Optimality of Matched-Pair Designs in Randomized Controlled Trials*

Yuehao Bai
Department of Economics
University of Michigan
yuehaob@umich.edu

November 23, 2020

Abstract

This paper 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 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.

KEYWORDS: Matched-pair design, stratified randomization, randomized controlled trial, ex-post bias, treatment effect, stratification, pilot experiment, matched pairs

JEL CLASSIFICATION CODES: C12, C13, C14, C90

*I am deeply grateful for the encouragement and guidance from my advisors Azeem Shaikh, Stephane Bonhomme, Alex Torgovitsky, and Leonardo Bursztyn. I thank Marinho Bertanha, Wooyong Lee, Joshua Shea, and Max Tabord-Meehan for extensive feedback on earlier drafts of the paper. I would also like to thank seminar participants at many institutions for helpful comments on the paper. I gratefully acknowledge the financial support from the William Rainey Harper/Provost Dissertation Year Fellowship.

1 Introduction

This paper 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 more than 600 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. 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 index function of the covariates and paired adjacently. The index function is defined by the sum of the expectations of potential outcomes if treated and not treated conditional on the covariates. To the best of our knowledge, our result is the first to characterize the optimal one among all stratified randomization procedures, and additionally, it holds under almost no assumption on the distributions of potential outcomes, and in particular, does not rely on the knowledge of conditional variances of the potential outcomes given the covariates. In some special cases, and for instance when there is only one covariate and the index function is monotonic in the covariate, we know the optimal stratification even without knowing the value of the index function. We further show that the optimality of matched-pair designs, though possibly in a different form, holds under any expected utility criterion, and even any criterion convex in the distribution of treatment status. See Remark 3.4 for more details. Although one could go one step further to deterministic treatment assignments, it will make the difference-in-means estimator neither unbiased nor consistent for the ATE and frequentist inference impossible. We also observe that very few, if any, experiments in the AEA RCT Registry use deterministic assignment, but many of them, listed in Appendix S.6, use matched-pair designs in various forms.

We then study the properties of empirical counterparts to this optimal stratification, in which we replace the unknown index 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 1.1. We first consider a plug-in procedure that estimates the index 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, or more precisely, that it is L^2-consistent for the index function, 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 requirement 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. But even then, researchers may well be content with the plug-in procedure because it results in smaller limiting variance of the difference-in-means estimator than many alternatives. That said, 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. Inference for matched-pair designs is challenging because of the difficulty of consistently estimating the limiting variance of the ATE. Indeed, this is the main reason why Athey and Imbens (2017) suggest not to use matched-pair designs. We get around this problem by a novel standard error adjustment and Lipschitz conditions that guarantee the smoothness of conditional expectations of potential outcomes given the covariates. This condition, together with the observation that paired observations become close in terms of the pairing covariate in the limit, enables us to estimate the limiting variance consistently. Therefore, 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

Bai et al. (2019) 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 where 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.

The remainder of the paper is organized as follows. In Section 2, we introduce the setup and notation. We study the optimal stratification in Section 3. In Section 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 5 includes asymptotic results and methods for inference for ATE. In Section 6, we illustrate the properties of different procedures in a small simulation study. Section 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 8 briefly discusses the minimax procedure, the details of which are included in Appendix S.5. We conclude with recommendations for empirical practice in Section 9.

1.1 Related literature

This paper is most closely related to Barrios (2013) and Tabord-Meehan (2020). In a closely related paper, Barrios (2013) considers minimizing the variance of the difference-in-means estimator. Despite having "optimal stratification" in the title of his paper, he only shows that a certain matched-pair design is optimal among all matched-pair designs, instead of all stratified randomization procedures. Although intuitively attractive, it is not always without loss of generality to restrict attention to matched-pair designs in the first place. Example S.5.7 shows that under a minimax criterion the optimal stratification might not be a matched-pair design. We show, however, that we could restrict attention to matched-pair designs if the criterion is MSE. In fact, we show that the optimality of matched-pair designs holds under any expected utility criterion, and even any criterion convex in the distribution of treatment status. See Remark 3.4 for more details. Moreover, Barrios (2013) assumes a homogeneous treatment effect and uses only information about untreated potential outcomes in his analysis, while our optimality result instead holds under heterogeneous treatment effects. Finally, we provide novel results relating the MSE to the ex-post bias, as well as novel results on the large sample properties of empirical counterparts to the optimal procedure and 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 5.9 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 limiting 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.

Our paper is also related to a series of paper studying regression adjustments in RCTs, including Wager et al. (2016) and Spiess (2020). In fact, the limiting variance under the optimal stratification, as well under the plug-in procedure with a large pilot, both coincide with Hahn (1998)'s semiparametric efficiency bound. We emphasize, however, that stratified randomization procedures have better properties in finite sample, and it is a pre-specification device to guard against data mining. See Glennerster and Takavarasha (2013) for more details. In fact, Section 5 of Athey and Imbens (2017) recommends caution for using regression adjustment, and say "in many cases the potential gains from regression adjustment can also be captured by careful ex-ante design, that is, through stratified randomized experiments ...without the potential costs associated with ex-post regression adjustment." We also want to emphasize that our optimality result holds in finite sample, and is known without any estimation when for instance the conditional expectations are monotonic in the scalar covariate. Even when we need to estimate the conditional expecations using pilot data, the difference-in-means estimator is unbiased, and our inference procedure is asymptotically valid regardless of the sample size of the pilot. Appendix S.2 further provides a straightforward way to extend our procedure in order to prespecify targeted subgroups. Finally, Remark 5.2 shows it is straightforward to combine stratification and regression adjustment, if additional covariates become available after the treatment assignment.

