

Fine Stratification of Survey Experiments*

Max Cytrynbaum[†]

June 11, 2026

Abstract

This paper studies a two-stage model of experimentation, where the researcher first samples representative experimental participants from an eligible pool, then assigns each sampled unit to treatment or control, using matched k -tuples randomization at both stages. To implement such designs, we develop a fast new algorithm for matching units into k -tuples for any $k \geq 2$ and any dimension of covariates. By surveying 200 recent experimental working papers, we estimate that our algorithm newly enables multivariate fine stratification with provable match quality guarantees for about 44% of experiments in economics. We show that finely stratified sampling and assignment both nonparametrically reduce the variance of treatment effect estimation, with the gains from stratified sampling increasing in the size of the eligible pool and how well covariates predict treatment effect heterogeneity. We develop new inference methods that fully exploit the efficiency gains from both design stages, allowing researchers to report smaller standard errors if they designed a representative experiment. An application to nine published experiments quantifies the efficiency gains.

Keywords: Matched Pairs, Blocking, Survey Sampling, Robust Standard Error, Treatment Effects.

JEL Codes: C10, C14, C90

*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

A key objective in the design of experiments is to construct an experimental sample that is externally valid in the sense that it is representative of a broader population of interest. Given such a sample, experimenters also want to randomize so that the treatment and control groups are finely balanced on covariates, improving internal validity. One way to achieve both goals is through a survey experiment that uses stratified randomization for both the sampling of participants and assignment of treatments.

To illustrate, consider the OpenResearch Unconditional Income Study (ORUS), a large-scale experiment which distributed over \$40 million in unconditional cash transfers in the United States in an effort to study the effects of guaranteed income on a range of social and economic outcomes (Broockman et al., 2024). OpenResearch recruited 14,573 individuals who were eligible and willing to participate, but its budget only allowed for the final enrollment of 3,000 participants in the experiment. To ensure representativeness, the authors drew a stratified sample of size 3,000 from the eligible pool. Among the participants, they also assigned treatments by stratified randomization, matching units into triples on income, race, and state, along with several dozen other covariates ψ , then assigning one unit in each triple to the \$1,000 per month transfer.

Beyond the ORUS study, the value of stratification for both sampling participants and assigning treatments is recognized by practitioners. The J-PAL guide to randomized experiments recommends stratifying the sample “based on observed covariates that are expected to moderate the treatment effect,” and observes that “stratifying at the sampling stage ... reduces the variance of the sample relative to the underlying population,” just as “stratifying at the treatment assignment stage creates treatment and control groups that are more similar to each other” (Kopper and Riis-Vestergaard, 2023).

Similarly, Muralidharan and Niehaus (2017) note that “the lowest-hanging fruit may be to make samples more representative of the populations about which we wish to learn” and suggest “researchers have devoted more effort to persuading their institutional partners to randomize (for internal validity) than to be representative (for external validity).”

Motivated by such considerations, we study a general family of stratified designs for survey experiments, in which participants are representatively sampled from an eligible pool, then assigned to a binary treatment. We allow the sampling proportions $q(\psi)$ to vary with covariates ψ , as in the ORUS study above, implementing these proportions using fine stratification. Such covariate-dependent sampling rates are a classical feature of

stratified survey sampling (Cochran, 1977) and may arise for a variety of reasons, including a desire to oversample subgroups of particular interest for heterogeneous treatment effects (as discussed in the J-PAL guide), heterogeneous costs of experimentation in different regions, or administrative and logistical constraints in constructing the sample. Likewise, we allow treatment probabilities $p \neq 1/2$, e.g. $p = 1/3$ as in the ORUS study.

To implement such designs, we develop a new matching algorithm, which produces high quality matched k -tuples for any $k \geq 2$ by matching on covariates ψ in any dimension. Our algorithm provably satisfies an asymptotic tight matching condition, which is needed to guarantee that fine stratification delivers nonparametric efficiency improvements at both the sampling and assignment stages.

Prior work has provided algorithms with such guarantees only in the special cases of matched pairs $k = 2$ (Bai et al., 2022) and univariate matched k -tuples produced by sorting (Bai, 2022). Our algorithm is the first to possess such match quality guarantees for tuples of any size $k \geq 2$ and $\dim(\psi) \geq 1$.

Leaving theory aside, even as a purely practical matter, the literature has few recommendations on how to form matched k -tuples for general $k \geq 2$ and multivariate ψ beyond greedy heuristics (Moore, 2012) and mixed-integer schemes that do not scale above several hundred units (Brixius, 2025). This is perhaps not surprising, since solving for the globally optimal matched k -tuples in Euclidean distance is NP-hard if $k \geq 3$ and $\dim(\psi) > 1$ (Pyatkin et al., 2017). We navigate around this hardness result with an algorithm that scales to tens of thousands of units and also satisfies the appropriate tight matching condition. A more basic version of the procedure with lower finite sample match quality scales to millions of units. Thus, our procedure also closes a practical gap.

To assess the impact of such designs on experiments in economics, we surveyed 200 recent NBER working papers involving randomized experiments from 2024 to 2026. We estimate that multivariate fine stratification with match quality guarantees is not feasible using existing methods for 44% of these experiments, either because they involve multiple treatment arms or treatment proportions $p \neq 1/2$, both of which require matched k -tuples with $k > 2$.¹ Thus, for 44% of recent working papers, provably high quality stratified treatment assignment with $\dim(\psi) > 1$ is newly enabled by our procedure.

This is before considering finely stratified sampling, which has propensity $q \neq 1/2$ unless the eligible units are exactly twice the size of the experiment, thus generically

¹The 44% additionally includes factorial designs with $k = 2^m$ ($m \geq 2$) and sample size $n \geq 10,000$, for which the iterated optimal pairing scheme advocated by Bai et al. (2024) is computationally intractable.

requiring matched k -tuples designs with $k > 2$. In our survey, we find that about 23% of these working papers already perform such a sampling step, either by randomly sampling participants from a larger eligible pool (15%) or using a third-party service such as Prolific or YouGov to recruit a representative sample (8%). Our results show that experimenters who construct a representative sample in this way can report smaller standard errors, providing an incentive for researchers to make their experiments representative.

In Section 2, we formally introduce a family of finely stratified designs that randomizes sampling and assignment variables within a matched partition \mathcal{G}_n . For matched groups $g \in \mathcal{G}_n$ and centroids $\bar{\psi}_g = |g|^{-1} \sum_{i \in g} \psi_i$, we require the tight matching condition:

$$\frac{1}{n} \sum_{g \in \mathcal{G}_n} \sum_{i \in g} |\psi_i - \bar{\psi}_g|_2^2 = o_p(1). \quad (1.1)$$

This generalizes similar match quality conditions that have appeared in [Bai et al. \(2022\)](#) and [Bai \(2022\)](#). In the remaining sections, we make the following contributions:

1. In Section 3, we develop a new matching algorithm, which uses a fast spatial sorting procedure as a warm start to initialize the balanced k -means procedure of [Malinen and Fränti \(2014\)](#). We show that the resulting partition \mathcal{G}_n satisfies the tight matching condition (1.1) under weak assumptions.
2. In Section 4, we provide asymptotic theory for finely stratified survey experiments with general sampling proportions $q(\psi)$ and assignment propensity $p \neq 1/2$. We show finely stratified sampling and assignment both provide nonparametric variance reductions, attaining the relevant semiparametric variance bound. Stratified sampling attenuates the variance due to treatment effect heterogeneity, with larger reductions as the size of the eligible pool grows. We characterize the optimal stratification variables, showing that for sampling one should stratify on a small set of covariates ψ most predictive of treatment effect heterogeneity, formally justifying the J-PAL recommendation above.
3. In Section 5, we apply these designs to formulate and solve a budget-constrained optimal sampling problem with heterogeneous costs, deriving the optimal sampling proportions $q^*(\psi) \propto C(\psi)^{-1/2} \sigma(\psi)$ for costs $C(\psi)$ and an appropriate residual variance function $\sigma^2(\psi)$. Under a homoskedasticity assumption, this design can be implemented using the known costs. We also propose versions that use available pilot or observational data to estimate the unknown $\sigma^2(\psi)$ term, showing these

asymptotically minimize variance over all budget-feasible proportions $q(\psi)$.

4. In Section 6, we develop new inference methods for experiments with joint finely stratified sampling and assignment. We enable asymptotically non-conservative inference on the population average treatment effect, as well as asymptotically valid inference on the average treatment effect among the eligible units.

In Section 7, we present an empirical application to nine experiments recently published in top economics journals, which demonstrates the value of our methods.

1.1 Related Literature

For overviews of experimental design theory, see [Rosenberger and Lachin \(2016\)](#), [Athey and Imbens \(2017\)](#), or [Bai et al. \(2025\)](#). For reviews of the classical literature on survey sampling, see [Cochran \(1977\)](#) and [Lohr \(2021\)](#).

The idea that trial findings should be extended from enrolled participants to the broader population of trial-eligible individuals originates in the clinical literature ([Rothwell, 2005](#)). A subsequent literature in statistics develops estimators that correct for non-representative experimental participation and imbalanced assignment by ex-post adjustment, as in [Dahabreh et al. \(2019\)](#) and [Li et al. \(2023\)](#). See [Degtiar and Rose \(2023\)](#) for a review of this area. By contrast, our finely stratified sampling and assignment designs prevent such imbalances from arising at the randomization stage.

Contemporaneous with the first version of our paper, [Yang et al. \(2023\)](#) propose a two-stage design using rerandomization for both sampling and assignment. Under rerandomization, difference-of-means estimation is asymptotically slightly less efficient than ex-post linear covariate adjustment. By contrast, we show that two-stage fine stratification is asymptotically equivalent to nonparametric covariate adjustment for the imbalances in both the sampling and assignment variables.

Recent work on stratified treatment assignment includes [Imai et al. \(2009\)](#), [Bugni et al. \(2018\)](#), [Fogarty \(2018\)](#), [Wang et al. \(2023\)](#), [Bai et al. \(2022\)](#), [de Chaisemartin and Ramirez-Cuellar \(2024\)](#), [Bai \(2022\)](#), [Tabord-Meehan \(2023\)](#), [Bai et al. \(2024\)](#), and [Bai et al. \(2026\)](#). For treatment assignment, our work is most closely related to [Bai et al. \(2022\)](#), who study matched pairs designs, and [Bai \(2022\)](#), who studies finely stratified designs with constant propensity $p = a/k$ and univariate stratification variables, $\dim(\psi) = 1$. Aside from stratification, other recent proposals for balanced treatment

assignment include [Kasy \(2016\)](#), [Kallus \(2018\)](#), [Li et al. \(2018\)](#), [Krieger et al. \(2019\)](#), and [Harshaw et al. \(2024\)](#).

From a computational perspective, finely stratified designs require the construction of high-quality matched groups. Multivariate fine stratification has previously relied on greedy algorithms such as that of [Moore \(2012\)](#), implemented in the `blockTools` package. For large experiments, the threshold blocking method of [Higgins et al. \(2016\)](#) forms blocks of uneven sizes bounded below by some chosen k . These methods do not deliver the tight matching guarantees that drive our efficiency results.

The most closely related previous work providing such guarantees is [Bai et al. \(2022\)](#), who bound the average squared Euclidean distance of the optimal non-bipartite matching ([Derigs, 1988](#)) using a space-filling path argument. [Bai \(2022\)](#) forms univariate k -tuples with a match quality guarantee by sorting units along their ψ_i values for $\dim(\psi) = 1$. We generalize the space-filling construction and adapt it into a practical algorithm, deploying this as the warm start for the balanced k -means procedure of [Malinen and Fränti \(2014\)](#). This results in a fast, practical algorithm that provably satisfies the appropriate tight matching condition for any $k \geq 2$ and $\dim(\psi) \geq 1$.

Our budget-constrained optimal sampling design extends the classical theory of optimal allocation in survey sampling ([Cochran, 1977](#)) to finely stratified sampling into an experiment. Our results on design using a pilot study are related to previous work in [Hahn et al. \(2011\)](#), [Bai \(2022\)](#), [Tabord-Meehan \(2023\)](#), and [Kasy and Sautmann \(2021\)](#). Our inference results build on the method of collapsed-strata in [Hansen et al. \(1953\)](#) and its modern “pairs of pairs” variants studied in [Abadie and Imbens \(2008\)](#), [Bai et al. \(2022\)](#), and [Bai et al. \(2026\)](#).

2 Setting and Designs

Consider running an experiment to estimate the effect of a binary treatment. There are n eligible units with potential outcomes $Y_i(d)$ for $d \in \{0, 1\}$ and observed baseline covariates ψ_i for $i = 1, \dots, n$. Let $T_i = 1$ if an eligible unit is sampled to participate in the experiment and $T_i = 0$ otherwise, so experiment size $n_T = \sum_i T_i$. In general, we may want to implement sampling proportions $q(\psi) = P(T = 1|\psi) \in (0, 1]$. Sampled units are assigned to treatment or control $D_i \in \{0, 1\}$ with $p(\psi) = P(D = 1|\psi)$. In practice, p and q will often be constant. Then outcomes are $Y_i = T_i[D_i Y_i(1) + (1 - D_i)Y_i(0)]$.

For example, [Shim et al. \(2026\)](#) study the effect of fiscal news on household spending,

sampling $n_T = 11,262$ respondents into their experiment from an eligible population of $n \approx 200,000$ Koreans in a survey panel, so that $q = n_T/n \approx 1/18$. Similarly, [Chioda et al. \(2026\)](#) evaluate an entrepreneurship training program, sampling $n_T = 4,402$ participants into the experiment from a pool of $n = 7,431$ Ugandan applicants so that $q \approx 3/5$.

Denote $W_i = (\psi_i, Y_i(0), Y_i(1))$ and suppose $(W_i)_{i=1}^n \stackrel{\text{iid}}{\sim} P$, modeling the eligible units as an iid sample from a broader superpopulation. One natural estimand in this case is the average treatment effect $\text{ATE} = E[Y(1) - Y(0)]$. For example, in the ORUS basic income study in the introduction, the ATE is the treatment effect among the broader population of low-income Americans from which the $n = 14,573$ eligible participants were recruited, using a combination of direct mailers and online advertisements.

Another possible estimand is the $\text{SATE} = n^{-1} \sum_{i=1}^n (Y_i(1) - Y_i(0))$ among the n eligible units that the experimenter physically samples from. In the ORUS study, the SATE is the treatment effect among these $n = 14,573$ eligible units, from which they sampled the ultimate experimental population of size $n_T = 3,000$. Note the SATE is sometimes alternatively defined more narrowly as the treatment effect among the n_T experimental participants, $n_T^{-1} \sum_{i=1}^n T_i (Y_i(1) - Y_i(0))$. This estimand has less policy relevance in general, though it is equivalent to our definition of the SATE in the special case $q = 1$. We mostly focus on the ATE in what follows, though for completeness we extend our main asymptotic and inference results to the SATE as well.

In either case, our goal is to sample a representative subset of the eligible units into the experiment, then assign them to treatment and control in a way that finely balances the baseline covariates $\psi_i \in \mathbb{R}^d$. We implement both the sampling and treatment assignment steps using a form of matched k -tuples randomization.

Matched k -Tuples. We first describe matched k -tuples in the context of random sampling. Suppose we want to sample $q = 1/5$ of the n eligible units into the experiment. To do so, we first match them into homogeneous groups of size $k = 5$ using the covariates $(\psi_i)_{i=1}^n$, then sample one out of every five units in each matched group into the experiment, uniformly at random. Doing so balances covariates ψ_i between the sampled and non-sampled units, $T_i = 1$ and $T_i = 0$. More generally, we introduce a family of matched k -tuples designs that allows for potentially varying sampling and assignment propensities $q(\psi) = P(T = 1|\psi)$ and $p(\psi) = P(D = 1|\psi)$:

Definition 2.1 (Local Randomization). Let $q : \mathbb{R}^d \rightarrow (0, 1]$ be a propensity with rational levels $q(\psi) \in \{a_l/k_l : l = 1, \dots, L\}$. We write $T_{1:n} \sim \text{Loc}(\psi, q(\psi))$ and say that the design $T_{1:n} = (T_i)_{i=1}^n$ is a *locally randomized* implementation of $q(\cdot)$ with respect to ψ if there

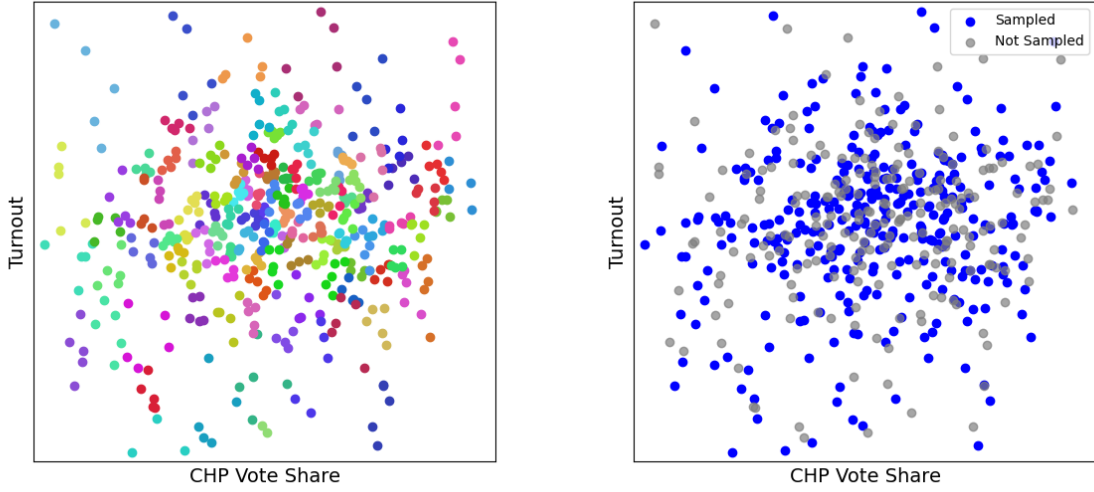


Figure 1: Sampling groups and variables for $T_{1:n} \sim \text{Loc}(\psi, q)$ with $q = 3/5$.

exists a partition of $\{1, \dots, n\}$ into matched groups $g \in \mathcal{G}_n$ such that:

- (1) *Tight matching.* The partition \mathcal{G}_n is determined only by $\psi_{1:n}$ and independent randomness π_n used to break ties in group formation. For centroids $\bar{\psi}_g \equiv |g|^{-1} \sum_{i \in g} \psi_i$, the partition has matching objective F with

$$F(\mathcal{G}_n) \equiv \frac{1}{n} \sum_{g \in \mathcal{G}_n} \sum_{i \in g} |\psi_i - \bar{\psi}_g|_2^2 = o_p(1). \quad (2.1)$$

- (2) *Complete randomization.* For each $g \in \mathcal{G}_n$, units $i, j \in g$ have equal sampling propensities $q(\psi_i) = q(\psi_j) = a_l/k_l$ for some $l = 1, \dots, L$. Conditional on $(W_{1:n}, \pi_n)$, each $(T_i)_{i \in g}$ is drawn uniformly over all binary vectors in $\{0, 1\}^{k_l}$ with exactly a_l out of k_l units having $T_i = 1$, independently between groups.

Condition (2) is the usual definition of stratified randomization. Note we may require one remainder group with $1 \leq |g| < k_l$ per propensity level due to integer effects. For example, if $n = 101$ and $q = 1/5$, the remainder group would have size $|g| = 1$, and we can draw $T_i \sim \text{Bernoulli}(q)$ for this unit.

