Optimal Stratification of Survey Experiments*

Max Cytrynbaum[†]

August 11, 2023

Abstract

This paper studies treatment effect estimation in a two-stage model of experimentation. In the first stage, using baseline covariate information, the researcher draws a representative sample of experimental participants from a pool of eligible units. In the second stage, they assign each sampled participant to treatment or control. To implement both representative sampling and balanced assignment, we introduce a new family of finely stratified designs, generalizing matched pairs randomization to propensities $p(x) \neq 1/2$. When used for representative sampling, our method nonparametrically dampens the estimator variance due to treatment effect heterogeneity. When used for balanced assignment, our designs make simple difference-of-means estimation behave like well-specified nonparametric regression adjustment, effectively doing nonparametric regression "by design." Building on these asymptotic results, we formalize and solve a two-stage optimal stratification problem with heterogeneous sampling costs and fixed budget constraint, providing simple heuristics for the optimal design. In settings with data available from a pilot or related previous experiment, we show that finely stratified implementation of a pilot estimate of this optimal design is efficient, achieving minimal asymptotic variance subject to the budget constraint. We provide novel asymptotically exact inference methods over this entire class of designs, allowing practitioners to fully exploit the efficiency gains from both finely stratified sampling and assignment. An empirical application to $N = 9$ papers recently published in top journals in economics demonstrates the value of our proposed methods.

*I wish to thank Alberto Abadie, Anna Mikusheva, and Victor Chernozhukov for their support and guidance during this project. This paper also benefited from conversations with Isaiah Andrews, Mert Demirer, and numerous seminar participants. The current work replaces an earlier version of my job market paper titled "Designing Representative and Balanced Experiments by Local Randomization."

[†]Yale Department of Economics. Correspondence: max.cytrynbaum@yale.edu

1 Introduction

Randomized controlled trials (RCTs) are now common in economics research, with over 7000 active experiments in the AEA RCT registry spanning a range of subfields. A long tradition in experimental design attempts to reduce the variance of treatment effect estimation, allowing researchers to make the most efficient use of their limited resources. One way to do this is by covariate-adaptive treatment assignment, which balances observed covariates between the treatment and control group at design-time. This paper contributes to theory of covariate-adaptive treatment assignment, but also models a new dimension of experimental design: the selection of the experimental participants.

The selection of participants, also known as the sampling frame, is an important step in designing an experiment. For example, Abaluck et al. (2021) run an experiment to estimate the effect of mask distribution on covid infection rates in Bangladesh. From a pool of 1000 eligible villages, they first randomly sample 600 to be included in the experiment, then assign the sampled villages to various interventions that promote mask usage. Similarly, Breza et al. (2021) estimate the effect of Facebook ads discouraging holiday travel on covid infection rates. Since their budget for running ads is finite, they first sample a small set of counties in which to run ads and collect outcome data, then randomly assign these sampled counties to low or high intensity of treatment. We show how to increase the efficiency of treatment effect estimation by sampling experimental units that are representative of the broader target population, and provide new inference methods that take full advantage of these precision gains.

To do so, this paper introduces a new family of finely stratified randomization procedures that can be used for both representative sampling of the experimental units and finely balanced treatment assignment. When used for assignment, our method generalizes the principle of matched pairs randomization to propensities $p(x) \neq 1/2$, allowing discrete or continuous stratification variables in general dimension. The basic building block is a new algorithm that matches the experimental units into homogeneous groups of k by minimizing an objective function directly linked to estimation efficiency. This matching algorithm also enables finely stratified sampling of the experimental participants. For example, suppose 1000 people respond to an advertisement to participate in an experiment, but the experimental budget only allows for 400 participants. Using observed covariate information, we match the units into homogeneous 5-tuples, sampling $q = 2/5$ of the units in each tuple to participate, uniformly at random. By finely representing the distribution of treatment effect heterogeneity in the population into our smaller experiment, this sampling procedure reduces the variance due to treatment effect heterogeneity. More generally, we provide finely stratified designs implementing heterogeneous sampling propensities $q(x)$.

We study a two-stage procedure that (1) samples participants then (2) assigns treatments to the sampled units, using finely stratified designs at both stages. Under finely stratified treatment assignment alone, the difference-of-means estimator $Y \sim 1 + D$ achieves the Hahn (1998) variance bound for the average treatment effect (ATE), effectively doing nonparametric regression adjustment "by design." Our analysis shows that finely stratified sampling provides an additional nonparametric variance reduction, dampening the variance due to treatment effect heterogeneity. In particular, representative sampling makes this variance component scale with the the number of units we sample from, rather than the smaller true experiment size, boosting the effective sample size for this component of the variance. Extending recent results in Bai (2022), we characterize the optimal stratification variables for sampling and assignment, showing that for sampling one wants to stratify on covariates that are most predictive of treatment effect heterogeneity. In an extension, we also study estimation of the sample average treatment effect (SATE) over the eligible population using design-based asymptotics.

Building on these asymptotic results, we formalize and solve an optimal design problem with fixed experimental budget and heterogeneous costs. In development economics, the cost of including a village in an experiment can vary widely based on observable characteristics such as its distance from the urban center, village size and so on. This forces applied researchers to choose a tradeoff between sample size and sample representativeness when they select where to experiment. We provide a new formalization of this tradeoff, deriving the jointly optimal sampling intensity $q^*(x)$ and assignment propensity $p^*(x)$ under finely stratified randomization. Under homoskedasticity, for instance, the optimal sampling propensity $q^*(x) \propto C(x)^{-1/2}$, where $C(x)$ is the cost of including a unit of type $X = x$ in the experiment. Our results give simple heuristics for optimal sampling of the experimental units, analogous to classical results on sample allocation for coarsely stratified survey design (Cochran (1977)). We show that an oracle design that implements discretizations of $q^*(x)$ and $p^*(x)$ using fine stratification minimizes the asymptotic variance of our estimator over all stratified designs, subject to the experimental budget constraint.

We also briefly investigate exact optimality in finite samples. For fixed propensity $p = 1/2$, we prove that the globally optimal covariate-adaptive randomization takes the form of novel "alternating design", assigning a certain optimal allocation vector $(d_i^*)_{i=1}^n$ and its mirror image $(1-d_i^*)_{i=1}^n$ each with probability $1/2$. The optimal allocation $(d_i^*)_{i=1}^n$ solves the well-known Max-Cut graph partitioning problem (Rendl et al. (2008)), with edge weights related to the smoothness of the outome functions.

We apply our optimal design results to the problem of two-wave design using data from a pilot experiment, providing the first fully efficient solution to this problem. The basic idea is to estimate the optimal sampling and assignment propensities $q^*(x)$ and $p^*(x)$ using the pilot, then implement these estimates in the main experiment using fine stratification. Under large pilot asymptotics, this strategy is as efficient as the oracle design, achieving the budget-constrained minimal asymptotic variance. We also provide results under fixed pilot asymptotics, and briefly discuss potential robustifications. In the case without sampling, the problem of design using a pilot has received considerable attention in the recent literature, see for instance Hahn et al. (2011), Tabord-Meehan (2022), and Bai (2022), and we give a detailed comparison with these results.

Finally, we provide novel asymptotically exact inference for the average treatment effect under joint finely stratified sampling and assignment, using a collapsed strata[1] type estimator (Hansen et al. (1953)). The use of non-constant sampling proportions $q(x)$ produces discontinuities in the propensity-weighted outcome functions, introducing new

[1] See also Abadie and Imbens (2008) and Bai et al. (2021) for related results in the context of matched pairs assignment.

technical complications that we overcome with our analysis. Simulations and an empirical application to $N = 9$ papers recently published in top journals in economics demonstrate the value of our proposed methods.

1.1 Related Literature

Our sampling model is related to the classical literature on survey sampling, e.g. as surveyed in Cochran (1977) and Lohr (2021). In contemporaneous work, Yang et al. (2021) propose a two-stage design using rerandomization for both sampling and assignment. Under rerandomization, difference of means estimation is asymptotically (almost) as efficient as ex-post linear adjustment. By contrast, we show that data-adaptive fine stratification is asymptotically equivalent to nonparametric adjustment for the imbalances in both sampling and assignment variables, see Proposition 3.12 for a formal statement.

For an overview of experimental design theory, see Rosenberger and Lachin (2016) or Athey and Imbens (2017). A representative sample of recent work on stratified treatment assignment includes Imai et al. (2009), Bugni et al. (2018), Fogarty (2018), Wang et al. (2021), Bai et al. (2021), de Chaisemartin and Ramirez-Cuellar (2021), Bai (2022), and Tabord-Meehan (2022). For treatment assignment, our work is most related to Bai (2022), who introduces finely stratified designs for constant propensity $p = a/k$ and univariate stratification variables. Aside from stratification, other recent proposals for balanced treatment assignment include Kasy (2016), Kallus (2017), Li et al. (2018), Krieger et al. (2019), and Harshaw et al. (2021). We explicitly compare with some of these methods in Remark 3.4.

Our results on design using a pilot study is related to previous results in Hahn et al. (2011), Bai (2022), Tabord-Meehan (2022), and Kasy and Sautmann (2021). We provide detailed comparisons in Section 5 below. Our inference results are related to the method of collapsed strata in Hansen et al. (1953) and its modern variants studied in Abadie and Imbens (2008) and Bai et al. (2021).

The rest of the paper is organized as follows. Section 2 introduces notation and discusses our matching algorithm. Section 3 states our main asymptotic results, including the equivalence with nonparametric adjustment. Section 4 formalizes and solves the optimal design problem. Section 5 discusses design using a pilot study. Section 6 provides our inference methods. Our empirical results are presented in Section 8.

2 Motivation and Description of Method

Consider running an experiment to estimate the average treatment effect (ATE). There are n eligible units, with observed baseline covariates $(X_i)_{i=1}^n$. We wish to sample proportion $q \in (0, 1]$ of these units to participate in the experiment. Denote $T_i = 1$ if a unit is sampled and $T_i = 0$ otherwise. Sampled units are then assigned to treatment or control $D_i \in \{0, 1\}$. Let $Y_i(d)$ denote the potential outcome of unit i for $d \in \{0, 1\}$. Since outcomes are only observed for participating units[2] we write

$$Y_i = T_i[D_i Y_i(1) + (1 - D_i)Y_i(0)]$$

We focus on estimation and inference for the population ATE $= E[Y(1)-Y(0)]$, modeling the n eligible units as a sample from a superpopulation of interest, $(X_i, Y_i(0), Y_i(1))_{i=1}^n \sim F$ iid. For example, the eligible units could be $n = 1000$ respondents to an online advertisement to participate in an experiment with a budget constraint of 100 participants. In this context, the variable $T_i \in \{0, 1\}$ models which of the $n = 1000$ units we choose to include in the experiment. In some applications, the eligible units may comprise the entire population of interest. For example, the units $i = 1, \ldots, n$ may be the entire population of villages in a country, which we sample to obtain the participating villages. To accommodate such applications, Section 10.1 in the appendix presents design-based versions of our main results, targeting the sample average treatment effect SATE $= n^{-1} \sum_{i=1}^n Y_i(1) - Y_i(0)$ in the eligible population. Here, we define the SATE over the entire eligible population $i \in [n]$ that we are allowed to sample from, not just the smaller set of units that are chosen to participate in the experiment $\{i : T_i = 1\}$.

Our goal is to sample a representative subset of the eligible units and assign them to treatment and control in a way that finely balances the baseline covariates $(X_i)_{i=1}^n$. To do so, we introduce a new family of finely stratified designs that generalize the principle of matched pairs randomization to arbitary propensity scores $p(x) = P(D = 1|X = x)$, with continuous or discrete covariates in general dimension. We also use these new designs for finely stratified sampling, allowing us to implement arbitrary heterogeneous sampling propensities $q(x) = P(T = 1|X = x)$, while finely balancing covariates between the sampled and non-sampled units. The basic building block of our method is a matched k-tuples design, which uses the baseline covariates to match units into homogeneous groups of k, randomly assigning a out of k units in each group to $T = 1$ during sampling or $D = 1$ during assigment. We formally define the method in the context of finely stratified sampling in the next definition.

Definition 2.1 (Local Randomization)**.** Let $q = a/k$ with $\gcd(a, k) = 1$. Partition the eligible units into groups with $|g| = k$, so that $\{1, \ldots, n\} = \bigcup_g g$ disjointly. In general, there may be one remainder group with $|g| < k$. Let $\psi(X) \in \mathbb{R}^d$ be a vector of stratification variables, and suppose that the groups are homogeneous in the sense that

$$n^{-1} \sum_g \sum_{i,j \in g} |\psi(X_i) - \psi(X_j)|_2^2 = o_p(1) \tag{2.1}$$

Require that the groups only depend on the stratification variable values $\psi_{1:n} = (\psi(X_i))_{i=1}^n$, and data-independent randomness π_n, so that $g = g(\psi_{1:n}, \pi_n)$. Independently over all groups with $|g| = k$, draw sampling variables $(T_i)_{i \in g}$ by setting $T_i = 1$ for exactly a out of k units, completely at random. For units in the remainder group with $|g| < k$, draw T_i iid with $P(T_i = 1) = a/k$. We say that such a design implements q locally with respect to $\psi(x)$, denoting $T_{1:n} \sim \text{Loc}(\psi, q)$.

Equation 2.1 generalizes a similar condition in Bai et al. (2021) for the case of matched pairs $k = 2$. We discuss matching algorithms and their associated homogeneity rates in

[2] If control outcomes are costlessly observed for all units, the sampling problem becomes trivial. We still contribute novel assignment designs in this case.

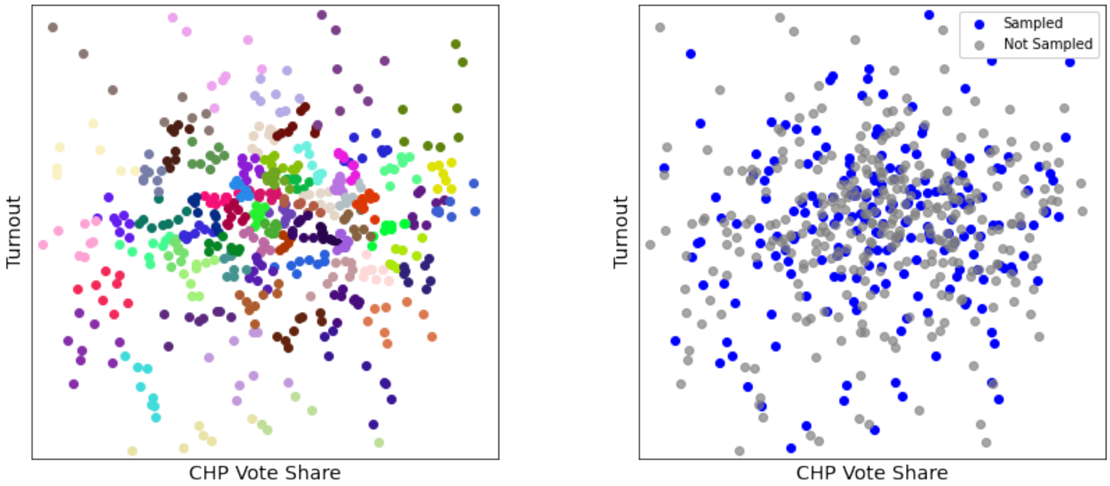


Figure 1: Sampling groups and variables for $T_{1:n} \sim \mathrm{Loc}(\psi, q)$ with $q = 3/8$.

Section 2.1 below. Consider sampling and assignment propensities $q = a/k$ and $p = a'/k'$. In the rest of the paper, we study treatment effect estimation under a two-stage procedure:

(1) Sample eligible units $T_{1:n} \sim \mathrm{Loc}(\psi, q)$.

(2) Assign treatments $D_{1:n} \sim \mathrm{Loc}(\psi, p)$ to the sampled units $\{i : T_i = 1\}$.

This two-stage procedure is illustrated in Figures 1 and 2, using data from an election experiment in Turkey reported in Baysan (2022). Each color represents a different group of units formed during sampling and assignment. For example, in Figure 1 we form groups of size $|g| = 8$, randomly sampling 3 out of 8 units from each group to "represent" that group in the experiment. The sampled units are shown in blue in the figure on the right. In figure 2, we match the sampled units into groups of $k' = 4$, assigning 3 out of 4 to $D = 1$ in each group.

Section 3.2 presents the most general version of our method, allowing different stratification variables ψ_1 and ψ_2 to be used for sampling and assignment, as well as varying sampling and assignment propensities $q(x)$, $p(x)$. Optimal choice of stratification variables for sampling and assignment is discussed in Section 3. Our framework enables unified asymptotics and inference for a wide variety of different designs, as shown in the examples below.

Example 2.2 (Matched Tuples)**.** Suppose $n = 1000$ individuals from a target population sign up to participate in an experiment, providing basic demographic information $(X_i)_{i=1}^n$. There are only resources for 300 units to be enrolled, so $q = 3/10$. Among these 300 units, $p = 1/4$ will be assigned to the more costly treatment and 3/4 to control. We sample using the design $T_{1:n} \sim \mathrm{Loc}(\psi, 3/10)$, which matches the 1000 eligible units into homogeneous groups of 10 and randomly sets $T_i = 1$ for 3 out of 10 units in each group. We assign treatments $D_{1:n} \sim \mathrm{Loc}(\psi, 1/4)$ to the $\sum_i T_i = 300$ sampled units, matching them into homogeneous tuples of four and assigning 1 out of 4 in each tuple to $D_i = 1$.

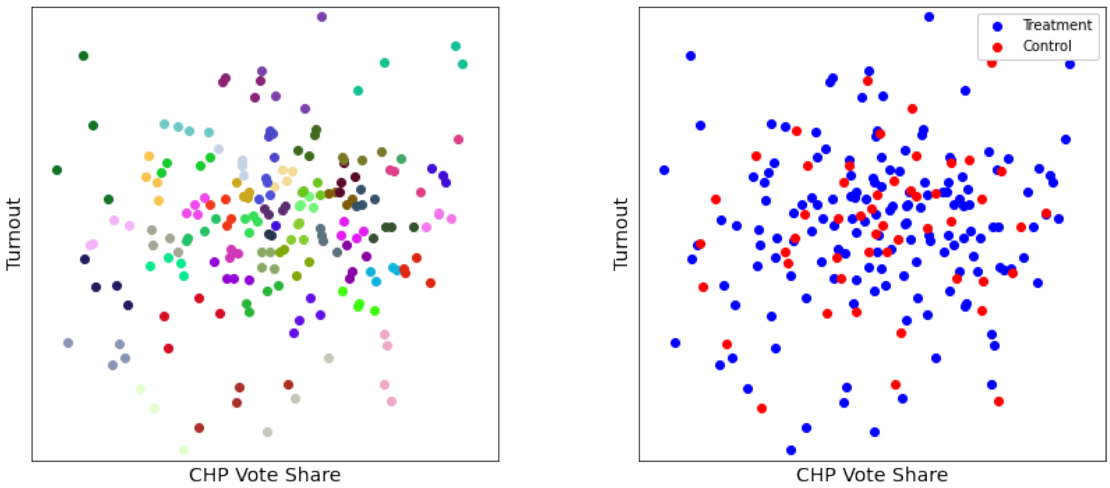


Figure 2: Assignment groups and variables for $D_{1:n} \sim \text{Loc}(\psi, p)$ with $p = 3/4$.

Example 2.3 (Complete Randomization)**.** We say that variables $T_{1:n}$ are completely randomized with probability q if $T_{1:n}$ is drawn uniformly from all vectors $t_{1:n}$ with $t_i = 1$ for exactly proportion q of the units. Formally, we have $P(T_{1:n} = t_{1:n}) = 1/\binom{n}{qn}$ for all such vectors. For sampling, we denote $T_{1:n} \sim \text{CR}(q)$ and $D_{1:n} \sim \text{CR}(p)$ for assignment. Complete randomization may be obtained in our framework by setting $\psi = 1$ and forming groups $|g| = k$ at random, which automatically satisfies Equation 2.1. For example, assigning 2 out of 3 units in each group to treatment gives a "random matched triples" representation of complete randomization with $p = 2/3$.

Example 2.4 (Coarse Stratification)**.** The procedure in Definition 2.1 produces n/k groups of k units that are tightly matched in $\psi(X)$ space, suggesting fine stratification. However, coarsely stratified designs with a fixed strata $S(X) \in \{1, \ldots, m\}$ can also be obtained in this framework by setting $\psi(X) = S(X)$ and matching units with the same $S(X)$ value into groups at random. Coarse stratification was previously studied using different methods in Bugni et al. (2018). We extend their results in Example 3.5 below, allowing coarse stratification at both the sampling and assignment stages.

Remark 2.5 (Sampling Centroids)**.** It's also possible to reverse the order of our two-stage procedure. For example, if $q = a/k$ and $p = a'/k'$ we can first match the eligible units $i = 1, \ldots, n$ into groups of size k', forming the group centroids $\bar{\psi}_g = (k')^{-1} \sum_{i \in g} \psi_i$. Next, we match these group centroids themselves into homogeneous groups of size k. For each group of k centroids, we randomly sample a of the centroids, and their corresponding groups of k' units, into the experiment. Finally, we assign a' out of k' units in each sampled group to treatment. Intuitively, this procedure allocates more of the finite "match quality" in the data set towards balanced assignment, making the assignment groups as tight as possible. We conjecture that this procedure is asymptotically equivalent to the one studied in this paper, but leave formal study to future work.

2.1 Matching Algorithms

One possibility is to treat the left hand side of Equation 2.1 as an objective function and minimize it over all partitions of the units into groups of k. Denoting $d(\psi) = \dim(\psi)$, Theorem 11.1 in the appendix shows that if $E[|\psi(X)|_2^\alpha] < \infty$ for some $\alpha > d(\psi) + 1$ then the optimal groups satisfy

$$\min_{(g)} n^{-1} \sum_g \sum_{i,j \in g} |\psi_i - \psi_j|_2^2 = O_p(n^{2/\alpha - 2/(d(\psi)+1)}) = o_p(1) \tag{2.2}$$

If $\psi(X)$ is bounded this becomes $O_p(n^{-2/(d(\psi)+1)})$, sharpening the rate achieved under a boundedness assumption in previous work on matched pairs (Bai et al. (2021)). For $k = 2$, the optimal groups are computable in $O(n^3)$ time using an algorithm due to Derigs (1988).[3] Efficient algorithms for computing the optimal groups for general k are not available, and we expect the problem to be NP-hard.[4]

Iterative Matching - Instead of calculating the optimal partition, for $k > 2$ we iteratively apply Derigs' algorithm to match units into larger groups. For fixed k, this procedure can be shown to satisfy the same rate in Equation 2.2 above. There are many ways to implement iterative pairing for each k. For example, $k = 5$ can be obtained by pairwise matching of 4-tuples to 1-tuples, or 3-tuples to 2-tuples and so on. This is shown in the figure below, where each level of the tree represents a call to the optimal pairing algorithm.

```
              5                                   5
          /       \                           /       \
        4           1                       2           3
       / \         / \                     / \         / \
     2     2     1     0                 1     1     2     1
    / \   / \   / \                     / \   / \   / \   / \
   1   1 1   1 1   0                   1   0 1   0 1   1 1   0
```

Before the jth algorithm call, we add a certain number of "empty centroids", represented by the 0's in the figure. We also prohibit certain types of matches in order to guarantee the desired sequence of group sizes. For example, at the second level of the tree on the left, we set the distance to $+\infty$ between groups of size $|g| = 2$ and $|g'| = 1$, size $|g| = 1$ and $|g'| = 1$, and size $|g| = 0$ and $|g'| = 0$. There are many choices of such cardinality trees for each k, not all of which can be feasibly implemented using these type of constraints. We provide a canonical way of generating such sequences, as well as the required constraints at each algorithm call, that is guaranteed to implement the desired group cardinality k. Technical details are provided in Section 10.5 in the appendix.

Large Experiments. This algorithm is highly tractable for small and medium experiment sizes. For example, matching $n = 500$ units into 5-tuples takes 23 seconds on a laptop computer, while $n = 2000$ takes about 24 minutes. However, larger experiments quickly become intractable. For example, Domurat et al. (2021) has $n = 87394$, which would take about 3.8 years to match using the algorithm described above. To enable fine stratification in larger experiments, one possibility is to exploit the global shape of the baseline covariate data to rule out matches between distant units. To do so, let v_1 be the

[3]We use the min-weight-matching implementation from the NetworkX 3.1 module in Python.
[4]See Karmakar (2022) for hardness results in a related problem.

first principal component of the stratification variables $(\psi_i)_{i=1}^n$ and consider the following procedure:

(1) Partition $(\psi_i)_{i=1}^n$ into K folds by $(1/K)$th quantile of the sorted projections $v_1'\psi_i$.

(2) Separately in each fold, run the iterative Derigs (1988) procedure above.

Intuitively, we use the first principal component to sort units by their projection along a natural "direction" through the dataset. This exploits the idea that good matches are unlikely between non-adjacent folds: Figure 4 in the appendix gives a visual representation. By parallelizing over $K = 80$ folds, a dataset of size $n = 87394$ can be matched about 5 minutes. The original version of our procedure and this "PCA folds" version are asymptotically equivalent for fixed K. We focus on the original version in the theory that follows.

3 Asymptotics and Optimal Designs

This section contains our main asymptotic results, showing nonparametric efficiency gains from both finely stratified sampling and assignment. First, we state our main assumption.

Assumption 3.1. *The moments $E[Y(d)^2] < \infty$ for $d = 0, 1$ and $E[|\psi(X)|_2^\alpha] < \infty$ hold for some $\alpha > \dim(\psi) + 1$.*

Previous work on fine stratification has required Lipschitz continuity of the outcome function $E[Y(d)|\psi(X) = \psi]$ and variance $\mathrm{Var}(Y(d)|\psi(X) = \psi)$, as well as boundedness of the stratification variables $\psi(X)$.[5] We provide a novel technical analysis that allows all of these assumptions to be removed.

Estimation. Let $\widehat{\theta}$ be the regression coefficient on D_i in $Y_i \sim 1 + D_i$, estimated in the sampled units $\{i : T_i = 1\}$. This is just the usual difference-of-means estimator. Before continuing to our asymptotic results, we state a variance decomposition for $\widehat{\theta}$ that will be used extensively in what follows. Let $c(\psi) = E[Y(1) - Y(0)|\psi(X) = \psi]$ denote the conditional average treatment effect (CATE) and $\sigma_d^2(\psi) = \mathrm{Var}(Y(d)|\psi(X) = \psi)$ the heteroskedasticity function. Define the *balance function*

$$b(\psi; p) = E[Y(1)|\psi(X)] \left(\frac{1-p}{p}\right)^{1/2} + E[Y(0)|\psi(X)] \left(\frac{p}{1-p}\right)^{1/2}. \tag{3.1}$$

Suppose that sampling and assignment are both completely randomized, $T_{1:n} \sim \mathrm{CR}(q)$ and $D_{1:n} \sim \mathrm{CR}(p)$, as in Example 2.3. Let $n_T = \sum_i T_i$ denote the experiment size. Our work shows that $\sqrt{n_T}(\widehat{\theta} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V)$ with variance

$$V = \mathrm{Var}(c(\psi)) + \mathrm{Var}(b(\psi; p)) + E\left[\frac{\sigma_1^2(\psi)}{p} + \frac{\sigma_0^2(\psi)}{1-p}\right]. \tag{3.2}$$

We think of the first term as the variance due to treatment effect heterogeneity, in particular, the heterogeneity predictable by the stratification variables. The second term is the variance due to random assignment, which arises from the chance covariate imbalances between treatment and control units created by complete randomization $D_{1:n} \sim \mathrm{CR}(p)$.

[5] For example, see Bai et al. (2021), Bai et al. (2023).

The results in the next section show how stratified sampling and assignment nonparametrically dampen the each component of this variance expansion.

3.1 Constant Sampling and Assignment Propensities

In this section we state a central limit theorem for ATE estimation for the simplest case where the sampling and assignment propensities $q = a/k$ and $p = a'/k'$ are constant. Section 3.2 below provides the most general version of this result. Remarks 3.7 and 3.9 draw connections between our findings and classical results on semiparametric efficiency in the analysis of observational data.

Theorem 3.2. *Require Assumption 3.1. If sampling and assignment designs are locally randomized $T_{1:n} \sim \mathrm{Loc}(\psi, q)$ and $D_{1:n} \sim \mathrm{Loc}(\psi, p)$, then $\sqrt{n_T}(\widehat{\theta} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V)$*

$$V = q\,\mathrm{Var}(c(\psi)) + E\left[\frac{\sigma_1^2(\psi)}{p} + \frac{\sigma_0^2(\psi)}{1-p}\right].$$

Comparing to Equation 3.2 above, the variance component $\mathrm{Var}(b(\psi; p))$ due to covariate imbalance between the treatment arms is now asymptotically negligible. The variance due to treatment effect heterogeneity $\mathrm{Var}(c(\psi))$ is now dampened by the sampling proportion $q \in (0, 1]$. Observe that the normalization $\sqrt{n_T}(\widehat{\theta} - \mathrm{ATE})$ effectively holds the experiment size n_T constant as we vary q. Holding n_T constant, the number of eligible units $n \approx n_T/q$ grows as $q \to 0$. For small q, there are many eligible units to sample from, allowing us to build a highly representative sample of experimental participants, and reducing the variance due to treatment effect heterogeneity. Another way to understand this is that under finely stratified sampling, the first variance component scales with the larger size n of eligible units, rather than the true experiment size n_T. This effectively "boosts" the experiment size for this component of the variance.

Example 3.3 (Matched Tuples)**.** In Example 2.2 above, we sampled $n_T = 300$ of $n = 1000$ eligible units using the stratified design $T_{1:n} \sim \mathrm{Loc}(\psi, 3/10)$ with $q = 3/10$. Next, we assigned $1/4$ of the sampled units to treatment by $D_{1:n} \sim \mathrm{Loc}(\psi, 1/4)$. Theorem 3.2 shows that $\sqrt{n_T}(\widehat{\theta} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V)$ with asymptotic variance

$$V = (3/10)\,\mathrm{Var}(c(\psi)) + E\left[\frac{\sigma_1^2(\psi)}{1/4} + \frac{\sigma_0^2(\psi)}{3/4}\right].$$

Nonparametric Regression by Design. If $q = 1$ or sampling is completely randomized $T_{1:n} \sim \mathrm{CR}(q)$[6] then the asymptotic variance in Theorem 3.2 is

$$V = \mathrm{Var}(c(\psi)) + E\left[\frac{\sigma_1^2(\psi)}{p} + \frac{\sigma_0^2(\psi)}{1-p}\right].$$

This is exactly the Hahn (1998) semiparametric variance bound for ATE with iid observations $(Y, D, \psi(X))$.[7] We achieve the semiparametric variance bound with a simple difference-of-means estimator, without the need for nonparametric re-estimation of the

[6] The latter statement follows from the more general results in Section 3.2.

[7] Recent work in Armstrong (2022) shows that this efficiency bound also holds in settings with covariate-adaptive randomization, including the designs considered here.

known propensity (Hirano et al. (2003)) or covariate adjustment with well-specified outcome models (Robins and Rotnitzky (1995)). We can interpret this result as saying that finely stratified treatment assignment $D_{1:n} \sim \text{Loc}(\psi, p)$ does nonparametric covariate adjustment "by design." See Proposition 3.12 below for a more formal equivalence result.

Representative Sampling. If sampling is locally randomized $T_{1:n} \sim \text{Loc}(\psi, q)$, then the variance due to treatment effect heterogeneity decreases from $\text{Var}(c(\psi))$ to $q\,\text{Var}(c(\psi))$. In this case, V can be strictly smaller than the classical semiparametric variance bound. Intuitively, by using the additional covariates $(\psi(X_i))_{i=1}^n$ to select a representative sample, we finely represent the distribution of treatment effect levels $(c(\psi_i))_{i=1}^n$ in the larger eligible population into our experiment. More formally, consider an oracle setting where we observe the treatment effect level $c(\psi_i)$ for each sampled unit $T_i = 1$, estimating the ATE by the sampled average $\widehat{\theta} = (1/n_T)\sum_i T_i c(\psi_i)$. Our analysis shows that if $T_{1:n} \sim \text{Loc}(\psi, q)$ then

$$(1/n_T)\sum_{i=1}^n T_i c(\psi_i) = E_n[c(\psi_i)] + o_p(n^{-1/2}).$$

Because of this, the sampled average $(1/n_T)\sum_i T_i c(\psi_i)$ behaves like the infeasible average $E_n[c(\psi_i)]$ over all eligible units, including those not sampled into the experiment. This nonparametrically dampens the variance due to treatment effect heterogeneity from $\text{Var}(c(\psi))$ to $q\,\text{Var}(c(\psi))$ for $q \in (0, 1]$.

Remark 3.4 (Comparison with Other Designs)**.** For the case without sampling, we can compare our results on fine stratification to other covariate-adaptive assignment designs. Li et al. (2018) shows that under rerandomized treatment assignment, $\widehat{\theta}$ is asymptotically (almost) as efficient as interacted linear regression adjustment, effectively doing linear regression "by design." Bugni et al. (2019) show that coarsely stratified assignment $\psi(X) \in \{1, \dots, m\}$ is asymptotically equivalent to an interacted linear regression adjustment that includes all strata indicators as covariates. Harshaw et al. (2021) suggest a novel Gram-Schmidt walk design with MSE bounded by a quantity related to linear ridge regression. By contrast, we show that the fine stratification $D_{1:n} \sim \text{Loc}(\psi, p)$ does nonparametric regression adjustment by design.

Extending Bugni et al. (2019), the next example proves an equivalence between coarse stratification and linear regression adjustment in the case where both sampling and assignment are both coarsely stratified.

Example 3.5 (Coarse Stratification)**.** If $T_{1:n} \sim \text{Loc}(S, q)$ and $D_{1:n} \sim \text{Loc}(S, p)$ for a fixed stratification $S(X) \in \{1, \dots, m\}$, Theorem 3.2 shows that $\sqrt{n_T}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V_S)$

$$V_S = q\,\text{Var}(c(S)) + E\left[\frac{\sigma_1^2(S)}{p} + \frac{\sigma_0^2(S)}{1-p}\right]. \tag{3.3}$$

Alternatively, suppose sampling $T_{1:n} \sim \text{CR}(q)$ and assignment $D_{1:n} \sim \text{CR}(p)$ are completely randomized (Example 2.3), with ex-post linear adjustment for the covariate imbalances due to both sampling and assignment. To define the adjustment, let $z_i = (\mathbb{1}(S_i = k))_{k=1}^{m-1}$ denote leave-one-out strata indicators and their de-meaned versions $\tilde{z}_i = z_i - E_n[z_i|T_i = 1]$. Consider the linear regression $Y \sim 1 + D + \tilde{z} + D\tilde{z}$ and let

$\widehat{\tau}$ denote the coefficient on D and $\widehat{\beta}$ the coefficient on $D\tilde{z}$. Define the sampled covariate mean $\bar{z}_{T=1} = E_n[z_i|T_i = 1]$ and eligible covariate mean $\bar{z} = E_n[z_i]$ and consider the doubly-adjusted estimator

$$\widehat{\theta}_{adj} = \widehat{\tau} - \widehat{\beta}'(\bar{z}_{T=1} - \bar{z}).$$

We study this estimator in the next proposition.

Proposition 3.6 (Regression Equivalence)**.** *Assume $E[Y(d)^2] < \infty$ and $P(S = k) > 0$ for all $k \in [m]$. If $T_{1:n} \sim \mathrm{CR}(q)$ and $D_{1:n} \sim \mathrm{CR}(p)$ then $\sqrt{n_T}(\widehat{\theta}_{adj} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V_S)$ with V_S as in Equation 3.5.*

This result shows that under coarsely stratified sampling and assignment, the simple difference-of-means estimator behaves like the doubly-adjusted estimator $\widehat{\theta}_{adj}$ under complete randomization.

