Optimal Stratification of Survey Experiments*

Max Cytrynbaum†

May 18, 2023

Abstract

This paper studies treatment effect estimation in a novel two-stage model of experimentation. In the first stage, using baseline covariates, the researcher samples units to participate in the experiment from a pool of eligible units. In the second stage, they assign each sampled unit to one of two treatment arms. We define a new family of matched tuples designs, significantly generalizing matched pairs, which can be used to implement finely stratified sampling and assignment with varying sampling intensity and non-constant treatment proportions. When used to sample representative units into the experiment, our method non-parametrically attenuates the variance due to treatment effect heterogeneity. When used to assign treatments, our designs make simple difference-of-means estimation asymptotically equivalent to well-specified doubly-robust estimation, effectively doing non-parametric regression adjustment by design. We derive the most efficient stratification variables for sampling and assignment, as well as the optimal heterogeneous sampling intensity and treatment proportions under fine stratification. We also provide asymptotically exact inference methods for such two-stage designs, allowing researchers to fully exploit the efficiency gains from finely stratified sampling and assignment. We apply our methods to the setting of two-wave design, where the researcher has access to a pilot study when designing the main experiment. We provide the first fully efficient solution to this problem, showing how to use pilot data to estimate the optimal sampling and assignment design for the main experiment.

*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.

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

1 Introduction

Randomized controlled trials (RCTs) are increasingly common in economics research. The AEA RCT registry currently lists over 5000 active experiments, spanning a range of different fields. Experimental design attempts to reduce the variance of causal effect estimates, increasing precision for a fixed experiment size. In the literature, design methodology often focuses on how researchers should assign treatments. For example, stratified block randomization tries to assign treatments so that covariates are balanced between the different treatment arms. This paper studies a new dimension of experimental design: sampling of the experimental participants. We show that it is possible to increase efficiency by sampling units that are *representative* of the broader population.

To implement representative sampling, we propose a new family of *local randomization* procedures, which use baseline covariates to randomly sample a representative subsample from a larger population of eligible units. Local randomization can also be used for treatment assignment, where it improves upon existing methods by randomizing within fine, data-adaptive strata that are optimally chosen to produce covariate balance. Applied researchers can use these tools to design a representative and balanced experiment, making the most efficient use of scarce experimental resources.

We study estimation of the average treatment effect (ATE) in a two-stage design model, where the researcher first *samples* units to participate in the experiment from among a pool of eligible units,[1] then *assigns* each sampled unit to one of two treatment arms. We propose a new family of local randomization methods to implement both stages of the design. Locally randomized sampling makes experimental participants more representative of the broader population, while locally randomized assignment finely balances covariates between treatment arms. We give novel asymptotically exact inference methods for locally randomized sampling and assignment, allowing experimenters to report smaller confidence intervals if they designed a representative experiment. We also apply our methods to the setting where the researcher has access to a pilot study when designing the main experiment, obtaining the first *fully efficient* solution to this problem.

In practice, sampling of participants is often an unavoidable part of designing an experiment. For example, in a recent paper Abaluck et al. (2021) consider the effect of mask promotion on village-level covid infection rates in Bangladesh. From a sample of 1000 villages, they first randomly sample 600 to be included in the experiment. Next, the sampled villages are randomly assigned to various interventions that promote mask usage. Similarly, Breza et al. (2021) run video ads on Facebook discouraging holiday travel, estimating the effect of low versus high intensity of ads on county-level covid infection rates. Since experimental resouces are finite, they first sample a small set of counties in which to run ads and collect outcome data. The sampled counties are then randomly assigned to either low or high intensity of treatment.[2] Can we do better than completely random sampling in these examples?

[1] We also allow the case where the researcher can sample any unit in the entire population, which is known. See Remark **??** for discussion.

[2] Alternatively, suppose 5000 job-seekers apply to a employment assistance program, as in Caria et al. (2021), but the budget only allows for 1000 participants. Who should we let into the program?

The sampling step is most applicable when the researcher has a sample of "eligible" units, any of which can be included in the experiment. As seen in the examples above, resource or logistical constraints, common in real life, may make it impossible to sample all of them. In some settings, however, the sampling step may be irrelevant or impossible. For instance, if there are no resource constraints, the researcher should just run the largest experiment possible, sampling all the eligible units. If the research question is such that one treatment arm is "do nothing" (control), and outcome data is always observed for the control units, then the sampling problem is trivial (sample everyone).

Representative sampling can significantly reduce estimator variance, while preserving finite-sample unbiasedness. To see why, consider estimating the effect of distributing surgical masks versus cloth masks on covid infection rates, as in Abaluck et al. (2021). Suppose average resident age predicts vulnerability to covid, so that older villages have larger treatment effects from switching to surgical masks from cloth masks. If sampling is completely random, the sampled villages may be older or younger than the full sample of eligible villages, just by random chance. If we randomly sample mostly old villages, our treatment effect estimate will be biased up (ex-post), and conversely it will be biased down if we sample villages with only healthy young residents. This illustrates how chance covariate imbalances created during sampling can increase estimator variance (ex-ante) when treatment effects are heterogeneous. Representative sampling avoids this problem, dampening this source of variance.

This paper implements both representative sampling and finely balanced treatment assignment using a new family of *local randomization* algorithms. We introduce the basic principle of local randomization using the example above. Suppose the researcher has a sample of 1000 eligible villages, but logistical constraints only allow sampling of 300 to participate in the experiment. Choose a set of baseline village-level covariates expected to be predictive of both outcomes (covid infection rates) and treatment effect heterogeneity. Partition the 1000 eligible villages into groups of 10 that are *maximally homogeneous*[3] in the predictive covariates. For instance, the villages in each group of 10 may be very similar in terms of age, housing density, and baseline infection rates. Randomly sample exactly 3 of 10 villages from each group into the experiment, leaving 300 villages. Suppose welfare considerations or logistical constraints require assigning 2/3 of the villages to surgical masks and 1/3 to cloth, as in Abaluck et al. (2021). Then during treatment assignment, again partition the 300 sampled villages into homogeneous groups of 3. Within each group, randomly assign exactly 2 villages to surgical masks and 1 to cloth.

We give novel algorithms and rate analysis for the combinatorial problem of constructing maximally homogeneous groups of units. Enforcing group homogeneity allows us to think of each group as consisting of units of a certain "covariate type." Sampling exactly 3 out of 10 units from each group ensures that each type of unit in the larger eligible sample is well-represented in the experiment. Under completely random sampling, by contrast, we may not sample any units from certain groups, while other groups may be over-represented in the experiment by random chance. Such fluctuations increase estimator variance.

Summarizing the two-stage procedure, during sampling we solve an optimization problem

[3]Defined formally in Section 2 below.

to match eligible units into homogeneous groups, randomly sampling a fixed proportion of the units in each group to participate in the experiment. During treatment assignment, we repeat the process among the sampled units, again forming homogeneous groups of sampled units and randomly assigning a fixed proportion from each group to one of two treatment arms. By introducing and studying this two-stage procedure, we make several contributions to the literature on experimental design.

Our first contribution is to extend the principle of matched-pairs randomization to arbitrary propensity scores $p(x) \neq 1/2$, producing a large family of novel designs. For example, consider an experiment where, due to welfare considerations, the researcher wants to assign older villages to surgical masks with probability 2/3, while younger villages receive surgical masks with lower probability 1/3. Local randomization provides a "matched triples" design with varying treatment proportions, implementing this design constraint while enforcing strong covariate balance. Other novel designs are given in the many examples throughout the paper.

This paper's second contribution is the sampling model. We show that *efficient sampling* requires finely balancing the covariates that are most predictive of treatment effect heterogeneity. Effectively, researchers should ensure that the different "types" of treatment response are well-represented among the sampled units. Local randomization methods give an efficient and practical implementation of this idea.

For our third contribution, we give a new central limit theorem for inverse propensity weighting (IPW) estimators of the ATE[4] in experiments with locally-randomized sampling and treatment assignment. We show that local randomization does *non-parametric* regression adjustment *by design*, reducing estimator variance without the need to actually specify or estimate a regression model ex-post. Similarly, locally randomized sampling reduces the variance due to treatment effect heterogeneity (non-parametrically), to the extent that the covariates used to form homogeneous groups explain this heterogeneity. For sampling, the optimal design locally randomizes with respect to the conditional average treatment effect (CATE). For treatment assignment, the optimal design locally randomizes with respect to the *balance function*, a weighted sum of conditional expectations defined below. The oracle design that locally randomizes with respect to the pair (CATE, balance function) in both stages of the design achieves full efficiency.

While matched pairs dates back at least to Fisher (1926), its asymptotic analysis is recent (Bai et al. (2021)), likely due to the complicated statistical dependencies created by solving a global optimization problem to form the pairs. Our analysis extends this work, proving a general CLT for randomization within fine, *data-adaptive* strata. This result can also be applied to obtain a new CLT for classical stratified block randomization with large, fixed strata, recently studied using entirely different methods in Bugni et al. (2018).

Our fourth contribution gives new methods for asymptotically *exact* inference on the ATE under locally randomized sampling and assignment. In particular, we allow researchers to report smaller confidence intervals if they use the methods in this paper to design a representative experiment. Our construction allows plug-in asymptotically exact inference over the entire class of locally randomized designs defined below. As

[4]We consider both the ATE and finite-population fixed-regressor estimands, as in Abadie et al. (2014).

a consequence, we get new asymptotically exact variance estimators for matched pairs and classical stratified block randomization. We also explore an interesting connection between our matching-based variance estimators and randomization inference.

We apply our methods to the problem of experimental design when the researcher has data from a pilot study, obtaining the first *fully efficient* design in this setting. To achieve full efficiency, we use pilot data to estimate the optimal propensity score[5] for the experiment, which assigns units to each treatment arm in proportion to that arm's conditional variance.[6] Intuitively, one can think of this as taking more measurements of the quantity that is harder to measure. We construct a balanced implementation of this propensity score by *double stratification*, first stratifying units by our pilot estimate of the optimal propensity score, then locally randomizing treatments within these propensity strata. Our analysis shows that this minimizes the estimator variance due to sampling and assignment, as well as the residual variance, *independently*, while existing methods make compromises between these different sources of variance.

We also propose a second method that uses the pilot to "estimate what to balance" during both sampling and assignment. We show that locally randomizing with respect to a consistent pilot estimate of (CATE, balance function), the optimal design mentioned above, is also asymptotically fully efficient. However, we argue that such consistency assumptions give poor guidance for practice in experimental design. Instead, we advocate a "robustified" approach that balances some key predictive covariates outright, in addition to balancing the pilot regressions. Finally, for settings where a pilot experiment is infeasible, we discuss matching on *proxy regressions*, estimated in a related experiment or observational data set. In clinical trials, our results allow exact inference for designs that balance a vector of "risk scores" from previous studies. We also compare our results with the recent approaches of Tabord-Meehan (2022) and Bai (2022) in this setting.

1.1 Related Literature

There is a large literature on experimental design, see e.g. Athey and Imbens (2017) or Rosenberger and Lachin (2016) for surveys. See Bruhn and McKenzie (2009) for evidence on experimental design as used in practice in development economics. The literature may be divided by assumptions about the timing of the experiment. Our first set of results assumes the classical timing: (1) pre-sampling baseline covariates are observed (2) units are sampled into the trial and (possibly) more covariates are observed (3) units are assigned to interventions and (4) all outcomes are observed. Alternatively, some papers assume sequential timing, with units arriving one-by-one and treatment decisions made "on the spot", as in Efron (1971) or Kapelner and Krieger (2014). Our results on design with a pilot fit into the adaptive design literature, with some outcomes observed during the treatment process. See Hu and Rosenberger (2006) for an overview.