Tight Matching Condition. The key requirement on the partition \mathcal{G}_n is the tight matching condition $F(\mathcal{G}_n) = o_p(1)$ in Equation 2.1, which asks that the matched groups be tightly clustered in ψ -space. We use it to guarantee strong control over

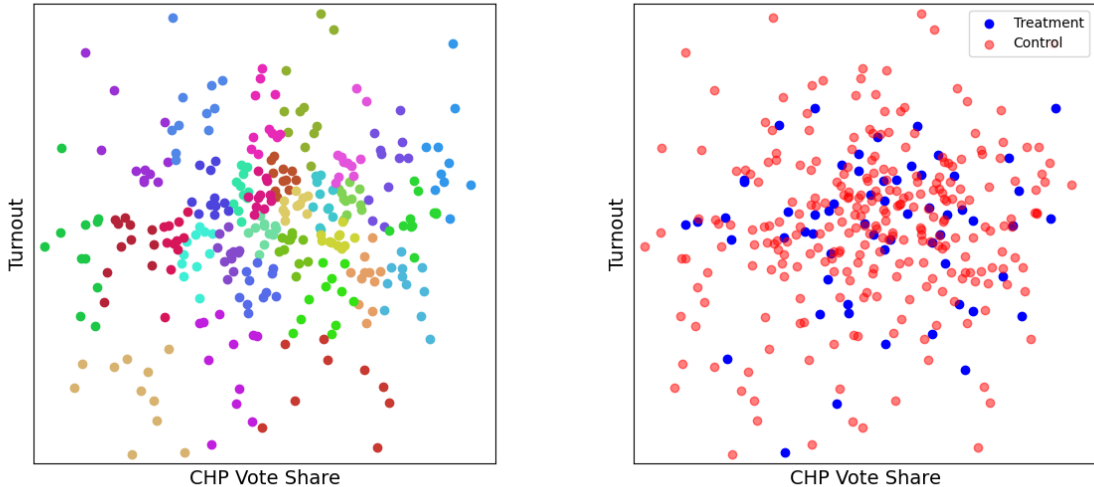


Figure 2: Assignment groups and variables for $D_{1:n} \sim \text{Loc}(\psi, p)$ with $p = 2/11$.

the in-sample covariate imbalances that may arise during randomization, allowing us to show that finely stratified sampling and assignment both provide nonparametric variance reductions for treatment effect estimation. This condition generalizes related match quality guarantees previously shown for optimal matched pairs in Bai et al. (2022) and univariate matched k -tuples in Bai (2022).

Constructing fast algorithms that provably satisfy this condition for any $k \geq 2$ and $\dim(\psi) \geq 1$ is a core contribution, enabling the finely stratified sampling and assignment designs we study in this paper. However, the value of this algorithm extends beyond the binary setting. For example, Bai et al. (2024) study finely stratified assignment for experiments with multiple treatment arms, which requires an asymptotically equivalent matching condition, but provide an algorithm only in the case $k = 2^m$. Our matching algorithm thus makes their asymptotic results applicable to designs outside this special case, for example experiments with three treatment arms, requiring matched triples with $k = 3$. We present the algorithm and its formal matching guarantee in Section 3 below.

Experiment Design. We consider a two-stage finely stratified design:

- (1) Sample eligible units $T_{1:n} \sim \text{Loc}(\psi, q(\psi))$.
- (2) Assign treatments $D_{1:n} \sim \text{Loc}(\psi, p(\psi))$ to the sampled units $\{i : T_i = 1\}$.

The most common use case of this design is to let the propensities be constant,

sampling $T_{1:n} \sim \text{Loc}(\psi, q)$ for $q = a/k$ of the eligible units, then assigning treatments $D_{1:n} \sim \text{Loc}(\psi, p)$ to the sampled units. In our survey of experiments in recent NBER working papers, 73 are binary-treatment RCTs, of which about 27% deliberately use a treatment proportion $p \neq 1/2$, requiring our matched k -tuples algorithm for multivariate fine stratification not only at the sampling stage, but also for treatment assignment.

More generally, the experimenter may desire non-constant propensities $q(\psi)$ or $p(\psi)$ for various budgetary or logistical reasons, or purely to improve statistical power. For example, in Section 5, we study stratified implementation of the variance minimizing sampling proportions $q^*(\psi)$ under a fixed budget constraint.

Example 2.2 (Matched Tuples). We illustrate the basic sampling and assignment procedure in Figures 1 and 2, using data from the electoral information campaign in [Baysan \(2022\)](#). Each color represents a different group of units in the partition \mathcal{G}_n . In Figure 1 we sample $q = 3/5$ of the units by forming groups of size $|g| = 5$, then randomly sampling $T_i = 1$ for exactly 3 out of 5 units in each group. By sampling in this way we ensure that units of each “type” in the space of covariates $\psi = (\text{vote share, turnout})$ are represented within the smaller experiment. [Baysan \(2022\)](#) assigned only $p = 2/11$ of the units to $D = 1$ due to the high cost of the treatment. In Figure 2, we implement this by matching sampled units $\{i : T_i = 1\}$ into groups of size $|g| = 11$, randomly assigning $D_i = 1$ to exactly 2 out of 11 units from each group.

Remark 2.3 (Coarse Stratification). The literature has also considered coarsely stratified designs with fixed strata $X_i \in \{1, \dots, m\}$ and stratum sizes $n_x = |\{i : X_i = x\}| \rightarrow \infty$ (e.g. [Bugni et al., 2018](#)). Such designs are nested in our framework by letting $\psi = X$ and matching units sharing the same discrete covariate value $X = x$ into groups at random. In this case, \mathcal{G}_n depends on both $\psi_{1:n}$ and an extra randomization device π_n , which we accommodate in our analysis. Such groups trivially satisfy the tight matching condition in Equation 2.1. Complete randomization is obtained by setting $\psi = 1$ and forming groups $|g| = k$ at random. Thus, we provide a unified asymptotic theory and inference methods for finely stratified, coarsely stratified, and completely randomized designs at both the sampling and assignment stages.

3 Algorithm and Match Quality Guarantee

In what follows, denote $[z] = \{1, \dots, z\}$ for any positive integer z . Let $a \wedge b = \min(a, b)$ and $|S|$ denote the cardinality of S . We begin by describing the algorithm for a single

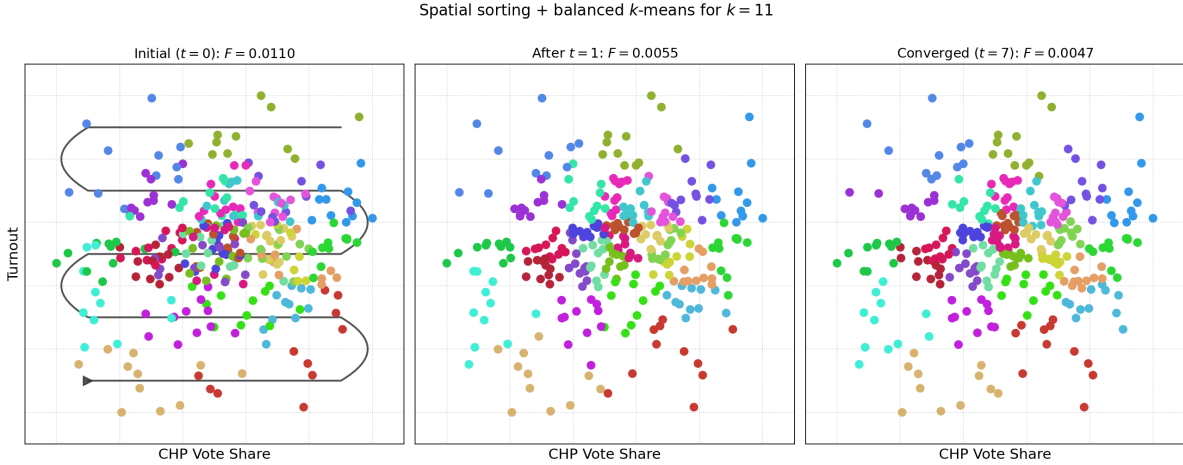


Figure 3: Matching algorithm of Section 3 applied to the [Baysan \(2022\)](#) data with tuple size $k = 11$. Preprocessed data and spatial sorting algorithm (left), one iteration of balanced k -means (middle), convergence after $t = 7$ iterations (right).

propensity level $q = a/k$, where a set of units $S \subseteq [n]$ is to be matched into groups of a common size k . For sampling, we apply the algorithm to the eligible units $S = [n]$, while we apply it to the participants $S = \{i : T_i = 1\}$ for assignment.

Our algorithm constructs a matched partition \mathcal{G}_n of the units $S \subseteq [n]$ as follows:

- (1) *Preprocessing.* Form a remainder group r of size $|r| < k$ units so that $S' = S \setminus r$ has cardinality $|S'|$ exactly divisible by the tuple size k . For the remaining units $i \in S'$, standardize covariates into the unit cube $\psi_i \in [0, 1]^{\dim(\psi)}$ by transforming $\psi_i^{(j)} \rightarrow (\psi_i^{(j)} - \min_{i' \in S'} \psi_{i'}^{(j)}) / (\max_{i' \in S'} \psi_{i'}^{(j)} - \min_{i' \in S'} \psi_{i'}^{(j)})$ for $j \in [\dim(\psi)]$.
- (2) *Spatial sorting.* Compute an index function $\rho : [0, 1]^{\dim(\psi)} \rightarrow \mathbb{N}$ returning the position of $\psi \in [0, 1]^{\dim(\psi)}$ along a coarsened space-filling curve. Sort units by their index values $(\rho(\psi_i))_{i \in S'}$, then match consecutive units into groups $g \in \mathcal{G}_n$ of size k , breaking ties at random.
- (3) *Balanced k -means polishing.* Refine the resulting partition \mathcal{G}_n by running the balanced k -means algorithm of [Malinen and Fränti \(2014\)](#) for $T \geq 0$ iterations. In practice, we run until convergence, which happens in finitely many iterations.

Readers primarily interested in treatment effect estimation can see the match quality guarantee in Theorem 3.1 and skip the technical construction below on a first reading.

For the preprocessing in step (1), note that if $\max_{i' \in S'} \psi_{i'}^{(j)} = \min_{i' \in S'} \psi_{i'}^{(j)}$ then covariate j has no variation and can be excluded from the procedure without changing

match quality. Then without loss, $\max_{i' \in S'} \psi_{i'}^{(j)} - \min_{i' \in S'} \psi_{i'}^{(j)} > 0$ for each dimension $j \in [\dim(\psi)]$. Also, our statistical guarantees are invariant to the choice of remainder group. In practice, we use the remainder to trim outliers, which improves the performance of the spatial sorting algorithm.

Spatial Sorting. We partition the unit cube $[0, 1]^{\dim(\psi)}$ into axis-aligned blocks of side length $1/m$ for grid size parameter m . This creates a grid with cells of the form $C(z) = \prod_{j=1}^{\dim(\psi)} [z_j/m, (z_j + 1)/m]$ for integers $z_j \in \{0, \dots, m - 1\}$, see the gray lines in Figure 3 for an illustration. For $S \subseteq [n]$, we require grid size $m \asymp (|S|/k)^{1/(\dim(\psi)+1)}$. In practice, we set $m = \lceil (|S|/(k \cdot \dim(\psi)))^{1/(\dim(\psi)+1)} \rceil$, which approximately optimizes an upper bound on the matching objective in our theoretical analysis.

The index function $\rho(\cdot)$ in step (2) enumerates these blocks along a space-filling path through $[0, 1]^{\dim(\psi)}$, such that any two consecutive cells share a $\dim(\psi) - 1$ dimensional face. The existence of such a path has also been used to study optimal matched pairs in Bai et al. (2022), see Remark 3.2 below for a detailed comparison. For data in dimension $v \geq 1$, denote the position of cell $C(z)$ along this path by $\rho_v(z) \in \mathbb{N}$. In the univariate case, we can set $\rho_1(z_1) = z_1$. In dimension $v \geq 2$, the position of cell $C(z)$ can be computed recursively as in Er (1984):

$$\rho_v(z_1, \dots, z_v) \equiv z_v \cdot m^{v-1} + \begin{cases} \rho_{v-1}(z_1, \dots, z_{v-1}) & z_v \text{ even,} \\ m^{v-1} - 1 - \rho_{v-1}(z_1, \dots, z_{v-1}) & z_v \text{ odd.} \end{cases} \quad (3.1)$$

To understand the recursion, note $\rho_{v-1}(\cdot)$ defines an order for traversing the $(v - 1)$ -dimensional subspace. If the coordinate z_v in the new dimension is even, we traverse the $v - 1$ dimensional layer in forward order, while if z_v is odd we go in reverse order, so that consecutive layers are traversed without jumps. See Figure 3 for an illustration.

We extend ρ_v as a map defined on vectors of integers to a map on the unit cube $\rho : [0, 1]^{\dim(\psi)} \rightarrow \mathbb{N}$ by setting $\rho(\psi) \equiv \rho_v(\lfloor m\psi \rfloor \wedge (m - 1))$ for $v = \dim(\psi)$, where the min and floor are applied componentwise. We sort the units in S' by their index values $\rho(\psi_i)$, matching consecutive units into groups of size k and breaking ties at random. This procedure yields an initial partition $\mathcal{G}_n^{(0)}$ of S' into groups with $|g| = k$.

Time Complexity. Overall, spatial sorting in step (2) requires $O(n \dim(\psi) + n \log n)$ operations, which is very tractable even for massive experiments. For $\dim(\psi) \leq 10$ this algorithm can match $n = 10$ million units into $k = 4$ tuples in under 20 seconds.

Balanced k -means polishing. Given the initial partition of S' into groups $g \in$

$\mathcal{G}_n^{(0)}$ with $|g| = k$, we use balanced k -means (Malinen and Fränti, 2014) to further decrease the matching objective defined by $F(\mathcal{G}) \equiv n^{-1} \sum_{g \in \mathcal{G}} \sum_{i \in g} |\psi_i - \bar{\psi}_g|_2^2$, subject to the equal-size constraint. At iteration t , let $\bar{\psi}_g^{(t)} = k^{-1} \sum_{i \in g} \psi_i$ denote the centroid of $g \in \mathcal{G}_n^{(t)}$, and update the partition of the units in S' into groups $|g| = k$ by

$$\mathcal{G}_n^{(t+1)} = \operatorname{argmin}_{\mathcal{G}: |g|=k} \sum_{g \in \mathcal{G}} \sum_{i \in g} |\psi_i - \bar{\psi}_g^{(t)}|_2^2. \quad (3.2)$$

The right-hand side of (3.2) is a balanced linear assignment problem that can be solved in polynomial time using the Hungarian algorithm (Malinen and Fränti, 2014). This objective is non-increasing, with $0 \leq F(\mathcal{G}_n^{(t+1)}) \leq F(\mathcal{G}_n^{(t)})$ for all $t \geq 0$ and strict decrease whenever the partition changes, until the algorithm converges to a fixed point $\mathcal{G}_n^{(t+1)} = \mathcal{G}_n^{(t)}$. Since there are finitely many such partitions, this convergence happens in finitely many iterations. The output of the single-level algorithm is the partition $\mathcal{G}_n = \mathcal{G}_n^{(T)} \cup \{r\}$, where $T \geq 0$ is the number of balanced k -means iterations.

Multiple Propensity Strata. In our general definition of local randomization, the propensity $q(\psi)$ may take several rational values $q(\psi) \in \{a_l/k_l : l \in [L]\}$. In this case, we partition the units into propensity strata $S_l = \{i : q(\psi_i) = a_l/k_l\}$ with size $n_l = |S'_l|$ for $S'_l = S_l \setminus r_l$ after removing the remainder. We run the algorithm above separately within each S'_l , matching the units into groups of size k_l . Each propensity stratum is handled exactly as above, with grid size $m_l = \lceil (n_l/(k_l \cdot \dim(\psi)))^{1/(\dim(\psi)+1)} \rceil$ and balanced k -means iteration counts $T_l \geq 0$. Writing \mathcal{G}_{nl} for the partition of S_l produced by the algorithm, the final output is the union $\mathcal{G}_n = \cup_{l=1}^L \mathcal{G}_{nl}$. Our main result in this section shows \mathcal{G}_n satisfies the tight matching condition in Equation 2.1.

Theorem 3.1 (Algorithm Guarantee). *Suppose $E[|\psi|_2^\alpha] < \infty$ for some $\alpha > \dim(\psi) + 1$. For each $l \in [L]$, let grid sizes $m_l \asymp (n_l/k_l)^{1/(\dim(\psi)+1)}$ and any number of balanced k -means iterations $T_l \geq 0$. Then the partition \mathcal{G}_n produced by the algorithm above satisfies*

$$n^{-1} \sum_{g \in \mathcal{G}_n} \sum_{i \in g} |\psi_i - \bar{\psi}_g|_2^2 = o_p(n^{2/\alpha - 2/(\dim(\psi)+1)}) = o_p(1).$$

See Theorem A.3 for our most general result, in asymptotics allowing the maximum group size $\bar{k}_n = \max_{l \in [L_n]} k_l$ and number of propensity levels L_n to grow with n . This generalization supports our results in Section 5, which consider a sampling design $T_{1:n} \sim \operatorname{Loc}(\psi, \hat{q}_n(\psi))$, where $\hat{q}_n(\psi)$ is a rational-valued estimate of the optimal sampling propensity $q^*(\psi)$, discretized with increasing fineness, so that $\bar{k}_n, L_n \rightarrow \infty$ as $n \rightarrow \infty$.

Remark 3.2 (Sorting Construction). [Bai et al. \(2022\)](#) were the first to use the existence of such space filling paths as an analytical device to bound the objective function value achieved by optimal matched pairs. In our notation, they proved $F(\mathcal{G}_n) = O(n^{-1/\dim(\psi)})$ for bounded ψ and $k = 2$. We show a slightly sharper result, which applies for general $k \geq 2$. In particular, [Theorem A.3](#) implies the improved rate $F(\mathcal{G}_n^*) \leq F(\mathcal{G}_n) = O(n^{-2/(\dim(\psi)+1)})$ for the optimal k -tuple matching \mathcal{G}_n^* , by a modified analysis.

Application to Survey Experiments. Recall in our two-stage design above we sample $T_{1:n} \sim \text{Loc}(\psi, q(\psi))$ and assign $D_{1:n} \sim \text{Loc}(\psi, p(\psi))$. By definition, such designs require partitions \mathcal{G}_n^T of the eligible units $[n]$ and \mathcal{G}_n^D of the sampled units $\{i : T_i = 1\}$ that both satisfy the tight matching condition. The following corollary shows that our algorithm provides such partitions.

Corollary 3.3 (Survey Experiments). *Suppose $E[|\psi|_2^\alpha] < \infty$ for some $\alpha > \dim(\psi) + 1$. Then the sampling and assignment partitions \mathcal{G}_n^T and \mathcal{G}_n^D produced by the algorithm have*

$$F(\mathcal{G}_n^T) = o_p(1), \quad F(\mathcal{G}_n^D) = o_p(1).$$

4 Asymptotic Theory

This section contains our main asymptotic results, showing nonparametric efficiency gains from both finely stratified sampling and assignment. Below, we denote $E_n[a_i] = n^{-1} \sum_{i=1}^n a_i$ for any array $(a_i)_{i=1}^n$. We assume the following throughout:

Assumption 4.1. *Moments $E[Y(d)^2] < \infty$ for $d = 0, 1$.*

Estimation. Recall in our definition of local randomization ([Definition 2.1](#)), we allowed sampling and assignment propensities $q(\psi)$ and $p(\psi)$ to have finitely many rational levels, such as $q(\psi) = a_l/k_l$ for $l \in [L]$. This may be required for logistical reasons, or to maximize statistical power subject to a budget constraint, as we study in [Section 5](#) below. To accommodate the most general designs, we state our main asymptotic results using the inverse propensity weighting (IPW) estimator

$$\hat{\theta} = E_n \left[\frac{T_i}{q(\psi_i)} \left(\frac{D_i Y_i}{p(\psi_i)} - \frac{(1 - D_i) Y_i}{1 - p(\psi_i)} \right) \right]. \quad (4.1)$$

In the special case with constant $p(\psi) = p$ and $q(\psi) = q$, we could use a standard regression estimator, like the coefficient on D_i in the regression $Y_i \sim 1 + D_i$, estimated

among the sampled units. Denoting this estimator by $\widehat{\theta}_{\text{ols}}$, we have $\widehat{\theta}_{\text{ols}} = \widehat{\theta} + O_p(n^{-1})$ under stratified sampling and assignment, so using OLS instead of IPW does not change our efficiency or inference results. Motivated by this, we restrict our attention to IPW in what follows. Note that this equivalence between OLS and IPW is a consequence of stratification and is not true for iid designs.

In what follows, let $c(\psi) = E[Y(1) - Y(0)|\psi]$ denote the conditional average treatment effect (CATE) and $\sigma_d^2(\psi) = \text{Var}(Y(d)|\psi)$ the heteroskedasticity function. Recall that $n_T = \sum_i T_i$ is the experiment size. Next we state our main asymptotic result.

Theorem 4.2 (CLT). *Assume sampling and assignment are locally randomized with $T_{1:n} \sim \text{Loc}(\psi, q(\psi))$ and $D_{1:n} \sim \text{Loc}(\psi, p(\psi))$. Then $\sqrt{n_T}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$*

$$V = E[q(\psi)] \left(\text{Var}(c(\psi)) + E \left[\frac{1}{q(\psi)} \left(\frac{\sigma_1^2(\psi)}{p(\psi)} + \frac{\sigma_0^2(\psi)}{1-p(\psi)} \right) \right] \right). \quad (4.2)$$

We begin with the most common case of constant sampling and assignment propensities $q(\psi) = q$ and $p(\psi) = p$. Then the variance in Theorem 4.2 simplifies to

$$V = q \text{Var}(c(\psi)) + E \left[\frac{\sigma_1^2(\psi)}{p} + \frac{\sigma_0^2(\psi)}{1-p} \right]. \quad (4.3)$$

Stratified Assignment. If $q = 1$ then $n_T = n$ and our framework recovers the conventional superpopulation sampling model. Setting $q = 1$, the variance V is precisely the [Hahn \(1998\)](#) semiparametric variance bound for estimating the ATE in this model. For iid treatments $D_i \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p)$, this variance can typically only be attained by nonparametrically estimating and weighting by the realized propensity $\widehat{p}(\psi)$ ([Hirano et al., 2003](#)), nonparametric regression adjustment ([Robins et al., 1994](#)), or a combination of the two ([Chernozhukov et al., 2017](#)).