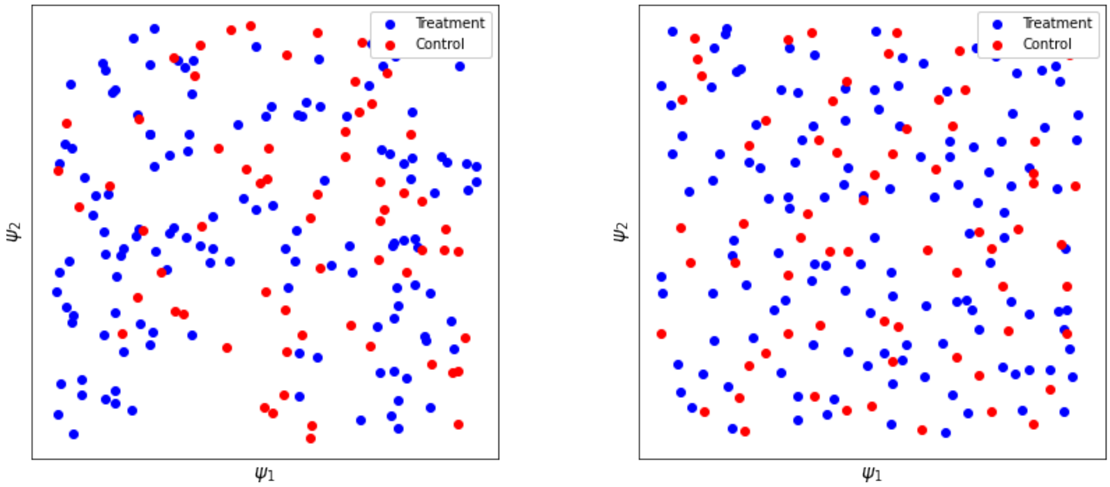
Proposition 3.12 in the next section generalizes this example to the case of fine stratification and nonparametric double adjustment for both sampling and assignment imbalances. The remainder of this subsection provides more intuition for Theorem 3.2, connecting our results on fine stratification to the previous literature on semiparametric efficiency.

Remark 3.7 (Efficient Influence Function)**.** Consider expanding the difference-of-means estimator $\widehat{\theta}$ about the efficient influence function for the ATE. Denote the nonparametric regression residuals $\epsilon_i^d = Y_i(d) - E[Y_i(d)|\psi_i]$. Under locally randomized sampling and assignment

$$\begin{aligned}\widehat{\theta} = E_n[c(\psi_i)] + E_n\left[\frac{D_i\epsilon_i^1}{p} + \frac{(1-D_i)\epsilon_i^0}{1-p}\middle| T_i = 1\right] \\ + \underbrace{\mathrm{Cov}_n(T_i, c(\psi_i))}_{\text{Sampling}}/q + \underbrace{\mathrm{Cov}_n(D_i, b(\psi_i)|\, T_i = 1)}_{\text{Assignment}}/c_p + O_p(n^{-1}).\end{aligned}$$

If $T_i = 1$ for all $i \in [n]$ then the first two terms are exactly the efficient influence function for the ATE. The third term is the estimator error due to correlation between the sampling variables T_i and treatment effect heterogeneity $c(\psi_i)$ among the eligible units. The fourth term is the estimator error due to correlation between treatment assignments D_i and outcome heterogeneity among the sampled units. Without stratification, the chance covariate imbalances produce by randomization contribute non-negligible asymptotic variance, and the errors $\sqrt{n}\,\mathrm{Cov}_n(T_i, c(\psi_i)) \Rightarrow \mathcal{N}(0, v)$ with $v > 0$. By contrast, we show that if $T_{1:n} \sim \mathrm{Loc}(\psi, q)$ then the sampling errors $\sqrt{n}\,\mathrm{Cov}_n(T_i, F(\psi_i)) = o_p(1)$ for any function $E[F(\psi)^2] < \infty$, and similarly for the assignment term. Because of this, the unadjusted estimator $\widehat{\theta}$ is first-order equivalent to the efficient influence function for the ATE under local randomization if $q = 1$. If $q < 1$, the situation is even better, and the first term behaves like the infeasible average $E_n[c(\psi_i)]$ over all eligible units.

Remark 3.8 (Table One)**.** Applied researchers often report tests of covariate balance in "table one." Consider testing for balance of a covariate $F(\psi_i)$. One common approach is to report a p-value for the test that $\beta = 0$ in the regression $F(\psi_i) = \widehat{a} + \widehat{\beta}D_i + e_i$, using the normal limit $\sqrt{n}\widehat{\beta} \Rightarrow \mathcal{N}(0, v)$. By contrast, if $D_{1:n} \sim \mathrm{Loc}(\psi, p)$ then $\sqrt{n}\widehat{\beta} = o_p(1)$ for any covariate $E[F(\psi)^2] < \infty$, showing that the level of such a test converges to zero. Intuitively, this shows that fine stratification with respect to $\psi(X)$ balances any square-integrable transformation $F(\psi)$ to order $o_p(n^{-1/2})$.



(a) Completely Randomized (b) Locally Randomized

Figure 3: Covariates $\psi_i \sim \text{Unif}([0,1]^2)$, $n = 1000$ and sampling and assignment propensities $q = 1/5$ and $p = 2/3$. For complete randomization, the realized propensities (Remark 3.9) widely diverge from q and p in certain parts of the space.

Remark 3.9 (Realized Propensity Score)**.** For any set A with $P(\psi \in A) > 0$ define the *realized* sampling proportions in A by $\widehat{q}_A = E_n[T_i|\psi_i \in A]$. If sampling is completely randomized, the discrepancy between expected and realized propensities $q - \widehat{q}_A = O_p(n^{-1/2})$, q is implemented with errors of order $1/\sqrt{n}$. This is illustrated in Figure 3, where the realized sampling and assignment propensities widely diverge from their nominal levels in certain regions of the space. Such fluctuations of $\widehat{q}_A$ about q increase estimator variance. One way to fix this problem is to nonparametrically re-estimate the realized sampling and assignment proportions $\widehat{q}(\psi)$ and $\widehat{p}(\psi)$ everywhere in the space, as in Hirano et al. (2003), using the propensity weighting

$$\widehat{\theta}_{ipw} = E_n\left[\frac{T_i(D_i - \widehat{p}(\psi_i))Y_i}{\widehat{q}(\psi_i)(\widehat{p} - \widehat{p}^2)(\psi_i)}\right]$$

For experiments, we provide a simpler solution, showing that fine stratification gets the realized propensities right at design-time. In particular, our analysis shows that if $T_{1:n} \sim \text{Loc}(\psi, q)$ then the gap between the realized and target propensities $q - \widehat{q}_{A_n} = o_p(n^{-1/2})$, even for a shrinking sequence of sets with $P(\psi \in A_n) \to 0$ slowly enough. Because of this, we think of the design $T_{1:n} \sim \text{Loc}(\psi, q)$ as a "local" implementation of the propensity q with respect to $\psi(X)$.

3.2 Varying Sampling and Assignment Propensities

This section describes the most general version of our method, providing finely stratified designs with heterogeneous sampling and assignment proportions $q(x)$, $p(x)$. The asymptotics developed in this section allows us to formulate and solve the problem of optimal stratification with heterogeneous costs in Section 4 below.

First, we formally define the procedure. Suppose that $q(x) \in \{a_l/k_l : l \in L\}$ for some finite index set L. Similarly, suppose $p(x) \in \{a'_l/k'_l : l \in L'\}$ with $|L'| < \infty$. Extending our definition, let $T_{1:n} \sim \mathrm{Loc}(\psi, q(x))$ denote the following double stratification procedure:

(1) Partition $\{1, \dots, n\}$ into propensity strata $S_l \equiv \{i : q(X_i) = a_l/k_l\}$.

(2) In each propensity stratum S_l, draw samples $(T_i)_{i \in S_l} \sim \mathrm{Loc}(\psi, a_l/k_l)$.

Equivalently, we partition each propensity stratum S_l into groups $g \subseteq S_l$ of size k_l such that $S_l = \bigsqcup_{g \in \mathcal{G}_l} g$ and the homogeneity condition

$$n^{-1} \sum_g \sum_{i,j \in g} |\psi_i - \psi_j|_2^2 = n^{-1} \sum_l \sum_{g \in \mathcal{G}_l} \sum_{i,j \in g} |\psi_i - \psi_j|_2^2 = o_p(1)$$

In practice, we simply run our matching algorithm separately in each propensity stratum S_l, and draw $(T_i)_{i \in g} \sim \mathrm{CR}(a_l/k_l)$ independently for each $g \in \mathcal{G}_l$. Treatment assignment $D_{1:n} \sim \mathrm{Loc}(\psi, p(x))$ is defined identically, partitioning only the units $\{i : T_i = 1\}$ sampled into the experiment.

Example 3.10 (Budget and Welfare Constraints)**.** Consider a village level experiment, where collecting outcome data is either high cost H or low cost L, depending on village proximity and labor costs. Due to budget constraints, we decide to sample $q(L) = 1/2$ of the low cost villages but only $q(H) = 1/10$ of the high cost villages. To do so, we match the high cost villages into 10-tuples and the low cost villages into pairs using publicly available covariates ψ_1. We randomly sample one village from each 10-tuple and one from each pair. Before assigning treatments, we collect additional survey covariates ψ_2 in each sampled village $T_i = 1$. We label the sampled villages as those likely to benefit most M and least L from the intervention according to our prior. Local policymakers insist on the targeted assignment propensity $p(M) = 2/3$ and $p(L) = 1/3$. We implement this assignment propensity using matched triples on ψ_2, assigning $D = 1$ to $2/3$ of the villages in each M-type triple and $1/3$ in each L-type triple.

Before stating our main result, we extend the definition of our estimator to accommodate varying propensities. Define the double IPW estimator

$$\widehat{\theta}_2 = E_n\left[\frac{T_i D_i Y_i}{q(\psi_i)p(\psi_i)}\right] - E_n\left[\frac{T_i(1-D_i)Y_i}{q(\psi_i)(1-p(\psi_i))}\right] \tag{3.4}$$

If $p(x) = p$ and $q(x) = q$ are constant, then $\widehat{\theta}_2 = \widehat{\theta} + O_p(n^{-1})$, where $\widehat{\theta}$ is the difference-of-means estimator studied in the previous section.[8] Then abusing notation we denote both estimators by $\widehat{\theta}$. The following theorem gives our asymptotic results for fine stratification with varying propensities, extending the fixed propensity results in Theorem 3.2 above. We begin with the special case $\psi_1 = \psi_2 = \psi$ and $q = q(\psi)$, $p = p(\psi)$, all non-random.

Theorem 3.11 (CLT)**.** *Suppose Assumption 3.1 holds. Assume sampling and assignment* $T_{1:n} \sim \mathrm{Loc}(\psi, q(\psi))$ *and* $D_{1:n} \sim \mathrm{Loc}(\psi, p(\psi))$*. Then* $\sqrt{n_T}(\widehat{\theta} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V)$

$$V = E[q(\psi)]\left(\mathrm{Var}(c(\psi)) + E\left[\frac{1}{q(\psi)}\left(\frac{\sigma_1^2(\psi)}{p(\psi)} + \frac{\sigma_0^2(\psi)}{1-p(\psi)}\right)\right]\right)$$

[8] This is because $E_n[D_i] = p + O(n^{-1})$ for stratified designs. It would be false for $D_i \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(p)$.

If $q = 1$ then this is exactly the Hahn (1998) semiparametric variance bound for ATE with iid observations $(Y, D, \psi(X))$ and propensity $p(\psi)$. This shows that under fine stratification the population IPW estimator is already semiparametrically efficient, with no need to nonparametrically re-esimate the known propensities $q(\psi)$ and $p(\psi)$ as in Hirano et al. (2003). Next consider the efficiency gain from stratified sampling. The overall sampling proportion n_T/n has $n_T/n = E[q(\psi)] + o_p(1)$. Then defining $\bar{q} = E[q(\psi)]$, local randomization reduces the variance due to treatment effect heterogeneity from $\mathrm{Var}(c(\psi))$ to $\bar{q}\,\mathrm{Var}(c(\psi))$ for $\bar{q} \in (0, 1]$, similar to before.

Regression Equivalence. Theorem 3.11 shows that if $q = 1$ then difference-of-means is semiparametrically efficient. To the best of our knowledge, no such efficiency bound is available for jointly finely stratified sampling and assignment $T_{1:n} \sim \mathrm{Loc}(\psi, q(\psi))$ and $D_{1:n} \sim \mathrm{Loc}(\psi, p(\psi))$ with $q \neq 1$. Instead, we show a direct equivalence between unadjusted estimation under fine stratification and nonparametric regression adjustment. In particular, the asymptotic variance V above is the same as that achieved by a doubly-robust estimator that adjusts for covariate imbalances during both sampling and assignment. To state the result, consider regression estimators $\widehat{m}_d$ for $m_d(\psi) = E[Y(d)|\psi]$. Define the doubly-augmented IPW (2-AIPW) estimator

$$\widehat{\theta}_{adj} = E_n[\widehat{m}_1(\psi_i) - \widehat{m}_0(\psi_i)] + E_n\left[\frac{T_i D_i (Y_i - \widehat{m}_1(\psi_i))}{q(\psi_i)p(\psi_i)} - \frac{T_i(1 - D_i)(Y_i - \widehat{m}_0(\psi_i))}{q(\psi_i)(1 - p(\psi_i))}\right]$$

$\widehat{\theta}_{adj}$ adjusts for covariate imbalances due to both sampling and assignment. We implement the estimator using cross-fitting, similar to Chernozhukov et al. (2017). See the proof for details. If $q = 1$ this reduces to the familiar AIPW estimator for the ATE. The next result provides an equivalence between doubly-robust nonparametric adjustment under an iid design and unadjusted estimation under a finely stratified design.

Proposition 3.12 (Regression Equivalence)**.** *Require Assumption 3.1. Suppose the estimators $|\widehat{m}_d - m_d|_{2,\psi} = o_p(1)$ are well-specified and consistent. If the design is iid $T_i \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(q(\psi_i))$ and $D_i \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(p(\psi_i))$ then $\sqrt{n_T}(\widehat{\theta}_{adj} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V)$ with the variance V given Theorem 3.11.*

Intuitively, Proposition 3.12 shows that finely stratified sampling and assignment makes the unadjusted estimator $\widehat{\theta}$ as efficient as a more complicated estimator that adjusts nonparametrically for covariate imbalances during both sampling and assignment.

Estimated Design Variables. We wish to formally accommodate the case where design variables $\psi, q(\psi), p(\psi)$ are estimated using previously collected data. To do so, define the random element $\xi \perp\!\!\!\perp (X_i, Y_i(1), Y_i(0))_{i=1}^n$ and let $\psi = \psi(X, \xi)$, $p = p(\psi, \xi)$, and $q = q(\psi, \xi)$. For example, we could let $\xi = (\widehat{m}_d)_{d=0,1}$ be regression estimates of $m_d(x) = E[Y(d)|X = x]$ from a pilot experiment and set $\psi(X, \xi) = (\widehat{m}_0(X), \widehat{m}_1(X))$. Define $c(\psi, \xi) = E[Y(1) - Y(0)|\psi, \xi]$ and $\sigma_d^2(\psi, \xi) = \mathrm{Var}(Y(d)|\psi, \xi)$. The proof of Theorem 3.11 shows that if $T_{1:n} \sim \mathrm{Loc}(\psi, q(\psi, \xi))$ and $D_{1:n} \sim \mathrm{Loc}(\psi, p(\psi, \xi))$ then $\sqrt{n_T}(\widehat{\theta} - \mathrm{ATE})|\xi \Rightarrow \mathcal{N}(0, V(\xi))$ with conditional asymptotic variance

$$V(\xi) = E[q(\psi, \xi)]\left(\mathrm{Var}(c(\psi, \xi)) + E\left[\frac{1}{q(\psi, \xi)}\left(\frac{\sigma_1^2(\psi, \xi)}{p(\psi, \xi)} + \frac{\sigma_0^2(\psi, \xi)}{1 - p(\psi, \xi)}\right)\right]\right) \tag{3.5}$$

All expectations and variances are conditional on ξ. The inference methods provided in Section 6 are asymptotically exact conditional on ξ. Note that marginally over both our experiment and the previous data ξ, the estimator is asymptotically mixed normal $\sqrt{n_T}(\widehat{\theta} - \text{ATE}) \Rightarrow Z$, where Z has characteristic function $E[\exp(-\frac{1}{2}t^2 V(\xi))]$.[9]

Collecting Data After Sampling. In practice, experimenters may want to use different stratification variables ψ_1 for sampling and ψ_2 for assignment with $\psi_1 \neq \psi_2$. For example, ψ_1 may include publicly available administrative data, while ψ_2 includes additional survey covariates collected after sampling units into the experiment. To accommodate this, next we state our most general version of Theorem 3.11. Suppose Assumption 3.1 holds and let sampling and assignment $T_{1:n} \sim \text{Loc}(\psi_1, q(x))$ and $D_{1:n} \sim \text{Loc}(\psi_2, p(x))$. If $q(x) = q(\psi_1)$, require $\psi_1 \subseteq \psi_2$. Otherwise, require $(\psi_1, q) \subseteq \psi_2$. Then $\sqrt{n_T}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, \bar{q}V)$ with $\bar{q} = E[q(X)]$ and asymptotic variance $V = V_1 + V_2$

$$
\begin{aligned}
V_1 &= \text{Var}(c(X)) + E\left[\frac{1}{q(X)}\left(\frac{\sigma_1^2(X)}{p(X)} + \frac{\sigma_0^2(X)}{1-p(X)}\right)\right] \\
V_2 &= E\left[\frac{1-q(X)}{q(X)}\text{Var}(c(X)|\psi_1, q)\right] + E\left[\frac{1}{q(X)}\text{Var}(b(X)|\psi_2, p)\right]
\end{aligned} \tag{3.6}
$$

The variance V does not have a simple form like in the special cases considered above if $\psi_1 \neq \psi_2$. Instead, we expand V relative to the efficient variance V_1, which we conjecture is the semiparametric efficiency bound in this setting. If $\psi_1 = \psi_2 = \psi$, and $q = q(\psi)$, $p = p(\psi)$ then V can be rearranged into the form in Theorem 3.11. The requirement that the stratification is increasing[10] $\psi_1 \subseteq \psi_2$ is subtle. We defer this technical discussion to Remark 10.2 in the appendix.

Optimal Stratification Variables. Inspecting the variance V_2 in Equation 3.6 shows that the minimal dimension efficient stratification variables are $\psi_1^* = c(X)$ and $\psi_2^* = (c(X), q(X), b(X))$. In this case, V_2 is identically zero, and $V = V_1$, the conjectured semiparametric efficiency bound. Since $b(X)$ and $c(X)$ are both propensity-weighted linear combinations of $m_d(X) = E[Y(d)|X]$, setting $\psi_2^* = (q, m_0, m_1)(X)$ is also optimal. We include $(c(X), q(X))$ in ψ_2^* to satisfy the subvector condition $\psi_1 \subseteq \psi_2$. If $q(x) = q(\psi_1)$, we can just take $\psi_2^* = (c(X), b(X))$. In practice, these optimal ψ_1^* and ψ_2^* are not known. The preceding discussion suggests letting $\psi_1(X)$ be a small subvector of the baseline covariates expected to be most predictive of treatment effect heterogeneity, and $\psi_2(X)$ a subvector expected to be predictive of outcomes.[11] If $c(X) = F(\psi_1(X))$ for some function F, then this choice is asymptotically optimal. If none of the baseline covariates predict treatment effect heterogeneity, so that $c(X)$ is constant, then there is no efficiency loss from completely randomized sampling with $\psi_1^* = 1$.

Remark 3.13 (Curse of Dimensionality). Setting $\psi_1(X) = \psi_2(X) = X$ in Equation 3.6 also minimizes the asymptotic variance V. However, there is a curse of dimensionality when matching on many baseline covariates. In particular, our analysis shows that if $E[Y(d)|X = x]$ is Lipschitz continuous, then the finite sample variance converges to the asymptotic limit at rate $n\,\text{Var}(\widehat{\theta}) = V + O_p(n^{-2/(\dim(\psi)+1)})$, which may be slow even

[9]This mixed normal limit was also observed by Cai and Rafi (2023) in a setting with iid treatments.
[10]In fact, we just require that $\psi_1 = g(\psi_2)$ for a measurable function g.
[11]Since $b(X; p) \propto E[Y(1)|X]/p + E[Y(0)|X]/(1-p)$, we should prioritize predicting outcomes in the arm assigned with lowest probability.

in moderate dimensions. Because of this, for fixed n the variance $\mathrm{Var}(\widehat{\theta})$ may be U-shaped in the dimension of the stratification variables, since matching on many irrelevant variables reduces match quality on the relevant variables. This motivates the search for stratification variables $\psi_1(x)$ and $\psi_2(x)$ of small dimension that minimize V.

Remark 3.14 (Sampling Subordinate Assignment)**.** Suppose the sampling propensity q is constant, $p = 1/2$ and consider a matched pair $g = \{i, j\}$ formed during treatment assignment. For a good match with $|\psi_i - \psi_j|_2$ small, the difference in balance functions $|b(\psi_i) - b(\psi_j)|$ will also be small as long as $b(\psi)$ is continuous, which reduces the variance due to random assignment. However, if sampling propensity $q(\psi)$ is not constant, then estimator variance is determined by the *weighted* balance function $b(\psi)/q(\psi)$. If $q(\psi_i) \neq q(\psi_j)$, e.g. because i and j lie across the boundary between different sampling propensity strata, then the weighted difference $|b(\psi_i)/q(\psi_i) - b(\psi_j)/b(\psi_j)|$ may be large even if $|\psi_i - \psi_j|_2$ is small. Such boundary effects are asymptotically negligible, as shown by our theory, but can significantly inflate finite sample variance in small experiments with many sampling strata and highly predictive covariates. One way to prevent this issue is to match units separately within each sampling stratum $\{i : q(\psi_i) = a/k\}$ at the assignment stage. We implement this modification in our empirical application in Section 8 below.

4 Optimal Stratified Designs

In this section we formulate and solve the problem of optimal stratification in survey experiments, characterizing the optimal sampling and assignment propensities for budget-constrained experimentation with heterogeneous costs. For simplicity, we restrict to the case with stratification variables $\psi_1 = \psi_2 = \psi$ and sampling and assignment propensities $q = q(\psi)$ and $p = p(\psi)$.

4.1 Budget-Constrained Sampling Problem

In Section 3, we presented asymptotic results of the form $\sqrt{n_T}(\widehat{\theta} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V)$, normalizing by the experiment size $n_T = \sum_i T_i$. This allowed for easy comparison with the previous literature on covariate-adaptive treatment assignment. However, the experiment size n_T varies with the sampling propensity $q(\psi)$, making this normalization unsuitable for studying the optimal sampling propensity. Because of this, in what follows we instead normalize by the number of eligible units n. Since $n_T/n \xrightarrow{p} E[q(\psi)]$, this just removes the multiplicative factor $E[q(\psi)]$ from our previous results, so that $\sqrt{n}(\widehat{\theta} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V(q, p))$ with

$$V(q, p) = \mathrm{Var}(c(\psi)) + E\left[\frac{1}{q(\psi)}\left(\frac{\sigma_1^2(\psi)}{p(\psi)} + \frac{\sigma_0^2(\psi)}{1 - p(\psi)}\right)\right] \tag{4.1}$$

With this fixed normalization in hand, consider minimizing Equation 4.1 over all sampling propensities $q(\psi)$. Clearly the unconstrained solution is $q^*(\psi) = 1$, making the experiment as large as possible. More generally, we can formalize the problem of experimentation subject to a budget constraint. To set up the problem, define the ex-ante heteroskedasticity function

$$\bar{\sigma}^2(\psi) = \frac{\sigma_1^2(\psi)}{p(\psi)} + \frac{\sigma_0^2(\psi)}{1 - p(\psi)} \tag{4.2}$$

We interpret $\bar{\sigma}^2(\psi)$ as the expected residual variance from sampling a unit with $\psi_i = \psi$, prior to the realization of its treatment assignment $D_i \in \{0, 1\}$. Fixing the propensity $p(\psi)$, the asymptotic variance can be written $V(q) = \text{Var}(c(\psi)) + E[\bar{\sigma}^2(\psi)/q(\psi)]$. Let $C(\psi)$ denote the known, potentially heterogeneous, cost of including a unit of type $\psi(X) = \psi$ in the experiment. More general costs $C(x)$ can be handled easily, and are discussed in Remark 4.2 below. Then the budget-constrained variance minimization problem with budget $\bar{B}$ can be written

$$\min_{0<q\leq 1} E\left[\frac{\bar{\sigma}^2(\psi)}{q(\psi)}\right] \quad \text{s.t.} \quad E[C(\psi)q(\psi)] = \bar{B}. \tag{4.3}$$

The next proposition characterizes the interior solutions to this problem. We assume that $\inf_\psi \sigma_d^2(\psi) \geq c > 0$ and costs $C(\psi) \in [C_l, C_u] \subseteq (0, \infty)$.

Proposition 4.1. *Define the candidate solution*

$$q^*(\psi) = \bar{B} \cdot \frac{\bar{\sigma}(\psi)C(\psi)^{-1/2}}{E[\bar{\sigma}(\psi)C(\psi)^{1/2}]}. \tag{4.4}$$

If $\sup_\psi q^*(\psi) \leq 1$, then q^* is optimal in Equation 4.3.

If the feasibility constraint $\sup_\psi q^*(\psi) \leq 1$ is violated, the optimal sampling propensity may not have a simple analytical form. Remark 5.5 below provides a rounding procedure that can be used to restore feasibility in this case. To build intuition for the form of the solution, consider the following special cases.

(a) Homoskedasticity. Suppose $\sigma_d^2(\psi) = \sigma_d^2$ constant for $d \in \{0, 1\}$ and $p(\psi) = p$. Then the optimal propensity $q^*(\psi) = \bar{B} \cdot C(\psi)^{-1/2}/E[C(\psi)^{1/2}]$ has $q^*(\psi) \propto 1/\sqrt{C(\psi)}$. This provides a simple heuristic for sample allocation with heterogeneous costs.

(b) Homogeneous costs. If $C(\psi) = 1$, then $E[q(\psi)] \leq \bar{B}$ constrains the total proportion of sampled units. In this case write $\bar{B} = \bar{q}$. The optimal solution has form $q^*(\psi) = \bar{q}\bar{\sigma}(\psi)/E[\bar{\sigma}(\psi)]$, with sampling propensity proportional to the ex-ante standard deviation. In particular, we would like to oversample ($q^*(\psi) > \bar{q}$) units of type $\psi(X) = \psi$ that have larger residual standard deviation than the average $E[\bar{\sigma}(\psi)]$, and undersample in the opposite case.

Optimal Spending. Under optimal sampling, the total amount spent on units of type ψ is $C(\psi)q^*(\psi)dP(\psi) \propto \sqrt{\bar{\sigma}^2(\psi)C(\psi)}dP(\psi)$. This shows that we should spend more on units with larger ex-ante variance and larger per unit cost. However, since the optimal propensity undersamples high cost units, spending grows as $\sqrt{C(\psi)}$, instead of linearly as it would if $q(\psi) = q$ were constant.

Remark 4.2 (General Costs)**.** More generally, let $C(x)$ denote the cost of including a unit of type $X = x$. Abusing notation, define the average cost $C(\psi) = E[C(X)|\psi]$. The conclusions of Proposition 4.1 and Theorem 4.4 below remain true as stated, with budget constraint $E[C(X)q(\psi)] = E[C(\psi)q(\psi)] \leq \bar{B}$.

4.2 Globally Optimal Stratification

The main result of this section studies implementation of the globally optimal stratified design, subject to the budget constraint. First, we characterize the optimal assignment propensity. For any fixed sampling propensity $q(\psi)$, the global minimizer of Equation 4.1 is the conditional Neyman allocation $p^*(\psi) = \sigma_1(\psi)/(\sigma_1(\psi) + \sigma_0(\psi))$. In some cases, we may only be interested in implementing a constant assignment propensity $p^* \in (0, 1)$. If q is also constant, then

$$p^* = \sqrt{E[\sigma_1^2(\psi)]} \left(\sqrt{E[\sigma_1^2(\psi)]} + \sqrt{E[\sigma_0^2(\psi)]}\right)^{-1} \tag{4.5}$$

Compare this to the classical Neyman allocation $\sigma_1/(\sigma_1 + \sigma_0)$ with $\sigma_d = \mathrm{SD}(Y(d))$. Here, only the residual variances $\sigma_d^2(\psi) = \mathrm{Var}(Y(d)|\psi)$ enter p^*, since the fluctuations of $Y(d)$ predictable by $\psi(X)$ do not contribute to first-order asymptotic variance under fine stratification.

The jointly optimal sampling and assignment propensities are obtained by plugging the conditional Neyman allocation $p^*(\psi) = \sigma_1(\psi)/(\sigma_1(\psi)+\sigma_0(\psi))$ into the formula for $q^*(\psi; p)$ above. This gives ex-ante variance $\bar{\sigma}^2(\psi) = (\sigma_1(\psi)+\sigma_0(\psi))^2$ and jointly optimal sampling and assignment propensities

$$p^*(\psi) = \frac{\sigma_1(\psi)}{\sigma_1(\psi) + \sigma_0(\psi)} \qquad q^*(\psi; p^*) = \bar{B}\frac{(\sigma_1(\psi) + \sigma_0(\psi))C(\psi)^{-1/2}}{E[(\sigma_1(\psi) + \sigma_0(\psi))C(\psi)^{1/2}]}. \tag{4.6}$$

The propensities $p^*(\psi)$ and $q^*(\psi)$ will generally need to be discretized in order to implement them using fine stratification. To do so, we provide novel asymptotics with both the number of distinct propensity levels as well as the number of units in each group g growing with the sample size n.

Definition 4.3 (Discretization)**.** Let $q_n^*(\psi)$ and $p_n^*(\psi)$ take values in the finite approximating propensity set $\{a_l/k_l : l \in L_n\}$ with levels L_n. Suppose that $|q_n^* - q^*|_\infty = o(1)$ and $|p_n^* - p^*|_\infty = o(1)$. Define maximum group size $\overline{k}_n = \max_{l \in L_n} k_l$ and require that $\overline{k}_n|L_n| = o(n^{1-\frac{d+1}{\alpha}})$ for $E[|\psi(X)|^\alpha] < \infty$.

If $\psi(X)$ is bounded, the final condition simplifies to $\overline{k}_n|L_n| = o(n)$. For example, one way to satisfy Definition 4.3 is to round the optimal propensities to the nearest a/k_n for some sequence $k_n \to \infty$, setting $q_n^*(\psi) = \mathrm{argmin}\{|q^*(\psi) - a/k_n| : 1 \le a \le k_n - 1, k_n = \lfloor n^{1/2-\epsilon} \rfloor\}$ and similarly for $p_n^*(\psi)$. The main theorem of this section shows that such discretizations are asymptotically efficient.

Theorem 4.4 (Optimal Stratification)**.** *Suppose Assumption 3.1. If $T_{1:n} \sim \mathrm{Loc}(\psi, q_n^*(\psi))$ and $D_{1:n} \sim \mathrm{Loc}(\psi, p_n^*(\psi))$. Then $\sqrt{n}(\widehat{\theta} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V^*)$*

$$V^* = \mathrm{Var}(c(\psi)) \;+ \min_{\substack{0<q,p\le 1\\ E[C(\psi)q(\psi)]=\bar{B}}} E\left[\frac{1}{q(\psi)}\left(\frac{\sigma_1^2(\psi)}{p(\psi)} + \frac{\sigma_0^2(\psi)}{1-p(\psi)}\right)\right]$$

The design in Theorem 4.4 minimizes the asymptotic variance over all sampling and assignment propensities, subject to the budget constraint. If we set $q = 1$, then $V^* = \min_{0\le p\le 1} V_H(p)$, minimizing the Hahn (1998) semiparametric efficiency bound for the ATE over all propensities scores. As noted above, Armstrong (2022) shows that this bound also applies to the designs in this paper for the case $q = 1$.

4.3 Finite Sample Optimality

In this final subsection, we briefly discuss exact optimality in finite samples. Consider the case $q = 1$ and $p(X) = a/k$ fixed and constant. In this setting, Bai (2022) shows that if the balance function $b(x)$ were known, then matching units into strata of size k according to their sorted $b(X_i)$ values minimizes $\mathrm{MSE}(\widehat{\theta}|X_{1:n})$ over all stratified designs. By contrast, here we show that if $b(x)$ were known, the class of stratified designs itself would generally be suboptimal. To see this, consider the case $p = 1/2$. Define the complete graph K_n with vertices $\{1, \ldots, n\}$ and edge weights $w_{ij} = b(X_i)b(X_j)$. The Max-Cut optimization problem asks for a partition of the vertices into disjoint sets $E_1 \cup E_0 = \{1, \ldots, n\}$ such that the weight of cut edges between E_1 and E_0 is maximized

$$\max_{E_0, E_1} \sum_{i,j} b(X_i)b(X_j)\mathbb{1}(i \in E_1, j \in E_0) \quad \text{s.t.} \quad E_0 \sqcup E_1 = \{1, \ldots, n\} \tag{4.7}$$

For example, see Rendl et al. (2008) for an overview. Let E_0^*, E_1^* solve the Max-Cut problem in Equation 4.7. Define the optimal treatment allocation $d_{1:n}^* = d_{1:n}^*(X_{1:n})$ by $d_i^* = \mathbb{1}(i \in E_1^*)$ for $1 \leq i \leq n$. Define the *alternating design* $P^*(D_{1:n}|X_{1:n})$ by

$$P^*(D_{1:n} = d_{1:n}^*|X_{1:n}) = P^*(D_{1:n} = 1 - d_{1:n}^*|X_{1:n}) = 1/2$$

with $D_{1:n} \perp\!\!\!\perp W_{1:n}|X_{1:n}$. Our next theorem shows that the the alternating design P^* is globally optimal over the set of all covariate-adaptive designs with fixed treatment probability $P(D_i = 1) = 1/2$. We denote this set of designs by $\mathcal{P}_{1/2} = \{P : P(D_i = 1) = 1/2,\ D_{1:n} \perp\!\!\!\perp W_{1:n}|X_{1:n}\}$. For simplicity, suppose $n = 2m$ for an integer m.

Theorem 4.5 (Optimal Design). *The design P^* has $\mathrm{MSE}_{P^*}(\widehat{\theta}|X_{1:n}) \leq \mathrm{MSE}_P(\widehat{\theta}|X_{1:n})$ for all $P \in \mathcal{P}_{1/2}$.*

The inequality is strict if Problem 4.7 has a unique solution up to permutation of set labels. In particular, note that P^* is not a matched pairs design. Nevertheless, $b(x)$ is not known, so neither the globally optimal design, nor the optimal stratified design from Bai (2022) are feasible. We also caution against plug-in approaches that use a pilot estimate of $b(X_i)$. Section 5.2 below shows that such approaches are equivalent to regression adjustment with regressions estimated in the pilot instead of the main sample, which may perform poorly if the pilot is small. For these reasons, we do not further pursue finite sample optimal designs in this paper.

5 Design with a Pilot Experiment

In this section, we study a procedure that uses pilot data to estimate and implement the solution to the optimal stratification problem derived in the previous section. We show that this feasible version of the optimal design is asymptotically efficient, achieving the budget-constrained optimal variance of Theorem 4.4. In particular, for the case $q = 1$ our procedure minimizes the semiparametric variance bound over all propensity scores, providing the first asymptotically efficient solution to the question of design using a pilot study (Hahn et al. (2011)). The methods in this section are also relevant when observational data from the experimental population or a related previous experiment are available. Small pilot considerations and potential robustifications are discussed in

Remark 5.3 below. Pilot estimation of the optimal stratification variables ψ^* is discussed in Section 5.2 below.

