)})| > \sqrt{\lambda/2}\} \\
&\quad + \mu_1^2(h_{\pi^h(4j)}) I\{|\mu_1(h_{\pi^h(4j)})| > \sqrt{\lambda/2}\} + \mu_0^2(h_{\pi^h(4j-2)}) I\{|\mu_0(h_{\pi^h(4j-2)})| > \sqrt{\lambda/2}\} \\
&\lesssim \frac{1}{2n} \sum_{1 \leq i \leq 2n} E[Y_i^2(1) I\{|Y_i(1) > \sqrt{\lambda/2}|\} | h(X_i)] + E[Y_i^2(0) I\{|Y_i(1) > \sqrt{\lambda/2}|\} | h(X_i)] \\
&\quad + E[Y_i^2(1) | h(X_i)] I\{E[Y_i^2(1) | h(X_i)] > \sqrt{\lambda/2}\} \\
&\quad + E[Y_i^2(0) | h(X_i)] I\{E[Y_i^2(0) | h(X_i)] > \sqrt{\lambda/2}\} \\
&\xrightarrow{P} E[Y_i^2(1) I\{|Y_i(1) > \sqrt{\lambda/2}|\}] \\
&\quad + E[Y_i^2(0) I\{|Y_i(1) > \sqrt{\lambda/2}|\}] + E[E[Y_i^2(1) | h(X_i)] I\{E[Y_i^2(1) | h(X_i)] > \sqrt{\lambda/2}\}] \\
&\quad + E[E[Y_i^2(0) | h(X_i)] I\{E[Y_i^2(0) | h(X_i)] > \sqrt{\lambda/2}\}] \ , \qquad \text{(B.28)}
\end{aligned}
$$

where the last line follows from WLLN and the law of iterated expectation. Since by Assumption 2.5.1 we have $E[Y_i^2(d)] < \infty$ and hence $E[E[Y_i(d)|h(X_i)]^2] < E[Y_i^2(d)]$ by Jensen's inequality, the limit as $\lambda \to \infty$ of the last line is 0, by the dominated convergence theorem. We finish the proof by arguing by contradiction. Suppose

$$
\hat{\rho}_n - E[\hat{\rho}_n | h^{(n)}]
$$

does not converge in probability to 0. There must then exist $\epsilon > 0$ and $\delta > 0$ and a subsequence, which for simplicity we again denote by $\{n\}$, such that

$$
P\{|\hat{\rho}_n - E[\hat{\rho}_n | h^{(n)}]| > \epsilon\} \to \delta \qquad \text{(B.29)}
$$

along this subsequence. But because of (B.28), there exists a further subsequence along which the condition in Lemma A.9.3 holds with probability one for $h^{(n)}$, but then along this subsequence $\hat{\rho}_n - E[\hat{\rho}_n|h^{(n)}] \xrightarrow{P} 0$ conditional on $h^{(n)}$ with probability one for $h^{(n)}$, i.e., for any $\epsilon > 0$, with probability one for $h^{(n)}$,

$$P\{|\hat{\rho}_n - E[\hat{\rho}_n|h^{(n)}]| > \epsilon|h^{(n)}\} \to 0 \ .$$

Since probabilities are bounded and hence uniformly integrable,

$$P\{|\hat{\rho}_n - E[\hat{\rho}_n|h^{(n)}]| > \epsilon\} \to 0$$

along the chosen subsequence, which implies a contradiction to (B.29). ■

Lemma B.3.6. *Suppose U_i, $1 \leq i \leq n$ are i.i.d. random variables where $E|U_i|^r < \infty$. Then*

$$n^{-1/r} \max_{1 \leq i \leq n} |U_i| \xrightarrow{P} 0 \ .$$

Proof of Lemma B.3.6. Note that for all $\epsilon > 0$,

$$\begin{aligned} P\left\{n^{-1/r} \max_{1 \leq i \leq n} |U_i| > \epsilon\right\} &= P\left\{\max_{1 \leq i \leq n} |U_i|^r > n\epsilon^r\right\} \\ &\leq nP\{|U_i|^r > n\epsilon^r\} \leq \frac{n}{n\epsilon^r} E[|U_i|^r I\{|U_i|^r > n\epsilon^r\}] = \frac{1}{\epsilon^r} E[|U_i|^r I\{|U_i|^r > n\epsilon^r\}] \to 0 \ , \end{aligned}$$

where the convergence follows because of the dominated convergence theorem and that $E|U_i|^r < \infty$. ■

Lemma B.3.7. *Suppose $E[h^2(X_i)] < \infty$. Then Assumptions B.2.1(c) and B.3.1 hold.*

Proof of Lemma B.3.7. We prove the case where $\tau = \frac{1}{2}$ and the results follow similarly

for any $\tau \in (0,1)$. Note that

$$\sum_{1\le s\le n} |h_{\pi^h(2s-1)} - h_{\pi^h(2s)}|^2 \le |h_{\pi^h(2n)} - h_{\pi^h(1)}|^2 \le 4 \max_{1\le i\le 2n} h^2(X_i) \ ,$$

where the first inequality follows from the definition of π^h and the second inequality follows by inspection, and therefore it follows from Lemma B.3.6 that

$$\frac{1}{n}\sum_{1\le s\le n} |h_{\pi^h(2s-1)} - h_{\pi^h(2s)}|^2 \le \frac{4}{n} \max_{1\le i\le 2n} h^2(X_i) \xrightarrow{P} 0 \ .$$

Assumption B.2.1(c) thus holds. To see Assumption B.3.1 holds, note that

$$\frac{1}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} |h_{\pi^h(4j-k)} - h_{\pi^h(4j-l)}|^2 \lesssim \frac{1}{n}|h_{\pi^h(2n)} - h_{\pi^h(1)}|^2 \ ,$$

and the result follows similarly as above. ■

Lemma B.3.8. *Suppose g satisfies Assumption 2.5.2(c) and $\hat{g}_m$ satisfies Assumption 2.5.3. Then, as $m, n \to \infty$,*

$$\frac{1}{n}\sum_{1\le s\le n} |g_{\pi^{\hat{g}_m}(2s-1)} - g_{\pi^{\hat{g}_m}(2s)}|^2 \xrightarrow{P} 0 \ ,$$

and

$$\frac{1}{n}\sum_{1\le j\le \lfloor \frac{n}{2}\rfloor} |g_{\pi^{\hat{g}_m}(4j-k)} - g_{\pi^{\hat{g}_m}(4j-l)}|^2 \xrightarrow{P} 0$$

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

Proof of Lemma B.3.8. We only prove the first conclusion as the second could be shown by similar arguments. We first show that Assumption 2.5.3 implies

$$\frac{1}{n}\sum_{1\le i\le 2n} |\hat{g}_i - g_i|^2 \xrightarrow{P} 0 \ . \tag{B.30}$$

Suppose Assumption 2.5.3 holds. For any $\epsilon > 0$, $\delta > 0$, there exists $M > 0$ such that for

$m > M$,

$$P\left\{\int_{\mathbf{R}^p} |\hat{g}_m(x) - g(x)|^2 \, Q_X(dx) > \frac{\epsilon\delta}{2}\right\} \leq \frac{\delta}{2} \; . \tag{B.31}$$

By Markov's inequality again, if

$$\int_{\mathbf{R}^p} |\hat{g}_m(x) - g(x)|^2 \, Q_X(dx) \leq \frac{\epsilon\delta}{2} \; ,$$

then by the independence of $\tilde{W}^{(m)}$ and $W^{(n)}$,

$$\begin{aligned} P\left\{\frac{1}{2n}\sum_{1\leq i\leq 2n} |\hat{g}_i - g_i|^2 > \epsilon \middle| \tilde{W}^{(m)}\right\} &\leq \frac{E\left[\frac{1}{2n}\sum_{1\leq i\leq 2n} |\hat{g}_i - g_i|^2 \middle| \tilde{W}^{(m)}\right]}{\epsilon} \\ &= \frac{\int_{\mathbf{R}^p} |\hat{g}_m(x) - g(x)|^2 \, Q_X(dx)}{\epsilon} \leq \frac{\delta}{2} \; . \end{aligned} \tag{B.32}$$

Then,

$$\begin{aligned} P\left\{\frac{1}{2n}\sum_{1\leq i\leq 2n} |\hat{g}_i - g_i|^2 > \epsilon\right\} &\leq P\left\{\frac{1}{2n}\sum_{1\leq i\leq 2n} |\hat{g}_i - g_i|^2 > \epsilon \middle| \tilde{W}^{(m)}\right\} \\ &\quad P\left\{\int_{\mathbf{R}^p} |\hat{g}_m(x) - g(x)|^2 \, Q_X(dx) \leq \frac{\epsilon\delta}{2}\right\} \\ &\quad + P\left\{\int_{\mathbf{R}^p} |\hat{g}_m(x) - g(x)|^2 \, Q_X(dx) > \frac{\epsilon\delta}{2}\right\} \\ &\leq \frac{\delta}{2}\left(1 - \frac{\delta}{2}\right) + \frac{\delta}{2} \leq \delta \; , \end{aligned}$$

where the first inequality follows by definition, and the second inequality follows from (B.31) and (B.32).

Next, note that since $|a+b|^2 \leq 2(a^2+b^2)$ for any $a, b \in \mathbf{R}$,

$$\frac{1}{n}\sum_{1\leq s\leq n} |g_{\pi^{\hat{g}_m}(2s-1)} - g_{\pi^{\hat{g}_m}(2s)}|^2$$

$$\lesssim \frac{1}{n} \sum_{1 \leq s \leq n} |\hat{g}_{\pi^{\hat{g}m}(2s-1)} - \hat{g}_{\pi^{\hat{g}m}(2s)}|^2 + \frac{1}{n} \sum_{1 \leq i \leq 2n} |\hat{g}_i - g_i|^2 \ . \tag{B.33}$$

Next, note that

$$\begin{aligned}
&\frac{1}{n} \sum_{1 \leq s \leq n} |\hat{g}_{\pi^{\hat{g}m}(2s-1)} - \hat{g}_{\pi^{\hat{g}m}(2s)}|^2 \\
&\leq \frac{1}{n} \max_{1 \leq i \leq 2n} |\hat{g}_i|^2 \\
&\lesssim \frac{1}{n} \max_{1 \leq i \leq 2n} |g_i|^2 + \frac{1}{n} \max_{1 \leq i \leq 2n} |\hat{g}_i - g_i|^2 \\
&\lesssim \frac{1}{n} \max_{1 \leq i \leq 2n} |g_i|^2 + \frac{1}{n} \sum_{1 \leq i \leq 2n} |\hat{g}_i - g_i|^2 \ .
\end{aligned} \tag{B.34}$$

The conclusion then follows from (B.30), (B.33), (B.34), Assumption 2.5.2(c) and an application of Lemma B.3.6. ■

B.3.1 Sufficient Conditions for Lipschitz Continuity

Let f denote the density function of X. Recall that $C^{(r)}$ is the class of functions which are rth continuously differentiable. We impose the following assumption on h in Assumption B.2.1 and f.

Assumption B.3.2. The function h and density function f satisfy the following conditions.

(a) $h \in C^{(2)}$.

(b) $\frac{\partial h(x)}{\partial x_p} \neq 0$ Lebesgue a.e.

(c) $f \in C^{(2)}$.