By contrast, here we achieve this variance bound using a simple IPW estimator plugging in the known propensity or OLS $Y_i \sim 1 + D_i$ if $p(\psi) = p$. This shows how fine stratification nonparametrically controls the covariate imbalances by design, without the need for ex-post adjustment. The original [Hahn \(1998\)](#) ATE variance bound was shown for iid data, but recent work in [Armstrong \(2025\)](#) has extended this bound to general forms of covariate-adaptive randomization, including the finely stratified designs here.

The original work of [Bai et al. \(2022\)](#) showed the variance V is attained by $\widehat{\theta}_{\text{ols}}$ under matched pairs randomization if the pairs satisfy a tight matching condition, corresponding to the case $p = 1/2$ in our framework. [Bai \(2022\)](#) extended this to the case

$p = a/k$ with $\dim(\psi) = 1$.

Stratified Sampling + Assignment. Next, consider finely stratified sampling into the experiment with $q < 1$. The first term in Equation 4.3 is then strictly smaller than the Hahn (1998) bound, with decreasing variance as $q = n_T/n \rightarrow 0$. This formalizes an observation of Muralidharan and Niehaus (2017) that “external validity is after all a continuous and not a binary concept, and all else equal, a sample representative of 10 million people does more for external validity than one that is representative of 10,000.”

This continuous variance reduction is enabled by stratified sampling, which allows us to tightly couple the distribution of covariates $(\psi_i)_{i=1}^n$ among the n eligible units to the covariate distribution in our smaller experiment of size n_T , reducing the variance due to treatment effect heterogeneity predictable by ψ by a factor $q = n_T/n$.

To see why, imagine an oracle setting where we exactly observe the treatment effect level $c(\psi_i)$ for each sampled unit $T_i = 1$, estimating the ATE by the sampled average $\hat{\theta} = (1/n_T) \sum_i T_i c(\psi_i)$. Our analysis shows that if $T_{1:n} \sim \text{Loc}(\psi, q)$, then $\hat{\theta}$ behaves like the infeasible average over all n eligible units:

$$(1/n_T) \sum_{i=1}^n T_i c(\psi_i) = E_n[c(\psi_i)] + o_p(n^{-1/2}). \quad (4.4)$$

Since $n_T/n \xrightarrow{p} q$, this accounts for the reduction from $\text{Var}(c(\psi))$ to $q \text{Var}(c(\psi))$ in the first term of Equation (4.3). In the general case with $q(\psi)$ non-constant, the treatment effect heterogeneity term in Equation (4.2) is attenuated to $\bar{q} \text{Var}(c(\psi))$ with $\bar{q} = E[q(\psi)]$.

The second term in Equation (4.3) captures the variance from treatment effect heterogeneity orthogonal to ψ , as well as random imbalances in the potential outcomes during second-stage treatment assignment. Such fluctuations cannot be affected by stratified sampling on ψ with constant q . To reduce this term, we need to change the relative sampling proportions $q(\psi)$, taking more samples of higher-variability units. This dependence on $q(\psi)$ is reflected in the residual variance term in Equation (4.2).

Adjustment for Sampling Imbalances. One could instead try to correct for sampling imbalances between the eligible and experimental populations using ex-post adjustment. Dahabreh et al. (2019) propose such a method in a related setting with iid sampling and assignment $T_i \stackrel{\text{iid}}{\sim} \text{Bernoulli}(q(\psi_i))$ and $D_i \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p(\psi_i))$. Their estimator adjusts for covariate imbalances during both sampling and assignment by estimating regressions $m_d(\psi) = E[Y(d)|\psi]$ and the realized propensities $\hat{p}(\psi)$ and $\hat{q}(\psi)$. Subsequent

work in [Li et al. \(2023\)](#) showed that this estimator attains the variance V in [Theorem 4.2](#), and also proved that V is the semiparametric efficiency bound in this iid setting.

We develop a cross-fit version of the [Dahabreh et al. \(2019\)](#) estimator, showing it attains the variance V in [Equation 4.2](#) without the need for re-estimation of the known propensities $p(\psi)$ and $q(\psi)$, though the regression functions $m_d(\psi)$ still need to be consistently estimated. This result and related discussion can be found in [Appendix B.1](#).

In contrast to these ex-post adjustment strategies, [Theorem 4.2](#) shows that under a more careful design with finely stratified sampling $T_{1:n} \sim \text{Loc}(\psi, q(\psi))$ and assignment $D_{1:n} \sim \text{Loc}(\psi, p(\psi))$, we can attain the variance V using a simple estimator $\hat{\theta}$, plugging in the known propensities without any ex-post modeling, or just running the linear regression $Y \sim 1 + D$ in the constant propensity case. This shows how fine stratification nonparametrically controls covariate imbalances during both the sampling and assignment, obviating the need for ex-post adjustment.

Remark 4.3 (Comparison with Other Designs). One alternative to fine stratification is rerandomization, where one redraws random treatment assignments or sampling variables until a covariate balance criterion is satisfied. [Li et al. \(2018\)](#) show that under rerandomized treatment assignment, $\hat{\theta}$ is asymptotically slightly less efficient than an interacted regression that adjusts linearly for ψ . Similarly, [Yang et al. \(2023\)](#) study a two-stage design that rerandomizes at both the sampling and assignment stages, showing it is asymptotically slightly less efficient than ex-post linear adjustment for covariate imbalances in ψ at both stages. By contrast, fine stratification using the algorithm in [Section 3](#) nonparametrically controls covariate imbalances at both stages.

Remark 4.4 (Smoothness Conditions). Previous work on efficiency of fine stratification has required Lipschitz continuity of the functions $E[Y(d)|\psi]$ and $\text{Var}(Y(d)|\psi)$, e.g. see [Bai et al. \(2022\)](#). We develop a novel technical analysis allowing these conditions to be removed, showing the universality of the efficiency gains from fine stratification under the mild moment condition $E[Y(d)^2] < \infty$.

4.1 Extensions

In this section, we extend our main result to allow for stratified sampling and assignment with different covariates ψ_1 and ψ_2 , characterizing the optimal covariates to stratify on at each stage. We also extend our asymptotic theory to target estimation of the $\text{SATE} = E_n[Y_i(1) - Y_i(0)]$ in the eligible population instead of the ATE.

It may be useful to stratify on different sets of variables during the sampling and assignment stages. For example, ψ_1 could contain publicly available administrative data, while ψ_2 contains further survey covariates collected after sampling. The next result extends Theorem 4.2 to this setting. We write the variance in a form that more clearly disambiguates the variance contributions from each sampling stage. To that end, let individual effect $\tau = Y(1) - Y(0)$ and define *outcome level* $\bar{Y} = (1 - p)Y(1) + pY(0)$.

Theorem 4.5 (CLT). *Let $\psi_1 = f(\psi_2)$ for some function $f(\cdot)$. If $T_{1:n} \sim \text{Loc}(\psi_1, q)$ and $D_{1:n} \sim \text{Loc}(\psi_2, p)$, then $\sqrt{n_T}(\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$ with*

$$V = q \text{Var}(\tau) + (1 - q)E[\text{Var}(\tau|\psi_1)] + \frac{E[\text{Var}(\bar{Y}|\psi_2)]}{p(1 - p)}. \quad (4.5)$$

Writing $V = V_1 + V_2 + V_3$, the term $V_1 = q \text{Var}(\tau)$ is the irreducible variance from iid sampling of the eligible units. The term $V_2 = (1 - q)E[\text{Var}(\tau|\psi_1)]$ is the variance from finely stratified sampling from the eligible pool, which is small if ψ_1 is highly predictive of treatment effect heterogeneity $\tau = Y(1) - Y(0)$. Finally, $V_3 \propto E[\text{Var}(\bar{Y}|\psi_2)]$ is the variance from random imbalances in the outcome levels \bar{Y} between $D = 1$ and $D = 0$, which is small if ψ_2 is highly predictive of \bar{Y} . In the case $\psi_1 = \psi_2 = \psi$, Equation (4.5) can be rearranged to yield the variance in our earlier theorem (4.2).

The technical condition $\psi_1 = f(\psi_2)$ is satisfied, for instance, if ψ_2 contains all of the variables originally included in ψ_1 . It is not immediately clear whether it can be removed with a finer analysis.

Finite Population Estimand. In some settings, the eligible population of size n may constitute the entire population of interest. In this case it is natural to instead target the sample average treatment effect $\text{SATE} = E_n[Y_i(1) - Y_i(0)]$ in the eligible population. Note this differs from alternative definitions of the SATE as the average of treatment effects within the experimental population $E_n[Y_i(1) - Y_i(0)|T_i = 1]$, which has less external validity, though they coincide if $q = 1$.

Corollary 4.6 (SATE). *In the setting of Theorem 4.5, $\sqrt{n_T}(\hat{\theta} - \text{SATE}) \Rightarrow \mathcal{N}(0, V_{\text{SATE}})$,*

$$V_{\text{SATE}} = (1 - q)E[\text{Var}(\tau|\psi_1)] + \frac{E[\text{Var}(\bar{Y}|\psi_2)]}{p(1 - p)}. \quad (4.6)$$

Targeting the SATE simply removes the variance from the first stage sampling of n eligible units from the superpopulation, resulting in $V_{\text{SATE}} = V - q \text{Var}(\tau)$. See

Appendix D.2 for more discussion about targeting the SATE vs. the ATE and the interpretation of the population measure P in this context.

For completeness, we also provide versions of Theorem 4.5 and Corollary 4.6 for general sampling and assignment propensities $q(\psi)$ and $p(\psi)$. See Theorem C.6 and Corollary C.7 in the appendix for the precise theorem statements.

Optimal Covariates. The results above reveal an important asymmetry between the optimal covariates to use for stratified sampling and assignment. For sampling, we want covariates ψ_1 that are highly predictive of individual treatment effect heterogeneity $\tau = Y(1) - Y(0)$. By contrast, for random assignment we want covariates ψ_2 that are highly predictive of outcome levels \bar{Y} . To see this even more clearly, observe that

$$E[\text{Var}(\tau|\psi_1)] = \min_{f(\cdot)} E[(\tau - f(\psi_1))^2] \quad \text{and} \quad E[\text{Var}(\bar{Y}|\psi_2)] = \min_{f(\cdot)} E[(\bar{Y} - f(\psi_2))^2].$$

In either case, it is important to choose a small set of such highly predictive covariates, since the finite sample variance $n \text{Var}(\hat{\theta})$ converges to the variance bound V slowly in high dimensions, with $n \text{Var}(\hat{\theta}) = V + O(n^{-2/(\dim(\psi)+1)})$ under conditions given Theorem 4.7 immediately below. It is an open question to determine exactly how many covariates to use in a specific study given experimenter beliefs. One simple heuristic is to only finely stratify on baseline versions of the important endline outcomes.

Uniform Convergence. Here we strengthen the asymptotic results above by presenting a finite sample bound quantifying how far the variance $n \text{Var}(\hat{\theta})$ can deviate from the efficient variance V , uniformly over a class of outcome distributions.

Let $\mathcal{P} = \mathcal{P}(B, M, K, \alpha)$ denote a class of distributions of $W = (\psi, Y(0), Y(1))$ such that for $d \in \{0, 1\}$ the conditional mean $\psi \mapsto E[Y(d)|\psi]$ is Lipschitz with constant at most B , $|Y(d)| \leq M$ almost surely, and $E[|\psi|_2^\alpha] \leq K$. Also require $\alpha > \dim(\psi) + 1$. For $P \in \mathcal{P}$, let $V(P)$ be the semiparametric variance bound (4.3) under P .

We fix the design with $q = 1$ and constant assignment propensity p for simplicity.

Theorem 4.7 (Uniform Convergence). *For $D_{1:n} \sim \text{Loc}(\psi, p)$ and \mathcal{P} as above,*

$$\sup_{P \in \mathcal{P}} |n \text{Var}_P(\hat{\theta}) - V(P)| \lesssim B^2 n^{2/\alpha - 2/(\dim(\psi)+1)} + M^2/n.$$

If we have a uniform bound $|\psi|_2 \leq K$ with probability one for all $P \in \mathcal{P}$, the bound simplifies to $B^2 n^{-2/(\dim(\psi)+1)} + M^2/n$. This result shows the effect of the curse of dimensionality in matching discussed above, suggesting fine stratification on a small

number of highly predictive covariates.

5 Optimal Sampling for Survey Experiments

In this section we apply our main asymptotic results to study the optimal sampling problem for budget-constrained experimentation with heterogeneous costs. After characterizing the oracle sampling design, we propose a feasible implementation that estimates unknown quantities using a available pilot or observational data, providing conditions under which this is asymptotically optimal.

5.1 Budget-Constrained Optimal Sampling

Our goal is to choose a sampling propensity $q(\psi)$ that minimizes the asymptotic variance of ATE estimation in an experiment with heterogeneous costs and fixed budget constraint. For simplicity, throughout this section we let assignment propensity $p(\psi) = p$ constant, e.g. with leading case $p = 1/2$, though our main results in the appendix allow for the case where $p(\psi)$ is varying as well.

To begin, recall in Section 4 we presented asymptotics of the form $\sqrt{n_T}(\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$, normalizing by the experiment size $n_T = \sum_i T_i$. This normalization is typical in the literature, allowing for easy comparison with previous results, such as the variance bound in Hahn (1998). However, the experiment size n_T varies with the sampling propensity $q(\psi)$, making this normalization unsuitable for our current task of optimizing over $q(\psi)$ given a fixed budget and eligible units $i \in [n]$.

Because of this, in what follows we instead normalize by the number of eligible units n . Since $n_T/n \xrightarrow{p} E[q(\psi)]$, this removes the multiplicative factor $E[q(\psi)]$ from our previous results. Then by Theorem 4.2, $\sqrt{n}(\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V(q))$ for

$$V(q) = \text{Var}(c(\psi)) + E \left[\frac{1}{q(\psi)} \left(\frac{\sigma_1^2(\psi)}{p} + \frac{\sigma_0^2(\psi)}{1-p} \right) \right]. \quad (5.1)$$

Similarly, by Corollary C.7 in the appendix $\sqrt{n}(\hat{\theta} - \text{SATE}) \Rightarrow \mathcal{N}(0, V_{\text{SATE}}(q))$ for $T_{1:n} \sim \text{Loc}(\psi, q(\psi))$, where $V_{\text{SATE}}(q) = V(q) - \text{Var}(\tau)$ for $\tau = Y(1) - Y(0)$. Since the two objectives differ by a constant that does not depend on $q(\cdot)$, the optimal sampling propensity is the same whether we target the SATE or the ATE. We therefore restrict attention to the ATE in what follows.

Consider minimizing Equation 5.1 over all sampling propensities $q(\psi)$. Clearly the unconstrained solution is $q^*(\psi) = 1$, making the experiment as large as possible. More generally, we can formalize a problem of sampling with heterogeneous costs subject to a fixed budget constraint.

Costs. Define $C(\psi)$ to be the known cost of including a unit of type ψ in the experiment. One natural specification is $C(\psi) = C_s(\psi) + pC_1(\psi) + (1-p)C_0(\psi)$, where $C_s(\psi)$ is the sampling cost and $C_1(\psi)$ and $C_0(\psi)$ are the marginal costs of assigning treatment and control, respectively. For example, in a development context $C_s(\psi)$ could be the cost of paying volunteers to collect outcome data in a village $\psi_i = \psi$. We can interpret $C(\psi)$ as the ex-ante cost of sampling a unit with $\psi_i = \psi$ into the experiment, prior to realization of its random treatment assignment $D_i \in \{0, 1\}$.

Define the ex-ante heteroskedasticity

$$\sigma_s^2(\psi) = \frac{\sigma_1^2(\psi)}{p} + \frac{\sigma_0^2(\psi)}{1-p}, \quad (5.2)$$

and let $\sigma_s(\psi) = \sigma_s^2(\psi)^{1/2}$ denote the corresponding ex-ante standard deviation. We interpret $\sigma_s^2(\psi)$ as the expected residual variance from sampling a unit with covariates $\psi_i = \psi$ into the experiment. Then $V(q) = \text{Var}(c(\psi)) + E[\sigma_s^2(\psi)/q(\psi)]$, so for a given budget B we would like to solve

$$\min_{0 < q(\cdot) \leq 1} E \left[\frac{\sigma_s^2(\psi)}{q(\psi)} \right] \quad \text{s.t.} \quad E[C(\psi)q(\psi)] = B. \quad (5.3)$$

The next theorem characterizes the interior solutions to this problem.

Theorem 5.1. *Assume $E[Y(d)^2] < \infty$ for $d = 0, 1$, $\inf_{\psi} \sigma_d^2(\psi) \geq c > 0$, and $C(\psi) \in [C_l, C_u] \subseteq (0, \infty)$. Define the candidate solution*

$$q^*(\psi) = B \cdot \frac{\sigma_s(\psi)C(\psi)^{-1/2}}{E[\sigma_s(\psi)C(\psi)^{1/2}]}. \quad (5.4)$$

If $\sup_{\psi} q^(\psi) \leq 1$, then q^* is optimal in Equation 5.3.*

To build intuition for the form of the solution, consider the following special cases.

- (a) *Homoskedasticity.* Suppose $\sigma_d^2(\psi) = \sigma_d^2$ constant for $d \in \{0, 1\}$. Then $\sigma_s^2(\psi)$ is constant and $q^*(\psi) = B \cdot C(\psi)^{-1/2} / E[C(\psi)^{1/2}]$. In particular, $q^*(\psi) \propto C(\psi)^{-1/2}$, so that more costly units are sampled with lower probability.

(b) *Homogeneous Costs.* If $C(\psi) = C$, then the budget constraint is $E[q(\psi)] \leq B/C$. Denoting $\bar{q} = B/C$, the optimal solution has form $q^*(\psi) = \bar{q}\sigma_s(\psi)/E[\sigma_s(\psi)]$. This shows how we would like to oversample units with above-average residual standard deviation, with $q^*(\psi) > \bar{q}$ when $\sigma_s(\psi) > E[\sigma_s(\psi)]$ and the reverse otherwise.

Under the homoskedasticity assumption in (a), the optimal propensity $q^*(\psi)$ only depends on the known costs $C(\psi)$ and distribution of ψ , so we do not require outcome data to estimate $q^*(\psi)$. However, heteroskedasticity is believed to be widely prevalent in economic data (e.g. [Romano and Wolf \(2017\)](#)), so the propensity in (a) may leave efficiency gains on the table. In the general case, we also require an estimate of the unknown residual variance $\sigma_s^2(\psi)$.

Note also that Equation 5.4 can be viewed as an analog of classical optimal allocation theory in survey sampling ([Cochran, 1977](#)), here applied to budget-constrained finely stratified sampling into an experiment.

5.2 Feasible Optimal Sampling

If data from a pilot experiment $(T_i, D_i, Y_i, \psi_i)_{i=1}^{n_{\text{pilot}}}$ is available, one can use the strategy of [Fan and Yao \(1998\)](#) to estimate the individual heteroskedasticity functions $\hat{\sigma}_d^2(\psi)$ for $d \in \{0, 1\}$ and form the composite $\hat{\sigma}_s^2(\psi) = \hat{\sigma}_1^2(\psi)/p + \hat{\sigma}_0^2(\psi)/(1-p)$.

To do so, one first forms the conditional-mean signal $S_i(1) = Y_i D_i T_i / (p_i q_i)$, where p_i and q_i are the pilot propensities, noting $E[S_i(1)|\psi_i] = E[Y_i(1)|\psi_i]$. Projecting $S_i(1)$ on ψ_i via a nonparametric regression yields $\hat{m}_1(\psi_i)$. A second projection of the IPW-weighted squared residuals $(Y_i - \hat{m}_1(\psi_i))^2 D_i T_i / (p_i q_i)$ on ψ_i then yields $\hat{\sigma}_1^2(\psi_i)$. The analogous construction with $1 - D_i$ in place of D_i and $1 - p_i$ in place of p_i gives $\hat{\sigma}_0^2(\psi)$.