With respect to solution concept, some papers emphasize "robustness", searching for minimax designs over a class of DGP's satisfying certain smoothness assumptions. For example, see Kallus (2017) or the discussion and references in Harshaw et al. (2021).

[5]For complete randomization, optimal constant treatment proportions are known as the Neyman allocation. We give a local randomization implementation of the *conditional* Neyman allocation.

[6]One can think of this as a design analogue of (optimal) feasible weighted least squares.

Alternatively, one may use a decision theoretic framework to characterize the optimal design, as in Kasy (2016). By contrast, this paper defines a new family of randomization procedures, studying their asymptotic efficiency pointwise in (design, DGP), over a class of DGP's satisfying certain regularity conditions.

Our sampling model is related to the literature on survey sampling, see e.g. Cochran (1977). The designs in this paper are examples of blocking, as in Fisher (1926), Higgins et al. (2015), Fogarty (2018), and Bai (2022). The family we study contains classical stratified block randomization (SBR) with fixed strata and matched pairs, as analyzed in Bugni et al. (2018) and Bai et al. (2021). Rerandomization, studied in Morgan and Rubin (2012) and Li et al. (2018), is not contained in this family. Follow up work in Cytrynbaum (2022) studies rerandomizing local designs, for instance giving results for rerandomized SBR and matched pairs, as well as rerandomized versions of all the other designs studied in this paper. The main result shows that local randomization generically dominates rerandomization, giving smaller asymptotic variance.

Related to our result that local randomization does non-parametric regression adjustment "by design", Li et al. (2018) show that rerandomization does (slightly worse than) linear regression by design. Similarly, Harshaw et al. (2021) bound the MSE of their "Grahm-Schmidt Walk" design by a quantity related to ridge regression. Cytrynbaum (2022) shows that rerandomized local designs do semiparametric regression by design.

Our first set of results is most related to Bai (2022), which we build on. Relative to this paper, we study two-stage designs, with both covariate-adaptive sampling and treatment assignment. For treatment assignment alone, the designs in Bai (2022) are a special case of our method with (1) $\dim(\psi) = 1$ for the local function defined below (function of covariates we match on) and (2) constant propensity $p = a/k$. Allowing $\dim(\psi) > 1$ is necessary to implement the oracle design for joint sampling and assignment, as well as its empirical analogue estimated using pilot data. More generally, our method allows balancing with respect to more than just a single covariate.

For $\dim(\psi) > 1$, the algorithm used in Bai that constructs groups of units by sorting their $\psi(X_i)$ values is no longer feasible. We study combinatorial optimization procedures (graph-partitioning) to form groups in higher dimensions, and give new analysis of their statistical rates. Our designs allow propensity $p(x)$ non-constant. This innovation is required for a fully efficient solution to the two-wave design problem, which uses the pilot to estimate, and "locally" implement, an optimally varying propensity score (Neyman allocation). Finally, Bai (2022) allows inference only for the case $p = 1/2$, using the results of Bai et al. (2021) on matched pairs. Our work gives novel, generic inference methods that cover the full family of locally randomized designs, including joint covariate-adaptive sampling and assignment.

Our solution to the problem of design with a pilot experiment follows work in Hahn et al. (2011), Tabord-Meehan (2022), and Bai (2022). More broadly, the two-wave setting is an example of response-adaptive design, as in Russo (2016) and Kasy and Sautmann (2021). A comparison with the large pilot results of Bai (2022) is given in Section 5. A detailed comparison with Tabord-Meehan (2022) is given in Remark 5.7.

The rest of this paper is organized as follows. In Section 2 we formally describe our method and preview some of our main results. Section 3 gives our main asymptotic results, examples of new designs produced by our framework, and discusses the optimal designs for sampling and assignment. Section 5 gives our results on optimal design with a pilot experiment. Section 6 gives our inference methods. Section 7 gives our empirical results. All proofs are given in the appendix.

2 Motivation and Description of Method

Consider running an experiment to estimate the average treatment effect (ATE). There are n eligible units, with observed 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 for unit i under $d \in \{0, 1\}$. Since outcomes are only observed for participating units[7] we write

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

We assume $(X_i, Y_i(0), Y_i(1))_{i=1}^n \overset{\text{iid}}{\sim} F$ and present results on estimation and inference on both the population ATE $= E[Y(1) - Y(0)]$ and sample average treatment effect SATE $= E_n[Y_i(1) - Y_i(0)]$. For external validity, the SATE is defined over the entire eligible population $i \in [n]$, not just the smaller set of experiment participants $\{i : T_i = 1\}$. For the ATE, we model the n eligible units as an independent sample from a superpopulation of interest, which we then subsample to choose the experiment participants. For instance, the eligible units could be $n = 1000$ respondents to an online advertisement to participate in an experiment with a budget constraint of $qn = 100$ participants. For the SATE, the n eligible units may be the entire population of villages in a country, which we sample to obtain the participating villages. For the SATE, our results are "design-based" in the sense that the asymptotics hold conditional on the data $W_{1:n} = (X_i, Y_i(0), Y_i(1))_{i=1}^n$, so that estimator variation is solely due to $T_{1:n}$ and $D_{1:n}$. In this case, the probability model F imposes homogeneity on the sequence of finite populations $W_{1:n}$ as n increases, so that sample moments converge to well-defined limits, but does not contribute to estimator variance. See Section 9.1 below for further discussion.

Our goal is to sample a representative subset of 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 fine stratifications, generalizing matched pairs designs to arbitary propensity scores $p(x) = P(D = 1|X = x)$, with continuous or discrete covariates in general dimension. The basic building block is a matched k-tuples design, which uses the baseline covariates to match units into homogeneous groups of k and randomly assigns a out of k units in each group to $T = 1$ during sampling or $D = 1$ during assigment. We introduce 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 $\{1, \ldots, n\} = \bigsqcup_g \{i \in g\}$. In general, there

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

may be a single 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 $T_{1:n} \sim \mathrm{Loc}(\psi, q)$.

Equation 2.1 generalizes a similar condition in Bai et al. (2021) for the case $k = 2$. We discuss matching algorithms to construct groups that satisfy these conditions below. Consider sampling and assignment propensities $q = a/k$ and $p = a'/k'$. We study a two-stage procedure that first samples eligible units $T_{1:n} \sim \mathrm{Loc}(\psi, q)$, then assigns treatments using $D_{1:n} \sim \mathrm{Loc}(\psi, p)$ to the sampled units $\{i : T_i = 1\}$. In Section 3.2, we extend Definition 2.1 to allow different stratification variables for sampling and assignment, as well as general propensities $q(x)$, $p(x)$. Finally, we let $q = 1$ denote the design with $T_i = 1$ for all $1 \le i \le n$.

Remark 2.2 (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, "coarse stratification" designs with a fixed number of large strata are also included in this framework. For example, we may set $\psi = 1$ and form groups $|g| = 3$ at random, automatically satisfying Equation 2.1. Assigning 2 out of 3 units in each group to treatment gives a "random matched triples" representation of complete randomization with $p = 2/3$. Similarly, coarse stratification[8] with fixed strata $S(X) \in \{1, \ldots, m\}$ can be obtained by setting $\psi(X) = S(X)$ and matching units with the same $S(X)$ value into groups at random. Because of this, our framework allows for unified asymptotics and inference procedures for a wide range of stratifications.

Example 2.3 (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 $T_{1:n} \sim \mathrm{Loc}(\psi, 3/10)$, matching the 1000 eligible units into homogeneous groups of 10 and randomly setting $T_i = 1$ 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 $D_i = 1$.

2.1 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. Denote $d(\psi) = \dim(\psi)$.

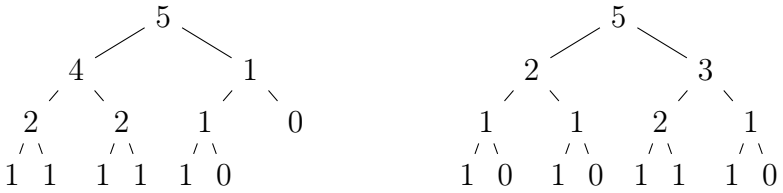
[8] Coarse stratification was previously studied using different methods in Bugni et al. (2018). We extend their results, e.g. allowing coarse stratification at both sampling and assignment stages.

Theorem 10.1 shows that if $E[|\psi(X)|_2^\alpha] < \infty$ for $\alpha > d(\psi) + 1$ then the optimal groups

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

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. For $k = 2$, the optimal groups are computable in $O(n^3)$ time using an algorithm due to Derigs (1988). Efficient algorithms for computing the optimal groups for general k are not available, and we expect the problem to be NP-hard.[9]

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.



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 cardinality trees for each k, not all of which can be feasibly implemented using such constraints. We provide two canonical ways of generating such sequences, their associated constraints, and the number of empty groups (zeros) to add at each level.

3 Asymptotics and Optimal Designs

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 imposed Lipschitz continuity on 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 $\psi(X)$. Our analysis shows that these assumptions can 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 just the usual difference-of-means estimator for the units sampled into the experiment. Before continuing to our asymptotic results,

[9] See Karmakar (2022) for hardness results in a related problem.

we state a variance decomposition that will be used extensively in what follows. Let $c(\psi) = E[Y(1)-Y(0)|\psi(X)]$ denote the conditional average treatment effect (CATE) and $\sigma_d^2(\psi) = \text{Var}(Y(d)|\psi(X))$ 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, $T_{1:n} \sim \text{CR}(q)$ and $D_{1:n} \sim \text{CR}(p)$. Our work shows that $\sqrt{n_T}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$ with variance

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

We think of the first term as the variance due to treatment effect heterogeneity.[10] We view the second term as the variance due to random assignment. It arises from the chance covariate imbalances between treatment and control units created by complete randomization $D_{1:n} \sim \text{CR}(p)$. 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 Propensity

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.

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

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

Comparing to Equation 3.2 above, the variance component $\text{Var}(b(\psi; p))$ due to covariate imbalance is now asymptotically negligible. The variance due to treatment effect heterogeneity $\text{Var}(c(\psi))$ is now dampened by the sampling proportion $q \in (0, 1]$. Observe that the normalization $\sqrt{n_T}(\widehat{\theta} - \text{ATE})$ effectively holds the experiment size n_T constant as we vary q. The number of eligible units $n \approx n_T/q$ grows as $q \to 0$. For small q, there are many units to sample from, allowing us to choose a very representative subsample to include in our experiment. Because of this, the first variance component scales with the larger sample size $n \approx n_T/q$ of eligible units, effectively "boosting" the experiment size for this component of the variance.

Example 3.3 (Matched Tuples)**.** In Example 2.3 above, we sampled $n_T = 300$ of $n = 1000$ individuals to treatment using the stratified design $T_{1:n} \sim \text{Loc}(\psi, 3/10)$ with $q = 3/10$. Next, we assigned $1/4$ of the sampled units to treatment by $D_{1:n} \sim \text{Loc}(\psi, 1/4)$. Theorem 3.2 shows that $\sqrt{n_T}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$ with asymptotic variance

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

[10] In particular, the treatment effect heterogeneity predictable by the baseline covariates.

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

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

This is exactly the Hahn (1998) semiparametric variance bound for ATE with iid observations $(Y, D, \psi(X))$.[11] We achieve the semiparametric variance bound with a simple difference-of-means estimator, without the need for nonparametric re-estimation of the 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."