Lemma B.3.9 (Theorem 24.4 of Munkres (1997))**.** *Let O be open in $\mathbf{R}^p$ and $f : O \to \mathbf{R}$ be of class $C^{(r)}$ for $r \geq 1$. Let M be the set of points x for which $f(x) = 0$ and N be the set of points x for which $f(x) \geq 0$. Suppose M is non-empty and $Df(x)$ has rank 1 at each point of M. Then N is a p-manifold in $\mathbf{R}^p$ and $\partial N = M$.*

Lemma B.3.10. *Suppose Assumption B.3.2(a)–(b) hold. Then* $M = \{x : h(x) = z\}$ *is a* $(p-1)$*-manifold in* $\mathbf{R}^p$*.*

Proof of Lemma B.3.10. For each $x \in M$, we aim at providing a coordinate patch on M about x. Indeed, by Assumption B.3.2(a)–(b) and Theorem 9.2 (implicit function theorem) of Munkres (1997), there exists an open set U containing $u = (x_1, \ldots, x_{p-1})$, an open ball $B(z)$ containing z and an open set O in $\mathbf{R}$ containing x_p, and a function $k : U \times B(z) \to \mathbf{R}^p$ of class $C^{(2)}$ such that $h(u, k(u, z')) = z'$ for all $u \in U$, $z' \in B(z)$ and $x \in O$. Moreover, $k(U \times B(z)) = O$. Define the coordinate patch $\alpha(u; z) = (u, k(u, z))$. The conclusion follows by Theorem 5-2 of Spivak (1965). ■

Note that $M = \{x : h(x) = z\}$ is a $(p-1)$-manifold by Lemmas B.3.9 and B.3.10. In what follows, we will need the definition of the integral of a function g over the manifold M. In order to do so, note that there exists a coordinate patch as $\{\alpha_j : U_j \subseteq \mathbf{R}^{p-1} \to V_j \subseteq M, j \in \mathcal{J}\}$, where $\alpha_j(u) = \alpha_j(u, z)$, and each $\alpha_j(u) = (u, k_j(u))$ for some function $k_j : U \to \mathbf{R}$ which is of class C^2, as shown in the proof of Lemma B.3.10, and $\alpha_j(U_j) = V_j$. Next, there exists a partition of unity $\{\phi_i : i \in \mathcal{I}\}$ dominated by the $\{V_j : j \in \mathcal{J}\}$. Moreover, both $\mathcal{I}$ and $\mathcal{J}$ could be chosen to be countable, according to Section 25 of Munkres (1997). The integral of a scalar function g over the manifold is written as

$$\int_M g \, \mathrm{d}V = \sum_{j \in \mathcal{J}} \sum_{i \in \mathcal{I}} \int_{U_j} [(g\phi_i) \circ \alpha_j] V(D\alpha_j) \ ,$$

where $V(A) = \sqrt{\det(A'A)}$ is the volume. We have

$$D\alpha_j = \begin{bmatrix} I_{p-1} & \dfrac{\partial k_j(u,z)}{\partial u} \end{bmatrix} \ ,$$

so that

$$V(D\alpha_j) = \sqrt{1 + \frac{\partial k_j(u,z)}{\partial u'} \frac{\partial k_j(u,z)}{\partial u}} = \frac{\|\nabla h(u, k_j(u,z))\|}{|D_p h(u, k_j(u,z))|} \ ,$$

where $D_p = \frac{\partial}{\partial x_p}$, by the implicit function theorem and matrix determinant lemma. Note that on one hand, for each $j \in \mathcal{J}$, only a finite number of ϕ_i is positive, and on the other hand, $\{\phi_i : i \in \mathcal{I}\}$ is dominated by the coordinate patch, which means that each ϕ_i is supported on a compact set inside a single V_j. As a result, the order of the above double sum could be interchanged.

By p.345 of Bogachev (2007), the conditional expectation of a function g on the manifold M is defined as

$$E[g(X)|M] = \lim_{t \to 0} \frac{E[g(X) I\{z \le h(X) \le z + t\}]}{P\{z \le h(X) \le z + t\}} .$$

Lemma B.3.11. *Suppose Assumption B.3.2(a)–(c) hold. Then*

$$E[g(X)|M] = \frac{\displaystyle\int_M \frac{fg}{\|\nabla h\|} \, \mathrm{d}V}{\displaystyle\int_M \frac{f}{\|\nabla h\|} \, \mathrm{d}V} . \tag{B.35}$$

For a continuously differentiable function $h : \mathbf{R}^p \to \mathbf{R}$, $x \in \mathbf{R}^p$ is a critical point of h if $\nabla h(x) = 0$, where $\nabla h(x)$ is the gradient of h at x; otherwise x is a regular point of h. A value z is a critical value of h if the set $\{x : h(x) = z\}$ contains at least one critical point; otherwise z is a regular value of h.

Proof of Lemma B.3.11. By L'Hospital's rule,

$$E[g(X)|M] = \frac{\displaystyle\lim_{t \to 0} \frac{E[g(X) I\{z \le h(X) \le z + t\}]}{t}}{\displaystyle\lim_{t \to 0} \frac{P\{z \le h(X) \le z + t\}}{t}} ,$$

and the lemma follows from Lemma A.1 of Chernozhukov et al. (2018). In particular, the denominator equals the one in (B.35) directly by that lemma, while for the numerator we merely need to redefine the 'density' function as fg and the same proof goes through. ■

Lemma B.3.12. *Suppose Assumption B.3.2(a)–(b) hold. Let $M = \{x : h(x) = z\}$, where z*

is a regular value of h on $\mathbf{R}^p$. Then for any $g \in C^{(2)}$,

$$\frac{\partial}{\partial z}\int_M g\,\mathrm{d}V = \int_M \frac{D_p g}{D_p h}\,\mathrm{d}V + \int_M g\frac{1}{\|\nabla h\|^2}\sum_{1\le i\le p}\frac{D_i h D_{ip} h}{D_p h}\,\mathrm{d}V - \int_M g\frac{D_{pp}h}{D_p^2 h}\,\mathrm{d}V\ . \quad \text{(B.36)}$$

Proof of Lemma B.3.12. To begin with, note that

$$\begin{aligned}
&\frac{\partial}{\partial z}\int_{U_j}[(g\phi_i)\circ\alpha_j]V(D\alpha_j)\\
&\quad = \int_{U_j} D_p(g\phi_i)\frac{\partial k_j(u,z)}{\partial z}\frac{\|\nabla h\|}{|D_p h|}\\
&\qquad + \int_{U_j} g\phi_i\frac{|D_p h|}{\|\nabla h\|}\frac{\partial k_j(u,z)}{\partial z}\frac{1}{D_p^4 h}\left(D_p^2 h\sum_{1\le i\le p} D_i h D_{ip} h - D_p h D_{pp} h\sum_{1\le i\le p} D_i^2 h\right)\ ,
\end{aligned} \quad \text{(B.37)}$$

where $D_{ij}h = \partial_i\partial_j h$ for any function $h \in C^{(2)}$. we have suppressed the arguments of h, being $(u, k_j(u,z))$. Note that it is legitimate to pass diffentiation inside the integral by the dominated convergence theorem. By the Implicit Function Theorem again,

$$\frac{\partial k_j(u,z)}{\partial z} = \frac{1}{D_p h(u, k_j(u,z))}\ . \quad \text{(B.38)}$$

By Theorem 7.17 of Rudin (1976), we know that $\frac{\partial}{\partial z}\int_M g(x)\,\mathrm{d}V$ is the sum over $i \in \mathcal{I}, j \in \mathcal{J}$ of the two terms in (B.37). Using (B.38), the sum of the first term is

$$\begin{aligned}
&\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}}\int_{U_j}(\phi_i D_p g + g D_p\phi_i)\frac{1}{D_p h}\frac{\|\nabla h\|}{|D_p h|}\\
&= \sum_j \int_{U_j}\frac{D_p g}{D_p h}V(D\alpha_j)\\
&= \int_M \frac{D_p g}{D_p h}\,\mathrm{d}V\ ,
\end{aligned} \quad \text{(B.39)}$$

because $\sum_{i\in\mathcal{I}}\phi_i = 1$ and hence $\sum_{i\in\mathcal{I}} D_p\phi_i = D_p\sum_{i\in\mathcal{I}}\phi_i = 0$. Again, the interchange of

differentiation and sum is allowed because the sum is actually over a finite number of terms, by definition of a partition of unity. The sum of the second term is

$$\begin{aligned}
&\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}}\int_{U_j} g\phi_i \frac{|D_p h|}{\|\nabla h\|}\frac{1}{D_p^4 h}\sum_{1\le i\le p}(D_i h D_p h D_{ip} h - D_i^2 h D_{pp} h)\\
&=\sum_{j\in\mathcal{J}}\int_{U_j} g\frac{D_p^2 h}{\|\nabla h\|^2}\frac{1}{D_p^4 h}\sum_{1\le i<p}(D_i h D_p h D_{ip} h - D_i^2 h D_{pp} h)V(D\alpha)\\
&=\int_M g\frac{1}{\|\nabla h\|^2 D_p^2 h}\sum_{1\le i\le p}(D_i h D_p h D_{ip} h - D_i^2 h D_{pp} h)\,\mathrm{d}V\\
&=\int_M g\frac{1}{\|\nabla h\|^2}\sum_{1\le i\le p}\frac{D_i h D_{ip} h}{D_p h}\,\mathrm{d}V - \int_M g\frac{D_{pp} h}{D_p^2 h}\,\mathrm{d}V\ . \qquad \text{(B.40)}
\end{aligned}$$

(B.36) now follows from (B.39) and (B.40). ■

Theorem B.3.1. *Suppose Assumption B.3.2 holds. If z is a regular value of h, then*

$$\frac{\partial}{\partial z}E[g(X)|M] = \frac{\displaystyle\int_M \frac{D_p(fg/D_p h)}{\|\nabla h\|}\,\mathrm{d}V\int_M \frac{f}{\|\nabla h\|}\,\mathrm{d}V - \int_M \frac{D_p(f/D_p h)}{\|\nabla h\|}\,\mathrm{d}V\int_M \frac{fg}{\|\nabla h\|}\,\mathrm{d}V}{\displaystyle\left[\int_M \frac{f}{\|\nabla h\|}\,\mathrm{d}V\right]^2}\ . \qquad \text{(B.41)}$$

Proof of Theorem B.3.1. To begin with, replace g in Lemma B.3.12 with $\frac{f}{\|\nabla h\|}$. We then have

$$\begin{aligned}
&\frac{\partial}{\partial z}\int_M \frac{f}{\|\nabla h\|}\,\mathrm{d}V\\
&=\int_M \frac{\|\nabla h\| D_p f - \dfrac{f\sum_{1\le i\le p} D_i h D_{ip} h}{\|\nabla h\|}}{\|\nabla h\|^2 D_p h}\,\mathrm{d}V\\
&\qquad+\int_M \frac{f}{\|\nabla h\|^3}\sum_{1\le i\le p}\frac{D_i h D_{ip} h}{D_p h}\,\mathrm{d}V - \int_M \frac{f D_{pp} h}{\|\nabla h\| D_p^2 h}\,\mathrm{d}V\\
&=\int_M \frac{D_p f D_p h - f D_{pp} h}{\|\nabla h\| D_p^2 h}\,\mathrm{d}V\\
&=\int_M \frac{D_p(f/D_p h)}{\|\nabla h\|}\,\mathrm{d}V\ . \qquad \text{(B.42)}
\end{aligned}$$