The choice of nonparametric regression is flexible. We tested multiple methods, including kernel ridge regression and random forests, and found random forests perform best across the settings in our empirical application. See Appendix B.2 for details.

However, experimenters do not always have the resources to run a pilot. Those that do often use a very small sample size, making estimation of $\hat{\sigma}_s^2(\psi)$ challenging. Even without a pilot, observational data $(T_i, Y_i, \psi_i)_{i=1}^{n_{\text{obs}}}$ for units under the control condition $Y_i = Y_i(0)$ for $T_i = 1$ may still be available. If we are willing to assume $\sigma_1^2(\psi) \approx \sigma_0^2(\psi)$ are similar between treatment arms, then $\sigma_s^2(\psi) \approx \sigma_0^2(\psi)/(p(1-p))$, which can be estimated consistently as above using only data on the control outcome $(T_i, Y_i, \psi_i)_{i=1}^{n_{\text{obs}}}$.

To see why this equal heteroskedasticity assumption is plausible, consider the survey

in Cai and Rafi (2024) of 10 recent RCTs published in the AER. They estimated the within-arm variance ratio σ_1^2/σ_0^2 for each measured outcome, finding $\sigma_1^2/\sigma_0^2 \approx 1$ for the vast majority of them. This provides strong empirical support for such an assumption.

Given such an estimator $\hat{\sigma}_s^2(\psi)$ of $\sigma_s^2(\psi)$, Theorem 5.1 suggests using fine stratification to implement a suitable discretization of the estimated optimal propensity

$$\hat{q}(\psi) = B \cdot \frac{\hat{\sigma}_s(\psi)C(\psi)^{-1/2}}{E_n[\hat{\sigma}_s(\psi_i)C(\psi_i)^{1/2}]}.$$
 (5.5)

Note that under homoskedasticity, the residual variance factors out of the expression above, and we can just set $\hat{q}(\psi) = B \cdot C(\psi)^{-1/2}/E_n[C(\psi_i)^{1/2}]$.

Discretization. To implement the propensity estimate $\hat{q}(\psi)$ using fine stratification, we construct a discretization $\hat{q}_n(\psi)$ of $\hat{q}(\psi)$ into the rational approximating set $\{a_l/k_l\}_{l=1}^{L_n}$ such that $E_n[(\hat{q}_n(\psi_i) - \hat{q}(\psi_i))^2] = o_p(1)$.

For example, $\hat{q}_n(\psi)$ can be obtained by rounding $\hat{q}(\psi)$ to the nearest a/k_n for some sequence $k_n \rightarrow \infty$. Larger k_n gives worse match quality, but better implementation of the estimated propensity $\hat{q}(\psi)$. Thus, in finite samples this parameter trades off between the variance due to random covariate imbalances in ψ versus optimization of the residual variance using $q(\psi)$. Below we provide conditions on the maximum tuple size $\bar{k}_n = \max_{l \in [L_n]} k_l$ and number of propensity levels L_n for this match-quality degradation to be asymptotically lower order. In practice, it is often possible to choose a reasonable level of discretization fineness by inspecting the estimated $\hat{q}(\psi_i)$ to see how many different propensity levels are needed.

Assumption 5.2 (Feasible Optimal Sampling). *Suppose the following:*

- (i) $T_{1:n} \sim \text{Loc}(\psi, \hat{q}_n(\psi))$ and $D_{1:n} \sim \text{Loc}(\psi, p)$. Variance estimates $\hat{\sigma}_d^2(\psi)$ for $d \in \{0, 1\}$ are computed from external data ξ_n , independent of the experimental units and design variables.
- (ii) Moments $E[Y(d)^4] < \infty$ for $d \in \{0, 1\}$ and $E[|\psi|_2^\alpha] < \infty$ for some $\alpha > \dim(\psi) + 1$. The heteroskedasticity is bounded below, $\inf_\psi \sigma_d^2(\psi) > 0$. The propensities satisfy $\hat{q}_n(\psi), q^*(\psi) \in (\delta, 1]$. The costs satisfy $C(\psi) \in [C_l, C_u] \subset (0, \infty)$.
- (iii) Heteroskedasticity estimates satisfy $E_\psi[(\hat{\sigma}_d^2(\psi) - \sigma_d^2(\psi))^2] = o_p(1)$ for $d \in \{0, 1\}$.
- (iv) The discretization has $E_n[(\hat{q}_n(\psi_i) - \hat{q}(\psi_i))^2] = o_p(1)$, with $\bar{k}_n L_n = o(n^{1 - (\dim(\psi) + 1)/\alpha})$ and $\bar{k}_n = o(n^{1/4})$ for $\bar{k}_n = \max_{l \in [L_n]} k_l$.

The consistency requirement on the heteroskedasticity estimates $\widehat{\sigma}_d^2(\psi)$ in part (iii) can be satisfied, for example, by the estimator of [Fan and Yao \(1998\)](#). See [Appendix B.2](#) for a more detailed discussion of our implementation. Part (iv) requires a discretization $\widehat{q}_n(\psi)$ of increasing fineness, subject to the upper bound $\bar{k}_n L_n = o(n^{1-(\dim(\psi)+1)/\alpha})$. For example, [Lemma D.25](#) shows that rounding $\widehat{q}(\psi)$ to the nearest a/k_n , for any $k_n \rightarrow \infty$ with $k_n = o(n^{(1-(\dim(\psi)+1)/\alpha)/2} \wedge n^{1/4})$, satisfies part (iv). Finally, the reason for the rate $\bar{k}_n L_n = o(n^{1-(\dim(\psi)+1)/\alpha})$ in part (iv) is to ensure the sequence of local randomizations in part (i) exist. See [Lemma C.3](#) for the formal result showing $F(\mathcal{G}_n^T) = o_p(1)$ and $F(\mathcal{G}_n^D) = o_p(1)$ under these conditions.

The conditions above ensure the variance due to random covariate imbalances is lower order, while also allowing asymptotic implementation of the optimal $q^*(\psi)$.

Theorem 5.3 (Feasible Optimal Sampling). *Impose [Assumption 5.2](#). Suppose that $T_{1:n} \sim \text{Loc}(\psi, \widehat{q}_n(\psi))$ and $D_{1:n} \sim \text{Loc}(\psi, p(\psi))$. Then $\sqrt{n}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V^*)$*

$$V^* = \text{Var}(c(\psi)) + \min_{\substack{0 < q(\cdot) \leq 1 \\ E[C(\psi)q(\psi)] = B}} E \left[\frac{\sigma_s^2(\psi)}{q(\psi)} \right].$$

In particular, this design minimizes the asymptotic variance $V(q)$ of [Equation 5.1](#) over all budget-feasible sampling propensities $q(\psi)$.

If the condition $\sup_{\psi} q^*(\psi) \leq 1$ fails, we may obtain infeasible propensity estimates $\widehat{q}(\psi_j) > 1$. To restore feasibility, we can iteratively set $\widehat{q}(\psi_j) = 1$ for such j and recompute the optimal propensity for the remaining units.

To do so, 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 on the remaining units

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

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

Remark 5.4 (Optimal Assignment Propensity). In an earlier version of this paper, we noted that for any fixed sampling propensity $q(\psi)$, the globally optimal assignment propensity is the conditional Neyman allocation $p^*(\psi) = \sigma_1(\psi)/(\sigma_1(\psi) + \sigma_0(\psi))$, pro-

viding a feasible version $D_{1:n} \sim \text{Loc}(\psi, \hat{p}_n(\psi))$ of this optimal design. However, recent work by [Cai and Rafi \(2024\)](#) has documented that σ_1^2 and σ_0^2 tend to be similar in experimental settings in economics, so the baseline $p = 1/2$ will often be approximately optimal under equal costs.

6 Inference Methods

This section provides new methods for asymptotically exact inference on the ATE and asymptotically conservative inference on the SATE under finely stratified sampling and assignment. We build on the pairs-of-pairs estimators originally introduced in [Abadie and Imbens \(2008\)](#) and studied in [Bai et al. \(2022\)](#) and [Bai et al. \(2026\)](#), among others.

We focus on the constant assignment propensity case $p = a/k$, so that each non-remainder assignment group $g \in \mathcal{G}_n$ has $|g| = k$. Throughout this section, we identify the partition \mathcal{G}_n with the non-remainder groups of full size k . For each assignment group $g \in \mathcal{G}_n$, define the group-level difference in means

$$\hat{\theta}_g = \frac{1}{a} \sum_{i \in g} D_i Y_i - \frac{1}{k-a} \sum_{i \in g} (1 - D_i) Y_i. \quad (6.1)$$

Then the IPW estimator (4.1) is $\hat{\theta} = |\mathcal{G}_n|^{-1} \sum_{g \in \mathcal{G}_n} \hat{\theta}_g$. Our variance estimators combine a between-group sample variance $\hat{S}^2 = \frac{1}{|\mathcal{G}_n|} \sum_{g \in \mathcal{G}_n} (\hat{\theta}_g - \hat{\theta})^2$ with a residual variance estimator \hat{P}^2 defined below, which captures variation conditional on ψ . For the simplest case $q(\psi) = q$, the estimators for inference on the SATE and ATE are

$$\hat{V}_{\text{SATE}} = k\hat{P}^2, \quad \hat{V} = q\hat{S}^2 + (k-q)\hat{P}^2. \quad (6.2)$$

We give two constructions of the residual estimator \hat{P}^2 . Both have the same probability limit, but differ in their finite-sample bias.

Within-group estimator. When $\min(a, k-a) \geq 2$, we can estimate the within- ψ residual variation directly inside each group. Let $\bar{Y}_{1,g} = a^{-1} \sum_{i \in g} D_i Y_i$ and $\bar{Y}_{0,g} = (k-a)^{-1} \sum_{i \in g} (1 - D_i) Y_i$ be the treated and control means in group g , so that $\hat{\theta}_g = \bar{Y}_{1,g} - \bar{Y}_{0,g}$, and define within-arm sample variances $s_{1,g}^2 = \frac{1}{a-1} \sum_{i \in g} D_i (Y_i - \bar{Y}_{1,g})^2$ and

$s_{0,g}^2 = \frac{1}{k-a-1} \sum_{i \in g} (1 - D_i)(Y_i - \bar{Y}_{0,g})^2$. The within-group estimator is

$$\widehat{P}_N^2 = \frac{1}{|\mathcal{G}_n|} \sum_{g \in \mathcal{G}_n} \left(\frac{s_{1,g}^2}{a} + \frac{s_{0,g}^2}{k-a} \right). \quad (6.3)$$

The summand $s_{1,g}^2/a + s_{0,g}^2/(k-a)$ is the Neyman variance estimator for the difference in means within group g , treating each matched group as a small completely randomized two-arm experiment (Neyman, 1990; Imbens and Rubin, 2015).

Pairs-of-pairs estimator. If $\min(a, k-a) = 1$, then at least one of the sample variances above is infeasible, necessitating the use of pairs-of-pairs style constructions (Abadie and Imbens, 2008). First, suppose that $|\mathcal{G}_n|$ is even. Let $\nu : \mathcal{G}_n \rightarrow \mathcal{G}_n$ be a bijective matching with $\nu(g) \neq g$ and $\nu^2 = \text{Id}$ such that the centroids $\bar{\psi}_g = |g|^{-1} \sum_{i \in g} \psi_i$ satisfy the pairwise tight matching condition $n^{-1} \sum_{g \in \mathcal{G}_n} |\bar{\psi}_g - \bar{\psi}_{\nu(g)}|^2 = o_p(1)$. In practice, ν is obtained by matching the centroids into pairs using the algorithm of Section 3, which guarantees this condition (Lemma D.23). The pairs-of-pairs residual estimator is

$$\widehat{P}_\nu^2 = \frac{1}{2|\mathcal{G}_n|} \sum_{g \in \mathcal{G}_n} (\widehat{\theta}_g - \widehat{\theta}_{\nu(g)})^2. \quad (6.4)$$

It compares estimates between tightly matched groups and applies for any $p = a/k$. If $|\mathcal{G}_n|$ is odd, we match one of the groups twice, contributing an extra term to \widehat{P}_ν^2 .

Bias Comparison. Inference on the SATE using $\widehat{V}_{\text{SATE}} = k\widehat{P}_\nu^2$ is conservatively biased in finite samples. In a design-based framework without stratified sampling, Bai et al. (2026) show that $E[\widehat{V}_{\text{SATE}}] = V_{\text{SATE}} + B_n$ for bias $B_n = \frac{k}{2|\mathcal{G}_n|} \sum_{g \in \mathcal{G}_n} (\theta_g - \theta_{\nu(g)})^2 \geq 0$, where $\theta_g = k^{-1} \sum_{i \in g} \tau_i$ is the average treatment effect in group g . The factor k shows this bias can be severe for the large tuples required by imbalanced assignment propensities p far from $1/2$. For example, this estimator performs very poorly for the Baysan (2022) example in our empirical application, with $p = 2/11$. By contrast, a calculation shows that the bias of $k\widehat{P}_N^2$ is of constant order in k , see Lemma D.20 for details.

We therefore set $\widehat{P}^2 = \widehat{P}_N^2$ when $\min(a, k-a) \geq 2$ and $\widehat{P}^2 = \widehat{P}_\nu^2$ otherwise. Both estimators are transparently non-negative for every realization: \widehat{S}^2 and \widehat{P}_ν^2 are sums of squares, and \widehat{P}_N^2 is a sum of sample variances.

Varying Sampling Propensity. When $q(\psi)$ takes finitely many values $q_l = a_l/k_l$, $l = 1, \dots, L$, we apply the constructions above within each propensity stratum $\{i : q(\psi_i) = q_l\}$. In this case, we assume sampling-subordinate matching of the assignment

partition, so that each assignment group lies within a single sampling stratum. Let sampling stratum size $n_l = |\{i : q(\psi_i) = q_l\}|$ and let $\hat{\theta}_l$, \hat{S}_l^2 , and \hat{P}_l^2 denote the within-stratum versions of $\hat{\theta}$, \hat{S}^2 , and \hat{P}^2 . The variance estimators are

$$\hat{V}_{\text{SATE}} = \frac{n_T}{n} \sum_{l=1}^L \frac{n_l}{n} \cdot \frac{k}{q_l} \hat{P}_l^2, \quad \hat{V} = \frac{n_T}{n} \sum_{l=1}^L \frac{n_l}{n} \left(\hat{S}_l^2 + \frac{k - q_l}{q_l} \hat{P}_l^2 + (\hat{\theta}_l - \hat{\theta})^2 \right). \quad (6.5)$$

The ATE estimator adds a cross-stratum correction $(\hat{\theta}_l - \hat{\theta})^2$ that recovers between-stratum variation in the conditional average treatment effect.

Theorem 6.1 (Inference). *Assume the conditions of Theorem 4.2 with assignment groups matched within sampling strata. Then*

$$\hat{V} \xrightarrow{p} V, \quad \hat{V}_{\text{SATE}} \xrightarrow{p} V_{\text{SATE}} + E[q(\psi)] E[\text{Var}(\tau|\psi)].$$

In the constant- q case with $L = 1$, the cross-stratum correction vanishes and the estimators reduce to (6.2), differing only by the converging factor $n_T/(qn) \xrightarrow{p} 1$.

By this theorem and the CLT in Section 4, the interval $\hat{C} = [\hat{\theta} \pm \hat{V}^{1/2} c_{1-\alpha/2}/\sqrt{n_T}]$ with critical value $c_\alpha = \Phi^{-1}(\alpha)$ is asymptotically non-conservative for the ATE, with $P(\text{ATE} \in \hat{C}) = 1 - \alpha + o(1)$.

By contrast, inference on the SATE is necessarily conservative, since the asymptotic variance V_{SATE} in Corollary 4.6 depends on $E[\text{Var}(\tau|\psi)]$, which is not identified because individual treatment effects τ_i are not observed. Combined with Corollary 4.6, the result above shows that the interval $\hat{C}_{\text{SATE}} = [\hat{\theta} \pm \hat{V}_{\text{SATE}}^{1/2} c_{1-\alpha/2}/\sqrt{n_T}]$ is asymptotically valid for the SATE, with $P(\text{SATE} \in \hat{C}_{\text{SATE}}) \geq 1 - \alpha + o(1)$.

Remark 6.2 (Prior Work). For finely stratified assignment without stratified sampling, the pairs-of-pairs construction of \hat{V}_{SATE} was originally introduced by [Abadie and Imbens \(2008\)](#) for $p = 1/2$. This is a direct analog of the classical collapsed-strata estimator from the survey-sampling literature, e.g. [Hansen et al. \(1953\)](#). [Bai et al. \(2026\)](#) extend the analysis of this estimator to the case $p = a/k$, providing finite-sample bias guarantees in a design-based framework. For inference on the ATE, our estimator \hat{V} can be shown to coincide with that of [Bai et al. \(2022\)](#) for matched pairs, $p = 1/2$.

7 Empirical Results

In this section, we quantify the finite sample performance of each of our designs on $N = 9$ real DGPs from experimental papers covering a range of fields in applied economics.

7.1 Designs and Empirical Papers

Our theoretical results showed separate variance reductions from (a) finely stratified treatment assignment, (b) finely stratified sampling, and (c) stratified implementation of the estimated optimal sampling propensity $q^*(\psi)$. To quantify the efficiency gain from each of these tools, we simulate unadjusted ATE estimation under the following designs:

CR: Complete randomization $T_{1:n} \sim \text{CR}(\hat{q}_n)$ and $D_{1:n} \sim \text{CR}(p)$, with \hat{q}_n a discretization of the budget-exhausting sampling propensity $\hat{q} = B/E_n[C(\psi_i)]$ and p the fixed assignment propensity from the original paper. In particular, we let $\hat{q}_n = a/k$, using the minimal k such that $\hat{q}_n \cdot E_n[C(\psi_i)] \in [.95B, 1.05B]$. The costs $C(\psi)$ and discretization rule are discussed in more detail below.

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

Loc: Stratified sampling and assignment $T_{1:n} \sim \text{Loc}(\psi, \hat{q}_n)$ and $D_{1:n} \sim \text{Loc}(\psi, p)$.

Hom: As in **Loc** but with sampling propensity $\hat{q}_n(\psi)$ a discretization of $\hat{q}(\psi) = B \cdot C(\psi)^{-1/2} / E_n[C(\psi_i)^{1/2}]$, the optimal sampling propensity assuming homoskedasticity. This is feasible but possibly misspecified.

Pilot S/L: As in **Loc**, but with $T_{1:n} \sim \text{Loc}(\psi, \hat{q}_n(\psi))$ where $\hat{q}_n(\psi)$ is the discretization of a pilot-based estimate $\hat{q}(\psi)$ formed from an experiment of size $n_{\text{pilot}} = n_T/5$ (S) or $n_{\text{pilot}} = 4n_T/5$ (L) with $p = 1/2$, $q = 1$, and matched-pair assignment. We estimate the ex-ante variance $\hat{\sigma}_s^2(\psi)$ from the pilot by random forest regression as in Appendix B.2, then form $\hat{q}(\psi)$ via the optimal sampling formula in Equation (5.5).

Obs: As in **Pilot**, but $\hat{q}(\psi)$ is estimated from observational data $(Y_i(0), \psi_i)_{i=1}^{n_{\text{obs}}}$ of size $n_{\text{obs}} = 4n_T$ under the working assumption $\sigma_1^2(\psi) = \sigma_0^2(\psi)$, as discussed in Section 5. We estimate $\hat{\sigma}_0^2(\psi)$ from $(Y_i(0), \psi_i)_{i=1}^{n_{\text{obs}}}$ as in Appendix B.2 and use the plug-in $\hat{\sigma}_s^2(\psi) = \hat{\sigma}_0^2(\psi)/(p(1-p))$ for the ex-ante variance $\sigma_s^2(\psi)$ in Equation (5.2), then form $\hat{q}(\psi)$ via Equation (5.5).

We evaluate each of these designs on data from experimental papers published in the AER between May 2021 and November 2022. We exclude papers for which data

is unavailable or that do not fit into our framework for various reasons, e.g. having multiple interventions on the same unit with a time series structure. The included papers are [Abebe et al. \(2021\)](#), [Baysan \(2022\)](#), [Casey et al. \(2021\)](#), [DellaVigna et al. \(2022\)](#), [Domurat et al. \(2021\)](#), [Hussam et al. \(2022\)](#), and [Lowe \(2021\)](#). We also include [Banerjee et al. \(2025\)](#), as well as [Finkelstein et al. \(2012\)](#), for a total of $N = 9$.