Representative Sampling. If sampling is also locally randomized $T_{1:n} \sim \text{Loc}(\psi, q)$, then the variance due to treatment effect heterogeneity becomes $q\,\text{Var}(c(\psi))$. In this case, V is strictly smaller than the semiparametric variance bound when $q < 1$ and $\text{Var}(c(\psi)) > 0$. 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 heterogeneity $(c(\psi_i))_{i=1}^n$ in the eligible population into our sample. More formally, consider an oracle setting where we observe $c(\psi_i)$ 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 (Other Designs)**.** For the case without sampling ($q = 1$), we may compare fine stratification to other covariate-adaptive designs. Li et al. (2018) show that $\widehat{\theta}$ is asymptotically almost as efficient as interacted linear regression adjustment under rerandomized treatment assignment, effectively doing linear regression "by design." Bugni et al. (2019) show that coarsely stratified assignment $\psi(X) \in \{1, \ldots, m\}$ is asymptotically equivalent to interacted linear regression adjustment, controlling for all strata indicators. Harshaw et al. (2021) suggest a novel 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.

The next example proves an equivalence between coarse stratification and linear regression adjustment in the case where sampling is also coarsely stratified. In particular, we show that $\widehat{\theta}$ is asymptotically equivalent to a doubly-adjusted estimator that adjusts linearly for covariate imbalances during both sampling and assignment.

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

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 $S(X) \in \{1, \ldots, m\}$, Theorem 3.2 shows that $\sqrt{n_T}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V_S)$

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

Compare this with a strategy that samples units $T_{1:n} \sim \text{CR}(q)$ and assigns treatment $D_{1:n} \sim \text{CR}(p)$ by complete randomization, with ex-post linear covariate adjustment for 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

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

One can show that $\sqrt{n_T}(\widehat{\theta}_{adj} - \text{ATE}) \Rightarrow \mathcal{N}(0, V_S)$, with the same variance V_S as in Equation 3.3. See Proposition 10.9 in the appendix for the proof. Intuitively, this result shows that under coarse stratification, the simple difference-of-means estimator is as precise as the doubly adjusted estimator $\widehat{\theta}_{adj}$. We generalize this example to arbitrary designs in Proposition 3.11 in the next section.

The remainder of this subsection provides more intuition for Theorem 3.2 by connecting our results to the previous literature on semiparametric efficiency.

Remark 3.6 (Efficient Influence Function)**.** Consider expanding the difference-of-means estimator $\widehat{\theta}$ about the efficient influence function for the ATE. Denote nonparametric residuals $\epsilon_i^d = Y_i(d) - E[Y_i(d) | \psi_i]$. Under locally randomized sampling and assignment

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

If $q = 1$ 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 $\sqrt{n} \operatorname{Cov}_n(T_i, c(\psi_i)) \Rightarrow \mathcal{N}(0, v)$ with $v > 0$. By contrast, we show that if $T_{1:n} \sim \text{Loc}(\psi, q)$ then the sampling errors $\sqrt{n} \operatorname{Cov}_n(T_i, f(\psi_i)) = o_p(1)$ 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 3.7 (Table One)**.** Applied researchers often report tests of covariate balance in "table one." 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 approximation $\sqrt{n}\widehat{\beta} \Rightarrow \mathcal{N}(0, v)$.

By contrast, if $D_{1:n} \sim \text{Loc}(\psi, q)$ 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 covariate $f(\psi)$ to order $o_p(n^{-1/2})$.

Remark 3.8 (Realized Propensity Score)**.** Suppose sampling and assignment are completely randomized. For any set A with $P(\psi \in A) > 0$ define the *realized* sampling proportions in the set A by $\widehat{q}_A = E_n[T_i|\psi_i \in A]$. The discrepancy between expected and realized propensities $q - \widehat{q}_A = O_p(n^{-1/2})$, so q is implemented with errors of order $1/\sqrt{n}$, and similarly for the treatment proportions p. This can be seen in Figure , where the expected and realized propensities widely diverge in certain regions of the covariate space. Such fluctuations increase estimator variance. As in Hirano et al. (2003), 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, using the propensity weighting

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

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 error $q - \widehat{q}_{A_n} = o_p(n^{-1/2})$, even for a slowly shrinking sequence of sets $P(\psi \in A_n) \to 0$. 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 Propensity

This section describes the most general version of our method, providing finely stratified designs for non-constant sampling and assignment proportions $q(x)$, $p(x)$. We allow for stratification on different covariate sets ψ_1 and ψ_2 during the sampling and assignment stages, respectively. We also formally accommodate the case when any of ψ, $p(x)$, or $q(x)$ are random, for instance estimated using data from a pilot or proxy study.

Motivation. There are a range of logistical and welfare constraints that may require experimenters to use non-constant propensities, see Example 3.9 below. By implementing varying $q(x)$, $p(x)$ using local randomization, we can satisfy such constraints while also taking advantage of the nonparametric efficiency gains from fine stratification. In fact, varying propensities allow us to do strictly increase efficiency relative to the previous section. For example, if the cost of sampling units is heterogeneous, we would like to oversample cheap units and undersample expensive units, minimizing the variance of our estimator subject to a budget constraint. Building on the work in this section, we formalize and solve the problem of optimal stratification with heterogeneous costs in Section 4 below. In Section 5, we show how to use data from a pilot or proxy study to estimate the optimal design for the main experiment, providing the first asymptotically efficient solution to this problem.

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 $T_{1:n} \sim \text{Loc}(\psi, q(x))$ denote the following double stratification procedure:

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

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

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 $(T_i)_{i\in g} \sim \text{CR}(a_l/k_l)$ independently for each $g \in \mathcal{G}_l$. Treatment assignment $D_{1:n} \sim \text{Loc}(\psi, p(x))$ is defined identically, partitioning only the units $\{i : T_i = 1\}$ sampled into the experiment.

Example 3.9 (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 draw $T_{1:n} \sim \text{Loc}(\psi_1, 1/10)$, randomly sampling one village from each 10-tuple of villages, matched on publicly available covariates ψ_1. Before assigning treatments, we collect additional survey covariates ψ_2 in each sampled village $T_i = 1$. We label the villages using as those likely to benefit most M and least L from the intervention according to our prior. Local policymakers require $p(M) = 2/3$ 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.[12] 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 3.10 (CLT)**.** *Suppose Assumption 3.1 holds. Assume sampling and assignment* $T_{1:n} \sim \text{Loc}(\psi, q(\psi))$ *and* $D_{1:n} \sim \text{Loc}(\psi, p(\psi))$*. Then* $\sqrt{n_T}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$

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

If $q = 1$ then this is exactly the Hahn (1998) semiparametric variance bound for ATE with iid observations $(Y, D, \psi(X))$ and propensity $p(\psi)$. Note that the population IPW estimator is already semiparametrically efficient, and don't 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

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

variance due to treatment effect heterogeneity from $\text{Var}(c(\psi))$ to $\bar{q}\,\text{Var}(c(\psi))$ for $\bar{q} \in (0, 1]$.

Regression Equivalence. Theorem 3.10 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 $T_{1:n} \sim \text{Loc}(\psi, q(\psi))$ and $D_{1:n} \sim \text{Loc}(\psi, p(\psi))$ 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 3.11 (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 $T_i \overset{\text{iid}}{\sim} \text{Bernoulli}(q(\psi_i))$ and $D_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p(\psi_i))$ then $\sqrt{n_T}(\widehat{\theta}_{adj} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$ with the variance V as in Theorem 3.10.*

Intuitively, Proposition 3.11 shows that local randomization makes the unadjusted estimator $\widehat{\theta}$ as efficient as a much more involved estimator, with well-specified nonparametric adjustment for covariate imbalances in 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 define $\psi = \psi(X, \xi)$, $p = p(\psi, \xi)$, and $q = q(\psi, \xi)$. For example, we could let $\xi = (\widehat{m}_d)_{d=0,1}$ be regression estimates of $E[Y(d)|X = x]$ from a previous experiment and set $\psi(X, \xi) = (\widehat{m}_0(X), \widehat{m}_1(X))$. Our asymptotic theory will condition on ξ, allowing the study of stratification under "small pilot asymptotics" in the previous example. Define $c(\psi, \xi) = E[Y(1) - Y(0)|\psi, \xi]$ and $\sigma_d^2(\psi, \xi) = \text{Var}(Y(d)|\psi, \xi)$. The proof of Theorem 3.10 shows that if $T_{1:n} \sim \text{Loc}(\psi, q(\psi, \xi))$ and $D_{1:n} \sim \text{Loc}(\psi, p(\psi, \xi))$ then $\sqrt{n_T}(\widehat{\theta} - \text{ATE})|\xi \Rightarrow \mathcal{N}(0, V(\xi))$

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

All expectations and variances are conditional on ξ. In this setting, 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 $E[\exp(-\frac{1}{2}t^2 V(\xi))]$.

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 the most general version of Theorem 3.10. Suppose Assumption 3.1 holds. Assume 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

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

If $\psi_1 \neq \psi_2$, the variance V does not have a simple form like in the special cases above. Instead, we expand V relative to the efficient variance for the full vector of baseline covariates X. For example, if $q = 1$ the first line is the semiparametric variance bound for iid observations (Y, D, X). If $\psi_1 = \psi_2 = \psi$, and $q = q(\psi)$, $p = p(\psi)$ then V can be rearranged into the form in Theorem 3.10. The requirement that the stratification is increasing[13] $\psi_1 \subseteq \psi_2$ is subtle. We defer this technical discussion to Remark 9.2 in the appendix.

Optimal Stratification Variables. Equation 3.6 has the advantage of clearly exposing the minimal efficient stratification variables. Setting $\psi_1^* = c(X)$ and $\psi_2^* = (c(X), q(X), b(X; p(X)))$, the third and fourth variance terms are identically zero, achieving the minimal variance

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

We include (c, q) in ψ_2^* to satisfy the subvector condition $\psi_1 \subseteq \psi_2$. Since $b(X; p(X))$ 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. If $q(x) = q(\psi_1)$, we may exclude $q(\psi_1)$ from ψ_2^*. In practice, 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 $\psi_2(X)$ to be a supervector of $\psi_1(X)$ expected to be predictive of outcomes.[14] If $c(X) = g(\psi_1(X))$ for some function g, so that $c(X)$ only varies through the subvector $\psi_1(X)$, then this choice is asymptotically optimal. If none of the baseline covariates predict treatment effect heterogeneity, so $c(X)$ is constant, then there is no efficiency loss from completely randomized sampling and $\psi_1^* = 1$.

Remark 3.12 (Curse of Dimensionality)**.** From Equation 3.6, setting $\psi_1(X) = \psi_2(X) = X$ also minimizes the asymptotic variance V in Theorem 3.10. However, our analysis shows that if $E[Y(d)|X = x]$ is Lipschitz continuous then the finite sample variance converges to the asymptotic variance at rate $n\,\text{Var}(\widehat{\theta}) = V + O_p(n^{-2/(\dim(\psi)+1)})$, which may be slow even in moderate dimensions if $\psi(X) = X$. In finite samples, the variance $\text{Var}(\widehat{\theta})$ may be U-shaped in the dimension of the stratification variables, ordered by covariate relevance, since matching on many irrelevant variables reduces match quality

[13] In fact, we just require that $\psi_1 = g(\psi_2)$ for a measurable function g.

[14] 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.

on the relevant variables. This motivates the choice of stratification variables $\psi_1(x)$ and $\psi_2(x)$ of small dimension that minimize V.

4 Optimal Stratified Designs

This section characterizes asymptotically optimal stratification in survey experiments, including the optimal propensities for both sampling and assignment. For simplicity, we restrict to the setting of Theorem 3.10 with $\psi_1 = \psi_2 = \psi$ and $q = q(\psi)$, $p = p(\psi)$.