By the same arguments,

$$\frac{\partial}{\partial z}\int_M \frac{fg}{\|\nabla h\|}\,\mathrm{d}V = \int_M \frac{D_p(fg/D_ph)}{\|\nabla h\|}\,\mathrm{d}V\ . \tag{B.43}$$

(B.41) now follows from (B.42) and (B.43) together with the quotient rule. ■

In general, by the Law of Iterated Expectation

$$E[Y_i^r(d)|h(X)=z] = E[E[Y_i^r(d)|X]|h(X)=z]\ .$$

Suppose h and the density function of X, $f(X)$ satisfy the smoothness conditions in Assumption B.3.2, the derivative

$$\frac{\partial}{\partial z}E[g(X)|h(X)=z]$$

is given in Theorem B.3.1, where $g(x) = E[Y_i^r(d)|X=x]$ for $r=1,2$ and $d=0,1$. In particular, it is equal to

$$\begin{aligned}&E\left[\frac{D_pg}{D_ph}+\frac{gD_pf}{fD_ph}-\frac{gD_{pp}h}{D_p^2h}\middle|h(X)=z\right]-E\left[\frac{D_pf}{fD_ph}-\frac{D_{pp}h}{D_p^2h}\middle|h(X)=z\right]E\left[g\middle|h(X)=z\right]\\&=E\left[\frac{D_pg}{D_ph}\middle|h(X)=z\right]+\mathrm{Cov}\left[\frac{D_pf}{fD_ph}-\frac{D_{pp}h}{D_p^2h},g\middle|h(X)=z\right]\ .\end{aligned} \tag{B.44}$$

Lemma B.3.13. *Each of the following conditions imply the boundedness of* (B.44)*.*

1. *h is linear, $\|D_pg\|_\infty<\infty$, $\|g\|_\infty<\infty$ and $\|D_p(\ln f)\|_\infty<\infty$.*
2. *h is linear, $\sup_{z\in\mathbf{R}}|E[D_pg|h(X)=z]|<\infty$, $\sup_{z\in\mathbf{R}}|E[g^2|h(X)=z]|<\infty$ and $\sup_{z\in\mathbf{R}}|E[D_p^2(\ln f)|h(X)=z]|<\infty$.*
3. *h includes linear and interaction terms, $\left\|\frac{D_pg}{D_ph}\right\|_\infty<\infty$, $\|g\|_\infty<\infty$ and $\left\|\frac{D_p(\ln f)}{D_ph}\right\|_\infty<\infty$.*

Proof of Lemma B.3.13. Follows from inspection. ■

B.4 Details of Penalized Matching

In this section, we consider the solution to the Bayesian problem in (2.33) a particular example that motivates the penalized matching procedure defined by (2.31). For simplicity, we focus on the special case under which and $Y_i(d) \sim N(X_i'\beta(d), \sigma^2)$ for $d \in \{0,1\}$. Note that the potential outcomes are homoskedastic conditional on the covariates. Define $\beta = \beta(1) + \beta(0)$, and we have $g(x) = x'\beta$. As before, we suppose $\tilde{W}^{(m)} = ((\tilde{Y}_j, \tilde{X}_j', \tilde{D}_j)' : 1 \leq j \leq m)$ is available from a pilot experiment. Suppose the prior on $\beta(d)$ is $G_d \overset{d}{=} N(\eta(d), \Sigma(d))$ for $d \in \{0,1\}$, being independent across $d \in \{0,1\}$. The prior distribution of β is then $G(d\beta) \overset{d}{=} N(\eta(1)+\eta(0), \Sigma(1)+\Sigma(0))$. We could show that the posterior distribution of $\beta(d)$ conditional on $\tilde{W}^{(m)}$ is

$$\bar{G}_d(d\beta|\tilde{W}^{(m)}) \overset{d}{=} N(\bar{\eta}, \bar{\Sigma}) \ ,$$

where for $d \in \{0,1\}$,

$$\bar{\eta}(d) = \left((\sigma^2)^{-1} \sum_{j:\tilde{D}_j=d} \tilde{X}_j\tilde{X}_j' + \Sigma^{-1}(d) \right)^{-1} \left((\sigma^2)^{-1} \sum_{j:\tilde{D}_j=d} \tilde{X}_j\tilde{Y}_j + \Sigma^{-1}(d)\eta(d) \right)$$

$$\bar{\Sigma}(d) = \left((\sigma^2)^{-1} \sum_{j:\tilde{D}_j=d} \tilde{X}_j\tilde{X}_j' + \Sigma^{-1}(d) \right)^{-1} \ .$$

Define $\bar{\eta} = \bar{\eta}(1) + \bar{\eta}(0)$ and $\bar{\Sigma} = \bar{\Sigma}(1) + \bar{\Sigma}(0)$. The posterior distribution for β is

$$\bar{G}(d\beta|\tilde{W}^{(m)}) \overset{d}{=} (\bar{\eta}, \bar{\Sigma}) \ ,$$

since $G_d(d\beta)$'s are independent across $d \in \{0,1\}$.

The next lemma provides the solution to the Bayesian problem in (2.33), where the choice set is over all measurable functions $u : (\tilde{w}^{(m)}, x^{(n)}) \mapsto \lambda \in \Lambda_n$.

Lemma B.4.1. *The solution to* (2.33) *maps each* $(\tilde{w}^{(m)}, x^{(n)})$ *to* $\lambda = \{\{\pi(2s-1), \pi(2s)\} :$

$1 \leq s \leq n/2\}$, *where* π *solves*

$$\min_{\pi \in \Pi_n} \sum_{1 \leq s \leq n} \bar{d}\left(x_{\pi(2s-1)}, x_{\pi(2s)}\right) \ ,$$

where

$$\bar{d}(x_1, x_2) = (x_1'\bar{\eta} - x_2'\bar{\eta})^2 + (x_1 - x_2)'\bar{\Sigma}(x_1 - x_2) \ . \tag{B.45}$$

Proof. First note that by (2.9) and (2.12), (2.33) is equivalent to

$$\min_u \iiint L(u(\tilde{w}^{(m)}, x^{(n)})|\beta, x^{(n)}) Q_X^n(dx^{(n)}) Q_{\tilde{W}}^m(d\tilde{w}^{(m)}) G(d\beta) \ . \tag{B.46}$$

Next, note that we could solve the problem pointwise for $\tilde{w}^{(m)}$ and $x^{(n)}$ since (B.46) is equivalent to

$$\min_u \bar{R}(u|\tilde{W}^{(m)}) \ , \tag{B.47}$$

where

$$\bar{R}(u|\tilde{W}^{(m)}) = \int L(u(\tilde{W}^{(m)}, x^{(n)})|\beta, x^{(n)}) \bar{G}(d\beta|\tilde{W}^{(m)}) \ .$$

To solve (B.47), first note that since $\bar{R}(u|\tilde{W}^{(m)})$ is linear in u, by Lemma B.3.1, it is solved by a matched-pair design. Next,

$$\bar{R}(u|\tilde{W}^{(m)}) = \sum_{1 \leq s \leq n} ((x'_{\pi(2s-1)}\bar{\eta} - x'_{\pi(2s)}\bar{\eta})^2 + (x_{\pi(2s-1)} - x_{\pi(2s)})'\bar{\Sigma}(x_{\pi(2s-1)} - x_{\pi(2s)})) \ .$$

As a result, minimizing it is equivalent to minimizing the sum of the distances defined in (B.45). ■

Finally, we want the prior to be irrelevant. For the purpose, suppose that $\Sigma = cI$ where I is an identity matrix. We let the constant $c \to \infty$, so that the prior diverges to a diffuse (uninformative) one. Then, $\bar{\eta}(d)$ converges to $\hat{\beta}_m(d)$ in (2.24) and $\bar{\Sigma}(d)$ converges to $\hat{\Sigma}_m(d)$ defined in (2.25). Therefore, we define $\hat{\beta}_m$ as in (2.26) and $\hat{\Sigma}_m$ as in (2.27). The metric

(B.45) converges to the metric defined in (2.32).

B.5 Minimax Matching

This section describes the minimax procedure in detail. First note that $L(\lambda|h, X^{(n)})$ depends on h only through $h^{(n)}$, and hence (2.48) is equivalent to

$$\min_{\lambda\in\Lambda} \max_{h^{(n)}\in G} L(\lambda|h^{(n)}) \ , \tag{B.48}$$

where

$$L(\lambda|h^{(n)}) = L(\lambda|h, X^{(n)})$$

and

$$G = \{h^{(n)} : h \in \mathcal{G}, h_1 = 0\} \ .$$

The restriction $h_1 = 0$ is a location normalization, since $L(\lambda|h^{(n)})$ only depends on $h^{(n)}$ through pairwise differences and is therefore shift-invariant. In order to solve (B.48) computationally, we impose the following requirement on G:

Assumption B.5.1. G is a bounded polyhedron in $\mathbf{R}^n$.

We now provide examples of G that satisfy Assumption B.5.1.

Example B.5.1. Consider the class of Lipschitz functions:

$$G = \{h^{(n)} : |h_i - h_j| \leq M\|X_i - X_j\| \text{ for } i \neq j, \ h_1 = 0\} \ . \tag{B.49}$$

G satisfies Assumption B.5.1. ■

Example B.5.2. When $p > 2$, i.e., X_i is multivariate, consider the class of functions which

are Lipschitz along each dimension:

$$G = \left\{ h^{(n)} : |h_i - h_j| \leq \sum_{1 \leq l \leq p} M_l |X_{il} - X_{jl}| \text{ for } i \neq j,\, h_1 = 0 \right\} .$$

G satisfies Assumption B.5.1. ■

Example B.5.3. Consider the class of functions Lipschitz in a known index. For a known function w, define

$$G = \left\{ h^{(n)} : |h_i - h_j| \leq M |\nu(X_i) - \nu(X_j)| \text{ for } i \neq j,\, h_1 = 0 \right\} . \tag{B.50}$$

G satisfies Assumption B.5.1. ■

Example B.5.4. Consider the class of linear functions with coefficients in a bounded polyhedron. For a bounded polyhedron $\mathcal{B}$ in $\mathbf{R}^p$, define

$$G = \{ X^{(n)}\beta - X_1'\beta \mathbf{1}_n : \beta \in \mathcal{B} \} .$$

G satisfies Assumption B.5.1. ■

Example B.5.5. Consider the class of monotonically increasing functions. Without loss of generality assume that $X_1 \leq \ldots \leq X_n$. For $M > 0$, define

$$G = \{ h^{(n)} : h_i \leq h_j \text{ for } i < j, h_n \leq M, h_1 = 0 \} .$$

Since G is bounded and defined by linear inequalities, it satisfies Assumption B.5.1. ■

Example B.5.6. Consider the class of convex functions. Without loss of generality assume that $X_1 \leq \ldots \leq X_n$. For $M > 0$, define

$$G = \left\{ h^{(n)} : h_i \leq \frac{X_{i+1} - X_i}{X_{i+1} - X_{i-1}} h_{i-1} + \frac{X_i - X_{i-1}}{X_{i+1} - X_{i-1}} h_{i+1}, 2 \leq i \leq 2n-1, |h_n| \leq M, h_1 = 0 \right\} .$$

Since G is bounded and defined by linear inequalities, it satisfies Assumption B.5.1. ■

Consider the minimax problem (B.48) with G defined in (B.50). The following theorem shows that without any information of how the covariate affects potential outcomes beyond the index, the best we could do is to match on the index itself.