For each paper, we use the observed data to impute a DGP to sample from. In particular, for $d = 0, 1$ we estimate the conditional mean $\widehat{m}_d(\psi) = \widehat{E}[Y(d)|\psi]$ and variance $\widehat{\sigma}_d^2(\psi) = \widehat{\text{Var}}(Y(d)|\psi)$, and take the outcome law to be $Y(d)|\psi \sim \mathcal{N}(\widehat{m}_d(\psi), \widehat{\sigma}_d^2(\psi))$. For the two papers with binary outcomes, [Casey et al. \(2021\)](#) and [Domurat et al. \(2021\)](#), we instead take $Y(d)|\psi \sim \text{Bernoulli}(\widehat{m}_d(\psi))$, where $\widehat{m}_d(\psi) = \widehat{P}(Y(d) = 1|\psi)$ is a fitted classifier. The regression model for imputation is selected separately for each paper by out-of-sample cross-validation between OLS, ridge, lasso with interactions, gradient-boosted trees, and random forest. See [Appendix B.3](#) for further details. Let N_0 denote the size of the original experiment.

Simulation Design. We do the following:

- (1) We draw $4N_0$ eligible units by sampling with replacement from the N_0 original units $(\psi_i)_{i=1}^{N_0}$, then perturbing each unit’s continuous coordinates by independent mean-zero noise, so that ψ_i is continuously distributed and exact ties $\psi_i = \psi_j$ have probability zero ([Appendix B.3](#)). For each unit, we form outcomes $\widetilde{Y}_i(d) = \widehat{m}_d(\psi_i) + \widehat{\sigma}_d^2(\psi_i)^{1/2}\epsilon_i^d$, $d = 0, 1$, with the residual $\epsilon_i^d \sim \mathcal{N}(0, 1)$ drawn independently across units and $d \in \{0, 1\}$. The budget-exhausting constant propensity $q = 1/4$ results in a representative experiment of size $n_T = N_0$.
- (2) Randomize $T_{1:n}$ and $D_{1:n}$ according to one of the designs above.
- (3) Reveal outcomes $\widetilde{Y}_i = T_i D_i \widetilde{Y}_i(1) + T_i(1 - D_i) \widetilde{Y}_i(0)$ and form the estimator $\widehat{\theta}$. Form confidence intervals $\widehat{C} = [\widehat{\theta} \pm \widehat{V}^{1/2} c_{1-\alpha/2} / \sqrt{n_T}]$ for the ATE and $\widehat{C}_{\text{SATE}} = [\widehat{\theta} \pm \widehat{V}_{\text{SATE}}^{1/2} c_{1-\alpha/2} / \sqrt{n_T}]$ for the SATE, with $\alpha = 0.05$ and \widehat{V} , $\widehat{V}_{\text{SATE}}$ as in [Section 6](#).

Since the ATE of the estimated population is known, we can compute the standard deviation (SD), coverage probabilities, and percent reduction in confidence interval length for each design on each DGP. We focus on the ATE in this section for brevity. Empirical results for the SATE are presented in [Appendix D.4](#).

Descriptions of each paper, including the treatment and outcome variables and our choice of stratification variables ψ are provided in [Appendix D.3](#). Experiment sizes n_T are as in the original papers, ranging from $n_T = 91$ for [Casey et al. \(2021\)](#) to $n_T = 1903$

| | Design, Paper | Abe. | Ban. | Bay. | Cas. | Del. | Dom. | Fin. | Hus. | Low. |
|---------------|---------------|------|------|------|------|------|------|------|------|------|
| % Δ SD | CR | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | CR, Loc | -5 | -43 | -47 | -1 | -38 | 0 | -8 | -2 | -10 |
| | Loc | -6 | -41 | -48 | 1 | -39 | 2 | -6 | -8 | -7 |
| | Hom. | -15 | -42 | -48 | -7 | -39 | -3 | -6 | -13 | -11 |
| | Pilot S | -12 | -49 | -45 | -7 | -41 | 2 | -19 | -13 | -13 |
| | Pilot L | -16 | -48 | -46 | -3 | -40 | -2 | -21 | -9 | -16 |
| | Obs | -14 | -51 | -47 | -2 | -41 | -5 | -22 | -12 | -15 |
| % Δ CI | CR | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | CR, Loc | -6 | -43 | -46 | -1 | -38 | 0 | -8 | -5 | -10 |
| | Loc | -6 | -41 | -46 | 0 | -38 | 0 | -8 | -5 | -9 |
| | Hom. | -13 | -41 | -46 | -6 | -40 | -3 | -9 | -9 | -9 |
| | Pilot S | -13 | -49 | -42 | -7 | -39 | 0 | -21 | -9 | -16 |
| | Pilot L | -13 | -50 | -44 | -6 | -40 | -3 | -21 | -9 | -18 |
| | Obs | -14 | -51 | -44 | -6 | -41 | -2 | -21 | -10 | -17 |
| Cover | CR | 0.95 | 0.95 | 0.94 | 0.95 | 0.94 | 0.95 | 0.95 | 0.94 | 0.95 |
| | CR, Loc | 0.95 | 0.95 | 0.94 | 0.95 | 0.95 | 0.95 | 0.95 | 0.92 | 0.95 |
| | Loc | 0.96 | 0.95 | 0.94 | 0.94 | 0.95 | 0.95 | 0.95 | 0.94 | 0.95 |
| | Hom. | 0.95 | 0.95 | 0.94 | 0.94 | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 |
| | Pilot S | 0.95 | 0.95 | 0.94 | 0.95 | 0.96 | 0.94 | 0.94 | 0.95 | 0.94 |
| | Pilot L | 0.96 | 0.94 | 0.94 | 0.94 | 0.95 | 0.95 | 0.95 | 0.93 | 0.94 |
| | Obs | 0.95 | 0.95 | 0.95 | 0.94 | 0.94 | 0.95 | 0.95 | 0.94 | 0.94 |
| | n_T | 1451 | 903 | 550 | 91 | 446 | 1000 | 1903 | 116 | 770 |
| | $\dim(\psi)$ | 8 | 7 | 3 | 3 | 3 | 4 | 7 | 4 | 3 |

Table 1: Empirical Results (ATE).

for [Finkelstein et al. \(2012\)](#). The one exception is [Domurat et al. \(2021\)](#) ($N_0 = 87394$), for which we set $n_T = 1000$ for Monte Carlo tractability. Baseline treatment proportions p are set to those of the original published designs: $p = 1/2$ in six of nine papers, $p = 2/3$ for [Lowe \(2021\)](#), $p = 1/3$ for [Finkelstein et al. \(2012\)](#), and $p = 2/11$ for [Baysan \(2022\)](#). The latter three papers with $p \neq 1/2$ are newly enabled by our matched k -tuples algorithm in Section 3. We use sampling-subordinate matching, as in Section 6.

Costs and Discretization. The marginal cost $C(\psi)$ of including different units in the experiment is not reported in the papers in our sample. To quantify variance reductions from representative sampling with $q^*(\psi)$ in the type of DGPs that occur in applied economics research, we specify $C(\psi) = \mathbb{1}(|\psi|_2 \leq \kappa) + 5\mathbb{1}(|\psi|_2 > \kappa)$ with κ the median of $|\psi_i|_2$ over the eligible pool and $B = 0.75$, yielding the budget-exhausting constant propensity $q = 1/4$ for the **CR**, **CR**, **Loc**, and **Loc** designs. This propensity discretizes to matched 4-tuples ($\hat{q}_n = a/k$ with $k = 4$), so the empirical application exercises the general matched k -tuples algorithm of Section 3 with $k > 2$ throughout, not only in the three papers with $p \neq 1/2$.

The varying- q designs (**Hom**, **Pilot S/L**, **Obs**) realize $\hat{q}_n(\psi) \in [0.1, 0.9]$, with empirical mean sampling rate $E_n[\hat{q}_n(\psi_i)] \approx 0.33$ averaged across these designs and papers. By spending more of the budget on cheaper units, these designs raise the average sampling rate above the constant \hat{q}_n of **CR**, **CR**, **Loc**, and **Loc**, which has $\hat{q}_n = 1/4$ in our cost specification. For each varying- q design (**Hom**, **Pilot S/L**, **Obs**), the discretized sampling propensity $\hat{q}_n(\psi)$ is formed by choosing values in $\{a/10 : a = 1, \dots, 10\}$ to minimize $E_n[(\hat{q}_n - \hat{q})^2(\psi_i)]$ subject to $L_n = |\text{Image}(\hat{q}_n)| \leq 2$ for $n_T < 500$, $L_n \leq 3$ for $500 \leq n_T < 1000$, and $L_n \leq 4$ for $n_T \geq 1000$.

7.2 Results

Our main results are presented in Table 1. The largest single SD reduction comes from finely stratified treatment assignment (**CR** vs. **CR**, **Loc**), with an average of -17% across our sample and reductions exceeding -30% in papers with highly predictive baseline covariates such as [Banerjee et al. \(2025\)](#), [Baysan \(2022\)](#), and [DellaVigna et al. \(2022\)](#). For these datasets, finely stratified sampling does not help on average (**CR**, **Loc** vs. **Loc**). This is expected from the theory, since stratified sampling reduces only the treatment effect heterogeneity component $\text{Var}(c(\psi))$ of the variance, and here the nonparametric R^2 of ψ for $Y(1) - Y(0)$ in our imputed DGP's averages only 4% across the nine papers. The gains from the varying-rate designs below instead arise from

heteroskedasticity and cost heterogeneity, which ψ does capture.

Averaging over the nine papers, the marginal reductions over **Loc** are -3.6% for **Hom**, -5.7% for **Pilot S**, -6.2% for **Pilot L**, and -7.5% for **Obs**. The homoskedasticity optimal design **Hom** has mixed effects across papers, while the feasible **Pilot S/L** and **Obs** designs deliver consistently larger reductions, though pilot estimation noise can offset the gain on some papers (e.g. **Pilot S** on [Domurat et al. \(2021\)](#)). Small pilots ($n_{\text{pilot}} = 0.2n_T$) and large pilots ($n_{\text{pilot}} = 0.8n_T$) perform similarly on average, and the observational-data design **Obs**, which uses only $(Y_i(0), \psi_i)$ observations under the assumption $\sigma_1^2(\psi) = \sigma_0^2(\psi)$, is empirically comparable to **Pilot L** on most papers. This provides further empirical support for this equal-heteroskedasticity assumption.

For inference, coverage is close to nominal across all 63 (paper, design) cells, with mean coverage 0.946 and range $[0.92, 0.96]$. The percent reduction in confidence interval length $\% \Delta \text{CI}$ closely tracks the percent reduction in standard deviation $\% \Delta \text{SD}$ across designs (within about one percentage point on average), so the confidence intervals are well calibrated and recover nearly all of the efficiency gains from finely stratified sampling and assignment. Empirical results for the SATE, presented in Section [D.4](#), are similarly well calibrated and yield slightly shorter confidence intervals than their ATE counterparts, as expected by Corollary [4.6](#).

8 Recommendations for Practice

First, we discuss the choice of matching algorithm. For basic matched pairs randomization, one can use the optimal matching of [Derigs \(1988\)](#), whose match quality guarantee was established by [Bai et al. \(2022\)](#). For fine stratification with $\dim(\psi) = 1$, one can simply sort the units by their ψ_i values, as in [Bai \(2022\)](#). For all other cases, or for large matched pairs experiments with $n \geq 10,000$, we recommend the new matching algorithm developed in Section [3](#). This is the first algorithm to deliver provable match quality guarantees for matched k -tuples of any size $k \geq 2$ and $\dim(\psi) \geq 1$. The spatial sorting step alone scales to massive experiments with millions of units, while the refined procedure adding balanced k -means remains tractable into the tens of thousands of units. These guarantees hold whether the stratification variable ψ is fully continuous, fully discrete, or a mix of the two, requiring only the weak moment condition of Theorem [3.1](#).

Next we turn to the design itself. Finely stratified assignment delivers the single largest marginal variance reduction in our empirical application of Section [7](#), averaging

–16% across the nine papers. Our results show an additional nonparametric variance reduction from finely stratified sampling into the experiment, which is larger the more eligible units n we have to choose from and the more predictive ψ is of treatment effect heterogeneity. This gives researchers an incentive to source a large pool of eligible units, then use fine stratification to sample representative experimental participants.

In particular, researchers should stratify on covariates expected to predict treatment effect heterogeneity, as has been advocated by the J-PAL guide on sampling for experiments (Kopper and Riis-Vestergaard, 2023). For treatment assignment, the optimal covariates are instead those most predictive of outcome levels, though baseline outcomes typically serve both roles well. If these covariates are indeed predictive, our inference methods report smaller standard errors.

For both sampling and assignment, one should stratify on a small set of such covariates. Indeed, the finite sample variance approaches the efficient bound at the rate $n^{-2/(\dim(\psi)+1)}$ of Theorem 4.7, which degrades quickly as $\dim(\psi)$ grows. Researchers should avoid finely stratifying on highly predictive covariates such as baseline outcomes together with many weakly predictive ones.

Finally, when costs are highly heterogeneous, using a varying sampling propensity $q^*(\psi)$ can reduce variance even under the basic design assuming homoskedasticity, which does not require pilot data. We found similar performance across small pilots, large pilots, and the observational-data design, each reducing standard deviation by roughly 6% over constant-propensity fine stratification. Since a large pilot is rarely available in practice, the observational-data design is an attractive default when control-arm data on $(Y(0), \psi)$ can be obtained at low cost. However, there is a finite sample penalty from reduced match quality, so when both costs and the residual variance are roughly homogeneous, it may be preferable to use a constant sampling propensity q .

In all cases, our inference methods performed well, giving close to nominal coverage for the ATE that reflects the efficiency gains from both finely stratified sampling and assignment, as well as valid coverage for the SATE.

References

- Abadie, A. and G. W. Imbens (2008). Estimation of the conditional variance in paired experiments. *Annales d’Economie et de Statistique* (91-92), 175–187.
- Abebe, G., A. S. Caria, and E. Ortiz-Ospina (2021). The selection of talent: Exper-

- imental and structural evidence from ethiopia. *American Economic Review* 111(6), 1757–1806.
- Armstrong, T. B. (2025). Asymptotic efficiency bounds for a class of experimental designs.
- Athey, S. and G. W. Imbens (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments*, Volume 1, pp. 73–140. Elsevier (North-Holland).
- Bai, Y. (2022). Optimality of matched-pair designs in randomized controlled trials. *American Economic Review* 112(12), 3911–3940.
- Bai, Y., X. Huang, J. P. Romano, A. M. Shaikh, and M. Tabord-Meehan (2026). A new design-based variance estimator for finely stratified experiments. arXiv:2503.10851.
- Bai, Y., J. Liu, A. M. Shaikh, and M. Tabord-Meehan (2026). On the efficiency of highly stratified experiments. arXiv:2307.15181.
- Bai, Y., J. Liu, and M. Tabord-Meehan (2024). Inference for matched tuples and fully blocked factorial designs. *Quantitative Economics* 15(2), 279–330.
- Bai, Y., J. P. Romano, and A. M. Shaikh (2022). Inference in experiments with matched pairs. *Journal of the American Statistical Association* 117(540), 1726–1737.
- Bai, Y., A. M. Shaikh, and M. Tabord-Meehan (2025). A primer on the analysis of randomized experiments and a survey of some recent advances. *Journal of Political Economy: Microeconomics*. Forthcoming; arXiv:2405.03910.
- Banerjee, A., A. G. Chandrasekhar, S. Dalpath, E. Duflo, J. Floretta, M. O. Jackson, H. Kannan, F. N. Loza, A. Sankar, A. Schrimpf, and M. Shrestha (2025). Selecting the most effective nudge: Evidence from a large-scale experiment on immunization. *Econometrica* 93(4), 1183–1223.
- Baysan, C. (2022). Persistent polarizing effects of persuasion: Experimental evidence from turkey. *American Economic Review* 112(11), 3528–3546.
- Billingsley, P. (1995). *Probability and Measure*. Wiley.
- Brixius, N. (2025). Optimal matched block design for multi-arm experiments. In *Learning and Intelligent Optimization – 19th International Conference, LION 19, Prague, Czech Republic, June 15–19, 2025, Proceedings, Part II*, Volume 15745, pp. 48–63. Springer.
- Broockman, D. E., E. Rhodes, A. W. Bartik, K. Dotson, S. Miller, P. K. Krause, and E. Vivaldi (2024). The causal effects of income on political attitudes and behavior: A randomized field experiment. NBER Working Paper No. 33214.

- Bugni, F. A., I. A. Canay, and A. M. Shaikh (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association* 113(524), 1784–1796.
- Cai, Y. and A. Rafi (2024). On the performance of the Neyman allocation with small pilots. *Journal of Econometrics* 242(1), 105793.
- Casey, K., A. Bakarr Kamara, and N. F. Meriggi (2021). An experiment in candidate selection. *American Economic Review* 111(5), 1575–1612.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review* 107(5), 261–265.
- Chioda, L., P. Gertler, D. Contreras-Loya, and D. R. Carney (2026). Making entrepreneurs: Long term returns to training youth in business skills. NBER Working Paper No. 34637.
- Cochran, W. G. (1977). *Sampling Techniques* (3 ed.). John Wiley and Sons.
- Dahabreh, I. J., S. E. Robertson, E. J. Tchetgen Tchetgen, E. A. Stuart, and M. A. Hernán (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics* 75(2), 685–694.
- de Chaisemartin, C. and J. Ramirez-Cuellar (2024). At what level should one cluster standard errors in paired and small-strata experiments? *American Economic Journal: Applied Economics* 16(1), 193–212.
- Degtiar, I. and S. Rose (2023). A review of generalizability and transportability. *Annual Review of Statistics and Its Application* 10, 501–524.
- DellaVigna, S., J. A. List, U. Malmendier, and G. Rao (2022). Estimating social preferences and gift exchange at work. *American Economic Review* 112(3), 1038–1074.
- Derigs, U. (1988). Solving non-bipartite matching problems via shortest path techniques. *Annals of Operations Research* 13, 225–261.
- Domurat, R., I. Menashe, and W. Yin (2021). The role of behavioral frictions in health insurance marketplace enrollment and risk: Evidence from a field experiment. *American Economic Review* 111(5), 1549–1574.
- Er, M. C. (1984). On generating the N-ary reflected Gray codes. *IEEE Transactions on Computers* C-33(8), 739–741.
- Fan, J. and Q. Yao (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* 85(3), 645–660.
- Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, and K. Baicker (2012). The oregon health insurance experiment: Evidence from the first year. *The Quarterly Journal of Economics* 127(3), 1057–1106.

- Fogarty, C. B. (2018). On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80(5), 1035–1056.
- Folland, G. B. (1999). *Real Analysis: Modern Techniques and Their Applications*. Wiley. 2nd edition.
- Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66(2), 315–332.
- Hahn, J., K. Hirano, and D. Karlan (2011). Adaptive experimental design using the propensity score. *Journal of Business and Economic Statistics* 29(1), 96–108.
- Hansen, M. H., W. N. Hurwitz, and W. G. Madow (1953). *Sample Survey Methods and Theory*. Wiley.
- Harshaw, C., F. Sävje, D. A. Spielman, and P. Zhang (2024). Balancing covariates in randomized experiments with the gram–schmidt walk design. *Journal of the American Statistical Association* 119(548), 2934–2946.
- Higgins, M. J., F. Sävje, and J. S. Sekhon (2016). Improving massive experiments with threshold blocking. *Proceedings of the National Academy of Sciences* 113(27), 7369–7376.
- Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4), 1161–1189.
- Hussam, R., E. M. Kelley, G. Lane, and F. Zahra (2022). The psychosocial value of employment: Evidence from a refugee camp. *American Economic Review* 112(11), 3694–3724.
- Imai, K., G. King, and C. Nall (2009). The essential role of pair matching in cluster-randomized experiments, with application to the mexican universal health insurance evaluation. *Statistical Science* 24(1), 29–53.
- Imbens, G. W. and D. B. Rubin (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*.
- Kallus, N. (2018). Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80(1), 85–112.
- Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis* 24(3), 324–338.
- Kasy, M. and A. Sautmann (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica* 89(1), 113–132.

- Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics* 17(2), 3008–3049.
- Klenke, A. (2020). *Probability Theory: A Comprehensive Course* (3 ed.). Universitext. Springer.
- Kopper, S. and M. Riis-Vestergaard (2023). Sampling. Abdul Latif Jameel Poverty Action Lab (J-PAL) Research Resources. <https://www.povertyactionlab.org/resource/sampling>, accessed 2026.
- Krieger, A. M., D. Azriel, and A. Kapelner (2019). Nearly random designs with greatly improved balance. *Biometrika* 106(3), 695–701.
- Li, F., H. Hong, and E. A. Stuart (2023). A note on semiparametric efficient generalization of causal effects from randomized trials to target populations. *Communications in Statistics – Theory and Methods* 52(16), 5767–5798.
- Li, X., P. Ding, and D. B. Rubin (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences* 115(37), 9157–9162.
- Lohr, S. L. (2021). *Sampling Design and Analysis*. Routledge.
- Lowe, M. (2021). Types of contact: A field experiment on collaborative and adversarial caste integration. *American Economic Review* 111(6), 1807–1844.
- Malinen, M. I. and P. Fränti (2014). Balanced K-means for clustering. In *Structural, Syntactic, and Statistical Pattern Recognition (S+SSPR)*, Volume 8621 of *Lecture Notes in Computer Science*, pp. 32–41. Springer.
- Moore, R. T. (2012). Multivariate continuous blocking to improve political science experiments. *Political Analysis* 20(4), 460–479.
- Muralidharan, K. and P. Niehaus (2017). Experimentation at scale. *Journal of Economic Perspectives* 31(4), 103–124.
- Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science* 5(4), 465–472. Translated and edited by D. M. Dabrowska and T. P. Speed from the Polish original (1923).
- Pyatkin, A., D. Aloise, and N. Mladenović (2017). NP-hardness of balanced minimum sum-of-squares clustering. *Pattern Recognition Letters* 97, 44–45.
- Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89(427), 846–866.
- Romano, J. P. and M. Wolf (2017). Resurrecting weighted least squares. *Journal of Econometrics* 197(1), 1–19.

- Rosenberger, W. F. and J. M. Lachin (2016). *Randomization in Clinical Trials: Theory and Practice*. John Wiley & Sons. 2nd edition.
- Rothwell, P. M. (2005). External validity of randomised controlled trials: “to whom do the results of this trial apply?”. *The Lancet* 365(9453), 82–93.
- Shim, M., K. H. Kim, M. A. Lee, S. Choi, S. Bae, O. Coibion, and Y. Gorodnichenko (2026). The effects of fiscal news on household expectations and spending: New causal evidence. NBER Working Paper No. 35009.
- Tabord-Meehan, M. (2023). Stratification trees for adaptive randomisation in randomised controlled trials. *The Review of Economic Studies* 90(5), 2646–2673.
- Wang, X., T. Wang, and H. Liu (2023). Rerandomization in stratified randomized experiments. *Journal of the American Statistical Association* 118(542), 1295–1304.
- Yang, Z., T. Qu, and X. Li (2023). Rejective sampling, rerandomization, and regression adjustment in survey experiments. *Journal of the American Statistical Association* 118(542), 1207–1221.

A Proof of Matching Algorithm Guarantee

In this section, we work with triangular arrays $(\psi_{i,n})_{i=1}^n \subseteq \mathbb{R}^{\dim(\psi)}$ and propensities $(q_{i,n})_{i=1}^n \subseteq \mathbb{Q}$ taking levels in $\{a_l/k_l\}_{l=1}^{L_n}$ with $\bar{k}_n = \max_l k_l$. For any subset $S_n \subseteq [n]$, the propensity stratum $S_{nl} = \{i \in S_n : q_{i,n} = a_l/k_l\}$ and $n_l = |S_{nl} \setminus r_l|$ counts the non-remainder units, where $\mathcal{R}_n = \{r_l : l \in [L_n]\}$ collects the remainder groups from preprocessing step (1) of the algorithm. Typical choices are $S_n = [n]$ at the sampling stage and $S_n = \{i : T_i = 1\}$ at the assignment stage. Define the squared maximal coordinate range $M_n^2 = \max_{j \leq \dim(\psi)} (\max_i \psi_{i,n}^{(j)} - \min_i \psi_{i,n}^{(j)})^2$.

This section proves the guarantees claimed for the algorithm in Section 3. Denote objective function $F(\mathcal{G}) = n^{-1} \sum_{g \in \mathcal{G}} \sum_{i \in g} |\psi_{i,n} - \bar{\psi}_g|_2^2$ for any collection of disjoint groups $\mathcal{G} \subseteq 2^{[n]}$, where centroid $\bar{\psi}_g = |g|^{-1} \sum_{i \in g} \psi_{i,n}$. Then $F(\mathcal{G} \cup \mathcal{G}') = F(\mathcal{G}) + F(\mathcal{G}')$ for disjoint $\mathcal{G}, \mathcal{G}'$. The proofs often use the objective function’s equivalent representation as a sum of within-group pairwise differences

$$F(\mathcal{G}) = \frac{1}{2n} \sum_{g \in \mathcal{G}} \frac{1}{|g|} \sum_{i,j \in g} |\psi_{i,n} - \psi_{j,n}|_2^2. \quad (\text{A.1})$$

This follows from the identity $\sum_{i,j \in g} |\psi_{i,n} - \psi_{j,n}|_2^2 = 2|g| \sum_{i \in g} |\psi_{i,n} - \bar{\psi}_g|_2^2$.

Assumption A.1 (Matching). *Require the following:*

- (a) $\max_{j \leq \dim(\psi)} (\max_{i \in S'_{nl}} \psi_{i,n}^{(j)} - \min_{i \in S'_{nl}} \psi_{i,n}^{(j)}) > 0$ for each $j \in \dim(\psi)$.
- (b) $\bar{k}_n L_n / n = o(1)$.

As in Section 3, Part (a) is without loss since if any covariate has zero range, it can be safely excluded from step (2) of the algorithm without changing the matching objective. Part (b) ensures that the remainder r_l removed from each S_{nl} to make stratum size divisible by k_l is negligible.

Theorem A.2 (Spatial Sorting). *Require A.1. The spatial sorting algorithm from Section 3 with grid sizes $m_l \asymp (n_l/k_l)^{1/(d+1)}$ returns $\mathcal{G}_n^{(0)} = \cup_{l=1}^{L_n} \mathcal{G}_{nl}^{(0)} \cup \mathcal{R}_n$ with*

$$F(\mathcal{G}_n^{(0)}) \leq M_n^2 \cdot O((n/\bar{k}_n L_n)^{-2/(d+1)}). \quad (\text{A.2})$$

Proof. We begin by reducing the problem to the case where all per-stratum points lie in $[0, 1]^d$ and the remainder groups have been removed. Additivity of F and group disjointness implies $F(\mathcal{G}_n^{(0)}) = \sum_{l=1}^{L_n} F(\mathcal{G}_{nl}^{(0)}) + F(\mathcal{R}_n)$. By Lemma A.7, the remainder satisfies $F(\mathcal{R}_n) \leq n^{-1} M_n^2 d \bar{k}_n L_n = O(M_n^2 (n/\bar{k}_n L_n)^{-2/(d+1)})$.

Then consider $\sum_{l=1}^{L_n} F(\mathcal{G}_{nl}^{(0)})$. For each $l \in [L_n]$ and $i \in S'_{nl}$, define $\phi_{i,n} \in [0, 1]^d$ to be the rescaled $\psi_{i,n}$ defined in the preprocessing step in part (1) of Section 3. By Lemma A.8 applied to $\mathcal{S} = S'_{nl}$, we have $F(\mathcal{G}_{nl}^{(0)}, \psi) \leq M_n^2 \cdot F(\mathcal{G}_{nl}^{(0)}, \phi)$. Summing over l , we obtain $\sum_{l=1}^{L_n} F(\mathcal{G}_{nl}^{(0)}, \psi) \leq M_n^2 \sum_{l=1}^{L_n} F(\mathcal{G}_{nl}^{(0)}, \phi)$, so it suffices to show $\sum_{l=1}^{L_n} F(\mathcal{G}_{nl}^{(0)}, \phi) = O((n/\bar{k}_n L_n)^{-2/(d+1)})$. Then we relabel ϕ as ψ and assume $\{\psi_{i,n} : i \in S'_{nl}\} \subseteq [0, 1]^d$ for each l for the remainder of the proof.

Fix $l \in [L_n]$ and grid size parameter $m_l \geq 1$. Let $\rho(\psi) = \rho_d(\lfloor m_l \psi \rfloor \wedge (m_l - 1))$ denote the sorting index and $B(0), \dots, B(m_l^d - 1)$ the block partition of $[0, 1]^d$ from Lemma A.5. For each unit $i \in S'_{nl}$ set $t(i) = \rho(\psi_{i,n}) \in \{0, \dots, m_l^d - 1\}$. Then $\psi_{i,n} \in B(t(i))$ holds since by definition $B(t) = \rho^{-1}(t)$. The partition $\mathcal{G}_{nl}^{(0)} = \{g_{l,s} : s = 1, \dots, K_l\}$ is formed by sorting the non-remainder units on $t(i)$, breaking ties at random, so the groups satisfy the following *monotonicity* property: for group indices $1 \leq s < s' \leq K_l$, we have

$$i \in g_{l,s}, j \in g_{l,s'} \implies t(i) \leq t(j). \quad (\text{A.3})$$

We will bound the terms in $F(\mathcal{G}_{nl}^{(0)})$ by a case analysis, depending on whether each pair $i, j \in g$ for $g \in \mathcal{G}_{nl}^{(0)}$ lies in the same block $t(i) = t(j)$, or different blocks $t(i) \neq t(j)$.

Case 1. Suppose $t(i) = t(j)$. Then both points lie in a common block of side length

$1/m_l$, so $|\psi_{i,n} - \psi_{j,n}|_2 \leq \sqrt{d}/m_l$ by item (2) of Lemma A.5. Then summing over $\mathcal{G}_{n_l}^{(0)}$,

$$\begin{aligned} & \sum_{g \in \mathcal{G}_{n_l}^{(0)}} \frac{1}{|g|} \sum_{i \neq j \in g} \mathbb{1}\{t(i) = t(j)\} |\psi_{i,n} - \psi_{j,n}|_2^2 \leq \frac{d}{m_l^2} \sum_{g \in \mathcal{G}_{n_l}^{(0)}} \frac{1}{|g|} \sum_{i \neq j \in g} \mathbb{1}\{t(i) = t(j)\} \\ & \leq \frac{d}{m_l^2} \sum_{g \in \mathcal{G}_{n_l}^{(0)}} \frac{|g|(|g| - 1)}{|g|} = \frac{d}{m_l^2} \sum_{g \in \mathcal{G}_{n_l}^{(0)}} (|g| - 1) \leq \frac{dn_l}{m_l^2}. \end{aligned} \quad (\text{A.4})$$

The final inequality holds since $g \in \mathcal{G}_{n_l}^{(0)}$ are disjoint subsets of S'_{n_l} with $|S'_{n_l}| = n_l$.

Case 2. Suppose $t(i) \neq t(j)$. Without loss, suppose $t(i) < t(j)$. By item (1) of Lemma A.5, for each index t with $t(i) < t < t(j)$ we can pick an arbitrary element $y_t \in B(t)$. Also set $y_{t(i)} = \psi_{i,n}$ and $y_{t(j)} = \psi_{j,n}$. Item (3) of Lemma A.5 gives the bound $|y_{t+1} - y_t|_2 \leq 2\sqrt{d}/m_l$ for any such t . Then by telescoping, triangle inequality, and since $|\psi_{i,n} - \psi_{j,n}|_2 \leq \sqrt{d}$ on $[0, 1]^d$, we have

$$|\psi_{i,n} - \psi_{j,n}|_2^2 \leq \sqrt{d} \cdot |\psi_{i,n} - \psi_{j,n}|_2 \leq \sqrt{d} \sum_{t=t(i)}^{t(j)-1} |y_{t+1} - y_t|_2 \leq \frac{2d}{m_l} |t(i) - t(j)|. \quad (\text{A.5})$$

Within each group $g_{l,s}$, enumerate the units as $g_{l,s} = \{u_{s,1}, \dots, u_{s,k_l}\}$. Re-indexing the inner sum by this enumeration and applying (A.5) termwise, we obtain

$$\begin{aligned} & \sum_{i,j \in g_{l,s}} \mathbb{1}\{t(i) \neq t(j)\} |\psi_{i,n} - \psi_{j,n}|_2^2 = \sum_{a,b=1}^{k_l} \mathbb{1}\{t(u_{s,a}) \neq t(u_{s,b})\} |\psi_{u_{s,a},n} - \psi_{u_{s,b},n}|_2^2 \\ & \leq \frac{2d}{m_l} \sum_{a,b=1}^{k_l} \mathbb{1}\{t(u_{s,a}) \neq t(u_{s,b})\} |t(u_{s,a}) - t(u_{s,b})| \leq \frac{2d}{m_l} \sum_{a,b=1}^{k_l} |t(u_{s,a}) - t(u_{s,b})|. \end{aligned}$$

For any $1 \leq s < K_l$ and $a, b \in \{1, \dots, k_l\}$, monotonicity (A.3) gives $t(u_{s,a}) \leq t(u_{s+1,b})$. Then the K_l intervals I_{ab}^s of the form $[t(u_{s,a}), t(u_{s,b})]$ or $[t(u_{s,b}), t(u_{s,a})]$ are disjoint up to endpoints with $\cup_{s=1}^{K_l} I_{ab}^s \subseteq [0, m_l^d - 1]$. With $\mathcal{L}(\cdot)$ Lebesgue measure, this implies

$$\sum_{s=1}^{K_l} |t(u_{s,a}) - t(u_{s,b})| = \sum_{s=1}^{K_l} \mathcal{L}(I_{ab}^s) = \mathcal{L}\left(\cup_{s=1}^{K_l} I_{ab}^s\right) \leq \mathcal{L}([0, m_l^d - 1]) \leq m_l^d.$$

Then aggregating using the bound above, we have

$$\begin{aligned} \sum_{s=1}^{K_l} \frac{1}{k_l} \sum_{i,j \in g_{l,s}} \mathbb{1}\{t(i) \neq t(j)\} |\psi_{i,n} - \psi_{j,n}|_2^2 &\leq \sum_{s=1}^{K_l} \frac{1}{k_l} \frac{2d}{m_l} \sum_{a,b=1}^{k_l} |t(u_{s,a}) - t(u_{s,b})| \\ &\leq \frac{1}{k_l} k_l (k_l - 1) \frac{2d}{m_l} m_l^d = 2d(k_l - 1) m_l^{d-1} \leq 2dk_l m_l^{d-1}. \end{aligned}$$

Combining the cases above and applying the pair-form representation (A.1), we obtain $F(\mathcal{G}_{nl}^{(0)}) \leq (2n)^{-1} [dn_l/m_l^2 + 2dk_l m_l^{d-1}]$. Setting $m_l \asymp (n_l/k_l)^{1/(d+1)}$ implies that $F(\mathcal{G}_{nl}^{(0)}) \leq n^{-1} C_d \cdot k_l^{2/(d+1)} n_l^{(d-1)/(d+1)}$ for some constant C_d . Summing over $l = 1, \dots, L_n$ and bounding $k_l \leq \bar{k}_n$, we obtain

$$\sum_{l=1}^{L_n} F(\mathcal{G}_{nl}^{(0)}) \leq n^{-1} C_d \bar{k}_n^{2/(d+1)} \sum_{l=1}^{L_n} n_l^{(d-1)/(d+1)} = n^{-1} C_d \bar{k}_n^{2/(d+1)} L_n \left(\frac{1}{L_n} \sum_{l=1}^{L_n} n_l^{(d-1)/(d+1)} \right).$$

Applying Jensen to $x \mapsto x^{(d-1)/(d+1)}$ and using $\sum_l n_l \leq n$, this is bounded above by

$$n^{-1} C_d \bar{k}_n^{2/(d+1)} L_n \left(\frac{1}{L_n} \sum_{l=1}^{L_n} n_l \right)^{(d-1)/(d+1)} \leq C_d (\bar{k}_n L_n)^{2/(d+1)} n^{-2/(d+1)}.$$

Combined with the reductions established at the start of the proof, this gives the bound $F(\mathcal{G}_n^{(0)}) \leq M_n^2 (C_d + d) (n/\bar{k}_n L_n)^{-2/(d+1)}$ for n sufficiently large, completing the proof. \square

Theorem A.3 (Algorithm Guarantee). *Require Assumption A.1. The algorithm in steps (1) through (3) in Section 3 with grid sizes $m_l \asymp (n_l/k_l)^{1/(d+1)}$ and per-stratum balanced- k -means iteration counts $T_l \geq 0$ returns a partition $\mathcal{G}_n = \cup_{l=1}^{L_n} \mathcal{G}_{nl}^{(T_l)} \cup \mathcal{R}_n$ with*

$$F(\mathcal{G}_n) \leq M_n^2 \cdot O((n/\bar{k}_n L_n)^{-2/(d+1)}). \quad (\text{A.6})$$

(1) *If $|\psi_i|_2 \leq c < \infty$ for $i = 1, \dots, n$, then $F(\mathcal{G}_n) = O((n/\bar{k}_n L_n)^{-2/(d+1)})$.*

(2) *If $(\psi_i)_{i=1}^n$ are iid and $E[|\psi|_2^\alpha] < \infty$ for some $\alpha > 0$, then*

$$F(\mathcal{G}_n) = o_p(n^{2/\alpha} (n/\bar{k}_n L_n)^{-2/(d+1)}).$$

Proof. Fix a level $l \in \{1, \dots, L_n\}$. Recall there are $K_l = n_l/k_l$ non-remainder groups in propensity stratum S'_{nl} , each of size k_l . We show $F(\mathcal{G}_{nl}^{(t+1)}) \leq F(\mathcal{G}_{nl}^{(t)})$ for every $t \geq 0$. The balanced assignment step (3.2) chooses $\mathcal{G}_{nl}^{(t+1)}$ to minimize $\sum_s \sum_{i \in g_s} |\psi_{i,n} - \bar{\psi}_{g_s}^{(t)}|_2^2$

over equal-size partitions. Since $\mathcal{G}_{nl}^{(t)}$ is feasible,

$$\sum_{s=1}^{K_l} \sum_{i \in g_s^{(t+1)}} |\psi_{i,n} - \bar{\psi}_{g_s}^{(t)}|_2^2 \leq \sum_{s=1}^{K_l} \sum_{i \in g_s^{(t)}} |\psi_{i,n} - \bar{\psi}_{g_s}^{(t)}|_2^2 = nF(\mathcal{G}_{nl}^{(t)}). \quad (\text{A.7})$$

For each s , expanding the square and using $\sum_{i \in g_s^{(t+1)}} (\psi_{i,n} - \bar{\psi}_{g_s}^{(t+1)}) = 0$ gives

$$\sum_{i \in g_s^{(t+1)}} |\psi_{i,n} - \bar{\psi}_{g_s}^{(t)}|_2^2 = \sum_{i \in g_s^{(t+1)}} |\psi_{i,n} - \bar{\psi}_{g_s}^{(t+1)}|_2^2 + k_l |\bar{\psi}_{g_s}^{(t+1)} - \bar{\psi}_{g_s}^{(t)}|_2^2 \geq \sum_{i \in g_s^{(t+1)}} |\psi_{i,n} - \bar{\psi}_{g_s}^{(t+1)}|_2^2.$$

Summing over s and combining with (A.7) yields the inequality $nF(\mathcal{G}_{nl}^{(t+1)}) \leq nF(\mathcal{G}_{nl}^{(t)})$, hence $F(\mathcal{G}_{nl}^{(t+1)}) \leq F(\mathcal{G}_{nl}^{(t)})$. Iterating, we have $F(\mathcal{G}_{nl}^{(T_l)}) \leq F(\mathcal{G}_{nl}^{(0)})$ for any $T_l \geq 0$. Summing over $l \in [L_n]$ and adding the remainder contribution we have

$$F(\mathcal{G}_n) = \sum_{l=1}^{L_n} F(\mathcal{G}_{nl}^{(T_l)}) + F(\mathcal{R}_n) \leq \sum_{l=1}^{L_n} F(\mathcal{G}_{nl}^{(0)}) + F(\mathcal{R}_n) = F(\mathcal{G}_n^{(0)}).$$

Theorem A.2 bounds $F(\mathcal{G}_n^{(0)})$ by the right side of (A.6), completing the proof of (A.6). For the bounded case, $|\psi_i|_2 \leq c$ for all i gives $M_n \leq 2c$, so $M_n^2 = O(1)$, and substituting into (A.6) yields item (1). For the moment case, Lemma A.9 applied to the iid sample gives $M_n^2 = o_p(n^{2/\alpha})$, and substituting into (A.6) yields item (2). \square

Corollary A.4 (Expected Matching Bound). *Let $(\psi_i)_{i=1}^n$ be iid with $E[|\psi|_2^\alpha] \leq K$ for some $\alpha > 2$, and let \mathcal{G}_n be the partition produced by the algorithm of Section 3. Then for a constant $C(\alpha, d)$ depending only on α and d*

$$E[F(\mathcal{G}_n)] \leq C(\alpha, d) K^{2/\alpha} \cdot n^{2/\alpha} \cdot (n/\bar{k}_n L_n)^{-2/(d+1)}.$$

Proof. By (A.6), $F(\mathcal{G}_n) \leq C_d M_n^2 \cdot (n/\bar{k}_n L_n)^{-2/(d+1)}$, with $M_n = \max_{j \leq d} (\max_i \psi_{i,n}^{(j)} - \min_i \psi_{i,n}^{(j)})$ and C_d depending only on d . The proof of Lemma A.9 gives the pathwise bound $M_n \leq 2 \max_i |\psi_i|_2$, so $M_n^2 \leq 4 \max_i |\psi_i|_2^2$. It remains to bound $E[\max_i |\psi_i|_2^2]$.