4.1 Budget Constrained Sampling Probem

Theorem 3.10 showed that $\sqrt{n_T}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$. Normalizing by the experiment size $n_T = \sum_i T_i$ allowed for easy comparison with the previous literature on iid sampling from a superpopulation. However, the experiment size n_T varies with $q(\psi)$, making this normalization unsuitable for minimization over $q(\psi)$ to study the optimal sampling propensity. In this section, we normalize instead by the number of eligible units n. Since $n_T/n \xrightarrow{p} E[q(\psi)]$, we have $\sqrt{n}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V(q,p))$

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

Clearly the optimal unconstrained sampling propensity is $q^*(x) = 1$, sampling everyone and making the experiment as large as possible. More generally, consider sampling with a budget constraint. To set up the problem, define the ex-ante heteroskedasticity function $\bar{\sigma}^2(\psi) = \sigma_1^2(\psi)/p(\psi) + \sigma_0^2(\psi)/(1-p(\psi))$. We interpret $\bar{\sigma}^2(\psi)$ as the expected residual variance from sampling a unit with $X_i = x$, prior to its treatment assignment $D_i \in \{0,1\}$ being realized. With this notation, 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. The budget-constrained variance minimization problem is

$$\min_{0<q\leq 1} E\left[\frac{\bar{\sigma}^2(\psi)}{q(\psi)}\right] \quad \text{s.t.} \quad E[C(\psi)q(\psi)] = \bar{C} \tag{4.2}$$

fixing the total budget $\bar{C}$. The next proposition characterizes the interior solutions to this problem.

Proposition 4.1. *Suppose* $\inf_x \sigma_d^2(\psi) \geq c > 0$. *Define the candidate solution*

$$q^*(\psi) = \bar{C} \cdot \frac{\bar{\sigma}(\psi)C(\psi)^{-1/2}}{E[\bar{\sigma}(\psi)C(\psi)^{1/2}]} \tag{4.3}$$

If $\sup_\psi q^*(\psi) \leq 1$, *then* q^* *is optimal in Equation 4.2.*

If the feasibility constraint $\sup_\psi q^*(\psi) \leq 1$ is violated, the optimal sampling propensity may not have a simple analytical form. Remark 5.6 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) Homogeneous costs. If $C(\psi) = 1$, then $E[q(\psi)] \leq \bar{C}$ constrains the total number of sampled units. In this case write $\bar{C} = \bar{q}$. The optimal propensity $q^*(\psi) = \bar{q}\bar{\sigma}(\psi)/E[\bar{\sigma}(\psi)]$ is proportional to the ex-ante standard deviation. In particular, $q^*(\psi) > \bar{q}$ (oversampling) if units of type $\psi(X) = \psi$ have larger ex-ante residual variance than the population average $E[\bar{\sigma}(\psi)]$, and undersampling in the opposite case.

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

Optimal Spending. Under optimal sampling, the total amount spent on units of type ψ is $C(\psi)q^*(\psi)dP(\psi) \propto \sqrt{\bar{\sigma}^2(\psi)C(\psi)}dP(\psi)$. We 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 $E[C(X)q(\psi)] = E[C(\psi)q(\psi)] \leq \bar{C}$.

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 $q(\psi)$, the global minimizer of $V(q, p)$ in Equation 4.1 is the conditional Neyman allocation $p^*(\psi) = \sigma_1(\psi)/(\sigma_1(\psi)+\sigma_0(\psi))$. We may also be interested in implementing a constant assignment propensity $p^* \in (0, 1)$. If q is constant, then

$$p^* = \sqrt{E[\sigma_1^2(\psi)]}\left(\sqrt{E[\sigma_1^2(\psi)]} + \sqrt{E[\sigma_0^2(\psi)]}\right)^{-1} \tag{4.4}$$

Compare this to the classical Neyman allocation $\sigma_1/(\sigma_1 + \sigma_0)$ with $\sigma_d = \mathrm{SD}(Y(d))$. By contrast, here only the residual variances $\sigma_d^2(\psi) = \mathrm{Var}(Y(d)|\psi)$ enter p^*. This is because the fluctuations of $Y(d)$ predictable by ψ 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 sampling and assignment optimal propensities

$$p^*(\psi) = \frac{\sigma_1(\psi)}{\sigma_1(\psi) + \sigma_0(\psi)} \qquad q^*(\psi; p^*) = \bar{C}\frac{(\sigma_1(\psi) + \sigma_0(\psi))C(\psi)^{-1/2}}{E[(\sigma_1(\psi) + \sigma_0(\psi))C(\psi)^{1/2}]}. \tag{4.5}$$

The propensities $p^*(\psi)$ and $q^*(\psi)$ will generally need to be discretized in order to implement them using fine stratification.

Definition 4.3 (Discretization)**.** Let $q_n^*(\psi)$ and $p_n^*(\psi)$ take values in the finite approximating propensity set $\{a_l/k_l : l \in L_n\}$ such that $|q_n^* - q^*|_\infty = o(1)$ and $|p_n^* - p^*|_\infty = o(1)$. Define $\overline{k}_n = \max_{l \in L_n} k_l$ and require $\overline{k}_n|L_n| = o(n^{1-\frac{d+1}{\alpha}})$ 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 $q_n^*(\psi) = \text{argmin}\{|q^*(\psi) - a/k_n| : 1 \leq a \leq k_n - 1, k_n = \lfloor n^{1/2-\epsilon} \rfloor\}$ 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* $T_{1:n} \sim \text{Loc}(\psi, q_n^*(\psi))$ *and* $D_{1:n} \sim \text{Loc}(\psi, p_n^*(\psi))$*. Then* $\sqrt{n}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V^*)$

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

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 $V^* = \min_{0\leq p\leq 1} V_H(p)$, where $V_H(p) = \text{Var}(c(\psi)) + E[\sigma_1^2(\psi)/p(\psi) + \sigma_0^2(\psi)/(1-p(\psi))]$ is the Hahn (1998) semiparametric efficiency bound for the ATE with propensity score $p(\psi)$ and iid treatments. As noted above, Armstrong (2022) shows that this bound also applies to the designs in this paper for $q = 1$.

4.3 Finite Sample Optimality

In this final subsection, we briefly discuss exact optimality in finite samples. Consider the case $q = 1$ and $p(X) = a/k$ constant. In this setting, Bai (2022) shows that if $b(X_i)$ were known, then matching units into strata of size k according to their sorted $b(X_i)$ values minimizes $\text{MSE}(\widehat{\theta}|X_{1:n})$ over all stratified designs. By contrast, we show that if $b(X_i)$ were known, the class of stratified designs itself is generally suboptimal. To see this, consider the case $p = 1/2$ and define the complete graph K_n with vertices $\{1, \ldots, n\}$ and edge weights $w_{ij} = b(X_i)b(X_j)$. Let $E_0^* \sqcup E_1^* = \{1, \ldots, n\}$ be a disjoint partition of the vertices that solves the Max-Cut optimization problem (e.g. Rendl et al. (2008)), maximizing the weight of cut edges between E_1 and E_0

$$\max_{E_0, E_1} \sum_{i,j} b(X_i)b(X_j)\mathbb{1}(i \in E_1, j \in E_0) \quad \text{s.t.} \quad E_0 \sqcup E_1 = \{1, \ldots, n\} \tag{4.6}$$

The optimal treatment allocation $d_{1:n}^* = d_{1:n}^*(X_{1:n})$ is $d_i^* = \mathbb{1}(i \in E_1^*)$ 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* $\text{MSE}_{P^*}(\widehat{\theta}|X_{1:n}) \leq \text{MSE}_P(\widehat{\theta}|X_{1:n})$ *for all* $P \in \mathcal{P}_{1/2}$*.*

The inequality is strict if Problem 4.6 has a unique solution up to permutation of set labels. In particular, note that P^* is not a matched pairs design. Nevertheless, $b(X_i)$ 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 5.1 below shows that such approaches are equivalent to regression adjustment with regressions estimated in the pilot instead of the main sample, which may be much noisier 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

Suppose that data from a pilot or proxy experiment with related outcomes is available, as in Hahn et al. (2011). The theory in Section 3 suggests using this data to (1) estimate the optimal stratification variables ψ^* for sampling and assignment or (2) estimate the efficient sampling and assignment propensities $q^*(x)$ and $p^*(x)$. We consider both of these approaches in turn. For simplicity, we restrict to the case with equal sampling and assignment variables $\psi_1 = \psi_2 \equiv \psi$.

5.1 Estimating Stratification Variables

In Section 3, we showed that the variables $\psi^* = (c, b)(X)$ and $\psi^* = (m_0, m_1)(X)$ were asymptotically efficient for both sampling and assignment. This suggests setting $\psi(X) = (\widehat{c}, \widehat{b})(X)$ or $\psi(X) = (\widehat{m}_0, \widehat{m}_1)(X)$, 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).

Comparison with Regression Adjustment. Our first result is negative, suggesting that such an approach would be dominated by throwing away the pilot data and using an iid design with regression adjustment in the main sample. For simplicity, consider $q = 1$ and let $D_{1:n} \sim \text{Loc}(\widehat{b}, p)$ be the design with estimated stratification variables. Let iid treatments $\check{D}_i \sim \text{Bernoulli}(p)$ and $\widehat{\theta}_{adj}$ be the cross-fit AIPW estimator of Proposition 3.11, estimated using the main experiment data $\check{W}_{1:n} = (\check{D}_i, X_i, Y_i(\check{D}_i))_{i=1}^n$, with regression estimators $\check{m}_d(\psi)$. The estimators $\widehat{\theta}$ and $\widehat{\theta}_{adj}$ have identical expansions

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

for a constant c_p. The residual terms $\sqrt{n}R_n, \sqrt{n}\check{R}_n \Rightarrow \mathcal{N}(0, v)$ for the same variance $v > 0$. Denote the middle terms by $B_n = E_n[(D_i - p)(b - \widehat{b})(X_i)]$ and $\check{B}_n = E_n[(\check{D}_i - p)(b - \check{b})(X_i)]$. These terms parameterize the error due to covariate imbalance between treatment and control units. Denote the 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 that $\sqrt{n}B_n = O_p(r_n^{pilot})$ while $\sqrt{n}\check{B}_n = O_p(r_n^{main})$. We expect the pilot estimation error is larger $r_n^{pilot} \gg r_n^{main}$ if the pilot is much smaller than the main sample. This shows that the design $D_{1:n} \sim \text{Loc}(\widehat{b}, p)$ behaves like a noisy version of the usual AIPW estimator $\widehat{\theta}_A$, with regressions estimated in the pilot instead of the main experiment.

Remark 5.1 (Model-Based Adjustment)**.** The preceding discussion shows that the design $D_{1:n} \sim \text{Loc}(\widehat{b}, p)$ makes the unadjusted estimator $\widehat{\theta}$ behave like an AIPW estimator, up to lower order factors. However, the AIPW regression adjustments are estimated using data from the pilot sample instead of the main experiment, leading to worse control

over the covariate imbalance term B_n. Instead, we suggest using a model-free strategy at design time, drawing $T_{1:n} \sim \text{Loc}(\psi, q(\psi))$ and $D_{1:n} \sim \text{Loc}(\psi, p(\psi))$ for a small set of covariates expected to be most predictive of outcomes and treatment effect heterogeneity. Model-based covariate adjustments such as AIPW or other regresssion adjustments can always be applied ex-post, using the main experiment data (or pooled data) to estimate the relevant regressions more precisely.