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. (2015, page 96), Duflo et al. (2015, 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), Kasy and Lehner (2020), and the references therein. See Appendix S.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 limiting variance of the estimator is hard to estimate. Our inference procedure solves this problem and therefore eliminates this cost. Besides Bai et al. (2019), inference under matched-pair designs has also been studied in Abadie and Imbens (2008), who consider another adjustment of standard error, in Fogarty (2018a) and Fogarty (2018b), who provides conservative estimators for the limiting variance, and de Chaisemartin and Ramirez-Cuellar (2019), under a sampling scheme different from that in Bai et al. (2019) 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. As we already mentioned, under a determistic assignment, the difference-in-means estimator is neither unbiased nor consistent, and frequentist inference is impossible. Furthermore, for Kasy (2016)'s results to hold, one needs to fully specify the prior including that of the conditional variances of the potential outcomes given the covariates, while our optimality result only relies on the conditional expectations. On the contrary, Wu (1981), Li (1983), and Hooper (1989), and Bai (2020) 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 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 $X_i = (X_{i,1}, \dots, X_{i,p})' \in \mathbf{R}^p$ 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 $A^{(n)} = (A_1, \dots, A_{2n})'$. Our parameter of interest is the average treatment effect (ATE) under Q:

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

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 $\{1, \dots, 2n\}$, i.e.,

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

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

Let Λ_n denote the set of all stratifications of $2n$ units. Many results in the paper will feature matched-pair designs. Recall that a permutation of $\{1, \dots, 2n\}$ is a function that maps $\{1, \dots, 2n\}$ onto itself.

Let Π_n denote the group of all permutations of $\{1, \ldots, 2n\}$. A matched-pair design is a stratified randomization with

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

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 $(\tau_s : 1 \le s \le S)$, the treatment assignment scheme is as follows: independently for $1 \le s \le S$, 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

$$(Y^{(n)}(0), Y^{(n)}(1)) \perp\!\!\!\perp D^{(n)} | X^{(n)} . \tag{2}$$

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

Assumption 2.1. The treatment assignment scheme satisfies $\tau_s \equiv \frac{1}{2}$.

Assumption 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 S.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

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

The difference-in-means estimator is defined as

$$\hat{\theta}_n = \hat{\mu}_n(1) - \hat{\mu}_n(0) . \tag{3}$$

The difference-in-means estimator is widely used because it is simple and transparent. Under Assumption 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).

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:

$$\mathrm{MSE}(\lambda|X^{(n)}) = E_\lambda[(\hat{\theta}_n - \theta)^2|X^{(n)}] \ . \tag{4}$$

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 (4) over the set of all stratifications:

$$\min_{\lambda \in \Lambda_n} \ \mathrm{MSE}(\lambda|X^{(n)}) \ . \tag{5}$$

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 4. By Assumption 2.1, other aspects of the stratified randomization procedure, especially the treated fractions, are fixed. Therefore, the stratification that solves (5) corresponds to an optimal stratified randomization procedure among all those satisfying Assumption 2.1.

In order to describe an important result that leads to the solution to (5), we define the ex-ante bias of $\hat{\theta}_n$ for θ conditional on $X^{(n)}$ as

$$\mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}) = E_\lambda[\hat{\theta}_n|X^{(n)}] - \theta \ , \tag{6}$$

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

$$\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)}) = E_\lambda[\hat{\theta}_n|X^{(n)}, D^{(n)}] - \theta \ . \tag{7}$$

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,

$$E_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})|X^{(n)}] = \mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}) \ , \tag{8}$$

i.e., the expectation of the ex-post bias over the distribution of treatment status equals the ex-ante bias. Note that by (3),

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

Under Assumption 2.1,

$$\mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}) = \frac{1}{2n} \sum_{1 \le i \le 2n} (E[Y_i(1)|X_i] - E[Y_i(0)|X_i]) - \theta \ , \tag{9}$$

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

To solve (5), we decompose the conditional MSE as follows. First, note that

$$\mathrm{MSE}(\lambda|X^{(n)}) = \mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)})^2 + \mathrm{Var}_\lambda[\hat{\theta}_n|X^{(n)}] \,. \tag{10}$$

Here, Var_λ indicates that the distribution of treatment status depends on λ. By (9), the first term on the right-hand side is identical across all $\lambda \in \Lambda_n$. Hence, (5) is equivalent to minimizing the second term on the right-hand side of (10), which could be further decomposed into

$$\mathrm{Var}_\lambda[\hat{\theta}_n|X^{(n)}] = E_\lambda[\mathrm{Var}[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] + \mathrm{Var}_\lambda[E[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] \,. \tag{11}$$

By (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 (11) equals

$$\begin{aligned} E_\lambda\left[\frac{1}{n^2}\sum_{1\le i\le 2n}(\mathrm{Var}[Y_i(1)|X_i]D_i + \mathrm{Var}[Y_i(0)|X_i](1-D_i))\middle| X^{(n)}\right] \\ = \frac{1}{2n^2}\sum_{1\le i\le 2n}(\mathrm{Var}[Y_i(1)|X_i] + \mathrm{Var}[Y_i(0)|X_i]) \,, \end{aligned} \tag{12}$$

which is also identical across all $\lambda \in \Lambda_n$. Here, we use (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, (5) is further equivalent to minimizing the second term on the right-hand side of (11), which equals

$$\mathrm{Var}_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})|X^{(n)}] \,. \tag{13}$$