Set $Y_i = |\psi_i|_2$, so $E[Y^\alpha] \leq K$. By the layer-cake formula and a union bound,

$$E[\max_i Y_i^2] = \int_0^\infty P(\max_i Y_i^2 > t) dt \leq \int_0^\infty \min(1, nP(Y > \sqrt{t})) dt.$$

By Markov, $P(Y > \sqrt{t}) \leq Kt^{-\alpha/2}$. Splitting the integral at $t_0 = (nK)^{2/\alpha}$, where

$nKt_0^{-\alpha/2} = 1$, the first piece contributes at most $t_0 = (nK)^{2/\alpha}$ and the second is

$$\int_{t_0}^{\infty} nKt^{-\alpha/2} dt = \frac{2nK}{\alpha-2} t_0^{1-\alpha/2} = \frac{2}{\alpha-2} (nK)^{2/\alpha}.$$

The second equality used $t_0^{1-\alpha/2} = (nK)^{2/\alpha-1}$ and $\alpha > 2$. Summing, $E[\max_i Y_i^2] \leq \frac{\alpha}{\alpha-2} (nK)^{2/\alpha}$, so $E[M_n^2] \leq \frac{4\alpha}{\alpha-2} (nK)^{2/\alpha}$. Substituting into the pathwise bound gives the claim with $C(\alpha, d) = 4C_d \alpha / (\alpha - 2)$. \square

Lemma A.5 (Block Sequence). *Fix $d, m \geq 1$. Let $\rho_d : \{0, \dots, m-1\}^d \rightarrow \{0, \dots, m^d-1\}$ and $\rho : [0, 1]^d \rightarrow \{0, \dots, m^d-1\}$ be the index functions defined in Section 3 in dimension $d \geq 1$. Define blocks $B(t) = \rho^{-1}(t)$ for $t = 0, \dots, m^d-1$. The following hold:*

- (1) ρ_d is a bijection and $\{B(t) : t = 0, \dots, m^d-1\}$ is a partition of $[0, 1]^d$.
- (2) For $t = 0, \dots, m^d-1$, if $x, y \in B(t)$ then $|x - y|_2 \leq \sqrt{d}/m$.
- (3) For $t = 0, \dots, m^d-2$, if $x \in B(t)$, $y \in B(t+1)$ then $|x - y|_2 \leq 2\sqrt{d}/m$.

Proof. First consider (1). We show bijectivity of ρ_d by induction. For $d = 1$, $\rho_1(z_1) = z_1$ is the identity, so this holds trivially. For $d \geq 2$, note that the second summand in (3.1) lies in $\{0, \dots, m^{d-1}-1\}$. Then we can recover $z_d = \lfloor \rho_d(z) / m^{d-1} \rfloor$. Moreover, (z_1, \dots, z_{d-1}) can be recovered from the parity of z_d and using inductive bijectivity of ρ_{d-1} . Then we have constructed an inverse for ρ_d , so it is a bijection. For any $\psi \in [0, 1]^d$, $\lfloor m\psi \rfloor \in \{0, \dots, m\}^d$, so $\lfloor m\psi \rfloor \wedge (m-1) \in \{0, \dots, m-1\}^d$, hence $\rho(\psi) = \rho_d(\lfloor m\psi \rfloor \wedge (m-1))$ is a well-defined function from $[0, 1]^d$ to $\{0, \dots, m^d-1\}$. Then the level sets $B(t) = \rho^{-1}(t)$ of any function are pairwise disjoint with union equal to its domain. Moreover, each $B(t)$ is non-empty since $\rho(m^{-1}\rho_d^{-1}(t)) = \rho_d(\rho_d^{-1}(t)) = t$ so $m^{-1}\rho_d^{-1}(t) \in B(t)$. Then $\{B(t) : t = 0, \dots, m^d-1\}$ partitions $[0, 1]^d$.

For item (2), define $z(t) = \rho_d^{-1}(t)$. If $\psi \in B(t) = \rho^{-1}(t)$, then $\rho_d(\lfloor m\psi \rfloor \wedge (m-1)) = t$, so $z(t) = \rho_d^{-1}(t) = \lfloor m\psi \rfloor \wedge (m-1)$. Then componentwise, by definition of the floor function, $z(t) \leq m\psi \leq z(t) + 1$. Then for any $x, y \in B(t)$, we have $|x_j - y_j| \leq 1/m$ for each j , so $|x - y|_2^2 \leq d/m^2$, establishing (2).

For item (3), we claim $|z(t+1) - z(t)|_1 = 1$ for $t = 0, \dots, m^d-2$. Assuming the claim, let $x \in B(t)$, $y \in B(t+1)$. By the key inequality in the proof of (2), $z(t) \leq mx \leq z(t) + 1$ and $z(t+1) \leq my \leq z(t+1) + 1$ componentwise. Subtracting inequalities, for each coordinate j we obtain $z(t)_j - z(t+1)_j - 1 \leq m(x_j - y_j) \leq z(t)_j - z(t+1)_j + 1$, so $m(x_j - y_j)$ lies in an interval of length 2 centered at $z(t)_j - z(t+1)_j$. The maximum absolute value attained on such an interval is $|z(t)_j - z(t+1)_j| + 1$, so $|m(x_j - y_j)| \leq |z(t)_j - z(t+1)_j| + 1$.

By the claim, exactly one coordinate of $z(t+1) - z(t)$ has absolute value 1 and the rest are 0, so squaring and summing over j gives $m^2|x - y|_2^2 \leq \sum_j (|z(t)_j - z(t+1)_j| + 1)^2 = (1+1)^2 + (d-1) \cdot 1 = d+3 \leq 4d$. Hence $|x - y|_2 \leq 2\sqrt{d}/m$, establishing (3).

To finish the proof, we show the key claim $|z(t+1) - z(t)|_1 = 1$ for $t = 0, \dots, m^d - 2$. Recall we defined $z(t) = \rho_d^{-1}(t)$. For $d = 1$, ρ_1 is the identity and the adjacency property $|z(t+1) - z(t)|_1 = 1$ is immediate. For $d \geq 2$, assume the adjacency property holds for ρ_{d-1} . Suppose $\rho_d(z') = \rho_d(z) + 1$. Denote $\bar{z} = (z_1, \dots, z_{d-1})$ and similarly for \bar{z}' . Write $\rho_d(z) = z_d m^{d-1} + r(\bar{z}, z_d)$, defining $r(\bar{z}, z_d) \in \{0, \dots, m^{d-1} - 1\}$ to be the second summand in (3.1). We consider two cases.

Case 1: Suppose $z'_d = z_d$. Then $r(\bar{z}', z_d) - r(\bar{z}, z_d) = 1$. If z_d is even, inspecting the formula this implies $\rho_{d-1}(\bar{z}') - \rho_{d-1}(\bar{z}) = +1$. If z_d is odd, then $\rho_{d-1}(\bar{z}') - \rho_{d-1}(\bar{z}) = -1$. The inductive hypothesis applied to (\bar{z}, \bar{z}') or (\bar{z}', \bar{z}) yields that \bar{z}, \bar{z}' differ in exactly one coordinate by exactly 1. Together with $z_d = z'_d$ this gives the claim.

Case 2: Suppose $z'_d \neq z_d$. Since $\rho_d(z') - \rho_d(z) = (z'_d - z_d)m^{d-1} + (r(\bar{z}', z'_d) - r(\bar{z}, z_d)) = 1$ and $|r(\bar{z}', z'_d) - r(\bar{z}, z_d)| \leq m^{d-1} - 1$, the only possibility is $z'_d = z_d + 1$, $r(\bar{z}, z_d) = m^{d-1} - 1$, $r(\bar{z}', z'_d) = 0$. If z_d is even then $\rho_{d-1}(\bar{z}) = m^{d-1} - 1$ and $z'_d = z_d + 1$ is odd so $\rho_{d-1}(\bar{z}') = m^{d-1} - 1$. Since inductively ρ_{d-1} is a bijection, we must have $\bar{z} = \bar{z}'$. Similarly, if z_d is odd then $\rho_{d-1}(\bar{z}) = 0$ and z'_d is even so $\rho_{d-1}(\bar{z}') = 0$ so that $\bar{z} = \bar{z}'$ by inductive bijectivity. Thus z, z' differ only in the d -th coordinate, by exactly 1. Then by induction, ρ_d satisfies the adjacency property $|z(t+1) - z(t)|_1 = 1$. \square

Remark A.6 (Convergence). Since each iteration strictly decreases F whenever the partition changes and the set of equal-size partitions of S'_{nl} into K groups is finite, the per-stratum sequence $(\mathcal{G}_{nl}^{(t)})_{t \geq 0}$ stabilizes in finitely many iterations at a partition $\mathcal{G}_{nl}^{(\infty)}$. Since this holds for each $l = 1, \dots, L_n$, the full partition $\mathcal{G}_n^{(t)} = \cup_{l=1}^{L_n} \mathcal{G}_{nl}^{(t)} \cup \mathcal{R}_n$ likewise converges in finitely many steps to $\mathcal{G}_n^{(\infty)} = \cup_{l=1}^{L_n} \mathcal{G}_{nl}^{(\infty)} \cup \mathcal{R}_n$.

Lemma A.7 (Remainder). *The remainder groups $\mathcal{R}_n = \{r_l : l = 1, \dots, L_n\}$, each of size $|r_l| < k_l \leq \bar{k}_n$, satisfy the deterministic bound $F(\mathcal{R}_n) \leq n^{-1} M_n^2 d \bar{k}_n L_n$. If in addition $\bar{k}_n L_n / n = o(1)$, then*

$$F(\mathcal{R}_n) = O(M_n^2 (n / \bar{k}_n L_n)^{-2/(d+1)}).$$

Proof. For each $l, i \in r_l$, and $j \in [d]$, writing $\psi_{i,n}^{(j)} - \bar{\psi}_{r_l}^{(j)} = |r_l|^{-1} \sum_{i' \in r_l} (\psi_{i,n}^{(j)} - \psi_{i',n}^{(j)})$ and applying the triangle inequality gives $|\psi_{i,n}^{(j)} - \bar{\psi}_{r_l}^{(j)}| \leq \max_{i' \in r_l} \psi_{i',n}^{(j)} - \min_{i' \in r_l} \psi_{i',n}^{(j)} \leq M_n$. Squaring and summing over j gives $|\psi_{i,n} - \bar{\psi}_{r_l}|_2^2 \leq \sum_{j=1}^d (\max_{i' \in r_l} \psi_{i',n}^{(j)} - \min_{i' \in r_l} \psi_{i',n}^{(j)})^2 \leq dM_n^2$, and summing over $i \in r_l$ gives $\sum_{i \in r_l} |\psi_{i,n} - \bar{\psi}_{r_l}|_2^2 \leq dM_n^2 |r_l| \leq dM_n^2 \bar{k}_n$. Sum-

ming across l gives the deterministic bound $F(\mathcal{R}_n) \leq n^{-1}M_n^2 d \bar{k}_n L_n$. For $d \geq 1$ and $x \geq 1$, $x^{-1} \leq x^{-2/(d+1)}$, so under the condition $\bar{k}_n L_n/n = o(1)$ we have $n^{-1}\bar{k}_n L_n = (n/\bar{k}_n L_n)^{-1} \leq (n/\bar{k}_n L_n)^{-2/(d+1)}$ for all n large enough, giving the second conclusion. \square

Lemma A.8 (Unit Cube Reduction). *Let $(\psi_{i,n})_{i=1}^n \subseteq \mathbb{R}^d$ be a triangular array, $M_n^2 = \max_{j \leq d} (\max_i \psi_{i,n}^{(j)} - \min_i \psi_{i,n}^{(j)})^2$, and fix any subset $\mathcal{S} \subseteq [n]$. For $j = 1, \dots, d$ and $i \in \mathcal{S}$, define rescaling within \mathcal{S} by $\phi_{i,n}^{(j)} = (\psi_{i,n}^{(j)} - \min_{i' \in \mathcal{S}} \psi_{i',n}^{(j)}) / (\max_{i' \in \mathcal{S}} \psi_{i',n}^{(j)} - \min_{i' \in \mathcal{S}} \psi_{i',n}^{(j)})$ so that $\phi_{i,n} \in [0, 1]^d$. Then for any partition \mathcal{G} of \mathcal{S} , $F(\mathcal{G}, \psi) \leq M_n^2 \cdot F(\mathcal{G}, \phi)$.*

Proof. Let $m_{n,j}(\mathcal{S}) = \max_{i \in \mathcal{S}} \psi_{i,n}^{(j)} - \min_{i \in \mathcal{S}} \psi_{i,n}^{(j)}$, so the affine relation $\psi_{i,n}^{(j)} = m_{n,j}(\mathcal{S})\phi_{i,n}^{(j)} + \min_{i' \in \mathcal{S}} \psi_{i',n}^{(j)}$ holds for $i \in \mathcal{S}$. Averaging within any $g \subseteq \mathcal{S}$ yields $\bar{\psi}_g^{(j)} = m_{n,j}(\mathcal{S})\bar{\phi}_g^{(j)} + \min_{i' \in \mathcal{S}} \psi_{i',n}^{(j)}$, so $\psi_{i,n}^{(j)} - \bar{\psi}_g^{(j)} = m_{n,j}(\mathcal{S})(\phi_{i,n}^{(j)} - \bar{\phi}_g^{(j)})$. Since $m_{n,j}(\mathcal{S}) \leq M_n$ for every j , $|\psi_{i,n} - \bar{\psi}_g|_2^2 = \sum_{j=1}^d m_{n,j}(\mathcal{S})^2 (\phi_{i,n}^{(j)} - \bar{\phi}_g^{(j)})^2 \leq M_n^2 \cdot |\phi_{i,n} - \bar{\phi}_g|_2^2$. Applying $n^{-1} \sum_{g \in \mathcal{G}} \sum_{i \in g}$ to both sides gives $F(\mathcal{G}, \psi) \leq M_n^2 \cdot F(\mathcal{G}, \phi)$. \square

Lemma A.9 (Coordinate Range). *Let $\psi_{i,n} \sim P$ iid $E_P[|\psi_{i,n}|_2^\alpha] < \infty$ for some $\alpha > 0$. Define $M_{j,n} = \max_i \psi_{i,n}^{(j)} - \min_i \psi_{i,n}^{(j)}$ and set $M_n^2 = \max_{j \leq d} M_{j,n}^2$. Then $M_n^2 = o_p(n^{2/\alpha})$.*

Proof. Note that for each j , we have $M_{j,n} \leq 2 \max_i |\psi_{i,n}^{(j)}|$, so

$$\max_j M_{j,n} \leq 2 \max_i \max_j |\psi_{i,n}^{(j)}| = 2 \max_i |\psi_{i,n}|_\infty \leq 2 \max_i |\psi_{i,n}|_2.$$

Applying Lemma D.9 with $X_i = |\psi_i|_2$ and $E[|\psi|_2^\alpha] < \infty$ yields $\max_i |\psi_{i,n}|_2 = o_p(n^{1/\alpha})$, hence $M_n^2 = \max_j M_{j,n}^2 = o_p(n^{2/\alpha})$. \square

B Additional Results and Implementation Details

B.1 DML Adjustment

In this section we develop a doubly-robust estimator that attains the asymptotic variance V of Theorem 4.2 through ex-post adjustment for covariate imbalances at both stages. Since the estimator $\hat{\theta}$ attains the same variance under fine stratification without any need for ex-post adjustment, this highlights how finely stratified designs nonparametrically control covariate imbalances at both design stages.

To state the result, consider regression estimators $\hat{m}_d(\psi)$ for the conditional means $m_d(\psi) = E[Y(d)|\psi]$, $d \in \{0, 1\}$. We construct the estimator using *cross-fitting*, as in Chernozhukov et al. (2017). Fix a number of folds $K \geq 2$, and let $[n] = I_1 \cup \dots \cup I_K$ be a partition into folds of size $|I_k| \asymp n$, drawn at random independently of the data. For each fold k , let $\hat{m}_d^{(-k)}$ denote a regression estimator for m_d computed using only the units outside fold k , and write $k(i)$ for the fold containing unit i . The cross-fit doubly-augmented IPW (2-AIPW) estimator is

$$\begin{aligned} \hat{\theta}_{adj} &= E_n \left[\hat{m}_1^{(-k(i))}(\psi_i) - \hat{m}_0^{(-k(i))}(\psi_i) \right] \\ &+ E_n \left[\frac{T_i D_i (Y_i - \hat{m}_1^{(-k(i))}(\psi_i))}{q(\psi_i) p(\psi_i)} - \frac{T_i (1 - D_i) (Y_i - \hat{m}_0^{(-k(i))}(\psi_i))}{q(\psi_i) (1 - p(\psi_i))} \right]. \end{aligned} \quad (\text{B.1})$$

The estimator $\hat{\theta}_{adj}$ adjusts for covariate imbalances due to both sampling and assignment, and if $q = 1$ it reduces to the familiar cross-fit AIPW estimator for the ATE.

Assumption B.1 (Double Adjustment). *Moments $E[Y(d)^2] < \infty$ hold for $d \in \{0, 1\}$, and the propensities $q(\psi) \in (\delta, 1]$ and $p(\psi) \in (\delta, 1 - \delta)$ for some $\delta > 0$. Sampling and assignment are conditionally independent $T_i \perp\!\!\!\perp D_i | \psi_i$. Suppose also that regressions $|\hat{m}_d^{(-k)} - m_d|_{2,\psi} = o_p(1)$ for $d \in \{0, 1\}$ and each fold k .*

Theorem B.2 (Regression Equivalence). *Suppose Assumption B.1 holds. Let the design be $T_i \stackrel{\text{iid}}{\sim} \text{Bernoulli}(q(\psi_i))$ and $D_i \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p(\psi_i))$. Then $\sqrt{n_T}(\hat{\theta}_{adj} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$, with the variance V the same as under fine stratification in Theorem 4.2.*

B.2 Heteroskedasticity Function Estimation

The theory in Section 5 requires heteroskedasticity function estimates $\hat{\sigma}_d^2(\psi)$ for $d \in \{0, 1\}$, as well as the ex-ante variance estimate $\hat{\sigma}_s^2(\psi)$ from Equation (5.2) used in the

optimal sampling formula in Equation (5.5). Our simulations and empirical application use a modification of the method in Fan and Yao (1998). In a regression model $Y = m(\psi) + \sigma^2(\psi)\epsilon$, they propose to (1) use local linear regression to estimate $m(\psi)$ and (2) use local linear regression to project estimated residuals $(Y - \widehat{m}(\psi))^2$ on ψ .

In our setting, we can form signal $S_i(1) = Y_i D_i T_i / (p_i q_i)$, noting that $E[S_i(1)|\psi_i] = E[Y_i(1)|\psi_i] = m_1(\psi_i)$. We then (1) project $S_i(1)$ on ψ_i to estimate $m_1(\psi_i)$. Next, (2) we project the IPW-weighted squared residuals $(Y_i - \widehat{m}_1(\psi_i))^2 D_i T_i / (p_i q_i)$ on ψ_i to estimate $\widehat{\sigma}_1^2(\psi_i)$, and similarly for $d = 0$.

For the ex-ante variance $\widehat{\sigma}_s^2(\psi)$ in Equation (5.2) used in the optimal sampling formula in Equation (5.5), two routes are available: (i) the plug-in $\widehat{\sigma}_s^2(\psi) = \widehat{\sigma}_1^2(\psi)/p + \widehat{\sigma}_0^2(\psi)/(1-p)$ from the per-arm estimates above, or (ii) a single regression on ψ_i of the combined IPW-weighted residual signal

$$S_i = (Y_i - \widehat{m}_1(\psi_i))^2 D_i T_i / (p_i^2 q_i) + (Y_i - \widehat{m}_0(\psi_i))^2 (1 - D_i) T_i / ((1 - p_i)^2 q_i)$$

This has the same conditional expectation as $\sigma_s^2(\psi_i)$ in Equation (5.2) and avoids combining per-arm fits with different smoothing. This is analogous to the DR-learner of Kennedy, 2023 in CATE estimation, where a single influence-function signal is regressed on covariates rather than separately combining per-arm fits. Our empirical implementation uses route (ii), which we found to be more numerically stable.