Residual Variance. Setting $\psi(X) = X$ in Equation 3.2 shows that under complete randomization $T_{1:n} \sim \text{CR}(q)$ and $D_{1:n} \sim \text{CR}(p)$ then $\sqrt{n_T}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$ with

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

The middle term is the variance due to covariate imbalance, and the third term is the residual variance. We can think of $\text{Var}(b(X; p))$ as the "easier" term. We can make this term zero by any one of the following: (1) fine stratification on $\psi(X) = X$ under weak assumptions, (2) ex-post propensity re-weighting, or (3) ex-post regression adjustment under well-specification, 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

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

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 propensity at design-time. Motivated by this, the next section studies pilot estimation of the optimal sampling and assignment propensities.

5.2 Feasible Optimal Stratification

Consider estimating the optimal stratified design for the budget-constrained sampling problem in Equation 4.2. As in Section 4, restrict to the case $\psi_1 = \psi_2 = \psi$ for simplicity. This amounts to using pilot 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 $\sigma_d(\psi)$. We show that the estimated optimal design is asymptotically optimal in the sense of Theorem 4.4. The performance of this procedure under fixed pilot asymptotics is described by Equation 3.5. We discuss more small sample considerations in Remark 5.4 below. Section 4 showed that the budget constrained efficient design problem in Equation 4.2 is solved by the propensities

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

provided the feasibility condition $\sup_\psi q^*(\psi) \leq 1$ is satisfied. Consider pilot heteroskedasticity estimates[15] $\widehat{\sigma}_d^2(\psi)$ for $d = 0, 1$. Define the propensity estimates[16]

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

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, introduced in Remark 5.6 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 5.2. *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* $|\sigma_d^2 - \widehat{\sigma}_d^2|_{2,\psi} = O_p(n^{-r})$ *for some* $r > 0$. *Require discretization rate* $\bar{k}_n|L_n| = o(n^{1-(\dim(\psi)+1)/\alpha_1})$ *and* $\bar{k}_n = \omega(n^{1/\alpha_2})$. *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 5.3 (Pilot Design)**.** *Impose Assumption 5.2. Suppose* $T_{1:n} \sim \text{Loc}(\psi, \widehat{q}_n(\psi))$ *and* $D_{1:n} \sim \text{Loc}(\psi, \widehat{p}_n(\psi))$. *Then* $\sqrt{n}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V^*)$

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

For intuition, it is also helpful to consider certain special cases of Theorem 5.3. If we fix $q = 1$, the design $D_{1:n} \sim \text{Loc}(\psi, \widehat{p}_n(\psi))$ asymptotically minimizes the Hahn (1998) variance bound over all propensity scores

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

If $\widehat{p}^*$ is a consistent pilot estimate of the optimal constant propensity p^* in Equation 4.4, then the design $D_{1:n} \sim \text{Loc}(\psi, \widehat{p}_n)$ has $\sqrt{n}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$ with

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

Remark 5.4 (Small Pilots)**.** For small pilots, the asymptotic results in Theorem 5.3 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

[15] For example see Fan and Yao (1998).

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

$\{D_i = d\}$. 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 estimated design variable version of Theorem 3.10, if $D_{1:n} \sim \text{Loc}(\psi, \widehat{p})$ then $\sqrt{n}(\widehat{\theta}-\text{ATE})|W_{1:m} \Rightarrow \mathcal{N}(0, V)$

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

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 5.5 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 5.5 (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 5.2 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 $\widehat{p} = 13/20$.

Remark 5.6 (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

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

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

Remark 5.7 (Optimal Stratification Trees)**.** Tabord-Meehan (2022) suggests using pilot data to estimate a stratification $\widehat{S}$ and assignment propensity $\widehat{p}(S)$ over a set of tree partitions $\mathcal{T}$ of the covariate space. In our notation, their Theorem 3.1 shows that if $D_{1:n} \sim \text{Loc}(\widehat{S}, \widehat{p}(S))$ then $\sqrt{n}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$ with asymptotic variance

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

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 globally 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).

6 Inference Methods

This section provides new methods for asymptotically exact inference on the ATE under locally randomized designs, allowing researchers to report smaller confidence intervals that fully reflect the efficiency gains from finely stratified samping and assignment. Restrict to the case with $\psi_1 = \psi_2 = \psi$ for simplicity. We define a generalized pairs-of-pairs estimator (Abadie and Imbens (2008)), accommodating finely stratified sampling and assignment with varying propensities $q(\psi), p(\psi)$. 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 $w_i^1 = (1 - p_i q_i)/(p_i q_i)^2$ and $w_i^0 = (1 - q_i(1-p_i))/(q_i(1-p_i))^2$. Finally, define the variance estimator components

$$\begin{aligned}
\widehat{v}_1 &= n^{-1} \sum_{g \in \mathcal{G}_n^\nu} \frac{1}{a(g) - 1} \sum_{i \neq j \in g} Y_i Y_j D_i D_j (w_i^1 w_j^1)^{1/2} \\
\widehat{v}_0 &= n^{-1} \sum_{g \in \mathcal{G}_n^\nu} \frac{1}{(k-a)(g) - 1} \sum_{i \neq j \in g} Y_i Y_j (1 - D_i)(1 - D_j)(w_i^0 w_j^0)^{1/2} \\
\widehat{v}_{10} &= n^{-1} \sum_{g \in \mathcal{G}_n} \frac{k}{a(k-a)}(g) \sum_{i,j \in g} Y_i Y_j D_i (1 - D_j)(q_i q_j)^{-1/2}
\end{aligned}$$

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 the 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 3.10, enabling asymptotically exact inference.

Theorem 6.1 (Inference)**.** *Assume the conditions of Theorem 3.10. If sampling* $T_{1:n} \sim \mathrm{Loc}(\psi, q(\psi))$ *and assignment* $D_{1:n} \sim \mathrm{Loc}(\psi, p(\psi))$*, then* $\widehat{V} = V + o_p(1)$.

Theorem 6.1 shows that the confidence interval $\widehat{C} = \widehat{\theta} \pm \widehat{V}^{1/2} c_{\alpha/2} / \sqrt{n}$ with $c_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ is asymptotically exact in the sense that $P(\mathrm{ATE} \in \widehat{C}) = 1 - \alpha + o(1)$. Note the scaling is by number of eligible units n, not the experiment size $n_T = \sum_i T_i \leq n$.

7 Empirical Results

This section examines the finite sample properties of the estimation and inference methods studied above through a series of monte carlo simulation studies.

7.1 Simulations

Our first set of simulations focuses on the efficiency gain and inference properties of locally randomized *selection* into the experiment. Let $X_i \in \mathbb{R}^2$ and consider quadratic potential outcome models of the form

$$Y_i(0) = X_i'\beta(0) + \sigma_1^2(X_i)\epsilon_i(0) \qquad Y_i(1) = X_i'\beta(1) + X_i'AX_i + \sigma_0^2(X_i)\epsilon_i(1)$$

We sample from the following DGP's

> **Model 1**: $\beta(0) = (1,1)$, $\beta(1) = (2,2)$, $A_{12} = A_{21} = 1/2$, $A_{11} = A_{00} = 0$, with residual variance $\sigma_1^2 = \sigma_0^2 = 0.1$
>
> **Model 2**: As in (1) but with $\beta(0) = (1,1)$, $\beta(1) = (2,2)$, $A = 0$
>
> **Model 3**: As in (1) but with $\beta(0) = \beta(1) = (1,1)$, $A = 0$

We let $X_{ij} \overset{\text{iid}}{\sim} 2(\text{Beta}(2,2) - 1/2)$ and $\epsilon_i(d) \sim \mathcal{N}(0,1)$.

Designs - In Table 1 we vary n, the number of units eligible for selection, fixing the experiment size $qn = \sum_i T_i = 100$. To highlight the marginal efficiency gains from representative selection, while also using locally randomized treatment assignment, we *fix* the assignment procedure to be "matched triples" on $\psi(x) = (x_1, x_2)$ with $p = 2/3$, as in Example 3.3. In our notation, we let selections $T_{1:n}$ and treatment assignments $D_{1:n}$ be given by

(1) $T_{1:n} \sim \text{Loc}_\text{n}(\psi, q)$ (selection) with $n = 100/q$

(2) $D_{1:n} \sim \text{Loc}_\text{n}(\psi, 2/3)$ (treatment assignment)

Evaluation Criteria - Table 1 presents metrics evaluating both the efficiency gains from locally-randomized selection, as well as the power and validity of our inference procedure. Note that $n = 100$ is the usual case of random selection. We evaluate each metric relative to this least efficient benchmark design using 1000 monte carlo replications drawn from the DGP's above. %ΔSD is the change in estimator standard deviation relative to random selection. ESS is the *effective sample size*. This is the size of an experiment with units selected completely at random needed to achieve the same variance as representative selection. Algebraically, $\text{ESS} = \frac{100}{1+\%\Delta\text{Var.}}$. For inference, %ΔLength is the reduction in confidence interval (CI) length, relative to the CI length under random selection.

Results - The variance reduction increases as the number of eligible units $n = 100/q$ is larger (q is smaller). This is expected from Theorem 3.10, which gives a reduction asymptotic variance of $-(1-q)\,\text{Var}(c(X))$ from local randomization. The boost in effective sample size (ESS) is significant relative to completely random selection. Note that this comes essentially "for free" by being slightly more careful about choosing which units to enroll in the experiment. In Model 3, $c(x) = 0$ identically, so the theory predicts no efficiency gain due to selection. The small variance reduction seen in practice is likely a

finite sample effect where balanced first-stage selection facilitates better matches during the second-stage assignment process.

		Efficiency		Inference	
Model	n	%ΔSD	ESS	Coverage	%ΔLength
1	100	0.0	100	95	0.0
	200	-17.4	147	97	-10.3
	300	-22.8	168	97	-13.6
	400	-26.4	185	97	-15.9
2	100	0.0	100	94	0.0
	200	-17.7	148	96	-9.5
	300	-25.1	178	98	-12.4
	400	-28.6	196	97	-14.5
3	100	0.0	100	96	0.0
	200	-0.4	101	95	-4.2
	300	-5.5	112	96	-4.9
	400	-10.6	125	97	-5.6

Table 1: Effect of Representative Selection

Table 1 suggests our inference procedure is generally slightly conservative in finite samples. This is likely due to the between-group matching used for inference in Section 6 giving slightly worse matches than the groups produced by the design itself. For CI length, finite-sample exact inference would have %ΔLength = %ΔSD. In finite samples, our inference procedure gives a significant reduction in CI length, reflecting the increase in estimator precision due to selection.

Our second set of simulations focuses on the choice of $\psi(x)$. Asymptotically, including more baseline covariates always weakly improves estimator efficiency. However, as discussed in Section 3, the within-group matching discrepancies converge slowly in high dimensions, giving balancing rate $r_n^\psi = O(n^{-2/(d+1)})$. This suggests preferentially including covariates that are predictive of outcomes and treatment effect heterogeneity (predict functions $b_n(x)$ and $c(x)$), excluding "noise covariates" that contain little additional information. Consider the quadratic model and variable distributions above, but with $X_i \in \mathbb{R}^{10}$ and

$$\beta(0) = (1, 1, 1, .4, .4, .4, 0, 0, 0, 0) \qquad \beta(1) = 2\beta(1) \qquad A = (1/20)(\mathbb{1}\mathbb{1}' - \text{diag}(\mathbb{1}))$$

The first three covariates are important. The next three are less important. The last four provide no information about outcomes. We also include nonlinear interaction terms. For each $\psi(x)$ considered below, we select $q = 1/2$ of the eligible units n by matched pairs, then use matched triples for treatment assignment. Formally, we have

(1) $T_{1:n} \sim \text{Loc}_\text{n}(\psi, 1/2)$ (selection)

(2) $D_{1:n} \sim \text{Loc}_n(\psi, 2/3)$ (treatment assignment)

In Table 2 below, $\psi^*(x) = (b_n(x), c(x))$, the optimal design for joint selection and assignment given in Section 4. The evaluation criteria are the same as above, comparing to the least efficient design with selection and assignment done by complete randomization (CR).