Theorem B.5.1. *The solution to* (B.48) *with* G *defined in* (B.50) *is* $\lambda^\nu = \{\{\pi^\nu(2s-1), \pi^\nu(2s)\} : 1 \le s \le n\}$ *where* $\nu_{\pi^\nu(1)} \le \ldots \le \nu_{\pi^\nu(2n)}$.

Proof of Theorem B.5.1. Without loss of generality, consider $p = 1$ and $\nu(x) = x$. The general case is proved in exactly the same way. We use another expression of (2.47). Define $\Delta_i = g_{\pi(i+1)} - g_{\pi(i)}$ for $i = 1, \ldots, 2n-1$. For $\lambda^0 = \{\{1, \ldots, 2n\}\}$,

$$
\begin{aligned}
& L(\lambda^0 | g, X^{(n)}) \\
&= \frac{1}{2n(2n-1)} \sum_{1 \le i \le 2n} \left[(2n-1) g_i - \sum_{j \ne i} g_j \right]^2 \\
&= \frac{1}{2n(2n-1)} \sum_{1 \le i \le 2n} \left[- \sum_{1 \le j \le i-1} j \Delta_j + \sum_{i \le j \le 2n-1} (2n-j) \Delta_j \right]^2 \\
&= \frac{1}{2n(2n-1)} \left[\sum_{1 \le i \le 2n-1} 2n(2n-i) i \Delta_i^2 + 2 \sum_{k < l \le 2n-1} 2n(2n-l) k \Delta_k \Delta_l \right] \\
&= \frac{1}{2n-1} \left[\sum_{1 \le i \le 2n-1} (2n-i) i \Delta_i^2 + 2 \sum_{k < l \le 2n-1} (2n-l) k \Delta_k \Delta_l \right] .
\end{aligned}
$$

As a result, for a general stratification λ, the loss function (2.47) equals

$$
L(\lambda | g, X^{(n)}) = \sum_{1 \le s \le S} \frac{1}{n_s - 1} \left[\sum_{1 \le i \le n_s - 1} (n_s - i) i \Delta_{i,s}^2 + 2 \sum_{k < l \le n_s - 1} (n_s - l) k \Delta_{k,s} \Delta_{l,s} \right] . \tag{B.51}
$$

Note that $g^{\mathrm{mm}}(x) = Mx$ simultaneously maximizes (B.51) for every λ. But we know the stratification that solves

$$
\min_{\lambda \in \Lambda} L(\lambda | g^{\mathrm{mm}}, X^{(n)})
$$

is the "optimal non-bipartite matching" of X on $\mathbf{R}$, i.e. λ^x. ■

For a prespecified $\theta_0 \in \mathbf{R}$, consider the problem of testing (2.35) at level $\alpha \in (0,1)$. We use the test in (2.38) by setting $\hat{g}_m = \nu$.

Corollary B.5.1. *Suppose the treatment assignment scheme satisfies Assumption 2.2.1 and Q satisfies Assumption 2.5.1 and $h = \nu$ satisfies Assumption B.2.1 with $\tau = \frac{1}{2}$. Then, for the problem of testing* (2.35) *at level $\alpha \in (0,1)$, ϕ_n^ν satisfies*

$$\lim_{n\to\infty} E[\phi_n^\nu(W^{(n)})] = \alpha \ ,$$

whenever Q additionally satisfies the null hypothesis, i.e., $\theta(Q) = \theta_0$.

For other specifications of G in (B.48), there does not exist a clean result as Theorem B.5.1, as illustrated by the following example.

Example B.5.7. Let $n = 4$ and $X_1 = (0,0)', X_2 = (1,0)', X_3 = (0,1)', X_4 = (1,1)'$. Let $n = 4$ and define

$$\begin{aligned}
\lambda^0 &= \{\{1,2,3,4\}\} \\
\lambda^1 &= \{\{1,2\},\{3,4\}\} \\
\lambda^2 &= \{\{1,3\},\{2,4\}\} \\
\lambda^3 &= \{\{1,4\},\{2,3\}\} \ .
\end{aligned}$$

Let G be as defined in (B.49) with $M = 1$. Then λ^0 solves (B.48). Indeed, for $\lambda = \lambda^1$, the worst case occurs at $h^{(n)} = (0, \sqrt{2}-1, \sqrt{2}-1, \sqrt{2})$, with the loss equal to 2. For $\lambda = \lambda^2$ or λ^3, the worst case occurs at $h^{(n)} = (0,1,1,0)$, with the loss equal to 2. In contrast, the worst case for $\lambda = \lambda^0$ occurs at $h^{(n)} = (0, \sqrt{2}-1, \sqrt{2}-1, \sqrt{2})$, and the loss is $(10 - 4\sqrt{2})/3 < 2$.

■

The key reason why (B.48) is hard to solve when $p > 1$ is that the choice set Λ is not

convex. In principle, we could convexify the problem by considering the co(Λ), the convex hull of Λ. That amounts to allowing for mixing over (potentially a large number of) matched-pair designs, which is hard to interpret and is almost never used in practice. Although Λ is not convex, we can still provide computational strategies to solve (B.48). Note that $L(\lambda|h^{(n)})$ is convex in $h^{(n)}$, which combined with Assumption B.5.1 implies that the inner maximum in (B.48) is attained on the vertices of G, which we denote by V. Then, the minimax problem is equivalent to

$$\min_{\lambda\in\Lambda}\max_{h^{(n)}\in V} L(\lambda|h^{(n)}) \ . \tag{B.52}$$

We now apply results from graph theory to reformulate (B.52) into Mixed Integer Linear Programs (MILPs). We first recall some definitions from the graph theory and connect them to the optimal stratification problem. For more details, see Bertsimas and Tsitsiklis (1997).

An undirected graph $\Gamma = (N, E)$ consists of a set of nodes N and a set of edges E. Each element of E is an unordered pair $\{i, j\}$ where $i \in N$ and $j \in N$. Define $q_e = 1$ if $e \in E$ and define $\mathbf{q} = (q_e)_{e\in E}$. Define $q_{ij} = q_{i,j}$. The degree of i is defined as $d_i = \sum_j q_{ij}$. The graph Γ is complete if $q_{ij} = 1$ for all $i \neq j$. A subset U of N is a clique in Γ if $\{i, j\} \in E$ for all $i, j \in U$. The set of induced edges by U is $E(U) = \{\{i, j\} \in E : i, j \in U, i \neq j\}$. A clique partition of Γ is $\Gamma^C = (N, E(U_1, \ldots, U_S))$ for $E(U_1, \ldots, U_S) = \cup_{s=1}^S E(U_s)$ where each U_s is a clique in Γ^C (and Γ), and $\{U_s\}_{s=1}^S$ is a partition of N, i.e., $N = \cup_{s=1}^S U_s$ and $U_s \bigcap U_t = \varnothing$ for $s \neq t$.

In terms of stratification, a unit is a node and an edge $\{i, j\} \in E$ if units i and j are in the same stratum. A stratum is a clique. A stratification $\lambda = \{\lambda_s\}_{s=1}^S$ of $N = \{1, \ldots, n\}$ induces a clique partition $\Gamma^\lambda = (N, E(\lambda_1, \ldots, \lambda_S))$ of $\Gamma = (N, E)$ for $E = \{\{i, j\} : i, j \in N, i \neq j\}$ where the size of each clique λ_s is even, or equivalently the degree of each node in Γ^λ is odd.

Define $c_e = (h_i - h_j)^2$ as the cost of edge $e = \{i, j\} \in E$, $\mathbf{c} = (c_e)_{e\in E}$ and $C = \{\mathbf{c} :$

$h^{(n)} \in V\}$. By (2.47),

$$L(\lambda|h^{(n)}) = L(\lambda|h, X^{(n)}) = \sum_{1\le s\le S} \frac{1}{n_s - 1} \sum_{i,j\in\lambda_s, i<j} (h_i - h_j)^2 \ .$$

If $n_s \equiv 2$, then it equals

$$\sum_{e\in E} c_e q_e \ .$$

If $n_s > 2$ for some s, then we need to introduce other binary variables to indicate n_s. The minimax problem (B.48) is equivalent to the following MILP which solves the cost minimization problem over size-bounded stratifications within Λ, i.e., λ with $n_s \le 2K$ for all s.

$$\min_{\mathbf{q}} \quad z \tag{B.53}$$

$$\begin{aligned}
\text{subject to} \quad & \sum_{e\in E} c_e \left(\sum_{1\le k\le K} \frac{u_{ik}}{2k-1} \right) I\{i \in e\} \le z, \text{ for all } \mathbf{c} \in C \ , \\
& \sum_{l\in N} q_{il} = \sum_{1\le k\le K} (2k-1) u_{ik}, \text{ for all } i \in N \ , \\
& u_{ik} \in \{0,1\}, \text{ for all } i \in N, \ 1 \le k \le K \ , \\
& q_{e_1} + q_{e_2} - q_{e_3} \le 1, \text{ for all } e_1, e_2, e_3 \in E \ , \qquad \text{(B.54)} \\
& q_e \in \{0,1\}, \text{ for all } e \in E \ .
\end{aligned}$$

We impose an upper bound on the size of each stratum, $2k$. $u_{ik}, k = 1, \ldots, K-1$ are binary indicators of whether the stratum of unit i has size $2k$. The first set of constraints express the loss function (2.47). The second set of constraints say the degree of each node is $2k-1$, the stratum size minus one. The third set of constraints restrict u_{ik} to be binary. The fourth and the most important set of constraints, (B.54), are called triangle inequalities in the clique partition literature. See Grötschel and Wakabayashi (1990). They ensure that the solution to (B.53) is indeed a clique partition, i.e., a stratification. However, our problem differs

from the standard clique partition problem in two ways: we only allow an even number of units within each clique; and the final weights on each edge in the total cost depends on the degrees of either of its nodes, rather than being a constant.

The program (B.53) is computationally intensive even when $k = 2$ and becomes prohibitive quickly as n increases. Therefore, we consider two relaxations of it. The first relaxation is to optimize over Λ^p instead of Λ. For a matched-pair design $\lambda = \{\{\pi(2s-1), \pi(2s)\} : 1 \leq s \leq n\}$,

$$L(\lambda|h, X^{(n)}) = \sum_{1 \leq s \leq n} (h_{\pi(2s-1)} - h_{\pi(2s)})^2 \ .$$

As a result, we introduce the program as

$$\begin{aligned} \min_{\mathbf{q}} \quad & z \\ \text{subject to} \quad & \sum_{e \in E} c_e q_e \leq z, \text{ for all } \mathbf{c} \in C \ , \\ & \sum_{j \in N} q_{ij} = 1, \text{ for all } i \in N \ , \\ & q_e \in \{0, 1\}, \text{ for all } e \in E \ . \end{aligned} \tag{B.55}$$

The solution to (B.55) is $\lambda^{\text{mm}} = \{e \in E : q_e = 1\}$. We define the permutation π^{mm} such that $\lambda^{\text{mm}} = \{\{\pi^{\text{mm}}(2s-1), \pi^{\text{mm}}(2s)\} : 1 \leq s \leq n\}$. (B.55) is feasible even when n is large and requires substantially less computational budget than (B.53). Moreover, as simulation evidence in Section in Table B.1 shows, the solution to (B.55) is frequently the same with (B.53) for a small K and (B.56).