Furthermore, we have

$$\begin{aligned}
&\mathrm{Var}_\lambda[E[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] \\
&= E_\lambda[(E[\hat{\theta}_n|X^{(n)}, D^{(n)}] - E[\hat{\theta}_n|X^{(n)}])^2|X^{(n)}] \\
&= E_\lambda[(\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)}) - \mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}))^2|X^{(n)}] \\
&= E_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})^2|X^{(n)}] - 2E_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})\,\mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)})|X^{(n)}] \\
&\qquad + \mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)})^2 && (14) \\
&= E_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})^2|X^{(n)}] - 2E_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})|X^{(n)}]\,\mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}) \\
&\qquad + \mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)})^2 && (15) \\
&= E_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})^2|X^{(n)}] - 2\,\mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)})^2 + \mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)})^2 \\
&= E_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n|X^{(n)}, D^{(n)})^2|X^{(n)}] - \mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)})^2 \,, && (16)
\end{aligned}$$

where the first equality follows from definition, the second follows from (6) and (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 (8) . By (9), $\mathrm{Bias}^{\mathrm{ante}}_{n,\lambda}(\hat{\theta}_n|X^{(n)})$ is the same across λ, and therefore it follows from (10)–(16) that (5) is equivalent to minimizing the first term in (16), i.e., the second moment of the ex-post bias. We summarize the results in the following lemma:

Lemma 3.1. *Suppose the treatment assignment scheme satisfies Assumption 2.1. Then, the set of solutions to (5) is the same as the set of solutions to*

$$\min_{\lambda \in \Lambda_n} E_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n | X^{(n)}, D^{(n)})^2 | X^{(n)}] \,, \tag{17}$$

and the set of solutions to

$$\min_{\lambda \in \Lambda_n} \mathrm{Var}_\lambda[\mathrm{Bias}^{\mathrm{post}}_{n,\lambda}(\hat{\theta}_n | X^{(n)}, D^{(n)}) | X^{(n)}] \,. \tag{18}$$

Remark 3.1. We have shown that minimizing the conditional MSE is equivalent to (17), i.e., minimizing the second moment of the ex-post bias, and (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 (9). (17) is more convenient for intuition, while (18) is easier to solve. ■

The following theorem contains our main result on optimal stratification, which shows that (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

$$g(x) = E[Y_i(1) + Y_i(0) | X_i = x] \,. \tag{19}$$

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

$$\lambda^g(X^{(n)}) = \{\{\pi^g(2s-1), \pi^g(2s)\} : 1 \leq s \leq n\} \,. \tag{20}$$

Theorem 3.1. *Suppose the treatment assignment scheme satisfies Assumption 2.1. Then, $\lambda^g(X^{(n)})$ defined in (20) solves (5).*

Remark 3.2. Figure 3 illustrates the optimal stratification in (20). The outline of the proof of Theorem 3.1 is as follows. Lemma S.3.1 shows that each stratification is a convex combination of matched-pair designs. Therefore, one of the solutions to (5) must be a "vertex" of these convex combinations, i.e., a matched-pair design. Using the second part of Lemma 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 (20), according to a variant of the Hardy-Littlewood-Pólya rearrangement inequality for non-bipartite matching. ■

Remark 3.3. Note from (19) that g_i is a scalar regardless of the dimension p of X_i. Moreover, (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

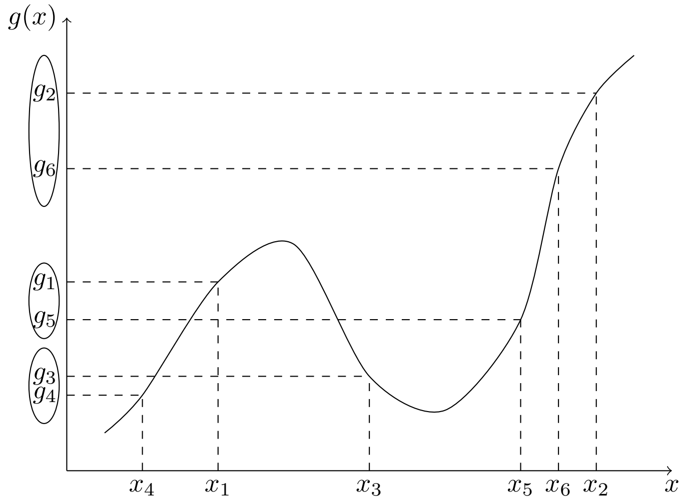

Figure 1: Illustration of the optimal stratification defined in (20). In the example, $p = 1$, i.e., X_i's are scalars. The optimal stratification is $\{\{3,4\},\{1,5\},\{2,6\}\}$.

units adjacently, regardless of the values of $g_i, 1 \le i \le 2n$. We further emphasize that the result in Theorem 3.1 does not rely on the knowledge of the conditional variances of $Y_i(1)$ and $Y_i(0)$ given X_i. ■

Remark 3.4. Using similar arguments as those used to establish Theorem 3.1, it is possible to show that if MSE in (5) is replaced by any expected utility criterion, then one of the solutions is a matched-pair design. It is further possible to show the same conclusion holds for any criterion that is convex in the distribution of treatment status. Therefore, the optimality of matched-pair designs holds quite generally. That said, it is nontrivial to characterize the form of the optimal matched-pair design in those general settings and this is left for future work. ■