		Efficiency		Inference	
(n, qn)	$\psi(x)$	%ΔSD	ESS	Coverage	%ΔLength
$(200, 100)$	CR	0.0	100	95	0.0
	x_1, x_2	-28.7	197	96	-23.5
	$x_1 \dots x_4$	-37.4	256	97	-28.2
	$x_1 \dots x_6$	-36.0	244	97	-23.5
	$x_1 \dots x_8$	-33.9	229	98	-19.6
	$\psi^*(x)$	-57.1	543	95	-53.0
$(400, 200)$	CR	0.0	100	95	0.0
	x_1, x_2	-27.2	188	96	-24.6
	$x_1 \dots x_4$	-40.4	282	98	-32.0
	$x_1 \dots x_6$	-36.5	248	98	-28.0
	$x_1 \dots x_8$	-33.2	224	98	-22.8
	$\psi^*(x)$	-56.7	534	97	-54.1

Table 2: Varying $\psi(x)$

Results - There are significant precision gains from locally randomized selection and assignment, as predicted by Theorem 3.10. Theorem 6.1 shows that our inference procedures are asymptotically exact, but we get slight conservativeness in finite samples, likely due to between-group match quality during inference being worse than the within-group matches. Note that, contrary to the asymptotic theory, including $x_1 \dots x_6$ is slightly *worse* than $x_1 \dots x_4$ in finite samples. Recall have the largest weight in the outcome model. Additionally including $x_4 \dots x_6$, which are less predictive of outcomes, degrades match quality for the important covariates $x_1 \dots x_3$, reducing efficiency in finite samples. Including noise covariates x_7, x_8 performs still worse. As in Theorem 4.4, the optimal design $\psi^*(x) = (b_n(x), c(x))$ gives the largest efficiency gains.

8 Conclusion

This paper proposes a flexible new family of designs that randomize within maximally homogeneous local groups. We show that experimenting on a representative sample of units increases estimator precision, providing a practical implementation by locally randomized sampling. We give novel asymptotically exact inference methods for covariate-adaptive sampling and assignment, allowing researchers to shrink their confidence intervals if they use our methods to design a representative experiment. We also applied our methods to the setting of design with a pilot experiment. By using pilot data to estimate the optimal

treatment proportions, then locally randomizing within estimated propensity strata, we construct the first asymptotically fully efficient design in this setting.

This line of research can be extended in several interesting directions. It is clear how to extend our methods to $k > 2$ treatments, though this case is not included in our analysis. Accommodating continuous treatments is more difficult, given the discrete nature of the assignment process. Doing so would enable more efficient estimation of marginal effects and dose-response curves. We implicitly allow cluster-randomized trials by treating each cluster as a separate experimental unit. It could be helpful to practitioners to study this example more formally, explicitly accounting for heterogeneous cluster size. The design ordering we propose may not be optimal. For instance, we could reverse the process, first matching units into assignment pairs, then selecting a representative sample of pairs into the experiment. The current ordering emphasizes representativeness at the cost of match quality during treatment assignment. We suspect these variations to be asymptotically equivalent, though they may differ in finite samples.

References

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

Abadie, A., G. W. Imbens, and F. Zheng (2014). Inference for misspecified models with fixed inference for misspecified models with fixed regressors. *Journal of the American Statistical Association 109*(508).

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.

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*.

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.

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

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*.

Caria, S. A., G. Gordon, M. Kasy, S. Quinn, S. Shami, and A. Teytelboym (2021). An adaptive targeted field experiment: Job search assistance for refugees in jordan. Working Paper.

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.

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

Cytrynbaum, M. (2022). Rerandomization in finely stratified experiments. Working Paper.

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

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

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika 58*(3), 403–417.

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

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain 33*, 503–513.

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.

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

Higgins, M. J., F. Sävje, and J. S. Sekhon (2015). Improving massive experiments with threshold blocking. *Proceedings of the National Academy of Sciences*.

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).

Hu, F. and W. F. Rosenberger (2006). *The Theory of Response-Adaptive Randomization in Clinical Trials*. John Wiley and Sons.

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

Kapelner, A. and A. Krieger (2014). Matching on-the-fly: Sequential allocation with higher power and efficiency. *Biometrics 70*, 378–388.

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).

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.

Morgan, K. L. and D. B. Rubin (2012). Rerandomization to improve covariate balance in experiments. *Annals of Statistics 40*(2).

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.

Russo, D. (2016). Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*.

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.

9 Appendix

9.1 Finite Population Estimands

If the eligible units $\{1, \ldots, n\}$ 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 $\sigma_\tau^2(X) = \text{Var}(Y(1) - Y(0)|X)$. 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

$$V_{\text{SATE}} = E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1-p} - q\sigma_\tau^2(X)\right]. \tag{9.1}$$

The final variance component $E[\sigma_\tau^2(X)]$ 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 $\sigma_\tau^2(X)$ 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

$$\widehat{\theta} - \text{SATE} = \text{Cov}_n(T_i, Y_i(1) - Y_i(0))/q + \text{Cov}_n(D_i, \bar{Y}_i \,|T_i = 1) \tag{9.2}$$

The first term parameterizes the errors due to correlation between the sampling variables and treatment effects, and naturally increases with $\sigma_\tau^2(X) = \text{Var}(Y(1) - Y(0)|X)$. The second term is increasing in $\text{Cov}(Y(1), Y(0)|X)$ (decreasing in $\sigma_\tau^2(X)$) and larger in mean square, resulting in a net negative dependence on $\sigma_\tau^2(X)$. As $q \to 0$, the relative variance[17] due to sampling increases, exactly cancelling the $E[\sigma_\tau^2(X)]$ 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 $\sigma_\tau^2(X) \geq \sigma_1^2(X) + \sigma_0^2(X) - 2\sigma_1(X)\sigma_0(X) \geq 0$. We implement such a strategy in Section 6.

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}.[18] 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 $V_c = E[\sigma_1^2(X)/p + \sigma_0^2(X)/(1-p)] \geq V_{\text{SATE}}$.

Remark 9.1 (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

[17] 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.

[18] 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.

9.2 Remarks

Remark 9.2 (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

$$P(|\psi_{1,i} - \psi_{1,j}| < \epsilon) > P(|\psi_{1,i} - \psi_{1,j}| < \epsilon \,|\, T_i = T_j = 1). \tag{9.3}$$

If $E[Y_i(d)|\psi_{1,i}] \neq 0$, such changes to the joint distribution will show up in the conditional variance $\text{Var}(\widehat{\theta}|X_{1:n}, T_{1:n})$ in complicated ways that depend on the details of the matching 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 9.3 are lower order, but leave the detailed study of this issue for specific matching procedures to future work.

10 Proofs

10.1 Matching

Theorem 10.1 (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*

$$n^{-1} \sum_{g \in \mathcal{G}_n} \frac{1}{|g|} \sum_{i,j \in g} |\psi_{i,n} - \psi_{j,n}|_2^2 \leq (1 \vee \max_{i=1}^{n} |\psi_{i,n}|_2^2) \cdot O((n/\overline{k}_n |L_n|)^{-2/(d+1)}). \tag{10.1}$$

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 10.1.*

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 $a = 1, \ldots, |L_n|$. By construction, this creates groups $\mathcal{G}_n = \{g_{a,s} : 1 \leq a \leq |L_n|, 1 \leq s \leq n\}$ with the ordering property

$$l(i) \leq l(j) \quad \forall i \in g_{a,s}, j \in g_{a,s'} \quad s < s' \quad a = 1, \ldots, |L_n| \tag{10.2}$$

Fix an indexing of all within-group pairs $(\mathfrak{p}_{a,s,t})_{t=1}^{k_a^2 - k_a} \equiv \{(i,j) : i \neq j;\ i,j \in g_{a,s}\}$, 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

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

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

$$n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a^{-1} \sum_{t=1}^{k_a^2 - k_a} d_{a,s,t}^2 \mathbb{1}(E_{a,s,t}) \mathbb{1}(g_{a,s} \neq \emptyset) \leq \frac{d}{nm^2} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a^{-1} \sum_{t=1}^{k_a^2 - k_a} \mathbb{1}(g_{a,s} \neq \emptyset)$$
$$\leq \frac{d}{nm^2} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a \mathbb{1}(g_{a,s} \neq \emptyset) \leq \frac{d}{m^2}$$

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

(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 $l(i_{a,s,t}) \leq l \leq l(j_{a,s,t})$, 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 $|x - y|_2 \leq 2\sqrt{d}/m$. Then by telescoping and triangle inequality, on the event $E_{a,s,t}^c$

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

Note also that if $x, y \in [0,1]^d$ then $|x-y|_2^2 = d(|x-y|_2/\sqrt{d})^2 \leq \sqrt{d}|x-y|_2$, using $c^2 \leq c$ for $0 \leq c \leq 1$. In particular, we have $d_{a,s,t}^2 \leq \sqrt{d} \cdot d_{a,s,t}$.

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

The second inequality follows by the ordering property in equation 10.2 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

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

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. □

10.2 CLT

Lemma 10.2 (Balancing). *Suppose $T_{1:n} \sim \text{Loc}(\psi_1, q(x))$ and $D_{1:n} \sim \text{Loc}(\psi_2, p(x))$ 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 $\mathcal{L}_n = \{g(Z) \in L_2(Z) : |g|_{lip} \vee |g|_\infty \le c_n\}$ and let $g_n \in \mathcal{L}_n$ be such that $E[(g_n(Z) - f(Z))^2] \le 2\inf_{g\in\mathcal{L}_n} E[(g(Z) - f(Z))^2]$. We claim that $|g_n - f|_{2,Z}^2 \to 0$. Let $\epsilon > 0$ and note that by Lemma 10.11 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 10.17 $E[A_n|\mathcal{F}_n] = 0$ and

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

The equality by conditional Markov (Lemma 10.10) 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 10.17, we have $E[B_n|\mathcal{F}_n] = 0$. Using the claim

and Lemma 10.17, we calculate

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

The second to last equality since $p_i = p_j$ for any $i,j \in g$, $\forall g$ and by Theorem 10.1, using the fact that $\max_{i=1}^n |\psi_{2,i}|^2 = o_p(n^{2/\alpha_2})$ by Lemma 10.14. 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 10.3 (CLT)**.** *Consider the following assumptions*

(A1) $E[Y(d)^2] < \infty$ *and* ψ_1, ψ_2, q, p *as in Theorem 3.10 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* $\widehat{q}_{i,n}, q_i \geq \delta$. *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 10.4 (CLT)**.** *(1) Suppose Assumption 10.3 (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* $\sqrt{n}E_n[T_i u_i]|\mathcal{F}_{0,n} \Rightarrow \mathcal{N}(0, V_1)$ *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 10.3 (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 $(\mathcal{H}_{k,n}, c_{1,k} - \text{ATE})_{k\geq 1}$ is an MDS. Adaptation is clear. For the MDS property, $E[c_{1,i}|\mathcal{F}_{i-1,n}] = E[c_{1,i}|\xi, (c_{1,k})_{k=1:i-1}] = E[c_{1,i}|\xi] = E[Y_i(1) - Y_i(0)|\xi] = \text{ATE}$ by tower law and independence. Similarly, $E[(c_{1,i} - \theta)^2|\mathcal{F}_{i-1,n}] = E[(c_{1,i} - \theta)^2|\xi] = \text{Var}(c_{1,i}|\xi)$. 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 10.6 $\sqrt{n}(E_n[c_{1,i}] - \text{ATE})|\xi \Rightarrow \mathcal{N}(0, V_1)$ 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 10.6. 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 10.17. For the second term B_n, we have

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

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 10.3. 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

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

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 10.6.