The second relaxation is the following hierarchical procedure.

Algorithm B.5.1.

1. Solve (B.55). Denote the solution by $\mathbf{q}^0$ and denote $\Lambda^0 = \{e \in E : q_e = 1\}$.

2. For $k \geq 0$, repeat steps (a) and (b) below.

(a) For $\mathbf{q}^k = (q^k_{AB})_{A,B\in\Lambda^k, A\neq B}$, solve

$$\begin{aligned} \min_{\mathbf{q}^k} \quad & z \\ \text{subject to} \quad & \sum_{A,B\in\Lambda^k} q_{AB}c_{AB} + \sum_{A\in\Lambda^k} c_A \leq z, \text{ for all } \mathbf{c}\in C\ , \\ & \sum_{B\in\Lambda^k} q^k_{AB} \leq 1, \quad \text{for all } A\in\Lambda^k\ , \\ & q^k_{AB}\in\{0,1\}, \text{ for all } A,B\in\Lambda^k\ , \end{aligned} \tag{B.56}$$

where $c_A = L(\lambda|g, X_A)$, for $X_A = \{X_i : i\in A\}$ and $c_{AB} = c_{A\bigcup B} - c_A - c_B$.

(b) Update

$$\Lambda^{k+1} = \{A\bigcup B : q^k_{AB} = 1\}\bigcup\{A : \sum_{B\in\Lambda^k} q^k_{AB} = 0\}$$

until $\Lambda^{k^*} = \Lambda^{k^*+1}$. Collect Λ^{k^*} as the solution.

Algorithm B.5.1 iteratively decides whether to merge pairs of strata or not. The algorithm stops when no paiswise merging of existing strata reduces the worst-case loss.

We now study the properties of minimax matching in a small simulation study. We compare both the actual and worst-case losses under different stratifications. In the following model, we construct a bounded polyhedron G around $g^{(n)}$. We then calculate both the actual losses $L(\lambda|g^{(n)})$ and worst-case losses $\max_{h^{(n)}\in G} L(\lambda|h^{(n)})$ across different stratifications. We set $g(x) = x'\beta$ and

$$G = \{X^{(n)}\beta : \beta\in\mathcal{B}\}\ ,$$

where $\mathcal{B}$ is a polyhedron such that $\beta\in\mathcal{B}$.

Model MM $2n = 24$; $p = 2$; $X_{i,1} = 0$ for $1\leq i\leq 8$, $X_{i,2} = 1$ for $9\leq i\leq 24$; $X_{i,2}\sim N(0,1)^2$ i.i.d. across i; $g(x) = x'\beta$, $\beta = (1,1)'$; $\mathcal{B} = \mathcal{B}_1\times\mathcal{B}_2$, $\mathcal{B}_1 = \beta_1 + \gamma_1\times[0,1]$, $\mathcal{B}_2 = \beta_2 + \gamma_2\times[-1,1]$; $\gamma\in\{(0.5,0.5)', (2,2)', (0,2)', (2,0)'\}$.

We randomly generate $X^{(n)}$ in 100 replications and summarize

(a) ratios of the values of the actual loss against those under infeasible optimal stratifications.

(b) ratios of the values of the worst-case loss against those under size-bounded minimax stratifications with $k = 2$.

We consider the following stratifications:

Oracle infeasible optimal stratification in (2.20).

by1 : $\lambda_1 = \{i : 1 \leq i \leq 8\}, \lambda_2 = \{i : 9 \leq i \leq 24\}$.

by2 two strata separated by the sample median of $X_{i,2}$.

2by2 four strata as the cross product of **by1** and **by2**.

MP2 matching on $X_{i,2}$ only, i.e., stratification in (2.23) with $\hat{g}_m(x) = x_1$.

MPcell within each value of $X_{i,1}$, optimal matched-pair design using $X_{i,2}$.

MMpair the minimax matched-pair design in (B.55).

MMbdd the size-bounded minimax stratification in (B.53) with $k = 2$.

MMhier results from the hierarchical procedure in Algorithm B.5.1.

In Model MM, **MMpair** and **MMhier** have the same solution with **MMbdd** (which we know weakly dominates **MMpair**) most of the time, while other stratifications which do not incorporate minimax consideration sometimes generate much larger worst-case losses.

B.6 AEA RCT Registry

The following experiments in the AEA RCT Registry use matched-pair designs: AEARCTR-0000086, 0000171, 0000293, 0000443, 0000481, 0000550, 0000578, 0000587, 0000644, 0000688, 0000721, 0000983, 0000986, 0001034, 0001097, 0001218, 0001370, 0001591, 0001607, 0001712, 0001714, 0001778, 0001992, 0001995, 0002010, 0002125, 0002132, 0002282, 0002585, 0002622,

0002664, 0002750, 0002776, 0003056, 0003076, 0003524, 0003581, 0003629, 0003648, 0003779, 0003814, 0003933, 0003994, 0004024, 0004042, 0004022.

		Oracle	**by1**	**by2**	**2by2**	**MP2**	**MPcell**	**MMpair**	**MMbdd**	**MMhier**
$\gamma = (0.5, 0.5)$	25%	**1.0000**	4.0109	2.3318	1.8158	1.0619	1.0256	1.0000	1.0000	1.0000
Actual	50%	**1.0000**	7.2394	3.8858	2.9807	1.2631	1.5291	1.0000	1.0000	1.0000
	75%	**1.0000**	13.7890	7.7959	6.5012	1.8567	4.4629	1.0001	1.0001	1.0001
	Mean	**1.0000**	13.6242	7.0691	7.6378	1.7480	5.5226	1.0346	1.0346	1.0346
Worst-case	25%	1.0000	4.0109	2.3381	1.7832	1.0481	1.0243	1.0000	**1.0000**	1.0000
	50%	1.0000	6.9420	3.6908	2.8469	1.1858	1.4011	1.0000	**1.0000**	1.0000
	75%	1.0003	11.9445	6.7125	5.9020	1.6146	3.9388	1.0000	**1.0000**	1.0000
	Mean	1.0212	10.3183	5.4894	5.6169	1.4884	3.8240	1.0000	**1.0000**	1.0000
$\gamma = (2, 2)$	25%	**1.0000**	4.4595	2.7007	2.1185	1.0700	1.0994	1.0000	1.0000	1.0000
Actual	50%	**1.0000**	9.2109	4.1580	3.5446	1.3348	1.8127	1.0096	1.0096	1.0096
	75%	**1.0000**	14.5268	6.8864	6.9304	1.8268	4.0986	1.3038	1.3038	1.3038
	Mean	**1.0000**	13.5257	7.3036	6.3795	1.7873	3.8773	1.2997	1.2997	1.2997
Worst-case	25%	1.0000	3.8897	2.2736	1.8008	1.0408	1.0315	1.0000	**1.0000**	1.0000
	50%	1.0126	6.2500	3.1816	2.6923	1.1604	1.4000	1.0000	**1.0000**	1.0000
	75%	1.2516	10.1048	5.0563	4.3542	1.6279	2.8357	1.0000	**1.0000**	1.0000
	Mean	1.2390	8.5778	4.9668	3.9661	1.4436	2.2735	1.0000	**1.0000**	1.0000
$\gamma = (0, 1)$	25%	**1.0000**	4.1720	2.5479	1.8497	1.0397	1.1857	1.0000	1.0000	1.0000
Actual	50%	**1.0000**	7.4458	3.8469	3.3647	1.2599	1.7892	1.0135	1.0135	1.0135
	75%	**1.0000**	14.1891	7.6734	6.3794	1.7666	3.1199	1.1199	1.1199	1.1199
	Mean	**1.0000**	12.4138	6.8864	5.5793	1.8987	2.8784	1.1301	1.1301	1.1301
Worst-case	25%	1.0000	4.3021	2.3348	1.8989	1.0012	1.2292	1.0000	**1.0000**	1.0000
	50%	1.0077	7.2928	3.4658	3.6051	1.0450	1.5861	1.0000	**1.0000**	1.0000
	75%	1.1138	16.6540	6.7290	6.8655	1.2165	3.7622	1.0000	**1.0000**	1.0000
	Mean	1.1128	12.0228	5.8405	5.4142	1.2350	2.8276	1.0000	**1.0000**	1.0000
$\gamma = (1, 0)$	25%	**1.0000**	3.5310	2.1679	2.0152	1.0654	1.0985	1.0000	1.0000	1.0000
Actual	50%	**1.0000**	8.5908	4.1682	3.8322	1.2567	1.9700	1.0481	1.0481	1.0481
	75%	**1.0000**	17.9252	8.6984	8.2448	1.8296	3.6598	1.5850	1.5850	1.5850
	Mean	**1.0000**	14.7115	8.3951	6.6366	1.6191	3.8705	1.7197	1.7197	1.7197
Worst-case	25%	1.0000	2.9528	2.4142	1.5418	1.1470	1.0000	1.0000	**1.0000**	1.0000
	50%	1.0435	4.6975	3.3215	2.1056	1.5634	1.0211	1.0000	**1.0000**	1.0000
	75%	1.6225	9.0650	5.4879	3.9089	2.6225	1.6384	1.0000	**1.0000**	1.0000
	Mean	1.6231	7.8535	6.4319	3.6442	2.3219	1.8804	1.0000	**1.0000**	1.0000

Table B.1: Ratios of values of the actual loss under all stratifications against those under the infeasible optimal stratifications (**Oracle**) and ratios of values of the worst-case loss under all stratifications against those under size-bounded minimax stratifications (**MMbdd**) in Model MM. Benchmarks are displayed in bold face.

APPENDIX C

APPENDIX FOR CHAPTER 3

C.1 Homogeneous Treatment Effects

Under homogeneous treatment effects, the notation in model (3.1) is simplified. For underlying units $i = 1, \ldots, n$, assume that $m_0(Z_i) = m_1(Z_i) = m(Z_i)$ and $\epsilon_i(0) = \epsilon_i(1) = \epsilon_i$ for $i = 1, \ldots, n$. Let $\sigma^2(Z_i) = \text{Var}[Y_i(a)|Z_i]$. The observed units have

$$Y_l = A_l'\mu + U_{\pi^{-1}(l)} \ , \tag{C.1}$$

with matrix form

$$\mathbf{Y} = \mathbf{A}\mu + \pi\mathbf{U} \ . \tag{C.2}$$

Here $U_i = m_0(Z_i) + \epsilon_i(0) = m_1(Z_i) + \epsilon_i(1)$ is a scalar and $\mathbf{U} = (U_1, \ldots, U_n)'$. Since I assume homogeneous treatment effect, $\mathbf{A}$ could be simplified to $\mathbf{A} = (A_1, \ldots, A_n)'$ where $A_l \in \{(1,0)', (0,1)'\}$ for $l = 1, \ldots, n$. Define $P_{\mathbf{U}} = \mathcal{L}(\mathbf{U}|\mathbf{Z}) \in \mathcal{P}_{\mathbf{U}}$ and impose Assumption 3.3.1 on the set $\mathcal{P}_{\mathbf{U}}$.