5.1 Feasible Optimal Stratification

Fix stratification variables $\psi_1 = \psi_2 = \psi$ and consider estimating the optimal design for the budget-constrained problem in Equation 4.3. This amounts to using pilot or proxy data to estimate the efficient sampling proportions $q^*(\psi)$ and treatment propensity $p^*(\psi)$. As a proof of concept, we first state our result under large pilot asymptotics, allowing consistent estimation of the heteroskedasticity functions $\sigma_d(\psi)$. We show that the feasible estimated optimal design is asymptotically efficient in the sense of Theorem 4.4. Fixed pilot asymptotics and small sample considerations are discussed in Remark 5.3 below.

In Section 4 we derived the optimal propensities for the budget-constrained design problem

$$q^*(\psi) = \bar{B}\frac{(\sigma_1(\psi) + \sigma_0(\psi))C(\psi)^{-1/2}}{E[(\sigma_1(\psi) + \sigma_0(\psi))C(\psi)^{1/2}]} \qquad p^*(\psi) = \frac{\sigma_1(\psi)}{\sigma_1(\psi) + \sigma_0(\psi)}.$$

The sampling propensity $q^*(\psi)$ is optimal provided the feasibility condition $\sup_\psi q^*(\psi) \leq 1$ is satisfied. Consider pilot heteroskedasticity estimates[12] $\widehat{\sigma}^2_d(\psi)$ for $d = 0, 1$. Define the propensity estimates[13]

$$\widehat{q}(\psi) = \bar{B}\frac{(\widehat{\sigma}_1(\psi) + \widehat{\sigma}_0(\psi))C(\psi)^{-1/2}}{E_n[(\widehat{\sigma}_1(\psi_i) + \widehat{\sigma}_0(\psi_i))C(\psi_i)^{1/2}]} \qquad \widehat{p}(\psi) = \frac{\widehat{\sigma}_1(\psi)}{\widehat{\sigma}_1(\psi) + \widehat{\sigma}_0(\psi)}.$$

In practice, we may find $\widehat{q}(\psi_j) > 1$ for some j. This could be because the condition $\sup_\psi q^*(\psi) \leq 1$ is violated and the optimal sampling problem does not have an interior solution, or just due to statistical error. However, we can transform $\widehat{q}(\psi)$ into an admissible sampling propensity by an iterative rounding procedure, described in Remark 5.5 below. Suppose we have done so and let $\widehat{q}_n(\psi)$ and $\widehat{p}_n(\psi)$ be a sequence of discretizations of $\widehat{q}(\psi)$ and $\widehat{p}(\psi)$, satisfying the conditions in Definition 4.3. We require the following technical conditions, including consistency of the pilot heteroskedasticity estimates.

Assumption 5.1. *Assume the interior solution condition $\sup_\psi q^*(\psi) \leq 1$. Require the variance regularity condition $\inf_\psi \sigma_d(\psi) > 0$ and $(\sigma_1/\sigma_0)(\psi) \in [c_l, c_u]$ with $0 < c_l < c_u < \infty$. Require pilot estimation rate $|\sigma^2_d - \widehat{\sigma}^2_d|_{2,\psi} = O_p(n^{-r})$ for some $r > 0$. Require discretization rate $\overline{k}_n|L_n| = o(n^{1-(\dim(\psi)+1)/\alpha_1})$ and $\overline{k}_n = \omega(n^{1/\alpha_2})$. Assume the moments $E[Y(d)^4] < \infty$, $E|\psi(X)|_2^{\alpha_1} < \infty$ for $\alpha_1 \geq \dim(\psi) + 1$, and $E[|m_d(\psi)|^{\alpha_2}] < \infty$ for $\alpha_2 \geq 1/r$. Assume costs $0 < C(\psi) < \infty$ for all ψ.*

Our main result shows that finely stratified implementation of the optimal propensity estimates is asymptotically fully efficient.

Theorem 5.2 (Pilot Design)**.** *Impose Assumption 5.1. Suppose $T_{1:n} \sim \mathrm{Loc}(\psi, \widehat{q}_n(\psi))$*

---

[12] In practice, we use a modification of Fan and Yao (1998) to estimate variance functions. See appendix section 10.4 for details.

[13] Note in $\widehat{q}(\psi)$ the average is taken over the main experiment covariates, allowing for covariate shift between pilot and main experiment.

and $D_{1:n} \sim \mathrm{Loc}(\psi, \widehat{p}_n(\psi))$. *Then* $\sqrt{n}(\widehat{\theta} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V^*)$

$$V^* = \mathrm{Var}(c(\psi)) \; + \min_{\substack{0<q,p\leq 1 \\ E[C(\psi)q(\psi)]=\bar{B}}} E\left[\frac{1}{q(\psi)}\left(\frac{\sigma_1^2(\psi)}{p(\psi)} + \frac{\sigma_0^2(\psi)}{1-p(\psi)}\right)\right]$$

For intuition, it is also helpful to consider certain special cases of Theorem 5.2. If we fix $q = 1$, the design $D_{1:n} \sim \mathrm{Loc}(\psi, \widehat{p}_n(\psi))$ asymptotically minimizes the Hahn (1998) variance bound over all propensity scores

$$V^* = \min_{0\leq p\leq 1} V_H(p) = \mathrm{Var}(c(\psi)) + \min_{0\leq p\leq 1} E\left[\frac{\sigma_1^2(\psi)}{p(\psi)} + \frac{\sigma_0^2(\psi)}{1-p(\psi)}\right]. \tag{5.1}$$

If $\widehat{p}^*$ is a consistent pilot estimate of the optimal constant propensity p^* in Equation 4.5, then the design $D_{1:n} \sim \mathrm{Loc}(\psi, \widehat{p}_n)$ has $\sqrt{n}(\widehat{\theta} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V)$ with

$$V^* = \mathrm{Var}(c(\psi)) + \min_{p\in(0,1)} E\left[\frac{\sigma_1^2(\psi)}{p} + \frac{\sigma_0^2(\psi)}{1-p}\right]$$

Remark 5.3 (Small Pilots)**.** For small pilots, the asymptotic results in Theorem 5.2 imposing consistent estimation of the variance function $\sigma_d^2(\psi)$ may not reflect finite sample performance. Instead consider an inconsistent variance approximation using the squared residuals from a linear regression. Let $\widehat{\epsilon}_i^{\,d}$ be residuals from regressions $Y(d) \sim 1 + \psi$ in $\{D_i = d\}$. For pilot data $(W_j)_{j=1}^m$ compute $\widehat{s}_d = E_m[(\widehat{\epsilon}_i^{\,d})^2]^{1/2}$ and form the pilot estimate $\widehat{p}^* = \widehat{s}_1/(\widehat{s}_1 + \widehat{s}_0)$, rounding it to a close rational number $\widehat{p} = a/k$. By the version of Theorem 3.11 in Equation 3.5, if $D_{1:n} \sim \mathrm{Loc}(\psi, \widehat{p})$ then $\sqrt{n}(\widehat{\theta} - \mathrm{ATE})|W_{1:m} \Rightarrow \mathcal{N}(0, V)$

$$V = \mathrm{Var}(c(\psi)) + E\left[\frac{\sigma_1^2(\psi)}{\widehat{p}} + \frac{\sigma_0^2(\psi)}{1-\widehat{p}}\middle| W_{1:m}\right]$$

The rounding of $\widehat{p}^*$ to $\widehat{p} = a/k$ adds robustness. For example, if $p^* = 1/2$ and our pilot estimate $\widehat{p}^* = .57$, we would round to $\widehat{p} = 1/2$ except in very large experiments. See the discussion of discretization in Remark 5.4 below. In practice, this procedure could be further robustified by constructing a confidence interval for p^* and checking that it excludes a baseline choice such as $p = 1/2$.

Remark 5.4 (Discretization)**.** Consider a pilot estimate $\widehat{p}(\psi) = .637$. This could be rounded to any of $\widehat{p} = 2/3,\ 3/5,\ 13/20,\ 63/100$ and so on. If ψ is bounded, Assumption 5.1 requires that $k_n = o(\sqrt{n})$ for the rounding scheme a/k_n. This condition gives some quantitative guidance about discretization fineness. For example, if $n = 400$ we would rule out $\widehat{p} = 13/20$. In our simulations and empirical application, it's often possible to choose the number of discretization levels just by inspecting the histogram of the estimated $\widehat{p}(\psi)$ and $\widehat{q}(\psi)$.

Remark 5.5 (Feasible Sampling)**.** In practice, we may find that $\widehat{q}(\psi_j) > 1$ for some j, violating the sampling constraint. To fix this, define an index set $J = \emptyset$ and implement the following iterative rounding procedure. (1) Find the largest $\widehat{q}(\psi_j) > 1$. Set $\widehat{q}(\psi_j) = 1$ and add j to J. (2) Recompute the sampling propensity according to

$$\widehat{q}(\psi_i) = \frac{\bar{B} - (1/n)\sum_{i\in J} C(\psi_i)}{1 - |J|/n} \frac{(\widehat{\sigma}_1(\psi_i) + \widehat{\sigma}_0(\psi_i))C(\psi_i)^{-1/2}}{E_n[(\widehat{\sigma}_1(\psi_l) + \widehat{\sigma}_0(\psi_l))C(\psi_l)^{1/2}|l \notin J]} \qquad \forall i \notin J.$$

If $\max_{i=1}^n \widehat{q}(\psi_i) \leq 1$, stop. Otherwise, return to (1). This procedure satisfies the in-sample budget constraint $E_n[\widehat{q}(\psi_i)C(\psi_i)] = \bar{B}$ after each iteration and terminates with $\max_{i=1}^n \widehat{q}(\psi_i) \leq 1$.

Remark 5.6 (Optimal Stratification Trees)**.** Tabord-Meehan (2022) suggests using pilot data to estimate a stratification $\widehat{S}$ and assignment propensity $\widehat{p}(\widehat{S})$ over a set of tree partitions $\widehat{S} \in \mathcal{T}$ of the covariate space. If $\widehat{S} = \widehat{S}(\psi)$ then in our notation their Theorem 3.1 implies that $\sqrt{n}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$

$$V = \text{Var}(c(\psi)) + \min_{S \in \mathcal{T}} \left(E[\text{Var}(b(\psi; p^*(S))|S)] + E\left[\frac{\sigma_1^2(\psi)}{p^*(S)} + \frac{\sigma_0^2(\psi)}{1 - p^*(S)}\right]\right).$$

The optimal propensity $p^*(S) = \sigma_1(S)/(\sigma_1(S) + \sigma_0(S))$ with $\sigma_d^2(S) = \text{Var}(Y(d)|S)$. The display shows that the optimal stratification tree chooses a compromise between two different forces. In the first term, it tries to minimize the variance due to covariate imbalance by choosing strata S that predict propensity-weighted outcomes well, minimizing $E[\text{Var}(b(\psi; p^*(S))|S)]$. In the second term, it tries to minimize the residual variance by choosing strata S such that $p^*(S)$ is close to the optimal propensity $p^*(\psi) = \sigma_1(\psi)/(\sigma_1(\psi) + \sigma_0(\psi))$. By contrast, we implement a discretized consistent estimate $\widehat{p}_n(\psi)$ of the optimal propensity $p^*(\psi)$ using fine stratification. This makes the middle term above asymptotically lower order and globally minimizes the residual term, without any first-order tradeoff (Equation 5.1). Of course, if $S = S(X)$ uses different covariates than our stratification variables ψ, then the efficiencies cannot be ranked.

5.2 Estimating Stratification Variables

Section 3 showed that the stratification variables $\psi_1^* = c(X)$ and $\psi^* = (c, b)(X)$ were asymptotically efficient for both sampling and assignment. This suggests setting $\psi_1(X) = \widehat{c}(X)$ and $\psi_2(X) = (\widehat{c}(X), \widehat{b}(X))$, using pilot estimates of the various regression functions. In our notation, the design $D_{1:n} \sim \text{Loc}(\widehat{b}, p)$ was proposed in Bai (2022) for the case with iid sampling.

Equivalence with Regression Adjustment. Our first result is negative, suggesting that for small pilots such an approach would be dominated by not using the pilot data at all at the design stage, drawing treatments iid and doing regression adjustment in the main sample. For simplicity, set $q = 1$ and let $D_{1:n} \sim \text{Loc}(\widehat{b}, p)$ be the design with pilot-estimated stratification variables. Let $\widehat{\theta}$ be the difference of means estimator formed using the data $W_{1:n} = (D_i, X_i, Y_i(D_i))_{i=1}^n$. Separately, define iid treatments $\check{D}_i \sim \text{Bernoulli}(p)$ and let $\widehat{\theta}_{adj}$ be the cross-fit AIPW estimator in Proposition 3.12, estimated using the alternate data $\check{W}_{1:n} = (\check{D}_i, X_i, Y_i(\check{D}_i))_{i=1}^n$ using regression estimators $\check{m}_d(\psi)$. Define $\check{b}(X)$ by plugging in $\check{m}_d(\psi)$ to the formula in Equation 3.1. Then the estimators $\widehat{\theta}$ and $\widehat{\theta}_{adj}$ have identical expansions

$$\widehat{\theta} = E_n[c(X_i)] + E_n[(D_i - p)(b - \widehat{b})(X_i)]/c_p + R_n + O_p(n^{-1})$$
$$\widehat{\theta}_{adj} = E_n[c(X_i)] + E_n[(\check{D}_i - p)(b - \check{b})(X_i)]/c_p + \check{R}_n$$

for a constant c_p. The residual terms $\sqrt{n}R_n, \sqrt{n}\check{R}_n \Rightarrow \mathcal{N}(0, v)$ for $v > 0$ and are mean-independent of the first term. Denote the imbalance terms $B_n = E_n[(D_i - p)(b - \widehat{b})(X_i)]$

and $\check{B}_n = E_n[(\check{D}_i - p)(b - \check{b})(X_i)]$. These terms control estimator error due to covariate imbalances between the treatment arms. With pilot regression error $r_n^{pilot} = \max_d \|\widehat{m}_d - m_d\|_{2,\psi}$ and main sample regression error $r_n^{main} = \max_d \|\check{m}_d - m_d\|_{2,\psi}$, it's easy to show

$$\sqrt{n}B_n = O_p(r_n^{pilot}) \quad \text{and} \quad \sqrt{n}\check{B}_n = O_p(r_n^{main}).$$

We expect pilot estimation error to be larger $r_n^{pilot} \gg r_n^{main}$ if the pilot is much smaller than the main sample.

Robust Approach. The discussion above showed that, for small pilots, the design $D_{1:n} \sim \text{Loc}(\widehat{b}, p)$ behaves like a noisy version of the cross-fit AIPW estimator $\widehat{\theta}_A$, with regression adjustments estimated using the pilot instead of the main experiment. However, the Bai (2022) approach could still dominate if e.g. $\widehat{b}$ is estimated consistently from a large observational dataset or a larger previous experiment with closely related covariates and potential outcomes. In any case, the large pilot asymptotics in Bai (2022) can be extended to show that the two-stage sampling and assignment design $T_{1:n} \sim \text{Loc}(\widehat{c}, q)$ and $D_{1:n} \sim \text{Loc}((\widehat{c}, \widehat{b}), p)$ achieves the optimal variance V_1 in Equation 3.6. Another natural idea is to robustify the Bai (2022) approach, setting $T_{1:n} \sim \text{Loc}((\widehat{c}, \widehat{b}, \psi'), q)$ and $D_{1:n} \sim \text{Loc}((\widehat{c}, \widehat{b}, \psi'), p)$ for stratification variables ψ' expected to be predictive of both treatment effects and outcomes ex-ante. This can then be combined with the methods in the previous section, setting $\psi = (\widehat{c}, \widehat{b}, \psi')$ and proceeding as in Section 5.1. The efficiency of such designs under fixed pilot asymptotics is described by Equation 3.5. Conditionally asymptotically exact inference, conditional on the pilot data, is available using the methods in Section 6.

Remark 5.7 (Imbalance Term vs. Residual Variance)**.** Equation 3.2 shows that under completely randomized sampling and assignment $\sqrt{n_T}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$ with

$$V = \text{Var}(c(X)) + \text{Var}(b(X)) + E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1-p}\right].$$

The middle term $\text{Var}(b(X))$ is the variance due to covariate imbalance, and the third term is the residual variance. We can think of $\text{Var}(b(X))$ as the "easier" term. We can make this term asymptotically negligible by any one of the following: (1) fine stratification on $\psi(X) = X$ under weak assumptions ($\dim(X)$ small) (2) ex-post propensity reweighting under a smoothness condition (3) ex-post regression adjustment under well-specification (4) the Bai (2022) design with a large enough pilot, or any combination of these methods. By contrast, after treatments have been assigned, neither regression adjustment nor propensity reweighting can help us further minimize the semiparametric variance bound

$$V_H(p) = \text{Var}(c(X)) + E\left[\frac{\sigma_1^2(X)}{p(X)} + \frac{\sigma_0^2(X)}{1-p(X)}\right]$$

In this sense, the residual variance in this expression is the "harder" quantity. To affect it, we need to change the law of the data-generating process by changing the treatment and sampling proportions at design-time, as we have implemented in the previous sections.

6 Inference Methods

This section provides new methods for asymptotically exact inference on the ATE under two-stage locally randomized designs. To do so, we generalize pairs-of-pairs[14] type methods to accommodate designs with both finely stratified sampling and assignment, as well as varying propensities $q(\psi), p(\psi)$. Our inference methods enable applied researchers to report smaller confidence intervals that fully reflect the efficiency gains from our proposed designs.

For each assignment group $g \in \mathcal{G}_n$, define the centroid $\bar{\psi}_g = |g|^{-1} \sum_{i\in g} \psi_i$. Let $\nu : \mathcal{G}_n \to \mathcal{G}_n$ be a bijective matching between groups satisfying $\nu(g) \neq g$, $\nu^2 = \mathrm{Id}$, and the homogeneity condition

$$\frac{1}{n} \sum_{g\in\mathcal{G}_n} |\bar{\psi}_g - \bar{\psi}_{\nu(g)}|_2^2 = o_p(1) \tag{6.1}$$

In practice, ν is obtained by matching the group centroids $\bar{\psi}_g$ into pairs using the algorithm in Section 2. Let $\mathcal{G}_n^\nu = \{g \cup \nu(g) : g \in \mathcal{G}_n\}$ be the unions of paired groups formed by this matching. Define $a(g) = \sum_{i\in g} D_i$ and $k(g) = |g|$. Define the propensity weights $w_i^1 = (1-p_iq_i)/(p_iq_i)^2$ and $w_i^0 = (1-q_i(1-p_i))/(q_i(1-p_i))^2$. Finally, define the variance estimator components

$$\widehat{v}_1 = n^{-1} \sum_{g\in\mathcal{G}_n^\nu} \frac{1}{a(g)-1} \sum_{i\neq j\in g} Y_iY_jD_iD_j(w_i^1w_j^1)^{1/2}$$
$$\widehat{v}_0 = n^{-1} \sum_{g\in\mathcal{G}_n^\nu} \frac{1}{(k-a)(g)-1} \sum_{i\neq j\in g} Y_iY_j(1-D_i)(1-D_j)(w_i^0w_j^0)^{1/2}$$
$$\widehat{v}_{10} = n^{-1} \sum_{g\in\mathcal{G}_n} \frac{k}{a(k-a)}(g) \sum_{i,j\in g} Y_iY_jD_i(1-D_j)(q_iq_j)^{-1/2}$$

Our inference strategy begins with the sample variance of the double-IPW estimator (Equation 3.4), which is consistent for the true asymptotic variance under an iid design, but too large under stratified designs. We correct this sample variance using the estimators above, which measure how well the stratification variables predict observed outcomes in local regions of the covariate space. Define the variance estimator

$$\widehat{V} = \mathrm{Var}_n\left(\frac{T_i(D_i - p(\psi_i))Y_i}{q(\psi_i)(p-p^2)(\psi_i)}\right) - \widehat{v}_1 - \widehat{v}_0 - 2\widehat{v}_{10}. \tag{6.2}$$

Our main result shows that $\widehat{V}$ is consistent for the limiting variance of Theorem 3.11, enabling asymptotically exact inference.

Theorem 6.1 (Inference)**.** *Assume the conditions of Theorem 3.11. If sampling* $T_{1:n} \sim \mathrm{Loc}(\psi, q(\psi))$ *and assignment* $D_{1:n} \sim \mathrm{Loc}(\psi, p(\psi))$*, then* $\widehat{V} = V + o_p(1)$.

By Theorem 6.1 and the CLT in Section 3, the confidence interval $\widehat{C} = [\widehat{\theta} \pm \widehat{V}^{1/2} c_{1-\alpha/2}/\sqrt{n}]$ with $c_\alpha = \Phi^{-1}(\alpha)$ is asymptotically exact in the sense that $P(\mathrm{ATE} \in \widehat{C}) = 1 - \alpha +$

[14] Also known as the method of collapsed strata, as in Hansen et al. (1953). See Abadie and Imbens (2008) and Bai et al. (2021) for recent analyses.

$o(1)$. Importantly, note that the scaling is by number of eligible units n, not the smaller experiment size $n_T = \sum_i T_i \le n$.

7 Simulations

This section presents simulations exhibiting the finite sample properties of our method. Our asymptotic results show separate variance reductions from each of the following: (a) finely treatment assignment, (b) finely stratified sampling and assignment, (c) using the optimal propensities $q^*(\psi)$ and $p^*(\psi)$ (generally infeasible), and (d) using pilot estimates $\widehat{q}(\psi)$ and $\widehat{p}(\psi)$ of the optimal propensities (feasible). In particular, designs (a)-(d) are weakly increasing in asymptotic efficiency. To quantify the marginal efficiency gain from each of our proposed methods in finite samples, we simulate unadjusted ATE estimation under the following designs:

CR: Complete randomization $T_{1:n} \sim \mathrm{CR}(q_k^*)$ and $D_{1:n} \sim \mathrm{CR}(p)$, with q_k^* a discretization of the budget-exhausting sampling propensity $q^* = \bar{B}/E[C(\psi)]$.[15]

CR, Loc: As in CR but with stratified assignment $D_{1:n} \sim \mathrm{Loc}(\psi, p)$.

Loc: Stratified sampling and assignment $T_{1:n} \sim \mathrm{Loc}(\psi, q_k^*)$ and $D_{1:n} \sim \mathrm{Loc}(\psi, p)$.

Hom: As in **Loc** but with $q_{hom,k}^*(\psi)$ a discretization of $q_{hom}^*(\psi) = \bar{B} \cdot C(\psi)^{-1/2}/E[C(\psi)^{1/2}]$, the optimal sampling propensity assuming homoskedasticity. This is feasible but may be misspecified.

Opt: As in **Loc** but with $q_k^*(\psi)$ and p_k^*[16] discretizations of the optimal propensities from Section 4, using oracle knowledge of the heteroskedasticity functions $\sigma_d^2(\psi)$ (infeasible).

Pilot S/L: As in **Opt**, replacing the unknown optimal propensities $q_k^*(\psi)$ and p_k^* with $\widehat{q}_k(\psi)$ and $\widehat{p}_k$ estimated from a pilot of size (S) $n_{pilot} = 100$ or (L) $n_{pilot} = 400$. Variance functions $\sigma_d^2(\psi)$ are estimated using a modification of Fan and Yao (1998), see appendix section 10.4 for details.

Discretization - In each of the designs with varying propensities we discretize by choosing $q_k(\cdot)$ to minimize $E_n[(q_k - q^*)^2(\psi_i)]$ over the set of discretizations $\{q : q(\psi) \in a/10 : a = 1, \dots, 10\}$ subject to number of propensity levels $L = |\mathrm{Image}(q)| \le 3$.

Theory Predictions. Our results predict that there will be large variance reduction from stratified assignment, design **CR** to **CR, Loc** in the notation above, if ψ predicts outcomes $Y(d)$ well. Similarly, there will be large variance reductions from stratified sampling (**CR, Loc** to **Loc**) if ψ predicts treatment effects $Y(1) - Y(0)$ well, from **Loc** to **Hom** if costs are heterogeneous and heteroskedasticity is mild, and from **Hom** to **Opt** if there is significant heteroskedasticity. Moreover, the confidence intervals from Section 6 should have close to nominal coverage. Let $Y_i(d) = m_d(\psi_i) + \sigma_d^2(\psi_i)\epsilon_i^d$ with $E[\epsilon_i^d|\psi_i] = 0$ and $\mathrm{Var}(\epsilon_i^d|\psi_i) = 1$ for $d = 0, 1$, and denote $\nu = \dim(\psi)$. To test the predictions of our theory, we simulate data from the following DGP's:

[15] In particular, we let $q_k^* = a/k$, using the minimal k such that $q_k^* \cdot E_n[C(\psi_i)] \in [.95\bar{B}, 1.05\bar{B}]$.

[16] For simplicity, we focus on the optimal constant assignment propensity under fine stratification (Equation 4.5).

Model 1: Outcomes $m_0(\psi) = \beta_0'\psi$ and $m_1(\psi) = \beta_1'\psi + \psi' Q\psi$ with $\beta_0 = 0$, $\beta_1 = 3 \cdot \mathrm{vec}(1/m : m \in [\nu])$, and $Q = (1/2)\mathbb{1}\mathbb{1}'$ with $\psi \sim \mathrm{Unif}([-1,1]^\nu)$. Costs $c(\psi) = \mathbb{1}(\psi_1 \leq 0) + 10 \cdot \mathbb{1}(\psi_1 > 0)$, budget constraint $\bar{B} = 4$, and baseline treatment proportions $p = 3/8$. Residuals $\epsilon^d \sim \mathcal{N}(0, I_\nu)$ with $\sigma_0^2 = 1$ and $\sigma_1^2 = 9$.

Model 2: As in Model 1, but $c(\psi) = \mathbb{1}(\psi_1 \leq 0) + 4 \cdot \mathbb{1}(\psi_1 > 0)$, budget constraint $\bar{B} = 1$, and $p = 1/2$. Heteroskedasticity functions $\sigma_0^2(\psi) = 5$ and $\sigma_1^2(\psi) = 5 + 30 \cdot \|\psi\|_2^2/\nu$ and $\epsilon^d \sim \mathrm{Unif}([-1,1])$.

Model 3: As in Model 1, but $c(\psi) = 1/2 + 10 \cdot \|\psi\|_2^2/\nu$, $\bar{B} = 2$, and $\sigma_d^2(\psi) = 2$.

Model 4: As in Model 1, but with $\beta_0 = 2 \cdot \mathrm{vec}(1/k : k \in [\nu]) = (2/3)\beta_1$, and $Q = 2\mathbb{1}\mathbb{1}'$. Costs as in Model 3 and heteroskedasticity as in Model 2.

Model 5: As in Model 1, but with $m_d(\psi) = 5 \cdot g(\beta_d'\psi)$ for $\beta_0 = \beta_1 = 10 \cdot \mathrm{vec}(1/m : m \in [\nu])$ and g the Cauchy CDF, $p = 1/2$, and $\sigma_d^2(\psi) = 2$.

Model 6: As in Model 1, but outcomes $m_1(\psi) = 4\sum_{m=1}^{\nu} \sin(\psi_m) + 2\mathbb{1}'\psi$ and $m_0(\psi) = 2\sum_{m=1}^{\nu} \cos(\psi_m)$ and $\psi \sim \mathrm{Unif}([-\pi,\pi]^\nu)$, with $p = 3/10$ and $\sigma_d^2(\psi) = 6$.

		$n = 800$, $\dim(\psi) = 2$						$n = 400$, $\dim(\psi) = 6$					
	Design, DGP	1	2	3	4	5	6	1	2	3	4	5	6
SD	CR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	CR, Loc	0.96	0.94	0.81	0.87	0.56	0.62	0.89	0.96	0.86	0.92	0.66	0.76
	Loc	0.90	0.93	0.74	0.84	0.54	0.58	0.89	0.94	0.81	0.86	0.67	0.73
	Hom.	0.93	0.91	0.74	0.89	0.55	0.58	0.88	0.93	0.81	0.88	0.66	0.74
	Opt.	0.80	0.87	0.72	0.79	0.56	0.55	0.81	0.90	0.68	0.76	0.64	0.59
	Pilot S	0.87	0.91	0.68	0.83	0.56	0.52	0.78	0.86	0.75	0.73	0.66	0.57
	Pilot L	0.82	0.87	0.70	0.77	0.56	0.52	0.78	0.90	0.74	0.70	0.65	0.56
%ΔCI	CR	0	0	0	0	0	0	0	0	0	0	0	0
	CR, Loc	-7	-3	-18	-12	-42	-35	-7	-3	-9	-6	-30	-18
	Loc	-9	-6	-24	-13	-42	-40	-9	-5	-11	-7	-29	-20
	Hom.	-14	-9	-26	-14	-45	-41	-13	-8	-10	-5	-33	-23
	Opt.	-21	-12	-30	-24	-45	-47	-21	-9	-24	-21	-33	-43
	Pilot S	-21	-12	-29	-22	-43	-46	-20	-11	-21	-22	-30	-40
	Pilot L	-21	-13	-30	-24	-44	-46	-21	-11	-23	-25	-32	-40
Cover	CR	0.95	0.95	0.95	0.95	0.95	0.95	0.94	0.94	0.94	0.95	0.95	0.94
	CR, Loc	0.95	0.95	0.94	0.95	0.96	0.96	0.96	0.95	0.96	0.96	0.97	0.97
	Loc	0.95	0.95	0.95	0.96	0.95	0.96	0.95	0.95	0.97	0.96	0.97	0.97
	Hom.	0.92	0.95	0.96	0.95	0.95	0.90	0.94	0.94	0.97	0.97	0.95	0.96
	Opt.	0.92	0.96	0.95	0.96	0.95	0.87	0.94	0.95	0.97	0.96	0.96	0.94
	Pilot S	0.94	0.95	0.96	0.95	0.95	0.95	0.96	0.96	0.95	0.96	0.96	0.96
	Pilot L	0.94	0.96	0.94	0.95	0.95	0.96	0.95	0.95	0.95	0.96	0.96	0.96

Table 1: Simulation Results

Table 1 presents our main results, computed using 1000 Monte Carlo repetitions, discretizing propensities to level $k = 8$ in all cases. "SD" denotes estimator standard deviation relative to design **CR**.[17] "%ΔCI" is percent change in confidence interval length relative to the length under **CR**, using the asymptotically exact confidence intervals from Section 6 for all designs.

[17]Note that $E[\widehat{\theta} - \mathrm{ATE}] = 0$ for all designs and estimators, so we only report standard deviation.

Results. Our simulation results agree with the theory predictions above. For example, Models 5 and 6 show large variance reductions from stratified assignment, but not from stratified sampling since ψ is not predictive of treamtent effects $Y(1) - Y(0)$ in these models. By contrast, stratified sampling significantly reduces variance in Models 1-3, where ψ is predictive of treatment effects. The simulation results also show efficiency gains from incorporating costs $C(\psi)$ into the sampling propensity $q^*_{hom}(\psi)$, even without knowledge of the true heteroskedasticity function, as in design **Hom**. Models 2 and 4 have significant heteroskedasticity, and show further variance reduction from incorporating oracle knowledge of $\sigma^2_d(\psi)$ into the globally optimal sampling propensity $q^*(\psi)$ in **Opt**. Design **Pilot S** is less efficient than **Opt** due to noisy estimation of $\sigma^2_d(\psi)$ using a small pilot, while **Pilot L** is almost optimal. The results also show approximately nominal coverage, and the confidence interval shrinkage in panel 2 reflects the efficiency gains in panel 1.

8 Empirical Results

In this section, we quantify the performance of each of our designs on $N = 9$ real DGP's from experimental papers, covering a range of fields in applied economics. We use data from 7 experimental papers published in AER between May 2021 and November 2022, excluding those for which data is unavailable or that don't fit into our framework for various reasons, e.g. having multiple interventions on the same unit with a time series structure. The included papers are Abebe et al. (2021), Baysan (2022), Casey et al. (2021), Dellavigna et al. (2022), Domurat et al. (2021), Hussam et al. (2022), and Lowe (2021). We additionally include data from Banerjee et al. (2021), a study with significant treatment effect heterogeneity as recently analyzed by Chernozhukov et al. (2023), as well as data from the Oregon health insurance experiment, reported in Finkelstein et al. (2012).

For each paper above, we impute missing potential outcomes for all units, setting $\tilde{Y}_i(d) = Y_i(d)\mathbb{1}(D_i = d) + \widehat{Y}_i(d)\mathbb{1}(D_i \neq d)$. Following the empirical exercise in Bai (2022), we use the matching-based imputation $\widehat{Y}_i(d) = Y_{j(i)}(d)$ with $j(i) = \text{argmin}_{j:D_j=d} |\psi_i - \psi_j|_2$.[18] Let N_0 denote the size of the original experiment. Using this full panel of imputed potential outcomes, we do the following:

(1) Draw $(\tilde{Y}_i(0), \tilde{Y}_i(1), \psi_i)$ for $i = 1, \ldots, n$ with replacement from $(\tilde{Y}_i(0), \tilde{Y}_i(1), \psi_i)_{i=1}^{N_0}$.

(2) Randomize $T_{1:n}$ and $D_{1:n}$ according to one of the designs (a) **CR** (b) **CR, Loc** (c) **Loc** (d) **Hom** and (e) **Pilot S/L** in Section 7 above, excluding the infeasible design **Opt**.

(3) Reveal outcomes $\tilde{Y}_i = T_i D_i \tilde{Y}_i(1) + T_i(1-D_i)\tilde{Y}_i(0)$ for $i = 1, \ldots, n$, form the estimator $\widehat{\theta}$ and confidence interval $\widehat{C} = [\widehat{\theta} \pm \widehat{V}^{1/2} c_{1-\alpha/2}/\sqrt{n}]$ for $\alpha = 0.05$.

Since the ATE for the imputed DGP $(\tilde{Y}_i(0), \tilde{Y}_i(1), \psi_i)_{i=1}^{N}$ is known, this approach allows us to compute the standard deviation, coverage probabilities, and percent reduction in confidence interval length for each DGP and design. In particular, this exercise allows us to quantify the marginal variance reduction from each component of our methods on the

[18] Model-based imputation of $Y_i(d) = m_d(\psi_i) + \sigma^2_d(\psi_i)\epsilon^d_i$ with $E[\epsilon^d_i|\psi_i] = 0$, $\text{Var}(\epsilon^d_i|\psi_i) = 1$, and $\epsilon^d_i \sim \mathcal{N}(0, 1)$ yields qualitatively similar results.

type of DGP's that occur in applied research, isolating the efficiency gain due to finely stratified assignment, finely stratified sampling, as well as finely stratified implementation of the optimal propensities from Section 4.

	Design, Paper	A.	Ban.	Bay.	C.	De.	Do.	F.	H.	L.
SD	CR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	CR, Loc	0.81	0.48	0.51	0.88	0.55	0.93	0.79	0.73	0.87
	Loc	0.75	0.45	0.39	0.84	0.47	0.88	0.74	0.70	0.76
	Hom.	0.72	0.46	0.44	0.79	0.58	0.86	0.85	0.68	0.90
	Pilot S	0.72	0.45	0.55	0.81	0.57	1.01	0.75	0.71	0.75
	Pilot L	0.70	0.42	0.52	0.81	0.52	0.87	0.65	0.68	0.71
%ΔCI	CR	0	0	0	0	0	0	0	0	0
	CR, Loc	-11	-49	-41	-6	-36	-4	-14	-10	-11
	Loc	-12	-48	-38	-5	-33	-4	-14	-15	-11
	Hom.	-21	-44	-47	-6	-40	-8	-8	-16	-10
	Pilot S	-19	-47	-23	-5	-23	3	-18	-15	-21
	Pilot L	-21	-52	-31	-5	-30	-9	-27	-18	-24
Cover	CR	0.95	0.95	0.95	0.94	0.96	0.95	0.94	0.94	0.96
	CR, Loc	0.96	0.96	0.99	0.96	0.98	0.96	0.96	0.98	0.96
	Loc	0.98	0.97	1.00	0.96	1.00	0.97	0.97	0.97	0.98
	Hom.	0.97	0.98	0.98	0.97	0.97	0.96	0.96	0.98	0.95
	Pilot S	0.97	0.98	0.99	0.96	0.99	0.95	0.97	0.98	0.97
	Pilot L	0.97	0.97	0.99	0.97	0.99	0.96	0.97	0.97	0.96
	n	1451	903	550	91	446	1000	1903	116	770
	$\dim(\psi)$	8	6	3	3	3	4	6	4	3

Table 2: Empirical Results

Descriptions of each paper, treatment, outcome, our choice of stratification variables ψ, level of aggregation, and other design parameters are provided in Section 10.2 in the appendix. Experiment sizes n are as in the original papers, ranging from $n = 91$ for Casey et al. (2021) to $n = 1903$ for Finkelstein et al. (2012). The one exception is Domurat et al. (2021) ($N_0 = 87394$), for which we set $n = 1000$ for Monte Carlo tractability. Baseline treatment proportions are set to $p = 1/2$, except for Lowe (2021) who used $p = 2/3$ and Finkelstein et al. (2012) with $p = 1/3$. We use the large experiment version of our method with folds of size 200 (Section 2.1) and implement sampling subordinate assignment (Remark 3.14).