Remark 3.5. Theorem S.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

$$g^\tau(X_i) = \frac{E[Y_i(1)|X_i]}{\tau} + \frac{E[Y_i(0)|X_i]}{1-\tau} . \tag{21}$$

Let π^{τ,g^τ} be a permutation of $\{1,\dots,kn\}$ such that $g^\tau_{\pi^{\tau,g^\tau}(1)} \le \dots \le g^\tau_{\pi^{\tau,g^\tau}(kn)}$. We show that (5) is solved by

$$\lambda^{\tau,g}(X^{(n)}) = \{\{\pi^{\tau,g^\tau}((s-1)k+1),\dots \pi^{\tau,g^\tau}(sk)\} : 1 \le s \le n\} , \tag{22}$$

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 3.1, and in particular (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 $1 \le i \le 2n$; 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, we plot the densities of the distributions of $\text{Bias}^{\text{post}}_{n,\lambda}(\hat\theta_n|X^{(n)}, D^{(n)})$

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

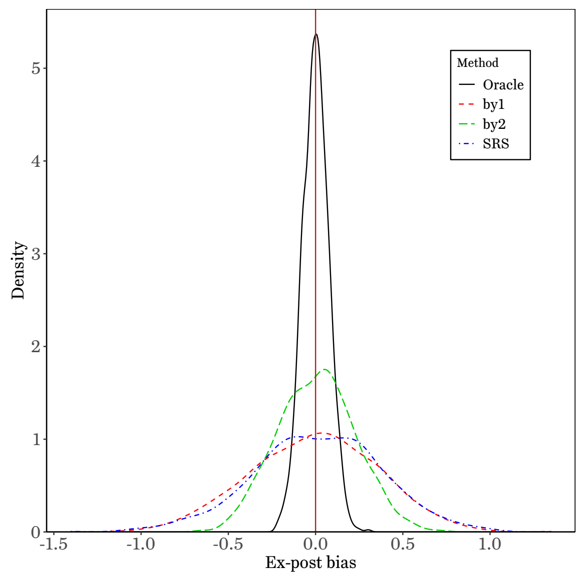

Figure 2: Densities of the distributions of the $\text{Bias}_{n,\lambda}^{\text{post}}(\hat{\theta}_n | X^{(n)}, D^{(n)})$ over 1000 draws of $X^{(n)}$ and $D^{(n)}$ under all treatment assignment schemes.

4 Empirical counterparts

The optimal procedure in (20) depends on the function g defined in (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 (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 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 $((\tilde{Y}_j(1), \tilde{Y}_j(0), \tilde{X}_j) : 1 \leq j \leq m)$ 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 (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:

$$\lambda^{\hat{g}_m}(X^{(n)}) = \{\{\pi^{\hat{g}_m}(2s-1), \pi^{\hat{g}_m}(2s)\} : 1 \leq s \leq n\} . \tag{23}$$

As Theorem 5.1 shows, the plug-in procedure enjoys the property that as the sample size of the pilot increases, the limiting variance of $\hat{\theta}_n$ in (3) is that same as that under the optimal procedure defined by (20). The key condition for the property is that $\hat{g}_m$ is consistent for g in a certain cense. See Assumption 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. But even then, researchers may well be content with the plug-in procedure because it results in smaller limiting variance of $\hat{\theta}_n$ than many alternatives. That said, we may be concerned that the plug-in estimator $\hat{g}_m$ is a poor approximation for g in (19), and as a result, that under the plug-in stratification defined in (23), the conditional MSE and the limiting variance of $\hat{\theta}_n$ is large. Therefore, we consider a penalized procedure under which, according to simulation studies in Section 6, the conditional MSE of $\hat{\theta}_n$ is often smaller than that under the stratification defined in (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.

We will describe the procedure first and then explain the intuititon why it is of this particular form. For $d \in \{0, 1\}$, define the least-square estimators based on the treated or control units as

$$\hat{\beta}_m(d) = \left(\sum_{1 \leq j \leq m : \tilde{D}_j = d} \tilde{X}_j \tilde{X}_j' \right)^{-1} \sum_{1 \leq j \leq m : \tilde{D}_j = d} \tilde{X}_j \tilde{Y}_j , \tag{24}$$

and the variance estimators assuming homoskedasticity as

$$\hat{\Sigma}_m(d) = \hat{\nu}_m^2(d) \left(\sum_{1 \leq j \leq m : \tilde{D}_j = d} \tilde{X}_j \tilde{X}_j' \right)^{-1} , \tag{25}$$

where

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

Further define

$$\hat{\beta}_m = \hat{\beta}_m(1) + \hat{\beta}_m(0) \tag{26}$$

$$\hat{\Sigma}_m = \hat{\Sigma}_m(1) + \hat{\Sigma}_m(0) . \tag{27}$$

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

$$R_m' R_m = \hat{\beta}_m \hat{\beta}_m' + \hat{\Sigma}_m , \tag{28}$$

and the following transformation of the covariates:

$$Z_i = R_m X_i . \tag{29}$$

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:

$$\min_{\pi \in \Pi_n} \frac{1}{n} \sum_{1 \leq s \leq n} \| Z_{\pi(2s-1)} - Z_{\pi(2s)} \| . \tag{30}$$

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

$$\lambda^{\text{pen}}(X^{(n)}) = \{\{\pi^{\text{pen}}(2s-1), \pi^{\text{pen}}(2s)\} : 1 \leq s \leq n\} . \tag{31}$$

(31) can be viewed as penalizing the plug-in procedure in (23) by the variance of the plug-in estimator.