Assumption C.1.1. $\mathcal{P}_{\mathbf{U}}$ is H-invariant, i.e. $\mathcal{P}_{\mathbf{U}}$ satisfies that $P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}} \Rightarrow \pi P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}$ for all $P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}$ and all $\pi \in H$ for H a group of permutations.

Since I use the squared loss, the relevant attribute of the distribution of $\mathbf{U}$ is its second moment $E[\mathbf{U}\mathbf{U}'|\mathbf{Z}]$. The set C in (3.11) simplifies to

$$C = \left\{ c \in \mathbb{R}^n : \sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} c'E[\mathbf{H}\mathbf{U}\mathbf{U}'\mathbf{H}'|\mathbf{Z}]c < \infty \right\} \tag{C.3}$$

and I define the projection matrix Q as before. Note that $\sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} E[\mathbf{U}\mathbf{U}'|\mathbf{Z}] < \infty \Leftrightarrow \sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} c'E[\mathbf{U}\mathbf{U}'|\mathbf{Z}]c < \infty$ for all $c \in \mathbb{R}^n$. The set C also includes constant vectors such that $c'\pi\mathbf{U}$ has bounded second moment over $P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}$ for each $\pi \in H$, even if $\sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} E[\mathbf{U}\mathbf{U}'|\mathbf{Z}] = \infty$.

Let S_n be the set of $n \times n$ positive semi-definite (p.s.d.) symmetric matrices. For $P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}$, define

$$V_{P_{\mathbf{U}}} = E[Q\mathbf{H}\mathbf{U}\mathbf{U}'\mathbf{H}'Q|\mathbf{Z}] \ .$$

Then, Assumption 3.3.2 simplifies as follows.

Assumption C.1.2. There exists a V such that

$$V_{P_{\mathbf{U}}} \leq V \text{ for all } P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}} \text{ and } \sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} \operatorname{tr} V_{P_{\mathbf{U}}} = \operatorname{tr} V \ . \tag{C.4}$$

The object of interest and loss function are (3.9) and (3.10) again. An assignment scheme is again the distribution $\mathcal{L}(\mathbf{A}, \beta, \pi)$ of the triple $(\mathbf{A}, \beta, \pi) \in \mathcal{A}_n \times \mathbb{R}^n \times H$. Suppose there exists a solution $(\mathbf{A}_0, \beta_0)$ to the problem

$$\min \quad \mathbf{A}V\mathbf{A}' \quad \text{subject to} \quad \beta'\mathbf{A} = \tau', Q\beta = \beta \ . \tag{C.5}$$

The following restricted version of Theorem 3.3.1 is due to Hooper (1989), so the proof follows from that of Theorem 3.3.1 as well.

Theorem C.1.1. *Suppose Assumptions 3.3.1 and C.1.2 hold and* $(\mathbf{A}_0, \beta_0)$ *solves* (C.5)*, then the scheme* $(\mathbf{A}_0, \mathbf{H}, \beta_0)$ *is minimax in that it solves*

$$\min_{\mathcal{L}(\mathbf{A},\beta,\pi)} \max_{(\mu, P_{\mathbf{U}}) \in \mathbb{R}^2 \times \mathcal{P}_{\mathbf{U}}} E[\|\beta'\mathbf{Y} - \tau'\mu\|^2 | \mathbf{Z}] \ . \tag{C.6}$$

Example C.1.1. Consider a block design in Example 3.2.2 under homogeneous treatment effect. Suppose further that there are B blocks and n_b observations in block b, where $\sum_{1 \leq b \leq B} n_b = n$. Write $i \in b$ if i is in block b. It is without loss of generality to assume that $Z_{1i} = b$ for $\sum_{1 \leq j \leq b} n_b + 1 \leq i \leq \sum_{1 \leq j \leq b+1} n_b$ and $1 \leq b \leq B$, i.e., the first n_1 observations are in block 1, the next n_2 in block 2, and so on. Define H as in Example 3.2.2, the group of permutations which permutes only within blocks but not across them. One could verify

that

$$\mathbb{R}^n = \bigoplus_{b=1}^{B} (W_{0,b} \oplus W_{1,b}) \ , \tag{C.7}$$

where $W_{0,b}$ is the 1-dimensional subspace spanned by $\mathbf{1}_b$, the $n \times 1$ dimensional vector with 1 for block b and 0 otherwise, and $W_{1,b}$ is the $n_b - 1$ dimensional subspace spanned by all vectors of the form $\mathbf{1}_i - \mathbf{1}_j$ for $i, j \in b$. These spaces are mutually orthogonal and $\oplus$ denotes the direct sum. Let J_b be the $n \times n$ matrix with $\{(i,k) : \sum_{1 \le j \le b} n_b + 1 \le i, k \le \sum_{1 \le j \le b+1} n_b\}$-th elements as 1 and 0 otherwise and let $I_b = \text{diag}(J_b)$. Then those spanning vectors of $W_{0,b}$ and $W_{1,b}$ are actually eigenvectors of $D_{0,b} = I_b$ and $D_{1,b} = J_b - I_b$. The orthogonal projection matrices onto the spaces in (C.7) are

$$\begin{aligned} Q_{0,b} &= \frac{1}{n_b} J_b \\ Q_{1,b} &= I_b - \frac{1}{n_b} J_b \ . \end{aligned}$$

See Bailey (2004) for more details.

Now, the set C in (C.3) could be chosen based on the designer's belief about relative magnitudes of variances. The key object in (C.3) is

$$O = E[\mathbf{H}\mathbf{U}\mathbf{U}'\mathbf{H}'|\mathbf{Z}] = \frac{1}{|H|} \sum_{\pi \in H} \pi E[\mathbf{U}\mathbf{U}'|\mathbf{Z}]\pi' \ .$$

Note that

$$\begin{aligned} \frac{1}{n_b^2} \mathbf{1}_b' O \mathbf{1}_b &= \frac{1}{n_b} \sum_{i \in b} E[U_i^2 | Z_i] + \frac{1}{n_b(n_b-1)} \sum_{i \neq j \in b} E[U_i U_j | Z_i, Z_j] \\ &= \frac{1}{n_b} \sum_{i \in b} (m(b, Z_{2i})^2 + \sigma^2(b, Z_{2i})) + \frac{1}{n_b(n_b-1)} \sum_{i \neq j \in b} m(b, Z_{2i}) m(b, Z_{2j}) \end{aligned} \tag{C.8}$$

Suppose the above quantity is unbounded across all $P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}$. Then, any vector $c \in$

$\oplus_{b=1}^{B} W_{0,b}$ has

$$\sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} c' E[\mathbf{H}\mathbf{U}\mathbf{U}'\mathbf{H}'|\mathbf{Z}]c = \infty$$

and hence doesn't belong to C in (C.3). I therefore restrict $C \subseteq \bigcup_{1 \leq b \leq B} W_{1,b}$ and $Q = \sum_{1 \leq b \leq B} Q_{1,b}$.

Now let j denote the double index $(1, b)$,

$$QOQ = \sum_j \frac{\operatorname{tr} Q_j O}{\operatorname{tr} Q_j} Q_j \ , \tag{C.9}$$

and since Q_j is idempotent,

$$\operatorname{tr} Q_j O = E[\|Q_j \mathbf{H}\mathbf{U}\|^2] \ .$$

One could then define

$$V = \sum_{1 \leq b \leq B} \sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} E[\|Q_{1,b}\mathbf{H}\mathbf{U}\|^2|\mathbf{Z}]$$

if the supremum in all B terms are reached at the same $P_{\mathbf{U}}$. This is a nontrivial requirement. If so, the diagonal of V for $i \in b$ is

$$\begin{aligned} V_b &= \frac{n_b - 1}{n_b^2} \sum_{i \in b} E[U_i^2|Z_i] - \frac{1}{n_b^2} \sum_{i \neq j \in b} E[U_i U_j|Z_i, Z_j] \\ &= \frac{1}{n_b} \sum_{i \in b} (m(b, Z_{2i})^2 + \sigma^2(b, Z_{2i})) - \frac{1}{n_b^2} \sum_{i,j \in b} m(b, Z_{2i}) m(b, Z_{2j}) \ , \end{aligned} \tag{C.10}$$

where the expectations are with respect to $P_{\mathbf{U}}$. The same expression appears in Li (1983).

Note that (C.5) requires $Q\beta = \beta$, which means that $\sum_{l \in b} \beta_l = 0$, so the coefficients have to sum up to zero within each block, and the estimator is a weighted average of within-block differences. $\beta'\mathbf{A} = \tau'$ is an unbiasedness condition that on average treatment and control has to be balanced. The researcher solves (C.5) and then completely randomizes the $(\mathbf{A}_0, \beta_0)$ within blocks, obtaining the minimax solution.

In particular, if all V_b's and n_b's' are the same across $b = 1, \ldots, B$, one could verify

the difference-in-means estimator with half of each block assigned to treatment is minimax-optimal. ■

C.2 Proofs of Theorems

Proof of Theorem 3.2.1. The proof follows Li (1983). It is straight forward to see that I need only prove that for any $s_0 \in S$,

$$\min_{\mathbf{G}\in\Phi}\max_{\pi\in H} f(r(\mathbf{G},\mathbf{A},\delta;\pi s_0)) = \max_{\pi\in H} f(r(\mathbf{H},\mathbf{A},\delta;\pi s_0)) \ , \tag{C.11}$$

i.e. I need only focus on the orbit of s_0. Note that the orbit could be defined because of Assumption 3.2.1. Next, for any $\tilde{\pi} \in H$,

$$\begin{aligned} r(\tilde{\pi},\mathbf{A},\delta;\pi s_0) = E_{\pi s_0}[L(\pi s_0,\delta(\mathbf{Y}^{\tilde{\pi}}))|\mathbf{Z}] &= E_{\pi s_0}[L(s_0,\delta(\mathbf{Y}^{\tilde{\pi}}))|\mathbf{Z}] \\ &= E_{s_0}[L(s_0,\delta(\mathbf{Y}^{\tilde{\pi}\pi}))|\mathbf{Z}] = r(\tilde{\pi}\pi,\mathbf{A},\delta;s_0) \ , \end{aligned}$$

where the first equality is definition, the second equality holds by Assumption 3.2.2, the third equality holds by (3.7), and the last is again definition. Therefore

$$r(\mathbf{G},\mathbf{A},\delta;\pi s_0) \overset{d}{=} r(\mathbf{G}\pi,\mathbf{A},\delta;s_0) \tag{C.12}$$

and

$$r(\mathbf{H},\mathbf{A},\delta;\pi s_0) \overset{d}{=} r(\mathbf{H}\pi,\mathbf{A},\delta;s_0) \overset{d}{=} r(\mathbf{H},\mathbf{A},\delta;s_0) \tag{C.13}$$

for any $\pi \in H$ since $\mathbf{H}$ is the uniform distribution on H. Note also that

$$\frac{1}{|H|}\sum_{\pi\in H} r(\mathbf{G}\pi,\mathbf{A},\delta;s_0) \overset{d}{=} r(\mathbf{G}\mathbf{H},\mathbf{A},\delta;s_0) \overset{d}{=} r(\mathbf{H},\mathbf{A},\delta;s_0) \ . \tag{C.14}$$

again because $\mathbf{H}$ is the uniform distribution on H. Combining (C.12), (C.13) and (C.14), we have that for any $\tilde{\pi} \in H$,

$$\begin{aligned}
&\max_{\pi \in H} f(r(\mathbf{G}, \mathbf{A}, \delta; \pi s_0)) \\
&\geq f\left(\frac{1}{|H|} \sum_{\pi \in H} r(\mathbf{G}, \mathbf{A}, \delta; \pi s_0)\right) \\
&= f(r(\mathbf{H}, \mathbf{A}, \delta; s_0)) \\
&= f(r(\mathbf{H}, \mathbf{A}, \delta; \tilde{\pi} s_0)) \qquad \text{(C.15)} \\
&= \max_{\pi \in H} f(r(\mathbf{H}, \mathbf{A}, \delta; \pi s_0))
\end{aligned}$$

for any $\tilde{\pi} \in H$, where the inequality holds by Assumption 3.2.3 and that $\mathbf{G}$ induces a measure on real line in terms of risk $r(\tilde{\pi}, \mathbf{A}, \delta; \pi s_0)$ across $\tilde{\pi} \in H$, the first equality holds by (C.12) and (C.14), the second equality holds by (C.13) and the last equality holds because (C.15) holds for each $\tilde{\pi}$ and so maximum over $\tilde{\pi}$ is equal to any one of them. ■