Costs and Discretization - Our theory shows that efficiency can be improved by (1) sourcing a large pool of units willing to participate in the experiment and (2) using fine stratification to choose a representative experimental subsample from this pool. If the marginal cost of including a unit is zero, step (2) is trivial, and we just take as many units as possible. The marginal cost $C(\psi)$ of sampling a unit of type ψ is not reported in the papers in our sample. To understand the potential marginal variance reduction from (1) representative sampling and (2) using the optimal propensity $q^*(\psi)$ in the types of DGP's that occur in applied research, we set $C(\psi) = \mathbb{1}(|\psi|_2 \leq \kappa) + 5\mathbb{1}(|\psi|_2 > \kappa)$ with

$\kappa = \text{Median}_{i=1}^n |\psi_i|_2$ and $\bar{B} = 1.5$. This results in feasible constant sampling proportions $q \approx 0.7$. For example, if ψ_i were village location relative to an urban center, this would correspond to higher cost of collecting data in rural villages. As in Section 7, we discretize $q^*_{hom}(\psi)$ and $q^*(\psi)$ by choosing $q_k(\cdot)$ to minimize discretization error $E_n[(q_k - q^*)^2(\psi_i)]$ over the set $\{q : q(\psi) \in a/10 : a = 1, \ldots, 10\}$ subject to a constraint on $L_n = |\text{Image}(q_k)|$, with $L_n \le 2$ for $n < 500$, $L_n \le 3$ for $500 \le n < 1000$ and $L_n \le 4$ for $1000 \le n \le 2000$.

Results. Our main results are presented in Table 2. Papers are listed by initials of the first author, and all designs are as in Section 7. The largest change standard deviation (SD) change is in the contrast between complete randomization **CR** and finely stratified assignment **CR, Loc**, with an average of -27% across the papers in our sample. The improvement is particulary striking in papers like Baysan (2022) and Banerjee et al. (2021) with highly predictive baseline covariates. The average marginal change in SD due to finely stratified sampling (from **CR, Loc** to **Loc**) is smaller at -6%. Using the optimal sampling proportions $q^*_{hom}(\psi)$, which assume homoskedasticity, reduces the variance for some studies but increases it for others, resulting in $+4\%$ change on average. This is not surprising considering that many of these studies have considerable heteroskedasticity. The change in SD between **Loc** and **Pilot S** is $+4\%$ on average, while for the case with a large pilot the change in SD from **Loc** to **Pilot L** is -5% on average. This shows that with a large pilot, closely related previous experiment, or observational data from the same population, the feasible estimates of the optimal designs in Section 4 can be used to increase efficiency. However, our empirical results suggest this may not be appropriate when only a small pilot study is available.

Next we discuss the performance of our inference methods (Section 6). Coverage is close to nominal, but somewhat conservative in finite samples. This is due to two different forces. First, match quality between groups g and $\nu(g)$ in the collapsed-strata variance estimators $\widehat{v}_1$ and $\widehat{v}_0$ is worse than match quality within groups, which results in $\widehat{v}_1$ and $\widehat{v}_0$ in Section 6 being conservative. This effect is most severe for designs with highly predictive covariates. Second, match quality at the sampling stage $T_{1:n} \sim \text{Loc}(\psi, q(\psi))$ is generally better than match quality at the assignment stage, since more units are available during sampling. However, our variance estimators can only only use the "thinned out" outcome data available for the units $T_i = 1$ included in the experiment, which underestimates match quality during sampling. This effect will be most severe for small sampling proportions $q \to 0$ and in DGP's with significant treatment effect heterogeneity.

The change in confidence interval length $\%\Delta$CI is slightly conservative but broadly reflects the efficiency gains in the first panel. This shows that our inference methods are able to take advantage of the reduction in variance from both finely stratified sampling and assignment, as well as designs with varying sampling proportions.

9 Recommendations for Practice

Our empirical results show robust variance reductions from fine stratification at both the sampling and assignment stages. When choosing stratification variables ψ, we recommend including baseline outcomes, if available, and a small set of other variables suspected to be predictive of outcomes and treatment effect heterogeneity. In particular, if experimenters have pre-registered measuring treatment effect heterogeneity with respect to a certain variable, then it is natural to include this variable in ψ. Fine stratification methods increase the value of collecting baseline survey data, insofar as extra investment in the baseline survey process allows us to measure variables expected to be most predictive of outcomes and treatment effect heterogeneity. Our theory in Section 3 showed that the efficiency gains from stratified sampling are larger the more eligible units we have, since this helps us build a more representative experimental sample. Because of this, sourcing a large pool of candidate units for the experiment can improve precision, even if the experimental budget constraints do not allow all of these units to ultimately participate.

The feasible sampling design $q^*_{hom}(\psi)$ (assuming homoskedasticity) reduced variance in our simulations, but had mixed effects in the empirical application. Relative to the simpler **Loc** design with constant propensity q, the $q^*_{hom}(\psi)$ design reduced variance for some DGP's but increased it for others. Aside from potential misspecification, there is a finite sample tradeoff between (B) better optimization of the residual variance by using varying $q(\psi)$ and (W) worse sampling and assignment matches due to having many different $q(\psi)$ strata. For experiments with highly predictive covariates, effect (W) may dominate, so that a design with constant $q(\psi) = q$ may be preferable. However, if costs are very heterogeneous, then the residual variance effect (B) will dominate, and $q^*_{hom}(\psi)$ can produce significant efficiency gains. The estimated optimal designs from Section 5 performed well in our empirical application for $n_{pilot} = 400$, but were generally too noisy for $n_{pilot} = 100$. In the absence of a very large pilot or related previous experiment, one alternative is to use observational data to calibrate the ex-ante variance function $\bar{\sigma}^2(\psi)$ appearing in the optimal design $q^*(\psi)$. This could improve on the design $q^*_{hom}(\psi)$, which unrealistically assumes perfect homoskedasticity, without the added noise associated with estimating the ex-ante variance $\bar{\sigma}^2(\psi)$ from a small pilot.

Finally, the inference methods in Section 6 were slightly conservative in finite samples, but still allow researchers to report smaller confidence intervals that reflect the efficiency gains from finely stratified sampling and assignment.

References

Abadie, A. and G. W. Imbens (2008). Estimation of the conditional variance in paired experiments. *Annales d'Economie et de Statistique*, 175–187.

Abaluck, J., L. H. Kwong, A. Styczynski, A. Haque, A. Kabir, E. Bates-Jeffries, E. Crawford, J. Benjamin-Chung, S. Raihan, S. Rahman, S. Benhachmi, N. Zaman, P. J. Winch, M. Hossain, H. Mahmud Reza, A. All Jaber, S. G. Momen, F. L. Bani, A. Rahman, T. S. Huq, S. P. Luby, and A. M. Mobarak (2021). The impact of community masking on covid-19: A cluster-randomized trial in bangladesh. Working Paper.

Abebe, G., S. A. Caria, and E. Ortiz-Ospina (2021). The selection of talent: Experimental and structural evidence from ethiopia. *American Economic Review*.

Armstrong, T. (2022). Asymptotic efficiency bounds for a class of experimental designs.

Armstrong, T. and M. Kolesár (2021). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica*.

Athey, S. and G. W. Imbens (2017). The econometrics of randomized experiments. *Handbook of Economic Field Experiments*.

Bai, Y. (2022). Optimality of matched-pair designs in randomized controlled trials. *American Economic Review*.

Bai, Y., J. P. Romano, and A. M. Shaikh (2021). Inference in experiments with matched pairs. *Journal of the American Statistical Association*.

Bai, Y., M. Tabord-Meehan, and J. Liu (2023). Inference for matched tuples and fully blocked factorial designs.

Banerjee, A., A. G. Chandrasekhar, S. Dalpath, E. Duflo, J. Floretta, M. O. Jackson, H. Kannan, F. N. Loza, A. Sankar, A. Schrimpf, and M. Shrestha (2021). Selecting the most effective nudge: Evidence from a large-scale experiment on immunization. *NBER Working Paper*.

Baysan, C. (2022). Persistent polarizing effects of persuasion: Experimental evidence from turkey. *American Economic Review*.

Billingsley, P. (1995). *Probability and Measure*. Wiley.

Breza, E., F. C. Stanford, M. Alsan, B. Alsan, A. Banerjee, A. G. Chandrasekhar, S. Eichmeyer, T. Glushko, P. Goldsmith-Pinkham, K. Holland, E. Hoppe, M. Karnani, S. Liegl, T. Loisel, L. Ogbu-Nwobodo, B. A. Olken, C. Torres, P.-L. Vautrey, E. Warner, S. Wootton, and E. Duflo (2021). Doctors' and nurses' social media ads reduced holiday travel and covid-19 infections: A cluster randomized controlled trial in 13 states. Working Paper.

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

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2019). Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics*.

Cai, Y. and A. Rafi (2023). On the performance of the neyman allocation with small pilots.

Casey, K., A. Bakarr Kamara, and N. F. Meriggi (2021). An experiment in candidate selection. *American Economic Review*.

Chernozhukov, V., D. Chetverikov, E. Duflo, C. Hansen, and W. Newey (2017). Double / debiased / neyman machine learning of treatment effects. *American Economics Review Papers and Proceedings 107*(5), 261–265.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernández-Val (2023). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Working Paper.

Cochran, W. G. (1977). *Sampling Techniques* (3 ed.). John Wiley and Sons.

de Chaisemartin, C. and J. Ramirez-Cuellar (2021). At what level should one cluster standard errors in paired and small-strata experiments?

Deeb, A. and C. de Chaisemartin (2022). Clustering and external validity in randomized controlled trials.

Dellavigna, S., J. A. List, U. Malmendier, and G. Rao (2022). Estimating social preferences and gift exchange at work. *American Economic Review*.

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

Domurat, R., I. Menashe, and W. Yin (2021). The role of behavioral frictions in health insurance marketplace enrollment and risk: Evidence from a field experimen. *American Economic Review*.

Fan, J. and Q. Yao (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika 85*(3), 645–660.

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, and K. Baicker (2012). The oregon health insurance experiment: Evidence from the first year. *QJE*, 1057–1106.

Fogarty, C. B. (2018). On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society: Series B 80*(5).

Folland, G. B. (1999). *Real Analysis: Modern Techniques and Their Applications*. Wiley.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*.

Hahn, J., K. Hirano, and D. Karlan (2011). Adaptive experimental design using the propensity score. *Journal of Business and Economic Statistics 29*(1), 96–108.

Hansen, M. H., W. N. Hurwitz, and W. G. Madow (1953). *Sample Survey Methods and Theory*. Wiley.

Harshaw, C., F. Sävje, D. A. Spielman, and P. Zhang (2021). Balancing covariates in randomized experiments with the gram-schmidt walk design. Working Paper.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4).

Hussam, R., E. M. Kelley, G. Lane, and F. Zahra (2022). The psychosocial value of employment: Evidence from a refugee camp. *American Economic Review*.

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

Kallus, N. (2017). Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B*.

Karmakar, B. (2022). An approximation algorithm for blocking of an experimental design. *Journal of the Royal Statistical Society: Series B*.

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis*, 1–15.

Kasy, M. and A. Sautmann (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica 89*(1).

Krieger, A., D. Azriel, and A. Kapelner (2019). Nearly random designs with greatly improved balance. *Biometrika*.

Li, X. and P. Ding (2018). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*.

Li, X., P. Ding, and D. B. Rubin (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics 7*(1), 295–318.

Lohr, S. L. (2021). *Sampling Design and Analysis*. Routledge.

Lowe, M. (2021). Types of contact: A field experiment on collaborative and adversarial caste integration. *American Economic Review*.

Rendl, F., G. Rinaldi, and A. Wiegele (2008). Solving max-cut to optimality by intersecting semidefinite and polyhedral relaxations. *Mathematical Programming*.

Robins, J. M. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association 90*, 122–129.

Rosenberger, W. F. and J. M. Lachin (2016). *Randomization in Clinical Trials*. Wiley.

Tabord-Meehan, M. (2022). Stratification trees for adaptive randomisation in randomised controlled trials. *The Review of Economic Studies*.

Wang, X., T. Wang, and H. Liu (2021). Rerandomization in stratified randomized experiments. *Journal of the American Statistical Association*. Working Paper.

Yang, Z., T. Qu, and X. Li (2021). Rejective sampling, rerandomization, and regression adjustment in survey experiments. *Journal of the American Statistical Association*.

10 Appendix

10.1 Finite Population Estimands

If the eligible units $\{1, \dots, n\}$ comprise the entire population of interest, then we may be interested in estimation and inference on the sample average treatment effect SATE $= E_n[Y_i(1) - Y_i(0)]$, or the average conditional treatment effect $E_n[c(X_i)]$, as in Armstrong and Kolesár (2021). Note that both estimands are defined over the full population of eligible units, not just the smaller set of experiment participants $\{i : T_i = 1\}$. For the SATE, suppose $T_{1:n} \sim \text{Loc}(X, q)$ and $D_{1:n} \sim \text{Loc}(X, p)$ and define the residual treatment effect variance $\sigma^2_\tau(X) = \text{Var}(Y(1) - Y(0)|X)$. Under the same conditions as Theorem 3.2, we have $\sqrt{n_T}(\widehat{\theta} - \text{SATE}) \Rightarrow \mathcal{N}(0, V_{\text{SATE}})$ with

$$V_{\text{SATE}} = E\left[\frac{\sigma^2_1(X)}{p} + \frac{\sigma^2_0(X)}{1-p} - q\sigma^2_\tau(X)\right]. \tag{10.1}$$

The final variance component $E[\sigma^2_\tau(X)]$ is not identified, as in the case of SATE estimation under complete randomization. Setting $X = 1$ and $q = 1$ recovers the classical results in that setting. Observe that V_{SATE} decreases as the residual treatment effect heterogeneity $\sigma^2_\tau(X)$ increases. The negative sign reflects a competition between two opposing forces. To see this, let $\bar{Y}_i = Y_i(1)/p + Y_i(0)/(1-p)$ and consider the error decomposition

$$\widehat{\theta} - \text{SATE} = \text{Cov}_n(T_i, Y_i(1) - Y_i(0))/q + \text{Cov}_n(D_i, \bar{Y}_i|T_i = 1) \tag{10.2}$$

The first term parameterizes the errors due to correlation between the sampling variables and treatment effects, and naturally increases with $\sigma^2_\tau(X) = \text{Var}(Y(1) - Y(0)|X)$. The second term is increasing in $\text{Cov}(Y(1), Y(0)|X)$ (decreasing in $\sigma^2_\tau(X)$) and larger in mean square, resulting in a net negative dependence on $\sigma^2_\tau(X)$. As $q \to 0$, the relative variance[19] due to sampling increases, exactly cancelling the $E[\sigma^2_\tau(X)]$ factor due to assignment in the limit. Conservative inference for the SATE under stratified sampling and assignment can be based on either of the lower bounds $\sigma^2_\tau(X) \geq \sigma^2_1(X) + \sigma^2_0(X) - 2\sigma_1(X)\sigma_0(X) \geq 0$.

The average conditional treatment effect ACTE $= E_n[c(X_i)]$ is more difficult to motivate from a policy perspective. One potential application is a structural model of the form $Y_{it}(1) - Y_{it}(0) = c(X_i) + \epsilon_{it}$, with systematic component $c(X_i)$ mediated solely through observables X and transitory shock component ϵ_{it}.[20] Intuitively, in this model the ACTE acts like a "denoised" version of the SATE. If the ϵ_{it} are uncorrelated, the ACTE estimated at time t is the best predictor of the SATE for a policy implemented at time $t + 1$. The proof of Theorem 3.2 shows that $\sqrt{n_T}(\widehat{\theta} - E_n[c(X_i)]) \Rightarrow \mathcal{N}(0, V_c)$ with identified variance $V_c = E[\sigma^2_1(X)/p + \sigma^2_0(X)/(1-p)] \geq V_{\text{SATE}}$.

Remark 10.1 (Design-based Theory)**.** It is interesting to compare our results with those obtained under a finite population "design-based" framework. The design based approach in statistics takes the sequence of finite populations $W_{1:n} = (X_i, Y_i(0), Y_i(1))_{i=1}^n$ as non-random, with all randomness due to treatment assignment. For example, see Li and Ding

[19] The normalization $\sqrt{n_T}(\widehat{\theta} - \text{SATE})$ by experiment size $n_T = \sum_i T_i$ holds the number of sampled units fixed.

[20] Inference on a more general denoised SATE parameter in a model with transitory shocks is studied in Deeb and de Chaisemartin (2022).

(2018) or Wang et al. (2021). By contrast, here we impose a nonparametric generative model $W_{1:n} \sim F$, requiring F to have certain finite moments in Assumption 3.1. However, our analysis shows that $\sqrt{n_T}(\widehat{\theta} - \text{SATE})|W_{1:n} \Rightarrow \mathcal{N}(0, V_{\text{SATE}})$ conditional on the data $W_{1:n}$. Since we condition on $W_{1:n}$, the generative model $W_{1:n} \sim F$ does not contribute to estimator variance. There is no hypothetical superpopulation, and all randomness is due to the sampling and assignment variables $T_{1:n}$ and $D_{1:n}$, just as in design-based theory. The distribution F imposes homogeneity on the sequence of finite populations $W_{1:n}$ as $n \to \infty$, ensuring that quantities such as $E_n[W_i]$ converge to well-defined limits. By contrast, in the finite population framework the convergence $E_n[W_i] \to W_\infty$ is assumed ex-post as a high-level condition on the sequence of populations $W_{1:n}$.

The finite population framework appears to be formally more general than our model, but also potentially more cumbersome. For example, the outcome function $E[Y(d)|X]$, which plays an important conceptual role in many parts of causal inference, does not appear to have a nonparametric design-based analogue. It would be interesting to characterize when this extra generality has "bite" and leads to substantively different prescriptions for estimation and inference than the SATE theory presented here.

10.2 Empirical Application Details

This section provides descriptions of each paper and implementation details for our empirical application in Section 8. In all cases, we normalize the stratification variables so that $\text{Var}(\psi_{ij}) = 1$ for $j = 1, \dots, \dim(\psi)$.

(1) Abebe et al. (2021) estimates the effect of an application incentive on the ability of applicants for clerical employment in Ethiopia. We let Y be the authors' index of cognitive ability, D be the application incentive, and $\psi(X)$ be gender, age, work experience in years, self-reported gpa and previous wage, and indicators for being born in Addis Ababa, speaking Amharic, and studying engineering.

(2) Banerjee et al. (2021) estimates the effect of various strategies to promote child-vaccination on the number of children completing the full vaccination sequence. We let Y_i be the number of children receiving the measles shot over the full trial period in village i and D_i whether the village received the SMS reminder intervention. We let ψ include village population, the proportion of individuals in the baseline survey of that village who were in a "scheduled caste", a "backward class", who received nursery education or less, the proportion of vaccine completions among older children, and proportion who had a vaccine card.

(3) Baysan (2022) estimates the effect of political information campaigns about concentration of executive power in Turkey on voter polarization. Data is at the ballot box level, while assignment to information campaigns is at the neighborhood level. We let Y_i be the vote share of "No" votes, averaged over ballot boxes in a neighborhood, in the 2017 referendum. We let D be whether the village was exposed to any information campaign, and ψ include the village-level average vote share for the CHP in the 2015 election, a measure of turnout, and the number of ballots collected.

(4) Casey et al. (2021) estimates the effect of increased information given to political parties about voter preferences over candidates on whether the most voter-preferred

candidate was selected to run by the party. We let D be assignment to the treatment package at the party-constituency level, Y whether the most preferred candidate was selected, and let ψ include competitiveness of the race, candidate professional qualifications index, and candidate public service motivation index.

(5) Dellavigna et al. (2022) estimates the effect of employer gifts and other "social preference" related interventions on worker productivity. We study the first experiment in the paper and let $D = 1$ if the worker received either a positive gift or an in-kind gift and $D = 0$ for no gift or a negative gift. We let Y be worker output in the last working period. We let ψ be total productivity during the first 8 periods of the trial (excluding the final gift period), gender, and age.

(6) Domurat et al. (2021) estimates the effect of informational interventions about Covered California insurance policies on takeup of insurance. We let $D = 0$ for the control group and $D = 1$ if assigned to any of the letter campaigns in arms 3, 4, or 5. Y is an indicator for insurance takeup. ψ includes a measure of household income, mean age, household subsidy size, and an indicator for being Latino.

(7) Finkelstein et al. (2012) reports the effect of winning the 2008 Oregon Medicaid lottery on various health and public service utilization outcomes. We use data from wave one, restricting to single person households. We estimate an ITT effect with Y the number of emergency department (ED) visits in the post-period and D an indicator for winning the lottery. We let ψ include gender, age, any visits to the ED in the pre-period, number of visits in the pre-period, total SNAP benefits in the pre-period, and indicators for ever being on SNAP or having a chronic condition.

(8) Hussam et al. (2022) estimates the value of employment on measures of psychosocial wellbeing. Treatment assignment is at the block level in the refugee camp, with a sample of five individuals chosen in each block. We aggregate by taking the mean of outcomes and covariates in each block. We let $D = 0$ if the block was randomized to cash only and $D = 1$ if it was randomized to employment. Y is the endline mental health index, and ψ includes the baseline mental health index, average gender (in $[0, 1]$), proportion who had a family member killed, and a measure of sociability.

(9) Lowe (2021) estimates the effect of collaborative and adversarial intergroup contact on cross-caste friendships using randomization to different teams in a cricket league in India. We let Y be number of other caste friends at endline, D be whether the person is assigned to a mixed caste team, and ψ include number of other caste friends at baseline, age, and a measure of cricket ability.

10.3 Remarks

Remark 10.2 (Increasing Stratification Condition)**.** At a high level, the condition $\psi_1 \subseteq \psi_2$ allows us to ignore the complicated effect of first-stage sampling on the joint distribution of sampled stratification variables $(\psi_{1,i})_{i:T_i=1}$. To see the problem, observe that if $q = 1/k$ then $T_i = T_j = 1$ implies that i, j cannot have been matched together during sampling. Then, for instance, we expect

$$P(|\psi_{1,i} - \psi_{1,j}| < \epsilon) > P(|\psi_{1,i} - \psi_{1,j}| < \epsilon \,|\, T_i = T_j = 1). \tag{10.3}$$

If $E[Y_i(d)|\psi_{1,i}] \neq 0$, such changes to the joint distribution will show up in the conditional variance $\mathrm{Var}(\widehat{\theta}|X_{1:n}, T_{1:n})$ in complicated ways that depend on the details of the matching

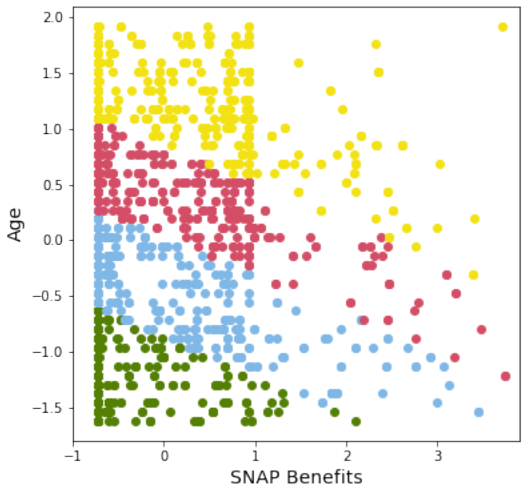


Figure 4: PCA Folds

algorithm. However, we can use the fact that the assignment stratification "partials out" $Y_i(d)$ so that, to first order, only the residuals $u_i = Y_i - E[Y_i(d)|\psi_{2,i}]$ enter this conditional variance. If $\psi_1 \subseteq \psi_2$ the residuals u_i are less affected by selection on $\psi_{1,i}$. For example, we can show $E[u_i u_j | T_i = T_j = 1] = E[u_i u_j] = 0$. We conjecture the theorem may be true without this condition if the effects in Equation 10.3 are lower order, but leave the detailed study of this issue for specific matching procedures to future work.

10.4 Heteroskedasticity Function Estimation

The theory in Section 5 requires heteroskedasticity function estimates $\widehat{\sigma}_d^2(\psi)$. Our simulations and empirical application using a modification of Fan and Yao (1998). In a setting without outcomes $Y = m(X) + \sigma^2(X)\epsilon$, they propose (1) use local linear regression to estimate $m(X)$ and (2) use local linear regression to project estimated residuals $(Y - \widehat{m}(X))^2$ on X. In our setting, we let $\tilde{Y}_i(1) = Y_i D_i T_i / p_i q_i$ and (1) project $\tilde{Y}_i(1) \sim \psi_i$ to estimate $m_1(\psi_i)$. Next, we (2) project $\tilde{\epsilon}_i(1)^2 = (Y_i - \widehat{m}(\psi_i))^2 D_i T_i / p_i q_i \sim \psi_i$ to estimate $\widehat{\sigma}_1^2(\psi_i)$, and similarly for $d = 0$. We tested linear ridge regression, RBF-kernel ridge regression, and random forests for each regression step, with hyperparameters chosen by cross-validation in all cases. Kernel ridge estimated $\sigma_d^2(\psi)$ the most precisely in dimensions $\dim(\psi) = 1, 2$, while forests were superior in higher dimensions. Our simulation and empirical results are presented using random forest regression. Pilot data is drawn from a stratified experiment of the specified size with $p = 1/2$.

10.5 Details of Matching Algorithm

Write $n = lk + \delta$ for some integer l and $0 \leq \delta < k$. First, match a remainder group of size δ and set it aside. Then suppose without loss that $n = lk$. Let J be the minimal positive integer such that $2^J \geq k$ and let $\sum_{j=0}^{J-1} a_j 2^j$ with $a_j \in \{0, 1\}$ be the binary representation of $2^J - k \geq 0$. Before the jth call of Derigs' algorithm, add l singleton groups of fake units $g = \{F_j\}$ of type j to the dataset if and only if $a_j = 1$. Let $R(g)$ denote the real units in a group $R(g) \subseteq [n]$ and $F(g)$ the fake units so that $g = R(g) \cup F(g)$. Before the jth call to Derig's algorithm, set $d(g, g') = +\infty$ if either of the following occur: (1) $F(g) \cap F(g') \neq \emptyset$ or (2) $|F(g) \cup F(g')| > 0$ but $|F(g) \cup F(g')| \neq \sum_{i=0}^{j} a_i$. Otherwise, set the distance $d(g, g') = |\bar{\psi}_{R(g)} - \bar{\psi}_{R(g')}|_2^2$. Compute the optimal pairing at each step. After J steps, remove all the fake units by setting $g = R(g)$. The binary representation trick is inspired by Karmakar (2022), though the algorithm in his paper does not seem to guarantee groups of the correct cardinality for larger k.

Figure 4 shows PCA Folds for the "large experiment" version of our algorithm. We use the data from Finkelstein et al. (2012), with $K = 4$ folds and $n = 1903$ samples.

11 Proofs

11.1 Matching

Theorem 11.1 (Matching)**.** *Consider triangular arrays $(\psi_{i,n})_{i=1}^n \subseteq \mathbb{R}^d$ and $(p_{i,n})_{i=1}^n \subseteq \mathbb{Q}$ with levels $p_{i,n} \in L_n = \{a_l/k_l\}$ and $\overline{k}_n = \max\{k_l : a_l/k_l \in L_n\}$. For any sequence of subsets $S_n \subseteq [n]$, there exists a sequence of partitions $(\mathcal{G}_n)_{n\geq 1}$ of S_n such that $\mathcal{G}_n = \cup_l \mathcal{G}_{nl}$ and $\mathcal{G}_{nl}$ partitions $S_n \cap \{i : p_{i,n} = a_l/k_l\}$ with $|g| = k_l$ for all but one non-empty group $g \in \mathcal{G}_{nl}$. The partition $\mathcal{G}_n$ satisfies the homogeneity rate*

$$n^{-1} \sum_{g\in\mathcal{G}_n} \frac{1}{|g|} \sum_{i,j\in g} |\psi_{i,n} - \psi_{j,n}|_2^2 \leq (1 \vee \max_{i=1}^n |\psi_{i,n}|_2^2) \cdot O((n/\overline{k}_n|L_n|)^{-2/(d+1)}). \tag{11.1}$$

Let $\mathcal{G}_{nl}^*$ be the optimal partition of $S_n \cap \{i : p_{i,n} = a_l/k_l\}$, solving the minimization

$$\mathcal{G}_{nl}^* = \operatorname*{argmin}_{\mathcal{G}_{nl}} \left[n^{-1} \sum_{g\in\mathcal{G}_{nl}} \sum_{i,j\in g} |\psi_{i,n} - \psi_{j,n}|_2^2 \right]$$

subject to the group size constraint $|g| = k_l$. Then $\mathcal{G}_n^* = \cup_l \mathcal{G}_{nl}^*$ satisfies the homogeneity rate in Equation 11.1.

Proof. First, note that for any partition $\mathcal{G}_n$

$$\frac{1}{n} \sum_{g\in\mathcal{G}_n} \frac{1}{|g|} \sum_{i,j\in g} |\psi_{i,n} - \psi_{j,n}|_2^2 \leq (1 \vee \max_{i=1}^n |\psi_{i,n}|_2^2) \cdot \frac{1}{n} \sum_{g\in\mathcal{G}_n} \frac{1}{|g|} \sum_{i,j\in g} \left| \frac{\psi_{i,n} - \psi_{j,n}}{1 \vee \max_{i=1}^n |\psi_{i,n}|_2} \right|_2^2$$

Clearly, $|\psi_{i,n}/(1 \vee \max_{i=1}^n |\psi_{i,n}|_2)|_2 \leq 1$ for all $1 \leq i \leq n$. Then by recentering, it suffices to show the claim for $\psi_{i,n} \in [0,1]^d$ for all $1 \leq i \leq n$ as $n \to \infty$. Consider blocks of the form $B = \{\sum_{k=1}^d x_k e_k : x_k \in [s_k/m, (s_k+1)/m]\}$ for indices $\{s_k\}_{k=1}^d \subseteq \{0, \ldots, m-1\}$. Fix an ordering of these blocks $(B_l)_{l=1}^{m^d}$ such that B_l and B_{l+1} are adjacent for all l. Intuitively, this forms a "block path" through $[0,1]^d$, see Bai et al. (2021) for a picture. Define $l(i) = \min_{l=1}^{m^d}\{l : \psi_{i,n} \in B_l\}$.

Algorithm - Fix $p_a \in L_n$ and form groups $(g_{a,s})_{s=1}^n$ of units $g_{a,s} \subseteq \{i : p_{i,n} = p_a = f_a/k_a\}$ by induction as follows. (1) Form groups of size k_a arbitrarily among $\{i : l(i) = 1$, and $p_n(X_i, \xi_n) = p_a$, possibly using external randomness π_n. This process results in at most one partially completed group with $|g_{a,s'}| < k_a$. While $|g_{a,s'}| < k_a$, do the following: (2) increment $l \to l+1$ and (3) add an unmatched unit i with $l(i) = l+1$ to this group.

If $|g_{a,s'}| = k_a$, stop. If $|g_{a,s'}| < k_a$ and there are unmatched units in B_{l+1}, goto (3). If $|g_{a,s'}| < k_a$ and there are no unmatched units in B_{l+1}, increment l and go to (1). Suppose that $g_{a,s'}$ is completed with a unit from $B_{l'}$. Then repeat the process above starting with the next group $g_{a,s'+1}$ and the units in block $B_{l'}$. Since there are $n < \infty$ units, this process terminates. Repeat this for each $a = 1, \dots, |L_n|$. By construction, this creates groups $\mathcal{G}_n = \{g_{a,s} : 1 \le a \le |L_n|, 1 \le s \le n\}$ with the ordering property

$$l(i) \le l(j) \quad \forall i \in g_{a,s}, j \in g_{a,s'} \quad s < s' \quad a = 1, \dots, |L_n| \tag{11.2}$$

Fix an indexing of all within-group pairs $(\mathfrak{p}_{a,s,t})_{t=1}^{k_a^2-k_a} \equiv \{(i,j) : i \ne j;\ i,j \in g_{a,s}\}$, and denote $\mathfrak{p}_{a,s,t} = (i_{a,s,t}, j_{a,s,t})$. Define $E_{a,s,t} = \{l(i_{a,s,t}) = l(j_{a,s,t})\}$, the event that a pair is in the same element of the block partition. With this notation, we have

$$\begin{aligned} n^{-1}\sum_{g\in\mathcal{G}_n}\frac{1}{|g|}\sum_{i,j\in g}|\psi_{i,n}-\psi_{j,n}|_2^2 &= n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}k_a^{-1}\sum_{i,j\in g_{a,s}}|\psi_{i,n}-\psi_{j,n}|_2^2 \\ &\le n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}k_a^{-1}\sum_{t=1}^{k_a^2-k_a}|\psi_{i_{a,s,t},n}-\psi_{j_{a,s,t},n}|_2^2\mathbb{1}(g_{a,s}\ne\emptyset) \end{aligned}$$

(1) Suppose $E_{a,s,t}$ occurs. Then $d_{a,s,t} \equiv |\psi_{i_{a,s,t},n} - \psi_{j_{a,s,t},n}|_2 \le \max_{l=1}^{m^d} \operatorname{diam}(B_l, |\cdot|_2) \le \sqrt{d}/m$ on this event. Then we may bound

$$\begin{aligned} n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}k_a^{-1}\sum_{t=1}^{k_a^2-k_a}d_{a,s,t}^2\mathbb{1}(E_{a,s,t})\mathbb{1}(g_{a,s}\ne\emptyset) &\le \frac{d}{nm^2}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}k_a^{-1}\sum_{t=1}^{k_a^2-k_a}\mathbb{1}(g_{a,s}\ne\emptyset) \\ \le \frac{d}{nm^2}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}k_a\mathbb{1}(g_{a,s}\ne\emptyset) \le \frac{d}{m^2} \end{aligned}$$

The final inequality since the double sum exactly counts the number of units in the sample (by group) $\sum_{a=1}^{|L_n|}\sum_{s=1}^{n} k_a \mathbb{1}(g_{a,s} \ne \emptyset) = |S_n| \le n$.