We now briefly explain the intuition behind (30). For simplicity, suppose $E[Y_i(d)|X_i] = X_i'\beta(d)$ for $d \in \{0, 1\}$. In addition, define $\beta = \beta(1) + \beta(0)$. (30) penalizes the the plug-in stratification by the standard error of the plug-in estimate. Indeed, the objective in (30) equals

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

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

$$\hat{d}(x_1, x_2) = (x_1' \hat{\beta}_m - x_2' \hat{\beta}_m)^2 + (x_1 - x_2)' \hat{\Sigma}_m (x_1 - x_2) . \tag{32}$$

If $\hat{\Sigma}_m = 0$, then (30) is solved by $\pi^{\hat{g}_m}$ in the plug-in stratification in (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 (32) dominates, and $\hat{g}_m$ contributes little to the solution to (30).

Remark 4.1. We now provide a further justification for (31) by discussing its optimality in a Bayesian framework. To begin with, note that the problem in (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 4 of Bai et al. (2019). 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 (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 $u : (\tilde{w}^{(m)}, x^{(n)}) \mapsto \lambda \in \Lambda_n$:

$$\min_u \iiint \mathrm{MSE}(u(\tilde{w}^{(m)}, x^{(n)})|g, x^{(n)}) Q_X^n(dx^{(n)}) Q_{\tilde{W}}^m(d\tilde{w}^{(m)}) F(dg) \,. \tag{33}$$

In Appendix S.4, we first show that the problem in (33) under any prior F is solved by a matched-pair design. 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 (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 (26) shrunk towards the prior mean. We show, however, that the solution to (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 (33) converges to the procedure defined by (30) with the squared norm $\|Z_{\pi(2s-1)} - Z_{\pi(2s)}\|^2$. ■

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 (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 (23) and (31) depend on data from the pilot experiment. Extending results from Bai et al. (2019), 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 5.1. $E[Y_i^2(d)] < \infty$ for $d \in \{0, 1\}$.

5.1 Asymptotic results for plug-in with large pilot

In this subsection, we study the properties of $\hat{\theta}_n$ defined in (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 (19).

Assumption 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 5.2(a)–(c) are conditions imposed on the target function g instead of the plug-in estimator $\hat{g}_m$. Assumption 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 5.2(c) is another mild moment restriction to ensure the pairs are "close" in the limit. New sufficient conditions for Assumption 5.2(b) are provided in Appendix S.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 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 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 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 5.3. (b) also implies Assumption 5.3 by Markov's inequality.

The next theorem reveals that under L^2-consistency of the estimator $\hat{g}_m$, the limiting variance of $\hat{\theta}_n$ under the plug-in procedure is the same with that under the infeasible optimal procedure defined by (20).

Theorem 5.1. *Suppose the treatment assignment scheme satisfies Assumption 2.1, Q satisfies Assumption 5.1, g satisfies Assumption 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

$$\varsigma_g^2 = \mathrm{Var}[Y_i(1)] + \mathrm{Var}[Y_i(0)] - \frac{1}{2}E[(g(X_i) - E[Y_i(1) + Y_i(0)])^2] . \tag{34}$$

In addition, suppose $\hat{g}_m$ satisfies Assumption 5.3. Then, under $\lambda^{\hat{g}_m}(X^{(n)})$ defined in (23), as $m, n \to \infty$,

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

Remark 5.1. Bai et al. (2019) studies the scenario where units are matched to minimize the sum of the Euclidean distance in terms of their covariates, and show that the limiting variance of $\hat{\theta}_n$ is equal to ς_g^2 in (34). The results there, though, are derived assuming that the number of covariates p is fixed. Instead, we could allow for p to increase with the sample size n, as long as $\hat{g}_m$ is L^2-consistent for g. ■

Remark 5.2. Since the limiting variance in (34) is the same as Hahn (1998)'s semiparametric efficiency bound, there is no additional benefit from doing regression adjustments. If more covariates become available after the treatment assignment, however, it is straightforward to combine stratification with covariate adjustments. By using a conventional argument in partitioned regression, one could show that this will lead to a weakly smaller limiting variance of $\hat{\theta}_n$. ■

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

$$H_0 : \theta(Q) = \theta_0 \text{ versus } H_1 : \theta(Q) \neq \theta_0 \tag{35}$$

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

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

Define

$$\hat{\rho}_n = \frac{2}{n} \sum_{1 \le j \le \lfloor \frac{n}{2} \rfloor} (Y_{\pi^{\hat{g}_m}(4j-3)} + Y_{\pi^{\hat{g}_m}(4j-2)})(Y_{\pi^{\hat{g}_m}(4j-1)} + Y_{\pi^{\hat{g}_m}(4j)}) \tag{36}$$

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

$$(\hat{\varsigma}_n^{\hat{g}_m})^2 = \hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0) - \frac{1}{2}\hat{\rho}_n + \frac{1}{2}(\hat{\mu}_n(1) + \hat{\mu}_n(0))^2 . \tag{37}$$

The test is

$$\phi_n^{\hat{g}_m}(W^{(n)}) = I\{|T_n^{\hat{g}_m}(W^{(n)})| > \Phi^{-1}(1-\frac{\alpha}{2})\} , \tag{38}$$

where

$$T_n^{\hat{g}_m}(W^{(n)}) = \frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\hat{\varsigma}_n^{\hat{g}_m}} , \tag{39}$$

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 (37) is possibly negative, its limit in probability must be positive under assumptions imposed below. By Remark 5.5 below, we could always adjust it to be positive. Therefore, we assume all quantities like (37) are positive for the rest of the paper.