Proof of Theorem 3.3.1.

$$E[\|\beta' \mathbf{Y}^{\pi} - \tau' \mu\|^2 | \mathbf{Z}] = E[\|(\beta' \mathbf{A} \mathbf{1}_n \otimes \mu - \tau' \mu) + \beta' \mathbf{A} \operatorname{vec}(\mathbf{U}' \mathbf{G}')\|^2 | \mathbf{Z}] \ .$$

Since the maximum is over $\mu \in \mathbb{R}^2$, I get unbounded maximum risk unless $\beta' \mathbf{A} \mathbf{1}_n \otimes \mu = \tau' \mu$ for all $\mu \in \mathbb{R}^2$. This is equivalent to the two conditions that

$$\begin{aligned}
\beta' \mathbf{A} \mathbf{1}_n \otimes (1, 0)' &= \tau_0 \\
\beta' \mathbf{A} \mathbf{1}_n \otimes (0, 1)' &= \tau_1 \ .
\end{aligned} \qquad \text{(C.16)}$$

Given that (C.16) holds, I have

$$\begin{aligned}
&E[\|\beta' \mathbf{A} \operatorname{vec}(\mathbf{U}' \mathbf{G}')\|^2 | \mathbf{Z}] \\
&= E[\|\beta' \mathbf{A} (\mathbf{G} \otimes I_2) \operatorname{vec}(\mathbf{U}')\|^2 | \mathbf{Z}]
\end{aligned}$$

$$= E[\beta' \mathbf{A}(\mathbf{G} \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\mathbf{G}' \otimes I_2)\mathbf{A}'\beta | \mathbf{Z}]$$

and therefore for any randomization scheme $\mathbf{G}$ a distribution on H,

$$\begin{aligned}
&\sup_{(\beta, P_{\mathbf{U}}) \in (\mathbb{R}^2 \times \mathcal{P}_{\mathbf{U}})} E[\|\beta' \mathbf{Y}^\pi - \tau' \mu\|^2 | \mathbf{Z}] \\
&= \sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} E[\beta' \mathbf{A}(\mathbf{G} \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\mathbf{G}' \otimes I_2)\mathbf{A}'\beta | \mathbf{Z}] \\
&= \sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} E[\beta' \mathbf{A}(\pi\mathbf{G} \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\mathbf{G}'\pi' \otimes I_2)\mathbf{A}'\beta | \mathbf{Z}]
\end{aligned}$$

for all $\pi \in H$ since $P_{\mathbf{U}}$ is H-invariant by Assumption 3.3.1. Therefore the above is equal to

$$\begin{aligned}
&\frac{1}{|H|} \sum_{\pi \in H} \sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} E[\beta' \mathbf{A}(\pi\mathbf{G} \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\mathbf{G}'\pi' \otimes I_2)\mathbf{A}'\beta | \mathbf{Z}] \\
&\geq \sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} \frac{1}{|H|} \sum_{\pi \in H} E[\beta' \mathbf{A}(\pi\mathbf{G} \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\mathbf{G}'\pi' \otimes I_2)\mathbf{A}'\beta | \mathbf{Z}] \\
&= \sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} E[\beta' \mathbf{A}(\mathbf{H}\mathbf{G} \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\mathbf{G}'\mathbf{H}' \otimes I_2)\mathbf{A}'\beta | \mathbf{Z}] \\
&= \sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} E[\beta' \mathbf{A}(\mathbf{H} \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\mathbf{H}' \otimes I_2)\mathbf{A}'\beta] \\
&= \sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} E[\|\beta' \mathbf{A} \operatorname{vec}(\mathbf{U}'\mathbf{H}')\|^2 | \mathbf{Z}] \qquad \text{(C.17)}
\end{aligned}$$

where the second equality holds because $\mathbf{H}$ is the uniform distribution on H so that $\mathbf{H}\mathbf{G} \overset{d}{=} \mathbf{H} \overset{d}{=} \mathbf{G}\mathbf{H}$. I have shown that for each scheme $\mathcal{L}(\mathbf{A}, \beta, \pi)$ there exists a fully randomized scheme that dominates it in terms of minimax risk. But I still need to choose $\mathcal{L}(\mathbf{A}, \beta, \pi)$ to minimize (C.17). To this end, note that $\mathbf{A}'\beta \in C$ for C in (3.11) because otherwise the maximum risk is infinite. So (C.17) becomes

$$\sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} E[\beta' \mathbf{A} Q \operatorname{vec}(\mathbf{H} \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\mathbf{H}' \otimes I_2) Q \mathbf{A}'\beta | \mathbf{Z}]$$

Since

$$\sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} E[Q \operatorname{vec}(\mathbf{H} \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\mathbf{H}' \otimes I_2) Q | \mathbf{Z}] = V \ ,$$

the conclusion follows. ■

Lemma C.2.1. *C is H-invariant, i.e., $c \in C \Rightarrow \pi c \in C$ for all $c \in \mathbb{R}^{2n}$ and $\pi \in H$.*

Proof of Lemma C.2.1. Suppose

$$\sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} c' E[(\mathbf{H} \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\mathbf{H}' \otimes I_2) | \mathbf{Z}] c < \infty$$

one need to prove that for any $\pi \in H$,

$$\sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} c' \pi' E[(\mathbf{H} \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\mathbf{H}' \otimes I_2) | \mathbf{Z}] \pi' c < \infty \tag{C.18}$$

However, since $\mathcal{P}_{\mathbf{U}}$ is H-invariant, (C.18) holds for $\mathbf{U}$ replaced by $\pi\mathbf{U}$ for any $\pi \in H$. Note that

$$\begin{aligned}
&E[(\mathbf{H} \otimes I_2) \operatorname{vec}(\mathbf{U}'\pi') \operatorname{vec}(\mathbf{U}'\pi')'(\mathbf{H}' \otimes I_2) | \mathbf{Z}] \\
&= E_{\tilde{\pi} \sim \mathbf{H}}[(\tilde{\pi} \otimes I_2)(\pi \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\pi' \otimes I_2)(\tilde{\pi}' \otimes I_2) | \mathbf{Z}] \\
&= E_{\tilde{\pi} \sim \mathbf{H}}[(\tilde{\pi}\pi \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\pi'\tilde{\pi}' \otimes I_2) | \mathbf{Z}] \\
&= E_{\tilde{\pi} \sim \mathbf{H}}[(\tilde{\pi} \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\tilde{\pi}' \otimes I_2) | \mathbf{Z}] \ ,
\end{aligned}$$

where the first equality holds because $\operatorname{vec}(ABC) = (C' \otimes A) \operatorname{vec}(B)$, the second equality holds since $(A \otimes B)(C \otimes D) = AB \otimes CD$, and the last holds since $\mathbf{H}$ is the uniform distribution on H. As a result, (C.18) holds. ■

Recall that a permutation matrix is orthogonal, i.e. $\pi'\pi = \pi\pi' = I$.

Lemma C.2.2. *$Q\pi = \pi Q$ or equivalently since π is orthogonal, $\pi Q \pi' = Q$.*

Proof of Lemma C.2.2. For $c \in C$, $\pi Q \pi^{-1} c = \pi \pi^{-1} c = c$ since C is H-invariant. For an element $b \in C^{\perp}$, $\pi Q b = 0$ since $Qb = 0$. Now, by definition of $C^{\perp}$, $b'c = 0$ for all $c \in C$, and therefore $b'\pi' c = 0$ since $\pi' c \in C$ for all $c \in C$. As a result, $\pi b \in C^{\perp}$ as well, and hence $Q\pi b = 0$. As a result, $\pi Q \pi^{-1} c = Qc$ for all $c \in \mathbb{R}^n$ and hence $\pi Q \pi^{-1} = Q$. ■

REFERENCES

Aker, J. C., Ksoll, C. and Lybbert, T. J. (2012). Can mobile phones improve learning? Evidence from a field experiment in Niger. *American Economic Journal: Applied Economics*, **4** 94–120.

Alatas, V., Banerjee, A., Hanna, R., Olken, B. A. and Tobias, J. (2012). Targeting the poor: Evidence from a field experiment in Indonesia. *American Economic Review*, **102** 1206–40.

Angrist, J. and Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American Economic Review*, **99** 1384–1414.

Armitage, P., Berry, G. and Matthews, J. N. S. (2008). *Statistical methods in medical research*. John Wiley & Sons.

Armstrong, T. B. and Kolesár, M. (2018). Optimal inference in a class of regression models. *Econometrica*, **86** 655–683.

Artin, M. (2013). *Algebra*. Pearson Education, Limited.

Ashraf, N., Berry, J. and Shapiro, J. M. (2010). Can higher prices stimulate product use? Evidence from a field experiment in Zambia. *American Economic Review*, **100** 2383–2413.

Ashraf, N., Karlan, D. and Yin, W. (2006). Deposit collectors. *Advances in Economic Analysis & Policy*, **5**.

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

Bailey, R. A. (2004). *Association schemes: Designed experiments, algebra and combinatorics*, vol. 84. Cambridge University Press.

Banerjee, A., Chassang, S., Montero, S. and Snowberg, E. (2019). A theory of experimenters.

Banerjee, A., Duflo, E., Glennerster, R. and Kinnan, C. (2015). The miracle of microfinance? Evidence from a randomized evaluation. *American Economic Journal: Applied Economics*, **7** 22–53.

Barrios, T. (2013). Optimal stratification in randomized experiments. Working paper.

Bellec, P. C., Dalalyan, A. S., Grappin, E., Paris, Q. and others (2018). On the prediction loss of the lasso in the partially labeled setting. *Electronic Journal of Statistics*, **12** 3443–3472.

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, **80** 2369–2429.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, **81** 608–650.

Berry, J., Karlan, D. and Pradhan, M. (2018). The impact of financial education for youth in Ghana. *World Development*, **102** 71–89.

Bertrand, M., Djankov, S., Hanna, R. and Mullainathan, S. (2007). Obtaining a driver's license in India: An experimental approach to studying corruption. *The Quarterly Journal of Economics*, **122** 1639–1676.