(2) Now consider the terms where $E_{a,s,t}$ does not occur. Fix any such pair $(i_{a,s,t}, j_{a,s,t})$. Without loss, suppose the block membership $l(i_{a,s,t}) < l(j_{a,s,t})$. For $l(i_{a,s,t}) \le l \le l(j_{a,s,t})$, define a sequence z_l as follows. $z_{l(i_{a,s,t})} = \psi_{i_{a,s,t},n}$, $z_{l(j_{a,s,t})} = \psi_{j_{a,s,t},n}$ and $z_l \in B_l$ chosen arbitrarily otherwise. Note that for $x \in B_l$ and $y \in B_{l+1}$, by construction of the contiguous blocks $|x-y|_2 \le 2\sqrt{d}/m$. Then by telescoping and triangle inequality, on the event $E_{a,s,t}^c$

$$d_{a,s,t} = |\psi_{i_{a,s,t},n} - \psi_{j_{a,s,t},n}|_2 \le \sum_{l=l(i_{a,s,t})}^{l(j_{a,s,t})-1} |z_{l+1} - z_l|_2 \le \frac{2\sqrt{d}}{m} \cdot [l(j_{a,s,t}) - l(i_{a,s,t})]$$

Note also that if $x, y \in [0,1]^d$ then $|x-y|_2^2 = d(|x-y|_2/\sqrt{d})^2 \le \sqrt{d}|x-y|_2$, using $c^2 \le c$

for $0 \leq c \leq 1$. In particular, we have $d_{a,s,t}^2 \leq \sqrt{d} \cdot d_{a,s,t}$.

$$
\begin{aligned}
&n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a^{-1} \sum_{t=1}^{k_a^2-k_a} d_{a,s,t}^2 \mathbb{1}(E_{a,s,t}^c) \leq \sqrt{d} n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a^{-1} \sum_{t=1}^{k_a^2-k_a} d_{a,s,t} \mathbb{1}(E_{a,s,t}^c) \\
&\sum_{a=1}^{|L_n|} \frac{2d}{mnk_a} \sum_{t=1}^{k_a^2-k_a} \sum_{s=1}^{n} [l(i_{a,s,t}) - l(j_{a,s,t})] \mathbb{1}(E_{a,s,t}^c) \leq \sum_{a=1}^{|L_n|} \frac{2\sqrt{d}}{mnk_a} \sum_{t=1}^{k_a^2-k_a} m^d \\
&\leq \sum_{a=1}^{|L_n|} \frac{2\sqrt{d} m^d}{mn} k_a \leq 2\sqrt{d} |L_n| \overline{k}_n n^{-1} m^{d-1}
\end{aligned}
$$

The second inequality follows by the ordering property in equation 11.2 above, since for each $t = 1, \ldots, k_a^2 - k_a$, the intervals $([l(i_{a,s,t}), l(j_{a,s,t})])_{s=1}^n$ are non-overlapping, and there are at most m^d blocks. Summarizing the above work, we have shown that

$$
\begin{aligned}
n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a^{-1} \sum_{i,j \in g_{a,s}} |\psi_{i,n} - \psi_{j,n}|_2^2 &\leq n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a^{-1} \sum_{t=1}^{k_a^2-k_a} d_{a,s,t}^2 \mathbb{1}(E_{a,s,t}) + d_{a,s,t}^2 \mathbb{1}(E_{a,s,t}^c) \\
&\leq \frac{d}{m^2} + \frac{2\sqrt{d} |L_n| \overline{k}_n m^{d-1}}{n}
\end{aligned}
$$

Setting $m \asymp \left(n/(|L_n|\overline{k}_n)\right)^{1/(d+1)}$ gives the rate

$$
n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a^{-1} \sum_{i,j \in g_{a,s}} |\psi_{i,n} - \psi_{j,n}|_2^2 = O\left(\left(n/(|L_n|\overline{k}_n)\right)^{-2/(d+1)}\right)
$$

This finishes the proof of the first statement. The second statement follows by optimality of $\mathcal{G}_{nl}^*$ and comparison with the groups just constructed. $\square$

11.2 CLT

Lemma 11.2 (Balancing)**.** *Suppose $T_{1:n} \sim \mathrm{Loc}(\psi_1, q(x))$ and $D_{1:n} \sim \mathrm{Loc}(\psi_2, p(x))$ with $(\psi_1(X), q(X)) \in \sigma(\psi_2(X))$, and ψ_1, ψ_2, q, and p possibly dependent on $\xi \perp\!\!\!\perp W_{1:n}$. Suppose $E[|\psi_1(X)|^{\alpha_1}] < \infty$ for $\alpha_1 > d(\psi_2) + 1$ and $E[|\psi_2(X)|^{\alpha_1}] < \infty$ for $\alpha_2 > d(\psi_2) + 1$. If $E[f(\psi_2(X), p(X), \xi)^2 | \xi] < \infty$ ξ-a.s. then*

$$
E_n[T_i(D_i - p(X_i)) f(\psi_2(X_i), p(X_i), \xi)] = o_p(n^{-1/2})
$$

If $E[g(\psi_1(X), q(X), \xi)^2|\xi] < \infty$ ξ-a.s. then $E_n[(T_i - q(X_i)) g(\psi_1(X_i), q(X_i), \xi)] = o_p(n^{-1/2})$. Suppose further that $\overline{k}_n |L_n| = o(n^{1-(d+1)/(\alpha_1 \wedge \alpha_2)})$. Then $E_n[T_i(D_i - p_n(X_i)) f(\psi_2(X_i), \xi)] = o_p(n^{-1/2})$ and $E_n[(T_i - q_n(X_i)) g(\psi_1(X_i), \xi)] = o_p(n^{-1/2})$ under the same conditions.

Proof. We begin with the first claim. Fix a sequence c_n with $c_n \to \infty$ and $c_n^2 = o((\overline{k}_n |L_n|)^{-2/(d+1)} n^{2/(d+1) - 2/\alpha_2})$. Work on the assumed almost sure event and fix ξ. Denote $Z = (\psi_2(X), p(X))$. Define $L_2(Z) = \{g(Z) : E[g(Z)^2|\xi] < \infty\}$. By assumption, $f(Z, \xi) \in L_2(Z)$. Our strategy is to approximate $f(Z, \xi)$ by Lipschitz functions. Define the spaces $\mathcal{L}_n = \{g(Z) \in L_2(Z) : |g|_{lip} \vee |g|_\infty \leq c_n\}$ and let $g_n \in \mathcal{L}_n$ be such that $E[(g_n(Z) - f(Z))^2] \leq 2 \inf_{g \in \mathcal{L}_n} E[(g(Z) - f(Z))^2]$. We claim that $|g_n - f|_{2,Z}^2 \to 0$. Let $\epsilon > 0$ and note that by Lemma 11.10 there exists a function $|h|_{lip} \vee |h|_\infty < \infty$ with

$|h-f|_{2,Z}^2 < \epsilon$. Since $c_n \to \infty$, there exists N such that $c_n \ge |h|_{lip} \vee |h|_\infty$ for all $n \ge N$. Then for $n \ge N$ we have $|g_n - f|_{2,Z}^2 \le 2\inf_{g\in\mathcal{L}_n} |g-f|_{2,Z}^2 \le 2|h-f|_{2,Z}^2 < 2\epsilon$ since $h \in \mathcal{L}_n$. Since ϵ was arbitrary, this shows the claim. Summarizing our work so far, we found $g_n(Z,\xi)$ such that $E[(f(Z,\xi)-g_n(Z,\xi))^2|\xi] = o(1)$, $|g_n(Z,\xi)-g_n(Z',\xi)| \le c_n|Z-Z'|_2$, and $|g_n(\cdot,\xi)|_\infty \le c_n$ ξ-a.s. Now we expand

$$E_n[T_i(D_i-p_i)f(Z_i,\xi)] = E_n[T_i(D_i-p_i)(f(Z_i,\xi)-g_n(Z_i,\xi))] + E_n[T_i(D_i-p_i)g_n(Z_i,\xi)]$$

Denote these terms A_n, B_n. Let $\mathcal{F}_n = \sigma((\psi_2)_{1:n}, (p_i)_{1:n}, \pi_n, \tau^t, \xi)$. Since $(\psi_1, q) \in \sigma(\psi_2)$, we have $\mathcal{G}_n, f(Z_i)_{1:n}, g_n(Z_i)_{1:n} \in \mathcal{F}_n$. Then by Lemma 11.16 $E[A_n|\mathcal{F}_n] = 0$ and

$$\mathrm{Var}(\sqrt{n}E_n[T_i(D_i-p_i)(f(Z_i,\xi)-g_n(Z_i,\xi))]|\mathcal{F}_n) \le 2E_n[(f(Z_i,\xi)-g_n(Z_i,\xi))^2] = o_p(1)$$

The equality by conditional Markov (Lemma 11.9) since $E[E_n[(f(Z_i,\xi)-g_n(Z_i,\xi))^2]|\xi] = |f-g_n|_{2,Z}^2 = o(1)$ ξ-a.s. as shown above. Then $A_n = o_p(n^{-1/2})$, again by conditional Markov. Next consider B_n. By Lemma 11.16, we have $E[B_n|\mathcal{F}_n] = 0$. Using the claim and Lemma 11.16, we calculate

$$\begin{aligned}
\mathrm{Var}(\sqrt{n}B_n|\mathcal{F}_n) &\le n^{-1}\sum_g \frac{1}{|g|}\sum_{i,j\in g}(g_n(Z_i,\xi)-g_n(Z_j,\xi))^2 + n^{-1}\overline{k}_n|L_n|\max_{i=1}^n g_n(Z_i,\xi)^2 \\
&\le c_n^2 n^{-1}\sum_g \frac{1}{|g|}\sum_{i,j\in g}|Z_i-Z_j|_2^2 + n^{-1}\overline{k}_n|L_n|c_n^2 \\
&= c_n^2 n^{-1}\sum_g \frac{1}{|g|}\sum_{i,j\in g}(|\psi_{2,i}-\psi_{2,j}|_2^2 + |p_i-p_j|^2) + n^{-1}\overline{k}_n|L_n|c_n^2 \\
&= c_n^2 o_p(n^{2/\alpha_2}(n/(\overline{k}_n|L_n|)^{-2/(d+1)}) + n^{-1}\overline{k}_n|L_n|c_n^2 = o_p(1)
\end{aligned}$$

The second to last equality since $p_i = p_j$ for any $i,j \in g$, $\forall g$ and by Theorem 11.1, using the fact that $\max_{i=1}^n |\psi_{2,i}|^2 = o_p(n^{2/\alpha_2})$ by Lemma 11.13. The final equality follows by our choice of c_n. Then $B_n = o_p(n^{-1/2})$, again by conditional Markov. This finishes the proof. The conclusion for sampling variables follows by the same proof, setting $T_i, q_i \to 1$, $D_i \to T_i$ and $p_i \to q_i$ and $\psi_2 \to \psi_1$. The second set of claims follows by setting $Z = \psi_2(X)$ and $p_i \to p_{i,n}$ in the above proof. □

Assumption 11.3. *Consider the following assumptions*

(A1) $E[Y(d)^2] < \infty$ *and* ψ_1, ψ_2, q, p *as in Theorem 3.11 with* $q_i, p_i \in [\delta, 1-\delta] \subseteq (0,1)$.

(A2) $E[Y(d)^4] < \infty$ *and propensities* $\widehat{p}_{i,n} = \widehat{p}_{i,n}(X_i, \xi_n)$, $\widehat{q}_{i,n} = \widehat{q}_{i,n}(X_i, \xi_n)$, $p_i = p(X_i)$, *and* $q_i = q(X_i)$ *for* $\xi_n \perp\!\!\!\perp \sigma(W_{1:n}, \tau^t, \tau^d, \pi_n)$. $\widehat{p}_{i,n}, p_i \in [\delta, 1-\delta] \subseteq (0,1)$ *and* $\widehat{q}_{i,n}, q_i \ge \delta$. *Also* $E_n[(\widehat{p}_{i,n}-p_i)^2] = o_p(1)$ *and* $E_n[(\widehat{q}_{i,n}-q_i)^2] = o_p(1)$.

Theorem 11.4 (CLT). *(1) Suppose Assumption 11.3 (A1) holds. Let* $T_{1:n} \sim \mathrm{Loc}(\psi_1, q(x))$ *and* $D_{1:n} \sim \mathrm{Loc}(\psi_2, p(x))$. *Define* $c_{1,i} = E[Y_i(1)-Y_i(0)|\psi_{1,i}, q_i, \xi]$. *Then* $\sqrt{n}(E_n[c_{1,i}] - \mathrm{ATE})|\xi \Rightarrow \mathcal{N}(0, V_1)$ *with* $V_1 = \mathrm{Var}(c_{1,i}|\xi)$. *Define*

$$u_i = q_i^{-1}(E[Y_i(1)-Y_i(0)|\psi_{1,i}, q_i, \xi] - E[Y_i(1)-Y_i(0)|\psi_{2,i}, p_i, \xi])$$

and $\mathcal{F}_{0,n} = \sigma(\xi, \pi_n, (\psi_{1,i})_{1:n}, (q_i)_{1:n}, \tau^t)$. *Then* $\sqrt{n}E_n[T_iu_i]|\mathcal{F}_{0,n} \Rightarrow \mathcal{N}(0, V_1)$ *with asymptotic variance*

$$V_2 = E[q_i^{-1}\mathrm{Var}(E[Y_i(1)-Y_i(0)|\psi_{2,i}, p_i, \xi]|\psi_{1,i}, q_i, \xi)|\xi].$$

Define the residuals $\epsilon_i^d = Y_i(d) - E[Y_i(d)|\psi_{2,i}, p_i, \xi]$ *and* $\mathcal{G}_n = \sigma(\xi, \pi_n, \tau^t, \tau^d, (\psi_{2,i})_{1:n}, (p_i)_{1:n},)$. *Then* $\sqrt{n}E_n[T_i D_i \epsilon_i^1/(p_i q_i) + T_i(1-D_i)\epsilon_i^0/((1-p_i)q_i)]|\mathcal{G}_{0,n} \Rightarrow \mathcal{N}(0, V_3)$ *with variance*

$$V_3 = E\left[\frac{\mathrm{Var}(Y_i(1)|\psi_{2,i}, p_i, \xi)}{q_i p_i} + \frac{\mathrm{Var}(Y_i(0)|\psi_{2,i}, p_i, \xi)}{q_i(1-p_i)}\middle|\xi\right].$$

(2) Alternatively, suppose Assumption 11.3 (A2) holds. Define the residuals $\epsilon_i^d = Y_i(d) - E[Y_i(d)|X]$. *Let* $\psi_{k,n} = \psi_{k,n}(X)$ *for* $k = 1, 2$. *Let* $T_{1:n} \sim \mathrm{Loc}(\psi_{1,n}, \widehat{q}_n(x))$ *and* $D_{1:n} \sim \mathrm{Loc}(\psi_{2,n}, \widehat{p}_n(x))$ *and define* $\mathcal{G}_{0,n} = \sigma(\pi_n, \xi_n, X_{1:n}, \tau^t, \tau^d)$. *Then* $\sqrt{n}E_n[T_i D_i \epsilon_i^1/(\widehat{p}_{i,n}\widehat{q}_{i,n}) + T_i(1-D_i)\epsilon_i^0/((1-\widehat{p}_{i,n})\widehat{q}_{i,n})]|\mathcal{G}_{0,n} \Rightarrow \mathcal{N}(0, V_3)$ *with variance*

$$V_3 = E\left[\frac{1}{q(X_i)}\left(\frac{\sigma_1^2(X_i)}{p(X_i)} + \frac{\sigma_0^2(X_i)}{1-p(X_i)}\right)\right].$$

Proof. First consider (1). Define $\mathcal{H}_{0,n} = \sigma(\xi)$ and $\mathcal{H}_{k,n} = \sigma(\xi, (c_{1,i})_{i=1:k})$ for $k \geq 1$. Claim that $(\mathcal{H}_{k,n}, c_{1,k} - \mathrm{ATE})_{k\geq 1}$ is an MDS. Adaptation is clear. For the MDS property, $E[c_{1,i}|\mathcal{F}_{i-1,n}] = E[c_{1,i}|\xi, (c_{1,k})_{k=1:i-1}] = E[c_{1,i}|\xi] = E[Y_i(1) - Y_i(0)|\xi] = \mathrm{ATE}$ by tower law and independence. Similarly, $E[(c_{1,i} - \theta)^2|\mathcal{F}_{i-1,n}] = E[(c_{1,i} - \theta)^2|\xi] = \mathrm{Var}(c_{1,i}|\xi)$. To finish, we show the Lindeberg condition. Note that by conditional Jensen inequality $Var(c_{1,i}|\xi) \leq E[c_{1,i}^2|\xi] \leq E[(Y_1 - Y_0)^2] < \infty$. Then we have $(c_{1,i} - \theta)^2\mathbb{1}((c_{1,i} - \theta)^2 > n\epsilon) \to 0$ ξ-a.s., so by dominated convergence as $n \to \infty$

$$E_n[E[(c_{1,i} - \theta)^2\mathbb{1}((c_{1,i} - \theta)^2 > n\epsilon^2)|\xi]] = E[(c_{1,i} - \theta)^2\mathbb{1}((c_{1,i} - \theta)^2 > n\epsilon^2)|\xi] \to 0$$

Then the Lindberg condition is satisfed, so by Theorem 11.6 $\sqrt{n}(E_n[c_{1,i}] - \mathrm{ATE})|\xi \Rightarrow \mathcal{N}(0, V_1)$ with the claimed limiting variance. Consider the second claim. Define $\mathcal{F}_{k,n} = \sigma(\mathcal{F}_{0,n}, (\psi_{2,i})_{i=1:k}, (p_i)_{i=1:k})$ for $k \geq 1$. Claim that $(T_k u_k, \mathcal{F}_{k,n})_{k\geq 1}$ is an MDS. Adaptation is clear from the definitions. Next we show the MDS property. Note that $T_{1:n} \in \mathcal{F}_{0,n}$ since $T_{1:n} = G(\tau^t, (\psi_{1,i})_{1:n}, (q_i)_{1:n}, \pi_n)$. Note the crucial fact that $(A, B) \perp\!\!\!\perp C \implies A \perp\!\!\!\perp C \,|\, B$. By randomization we have $(u_{k+1}, (\psi_{1,i})_{1:n}, (q_i)_{1:n}, \xi, (\psi_{2,i})_{i=1:k}, (p_i)_{i=1:k}) \perp\!\!\!\perp (\tau^t, \pi_n)$. Combining this with the crucial fact, gives conditional independence from (π_n, τ^t)

$$\begin{aligned}
&E[T_{k+1}u_{k+1}|\mathcal{F}_{k,n}] = T_{k+1}E[u_{k+1}|\tau^t, (\psi_{1,i})_{1:n}, (q_i)_{1:n}, \pi_n, (\psi_{2,i})_{i=1:k}, (p_i)_{i=1:k}, \xi] \\
&= T_{k+1}E[u_{k+1}|(\psi_{1,i})_{1:n}, (q_i)_{1:n}, (\psi_{2,i})_{i=1:k}, (p_i)_{i=1:k}, \xi]
\end{aligned}$$

Next, note that since $(X_i)_{i=1}^n$ are iid and $\xi \perp\!\!\!\perp X_{1:n}$, we have $f(X_{k+1}, \xi) \perp\!\!\!\perp g(X_{-(k+1)}, \xi)|\xi$ for any functions f, g. Let $g(X_{-(k+1)}, \xi) = ((\psi_{1,i})_{i\neq k+1}, (q_i)_{i\neq k+1}, \xi, (\psi_{2,i})_{i=1:k}, (p_i)_{i=1:k})$ and $f(X_{k+1}, \xi) = u_{k+1}$, giving conditional independence

$$\begin{aligned}
&E[u_{k+1}|(\psi_{1,i})_{1:n}, (q_i)_{1:n}, (\psi_{2,i})_{i=1:k}, (p_i)_{i=1:k}, \xi] = E[u_{k+1}|\psi_{1,k+1}, q_{k+1}, \xi] \\
&= q_{k+1}^{-1}E[Y_{k+1}(1) - Y_{k+1}(0)|\psi_{1,k+1}, q_{k+1}, \xi] \\
&- q_{k+1}^{-1}E[E[Y_{k+1}(1) - Y_{k+1}(0)|\psi_{2,k+1}, p_{k+1}, \xi]|\xi, \psi_{1,k+1}, q_{k+1}] = 0
\end{aligned}$$

The final equality by tower law since $\sigma(\xi, \psi_{1,k+1}, q_{k+1}) \subseteq \sigma(\psi_{2,k+1}, p_{k+1}, \xi)$ by the increasing stratification assumption. This shows the MDS property. Next, we compute the variance process $\Sigma_{k,n} = \sum_{i=1}^k E[(T_i u_i)^2|\mathcal{F}_{i-1,n}]$. By the exact argument above $E[T_i u_i^2|\mathcal{F}_{i-1,n}] = T_i E[u_i^2|\psi_{1,i}, q_i, \xi]$, so that $E[T_i u_i^2|\mathcal{F}_{i-1,n}] = T_i E[E[u_i^2|\mathcal{F}_{i-1,n}]|\mathcal{F}_{i-1,n}] = T_i E[u_i^2|\psi_{1,i}, q_i, \xi]$, since $(\psi_{1,i}, q_i, \xi) \in \mathcal{F}_{0,n} \subseteq \mathcal{F}_{i-1,n}$. In particular, we have shown that $\Sigma_{k,n} \in \mathcal{F}_{0,n}$ for all

k, n, satisfying the variance condition of Proposition 11.6. We have

$$\Sigma_n = E_n[T_i E[u_i^2|\psi_{1,i}, q_i, \xi]] = E_n[(T_i - q_i)E[u_i^2|\psi_{1,i}, q_i, \xi]] + E_n[q_i E[u_i^2|\psi_{1,i}, q_i, \xi]]$$

Call these terms A_n, B_n. By conditional Jensen and Young's, we have $E[|E[u_i^2|\psi_{1,i}, q_i, \xi]|] \leq E[u_i^2] \lesssim E[(Y_i(1) - Y_i(0))^2] \lesssim \sum_{d=0,1} E[Y_i(d)^2] < \infty$. Then $A_n = o_p(1)$ by Lemma 11.16. For the second term B_n, we have

$$\begin{aligned} E[B_n|\xi] &= E[E_n[q_i E[u_i^2|\psi_{1,i}, q_i, \xi]]|\xi] = E[q_i E[u_i^2|\psi_{1,i}, q_i, \xi]|\xi] \\ &= E[q_i^{-1} \operatorname{Var}(E[Y_i(1) - Y_i(0)|\psi_{2,i}, p_i, \xi]|\psi_{1,i}, q_i, \xi)|\xi] \end{aligned}$$

Moreover, similar to the calculation for A_n, conditional Jensen and Young's show that $E[|E[u_i^2|\psi_{1,i}, q_i, \xi]|\xi] \lesssim \sum_{d=0,1} E[Y_i(d)^2] < \infty$, so $B_n - E[B_n|\xi] = o_p(1)$ by conditional WLLN. Then we have shown $\Sigma_n - E[B_n|\xi] = o_p(1)$, with the claimed limit. Finally, we show the Lindberg condition in Equation 11.3. Let $\Delta_i = Y_i(1) - Y_i(0)$. By the propensity lower bound and Young's inequality $T_i^2 u_i^2 \leq 2\delta^{-2}(E[\Delta_i|\psi_{1,i}, q_i, \xi]^2 + E[\Delta_i|\psi_{2,i}, p_i, \xi]^2) \equiv v_i$. Note also that $E[v_i] \lesssim E[E[\Delta_i|\psi_{1,i}, q_i, \xi]^2 + E[\Delta_i|\psi_{2,i}, p_i, \xi]^2] \leq 2E[\Delta_i^2] \lesssim E[Y(1)^2 + Y(0)^2] < \infty$. The second inequality is conditional Jensen. Then for $\epsilon > 0$

$$n^{-1} \sum_{i=1}^n E[T_i^2 u_i^2 \mathbb{1}(T_i^2 u_i^2 > n\epsilon^2)|\mathcal{F}_{0,n}] \leq n^{-1} \sum_{i=1}^n E[v_i \mathbb{1}(v_i > n\epsilon^2)|\mathcal{F}_{0,n}] \equiv C_n$$

As in the conditional independence arguments above, we have

$$\begin{aligned} n^{-1} \sum_{i=1}^n E[v_i \mathbb{1}(v_i > n\epsilon^2)|\mathcal{F}_{0,n}] &= n^{-1} \sum_{i=1}^n E[v_i \mathbb{1}(v_i > n\epsilon^2)|\tau^t, \pi_n, (\psi_{1,i})_{1:n}, (q_i)_{1:n}, \xi] \\ = n^{-1} \sum_{i=1}^n E[v_i \mathbb{1}(v_i > n\epsilon^2)|(\psi_{1,i})_{1:n}, (q_i)_{1:n}, \xi] &= n^{-1} \sum_{i=1}^n E[v_i \mathbb{1}(v_i > n\epsilon^2)|\psi_{1,i}, q_i, \xi] \end{aligned}$$

Then $E[C_n] = E[E[C_n|\xi]] = E[v_i \mathbb{1}(v_i > n\epsilon^2)] \to 0$ as $n \to \infty$ since $v_i \geq 0$ and $E[v_i] < \infty$. Then $C_n \xrightarrow{p} 0$ by Markov inequality. This finishes the proof of the Lindberg condition. We skip the proof of the final claim, since it is similar to the proof of claim (2).

Next we show claim (2). Define $\mathcal{G}_{k,n} = \sigma(\mathcal{G}_{0,n}, (\epsilon_i^0, \epsilon_i^1)_{i=1:k})$. Then claim that $(z_{i,n}, \mathcal{G}_{i,n})_{i\geq 1}$ is an MDS with $z_{i,n} = T_i D_i \epsilon_i^1/(\widehat{p}_{i,n}\widehat{q}_{i,n}) + T_i(1 - D_i)\epsilon_i^0/((1 - \widehat{p}_{i,n})\widehat{q}_{i,n})$. Adaptation is clear. Next we show the MDS property. It suffices to show $E[\epsilon_i^d|\mathcal{G}_{i-1,n}] = 0$. Again we use the fact $(A, B) \perp\!\!\!\perp C \implies A \perp\!\!\!\perp C \,|\, B$. Then note that $E[\epsilon_i^d|\mathcal{G}_{i-1,n}]$ is equal to

$$E[\epsilon_i^d|\pi_n, \xi_n, X_{1:n}, \tau^t, \tau^d, (\epsilon_i^0, \epsilon_i^1)_{i=1:i-1}] = E[\epsilon_i^d|X_{1:n}, (\epsilon_i^0, \epsilon_i^1)_{i=1:i-1}] = E[\epsilon_i^d|X_i] = 0.$$

The second equality by the fact with $A = \epsilon_i^d$, $B = (X_{1:n}, (\epsilon_i^0, \epsilon_i^1)_{i=1:i-1})$, and $C = (\pi_n, \xi_n, \tau^t, \tau^d)$. The third equality by setting $A = \epsilon_i^d$, $B = X_i$, and $C = (X_{-i}, (\epsilon_i^0, \epsilon_i^1)_{i=1:i-1}))$. This proves the MDS property. Next we analyze $\Sigma_{k,n} = n^{-1} \sum_{i=1}^k E[z_{i,n}^2|\mathcal{G}_{i-1,n}]$, the variance process. By the exact reasoning above, $E[(\epsilon_i^d)^2|\mathcal{G}_{i-1,n}] = E[(\epsilon_i^d)^2|X_i] = \sigma_d^2(X_i)$.

$$\Sigma_{k,n} = n^{-1} \sum_{i=1}^k T_i D_i \sigma_1^2(X_i)/(\widehat{p}_{i,n}\widehat{q}_{i,n})^2 + T_i(1 - D_i)\sigma_0^2(X_i)/((1 - \widehat{p}_{i,n})\widehat{q}_{i,n})^2$$

so that $\Sigma_{k,n} \in \mathcal{G}_{0,n}$ for all k, n, satisfying the variance condition of Proposition 11.6.

$$E_n[T_i D_i \sigma_1^2(X_i)/(\widehat{p}_{i,n}\widehat{q}_{i,n})^2] = E_n[((T_i - \widehat{q}_{i,n}) + \widehat{q}_{i,n})D_i\sigma_1^2(X_i)/(\widehat{p}_{i,n}\widehat{q}_{i,n})^2] = R_n^1 + R_n^2$$

By our propensity lower bounds we have $D_i\sigma_1^2(X_i)/(\widehat{p}_{i,n}\widehat{q}_{i,n})^2 \le \delta^{-4}\sigma_1^2(X_i)$ with $E[\sigma_1^2(X)] < \infty$, so $D_i\sigma_1^2(X_i)/(\widehat{p}_{i,n}\widehat{q}_{i,n})^2$ is uniformly integrable. Then $R_n^1 = o_p(1)$ by Lemma 11.16. Similarly, $R_n^2 = E_n[((D_i - \widehat{p}_{i,n}) + \widehat{p}_{i,n})\sigma_1^2(X_i)/\widehat{p}_{i,n}^2\widehat{q}_{i,n}] = R_n^3 + R_n^4$ and $R_n^3 = o_p(1)$ by the same argument. Then $\Sigma_n = E_n[\sigma_1^2(X_i)/\widehat{p}_{i,n}\widehat{q}_{i,n}] + o_p(1)$.

$$\begin{aligned}
&|E_n[\sigma_1^2(X_i)/\widehat{p}_{i,n}\widehat{q}_{i,n} - \sigma_1^2(X_i)/p_iq_i]| \le \delta^{-4}|E_n[\sigma_1^2(X_i)((q_i - \widehat{q}_{i,n})p_i + \widehat{q}_{i,n}(p_i - \widehat{p}_{i,n}))]| \\
&\le E_n[\sigma_1^2(X_i)^2]^{1/2}\left(E_n[(q_i - \widehat{q}_{i,n})^2]^{1/2} + E_n[(p_i - \widehat{p}_{i,n})^2]^{1/2}\right) = O_p(1)o_p(1) = o_p(1)
\end{aligned}$$

Then $E_n[T_iD_i\sigma_1^2(X_i)/(\widehat{p}_{i,n}\widehat{q}_{i,n})^2] = E_n[\sigma_1^2(X_i)/p_iq_i] + o_p(1) = E[\sigma_1^2(X_i)/p_iq_i] + O_p(n^{-1/2})$ by Chebyshev inequality. By symmetry, the same holds for the $D = 0$ term. Putting this together, we have shown that $\Sigma_n \xrightarrow{p} E[\sigma_1^2(X_i)/(p_iq_i) + \sigma_0^2(X_i)/(1-p_i)q_i]$ as claimed. Finally, we show the Lindberg condition. Note the bound $|z_{i,n}|^2 \le 2\delta^{-4}((\epsilon_i^1)^2 + (\epsilon_i^0)^2) \equiv v_i$, with $E[v_i] \lesssim \mathrm{Var}(Y(1)) + \mathrm{Var}(Y(0)) < \infty$. Then for $\epsilon > 0$ we have

$$\begin{aligned}
L_n &= n^{-1}\sum_{i=1}^n E[z_{i,n}^2 \mathbb{1}(z_{i,n}^2 > n\epsilon^2)|\mathcal{G}_{0,n}] \le n^{-1}\sum_{i=1}^n E[v_i\mathbb{1}(v_i > n\epsilon^2)|\mathcal{F}_{0,n}] \\
&= n^{-1}\sum_{i=1}^n E[v_i\mathbb{1}(v_i > n\epsilon^2)|\pi_n, \xi_n, X_{1:n}, \tau^t, \tau^d] = n^{-1}\sum_{i=1}^n E[v_i\mathbb{1}(v_i > n\epsilon^2)|X_i]
\end{aligned}$$

The inequality from the upper bound just stated. The second equality from the conditional independence reasoning above. Then $E[L_n] = E[v_i\mathbb{1}(v_i > n\epsilon^2)] \to 0$ as $n \to \infty$ since $E[v_i] < \infty$. Then by Markov inequality $L_n \xrightarrow{p} 0$, which finishes the proof. □

Proof of Theorem 3.11. Define $c_{1,i} = E[Y(1) - Y(0)|\psi_{1,i}, q_i, \xi]$ and $c_{2,i} = E[Y(1) - Y(0)|\psi_{2,i}, p_i, \xi]$. Define $b_{2,i} = E[Y(1)|\psi_{2,i}, p_i, \xi]((1-p_i)/p_i)^{1/2} + E[Y(0)|\psi_{2,i}, p_i, \xi](p_i/(1-p_i))^{1/2}$ and $\epsilon_i^d = Y_i(d) - E[Y_i(d)|\psi_{2,i}, p_i, \xi]$. Define the σ-algebras $\mathcal{F}_{k,n}$ for $1 \le k \le 4$

$$\mathcal{F}_{1,n} = \sigma(\psi_{1,1:n}, q_{1:n}, \xi, \pi_n) \qquad \mathcal{F}_{2,n} = \sigma(\mathcal{F}_{1,n}, T_{1:n}) \qquad \mathcal{F}_{3,n} = \sigma(\mathcal{F}_{2,n}, \psi_{2,1:n}, p_{1:n})$$

and $\mathcal{F}_{4,n} = \sigma(\mathcal{F}_{3,n}, D_{1:n})$. We expand our estimator by projection

$$\begin{aligned}
\theta - \widehat{\theta} &= (\theta - E[\widehat{\theta}|\mathcal{F}_{1,n}]) + (E[\widehat{\theta}|\mathcal{F}_{1,n}] - E[\widehat{\theta}|\mathcal{F}_{2,n}]) + (E[\widehat{\theta}|\mathcal{F}_{2,n}] - E[\widehat{\theta}|\mathcal{F}_{3,n}]) \\
&\quad + (E[\widehat{\theta}|\mathcal{F}_{3,n}] - E[\widehat{\theta}|\mathcal{F}_{4,n}]) + (E[\widehat{\theta}|\mathcal{F}_{4,n}] - \widehat{\theta}) \\
&= E_n[c_{1,i} - \theta] + E_n\left[\frac{T_i - q_i}{q_i}c_{1,i}\right] + E_n\left[\frac{T_i}{q_i}(c_{2,i} - c_{1,i})\right] + E_n\left[\frac{(D_i - p_i)T_i}{q_i\sqrt{p_i - p_i^2}}b_{2,i}\right] \\
&\quad + E_n\left[\frac{D_iT_i\epsilon_i^1}{p_iq_i} + \frac{(1-D_i)T_i\epsilon_i^0}{(1-p_i)q_i}\right] \equiv A_n + B_n + C_n + S_n + R_n
\end{aligned}$$

We wish to apply Lemma 11.2 to show that $\sqrt{n}B_n, \sqrt{n}S_n = o_p(1)$. It suffices to check integrability. In the notation of the lemma, define $g(\psi_{1,i}, q_i, \xi) = c_{1,i}/q_i$, and note that by our propensity bound, Young's inequality, and contraction

$$E[E[c_{1,i}^2/q_i^2|\xi]] \le \delta^{-2}E[c_{1,i}^2] \lesssim E[\sum_{d=0,1} E[Y(d)|\psi_{1,i},q_i,\xi]^2] \le \sum_{d=0,1} E[Y(d)^2]$$

This shows $E[c_{1,i}^2/q_i^2|\xi] < \infty$ ξ-a.s. Similarly, $E[b_{2,i}^2 q_i^{-2}(p_i(1-p_i))^{-1}|\xi] < \infty$ ξ-a.s. Then $\sqrt{n}B_n, \sqrt{n}S_n = o_p(1)$ by the lemma.

The conditions of Theorem 11.4 are satisfied by assumption. Then $\sqrt{n}A_n|\xi \Rightarrow \mathcal{N}(0,V_1)$ with $V_1 = \operatorname{Var}(c_{1,i}|\xi)$, $\sqrt{n}C_n|\mathcal{F}_{2,n} \Rightarrow \mathcal{N}(0,V_2)$ with $V_2 = E[q_i^{-1}\operatorname{Var}(c_{2,i}|\psi_{1,i},q_i,\xi)|\xi]$, and $\sqrt{n}R_n|\mathcal{F}_{4,n} \Rightarrow \mathcal{N}(0,V_3)$ with $V_3 = E[q_i^{-1}(\sigma_{1,i}^2 p_i^{-1} + \sigma_{0,i}^2(1-p_i)^{-1})|\xi]$ for $\sigma_{d,i}^2 = \operatorname{Var}(Y(d)|\psi_{2,i},p_i,\xi)$. Then by Slutsky and Lemma 11.11, we have $\sqrt{n}(\widehat{\theta} - \text{ATE})|\xi \Rightarrow \mathcal{N}(0, V_1+V_2+V_3)$.