We start by studying the limiting behavior of the test defined in (38) with a large pilot. The following theorem shows that the test defined in (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 5.2. *Suppose the treatment assignment scheme satisfies Assumption 2.1, Q satisfies Assumption 5.1, g satisfies Assumption 5.2, and $\hat{g}_m$ satisfies Assumption 5.3. Then, under $\lambda^{\hat{g}_m}(X^{(n)})$ defined in (23), as $m, n \to \infty$,*

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

Thus, for the problem of testing (35) at level $\alpha \in (0,1)$, $\phi_n^{\hat{g}_m}(W^{(n)})$ defined in (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 5.3. The studentization by (37) is crucial for the asymptotic exactness of (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 Bai et al. (2019) for more details. ■

Remark 5.4. In order for $\hat{g}_m$ to satisfy Assumption 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 Bai et al. (2019). ■

Remark 5.5. In finite sample one might be worried that the right hand side of (37) is negative. Furthermore, we always have access to an asymptotically conservative estimator for the limiting variance, for example, $\hat{\varsigma}_n^2 = \hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0)$, whose probability limit is weakly greater than ς_g^2. So even though the right hand side of (37) is positive, it might be larger than $\hat{\varsigma}_n^2$ in finite sample. To get over both problems, we could simply redefine the variance estimator to be $\hat{\varsigma}_n^2$ if the right hand side of (37) is less than or equal to 0, and the smaller one of the right hand side of (37) and $\hat{\varsigma}_n^2$ otherwise. ■

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 (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 5.1:

Assumption 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 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 5.4 contains the same set of conditions as those in Assumption 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 S.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 (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 5.3. *Suppose the treatment assignment scheme satisfies Assumption 2.1, Q satisfies Assumption 5.1, and $\hat{g}_m$ satisfies Assumption 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 (23), for the problem of testing (35) at level $\alpha \in (0, 1)$, $\phi_n^{\hat{g}_m}(W^{(n)})$ defined in (38) satisfies*

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

Remark 5.6. Note that we use the same test $\phi_n^{\hat{g}_m}$ with large (Theorem 5.2) and small (Theorem 5.3) pilots, and it is asymptotically exact either way. When m increases at a rate such that Assumption 5.3 is satisfied, the limiting variance of $\hat{\theta}_n$ as $m, n \to \infty$ is ς_g^2, which equals the limiting variance under the infeasible optimal procedure defined by (20). Yet when m is fixed, the limiting 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 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 5.2 and Assumption 5.3 could potentially allow for more complicated estimation methods but require $\hat{g}_m$ to be L^2-consistent for g. ■

Remark 5.7. In fact, the asymptotic exactness of $\phi_n^{\hat{g}_m}(W^{(n)})$ holds conditional on data from the pilot experiment, i.e.,

$$\lim_{n \to \infty} E[\phi_n^{\hat{g}_m}(W^{(n)}) | \tilde{W}^{(m)}] = \alpha \tag{40}$$

with probability one for $\tilde{W}^{(m)}$. See the proof of Theorem 5.3 in the appendix for more details. Furthermore, it follows from the proof that the test is also asymptotically exact under

$$\lambda^h(X^{(n)}) = \{\{\pi^h(2s-1), \pi^h(2s)\} : 1 \le s \le n\}, \tag{41}$$

where $h_{\pi^h(1)} \le \ldots \le h_{\pi^h(2n)}$ and h is a fixed function satisfying $h \in \mathbf{H}$ for $\mathbf{H}$ defined in (5.4). ■

Remark 5.8. As an intermediate step in the proof of Theorem 5.3, we derive the limiting variance of $\hat{\theta}_n$ under $\lambda^h(X^{(n)})$ defined in (41), where h is a fixed function satisfying $h \in \mathbf{H}$. The limiting variance equals

$$\varsigma_h^2 = \mathrm{Var}[Y_i(1)] + \mathrm{Var}[Y_i(0)] - \frac{1}{2} E[(E[Y_i(1) + Y_i(0) | h(X_i)] - E[Y_i(1) + Y_i(0)])^2]. \tag{42}$$

Comparing (42) with (34), we could show the minimum of ς_h^2 over $h \in \mathbf{H}$ occurs when $h = g$, and the minimum is unique unless there exists and $h \in \mathbf{H}$ for which $E[Y_i(1) + Y_i(0) | h(X_i)] = E[Y_i(1) + Y_i(0) | X_i]$ with probability one. This result enables us to compare the limiting variance of $\hat{\theta}_n$ across a large class of stratifications, and in particular, all stratifications with a fixed number of large strata. Indeed, all such stratifications could be defined by a discrete-valued function $h : \mathbf{R}^p \to \{1, \ldots, R\}$ for a fixed integer R, and therefore $\varsigma_h^2 \ge \varsigma_g^2$ unless $E[Y_i(1) + Y_i(0) | h(X_i) = r] = E[Y_i(1) + Y_i(0) | X_i]$ with probability one, i.e, when $E[Y_i(1) + Y_i(0) | X_i]$ is the same within each stratum. Another corollary is that if $h \in \mathbf{H}$ and h_c is a constant function, then the stratification $\lambda^{h_c}(X^{(n)}) = \{\{1, \ldots, 2n\}\}$ with all units in one stratum satisfies $\varsigma_{h_c}^2 \ge \varsigma_h^2$, unless again the degeneracy condition holds, this time requiring $E[Y_i(1) + Y_i(0) | h(X_i)]$ to be a constant. Any $\hat{g}_m$ with $Q\{\hat{g}_m \in \mathbf{H}\} = 1$ is a constant function in $\mathbf{H}$ conditional on the pilot data $\tilde{W}^{(m)}$, so in this sense, almost all stratifications are better than not stratifying at all, because it results in a weakly smaller and typically strictly smaller limiting variance of $\hat{\theta}_n$. See Theorem S.2.2 for more details. By direct calculation we could also show that for any $h \in \mathbf{H}$, ς_h^2 is weakly less than and typically strictly less than the limiting variance of $\hat{\theta}_n$ under simple random sampling, i.e., when treatment status is determined by i.i.d. coin flips. ■