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

Bertsimas, D. and Tsitsiklis, J. N. (1997). *Introduction to linear optimization*, vol. 6.

Bhargava, S. and Manoli, D. (2015). Psychological frictions and the incomplete take-up of social benefits: Evidence from an IRS field experiment. *American Economic Review*, **105** 3489–3529.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, **13** 1063–1095.

Bogachev, V. I. (2007). *Measure theory*, vol. 1. Springer Science & Business Media.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A. and Sandefur, J. (2018). Experimental evidence on scaling up education reforms in Kenya. *Journal of Public Economics*, **168** 1–20.

Bruhn, M., Leão, L. d. S., Legovini, A., Marchetti, R. and Zia, B. (2016). The impact of high school financial education: Evidence from a large-scale evaluation in Brazil. *American Economic Journal: Applied Economics*, **8** 256–295.

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. Publisher: American Economic Association.

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. *Quantitative Economics*. Forthcoming.

Bursztyn, L., Ferman, B., Fiorin, S., Kanz, M. and Rao, G. (2018). Status goods: Experimental evidence from platinum credit cards. *The Quarterly Journal of Economics*, **133** 1561–1595.

Bursztyn, L., Fiorin, S., Gottlieb, D. and Kanz, M. (2019). Moral incentives in credit card debt repayment: Evidence from a field experiment. *Journal of Political Economy*. Forthcoming.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer Science & Business Media.

Callen, M., Gulzar, S., Hasanain, A., Khan, M. Y. and Rezaee, A. (2018). Data and policy decisions: Experimental evidence from Pakistan. Working paper.

Callen, M., Isaqzadeh, M., Long, J. D. and Sprenger, C. (2014). Violence and risk preference: Experimental evidence from Afghanistan. *American Economic Review*, **104** 123–48.

Carneiro, P., Lee, S. and Wilhelm, D. (2019). Optimal data collection for randomized control trials. *The Econometrics Journal*. Forthcoming.

Casaburi, L. and Macchiavello, R. (2019). Demand and supply of infrequent payments as a commitment device: Evidence from Kenya. *American Economic Review*, **109** 523–55.

Chambaz, A., van der Laan, M. J. and Zheng, W. (2015). Targeted covariate-adjusted response-adaptive LASSO-based randomized controlled trials. *Modern Adaptive Randomized Clinical Trials: Statistical, Operational, and Regulatory Aspects* 345–368.

Chatterjee, S. (2013). Assumptionless consistency of the Lasso. *arXiv preprint arXiv:1303.5817*.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. vol. 6 of *Handbook of Econometrics*. Elsevier, 5549 – 5632.

Chen, X. and White, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, **45** 682–691.

Chen, Y. and Yang, D. Y. (2019). The impact of media censorship: 1984 or Brave New World? *American Economic Review*, **109** 2294–2332.

Chernozhukov, V., Fernández-Val, I. and Luo, Y. (2018). The sorted effects method: Discovering heterogeneous effects beyond their averages. *Econometrica*, **86** 1911–1938.

Chernozhukov, V., Newey, W. and Santos, A. (2015). Constrained conditional moment restriction models.

Chetverikov, D., Santos, A. and Shaikh, A. M. (2018). The econometrics of shape restrictions. *Annual Review of Economics*, **10** 31–63.

Chong, A., Cohen, I., Field, E., Nakasone, E. and Torero, M. (2016). Iron deficiency and schooling attainment in Peru. *American Economic Journal: Applied Economics*, **8** 222–55.

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

Cox, D. and Reid, N. (2000). *The theory of the design of experiments*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, CRC Press.

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.

de Chaisemartin, C. and Ramirez-Cuellar, J. (2019). At what level should one cluster standard errors in paired experiments? *arXiv preprint arXiv:1906.00288*.

DellaVigna, S. and Pope, D. (2018). What motivates effort? Evidence and expert forecasts. *The Review of Economic Studies*, **85** 1029–1069.

Denil, M., Matheson, D. and De Freitas, N. (2014). Narrowing the gap: Random forests in theory and in practice. In *International Conference on Machine Learning*. 665–673.

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

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. Publisher: Taylor & Francis.

Dizon-Ross, R. (2019). Parents' beliefs about their children's academic ability: Implications for educational investments. *American Economic Review*, **109** 2728–2765.

Donoho, D. L. (1994). Statistical estimation and optimal recovery. *The Annals of Statistics*, **22** 238–270.

Duflo, E. and Banerjee, A. (2017). *Handbook of field experiments*. Elsevier Science.

Duflo, E., Dupas, P. and Kremer, M. (2015a). Education, HIV, and early fertility: Experimental evidence from Kenya. *American Economic Review*, **105** 2757–97.

Duflo, E., Dupas, P. and Kremer, M. (2015b). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools. *Journal of Public Economics*, **123** 92–110.

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.

Dupas, P., Karlan, D., Robinson, J. and Ubfal, D. (2018). Banking the unbanked? Evidence from three countries. *American Economic Journal: Applied Economics*, **10** 257–97.

Dupas, P. and Robinson, J. (2013). Savings constraints and microenterprise development: Evidence from a field experiment in Kenya. *American Economic Journal: Applied Economics*, **5** 163–92.

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.

Farrell, M. H., Liang, T. and Misra, S. (2018). Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. *arXiv preprint arXiv:1809.09953*.

Fogarty, C. B. (2018a). On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80** 1035–1056.

Fogarty, C. B. (2018b). Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika*, **105** 994–1000.

Freedman, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, **40** 180–193.

Fryer, J., Roland G, Devi, T. and Holden, R. T. (2017). Vertical versus horizontal incentives in education: Evidence from randomized trials. Working paper.

Fryer, R. (2017). Management and student achievement: Evidence from a randomized field experiment. Working paper.

Fryer, R. (2018). The "pupil" factory: Specialization and the production of human capital in schools. *American Economic Review*, **108** 616–656.

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.

Glewwe, P., Park, A. and Zhao, M. (2016). A better vision for development: Eyeglasses and academic performance in rural primary schools in China. *Journal of Development Economics*, **122** 170–182.

Groh, M. and McKenzie, D. (2016). Macroinsurance for microenterprises: A randomized experiment in post-revolution Egypt. *Journal of Development Economics*, **118** 13–25.

Grötschel, M. and Wakabayashi, Y. (1990). Facets of the clique partitioning polytope. *Mathematical Programming*, **47** 367–387.

Györfi, L., Kohler, M., Krzyzak, A. and Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Hahn, J., Hirano, K. and Karlan, D. (2011). Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics*, **29** 96–108.

Heard, K., O'Toole, E., Naimpally, R. and Bressler, L. (2017). *Real world challenges to randomization and their solutions*. Boston, MA: Abdul Latif Jameel Poverty Action Lab.

Heckman, J. J., Pinto, R., Shaikh, A. M. and Yavitz, A. (2011). Inference with imperfect randomization: The case of the Perry Preschool Program. Working paper.

Hooper, P. M. (1989). Minimaxity of randomized optimal designs. *The Annals of Statistics*, **17** 1315–1324.

Horton, J. J., Rand, D. G. and Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, **14** 399–425.

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

Imai, K., King, G., Nall, C. and others (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.

Imbens, G. W. (2011). Experimental design for unit and cluster randomized trials.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

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

Johansson, P., Schultzberg, M. a. and Rubin, D. (2019). On optimal rerandomization designs. Working paper.

Kallus, N. (2018). Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80** 85–112.

Karlan, D. and Appel, J. (2016). *Failing in the field: What we can learn when field research goes wrong*. Princeton University Press.

Karlan, D. and Wood, D. H. (2017). The effect of effectiveness: Donor response to aid effectiveness in a direct mail fundraising experiment. *Journal of Behavioral and Experimental Economics*, **66** 1–8.

Karlan, D. S. and Zinman, J. (2008). Credit elasticities in less-developed economies: Implications for microfinance. *American Economic Review*, **98** 1040–68.

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis*, **24** 324–338.

Khan, A. Q., Khwaja, A. I. and Olken, B. A. (2019). Making moves matter: Experimental evidence on incentivizing bureaucrats through performance-based postings. *American Economic Review*, **109** 237–70.

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica*, **86** 591–616.

Kuziemko, I., Norton, M. I., Saez, E. and Stantcheva, S. (2015). How elastic are preferences for redistribution? Evidence from randomized survey experiments. *American Economic Review*, **105** 1478–1508.

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.

Li, K.-C. (1983). Minimaxity for randomized designs: Some general results. *The Annals of Statistics*, **11** 225–239.

Li, Q. and Racine, J. S. (2007). *Nonparametric econometrics: Theory and practice*. Princeton University Press.

Li, X., Ding, P. and Rubin, D. B. (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, **115** 9157–9162.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, **7** 295–318.

Lin, Y., Zhu, M. and Su, Z. (2015). The pursuit of balance: An overview of covariate-adaptive randomization techniques in clinical trials. *Contemporary Clinical Trials*, **45** 21–25.

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

Litman, L., Robinson, J. and Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, **49** 433–442.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*, **72** 1221–1246.

Mason, W. and Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, **44** 1–23.

Mbakop, E. and Tabord-Meehan, M. (2018). Model selection for treatment choice: Penalized welfare maximization. Working paper.

Morgan, K. L. and Rubin, D. B. (2015). Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association*, **110** 1412–1421.

Morgan, K. L., Rubin, D. B. and others (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, **40** 1263–1282.

Moses, L. E. (2006). Matched pairs t-tests. In *Encyclopedia of Statistical Sciences*. American Cancer Society.

Munkres, J. R. (1997). *Analysis on manifolds*. Westview Press.

Muralidharan, K., Singh, A. and Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review*, **109** 1426–60.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, **79** 147–168.

Olken, B. A. (2007). Monitoring corruption: Evidence from a field experiment in Indonesia. *Journal of political Economy*, **115** 200–249.

Panagopoulos, C. and Green, D. P. (2008). Field experiments testing the impact of radio advertisements on electoral competition. *American Journal of Political Science*, **52** 156–168.

Paolacci, G. and Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, **23** 184–188.

Peters, J., Langbein, J. and Roberts, G. (2016). Policy evaluation, randomized controlled trials, and external validity—A systematic review. *Economics Letters*, **147** 51–54.

Pukelsheim, F. (2006). *Optimal design of experiments*. Classics in Applied Mathematics, Society for Industrial and Applied Mathematics.

Riach, P. A. and Rich, J. (2002). Field experiments of discrimination in the market place. *The Economic Journal*, **112** F480–F518.

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

Rudin, W. (1976). *Principles of mathematical analysis*, vol. 3. McGraw-hill New York.

Schultzberg, M. a. and Johansson, P. (2019). Optimal designs and asymptotic inference. Working paper.

Scornet, E., Biau, G., Vert, J.-P. and others (2015). Consistency of random forests. *The Annals of Statistics*, **43** 1716–1741.

Sondheimer, R. M. and Green, D. P. (2010). Using experiments to estimate the effects of education on voter turnout. *American Journal of Political Science*, **54** 174–189.

Spivak, M. (1965). *Calculus on manifolds*.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Tabord-Meehan, M. (2020). Stratification trees for adaptive randomization in randomized controlled trials. Working paper.

Wager, S. and Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.

White, H. (1990). Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, **3** 535–549.

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

Wu, C.-F. (1981). On the robustness and efficiency of some randomized designs. *The Annals of Statistics*, **9** 1168–1177.