To finish the proof, we use algebra to reformulate the variance $V = V_1 + V_2 + V_3$. By the law of total variance $\operatorname{Var}(c(X)|\xi) = \operatorname{Var}(E[c(X)|\psi_1,q,\xi]|\xi) + E[\operatorname{Var}(c(X)|\psi_1,q,\xi)|\xi]$, with $E[c(X_i)|\psi_{1,i},q_i,\xi] = c_{1,i}$. Then since $\xi \perp\!\!\!\perp W_{1:n}$

$$V_1 = \operatorname{Var}(c(X)) - E[\operatorname{Var}(c(X)|\psi_1,q,\xi)|\xi].$$

Next, by the law of total variance

$$\operatorname{Var}(c(X_i)|\psi_{1,i},q_i,\xi) = E[\operatorname{Var}(c(X)|\psi_{2,i},p_i,\xi)|\psi_{1,i},q_i,\xi] + \operatorname{Var}(E[c(X)|\psi_{2,i},p_i,\xi]|\psi_{1,i},q_i,\xi)$$

Using this equality and applying tower law gives

$$\begin{aligned}
V_2 &= E[q_i^{-1}\operatorname{Var}(c_{2,i}|\psi_{1,i},q_i,\xi)|\xi] = E[q_i^{-1}\operatorname{Var}(E[c(X)|\psi_{2,i},p_i,\xi]|\psi_{1,i},q_i,\xi)|\xi] \\
&= E[q_i^{-1}\operatorname{Var}(c(X_i)|\psi_{1,i},q_i,\xi)|\xi] - E[q_i^{-1}E[\operatorname{Var}(c(X)|\psi_{2,i},p_i,\xi)|\psi_{1,i},q_i,\xi]|\xi] \\
&= E[q_i^{-1}\operatorname{Var}(c(X_i)|\psi_{1,i},q_i,\xi)|\xi] - E[q_i^{-1}\operatorname{Var}(c(X)|\psi_{2,i},p_i,\xi)|\xi]
\end{aligned}$$

Also, $\operatorname{Var}(Y_i(d)|\psi_{2,i},p_i,\xi) = \operatorname{Var}(E[Y_i(d)|X_i,\xi]|\psi_{2,i},p_i,\xi)] + E[\operatorname{Var}(Y_i(d)|X_i,\xi)|\psi_{2,i},p_i,\xi]$. Using this fact gives

$$\begin{aligned}
V_3 &= E[(q_ip_i)^{-1}\sigma_{1,i}^2|\xi] + E[(q_i(1-p_i))^{-1}\sigma_{0,i}^2|\xi] \\
&= E[(q_ip_i)^{-1}\operatorname{Var}(m_1(X_i)|\psi_{2,i},p_i,\xi)|\xi] + E[(q_ip_i)^{-1}E[\sigma_1^2(X_i)|\psi_{2,i},p_i,\xi]|\xi] \\
&+ E[(q_i(1-p_i))^{-1}\operatorname{Var}(m_0(X_i)|\psi_{2,i},p_i,\xi)|\xi] + E[(q_i(1-p_i))^{-1}E[\sigma_0^2(X_i)|\psi_{2,i},p_i,\xi]|\xi] \\
&= E[(q_ip_i)^{-1}\operatorname{Var}(m_1(X_i)|\psi_{2,i},p_i,\xi)|\xi] + E[(q_i(1-p_i))^{-1}\operatorname{Var}(m_0(X_i)|\psi_{2,i},p_i,\xi)|\xi] \\
&+ E[(q_ip_i)^{-1}\sigma_1^2(X_i)|\xi] + E[(q_i(1-p_i))^{-1}\sigma_0^2(X_i)|\xi]
\end{aligned}$$

The final equality by tower law and since $q_i \in \sigma(\psi_{2,i},p_i,\xi)$ by construction. Putting this all together, we get total variance

$$\begin{aligned}
V_1 + V_2 + V_3 &= \operatorname{Var}(c(X_i)) - E[\operatorname{Var}(c(X_i)|\psi_{1,i},q_i,\xi)|\xi] \\
&+ E[q_i^{-1}\operatorname{Var}(c(X_i)|\psi_{1,i},q_i,\xi)|\xi] - E[q_i^{-1}\operatorname{Var}(c(X)|\psi_{2,i},p_i,\xi)|\xi] \\
&+ E[(q_ip_i)^{-1}\operatorname{Var}(m_1(X_i)|\psi_{2,i},p_i,\xi)|\xi] + E[(q_i(1-p_i))^{-1}\operatorname{Var}(m_0(X_i)|\psi_{2,i},p_i,\xi)|\xi] \\
&+ E[(q_ip_i)^{-1}\sigma_1^2(X_i)|\xi] + E[(q_i(1-p_i))^{-1}\sigma_0^2(X_i)|\xi] \\
&= T_1 + T_2 + T_3 + T_4 + T_5 + T_6 + T_7 + T_8
\end{aligned}$$

Note that $T_2 + T_3 = E[q_i^{-1}(1-q_i)\operatorname{Var}(c(X_i)|\psi_{1,i}, q_i, \xi)|\xi]$. Also note that

$$
\begin{aligned}
&- \operatorname{Var}(c(X_i)|\psi_{2,i}, p_i, \xi) + p_i^{-1}\operatorname{Var}(m_1(X_i)|\psi_{2,i}, p_i, \xi) + (1-p_i)^{-1}\operatorname{Var}(m_0(X_i)|\psi_{2,i}, p_i, \xi) \\
&= p_i^{-1}(1-p_i)\operatorname{Var}(m_1(X_i)|\psi_{2,i}, p_i, \xi) + p_i(1-p_i)^{-1}\operatorname{Var}(m_0(X_i)|\psi_{2,i}, p_i, \xi) \\
&+ 2\operatorname{Cov}(m_1(X_i), m_0(X_i)|\psi_{2,i}, p_i, \xi) = \operatorname{Var}(b(X_i; p_i)|\psi_{2,i}, p_i, \xi)
\end{aligned}
$$

Then we have $T_4 + T_5 + T_6 = E[q_i^{-1}\operatorname{Var}(b(X_i; p_i)|\psi_{2,i}, p_i, \xi)|\xi]$. Then as claimed

$$
\begin{aligned}
V = \sum_{k=1}^{8} T_k &= \operatorname{Var}(c(X)) + E[q_i^{-1}(1-q_i)\operatorname{Var}(c(X_i)|\psi_{1,i}, q_i, \xi)|\xi] \\
&+ E[q_i^{-1}\operatorname{Var}(b(X_i; p_i)|\psi_{2,i}, p_i, \xi)|\xi] + E\left[q_i^{-1}\left(\frac{\sigma_1^2(X_i)}{p_i} + \frac{\sigma_0^2(X_i)}{1-p_i}\right)\middle|\xi\right].
\end{aligned}
$$

Finally, note $E_n[T_i] = E_n[T_i - q_i] + E_n[q_i] = o_p(n^{-1/2}) + E_n[q_i] = E[q_i|\xi] + O_p(n^{-1/2})$. The second equality by Lemma 11.2, and the third by conditional Chebyshev (Lemma 11.9). Then $\sqrt{n_T}(\widehat{\theta} - \text{ATE}) = \sqrt{n}(E_n[T_i])^{1/2}(\widehat{\theta} - \text{ATE}) = E[q_i|\xi]^{1/2}\sqrt{n}(\widehat{\theta} - \text{ATE}) + o_p(1) \Rightarrow N(0, E[q(X,\xi)] \cdot V)$ by continuous mapping theorem and Slutsky. This finishes the proof $\square$

Definition 11.5 (Conditional Weak Convergence)**.** For random variables $A_n, A \in \mathbb{R}^d$ and σ-algebras $(\mathcal{F}_n)_n$, $\mathcal{G}$ define conditional weak convergence

$$
A_n|\mathcal{F}_n \Rightarrow A|\mathcal{G} \quad \iff \quad E[e^{it'A_n}|\mathcal{F}_n] = E[e^{it'A}|\mathcal{G}] + o_p(1) \quad \forall t \in \mathbb{R}^d
$$

We require a slight modification of the martingale difference CLT in Billingsley, allowing the weak limit to be a mixture of normals.

Proposition 11.6 (MDS-CLT)**.** *Consider probability spaces $(\Omega_n, \mathcal{G}_n, P_n)$ each equipped with filtration $(\mathcal{F}_{k,n})_{k\ge 0}$. Suppose $(Y_{k,n})_{k=1}^n$ is adapted to $(\mathcal{F}_{k,n})_{k\ge 0}$ and has $E[Y_{k,n}|\mathcal{F}_{k-1,n}] = 0$ for all $k \ge 1$ with $n \to \infty$. Make the following definitions*

$$
S_{k,n} = \sum_{j=1}^{k} Y_{k,n} \qquad \sigma_{k,n}^2 = E[Y_{k,n}^2|\mathcal{F}_{k-1,n}] \qquad \Sigma_{k,n} = \sum_{j=1}^{k} \sigma_{k,n}^2
$$

Denote $S_n \equiv S_{n,n}$ and $\Sigma_n \equiv \Sigma_{n,n}$. Suppose that $\sigma_{k,n}^2 \in \mathcal{F}_{0,n}$ for all k, n and $\Sigma_n = \sigma^2 + o_p(1)$ with $\sigma^2 \in \mathcal{F}_{0,n}$. Also, suppose for each $\epsilon > 0$

$$
L_n^\epsilon = \sum_{k=1}^{n} E[Y_{k,n}^2 \mathbb{1}(|Y_{k,n}| \ge \epsilon)|\mathcal{F}_{0,n}] = o_p(1) \tag{11.3}
$$

Then $E[e^{itS_n}|\mathcal{F}_{0,n}] = e^{-\frac{1}{2}t^2\sigma^2} + o_p(1)$.

Proof. We modify the argument in Theorem 35.12 of Billingsley (1995).

$$
\begin{aligned}
E\left[e^{itS_n} - e^{-\frac{1}{2}t^2\sigma^2}|\mathcal{F}_{0,n}\right] &= E[e^{itS_n}(1 - e^{\frac{1}{2}t^2\Sigma_n}e^{-\frac{1}{2}t^2\sigma^2})|\mathcal{F}_{0,n}] \\
&+ E[e^{-\frac{1}{2}t^2\sigma^2}(e^{\frac{1}{2}t^2\Sigma_n}e^{itS_n} - 1)|\mathcal{F}_{0,n}]
\end{aligned}
$$

For the first term, by conditional Jensen inequality

$$\begin{aligned}|E[e^{itS_n}(1-e^{\frac{1}{2}t^2\Sigma_n}e^{-\frac{1}{2}t^2\sigma^2})|\mathcal{F}_{0,n}]| &\leq E[|(1-e^{\frac{1}{2}t^2\Sigma_n}e^{-\frac{1}{2}t^2\sigma^2})||\mathcal{F}_{0,n}]\\ &= |(1-e^{\frac{1}{2}t^2\Sigma_n}e^{-\frac{1}{2}t^2\sigma^2})| = o_p(1)\end{aligned}$$

The first equality since $\Sigma_n, \sigma^2 \in \mathcal{F}_{0,n}$. Since $\Sigma_n = \sigma^2 + o_p(1)$, the second equality follows by continuous mapping. The second term has

$$\begin{aligned}&|E[e^{-\frac{1}{2}t^2\sigma^2}(e^{\frac{1}{2}t^2\Sigma_n}e^{itS_n}-1)|\mathcal{F}_{0,n}]| = e^{-\frac{1}{2}t^2\sigma^2}|E[(e^{\frac{1}{2}t^2\Sigma_n}e^{itS_n}-1)|\mathcal{F}_{0,n}]|\\ &= e^{-\frac{1}{2}t^2\sigma^2}\left|\sum_{k=1}^{n}E[e^{itS_{k-1,n}}e^{\frac{1}{2}t^2\Sigma_{k,n}}(e^{itY_{k,n}}-e^{-\frac{1}{2}t^2\sigma^2_{k,n}})|\mathcal{F}_{0,n}]\right|\\ &\leq e^{-\frac{1}{2}t^2\sigma^2}e^{\frac{1}{2}t^2\Sigma_n}\sum_{k=1}^{n}E[|e^{itS_{k-1,n}}E[e^{itY_{k,n}}-e^{-\frac{1}{2}t^2\sigma^2_{k,n}}|\mathcal{F}_{k-1,n}]||\mathcal{F}_{0,n}]\\ &= e^{-\frac{1}{2}t^2\sigma^2}e^{\frac{1}{2}t^2\Sigma_n}\sum_{k=1}^{n}E[|E[e^{itY_{k,n}}-e^{-\frac{1}{2}t^2\sigma^2_{k,n}}|\mathcal{F}_{k-1,n}]||\mathcal{F}_{0,n}] \equiv o_p(1)Z_n\end{aligned}$$

The first equality since $\sigma^2 \in \mathcal{F}_{0,n}$. The second equality by telescoping. The first inequality by triangle inequality and since $\Sigma_{k,n} \in \mathcal{F}_{0,n}$, $\Sigma_{k,n} \leq \Sigma_n$, and $S_{k-1,n} \in \mathcal{F}_{k-1,n}$ for $1 \leq k \leq n$. The final equality by continuous mapping since $\Sigma_n \xrightarrow{p} \sigma^2$. We want to show that $Z_n = O_p(1)$. Fix $\epsilon > 0$ and let $I_{k,n} = \mathbb{1}(|Y_{k,n}| > \epsilon)$. Note the facts $|e^{ix} - (1 + ix - (1/2)x^2)| \leq (1/6)|x|^3 \wedge |x|^2$ and $|e^z - (1 + z)| \leq |z|^2 e^{|z|}$ for real x, complex z. By the MDS property and $E[Y^2_{k,n}|\mathcal{F}_{k-1,n}] = \sigma^2_{k,n}$, combined with these facts

$$\begin{aligned}&|E[e^{itY_{k,n}}-e^{-\frac{1}{2}t^2\sigma^2_{k,n}}|\mathcal{F}_{k-1,n}]| \leq E[|tY_{k,n}|^3 \wedge |tY_{k,n}|^2 + (1/4)t^4\sigma^4_{k,n}e^{\frac{1}{2}t^2\sigma^2_{k,n}}|\mathcal{F}_{k-1,n}]\\ &\leq (t^2+t^4+|t|^3+e^{\frac{1}{2}t^2\Sigma_n})E[(\epsilon|Y_{k,n}|^2+|Y_{k,n}|^2I_{k,n}+\sigma^4_{k,n})|\mathcal{F}_{k-1,n}]\\ &\equiv A_{n,t}(\epsilon\sigma^2_{k,n}+E[|Y_{k,n}|^2I_{k,n}|\mathcal{F}_{k-1,n}]+E[\sigma^4_{k,n}|\mathcal{F}_{k-1,n}])\end{aligned}$$

Then we have

$$\begin{aligned}Z_n &\leq A_{n,t}\sum_{k=1}^{n}E[\epsilon\sigma^2_{k,n}+E[|Y_{k,n}|^2I_{k,n}|\mathcal{F}_{k-1,n}]+E[\sigma^4_{k,n}|\mathcal{F}_{k-1,n}]|\mathcal{F}_{0,n}]\\ &= A_{n,t}(\epsilon\Sigma_n+L^\epsilon_n)+A_{n,t}\sum_{k=1}^{n}E[\sigma^4_{k,n}|\mathcal{F}_{0,n}] \leq A_{n,t}(\epsilon\Sigma_n+L^\epsilon_n+\Sigma_n(\epsilon^2+L^\epsilon_n))\end{aligned}$$

To see the final inequality, note that $\sigma^4_{k,n} \leq \sigma^2_{k,n}\max_{k=1}^n \sigma^2_{k,n}$ and We have $\sigma^2_{k,n} = E[Y^2_{k,n}|\mathcal{F}_{k-1,n}] \leq \epsilon^2 + E[Y^2_{k,n}I_{k,n}|\mathcal{F}_{k-1,n}] \leq \epsilon^2 + \sum_{j=1}^n E[Y^2_{j,n}I_{j,n}|\mathcal{F}_{j-1,n}]$. Taking $\max_{k=1}^n$ on both sides gives $\max_{k=1}^n \sigma^2_{k,n} \leq \epsilon^2 + \sum_{j=1}^n E[Y^2_{j,n}I_{j,n}|\mathcal{F}_{j-1,n}]$. Then $\sum_{k=1}^n E[\sigma^4_{k,n}|\mathcal{F}_{0,n}] \leq \sum_{k=1}^n E[\sigma^2_{k,n}(\epsilon^2 + \sum_{j=1}^n E[Y^2_{j,n}I_{j,n}|\mathcal{F}_{j-1,n}])|\mathcal{F}_{0,n}] = \Sigma_n(\epsilon^2 + L^\epsilon_n)$. Note that since $\Sigma_n \xrightarrow{p} \sigma^2$, we have $A_{n,t}, \Sigma_n = O_p(1)$ and $L^\epsilon_n = o_p(1)$ by assumption. Since ϵ was arbitrary, this shows $Z_n = o_p(1)$. □

11.3 Optimal Stratification and Pilot Design

Proof of Proposition 4.1. Define $V(q) = E[\bar{\sigma}^2(\psi)/q(\psi)]$. Define the sets

$$\mathcal{Q}' = \{q \in \mathbb{R}^{\psi} : |q|_\infty < \infty, q > 0, E[C(\psi)q(\psi)] \leq \bar{B}, V(q) < \infty\} \qquad \mathcal{Q} = \mathcal{Q}' \cap \{0 < q \leq 1\}$$

and recall the candidate optimal solution $q^*(\psi) = \bar{B} \cdot \bar{\sigma}(\psi)C(\psi)^{-1/2}/E[\bar{\sigma}(\psi)C(\psi)^{1/2}]$. Let $t \in [0,1]$ and $q_1, q_2 \in \mathcal{Q}'$. By convexity of $y \to 1/y$ on $(0,\infty)$, for each $\psi \in \psi$ we have

$$\frac{\bar{\sigma}^2(\psi)}{tq_1(\psi) + (1-t)q_2(\psi)} \leq t\frac{\bar{\sigma}^2(\psi)}{q_1(\psi)} + (1-t)\frac{\bar{\sigma}^2(\psi)}{q_1(\psi)}$$

Taking expectations of both sides gives $V(tq_1 + (1-t)q_2) \leq tV(q_1) + (1-t)V(q_2)$, so V is convex on $\{q > 0\}$ and $\mathcal{Q}'$ is convex. We claim that $q^* \in \mathcal{Q}'$. First, $q^* \in (0,1]$ by assumption. Since $\sup_{\psi\in\psi} q^* \leq 1$ we have $E[\bar{\sigma}(\psi)C(\psi)^{1/2}] > 0$. By Holder $E[\bar{\sigma}(\psi)C(\psi)^{1/2}] \leq E[\bar{\sigma}^2(\psi)]^{1/2}E[C(\psi)]^{1/2} < \infty$. Then we have $V(q^*) = \bar{B}^{-1}E[\bar{\sigma}(\psi)C(\psi)^{1/2}]^2 < \infty$. Also clearly $E[C(\psi)q^*(\psi)] = \bar{B}$. This shows the claim. Next, suppose that $q^* + t\Delta \in \mathcal{Q}'$ for some $t \in [0,1]$ and $|\Delta|_\infty < M$. Since $q^* \in \mathcal{Q}'$, by convexity of $\mathcal{Q}'$, $q^* + t'\Delta \in \mathcal{Q}'$ for all $0 \leq t' \leq t$. Then $dV_{q^*}[\Delta] \equiv \limsup_{t\to 0^+} t^{-1}(V(q^*+t\Delta) - V(q^*))$ is well-defined. We claim that this limit exists. Under our assumptions $q^* > q_l \geq 0$ for some q_l, so that for any ψ and all $t \leq q_l^2/2M < \infty$

$$\left|t^{-1}\left(\frac{\bar{\sigma}^2(\psi)}{q^*(\psi) + t\Delta(\psi)} - \frac{\bar{\sigma}^2(\psi)}{q^*(\psi)}\right)\right| = \frac{|\bar{\sigma}^2(\psi)\Delta(\psi)|}{((q^*)^2 + q^*t\Delta)(\psi)} \leq \frac{M|\bar{\sigma}^2(\psi)|}{q_l^2 - tM} \leq \frac{M|\bar{\sigma}^2(\psi)|}{2q_l^2}$$

Since $\|\bar{\sigma}^2\|_1 < \infty$, dominated convergence implies

$$\begin{aligned} dV_{q^*}[\Delta] &= \lim_{t\to 0^+} t^{-1}(V(q^* + t\Delta) - V(q^*)) = E\left[\lim_{t\to 0^+} \frac{-\bar{\sigma}^2(\psi)\Delta(\psi)}{((q^*)^2 + q^*t\Delta)(\psi)}\right] \\ &= -E\left[\frac{\bar{\sigma}^2(\psi)\Delta(\psi)}{(q^*)^2(\psi)}\right] = (1/\bar{q})^2 E\left[C(\psi)\Delta(\psi)\right] E\left[\bar{\sigma}(\psi)C(\psi)^{1/2}\right]^2 = 0 \end{aligned}$$

The last line since $q^* + t\Delta \in \mathcal{Q}'$ implies $\bar{B} = E[q^*(\psi)C(\psi) + t\Delta(\psi)C(\psi)] = \bar{B} + tE[\Delta(\psi)C(\psi)]$, so that $E[\Delta(\psi)C(\psi)] = 0$. Let $q \in \mathcal{Q}'$, so that $\Delta = q - q^*$ has $|\Delta|_\infty < \infty$. Then by convexity $V(q) - V(q^*) \geq dV_{q^*}[q - q^*] = 0$, showing that $q^* = \text{argmin}_{q\in\mathcal{Q}'} V(q)$. Since $q^* \in \mathcal{Q}$ by assumption, it is optimal over $\mathcal{Q} \subseteq \mathcal{Q}'$ as well.

□

Proof of Theorem 5.2. Suppose that $T_{1:n} \sim \text{Loc}(\psi, \widehat{q}_n(\psi))$ and $D_{1:n} \sim \text{Loc}(\psi, \widehat{p}_n(\psi))$. Similar to the proof of Theorem 3.11, define σ-algebras $\mathcal{F}_{1,n} = \sigma(\psi_{1:n}, \pi_n, \xi_n)$, $\mathcal{F}_{2,n} = \sigma(\mathcal{F}_{1,n}, T_{1:n})$, and $\mathcal{F}_{3,n} = \sigma(\mathcal{F}_{2,n}, D_{1:n})$. Next, expand the estimator

$$\begin{aligned} \widehat{\theta} - \theta &= E_n[c(\psi_i) - \theta] + E_n\left[\frac{T_i - \widehat{q}_{i,n}}{\widehat{q}_{i,n}} c(\psi_i)\right] + E_n\left[\frac{T_i(D_i - \widehat{p}_{i,n})}{\widehat{q}_{i,n}\sqrt{\widehat{p}_{i,n}(1 - \widehat{p}_{i,n})}} b(\psi_i; \widehat{p}_n)\right] \\ &+ E_n\left[\frac{D_i T_i \epsilon_i^1}{\widehat{p}_{i,n}\widehat{q}_{i,n}} + \frac{(1 - D_i)T_i\epsilon_i^0}{(1 - \widehat{p}_{i,n})\widehat{q}_{i,n}}\right] \equiv A_n + B_n + C_n + R_n \end{aligned}$$

Our main task is to show that $\sqrt{n}B_n, \sqrt{n}C_n = o_p(1)$. We do this by modifying the proof of Lemma 11.2.

(1) Balancing Argument - Note that the term C_n has

$$C_n = E_n\left[\frac{T_i(D_i - \widehat{p}_n(\psi_i))}{\widehat{q}_n(\psi_i)}\left(\frac{m_1(\psi_i)}{\widehat{p}_n(\psi_i)} + \frac{m_0(\psi_i)}{1-\widehat{p}_n(\psi_i)}\right)\right] = C_{n1} + C_{n2}.$$

Recall $q^*(\psi) = \bar{B}(\sigma_1(\psi) + \sigma_0(\psi))C(\psi)^{-1/2}/E[(\sigma_1(\psi) + \sigma_0(\psi))C(\psi)^{1/2}]$. Under our assumptions, it's clear that $\inf_\psi q^*(\psi) > 0$. Consider the term above involving $m_1(\psi)$. We have $E[m_1(\psi)^2/q^*(\psi)^2] \lesssim E[m_0(\psi)^2] < \infty$. Then the construction in Lemma 11.2 provides functions $h_n \in L_2(\psi)$ with $|h_n|_{lip} \vee |h_n|_\infty \le c_n$ and $|h_n - (m_1/q^*)|_{2,\psi} \to 0$. Define $\mathcal{F}_n = \sigma(\psi_{1:n}, \xi_n, \pi_n, \tau^t)$. Note that $\mathcal{G}_n, (q^*(\psi_i))_{1:n}, (\widehat{q}_n(\psi_i))_{1:n}, (\widehat{p}_n(\psi_i))_{1:n} \in \mathcal{F}_n$. Then by Lemma 11.16 $E[C_{n1}|\mathcal{F}_n] = 0$. We can expand C_{n1} as

$$E_n\left[\frac{T_i(D_i - \widehat{p}_n(\psi_i))}{\widehat{p}_n(\psi_i)}\left(\left(\frac{m_1(\psi_i)}{\widehat{q}_n(\psi_i)} - \frac{m_1(\psi_i)}{\widehat{q}(\psi_i)}\right) + \left(\frac{m_1(\psi_i)}{\widehat{q}(\psi_i)} - \frac{m_1(\psi_i)}{q^*(\psi_i)}\right) + \frac{m_1(\psi_i)}{q^*(\psi_i)}\right)\right]$$

and write $C_{n1} = T_{n1} + T_{n2} + T_{n3}$. Consider the first term. By construction, $|\widehat{q}_n - \widehat{q}|_\infty \le 1/\overline{k}_n$. Then by Lemma 11.16 we have $E[B_{n1}|\mathcal{F}_n] = 0$ and

$$\mathrm{Var}(\sqrt{n}T_{n1}|\mathcal{F}_n) \lesssim E_n[m_1(\psi_i)^2(\widehat{q}_n - \widehat{q})^2(\psi_i)/\widehat{p}_n(\psi_i)^2] \lesssim o(1)E_n[m_1(\psi_i)^2] = o_p(1)$$

Again by Lemma 11.16 we have

$$\mathrm{Var}(\sqrt{n}T_{n2}|\mathcal{F}_n) \lesssim E_n[m_1(\psi_i)^2(\widehat{q} - q^*)^2(\psi_i)/\widehat{p}_n(\psi_i)^2] \lesssim \max_{i=1}^n m_1(\psi_i)^2 E_n[(\widehat{q} - q^*)^2(\psi_i)]$$

The final expression is $o_p(n^{2/\alpha_2 - 2r}) + O(n^{2/\alpha_2}/\overline{k}_n^2)$, using Lemma 11.13 and the bound from Lemma 11.7. This is $o_p(1)$ under our assumptions. Consider the final term T_{n3}

$$T_{n3} = E_n\left[\frac{T_i(D_i - \widehat{p}_n(\psi_i))}{\widehat{p}_n(\psi_i)}\left(\left(\frac{m_1(\psi_i)}{q^*(\psi_i)} - h_n(\psi_i)\right) + h_n(\psi_i)\right)\right] = T_{n4} + T_{n5}$$

As above $E[T_{n4}|\mathcal{F}_n] = 0$ and $\mathrm{Var}(\sqrt{n}T_{n4}|\mathcal{F}_n) \lesssim E_n[(m_1(\psi_i)/q^*(\psi_i) - h_n(\psi_i))^2] = o_p(1)$ by Markov inequality since $|h_n - (m_1/q^*)|_{2,\mathcal{X}} \to 0$. Finally, by the last part of Lemma 11.16 we have $E[T_{n5}|\mathcal{F}_n] = 0$ and

$$\begin{aligned}
&\mathrm{Var}(\sqrt{n}T_{n5}|\mathcal{F}_n) = n^{-1}\sum_{g\in\mathcal{G}_n}\frac{1}{|g|}\sum_{i,j\in g}(h_n(\psi_i) - h_n(\psi_i))^2 + n^{-1}\overline{k}_n|L_n|\max_{i=1}^n h_n(\psi_i)^2 \\
&\lesssim c_n^2 n^{-1}\sum_{g\in\mathcal{G}_n}\frac{1}{|g|}\sum_{i,j\in g}|\psi_i - \psi_j|_2^2 + n^{-1}\overline{k}_n|L_n|\max_{i=1}^n h_n(\psi_i)^2 \\
&\lesssim c_n^2\max_{i=1}^n|\psi_i|_2^2\cdot O((n/\overline{k}_n|L_n|)^{-2/(d+1)}) + n^{-1}\overline{k}_n|L_n|O_p(c_n^2) \\
&= o_p(c_n^2 n^{2/\alpha_1 - 2/(d+1)}(\overline{k}_n|L_n|)^{2/(d+1)}) = o_p(1)
\end{aligned}$$

The first inequality by Lipschitz continuity, the second inequality by Theorem 11.1. The final equality holds if $\overline{k}_n|L_n| = o(n^{1-\frac{d+1}{\alpha_1}})$ and $c_n = O(n^{2/\alpha_1 - 2/(d+1)}(\overline{k}_n|L_n|)^{2/(d+1)})$ with $c_n \to \infty$. This finishes the proof that $\sqrt{n}B_n = o_p(1)$. An identical proof shows that $\sqrt{n}C_n = o_p(1)$.

(2) CLT Argument - By the previous argument, we have $\sqrt{n}(\widehat{\theta} - \theta) = \sqrt{n}A_n + \sqrt{n}R_n +$

$o_p(1)$. $\sqrt{n}A_n \Rightarrow \mathcal{N}(0, V_1)$ with $V_1 = \mathrm{Var}(c(\psi))$ by vanilla CLT. We want to apply Theorem 11.4 to the term $\sqrt{n}R_n$. The moment bound is satisfied by assumption. For condition (a2), set $\psi_{k,n}(\psi) = \psi$, $p_i = p_i^*$, and $q_i = q_i^*$. The propensity lower and upper bounds follow from our assumptions on $\sigma_d(\psi)$. Condition (c) is shown in Lemma 11.7. Then by Theorem 11.4 $\sqrt{n}R_n|\mathcal{F}_{3,n} \Rightarrow \mathcal{N}(0, V_2)$ with $V_2 = E[(q_i^* p_i^*)^{-1}\sigma_1^2(\psi_i) + (q_i^*(1-p_i^*))^{-1}\sigma_0^2(\psi_i)]$ by Theorem 11.4. Then by Lemma 11.11, $\sqrt{n}(A_n + R_n) \Rightarrow \mathcal{N}(0, V_1 + V_2)$. Applying Lemma 11.16 with $h_n = 1$ shows that $\mathrm{Var}(E_n[(T_i - \widehat{q}_{i,n})]|\mathcal{F}_{1,n}) \le n^{-1}\overline{k}_n|L_n|$. Then $E_n[T_i] = E_n[\widehat{q}_{i,n}] + O_p((\overline{k}_n|L_n|/n)^{1/2})$ and $E_n[\widehat{q}_{i,n}] = E_n[\widehat{q}_i] + O(1/\overline{k}_n)$, since $|\widehat{q} - \widehat{q}_n|_\infty = O(1/\overline{k}_n)$ by discretization. Finally, $E_n[\widehat{q}_i] = E_n[q_i^*] + o_p(1)$ by Lemma 11.7. Putting this all together, we have $E_n[T_i] = E_n[q_i^*] + o_p(1) = E[q_i^*] + o_p(1)$, so that $\sqrt{n_T}(\widehat{\theta} - \mathrm{ATE}) = \sqrt{n}(E_n[T_i])^{1/2}(\widehat{\theta} - \mathrm{ATE}) = E[q^*(\psi_i)]^{1/2}\sqrt{n}(\widehat{\theta} - \mathrm{ATE}) + o_p(1) \Rightarrow N(0, E[q^*(\psi_i)] \cdot V)$ by continuous mapping theorem and Slutsky. This finishes the proof. The conclusion for Theorem 4.4 follows by setting $\widehat{q} = q^*$ and $\widehat{q}_n = q_n^*$, so that the error T_{n2} above is identically zero. □

Proof of Theorem 4.5. Fix $P \in \mathcal{P}_{1/2}$. By the fundamental expansion of the IPW estimator, $\widehat{\theta} = E_n[c(X_i)] + 2E_n[(D_i - 1/2)b(X_i)] + 2E_n[D_i\epsilon_i^1 + (1-D_i)\epsilon_i^0] \equiv A_n + B_n + C_n$. Then by the law of total variance

$$\begin{aligned}&\mathrm{Var}(\widehat{\theta}|X_{1:n}) = \mathrm{Var}(B_n + C_n|X_{1:n}) = E[\mathrm{Var}(B_n + C_n|X_{1:n}, D_{1:n})|X_{1:n}]\\&+ \mathrm{Var}(E[B_n + C_n|X_{1:n}, D_{1:n}]|X_{1:n}) = E[\mathrm{Var}(C_n|X_{1:n}, D_{1:n})|X_{1:n}] + \mathrm{Var}(B_n|X_{1:n})\end{aligned}$$

The last line follows since $E[C_n|X_{1:n}, D_{1:n}] = 0$. One can show $E[\mathrm{Var}(C_n|X_{1:n}, D_{1:n})|X_{1:n}] = (2/n)E_n[\sigma_1^2(X_i) + \sigma_0^2(X_i)]$, which does not depend on P. Then $\mathrm{argmin}_{P\in\mathcal{P}_{1/2}} \mathrm{Var}(\widehat{\theta}|X_{1:n}) = \mathrm{argmin}_{P\in\mathcal{P}_{1/2}} \mathrm{Var}(B_n|X_{1:n})$. Denote $Z_i = D_i - 1/2$. Then $n^2\,\mathrm{Var}(B_n|X_{1:n})$ is equal to

$$E[(Z_{1:n}'b_{1:n})^2] = \sum_{d_{1:n}\in\{0,1\}^n} P(d_{1:n}|X_{1:n})(Z_{1:n}'b_{1:n})^2 \ge \min_{d_{1:n}\in\{0,1\}^n}((D_i - 1/2)_{1:n}'b_{1:n})^2$$

Let $d_{1:n}^* = d_{1:n}^*(X_{1:n})$ be a vector achieving the RHS minimum. Define $P^* \in \mathcal{P}_{1/2}$ by $P^*(d_{1:n}^*|X_{1:n}) = P^*(1 - d_{1:n}^*|X_{1:n}) = 1/2$. Then

$$\begin{aligned}n^2 \min_{P\in\mathcal{P}_{1/2}} \mathrm{Var}(B_n|X_{1:n}) &\ge ((d_{1:n}^* - (1/2)\mathbb{1}_n)'b_{1:n})^2 = E_{P^*}[(D_{1:n} - (1/2)\mathbb{1}_n)'b_{1:n})^2]\\&= n^2\,\mathrm{Var}_{P^*}(B_n|X_{1:n}) \ge n^2 \min_{P\in\mathcal{P}_{1/2}} \mathrm{Var}(B_n|X_{1:n})\end{aligned}$$

The final equality since $P^* \in \mathcal{P}_{1/2}$ by construction. Then equality holds throughout, and $P^* \in \mathrm{argmin}_{P\in\mathcal{P}_{1/2}} \mathrm{Var}(B_n|X_{1:n}) = \mathrm{argmin}_{P\in\mathcal{P}_{1/2}} \mathrm{Var}(\widehat{\theta}|X_{1:n})$. Next, we characterize the optimal vector $d_{1:n}^*$. Observe that $(d_i - 1/2)(d_j - 1/2) = (d_i - 1/2)(d_j - 1/2) - 1/4 + 1/4 = -(1/2)\mathbb{1}(d_i \ne d_j) + 1/4$. Then $\mathrm{argmin}_{d_{1:n}\in\{0,1\}^n}((d_{1:n} - (1/2)\mathbb{1}_n)'b_{1:n})^2$ is equal to

$$\mathop{\mathrm{argmin}}_{d_{1:n}\in\{0,1\}^n} \sum_{i,j=1}^n (d_i - 1/2)(d_j - 1/2)b_ib_j = \mathop{\mathrm{argmax}}_{d_{1:n}\in\{0,1\}^n} \sum_{i\ne j}^n \mathbb{1}(d_i \ne d_j)b_ib_j$$