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

By our propensity lower bounds we have $D_i\sigma_1^2(X_i)/(\widehat{p}_{i,n}\widehat{q}_{i,n})^2 \leq \delta^{-4}\sigma_1^2(X_i)$ 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 10.17. 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 $\Sigma_n = E_n[\sigma_1^2(X_i)/\widehat{p}_{i,n}\widehat{q}_{i,n}] + o_p(1)$.

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

Then $E_n[T_i D_i \sigma_1^2(X_i)/(\widehat{p}_{i,n}\widehat{q}_{i,n})^2] = E_n[\sigma_1^2(X_i)/p_i q_i] + o_p(1) = E[\sigma_1^2(X_i)/p_i q_i] + O_p(n^{-1/2})$ by Chebyshev inequality. By symmetry, the same holds for the $D = 0$ term. Putting this together, we have shown that $\Sigma_n \xrightarrow{p} E[\sigma_1^2(X_i)/(p_i q_i) + \sigma_0^2(X_i)/(1-p_i)q_i]$ as claimed.

Finally, we show the Lindberg condition. Note the bound $|z_{i,n}|^2 \leq 2\delta^{-4}((\epsilon_i^1)^2+(\epsilon_i^0)^2) \equiv v_i$, with $E[v_i] \lesssim \text{Var}(Y(1)) + \text{Var}(Y(0)) < \infty$. Then for $\epsilon > 0$ we have

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

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. □

Suppose that $E[Y(d)^2] < \infty$.

Proof of Theorem 3.10. 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 $1 \leq k \leq 4$

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

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

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

We wish to apply Lemma 10.2 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

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

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 10.4 are satisfied by assumption. Then $\sqrt{n}A_n|\xi \Rightarrow \mathcal{N}(0, V_1)$ with $V_1 = \text{Var}(c_{1,i}|\xi)$, $\sqrt{n}C_n|\mathcal{F}_{2,n} \Rightarrow \mathcal{N}(0, V_2)$ with $V_2 = E[q_i^{-1}\,\text{Var}(c_{2,i}|\psi_{1,i}, q_i, \xi)|\xi]$, 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 $\sigma_{d,i}^2 = \text{Var}(Y(d)|\psi_{2,i}, p_i, \xi)$. Then by Slutsky and Lemma 10.12, 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 $\text{Var}(c(X)|\xi) = \text{Var}(E[c(X)|\psi_1, q, \xi]|\xi) + E[\text{Var}(c(X)|\psi_1, q, \xi)|\xi]$,

with $E[c(X_i)|\psi_{1,i}, q_i, \xi] = c_{1,i}$. Then since $\xi \perp\!\!\!\perp W_{1:n}$

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

Next, by the law of total variance

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

Using this equality and applying tower law gives

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

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

$$\begin{aligned}
V_3 &= E[(q_i p_i)^{-1} \sigma^2_{1,i}|\xi] + E[(q_i(1-p_i))^{-1} \sigma^2_{0,i}|\xi] \\
&= E[(q_i p_i)^{-1} \text{Var}(m_1(X_i)|\psi_{2,i}, p_i, \xi)|\xi] + E[(q_i p_i)^{-1} E[\sigma^2_1(X_i)|\psi_{2,i}, p_i, \xi]|\xi] \\
&+ E[(q_i(1-p_i))^{-1} \text{Var}(m_0(X_i)|\psi_{2,i}, p_i, \xi)|\xi] + E[(q_i(1-p_i))^{-1} E[\sigma^2_0(X_i)|\psi_{2,i}, p_i, \xi]|\xi] \\
&= E[(q_i p_i)^{-1} \text{Var}(m_1(X_i)|\psi_{2,i}, p_i, \xi)|\xi] + E[(q_i(1-p_i))^{-1} \text{Var}(m_0(X_i)|\psi_{2,i}, p_i, \xi)|\xi] \\
&+ E[(q_i p_i)^{-1} \sigma^2_1(X_i)|\xi] + E[(q_i(1-p_i))^{-1} \sigma^2_0(X_i)|\xi]
\end{aligned}$$

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

$$\begin{aligned}
&V_1 + V_2 + V_3 = \text{Var}(c(X_i)) - E[\text{Var}(c(X_i)|\psi_{1,i}, q_i, \xi)|\xi] \\
&+ E[q_i^{-1} \text{Var}(c(X_i)|\psi_{1,i}, q_i, \xi)|\xi] - E[q_i^{-1} \text{Var}(c(X)|\psi_{2,i}, p_i, \xi)|\xi] \\
&+ E[(q_i p_i)^{-1} \text{Var}(m_1(X_i)|\psi_{2,i}, p_i, \xi)|\xi] + E[(q_i(1-p_i))^{-1} \text{Var}(m_0(X_i)|\psi_{2,i}, p_i, \xi)|\xi] \\
&+ E[(q_i p_i)^{-1} \sigma^2_1(X_i)|\xi] + E[(q_i(1-p_i))^{-1} \sigma^2_0(X_i)|\xi] \\
&= T_1 + T_2 + T_3 + T_4 + T_5 + T_6 + T_7 + T_8
\end{aligned}$$

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

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

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

$$\begin{aligned}
V = \sum_{k=1}^{8} T_k &= \text{Var}(c(X)) + E[q_i^{-1}(1-q_i) \text{Var}(c(X_i)|\psi_{1,i}, q_i, \xi)|\xi] \\
&+ E[q_i^{-1} \text{Var}(b(X_i; p_i)|\psi_{2,i}, p_i, \xi)|\xi] + E\left[q_i^{-1} \left(\frac{\sigma^2_1(X_i)}{p_i} + \frac{\sigma^2_0(X_i)}{1-p_i}\right) \middle| \xi\right].
\end{aligned}$$

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 10.2, and the third by conditional Chebyshev (Lemma

10.10). 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 □

Definition 10.5 (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 10.6 (MDS-CLT)**.** *Consider probability spaces $(\Omega_n, \mathcal{G}_n, P_n)$ each equipped with filtration $(\mathcal{F}_{k,n})_{k\geq 0}$. Suppose $(Y_{k,n})_{k=1}^n$ is adapted to $(\mathcal{F}_{k,n})_{k\geq 0}$ and has $E[Y_{k,n}|\mathcal{F}_{k-1,n}] = 0$ for all $k \geq 1$ with $n \to \infty$. Make the following definitions*

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

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

$$L^\epsilon_n = \sum_{k=1}^{n} E[Y^2_{k,n}\mathbb{1}(|Y_{k,n}| \geq \epsilon)|\mathcal{F}_{0,n}] = o_p(1) \tag{10.3}$$

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).

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

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 $E[Y_{k,n}^2 | \mathcal{F}_{k-1,n}] = \sigma_{k,n}^2$, combined with these facts

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

Then we have

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

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

10.3 Optimal Stratification and Pilot Design

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

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

and recall the candidate optimal solution $q^*(\psi) = \bar{C} \cdot \bar{\sigma}(\psi)C(\psi)^{-1/2} / E[\bar{\sigma}(\psi)C(\psi)^{1/2}]$. 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 $V(q^*) = \bar{C}^{-1} E[\bar{\sigma}(\psi)C(\psi)^{1/2}]^2 < \infty$. Also clearly $E[C(\psi)q^*(\psi)] = \bar{C}$. 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 $\bar{C} = E[q^*(\psi)C(\psi) + t\Delta(\psi)C(\psi)] = \bar{C} + tE[\Delta(\psi)C(\psi)]$, 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.

□

Suppose that $\inf_\psi \sigma_d^2(\psi) \wedge \inf_\psi \widehat{\sigma}_d^2(\psi) \geq c_l > 0$. Assume $\text{Var}(Y(d)|\psi) \vee \widehat{\sigma}_d^2(\psi) < c_u < \infty$. Assume $|C|_\infty < \infty$ and $\inf_\psi C(\psi) > 0$.

Oracle - Assume $\inf_\psi \bar{\sigma}^2(\psi) > 0$. Assume that $\overline{k}_n |L_n| = o(n^{1 - \frac{d+1}{\alpha_1}})$. Assume that $E[Y(d)^4] < \infty$. Assume that $\sigma_1(\psi)/\sigma_0(\psi) \in [c_l, c_u] \subseteq (0, \infty)$ a.s. Assume that $E|\psi(X)|_2^\alpha < \infty$ for $\alpha \geq d + 1$. Assume $|C|_\infty < \infty$ and $\inf_\psi C(\psi) > 0$.

Pilot - Assume $\inf_\psi \bar{\sigma}^2(\psi) > 0$. Assume that $E[Y(d)^4] < \infty$. Assume that $\sigma_1(\psi)/\sigma_0(\psi) \in [c_l, c_u] \subseteq (0, \infty)$ a.s. Assume that $E|\psi(X)|_2^{\alpha_1} < \infty$ for $\alpha_1 \geq d+1$. Assume that $|\sigma_d^2 - \widehat{\sigma}_d^2|_2 = O_p(n^{-r})$. Assume that $E[|m_d(\psi)|^{\alpha_2}] < \infty$ for $a_2 \geq 1/r$. Assume that $\overline{k}_n > n^{1/\alpha_2}$. Assume $|C|_\infty < \infty$ and $\inf_\psi C(\psi) > 0$. Assume that $\overline{k}_n |L_n| = o(n^{1 - \frac{d+1}{\alpha_1}})$

Proof of Theorem 5.3. 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 3.10, 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

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

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 10.2.

(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 $q^*(\psi) = \bar{C}(\sigma_1(\psi) + \sigma_0(\psi))C(\psi)^{-1/2}/E[(\sigma_1(\psi) + \sigma_0(\psi))C(\psi)^{1/2}]$. 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 10.2 provides functions $h_n \in L_2(\psi)$ with $|h_n|_{lip} \vee |h_n|_\infty \leq c_n$ 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 10.17 $E[C_{n1}|\mathcal{F}_n] = 0$. 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, $|\widehat{q}_n - \widehat{q}|_\infty \leq 1/\overline{k}_n$. Then by Lemma 10.17 we have $E[B_{n1}|\mathcal{F}_n] = 0$ and

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

Again by Lemma 10.17 we have

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

The final expression is $o_p(n^{2/\alpha_2 - 2r}) + O(n^{2/\alpha_2}/\overline{k}_n^2)$, using Lemma 10.14 and the bound from Lemma 10.7. 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 $\text{Var}(\sqrt{n}T_{n4}|\mathcal{F}_n) \lesssim E_n[(m_1(\psi_i)/q^*(\psi_i) - h_n(\psi_i))^2] = o_p(1)$ by Markov inequality since $|h_n - (m_1/q^*)|_{2,\mathcal{X}} \to 0$. Finally, by the last part of Lemma 10.17 we have $E[T_{n5}|\mathcal{F}_n] = 0$ and

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

The first inequality by Lipschitz continuity, the second inequality by Theorem 10.1. 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 $V_1 = \text{Var}(c(\psi))$ by vanilla CLT. We want to apply Theorem 10.4 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 10.7. Then by Theorem 10.4 $\sqrt{n}R_n|\mathcal{F}_{3,n} \Rightarrow \mathcal{N}(0, V_2)$ 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 10.4. Then by Lemma 10.12, $\sqrt{n}(A_n + R_n) \Rightarrow \mathcal{N}(0, V_1 + V_2)$. Applying Lemma 10.17 with $h_n = 1$ shows that $\text{Var}(E_n[(T_i - \widehat{q}_{i,n})]|\mathcal{F}_{1,n}) \leq n^{-1}\overline{k}_n|L_n|$. 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 10.7. 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 $\sqrt{n_T}(\widehat{\theta} - \text{ATE}) = \sqrt{n}(E_n[T_i])^{1/2}(\widehat{\theta} - \text{ATE}) = E[q^*(\psi_i)]^{1/2}\sqrt{n}(\widehat{\theta} - \text{ATE}) + o_p(1) \Rightarrow N(0, E[q^*(\psi_i)] \cdot V)$ 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

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

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

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

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

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

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

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

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

Lemma 10.7 (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 $|\widehat{p}_{i,n} - p_i^*|_{2,n}^2 \leq 2|\widehat{p}_{i,n} - \widehat{p}_i|_{2,n}^2 + 2|\widehat{p}_i - p_i^*|_{2,n}^2 \leq O(1/\overline{k}_n^2) + |\widehat{p}_i - p_i^*|_{2,n}^2$ 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

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

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

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

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

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

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 $v_i = C_i^{-1/2}(\sigma_{1i}+\sigma_{0i})$.