Remark 5.9. 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 S.2, $\hat{\theta}_n$ is no longer consistent for θ in (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., $\hat{\theta}_n^{\mathrm{sat}}$ defined in (S.15). Appendix S.2 discusses straightforward extensions of the optimality result in Theorem 3.1 and empirical counterparts including that in (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 S.2.1 discusses a way to combine his procedure and procedures in this paper, under which the limiting variance of $\hat{\theta}_n^{\mathrm{sat}}$ is no greater than and typically strictly less than that under his procedure alone. ■

5.3 Inference under penalized procedure

We now consider inference under the penalized procedure defined by (31) with a small pilot. This subsection follows closely the exposition in Section 4 of Bai et al. (2019). Since in general Z defined in (29) is not a scalar, the correction term in (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

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

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

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

In other words, $\tilde{\pi}^{\mathrm{pen}}$ matches the pairs defined by π^{pen} based on the midpoints of pairs. Since $\tilde{\pi}^{\mathrm{pen}}$ rearranges π^{pen} in (31) while preserving the units in each stratum, it follows that for $\lambda^{\mathrm{pen}}(X^{(n)})$ defined in (31), we have

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

We then define the test similarly to (38), with $\pi^{\hat{g}_m}$ replaced by $\tilde{\pi}^{\mathrm{pen}}$. In particular, define

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

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

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

The test is

$$\phi_n^{\text{pen}}(W^{(n)}) = I\{|T_n^{\text{pen}}(W^{(n)})| > \Phi^{-1}(1 - \frac{\alpha}{2})\}\ , \tag{43}$$

where

$$T_n^{\text{pen}}(W^{(n)}) = \frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\hat{\varsigma}_n^{\text{pen}}}\ , \tag{44}$$

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 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 5.5(a)–(b) are the counterparts to Assumption 2.1(a) and (c) of Bai et al. (2019). Assumption 5.5(c) is also imposed in Section 4 of Bai et al. (2019). The following theorem establishes the asymptotic exactness of the test defined in (43), 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 5.4. *Suppose the treatment assignment scheme satisfies Assumption 2.1 and Q satisfies Assumptions 5.1 and 5.5. Suppose Q additionally satisfies the null hypothesis, i.e., $\theta(Q) = \theta_0$. Then, under $\lambda^{\text{pen}}(X^{(n)})$ defined in (31), for the problem of testing (35) at level $\alpha \in (0, 1)$, $\phi_n^{\text{pen}}(W^{(n)})$ defined in (38) satisfies*

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

Remark 5.10. 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). ■

5.4 Inference with pooled data

So far we have disregarded data from the pilot experiment in the test defined in (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

$$\phi_n^{\text{combined}}(W^{(n)}, \tilde{W}^{(m)}) = I\{|T_n^{\text{combined}}(W^{(n)}, \tilde{W}^{(m)})| > \Phi^{-1}(1 - \frac{\alpha}{2})\}\,, \tag{45}$$

where

$$T_n^{\text{combined}}(W^{(n)}, \tilde{W}^{(m)}) = \frac{\sqrt{m+2n}(\hat{\theta}_n^{\text{combined}} - \theta_0)}{\sqrt{\frac{m}{m+2n}\tilde{\varsigma}^2_{\text{pilot},m} + \frac{2n}{m+2n}2(\hat{\varsigma}_n^{\hat{g}_m})^2}}\,, \tag{46}$$

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 (45) 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 5.5. *Suppose the treatment assignment scheme satisfies Assumption 2.1, Q satisfies Assumptions 5.1, g satisfies Assumption 5.2, and $\hat{g}_m$ satisfies Assumption 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 (23), as $m, n \to \infty$,

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

Thus, for the problem of testing (35) at level $\alpha \in (0,1)$, $\phi_n^{\text{combined}}(W^{(n)}, \tilde{W}^{(m)})$ in (45) satisfies

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

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

Remark 5.11. Although Theorem 5.5 is stated under $\lambda^{\hat{g}_m}(X^{(n)})$ in (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., (31). ■

6 Simulation

In this section, we examine the properties of the procedures discussed in Section 4 in a small simulation study. For $d \in \{0,1\}$ and $1 \le i \le 2n$, 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 (20).

Plug-in matched-pair design with the plug-in stratification in (23) with $\hat{g}_m(x) = x'\hat{\beta}_m$ for $\hat{\beta}_m$ in (26).

Pen matched-pair design with the penalized stratification in (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 (23) with $\hat{g}_m(x) = x_1$.

MP2 matched-pair design using $X_{i,2}$ only, i.e., stratification in (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 (4). In these results, we set $\mu(1) = \mu(0) = 0$, so that $\theta(Q) = 0$ as well. By Lemma 3.1 and in particular (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 (19), we define a constant multiple of the objective in (18) as the loss:

$$L(\lambda|g, X^{(n)}) = 4n^2 \operatorname{Var}_\lambda[E[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] . \tag{47}$$

Table 1 displays the summary statistics of the values of the loss defined in (47) 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 $X^{(n)}$.