The latter problem is equivalent to Max-Cut with edge weights $w_{ij} = b_ib_j$, as claimed. □

Lemma 11.7 (Propensity Convergence). *Suppose that $|\widehat{\sigma}_d^2 - \sigma_d^2|_{2,\psi}^2 = O_p(n^{-r})$ for $d = 0, 1$. Then $|\widehat{p}_{i,n} - p_i^*|_{2,n}^2 \vee |\widehat{q}_{i,n} - q_i^*|_{2,n}^2 = O(1/\overline{k}_n^2) + O_p(n^{-r})$.*

Proof. We have $|\widehat{p}_{i,n} - p_i^*|_{2,n}^2 \le 2|\widehat{p}_{i,n} - \widehat{p}_i|_{2,n}^2 + 2|\widehat{p}_i - p_i^*|_{2,n}^2 \le O(1/\overline{k}_n^2) + |\widehat{p}_i - p_i^*|_{2,n}^2$ by discretization, and similarly for $|\widehat{q}_{i,n} - q_i^*|_{2,n}^2$. Then it suffices to bound $|\widehat{p}_i - p_i^*|_{2,n}^2$. Note

$$|\widehat{\sigma}_d - \sigma_d|_{2,\psi}^2 = \int_\psi \frac{(\widehat{\sigma}_d^2 - \sigma_d^2)^2(\psi)}{(\widehat{\sigma}_d + \sigma_d)^2(\psi)} dP(\psi) \le c_l^{-1} |\widehat{\sigma}_d^2 - \sigma_d^2|_{2,\psi}^2 = O_p(n^{-r}).$$

Then $|\widehat{p}_i - p_i^*|_{2,n}^2$ is equal to

$$\begin{aligned} &E_n[(\widehat{\sigma}_{1i}/(\widehat{\sigma}_{1i} + \widehat{\sigma}_{0i}) - \sigma_{1i}/(\sigma_{1i} + \sigma_{0i}))^2] \le 4c_l^{-2} E_n[((\sigma_{1i} - \widehat{\sigma}_{1i})\widehat{\sigma}_{0i} + \widehat{\sigma}_{1i}(\widehat{\sigma}_{0i} - \sigma_{0i}))^2] \\ &\lesssim E_n[(\sigma_{1i} - \widehat{\sigma}_{1i})^2 \widehat{\sigma}_{0i}^2] + E_n[\widehat{\sigma}_{1i}^2(\widehat{\sigma}_{0i} - \sigma_{0i}))^2] \lesssim |\widehat{\sigma}_0 - \sigma_0|_{2,n}^2 + |\widehat{\sigma}_1 - \sigma_1|_{2,n}^2 \end{aligned}$$

The last expression is equal to $\sum_{d=0,1} |\widehat{\sigma}_d - \sigma_d|_{2,\psi}^2 + O_p(n^{-1/2}) = O_p(n^{-r}) + O_p(n^{-1/2})$ by conditional Markov inequality. Next consider $|\widehat{q} - q^*|_{2,n}^2$. Denote $C_i = C(\psi_i)$. Define $\mu_n = E_n[C_i^{1/2}(\widehat{\sigma}_{1i} + \widehat{\sigma}_{0i})]$ and $\mu = E[C_i^{1/2}(\sigma_{1i} + \sigma_{0i})]$. Using $\xi_n \perp\!\!\!\perp W_{1:n}$, by conditional Chebyshev $E_n[C_i^{1/2}\widehat{\sigma}_{di}] = E[C_i^{1/2}\widehat{\sigma}_{di}|\xi_n] + O_p(n^{-1/2})$. Then we have

$$\begin{aligned} &(\mu_n - \mu)^2 \lesssim \sum_{d=0,1} (E_n[C_i^{1/2}\widehat{\sigma}_{di}] - E[C_i^{1/2}\widehat{\sigma}_{di}|\xi_n] + E[C_i^{1/2}\widehat{\sigma}_{di}|\xi_n] - E[C_i^{1/2}\sigma_{di}])^2 \\ &\lesssim \sum_{d=0,1} (E_n[C_i^{1/2}\widehat{\sigma}_{di}] - E[C_i^{1/2}\widehat{\sigma}_{di}|\xi_n])^2 + |C|_\infty E_\psi[(\widehat{\sigma}_{di} - \sigma_{di})^2] = O_p(n^{-1}) + O_p(n^{-r}) \end{aligned}$$

The first inequality is Young's, the second by Young's and Jensen. Define $\widehat{v}_i = C_i^{-1/2}(\widehat{\sigma}_{1i} + \widehat{\sigma}_{0i})$ and $v_i = C_i^{-1/2}(\sigma_{1i} + \sigma_{0i})$.

$$\begin{aligned} |\widehat{q} - q^*|_{2,n}^2 &= \overline{q}^2 E_n[(\widehat{v}_i/\mu_n - v_i/\mu)^2] \lesssim (\mu_n \mu)^{-2}(E_n[\widehat{v}_i^2](\mu - \mu_n) + \mu_n E_n[(\widehat{v}_i - v_i)^2]) \\ &\lesssim (\mu - \mu_n)^2 + E_n[(\widehat{v}_i - v_i)^2] = O_p(n^{-r}) + O_p(n^{-1/2}) \end{aligned}$$

The first inequality is Young's, the second using our bounded variance assumption. The final equality follows from work above. This finishes the proof. □

11.4 Inference

Proof of Theorem 6.1. First, consider the sample variance. The second moment is

$$E_n\left[\left(\frac{T_i(D_i - p(\psi_i))Y_i}{q(\psi_i)(p - p^2)(\psi_i)}\right)^2\right] = E_n\left[\frac{T_i D_i Y_i(1)^2}{p_i^2 q_i^2}\right] + E_n\left[\frac{T_i(1 - D_i)Y_i(0)^2}{(1 - p_i)^2 q_i^2}\right]$$

Consider the first term

$$\begin{aligned} &E_n[T_i D_i Y_i(1)^2/(p_i^2 q_i^2)] = E_n[T_i(D_i - p_i)Y_i(1)^2/(p_i^2 q_i^2)] + E_n[(T_i - q_i)Y_i(1)^2/(p_i q_i^2)] \\ &+ E_n[Y_i(1)^2/(p_i q_i)] = E_n[Y_i(1)^2/(p_i q_i)] + o_p(1) = E[Y_i(1)^2/(p_i q_i)] + o_p(1) \end{aligned}$$

The first equality is by Lemma 11.16, and the second by WLLN. Then by symmetry $E_n[T_i(1-D_i)Y_i(0)^2/(1-p_i)^2 q_i^2] = E[Y_i(0)^2/(1-p_i)q_i] + o_p(1)$. Then the sample variance in the theorem converges in probability to $v = E[Y_i(1)^2/(p_i q_i)] + E[Y_i(0)^2/(1-p_i)q_i] - \text{ATE}^2$.

With V the limiting variance of Theorem 3.11, simple algebra shows that

$$v - V = E[m_{1i}^2(1 - p_iq_i)/p_iq_i] + E[m_{0i}^2(1 - q_i(1 - p_i))/(q_i(1 - p_i))] + 2E[m_{1i}m_{0i}].$$

Then the conclusion follows from the consistency results in Lemma 11.8. □

Lemma 11.8 (Inference)**.** *Suppose that $E[Y(d)^2] < \infty$ for $d = 0, 1$. Replace w_i^d in the variance estimators in Section 6 with a bounded non-negative function $a_i = a(\psi_i)$. Then we have $\widehat{v}_1 \xrightarrow{p} E[q_ip_im_{1i}^2a_i]$, $\widehat{v}_0 \xrightarrow{p} E[q_i(1 - p_i)m_{0i}^2a_i]$, and $\widehat{v}_{10} \xrightarrow{p} E[q_im_{1i}m_{0i}a_i]$.*

Proof. First, we show the homogeneity condition $n^{-1}\sum_{g\in\mathcal{G}_n^\nu}\sum_{i,j\in g}|\psi_i - \psi_j|_2^2 = o(1)$ also holds for the matched groups $\mathcal{G}_n^\nu = \{g \cup \nu(g) : g \in \mathcal{G}_n\}$. Note that we may write

$$n^{-1}\sum_{g\in\mathcal{G}_n^\nu}\sum_{i,j\in g}|\psi_i - \psi_j|_2^2 = (2n)^{-1}\sum_{g\in\mathcal{G}_n}\sum_{i,j\in g\cup\nu(g)}|\psi_i - \psi_j|_2^2$$
$$= (2n)^{-1}\sum_{g\in\mathcal{G}_n}\sum_{i\in g, j\in\nu(g)}|\psi_i - \psi_j|_2^2 + (2n)^{-1}\sum_{g\in\mathcal{G}_n}\sum_{i\neq j}|\psi_i - \psi_j|_2^2\mathbb{1}(ij \in g \text{ or } ij \in \nu(g))$$

The second term is $o(1)$ by Theorem 11.1. By Young's inequality, the first term is

$$\lesssim n^{-1}\sum_{g\in\mathcal{G}_n}\sum_{i\in g, j\in\nu(g)}|\psi_i - \bar{\psi}_g|_2^2 + |\bar{\psi}_g - \bar{\psi}_{\nu(g)}|_2^2 + |\bar{\psi}_{\nu(g)} - \psi_j|_2^2$$
$$\leq 2n^{-1}\sum_{g\in\mathcal{G}_n}k(g)\sum_{i\in g}|\psi_i - \bar{\psi}_g|_2^2 + \bar{k}^2n^{-1}\sum_{g\in\mathcal{G}_n}|\bar{\psi}_g - \bar{\psi}_{\nu(g)}|_2^2$$

The second term is $o(1)$ again by Theorem 11.1. The first term is

$$2n^{-1}\sum_{g\in\mathcal{G}_n}k(g)\sum_{i\in g}|k(g)^{-1}\sum_{j\in g}(\psi_i - \psi_j)|_2^2 \leq 2n^{-1}\sum_{g\in\mathcal{G}_n}\sum_{i,j\in g}|\psi_i - \psi_j|_2^2 = o(1)$$

again by triangle inequality, Jensen, and Theorem 11.1. This finishes our proof of the claim.

Denote $v_i = m_1(\psi_i)\sqrt{a(\psi_i)}$. By Lemma 11.10, there exists a sequence $(z_n)_{n\geq1}$ of functions with $|z_n|_{lip} \leq c_n$ and $|z_n - m_1(\psi)^2a(\psi)|_{2,\psi} = o(1)$. Observe that $|z_n|_{2,\psi} = O(1)$ since $m_1(\psi)^2a(\psi) \in L_2(\psi)$. For matched groups $g = g_1 \sqcup g_2$ with assignment propensities a_1/k_1 and a_2/k_2 in each group, define the group weight $w_g = 1/(a_1 + a_2 - 1)$ and variance estimator

$$\widehat{v}_1 = n^{-1}\sum_{g\in\mathcal{G}_n^\nu}w_g\sum_{i,j\in g}Y_i(1)Y_j(1)(a_ia_j)^{1/2}D_iD_j\mathbb{1}(i \neq j)$$
$$= n^{-1}\sum_{g\in\mathcal{G}_n^\nu}w_g\sum_{\substack{i,j\in g\\ i\neq j}}(m_1(\psi_i) + \epsilon_i^1)(m_1(\psi_j) + \epsilon_j^1)(a_ia_j)^{1/2}D_iD_j = A_n + B_n + C_n + R_n.$$

Consider the first term A_n. Noting that $ab = -(1/2)[(a-b)^2 - a^2 - b^2]$, we have

$$A_n = n^{-1} \sum_{g \in \mathcal{G}_n^\nu} w_g \sum_{i,j \in g} v_i v_j D_i D_j \mathbb{1}(i \neq j) = -(2n)^{-1} \sum_{g \in \mathcal{G}_n^\nu} w_g \sum_{i,j \in g} (v_i - v_j)^2 D_i D_j$$
$$+ (2n)^{-1} \sum_{g \in \mathcal{G}_n^\nu} w_g \sum_{i,j \in g} (v_i^2 + v_j^2) D_i D_j \mathbb{1}(i \neq j) \equiv A_n^1 + A_n^2$$

Denote $z_{in} = z_n(\psi_i)$ and let $\bar{k} = \max_{g \in \mathcal{G}} |g|$. Observe that by Jensen's inequality

$$|A_n^1| \leq (2n)^{-1} \sum_{g \in \mathcal{G}_n^\nu} \sum_{\substack{i,j \in g \\ i \neq j}} (v_i - v_j)^2 = (2n)^{-1} \sum_{g \in \mathcal{G}_n^\nu} \sum_{\substack{i,j \in g \\ i \neq j}} (v_i - z_{in} + z_{in} - z_{jn} + z_{jn} - v_j)^2$$
$$\leq 3(2n)^{-1} \sum_{g \in \mathcal{G}_n^\nu} \sum_{\substack{i,j \in g \\ i \neq j}} |v_i - z_{in}|^2 + |z_{in} - z_{jn}|^2 + |z_{jn} - v_j|^2$$
$$\lesssim n^{-1} \sum_{g \in \mathcal{G}_n^\nu} (|g| - 1) \sum_{i \in g} |v_i - z_{in}|^2 + c_n^2 \cdot n^{-1} \sum_{g \in \mathcal{G}_n^\nu} \sum_{i,j \in g} |\psi_i - \psi_j|_2^2$$
$$\leq (\bar{k} - 1) n^{-1} E_n[T_i |v_i - z_{in}|^2] + o_p(1) \lesssim E_n[|v_i - z_{in}|^2] + o_p(1) = o_p(1)$$

For the second to last inequality, set $c_n = O(r_n^{-2})$ for $n^{-1} \sum_{g \in \mathcal{G}_n^\nu} \sum_{i,j \in g} |\psi_i - \psi_j|_2^2 = o(r_n)$, as guaranteed by the fact above. For the final equality, note that $E[E_n[|v_i - z_{in}|^2]] = |v_i - z_{in}|_{2,\psi}^2 = o(1)$, so that $E_n[|v_i - z_{in}|^2] = o_p(1)$ by Markov inequality. Next,

$$A_n^2 = (2n)^{-1} \sum_{g \in \mathcal{G}_n^\nu} w_g \sum_{\substack{i,j \in g \\ i \neq j}} (v_i^2 + v_j^2) D_i D_j = 2 \cdot (2n)^{-1} \sum_{g \in \mathcal{G}_n^\nu} w_g \sum_{i \in g} v_i^2 D_i \sum_{\substack{j \in g \\ j \neq i}} D_j$$
$$= 2 \cdot (2n)^{-1} \sum_{g \in \mathcal{G}_n^\nu} \sum_{i \in g} v_i^2 D_i w_g (a_1 + a_2 - 1) = n^{-1} \sum_{g \in \mathcal{G}_n^\nu} \sum_{i \in g} v_i^2 D_i = E_n[v_i^2 D_i T_i]$$

Finally, note that $E_n[v_i^2 D_i T_i] = E[v_i^2 p_i q_i] + o_p(1)$ by Lemma 11.16. Next we claim that $R_n = o_p(1)$. It suffices to verify the conditions of Lemma 11.12 for

$$u_g = w_g \sum_{i,j \in g} \epsilon_i^1 \epsilon_j^1 (a_i a_j)^{1/2} D_i D_j \mathbb{1}(i \neq j) = w_g \sum_{i \neq j} \epsilon_i^1 \epsilon_j^1 (a_i a_j)^{1/2} D_i D_j \mathbb{1}(ij \in g)$$

Define $\mathcal{F}_n = \sigma(\psi_{1:n}, \tau^t, \pi_n)$, so that $\mathcal{G}_n^\nu, D_{1:n} \in \mathcal{F}_n$. Note that $u_g \perp\!\!\!\perp u_{g'} | \mathcal{F}_n$ for $g \neq g'$ by Lemma 11.14. We also need to show $E[u_g | \mathcal{F}_n] = 0$ for each $g \in \mathcal{G}_n^\nu$. Note that $E[u_g | \mathcal{F}_n] = E[E[u_g | \mathcal{F}_n'] \mathcal{F}_n]$ with $\mathcal{F}_n' = \sigma(\mathcal{F}_n, \tau^d)$ and $D_{1:n} \in \mathcal{F}_n'$. Repeatedly using the fact that $(A, B) \perp\!\!\!\perp C \implies A \perp\!\!\!\perp C | B$, we have $E[\epsilon_i^1 \epsilon_j^1 | \mathcal{F}_n'] = E[\epsilon_i^1 \epsilon_j^1 | \psi_i, \psi_j] = E[\epsilon_j^1 E[\epsilon_i^1 | \epsilon_j^1, \psi_i, \psi_j] | \psi_i, \psi_j] = E[\epsilon_j^1 E[\epsilon_i^1 | \psi_i] | \psi_i, \psi_j] = 0$. Then apparently $E[u_g | \mathcal{F}_n] = 0$ for all $g \in \mathcal{G}_n^\nu$. Finally, Lemma 11.12 requires the condition $\frac{1}{n} \sum_{g \in \mathcal{G}_n^\nu} E[|u_g| \mathbb{1}(|u_g| > c_n) | \mathcal{F}_n] = o_p(1)$. Observe that for any positive constants $(a_k)_{k=1}^m$ we have $\sum_k a_k \mathbb{1}(\sum_k a_k > c) \leq m \sum_k a_k \mathbb{1}(a_k > c/m)$ and $ab\mathbb{1}(ab > c) \leq a^2 \mathbb{1}(a^2 > c) + b^2 \mathbb{1}(b^2 > c)$. Let $c_n \to \infty$ and let $\bar{k} = \max_{g \in \mathcal{G}_n^\nu} |g|$. Using the indicator function facts above, for any group $g \in \mathcal{G}_n^\nu$,

$E[|u_g|\mathbb{1}(|u_g| > c_n)|\mathcal{F}_n]$ is bounded above by

$$
\begin{aligned}
&E\left[\sum_{i,j\in g} |\epsilon_i^1||\epsilon_j^1|(a_i a_j)^{1/2}\mathbb{1}(i \neq j)\mathbb{1}\left(\sum_{i,j\in g} |\epsilon_i^1||\epsilon_j^1|(a_i a_j)^{1/2}\mathbb{1}(i\neq j) > c_n\right)\middle|\mathcal{F}_n\right] \\
&\leq \bar{k}^2 \sum_{\substack{i,j\in g\\ i\neq j}} E\left[|\epsilon_i^1||\epsilon_j^1|(a_i a_j)^{1/2}\mathbb{1}\left(|\epsilon_i^1||\epsilon_j^1|(a_i a_j)^{1/2} > c_n/\bar{k}^2\right)\middle|\mathcal{F}_n\right] \\
&\leq 2\bar{k}^3 \sum_{i\in g} E\left[a_i(\epsilon_i^1)^2\mathbb{1}\left(a_i(\epsilon_i^1)^2 > c_n/\bar{k}^2\right)\middle|\mathcal{F}_n\right] = 2\bar{k}^3 \sum_{i\in g} E\left[a_i(\epsilon_i^1)^2\mathbb{1}\left(a_i(\epsilon_i^1)^2 > c_n/\bar{k}^2\right)\middle|\psi_i\right]
\end{aligned}
$$

The first line by triangle inequality and since $w_g \leq 1$. The second and third inequalities use the fact about indicator functions above. The last line again uses the fact that $(A, B) \perp\!\!\!\perp C \implies A \perp\!\!\!\perp C|B$, combined with iid sampling and independence of (τ^t, π_n) from the data $W_{1:n}$. Then since $\mathcal{G}_n^\nu$ is a partition of $\{T_i = 1\} \subseteq [n]$, we have

$$\frac{1}{n}\sum_{g\in\mathcal{G}_n^\nu} E[|u_g|\mathbb{1}(|u_g| > c_n)|\mathcal{F}_n] \lesssim \frac{2\bar{k}^3}{n}\sum_{i=1}^n E\left[a_i(\epsilon_i^1)^2\mathbb{1}\left(a_i(\epsilon_i^1)^2 > c_n/\bar{k}^2\right)\middle|\psi_i\right]$$

Taking an expectation of the RHS gives $2\bar{k}^3 E[a_i(\epsilon_i^1)^2\mathbb{1}\left(a_i(\epsilon_i^1)^2 > c_n/\bar{k}^2\right)] = o(1)$ as $n \to \infty$ since $E[a_i(\epsilon_i^1)^2] \leq |a|_\infty E[\mathrm{Var}(Y(1)|\psi)] \lesssim \mathrm{Var}(Y(1)) < \infty$. Then by Markov inequality $\frac{1}{n}\sum_{g\in\mathcal{G}_n^\nu} E[|u_g|\mathbb{1}(|u_g| > c_n)|\mathcal{F}_n] = o_p(1)$, so $R_n = o_p(1)$ by Lemma 11.12. An identical argument shows that $B_n, C_n = o_p(1)$. Then we have shown that $\widehat{v}_1 = E[v_i^2 p_i q_i] + o_p(1)$. By symmetry, $\widehat{v}_0 = E[v_i^2(1-p_i)q_i] + o_p(1)$.

Next, consider the cross-term estimator.

$$
\begin{aligned}
\widehat{v}_{10} &= n^{-1}\sum_{g\in\mathcal{G}_n} w_g \sum_{i,j\in g} Y_i(1)Y_j(0)(a_i a_j)^{1/2} D_i(1-D_j) \\
&= n^{-1}\sum_{g\in\mathcal{G}_n} w_g \sum_{i,j\in g} (m_{1i} + \epsilon_i^1)(m_{0j} + \epsilon_j^0)(a_i a_j)^{1/2} D_i(1-D_j) = A_n + B_n + C_n + R_n
\end{aligned}
$$

Denote $v_i^d = m_{di} a_i^{1/2}$ and let $\mathcal{F}_n = \sigma(\psi_{1:n}, \tau^t, \pi_n)$ as before. Note the equality $a_i b_j + b_i a_j = -(a_i - a_j)(b_i - b_j) + a_i b_i + a_j b_j$. Then the conditional expectation $E[A_n|\mathcal{F}_n]$ is equal to

$$n^{-1}\sum_{g\in\mathcal{G}_n} w_g \sum_{i,j\in g} E[v_i^1 v_j^0 D_i(1-D_j)|\mathcal{F}_n] = n^{-1}\sum_{g\in\mathcal{G}_n} w_g \sum_{i,j\in g} v_i^1 v_j^0 \frac{a(k-a)}{k(k-1)}\mathbb{1}(i\neq j)$$

Expanding using the fact above gives

$$
\begin{aligned}
&= n^{-1}\sum_{g\in\mathcal{G}_n} w_g \frac{a(k-a)}{k(k-1)} \sum_{i<j\in g}(v_i^1 v_j^0 + v_j^1 v_i^0) = -n^{-1}\sum_{g\in\mathcal{G}_n} w_g \frac{a(k-a)}{k(k-1)} \sum_{i<j\in g}(v_i^1 - v_j^1)(v_i^0 - v_j^0) \\
&+ n^{-1}\sum_{g\in\mathcal{G}_n} w_g \frac{a(k-a)}{k(k-1)} \sum_{i<j\in g}(v_i^1 v_i^0 + v_j^1 v_j^0) \equiv T_{n1} + T_{n2}
\end{aligned}
$$

First consider T_{n1}. By Young's inequality we have

$$|T_{n1}| \lesssim n^{-1} \sum_{g\in\mathcal{G}_n} \sum_{i<j\in g} |v_i^1 - v_j^1||v_i^0 - v_j^0| \le (2n)^{-1} \sum_{g\in\mathcal{G}_n} \sum_{i,j\in g} |v_i^1 - v_j^1|^2 + |v_i^0 - v_j^0|^2$$

The form of each expression is identical to our analysis of A_n^1 above. Then $T_{n1} = o_p(1)$ by the same argument. Next consider T_{n2}. By counting, it's easy to see that $\sum_{i<j\in g}(v_i^1 v_i^0 + v_j^1 v_j^0) = (k-1)\sum_{i\in g} v_i^1 v_i^0$. Then setting $w_g = k/(a(k-a))$ gives

$$\begin{aligned} T_{n2} &= n^{-1} \sum_{g\in\mathcal{G}_n} w_g \frac{a(k-a)}{k(k-1)}(k-1) \sum_{i\in g} v_i^1 v_i^0 = n^{-1} \sum_{g\in\mathcal{G}_n} \sum_{i\in g} v_i^1 v_i^0 \\ &= E_n[T_i v_i^1 v_i^0] = E_n[q_i v_i^1 v_i^0] + E_n[(T_i - q_i) v_i^1 v_i^0] = E[q_i m_{1i} m_{0i} a_i] + o_p(1) \end{aligned}$$

The third equality since $\mathcal{G}_n$ partitions the units $\{T_i = 1\}$. The final equality follows from Lemma 11.16. Next consider

$$A_n - E[A_n|\mathcal{F}_n] = n^{-1} \sum_{g\in\mathcal{G}_n} \sum_{i,j\in g} w_g v_i^1 v_j^0 \left(D_i(1-D_j) - \frac{a(k-a)}{k(k-1)} \right) \mathbb{1}(i \ne j) \equiv n^{-1} \sum_{g\in\mathcal{G}_n} u_g$$

We have $E[u_g|\mathcal{F}_n] = 0$ for all $g \in \mathcal{G}_n$ by the calculation of $E[A_n|\mathcal{F}_n]$ above. Also, note $u_g \perp\!\!\!\perp u_{g'}|\mathcal{F}_n$ for all $g \ne g'$ by Lemma 11.14. Then by Lemma 11.12, it suffices to show that $\frac{1}{n}\sum_{g\in\mathcal{G}_n^v} E[|u_g|\mathbb{1}(|u_g| > c_n)|\mathcal{F}_n] = o_p(1)$ for some $c_n \to \infty$ with $c_n = o(n^{1/2})$. It's easy to see that $w_g \le 2$, so we have

$$|u_g| \le \sum_{i,j\in g} w_g |v_i^1||v_j^0| \left| D_i(1-D_j) - \frac{a(k-a)}{k(k-1)} \right| \mathbb{1}(i \ne j) \le \sum_{i,j\in g} |v_i^1||v_j^0|$$

Then using the indicator function bounds in our analysis above, $E[|u_g|\mathbb{1}(|u_g| > c_n)|\mathcal{F}_n]$ is bounded above by

$$\begin{aligned} &E\left[\sum_{i,j\in g} |v_i^1||v_j^0| \mathbb{1}\left(\sum_{i,j\in g} |v_i^1||v_j^0| > c_n \right) |\mathcal{F}_n \right] \le \bar{k}^2 \sum_{i,j\in g} E\left[|v_i^1||v_j^0| \mathbb{1}\left(|v_i^1||v_j^0| > c_n/(\bar{k})^2 \right) |\mathcal{F}_n \right] \\ &\le \bar{k}^3 \sum_{i\in g} \left(E\left[|v_i^1|^2 \mathbb{1}\left(|v_i^1|^2 > c_n/(\bar{k})^2 \right) |\mathcal{F}_n \right] + E\left[|v_i^0|^2 \mathbb{1}\left(|v_i^0|^2 > c_n/(\bar{k})^2 \right) |\mathcal{F}_n \right] \right) \\ &\le \bar{k}^3 \sum_{i\in g} \left(E\left[|v_i^1|^2 \mathbb{1}\left(|v_i^1|^2 > c_n/(\bar{k})^2 \right) |\psi_i \right] + E\left[|v_i^0|^2 \mathbb{1}\left(|v_i^0|^2 > c_n/(\bar{k})^2 \right) |\psi_i \right] \right) \end{aligned}$$

Then $\frac{1}{n}\sum_{g\in\mathcal{G}_n^v} E[|u_g|\mathbb{1}(|u_g| > c_n)|\mathcal{F}_n]$ is bounded above by

$$\begin{aligned} &\bar{k}^3 \frac{1}{n} \sum_{g\in\mathcal{G}_n^v} \sum_{i\in g} \left(E\left[|v_i^1|^2 \mathbb{1}\left(|v_i^1|^2 > c_n/(\bar{k})^2 \right) |\psi_i \right] + E\left[|v_i^0|^2 \mathbb{1}\left(|v_i^0|^2 > c_n/(\bar{k})^2 \right) |\psi_i \right] \right) \\ &= \bar{k}^3 E_n[T_i E[|v_i^1|^2 \mathbb{1}(|v_i^1|^2 > c_n/(\bar{k})^2)|\psi_i]] + E_n[T_i E[|v_i^0|^2 \mathbb{1}(|v_i^0|^2 > c_n/(\bar{k})^2)|\psi_i]] \\ &\le \bar{k}^3 E_n[E[|v_i^1|^2 \mathbb{1}(|v_i^1|^2 > c_n/(\bar{k})^2)|\psi_i]] + E_n[E[|v_i^0|^2 \mathbb{1}(|v_i^0|^2 > c_n/(\bar{k})^2)|\psi_i]] \end{aligned}$$

Taking expectation of the final line gives $\sum_{d=0,1} E[|v_i^d|^2 \mathbb{1}(|v_i^d|^2 > c_n/(\bar{k})^2)] \to 0$ as $n \to \infty$ by dominated convergence. Then by Markov inequality, we have shown that $\frac{1}{n}\sum_{g\in\mathcal{G}_n} E[|u_g|\mathbb{1}(|u_g| > c_n)|\mathcal{F}_n]$, so $A_n - E[A_n|\mathcal{F}_n] = o_p(1)$ by Lemma 11.12. Next, note that $B_n, C_n, R_n = o_p(1)$ by the exact argument used in our analysis of $\widehat{v}_1$. Then we have shown that $\widehat{v}_{10} \xrightarrow{p} E[q_i m_{1i} m_{0i} a_i]$ as claimed. $\square$

11.5 Regression Equivalence Propositions

Proof of Proposition 3.6. First, consider the regression $Y \sim 1 + D + \tilde{z} + D\tilde{z}$ defining $\widehat{\tau}$ and $\widehat{\beta}$. Define the within-arm regression coefficients $\widehat{\alpha}_d = \mathrm{Var}_n(\tilde{z}_i|D_i = d, T_i = 1)^{-1}\,\mathrm{Cov}_n(\tilde{z}_i, Y_i(d)|D_i = d, T_i = 1)$ for $d \in \{0,1\}$. E.g. by the calculations in Lin (2013), we have $\widehat{\beta} = \widehat{\alpha}_1 - \widehat{\alpha}_0$ and

$$\widehat{\tau} = \widehat{\theta} - (\widehat{\alpha}_1/p + \widehat{\alpha}_0/(1-p))\, E_n\left[(D_i - p)\tilde{z}_i|T_i = 1\right]$$

Consider $\widehat{\alpha}_1$, for instance.

$$\mathrm{Var}_n(\tilde{z}_i|D_i = 1, T_i = 1) = \frac{E_n[\tilde{z}_i\tilde{z}_i' D_i T_i]}{E_n[D_i T_i]} - \frac{E_n[\tilde{z}_i D_i T_i]E_n[\tilde{z}_i' D_i T_i]}{E_n[D_i T_i]^2}$$

By Definition 2.1, we have $E_n[D_iT_i] = E_n[(D_i - p)T_i] + pE_n[T_i - q] + pq = pq + O(1/n)$. Similarly, we expand $E_n[\tilde{z}_i\tilde{z}_i'D_iT_i] = E_n[\tilde{z}_i\tilde{z}_i'(D_i - p)T_i] + pE_n[\tilde{z}_i\tilde{z}_i'(T_i - q)] + pqE_n[\tilde{z}_i\tilde{z}_i']$ and $E_n[\tilde{z}_i'D_iT_i] = E_n[\tilde{z}_i(D_i-p)T_i] + pE_n[\tilde{z}_i(T_i-q)] + pqE_n[\tilde{z}_i]$. It's easy to check that the conditions of Lemma 11.16 are satisfied, so that $E_n[\tilde{z}_i\tilde{z}_i'D_iT_i] = pqE_n[\tilde{z}_i\tilde{z}_i'] + O_p(n^{-1/2})$ and $E_n[\tilde{z}_iD_iT_i] = pqE_n[\tilde{z}_i] + O_p(n^{-1/2})$. Putting these facts together gives $\mathrm{Var}_n(\tilde{z}_i|D_i = 1, T_i = 1) = E_n[\tilde{z}_i\tilde{z}_i'] - E_n[\tilde{z}_i]E_n[\tilde{z}_i'] + O_p(n^{-1/2})$. Similarly, $E_n[z_i|T_i = 1] = E_n[z_i] + O_p(n^{-1/2})$ so

$$\begin{aligned} E_n[\tilde{z}_i\tilde{z}_i'] - E_n[\tilde{z}_i]E_n[\tilde{z}_i'] &= E_n[z_iz_i'] - E_n[z_i]E_n[z_i'|T_i = 1] - E_n[z_i|T_i = 1]E_n[z_i'] \\ &\quad + E_n[z_i|T_i = 1]E_n[z_i'|T_i = 1] = E_n[z_iz_i'] - E_n[z_i]E_n[z_i'] + O_p(n^{-1/2}) \\ &= \mathrm{Var}(z) + O_p(n^{-1/2}) \end{aligned}$$

Similar reasoning shows that $\mathrm{Cov}_n(\tilde{z}_i, Y_i(d)|D_i = d, T_i = 1) = \mathrm{Cov}(z, Y(d)) + O_p(n^{-1/2})$ if $E[Y(d)^2] < \infty$. Note that $\mathrm{Var}(z) \succ 0$ by the assumption that $P(S = k) > 0\ \forall k \in [m]$. Then we have $\widehat{\alpha}_d = \mathrm{Var}(z)^{-1}\,\mathrm{Cov}(z, Y(d)) + O_p(n^{-1/2})$. Recall $z = (\mathbb{1}(S = k))_{k=1}^{m-1}$. For a function $f(X) \in L_1(X)$ define $\gamma(f) = \mathrm{Var}(z)^{-1}\,\mathrm{Cov}(z, f(X))$. Then it's easy to see that $\gamma(f)'z = E[f(X)|S] - E[f(X)|S = m]$. Now by above work $\widehat{\beta} = \widehat{\alpha}_1 - \widehat{\alpha}_0 = \mathrm{Var}(z)^{-1}\,\mathrm{Cov}(z, Y(1) - Y(0)) + O_p(n^{-1/2}) = \gamma(c) + O_p(n^{-1/2})$. Similarly, $\widehat{\alpha}_1/p + \widehat{\alpha}_0/(1 - p) = \mathrm{Var}(z)^{-1}\,\mathrm{Cov}(z, Y(1)/p + Y(0)/(1-p)) + O_p(n^{-1/2}) = \gamma(b)(p-p)^{-1/2} + O_p(n^{-1/2})$. Then by the fundamental expansion of the IPW estimator, we have

$$\begin{aligned} \check{\theta} &= \widehat{\theta} - (\widehat{\alpha}_1/p + \widehat{\alpha}_0/(1-p))\, E_n\left[(D_i - p)\tilde{z}_i|T_i = 1\right] - \widehat{\beta}' E_n\left[\frac{z_i(T_i - q)}{q}\right] + O_p(n^{-1}) \\ &= \widehat{\theta} - \gamma(b)' E_n\left[\frac{(D_i - p)z_i}{q\sqrt{p - p^2}}\right] - \gamma(c)' E_n\left[\frac{z_i(T_i - q)}{q}\right] + O_p(n^{-1}) = E_n\left[c(X_i)\right] \\ &+ E_n\left[\frac{T_i - q}{q}(c(X_i) - \gamma(c)'z_i)\right] + E_n\left[\frac{T_i(D_i - p)}{q\sqrt{p - p^2}}(b(X_i) - \gamma(b)'z_i)\right] + R_n + O_p(n^{-1}) \\ &= E_n\left[c(X_i)\right] + E_n\left[\frac{T_i - q}{q}(c(X_i) - E[c(X_i)|S_i])\right] + E_n\left[\frac{T_i(D_i - p)}{q\sqrt{p - p^2}}(b(X_i) - E[b|S_i])\right] \\ &+ R_n + O_p(n^{-1}) \end{aligned}$$