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

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

10.4 Inference

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

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

Consider the first term

$$E_n[T_iD_iY_i(1)^2/(p_i^2q_i^2)] = E_n[T_i(D_i-p_i)Y_i(1)^2/(p_i^2q_i^2)] + E_n[(T_i-q_i)Y_i(1)^2/(p_iq_i^2)]$$
$$+ E_n[Y_i(1)^2/(p_iq_i)] = E_n[Y_i(1)^2/(p_iq_i)] + o_p(1) = E[Y_i(1)^2/(p_iq_i)] + o_p(1)$$

The first equality is by Lemma 10.17, and the second by WLLN. Then by symmetry $E_n[T_i(1-D_i)Y_i(0)^2/(1-p_i)^2q_i^2] = E[Y_i(0)^2/(1-p_i)q_i] + o_p(1)$. Then the sample variance in the theorem converges in probability to $v = E[Y_i(1)^2/(p_iq_i)] + E[Y_i(0)^2/(1-p_i)q_i] - \text{ATE}^2$. With V the limiting variance of Theorem 3.10, 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 10.8. □

Lemma 10.8 (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 $\widehat{v}_1 \xrightarrow{p} E[m_{1i}^2a_i]$, $\widehat{v}_0 \xrightarrow{p} E[q_i(1-p_i)m_{0i}^2a_i]$, 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

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

The second term is $o(1)$ by Theorem 10.1. By Young's inequality, the first term is

$$\begin{aligned}
&\lesssim n^{-1}\sum_{g\in\mathcal{G}_n}\sum_{i\in g,j\in\nu(g)}|\psi_i-\bar\psi_g|_2^2+|\bar\psi_g-\bar\psi_{\nu(g)}|_2^2+|\bar\psi_{\nu(g)}-\psi_j|_2^2\\
&\leq 2n^{-1}\sum_{g\in\mathcal{G}_n}k(g)\sum_{i\in g}|\psi_i-\bar\psi_g|_2^2+\bar k^2n^{-1}\sum_{g\in\mathcal{G}_n}|\bar\psi_g-\bar\psi_{\nu(g)}|_2^2
\end{aligned}$$

The second term is $o(1)$ again by Theorem 10.1. 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 10.1. This finishes our proof of the claim.

Denote $v_i=m_1(\psi_i)\sqrt{a(\psi_i)}$. By Lemma 10.11, 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

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

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

$$\begin{aligned}
A_n&=n^{-1}\sum_{g\in\mathcal{G}_n^\nu}w_g\sum_{i,j\in g}v_iv_jD_iD_j\mathbb{1}(i\neq j)=-(2n)^{-1}\sum_{g\in\mathcal{G}_n^\nu}w_g\sum_{i,j\in g}(v_i-v_j)^2D_iD_j\\
&\quad+(2n)^{-1}\sum_{g\in\mathcal{G}_n^\nu}w_g\sum_{i,j\in g}(v_i^2+v_j^2)D_iD_j\mathbb{1}(i\neq j)\equiv A_n^1+A_n^2
\end{aligned}$$

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

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

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,

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

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 10.17. Next we claim that $R_n = o_p(1)$. It suffices to verify the conditions of Lemma 10.13 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 10.15. 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 10.13 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 10.13. 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

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

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 10.17. Next consider

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

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 $g \neq g'$ by Lemma 10.15. Then by Lemma 10.13, 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 $w_g \leq 2$, so we have

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

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

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

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

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

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 10.13. 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. □

10.5 Regression Equivalence Propositions

Proposition 10.9 (Coarse Stratification)**.** *Assume $E[Y(d)^2] < \infty$ and $P(S = k) > 0$ for all $k \in [m]$. Suppose that $T_{1:n} \sim \mathrm{CR}(q)$ and $D_{1:n} \sim \mathrm{CR}(p)$. The estimator $\check{\theta} = \widehat{\tau} - \widehat{\beta}'(\bar{z}_{T=1} - \bar{z})$ defined in Example 3.5 has $\sqrt{n_T}(\check{\theta} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V)$ with the claimed variance.*

Proof. 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 $E_n[D_i T_i] = E_n[(D_i - p)T_i] + pE_n[T_i - q] + pq = pq + O(1/n)$. Similarly, we expand $E_n[\tilde{z}_i \tilde{z}_i' D_i T_i] = E_n[\tilde{z}_i \tilde{z}_i'(D_i - p)T_i] + pE_n[\tilde{z}_i \tilde{z}_i'(T_i - q)] + pqE_n[\tilde{z}_i \tilde{z}_i']$ and $E_n[\tilde{z}_i' D_i T_i] = E_n[\tilde{z}_i(D_i - p)T_i] + pE_n[\tilde{z}_i(T_i - q)] + pqE_n[\tilde{z}_i]$. It's easy to check that the conditions of Lemma 10.17 are satisfied, so that $E_n[\tilde{z}_i \tilde{z}_i' D_i T_i] = pqE_n[\tilde{z}_i \tilde{z}_i'] + O_p(n^{-1/2})$ and $E_n[\tilde{z}_i D_i T_i] = pqE_n[\tilde{z}_i] + O_p(n^{-1/2})$. 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

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

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

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

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 3.10. The conclusion then follows by the same argument. □

Proof of Proposition 3.11. 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

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

Continuing the calculation, this is

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

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 10.15 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 10.15

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

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 $n_T/n \xrightarrow{p} E[q_i]$. □

10.6 Lemmas

Lemma 10.10 (Conditional Convergence)**.** *Let $(\mathcal{G}_n)_{n\geq 1}$ and $(A_n)_{n\geq 1}$ 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 10.11 (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 10.12 (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 10.5. Then $(X_{n,1}, \ldots, X_{n,m})|\mathcal{F} \Rightarrow (L_1, \ldots, L_m)|\mathcal{F}$, with jointly independent limit.*

Proof. By Levy continuity, it suffices to show $E[e^{it'(X_{1,n},\ldots X_{m,n})}|\mathcal{F}] \to \prod_{k=1}^m E[e^{it_k L_k}|\mathcal{F}]$ 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 $1 \leq k \leq k' \leq m$. 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 $|e^{i\sum_{k=1}^{k'} t_k X_{n,k}} (E[e^{it_{k'+1}X_{n,k'+1}}|\mathcal{F}_{n,k'}] - E[e^{it_{k'+1}L_{k'+1}}|\mathcal{F}])| \leq |E[e^{it_{k'+1}X_{n,k'+1}}|\mathcal{F}_{n,k'}] - E[e^{it_{k'+1}L_{k'+1}}|\mathcal{F}]| \xrightarrow{p} 0$. 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 10.13 (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 $(\mathcal{F}_n)_{n\geq 1}$ be σ-algebras such that $\mathcal{G}_n$ is $\mathcal{F}_n$-measurable, $E[u_g|\mathcal{F}_n] = 0$, and for all $g \neq g' \in \mathcal{G}_n$ $u_g \perp\!\!\!\perp u_{g'}|\mathcal{F}_n$. 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 $|\mathcal{G}_n| \leq n$. Fix an indexing $\mathcal{G}_n = \{g_s : 1 \leq s \leq n\}$, possibly with $g_s = \emptyset$ for some s. Define $\bar{u}_{sn} = u_s\mathbb{1}(|u_s| \leq c_n)$ and $\bar{\mu}_{sn} = E[u_s\mathbb{1}(|u_s| \leq c_n)|\mathcal{F}_n]$. 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 $E[|T_{n1}||\mathcal{F}_n] \leq (1/n)\sum_{s=1}^n E[|u_s|\mathbb{1}(|u_s| > c_n)|\mathcal{F}_n] = o_p(1)$ by assumption. Then $T_{n1} = o_p(1)$ by conditional Markov (Lemma 10.10). Next consider T_{n2}. Note that by definition $E[\bar{u}_{sn} - \bar{\mu}_{sn}|\mathcal{F}_n] = 0$ for each $1 \leq s \leq n$. Note that for $s \neq s'$ $\mathrm{Cov}(\bar{u}_{sn}, \bar{u}_{s'n}|\mathcal{F}_n) = 0$ by the conditional independence assumption. Then $\mathrm{Var}(T_{n2}|\mathcal{F}_n) = n^{-2}\sum_{s=1}^n \mathrm{Var}(\bar{u}_{sn}|\mathcal{F}_n) \leq n^{-2}\sum_{s=1}^n E[\bar{u}_{sn}^2|\mathcal{F}_n] \leq n^{-1}c_n^2 = o(1)$, 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 $E[|T_{n3}||\mathcal{F}_n] \leq (1/n)\sum_{s=1}^n E[|u_s|\mathbb{1}(|u_s| > c_n)|\mathcal{F}_n] = o_p(1)$, so that $T_{n3} = o_p(1)$ by conditional Markov as before. This finishes the proof. □

Lemma 10.14. *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 $P(\max_{i=1}^n |X_i| > \epsilon n^{1/p}) \leq nP(|X_i| > \epsilon n^{1/p}) = nP(|X_i|^p > \epsilon^p n) \leq n(\epsilon^p n)^{-1}E[|X_i|^p\mathbb{1}(|X_i|^p > \epsilon^p n)] \lesssim E[|X_i|^p\mathbb{1}(|X_i|^p > \epsilon^p n)] \to 0$. 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 10.15 (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 $s \neq r$ 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 10.16 (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

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

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 $1 \le j \ne i \le n$

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

By σ-algebra independence and joint independence of groupwise randomizations

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

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

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

The bounds follow by inspection. □

Lemma 10.17 (Stochastic Balance)**.** *Let $(\mathcal{F}_n)_{n\geq 1}$ such that treatment groups $\mathcal{G}_n, (h_n(W_i))_{i=1}^n \in \mathcal{F}_n$ for a sequence of functions $(h_n)_{n\geq 1}$ 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) $\mathrm{Var}(E_n[T_i(D_i - p_n(X_i)))h_n(W_i)]|\mathcal{F}_n) \leq 2n^{-1}E_n[T_i h_n(W_i)^2] \leq 2n^{-1}E_n[h_n(W_i)^2]$

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

(4) The variance $\mathrm{Var}(\sqrt{n}E_n[T_i(D_i - p_n(X_i)))h_n(W_i)]|\mathcal{F}_n)$ 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 10.16 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 10.16, $\mathrm{Var}(E_n[(D_i - p_{i,n})h_n(W_i)]|\mathcal{F}_n)$ 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 10.16. By triangle inequality and the covariance bound in Lemma 10.16, this is bounded above by

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

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

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

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 10.10). 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 10.16 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

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

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

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

This finishes the proof. □