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 1: Summary statistics for ratios of the values of the loss in (47) under all stratifications against those under the infeasible optimal stratifications (**Oracle**), over 1000 draws of $X^{(n)}$, in Models 1–6.

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 (19) is a monotonic function of the first covariate, so that **MP1** solves (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 (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 (38) with $\hat{g}_m = g$ for g defined in (19).

Plug-in: test in (38) with $\hat{g}_m(x) = x'\hat{\beta}_m$ for $\hat{\beta}_m$ defined in (26).

Pen: test in (43).

MPeuc: test in (43) with Z replaced by X.

by1: test in (38) with $\hat{g}_m(x) = I\{x_1 > \text{med}(X_{i,1} : 1 \leq i \leq 2n)\}$.

by2: test in (38) with $\hat{g}_m(x) = I\{x_2 > \text{med}(X_{i,2} : 1 \leq i \leq 2n)\}$.

MP1: test in (38) with $\hat{g}_m(x) = x_1$.

MP2: test in (38) with $\hat{g}_m(x) = x_2$.

Table 2 displays the rejection probabilities for Models 1–6 under all stratifications using tests described above. Note that loss properties in Table 1 translate into power properties in Table 2. Indeed, while all tests under all stratifications have correct sizes, the test in (43) under the penalized stratification in (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 (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**.

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

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 (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 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 (31), and conduct inference on the ATE in two ways: disregarding data from the pilot experiment as in (43), and combining data from the pilot and main experiments as in (45). 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 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 (31) and the test statistic in (44).

Combined penalized stratification in (31) and the test statistic in (46).

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

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 3.1 is optimal in population, which translates into optimality with large pilots in Theorem 5.1, while the penalized procedure in (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 S.5. By Lemma 3.1 and in particular (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 (47). Moreover, we have

$$L(\lambda|g, X^{(n)}) = 4n^2 \operatorname{Var}_\lambda[E[\hat{\theta}_n|X^{(n)}, D^{(n)}]|X^{(n)}] = \sum_{1\le s\le S} \frac{1}{n_s - 1} \sum_{i,j\in\lambda_s, i<j} (g_i - g_j)^2 . \tag{48}$$

Consider the following minimax problem to find the stratification λ that has the best worst-case performance in terms of the loss in (48), where the worst-case is among a class of functions $\mathcal{G}$.

$$\min_{\lambda\in\Lambda} \max_{h\in\mathcal{G}} L(\lambda|h, X^{(n)}) . \tag{49}$$

Our framework requires $\mathcal{G}$ to have a bounded polyhedron structure, in the sense made precise by Assumption S.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, (49) 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 S.5.1. Unfortunately, such a result no longer holds when $p > 1$. Indeed, Example S.5.7 shows that matched-pair designs may not even be minimax-optimal. We show, however, that under Assumption S.5.1 it is possible to reformulate (49) into a mixed-integer linear program. The reformulation is based on the special structure in (48), which enables us to rewrite (49) 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 (49) 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 S.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.

9 Conclusion and recommendations for empirical practice

This paper 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 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 5.4. Therefore, we recommend researchers to consider using matched-pair designs, or corresponding procedures in Appendix S.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 5.2, we recommend researchers to use flexible nonparametric estimation methods to estimate the target function in (19) when the pilot is large. When the pilot is small, researchers could still use the plug-in procedure with simple estimators such as least squares, but could also consider the penalized procedure.

References

ABADIE, A. and IMBENS, G. W. (2008). Estimation of the Conditional Variance in Paired Experiments. *Annales d'Économie et de Statistique* 175–187.

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.

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.

BAI, Y. (2020). Why randomize? Minimax optimality under permutation invariance. Working paper.

BAI, Y., SHAIKH, A. and ROMANO, J. P. (2019). Inference in experiments with matched pairs. Working paper.

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

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.

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*, **10** 1747–1785.

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. In *Handbook of Econometrics* (J. J. Heckman and E. E. Leamer, eds.), vol. 6. 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., 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** 222–255.

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.

DIZON-ROSS, R. (2019). Parents' beliefs about their children's academic ability: Implications for educational investments. *American Economic Review*, **109** 2728–2765.

DUFLO, E. and BANERJEE, A. (2017). *Handbook of field experiments*. Elsevier Science.

DUFLO, E., DUPAS, P. and KREMER, M. (2015). Education, HIV, and early fertility: Experimental evidence from Kenya. *American Economic Review*, **105** 2757–97.

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** 163–192.

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.

GYÖRFI, L., KOHLER, M., KRZYZAK, A. and WALK, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

HAHN, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, **66** 315–331.

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.

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.

JOHANSSON, P., SCHULTZBERG, M. A. and RUBIN, D. (2019). On optimal re-randomization 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.

KASY, M. and LEHNER, L. (2020). Employing the unemployed of Marienthal: Evaluation of a guaranteed job program. *AEA RCT Registry*. URL https://doi.org/10.1257/rct.6706-1.0.

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.

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.

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.

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.

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.

SPIESS, J. (2020). Optimal Estimation when Researcher and Social Preferences are Misaligned.

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., DU, W., TAYLOR, J. and TIBSHIRANI, R. J. (2016). High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, **113** 12673–12678. Publisher: National Academy of Sciences Section: Physical Sciences.

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.