The first line uses $E_n[z_i|T_i = 1] = E_n[z_i]/q + O(n^{-1})$ and $E_n[(D_i - p)E_n[z_i|T_i = 1]] = O(n^{-1})$ by stratification. The second line uses the probability limits established above and that $E_n[(D_i - p)z_i] = O_p(n^{-1/2})$ and $E_n[(T_i - q)z_i] = O_p(n^{-1/2})$. The final equality follows from the characterization of $\mathrm{Var}(z)^{-1}\,\mathrm{Cov}(z, f(X))$ above and since $E_n[(D_i - p)E[b(X_i)|S_i = m]] = O(n^{-1})$ and $E[(T_i - q)E[c(X_i)|S_i = m]] = O(n^{-1})$ by stratification. The final expansion is identical to that in the proof of Theorem 3.11. The conclusion then follows by the same argument. □

Proof of Proposition 3.12. First we discuss cross-fitting. Denote $Z_i = (W_i, D_i, T_i)$. Let $I_1 \sqcup I_2 = [n]$ be a random partition with $|I_1| \asymp n$ and $(I_1, I_2) \in \sigma(\pi_n)$ for randomness $\pi_n \perp\!\!\!\perp Z_{1:n}$. The estimator has form $E_n[F(Z_i, \widehat{m}_{I(i)})]$. Use the cross-fitting pattern $I(i) = I_1$ if $i \in I_2$ and similarly for $i \in I_1$. We may rearrange the summand of $\widehat{\theta}_{adj}$ as

$$\begin{aligned}
&\frac{T_iD_iY_i}{q_ip_i} - \frac{Y_i(1-D_i)T_i}{q_i(1-p_i)} + \widehat{m}_1(\psi_i) - \widehat{m}_0(\psi_i) - \frac{T_i}{q_i}\left(\frac{D_i\widehat{m}_1(\psi_i)}{p_i} - \frac{(1-D_i)\widehat{m}_0(\psi_i)}{(1-p_i)}\right) \\
&= \frac{T_iD_iY_i}{q_ip_i} - \frac{Y_i(1-D_i)T_i}{q_i(1-p_i)} + \widehat{m}_1(\psi_i) - \widehat{m}_0(\psi_i) - \frac{T_i}{q_i}\left(\widehat{m}_1(\psi_i)) - \widehat{m}_0(\psi_i)\right)) \\
&- \frac{T_i(D_i-p_i)}{q_i}\left(\frac{\widehat{m}_1(\psi_i))}{p_i} - \frac{\widehat{m}_0(\psi_i))}{(1-p_i)}\right)
\end{aligned}$$

Continuing the calculation, this is

$$\begin{aligned}
&= \frac{T_iD_iY_i}{q_ip_i} - \frac{Y_i(1-D_i)T_i}{q_i(1-p_i)} - \frac{(T_i-q_i)}{q_i}\widehat{c}(\psi_i) - \frac{T_i(D_i-p_i)}{q_i(p_i-p_i^2)^{1/2}}\widehat{b}(\psi_i) \\
&= c(\psi_i) + \frac{(T_i-q_i)}{q_i}(c-\widehat{c})(\psi_i) - \frac{T_i(D_i-p_i)}{q_i(p_i-p_i^2)^{1/2}}(b-\widehat{b})(\psi_i) + \frac{D_iT_i\epsilon_i^1}{p_iq_i} + \frac{(1-D_i)T_i\epsilon_i^0}{(1-p_i)q_i}
\end{aligned}$$

Consider the term $T_{1,n} = n^{-1}\sum_{i\in I_1} q_i^{-1}(T_i - q_i)(c - \widehat{c})(\psi_i, Z_{I_2})$. By Lemma 11.14 with $m = 2$, $g = I$ and $h, \tau_s = 1$ we have $Z_{I_1} \perp\!\!\!\perp Z_{I_2}|\pi_n$. Then applying Lemma 11.14

$$\begin{aligned}
E[nT_{1,n}^2|\pi_n, Z_{I_2}] &= n^{-1}\sum_{i\in I_1} E[q_i^{-2}(T_i-q_i)^2(c-\widehat{c})(\psi_i, Z_{I_2})^2|\pi_n, Z_{I_2}] \\
&= n^{-1}\sum_{i\in I_1} E[q_i^{-1}(1-q_i)(c-\widehat{c})(\psi_i, Z_{I_2})^2|\pi_n, Z_{I_2}] \\
&= n^{-1}|I_1|\int(1-q)(\psi)/q(\psi)(c-\widehat{c})(\psi, Z_{I_2})^2dP(\psi) \lesssim |c - \widehat{c}(Z_{I_2})|_{2,\psi} = o_p(1)
\end{aligned}$$

Then $\sqrt{n}T_{1,n} = o_p(1)$. Then by symmetry $\sqrt{n}E_n[q_i^{-1}(T_i - q_i)(c - \widehat{c})(\psi_i)] = o_p(1)$. Similarly, $\sqrt{n}E_n[\frac{T_i(D_i-p_i)}{q_i(p_i-p_i^2)^{1/2}}(b - \widehat{b})(\psi_i)] = o_p(1)$. The claim now follows from vanilla CLT, noting that $n_T/n \xrightarrow{p} E[q_i]$. □

11.6 Lemmas

Lemma 11.9 (Conditional Convergence)**.** *Let $(\mathcal{G}_n)_{n\geq1}$ and $(A_n)_{n\geq1}$ a sequence of σ-algebras and RV's. Define conditional convergence*

$$\begin{aligned}
A_n = o_{p,\mathcal{G}_n}(1) &\iff P(|A_n| > \epsilon|\mathcal{G}_n) = o_p(1) \quad \forall\epsilon > 0 \\
A_n = O_{p,\mathcal{G}_n}(1) &\iff P(|A_n| > s_n|\mathcal{G}_n) = o_p(1) \quad \forall s_n \to \infty
\end{aligned}$$

Then the following results hold

(i) $A_n = o_p(1) \iff A_n = o_{p,\mathcal{G}_n}(1)$ *and* $A_n = O_p(1) \iff A_n = O_{p,\mathcal{G}_n}(1)$

(ii) $E[|A_n||\mathcal{G}_n] = o_p(1)/O_p(1) \implies A_n = o_p(1)/O_p(1)$

(iii) $\mathrm{Var}(A_n|\mathcal{G}_n) = o_p(c_n^2)/O_p(c_n^2) \implies A_n - E[A_n|\mathcal{G}_n] = o_p(c_n)/O_p(c_n)$.

(iv) If $(A_n)_{n\geq 1}$ *has* $A_n \leq \bar{A} < \infty$ $\mathcal{G}_n$*-a.s.* $\forall n$ *and* $A_n = o_p(1) \implies E[|A_n||\mathcal{G}_n] = o_p(1)$

Proof. (i) Consider that for any $\epsilon > 0$

$$P(|A_n| > \epsilon) = E[\mathbb{1}(|A_n| > \epsilon)] = E[E[\mathbb{1}(|A_n| > \epsilon)|\mathcal{G}_n]] = E[P(|A_n| > \epsilon|\mathcal{G}_n)]$$

If $A_n = o_p(1)$, then $E[P(|A_n| > \epsilon|\mathcal{G}_n)] = o(1)$, so $P(|A_n| > \epsilon|\mathcal{G}_n) = o_p(1)$ by Markov inequality. Conversely, if $P(|A_n| > \epsilon|\mathcal{G}_n) = o_p(1)$, then $E[P(|A_n| > \epsilon|\mathcal{G}_n)] = o(1)$ since $(P(|A_n| > \epsilon|\mathcal{G}_n))_{n\geq 1}$ is uniformly bounded, hence UI. Then $P(|A_n| > \epsilon) = o(1)$. The second equivalence follows directly from the first. (ii) follows from (i) and conditional Markov inequality. (iii) is an application of (ii). For (iv), note that for any $\epsilon > 0$

$$E[|A_n||\mathcal{G}_n] \leq \epsilon + E[|A_n|\mathbb{1}(|A_n| > \epsilon)|\mathcal{G}_n] \leq \epsilon + \bar{A}P(|A_n| > \epsilon|\mathcal{G}_n) = \epsilon + o_p(1)$$

The equality is by (i) and our assumption. Since $\epsilon > 0$ was arbitrary $E[|A_n||\mathcal{G}_n] = o_p(1)$. □

Lemma 11.10 (Lipschitz Approximation)**.** *Let* $Z \in \mathbb{R}^d$ *be a random variable. Define* $\mathcal{L} = \{g(Z) \in L_2(Z) : |g|_{lip} \vee |g|_\infty < \infty\}$. *Then* $\mathcal{L}$ *is dense in* $L_2(Z)$.

Proof. Let $\mathbb{1}(Z \in A)$ P-measurable for non-empty A. Define $f_n(z) = (1 + nd(z, A))^{-1}$ with $d(z, A) = \inf_{y\in A} |z - y|_2$. The function $z \to d(z, A)$ is 1-Lipschitz (e.g. reverse triangle inequality). Then $|f_n(z) - f_n(y)| \leq n|z - y|_2$ for any $z, y \in \mathbb{R}^d$ and $|f_n|_\infty \leq 1$ so $f_n(Z) \in \mathcal{L}$. Observe that $f_n(z) \to \mathbb{1}(z \in A)$ pointwise as $n \to \infty$. Then by dominated convergence $\int (f_n(z) - \mathbb{1}(z \in A))^2 dP(z) \to 0$, since the integrand converges pointwise and is dominated by $2 \in L_1(Z)$. Then bounded Lipschitz functions are dense in the set of indicator functions of measurable sets. Next, consider a simple function $\sum_k a_k \mathbb{1}(Z \in A_k)$ with $|a_k| < \infty$ for all k. If $g_{nk}(z) = (1 + nd(z, A_k))^{-1}$ the same argument shows that $\sum_k a_k g_{nk}(Z) - \sum_k a_k \mathbb{1}(Z \in A_k) \to 0$ in $L_2(Z)$. The left hand sum is bounded Lipschitz, showing that the Lipschitz functions are dense in the set of simple functions in $L_2(Z)$. The bounded simple functions are dense in $L_2(Z)$ (e.g. Folland (1999)), so bounded Lipschitz functions are dense in $L_2(Z)$ by transitivity. □

Lemma 11.11 (Asymptotic Independence)**.** *Consider a probability space* $(\Omega, \mathcal{G}, P)$ *with* σ*-algebras* $\mathcal{F} = \mathcal{F}_{n,0}$ *and* $\mathcal{F}_{n,k} \subseteq \mathcal{F}_{n,k+1} \subseteq \mathcal{G}$ *for all* $n \geq 1$ *and* $0 \leq k \leq m - 1$. *Let* $X_{n,k}$ *be* $\mathcal{F}_{n,k}$*-measurable random variables for all* n *and* $1 \leq k \leq m$. *Suppose that* $X_{n,k}|\mathcal{F}_{n,k-1} \Rightarrow L_k|\mathcal{F}$*, as in Definition 11.5. Then* $(X_{n,1}, \dots, X_{n,m})|\mathcal{F} \Rightarrow (L_1, \dots, L_m)|\mathcal{F}$*, with jointly independent limit.*

Proof. By Levy continuity, it suffices to show $E[e^{it'(X_{1,n},\dots X_{m,n})}|\mathcal{F}] \to \prod_{k=1}^m E[e^{it_k L_k}|\mathcal{F}]$ for all $t \in \mathbb{R}^m$. We work by induction on k. By assumption, $E[e^{itX_{n,1}}|\mathcal{F}] \xrightarrow{p} E[e^{itL_1}|\mathcal{F}]$ for all

$t \in \mathbb{R}$. Assume by induction that the conclusion holds for $1 \le k \le k' \le m$. Then

$$\begin{aligned} &E[e^{i\sum_{k=1}^{k'+1} t_k X_{n,k}}|\mathcal{F}] = E[e^{i\sum_{k=1}^{k'} t_k X_{n,k}} E[e^{it_{k'+1}X_{n,k'+1}}|\mathcal{F}_{n,k'}]|\mathcal{F}] \\ &= E[e^{i\sum_{k=1}^{k'} t_k X_{n,k}}(E[e^{it_{k'+1}X_{n,k'+1}}|\mathcal{F}_{n,k'}] - E[e^{it_{k'+1}L_{k'+1}}|\mathcal{F}])|\mathcal{F}] \\ &+ E[e^{it_{k'+1}L_{k'+1}}|\mathcal{F}]E[e^{i\sum_{k=1}^{k'} t_k X_{n,k}}|\mathcal{F}] = \prod_{k=1}^{k'+1} E[e^{it_{k'+1}L_{k'+1}}|\mathcal{F}] + o_p(1) \end{aligned}$$

The first equality is by tower law and our measurability and increasing σ-algebra assumption. For the final equality, note that $|e^{i\sum_{k=1}^{k'} t_k X_{n,k}}(E[e^{it_{k'+1}X_{n,k'+1}}|\mathcal{F}_{n,k'}] - E[e^{it_{k'+1}L_{k'+1}}|\mathcal{F}])| \le |E[e^{it_{k'+1}X_{n,k'+1}}|\mathcal{F}_{n,k'}] - E[e^{it_{k'+1}L_{k'+1}}|\mathcal{F}]| \xrightarrow{p} 0$. Then the first term in the sum above is $o_p(1)$ since the integrand converges in probability and is bounded, hence UI. The final equality also uses our inductive hypothesis. This finishes the proof. □

Lemma 11.12 (LLN). *Consider $A_n = n^{-1}\sum_{g\in\mathcal{G}_n} u_g$, with $\mathcal{G}_n$ a collection of disjoint subsets of $[n]$. Let $(\mathcal{F}_n)_{n\ge1}$ be σ-algebras such that $\mathcal{G}_n$ is $\mathcal{F}_n$-measurable, $E[u_g|\mathcal{F}_n] = 0$, and for all $g \ne g' \in \mathcal{G}_n$ $u_g \perp\!\!\!\perp u_{g'}|\mathcal{F}_n$. If $n^{-1}\sum_{g\in\mathcal{G}_n} E[|u_g|\mathbb{1}(|u_g| > c_n)|\mathcal{F}_n] \xrightarrow{p} 0$ for $c_n = \omega(1)$, $c_n = o(n^{1/2})$, then $A_n \xrightarrow{p} 0$.*

Proof. By disjointness $|\mathcal{G}_n| \le n$. Fix an indexing $\mathcal{G}_n = \{g_s : 1 \le s \le n\}$, possibly with $g_s = \emptyset$ for some s. Define $\bar{u}_{sn} = u_s\mathbb{1}(|u_s| \le c_n)$ and $\bar{\mu}_{sn} = E[u_s\mathbb{1}(|u_s| \le c_n)|\mathcal{F}_n]$. Expand

$$\frac{1}{n}\sum_{g\in\mathcal{G}_n} u_g = \frac{1}{n}\sum_{s=1}^n u_s = \frac{1}{n}\sum_{s=1}^n [(u_s - \bar{u}_{sn}) + (\bar{u}_{sn} - \bar{\mu}_{sn}) + \bar{\mu}_{sn}] = T_{n1} + T_{n2} + T_{n3}$$

Observe that $E[|T_{n1}||\mathcal{F}_n] \le (1/n)\sum_{s=1}^n E[|u_s|\mathbb{1}(|u_s| > c_n)|\mathcal{F}_n] = o_p(1)$ by assumption. Then $T_{n1} = o_p(1)$ by conditional Markov (Lemma 11.9). Next consider T_{n2}. Note that by definition $E[\bar{u}_{sn} - \bar{\mu}_{sn}|\mathcal{F}_n] = 0$ for each $1 \le s \le n$. Note that for $s \ne s'$ $\mathrm{Cov}(\bar{u}_{sn}, \bar{u}_{s'n}|\mathcal{F}_n) = 0$ by the conditional independence assumption. Then $\mathrm{Var}(T_{n2}|\mathcal{F}_n) = n^{-2}\sum_{s=1}^n \mathrm{Var}(\bar{u}_{sn}|\mathcal{F}_n) \le n^{-2}\sum_{s=1}^n E[\bar{u}_{sn}^2|\mathcal{F}_n] \le n^{-1}c_n^2 = o(1)$, so that $T_{n2} = o_p(1)$ by conditional Chebyshev. Finally, since $E[u_s|\mathcal{F}_n] = 0$, we have $\bar{\mu}_{sn} = -E[u_s\mathbb{1}(|u_s| > c_n)|\mathcal{F}_n]$. Then $E[|T_{n3}||\mathcal{F}_n] \le (1/n)\sum_{s=1}^n E[|u_s|\mathbb{1}(|u_s| > c_n)|\mathcal{F}_n] = o_p(1)$, so that $T_{n3} = o_p(1)$ by conditional Markov as before. This finishes the proof. □

Lemma 11.13. *Suppose $E[|X|^p] < \infty$ for $p > 0$. Then $\max_{i=1}^n |X_i| = o_p(n^{1/p})$.*

Proof. For $\epsilon > 0$ we have $P(\max_{i=1}^n |X_i| > \epsilon n^{1/p}) \le nP(|X_i| > \epsilon n^{1/p}) = nP(|X_i|^p > \epsilon^p n) \le n(\epsilon^p n)^{-1}E[|X_i|^p\mathbb{1}(|X_i|^p > \epsilon^p n)] \lesssim E[|X_i|^p\mathbb{1}(|X_i|^p > \epsilon^p n)] \to 0$. The first inequality by union bound, the equality by monotonicity of $x \to x^p$. The second inequality is Markov's, and the final statement by dominated convergence, since $E[|X_i|^p] < \infty$. □

Lemma 11.14 (Random Partitions). *Let $\mathcal{G}_n = \{g_s\}_{s=1}^m$ a random collection of disjoint subsets of $[n]$. Let variables $((\tau_s)_s, W_{1:n}, \pi)$ jointly independent for some random elements π and $(\tau_s)_{s=1}^m$. Let $h_i = h(W_i)$ for a fixed measurable function h and suppose that the partition $\mathcal{G}_n$ is $\mathcal{F}_n$-measurable for $\mathcal{F}_n = \sigma(h_{1:n}, \pi)$. Then for $s \ne r$ and measurable F_s, F_r, we have $F_s((W_i)_{i\in g_s}, \tau_s) \perp\!\!\!\perp F_r((W_i)_{i\in g_r}, \tau_r)|\mathcal{F}_n$.*

Proof. Since F_s, F_r are arbitrary, it suffices to show $E[F_sF_r|\mathcal{F}_n] = E[F_s|\mathcal{F}_n]E[F_s|\mathcal{F}_n]$. Denote $W_I = (W_i)_{i\in I}$ and similarly for h_I. Then $F_s = \sum_{I\in 2^{[n]}} \mathbb{1}(g_s = I)F_s(W_I, \tau_s)$. We claim that it suffices to show

$$E[F_s(W_I,\tau_s)F_r(W_J,\tau_r)|\mathcal{F}_n] = E[F_s(W_I,\tau_s)|\mathcal{F}_n]E[F_r(W_J,\tau_r)|\mathcal{F}_n]$$

for all disjoint $I \cap J = \emptyset$. To see this, note that in this case by measurability of $\mathcal{G}_n$ with respect to $\mathcal{F}_n$ and disjointness of the groups we would have

$$E[F_sF_r|\mathcal{F}_n] = E\left[\sum_{I\in 2^{[n]}}\sum_{J\in 2^{[n]}} \mathbb{1}(g_s = I)\mathbb{1}(g_r = J)F_s(W_I,\tau_s)F_r(W_J,\tau_r)|\mathcal{F}_n\right]$$
$$= \sum_{\substack{I,J\in 2^{[n]}\\ I\cap J=\emptyset}} \mathbb{1}(g_s = I)\mathbb{1}(g_r = J)E[F_s(W_I,\tau_s)|\mathcal{F}_n]E[F_r(W_J,\tau_r)|\mathcal{F}_n] = E[F_s|\mathcal{F}_n]E[F_r|\mathcal{F}_n]$$

Then consider such $I\cap J = \emptyset$. Note the fact (1) if $(A,B) \perp\!\!\!\perp C$ then $A \perp\!\!\!\perp C|B$. By applying (1) with $A = (W_I,\tau_s,W_J,\tau_r)$, $B = h_{1:n}$ and $C = \pi$, we have $E[F_s(W_I,\tau_s)F_r(W_J,\tau_r)|\mathcal{F}_n] = E[F_s(W_I,\tau_s)F_r(W_J,\tau_r)|h_{1:n}]$. Then it suffices to show that (a) $(W_I,\tau_s) \perp\!\!\!\perp (W_J,\tau_r)|h_{1:n}$. By fact (1) to show (a) it suffices to prove (b) $(W_I,\tau_s,h_I) \perp\!\!\!\perp (W_J,\tau_r)|h_{I^c}$. By fact (1) again, it suffices to show (c) $(W_I,\tau_s,h_I) \perp\!\!\!\perp (W_J,\tau_r,h_{I^c})$. This is true by disjointness and joint independence of the $(\tau_s)_s$, which finishes the proof. □

Lemma 11.15 (Design Properties)**.** *Let $D_{1:n} \sim \mathrm{Loc}(\psi_n,p_n)$. Denote group randomization variables $\tau^d = (\tau^d_{a,s})_{a,s}$, jointly independent with $(\tau^d_{a,s,\ell})_{\ell=1}^{k_a} \sim \mathrm{CR}(q_a/k_a)$ for $1\le s\le n-1$ and remainder group $(\tau^d_{a,n,\ell})_{\ell=1}^{k_a} \sim \mathrm{SRS}(q_a/k_a)$. Let $(\mathcal{F}_n)_{n\ge 0}$ a sequence of σ-algebras with $\mathcal{G}_n \subseteq \mathcal{F}_n$ and $\mathcal{F}_n \perp\!\!\!\perp \tau^d$. Then the following hold*

(i) For each $i\in[n]$ we have $E[D_i|\mathcal{F}_n] = \sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\mathbb{1}(i\in g_{a,s})\cdot p_a$. In particular, $E[D_i\mathbb{1}(i\in g_{a,s})|\mathcal{F}_n] = \mathbb{1}(i\in g_{a,s})\cdot p_a$.

(ii) For $1\le i\le n$ and $1\le s\le n$ we have $\mathrm{Var}(D_i|\mathcal{F}_n)\mathbb{1}(i\in g_{a,s}) = p_a(1-p_a)\mathbb{1}(i\in g_{a,s})$. For $1\le i,j\le n$ distinct indices and $1\le s\le n-1$

$$\mathrm{Cov}(D_i,D_j|\mathcal{F}_n)\mathbb{1}(i,j\in g_{a,s}) = -\frac{q_a(k_a-q_a)}{k_a^2(k_a-1)}\mathbb{1}(i,j\in g_{a,s})$$

In particular, $|\mathrm{Cov}(D_i,D_j|\mathcal{F}_n)\mathbb{1}(i,j\in g_{a,s})| \le k_a^{-1}\mathbb{1}(i,j\in g_{a,s})\mathbb{1}(s\ne n)$. Moreover, $\mathrm{Cov}(D_i,D_j|\mathcal{F}_n)\mathbb{1}(g(i)\ne g(j)) = 0$.

Proof. For the first statement, note that

$$D_i = \sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{\ell=1}^{k_a} D_i\mathbb{1}(i = g_{a,s,l}) = \sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{\ell=1}^{k_a}\tau^d_{a,s,\ell}\mathbb{1}(i = g_{a,s,l})$$

Then since $\mathcal{G}_n \in \mathcal{F}_n$ and $\tau^d_{a,s,\ell} \perp\!\!\!\perp \mathcal{F}_n$ we have

$$E[D_i|\mathcal{F}_n] = \sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{\ell=1}^{k_a} E[\mathbb{1}(i = g_{a,s,\ell})\tau^d_{a,s,\ell}|\mathcal{F}_n] = \sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{\ell=1}^{k_a}\mathbb{1}(i = g_{a,s,\ell})E[\tau^d_{a,s,\ell}|\mathcal{F}_n]$$
$$= \sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{\ell=1}^{k_a}\mathbb{1}(i = g_{a,s,\ell})E[\tau^d_{a,s,\ell}] = \sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{\ell=1}^{k_a}\mathbb{1}(i = g_{a,s,\ell})p_a$$

To finish, note that $\sum_{\ell=1}^{k_a} \mathbb{1}(i = g_{a,s,\ell}) = \mathbb{1}(i \in g_{a,s})$ by definition. For (ii), by the decomposition above and measurability assumption, for $1 \le j \neq i \le n$

$$\operatorname{Cov}(D_i, D_j|\mathcal{F}_n) = \sum_{a,a'=1}^{|L_n|} \sum_{s,s'=1}^{n} \sum_{\ell=1}^{k_a} \sum_{\ell'=1}^{k_{a'}} \mathbb{1}(i = g_{a,s,\ell}) \mathbb{1}(j = g_{a',s',\ell'}) \operatorname{Cov}(\tau^d_{a,s,\ell}, \tau^d_{a',s',\ell'}|\mathcal{F}_n)$$

By σ-algebra independence and joint independence of groupwise randomizations

$$\begin{aligned}\operatorname{Cov}(\tau^d_{a,s,\ell}, \tau^d_{a',s',\ell'}|\mathcal{F}_n) &= \operatorname{Cov}(\tau^d_{a,s,\ell}, \tau^d_{a',s',\ell'}) \\ &= \begin{cases} 0 & (a,s) \neq (a',s') \\ p_a - p_a^2 & (a,s,\ell) = (a',s',\ell') \\ -\frac{q_a(k_a - q_a)}{k_a^2(k_a-1)} & (a,s) = (a',s'); \quad \ell \neq \ell' \quad 1 \le s < n \\ 0 & (a,s) = (a',s'); \quad \ell \neq \ell' \quad s = n \end{cases}\end{aligned}$$

The third line follows since by definition of $\mathrm{CR}(q_a/k_a)$, for $(a,s) = (a',s')$ we have

$$\begin{aligned}\operatorname{Cov}(\tau^d_{a,s,\ell}, \tau^d_{a',s',\ell'}) &= P(\tau^d_{a,s,\ell} = \tau^d_{a,s,\ell'} = 1) - (q_a/k_a)^2 = \binom{k_a}{q_a}^{-1} \binom{k_a - 2}{q_a - 2} - (q_a/k_a)^2 \\ &= \frac{q_a(q_a-1)}{k_a(k_a-1)} - (q_a/k_a)^2 = -\frac{q_a(k_a - q_a)}{k_a^2(k_a-1)}\end{aligned}$$

The bounds follow by inspection. □

Lemma 11.16 (Stochastic Balance)**.** *Let $(\mathcal{F}_n)_{n\ge1}$ such that treatment groups $\mathcal{G}_n, (h_n(W_i))_{i=1}^n \in \mathcal{F}_n$ for a sequence of functions $(h_n)_{n\ge1}$ and $\mathcal{F}_n \perp\!\!\!\perp \tau^d$. Let $D_{1:n} \sim \mathrm{Loc}(\psi_n, p_n)$ and $T_{1:n} \in \{0,1\}^n$ such that $\{i : T_i = 1\} = \sqcup_{g\in\mathcal{G}_n} g$. The following hold*

(1) $E[E_n[T_i(D_i - p_n(X_i))h_n(W_i)]|\mathcal{F}_n] = 0$

(2) $\operatorname{Var}(E_n[T_i(D_i - p_n(X_i)))h_n(W_i)]|\mathcal{F}_n) \le 2n^{-1}E_n[T_i h_n(W_i)^2] \le 2n^{-1}E_n[h_n(W_i)^2]$

(3) Suppose $\exists$ $(\mathcal{F}_n)_{n\ge1}$ satisfying the conditions above. If $\sup_{n\ge1} E[h_n(W)^2] < \infty$ then $E_n[T_i(D_i - p_n(X_i))h_n(W_i)] = O_p(n^{-1/2})$. If $(h_n(W))_{n\ge1}$ is uniformly integrable then $E_n[T_i(D_i - p_n(X_i))h_n(W_i)] = o_p(1)$.

(4) The variance $\operatorname{Var}(\sqrt{n}E_n[T_i(D_i - p_n(X_i)))h_n(W_i)]|\mathcal{F}_n)$ is bounded above by

$$n^{-1}\sum_{g\in\mathcal{G}_n} |g|^{-1} \sum_{i,j\in g} (h_n(W_i) - h_n(W_j))^2 + n^{-1}\overline{k}_n|L_n| \cdot \max_{i=1}^{n} h_n(W_i)^2$$

Proof. First, by assumption $\{i : T_i = 1\} = \bigsqcup_{g\in\mathcal{G}} g$, so $E_n[T_i(D_i - p_{i,n})h_n(W_i)]$ is equal to

$$n^{-1}\sum_{g\in\mathcal{G}_n}\sum_{i\in g}(D_i - p_{i,n})h_n(W_i) = n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{i=1}^{n}(D_i - p_a)h_n(W_i)\mathbb{1}(i \in g_{a,s})$$

By Lemma 11.15 and our measurability assumptions

$$E[(D_i - p_a)h_n(W_i)\mathbb{1}(i \in g_{a,s})|\mathcal{F}_n] = h_n(W_i)E[(D_i - p_a)\mathbb{1}(i \in g_{a,s})|\mathcal{F}_n] = 0$$

By linearity, this shows the claim. By Lemma 11.15, $\mathrm{Var}(E_n[(D_i - p_{i,n})h_n(W_i)]|\mathcal{F}_n)$ is

$$
\begin{aligned}
& n^{-2}\sum_{a,a'}^{|L_n|}\sum_{s,s'=1}^{n}\sum_{i,j}^{n} \mathrm{Cov}((D_i - p_a)h_n(W_i)\mathbb{1}(i \in g_{a,s}), (D_j - p_{a'})h_n(W_j)\mathbb{1}(j \in g_{a',s'})|\mathcal{F}_n) \\
&= n^{-2}\sum_{a,a'}^{|L_n|}\sum_{s,s'=1}^{n}\sum_{i,j}^{n} h_n(W_i)h_n(W_j)\mathbb{1}(i \in g_{a,s})\mathbb{1}(j \in g_{a',s'})\,\mathrm{Cov}(D_i, D_j|\mathcal{F}_n) \\
&= n^{-2}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{i,j}^{n} h_n(W_i)h_n(W_j)\mathbb{1}(i, j \in g_{a,s})\,\mathrm{Cov}(D_i, D_j|\mathcal{F}_n)
\end{aligned}
$$

The final equality follows from Lemma 11.15. By triangle inequality and the covariance bound in Lemma 11.15, this is bounded above by

$$
\begin{aligned}
& n^{-2}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\left[\sum_{i=1}^{n} h_n(W_i)^2\mathbb{1}(i \in g_{a,s}) + \sum_{i\neq j}^{n} |h_n(W_i)||h_n(W_j)|\mathbb{1}(i, j \in g_{a,s})k_a^{-1}\mathbb{1}(s \neq n)\right] \\
&\leq n^{-1}E_n[T_i h_n(W_i)^2] + \sum_{a=1}^{|L_n|} k_a^{-1}\sum_{s=1}^{n}\sum_{i\neq j}^{n} |h_n(W_i)||h_n(W_j)|\mathbb{1}(i, j \in g_{a,s}) \\
&\leq n^{-1}E_n[T_i h_n(W_i)^2] + n^{-2}\sum_{a=1}^{|L_n|} k_a^{-1}\sum_{s=1}^{n}\left(\sum_{i=1}^{n} |h_n(W_i)|\mathbb{1}(i \in g_{a,s})\right)^2
\end{aligned}
$$

Continuing the calculation, by Jensen's inequality the second term is equal to

$$
n^{-2}\sum_{a=1}^{|L_n|} k_a \sum_{s=1}^{n}\left(k_a^{-1}\sum_{i\in g_{a,s}} |h_n(W_i)|\right)^2 \leq n^{-2}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n} k_a k_a^{-1}\sum_{i\in g_{a,s}} |h_n(W_i)|^2 = n^{-1}E_n[T_i h_n(W_i)^2]
$$

This completes the proof of (2). Claim (3) follows by applying (1), (2) with $(\mathcal{F}_n)_{n\geq 1}$ any sequence satisfying the conditions, followed by Markov inequality (Lemma 11.9). For the final part of (3), let $c_n \to \infty$ with $c_n = o(\sqrt{n})$ and define $\bar{h}_{in} = h_n(W_i)\mathbb{1}(|h_n(W_i)| \leq c_n)$. Consider the expansion $E_n[T_i(D_i - p_{i,n})h_n(W_i)] = E_n[T_i(D_i - p_{i,n})(h_n(W_i) - \bar{h}_n(W_i))] + E_n[T_i(D_i - p_{i,n})\bar{h}_n(W_i)] \equiv A_n + B_n$. We may write $|A_n| \leq E_n[|h_n(W_i) - \bar{h}_n(W_i)|] = E_n[|h_n(W_i)|\mathbb{1}(|h_n(W_i)| > c_n)]$. Then we have $E[|A_n|] \leq E[|h_n(W_i)|\mathbb{1}(|h_n(W_i)| > c_n)] \to 0$ as $n \to \infty$ by uniform integrability. Then $A_n = o_p(1)$ by Markov inequality. By the first part of (3), $\mathrm{Var}(B_n|\mathcal{F}_n) \leq 2n^{-1}E_n[\bar{h}_n(W_i)^2] \leq 2n^{-1}c_n^2 = o(1)$. Then $B_n = o_p(1)$ by conditional Chebyshev.

For the final identity (5), note that from Lemma 11.15 for $s \neq n$ and $p_a = q_a/k_a$

$$
\begin{aligned}
\sum_{i,j\in g_{a,s}} h_n(W_i)h_n(W_j)\,\mathrm{Cov}(D_i, D_j|\mathcal{F}_n) &= \frac{q_a(k_a - q_a)}{k_a^2}\sum_{i\in g_{a,s}} h_n(W_i)^2 \\
&\quad + \frac{q_a(k_a - q_a)}{k_a^2(k_a - 1)}\sum_{i,j\in g_{a,s}} (-2)h_n(W_i)h_n(W_j)
\end{aligned}
$$

Note that $-2ab = (a-b)^2 - a^2 - b^2$. Then the second sum is

$$\sum_{i,j \in g_{a,s}} (-2) h_n(W_i) h_n(W_j) = \sum_{i,j \in g_{a,s}} [(h_n(W_i) - h_n(W_j))^2 - h_n(W_i)^2 - h_n(W_j)^2]$$
$$= \sum_{i,j \in g_{a,s}} (h_n(W_i) - h_n(W_j))^2 - (k_a - 1) \sum_{i \in g_{a,s}} h_n(W_i)^2$$

Substituting in the first display above, the diagonal terms cancel. For the claimed constant, note that $\max_{p \in (0,1)} p(1-p) \leq 1/4$ and $\max_{k \geq 2} \frac{k}{k-1} \leq 2$, so $\frac{q_a(k_a - q_a)}{k_a^2(k_a - 1)} \leq k_a^{-1}$. Aggregating over (a, s) gives

$$n^{-1} \sum_{a=1}^{|L_n|} \left[\sum_{s=1}^{n-1} \sum_{i,j \in g_{a,s}}^{n} (h_n(W_i) - h_n(W_j))^2 + \sum_{i=1}^{n} h_n(W_i)^2 \mathbb{1}(i \in g_{a,n}) \right]$$

The second term is

$$n^{-1} \sum_{a=1}^{|L_n|} \sum_{i=1}^{n} h_n(W_i)^2 \mathbb{1}(i \in g_{a,n}) \leq n^{-1} \max_{i=1}^{n} h_n(W_i)^2 \sum_{a=1}^{|L_n|} \sum_{i=1}^{n} \mathbb{1}(i \in g_{a,n})$$
$$\leq n^{-1} \overline{k}_n |L_n| \cdot \max_{i=1}^{n} h_n(W_i)^2$$

This finishes the proof. $\square$