We tested linear ridge regression, RBF-kernel ridge regression, and random forests for each regression step, with hyperparameters chosen by cross-validation in all cases. Kernel ridge estimated $\sigma_d^2(\psi)$ the most precisely in dimensions $\dim(\psi) = 1, 2$, while forests were superior in higher dimensions. Our simulation and empirical results are presented using random forest regression.

B.3 Imputation of Potential Outcomes and Simulation DGP

For each paper and arm $d \in \{0, 1\}$, we fit $\widehat{m}_d(\psi) = \widehat{E}[Y|\psi, D = d]$ on the observed-arm units and estimate $\widehat{\sigma}_d^2(\psi)$ from cross-fitted residuals: split arm- d units into $K = 5$ folds, fit $\widehat{m}_d^{(-k)}$ on $K - 1$ folds, form the held-out residuals $r_i = Y_i - \widehat{m}_d^{(-k)}(\psi_i)$ for i in the held-out fold, then regress r_i^2 on ψ_i to obtain $\widehat{\sigma}_d^2(\psi)$. The fitted \widehat{m}_d and $\widehat{\sigma}_d^2$ define the population we resample from. In each Monte Carlo replication, for each sampled unit i we draw $\widetilde{Y}_i(d) = \widehat{m}_d(\psi_i) + \widehat{\sigma}_d^2(\psi_i)^{1/2} \epsilon_i^d$ with $\epsilon_i^d \sim \mathcal{N}(0, 1)$ independently across replications, units, and arms. This gives $\text{Var}(\widetilde{Y}(d)|\psi) = \widehat{\sigma}_d^2(\psi)$ in the simulated population. For the

two binary-outcome papers we instead draw $\tilde{Y}_i(d) \sim \text{Bernoulli}(\hat{m}_d(\psi_i))$.

To sample from the imputed DGP, we draw a unit with replacement from $(\psi_i)_{i=1}^{N_0}$ and add independent mean-zero Gaussian noise to its continuous coordinates, with per-coordinate standard deviation equal to the average nearest-neighbor distance of that coordinate in the panel. Discrete coordinates are left unperturbed. Equivalently, ψ is drawn from the empirical measure convolved with a product kernel, so ψ is continuously distributed and the probability that two sampled units share the same covariate value is zero. This avoids the artificial perfect matches that arise when resampling with replacement from a finite panel of continuous covariates.

The regression family \mathcal{R} used for both \hat{m}_d and $\hat{\sigma}_d^2$ is selected per paper by K -fold cross-validated R^2 on the observed arms, averaged across arms; we pick \mathcal{R} from OLS, ridge, lasso with degree-2 interactions, gradient-boosted trees, and random forest. Hyperparameters are chosen by inner cross-validation where applicable. The covariates are the same ψ used for stratification in the design step. Table 1 reports $\dim(\psi)$, Appendix D.3 lists ψ paper by paper.

Some previous work, such as Bai (2022) and earlier versions of this paper, instead uses a matching-based imputation $\hat{Y}_i(d) = Y_{j(i)}(d)$ with $j(i) = \operatorname{argmin}_{j:D_j=d} |\psi_i - \psi_j|_2$. This can artificially inflate efficiency since nearest-neighbor pairs of units end up with identical imputed potential outcomes, making the within-stratum variance exactly zero with positive probability over the sampling distribution. We thank an anonymous referee for pointing this out.

C Proofs

C.1 Notation

We denote $[n] = \{1, \dots, n\}$ and use $a \wedge b$ for the minimum and $\mathbb{1}(\cdot)$ for the indicator. We write $E_n[a_i] = n^{-1} \sum_{i=1}^n a_i$. For $v \in \mathbb{R}^{\dim(\psi)}$ with $\dim(\psi) = \dim(\psi)$, $|v|_2$ denotes the Euclidean norm and $|v|_\infty$ the sup norm. For $f : \mathbb{R}^{\dim(\psi)} \rightarrow \mathbb{R}$, $|f|_{\text{lip}}$ is the Lipschitz seminorm and $\|f\|_{\psi,2}^2 = E[f(\psi)^2]$ the $L_2(\psi)$ norm. We write $X_n \xrightarrow{p} X$ for convergence in probability, $X_n \Rightarrow X$ for weak convergence, and use the usual stochastic orders $O_p(\cdot)$ and $o_p(\cdot)$. Notation $a_n \asymp b_n$ means a_n/b_n and b_n/a_n are both bounded and $a_n \lesssim b_n$ means $a_n \leq Cb_n$ for some constant C . Independence is denoted $\perp\!\!\!\perp$ and $E[\cdot|\mathcal{F}]$ is conditional expectation given a σ -algebra \mathcal{F} .

C.2 Asymptotics

This subsection develops a self-contained proof of the extension results in Section 4.1, from which the main theorem (Theorem 4.2) of Section 4 is derived. We allow different sampling and assignment stratification variables ψ_1 and ψ_2 and sampling and assignment propensities $q(\psi_1)$ and $p(\psi_2)$ taking finitely many rational levels. This subsumes the results presented in Theorem 4.5 and Corollary 4.6, which assume both constant.

First, we collect some notation.

- (1) *Filtrations.* At the assignment stage, let $\mathcal{F}_n^D = \sigma(\psi_{2,1:n}, \pi_n, T_{1:n})$ denote the information used to construct the assignment partition and $\mathcal{H}_n^D = \sigma(W_{1:n}, \pi_n, T_{1:n})$ the corresponding filtration carrying the full data $W_i = (Y_i(0), Y_i(1), \psi_{2,i})$ (recall $\psi_{1,i} = f(\psi_{2,i})$ by Assumption C.2(i)). The analogs for the sampling design are $\mathcal{F}_n^T = \sigma(\psi_{1,1:n}, \pi_n)$ and $\mathcal{H}_n^T = \sigma(W_{1:n}, \pi_n)$. Note the key relations $\mathcal{F}_n^T \subseteq \mathcal{H}_n^T$ and $\mathcal{F}_n^D \subseteq \mathcal{H}_n^D$. In the technical lemma section, we sometimes use a generic σ -algebra \mathcal{A}_n , which we set to be either $\mathcal{A}_n = \mathcal{F}_n^D$ or $\mathcal{A}_n = \mathcal{H}_n^D$ depending on the required context.
- (2) *Randomness.* Recall π_n is independent randomness used to break ties during matching. In what follows, ξ_n denotes an independent auxiliary data set, such as a pilot or previous observational data. We use η for the complete randomizations in the design itself, see the next definition for details. When necessary, we denote η^T for the sampling complete randomizations and η^D for the assignment randomizations, obeying the key relations $\eta^D \perp\!\!\!\perp \mathcal{H}_n^D$ and $\eta^T \perp\!\!\!\perp \mathcal{H}_n^T$.

We formalize the construction of the local randomization design from Definition 2.1. We state results for the assignment design $D_{1:n} \sim \text{Loc}(\psi_2, p(\psi_2))$. Our formalization also applies to the sampling design after the substitution $D_{1:n} \rightarrow T_{1:n}$, $T_{1:n} \rightarrow 1$, $p \rightarrow q$, $\psi_2 \rightarrow \psi_1$, under which the unit index set $I = \{i : T_i = 1\}$ becomes $I = [n]$.

Definition C.1 (Design Formalization). Let $D_{1:n} \sim \text{Loc}(\psi_2, p(\psi_2))$ with associated \mathcal{F}_n^D -measurable partition \mathcal{G}_n of index set $I = \{i : T_i = 1\} \subseteq [n]$. Recall $p(\psi_2) \in \{a_l/k_l : l \in [L]\}$, with k_l the interior group size at level l . Define a jointly independent array representing the design randomizations $(\eta_{l,j})_{l \in [L], j \in [n+1]}$ with $\eta \perp\!\!\!\perp (W_{1:n}, \pi_n, T_{1:n})$. Let $\eta_{l,j} \sim \text{CR}(a_l/k_l)$ for each $l \in [L]$ and $j \in [n]$ and $\eta_{l,n+1}$ be a k_l vector with iid Bernoulli(a_l/k_l) components. In particular, $n+1$ is a loose, fixed upper bound on the number of groups required in stratum $l \in [L]$. We define a function $s(\cdot)$ that specifies which randomization $\eta_{s(g)}$ is assigned to group g . In particular, let $s : \mathcal{G}_n \rightarrow [L] \times [n+1]$

be an injection sending each level- l interior group to some (l, j) with $j \in [n]$ and the level- l remainder group to $(l, n + 1)$. Note this function will not be surjective.

We can take $s(\cdot)$ to be \mathcal{F}_n^D -measurable since only the information in \mathcal{F}_n^D is used to construct the interior (full size k_l) and remainder groups $g \in \mathcal{G}_n$ and evaluate the group propensities $p(\psi_{2,i})$ for $i \in g$. For interior groups, we simply set $(D_i)_{i \in g} = \eta_{s(g)}$. More generally, let $(D_i)_{i \in g}$ be the first $|g|$ components of $\eta_{s(g)}$, accommodating remainder groups. Formally, if $f(g, x) = x_{1:|g|}$ is truncation, we can write for $g \in \mathcal{G}_n$

$$(D_i)_{i \in g} = f(g, \eta_{s(g)}). \quad (\text{C.1})$$

Strategy. We decompose $\widehat{\theta}$ -ATE into three terms by projection onto an increasing chain of conditioning σ -algebras: an assignment term, a sampling term, and a superpopulation term. We establish a CLT for each, then combine them using a characteristic function argument. Define $\tau_i = Y_i(1) - Y_i(0)$, $\bar{Y}_i = (1 - p(\psi_{2,i}))Y_i(1) + p(\psi_{2,i})Y_i(0)$, and $H_i = (D_i - p(\psi_{2,i}))/[p(\psi_{2,i})(1 - p(\psi_{2,i}))]$. Note the algebraic identity $H_i Y_i = \tau_i + H_i \bar{Y}_i$, which holds pointwise. Then the estimator can be expanded as

$$\begin{aligned} \widehat{\theta} - \text{ATE} &= E_n [T_i H_i Y_i / q(\psi_{1,i})] - \text{ATE} = E_n [T_i (\tau_i + H_i \bar{Y}_i) / q(\psi_{1,i})] - \text{ATE} \\ &= E_n [\tau_i - \text{ATE}] + E_n [\tau_i (T_i - q(\psi_{1,i})) / q(\psi_{1,i})] + E_n [T_i H_i \bar{Y}_i / q(\psi_{1,i})] \\ &\equiv A_n + B_n + C_n. \end{aligned}$$

Our approach is to prove a marginal CLT for $\sqrt{n}A_n$ and conditional CLTs for $\sqrt{n}B_n | \mathcal{H}_n^T$ and $\sqrt{n}C_n | \mathcal{H}_n^D$ given $\mathcal{H}_n^T = \sigma(W_{1:n}, \pi_n)$ and $\mathcal{H}_n^D = \sigma(W_{1:n}, \pi_n, T_{1:n})$.

Assumption C.2. *The following hold:*

- (i) $T_{1:n} \sim \text{Loc}(\psi_1, \widehat{q}_n(\psi_1))$ and $D_{1:n} \sim \text{Loc}(\psi_2, \widehat{p}_n(\psi_2))$, where \widehat{q}_n is $\sigma(\xi_n, \psi_{1,1:n})$ -measurable and \widehat{p}_n is $\sigma(\xi_n, \psi_{2,1:n})$ -measurable, with $\xi_n \perp\!\!\!\perp (W_{1:n}, \pi_n, \eta^D, \eta^T)$ for the assignment and sampling randomness η^D, η^T of Definition C.1. Also, $\psi_1 = f(\psi_2)$ for some measurable function $f(\cdot)$.
- (ii) Propensities $\widehat{q}_n(\psi_1) \in (\delta, 1]$ and $\widehat{p}_n(\psi_2) \in (\delta, 1 - \delta)$, and each takes at most L_n rational levels with denominators at most \bar{k}_n , for deterministic sequences \bar{k}_n, L_n with $\bar{k}_n L_n = o(n)$.
- (iii) Convergence $E_n[(\widehat{q}_n(\psi_{1,i}) - q(\psi_{1,i}))^2] \xrightarrow{P} 0$ and $E_n[(\widehat{p}_n(\psi_{2,i}) - p(\psi_{2,i}))^2] \xrightarrow{P} 0$ for fixed $q(\psi_1) \in (\delta, 1]$ and $p(\psi_2) \in (\delta, 1 - \delta)$.

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

Assumption C.2(i) requires $T_{1:n}$ and $D_{1:n}$ to be locally randomized designs. By Definition 2.1, this requires the existence of group partitions \mathcal{G}_n^T and \mathcal{G}_n^D that satisfy the tight matching conditions $F(\mathcal{G}_n^T) = o_p(1)$ and $F(\mathcal{G}_n^D) = o_p(1)$.

We can constructively show the existence of such partitions using the algorithms developed in Section 3. In particular, at the sampling stage we apply this algorithm to the set of units $S_n = [n]$ with covariates $(\psi_{1,i})_{i=1}^n \subseteq \mathbb{R}^{d_1}$ and propensity $\hat{q}_n(\cdot)$, producing \mathcal{G}_n^T with per-stratum grid sizes $m_l \asymp (n_l/k_l)^{1/(d_1+1)}$.

At the assignment stage, we apply the algorithm to $S_n = \{i : T_i = 1\}$ with covariates $(\psi_{2,i})_{i \in S_n} \subseteq \mathbb{R}^{d_2}$ and propensity $\hat{p}_n(\cdot)$, producing \mathcal{G}_n^D with per-stratum grid sizes $m_l \asymp (n_l/k_l)^{1/(d_2+1)}$. Here $d_1 = \dim(\psi_1)$ and $d_2 = \dim(\psi_2)$. The next lemma gives conditions sufficient for tight matching under our algorithm.

Lemma C.3 (Tight Matching). *Suppose $E[|\psi_j|_2^{\alpha_j}] < \infty$ for some $\alpha_j > \dim(\psi_j) + 1$, $j = 1, 2$. Suppose also $\bar{k}_n L_n = o(n^{1 - (\dim(\psi_j) + 1)/\alpha_j})$ for $j = 1, 2$. Then*

$$F(\mathcal{G}_n^T) = o_p(1), \quad F(\mathcal{G}_n^D) = o_p(1). \quad (\text{C.2})$$

Proof. First consider the sampling partition. Apply Theorem A.3 item (2) to $S_n = [n]$ with $\psi = \psi_1$, $d = d_1$, $\alpha = \alpha_1$. Assumption A.1(a) holds WLOG and Assumption A.1(b) $\bar{k}_n L_n = o(n)$ is implied by the rate hypothesis. Item (2) of the theorem gives $F(\mathcal{G}_n^T) = o_p(n^{2/\alpha_1} (n/\bar{k}_n L_n)^{-2/(d_1+1)})$. This is $o_p(1)$ by the rate hypothesis, hence $F(\mathcal{G}_n^T) = o_p(1)$.

The same argument applies to the assignment partition with $S_n = \{i : T_i = 1\}$, $\psi = \psi_2$, $d = d_2$, $\alpha = \alpha_2$, giving $F(\mathcal{G}_n^D) = o_p(n^{2/\alpha_2} (n/\bar{k}_n L_n)^{-2/(d_2+1)}) = o_p(1)$. \square

The next lemma is the key result showing that local randomization achieves non-parametric control over the imbalances predictable by ψ . In particular, it annihilates the imbalances in any square integrable function $b(\psi)$ of the covariates used for stratification.

Lemma C.4 (Nonparametric Balance). *Suppose Assumption C.2 holds. For any function $E[b(\psi_2)^2] < \infty$, then $E_n[T_i(D_i - \hat{p}_n(\psi_{2,i}))b(\psi_{2,i})] = o_p(n^{-1/2})$. If $E[b(\psi_1)^2] < \infty$, then $E_n[(T_i - \hat{q}_n(\psi_{1,i}))b(\psi_{1,i})] = o_p(n^{-1/2})$.*

Proof. We begin with the first claim. By Definition 2.1, the assignment partition satisfies $F(\mathcal{G}_n) = n^{-1} \sum_g \sum_{i \in g} |\psi_{2,i} - \bar{\psi}_{2,g}|_2^2 = o_p(1)$, where $\bar{\psi}_{2,g} = |g|^{-1} \sum_{i \in g} \psi_{2,i}$. By Lemma D.10 there is a deterministic $\mu_n \rightarrow \infty$ with $\mu_n F(\mathcal{G}_n) = o_p(1)$. Since $\bar{k}_n L_n = o(n)$

by Assumption C.2, the deterministic sequence $n^{-1}\bar{k}_n L_n \rightarrow 0$, so Lemma D.10 also provides a deterministic $\mu'_n \rightarrow \infty$ with $\mu'_n n^{-1}\bar{k}_n L_n \rightarrow 0$. Set $c_n = (\mu_n \wedge \mu'_n)^{1/2}$, so $c_n \rightarrow \infty$, $c_n^2 F(\mathcal{G}_n) = o_p(1)$, and $c_n^2 n^{-1}\bar{k}_n L_n = o(1)$.

By hypothesis $b(\psi_2) \in L_2(\psi_2)$, and we approximate it by Lipschitz functions. Define $\mathcal{L}_n = \{b'(\psi_2) \in L_2(\psi_2) : |b'|_{lip} \vee |b'|_\infty \leq c_n\}$ and let $b_n \in \mathcal{L}_n$ satisfy $|b_n - b|_{2,\psi_2}^2 \leq 2 \inf_{b' \in \mathcal{L}_n} |b' - b|_{2,\psi_2}^2$. We claim $|b_n - b|_{2,\psi_2}^2 \rightarrow 0$. Let $\epsilon > 0$. By Lemma D.6 there is a function h with $|h|_{lip} \vee |h|_\infty < \infty$ and $|h - b|_{2,\psi_2}^2 < \epsilon$. Since $c_n \rightarrow \infty$, $h \in \mathcal{L}_n$ for all n large enough, so $|b_n - b|_{2,\psi_2}^2 \leq 2|h - b|_{2,\psi_2}^2 < 2\epsilon$ for such n . Since ϵ was arbitrary, this shows the claim. Now expand

$$\begin{aligned} E_n[T_i(D_i - \hat{p}_n(\psi_{2,i}))b(\psi_{2,i})] &= E_n[T_i(D_i - \hat{p}_n(\psi_{2,i}))(b - b_n)(\psi_{2,i})] \\ &\quad + E_n[T_i(D_i - \hat{p}_n(\psi_{2,i}))b_n(\psi_{2,i})] \equiv A_n + B_n. \end{aligned}$$

We invoke Lemmas D.13, D.14, and D.15 with the generic \mathcal{A}_n taken to be \mathcal{F}_n^D . The lemma hypotheses $\mathcal{F}_n^D \subseteq \mathcal{A}_n$ and $\mathcal{A}_n \perp\!\!\!\perp \eta$ hold trivially and by Definition C.1. By Assumption C.2(i), $\hat{p}_n(\cdot)$ is $\sigma(\xi_n, \psi_{2,1:n})$ -measurable with $\xi_n \perp\!\!\!\perp (W_{1:n}, \pi_n, \eta^T, \eta^D)$. We subsume ξ_n in π_n , so that \hat{p}_n is $\sigma(\psi_{2,1:n}, \pi_n) \subseteq \mathcal{F}_n^D$ -measurable. Hence $\hat{p}_n(\psi_{2,i})_{1:n}$, \mathcal{G}_n , $b(\psi_{2,i})_{1:n}$, and $b_n(\psi_{2,i})_{1:n}$ are \mathcal{F}_n^D -measurable.

Then by Lemma D.13, $E[A_n | \mathcal{F}_n^D] = 0$. By the variance bound in the proof of Lemma D.14, also $\text{Var}(\sqrt{n}A_n | \mathcal{F}_n^D) \leq 2E_n[(b - b_n)^2(\psi_{2,i})]$. Since $E[E_n[(b - b_n)^2(\psi_{2,i})]] = |b - b_n|_{2,\psi_2}^2 = o(1)$ by work above, conditional Markov (Lemma D.1) gives $\text{Var}(\sqrt{n}A_n | \mathcal{F}_n^D) = o_p(1)$, and then $A_n = o_p(n^{-1/2})$, again by conditional Markov.

Next consider B_n . By Lemma D.13, $E[B_n | \mathcal{F}_n^D] = 0$. By Lemma D.15,

$$\begin{aligned} \text{Var}(\sqrt{n}B_n | \mathcal{F}_n^D) &\leq n^{-1} \sum_g |g|^{-1} \sum_{i,j \in g} (b_n(\psi_{2,i}) - b_n(\psi_{2,j}))^2 + n^{-1}\bar{k}_n L_n \max_{i=1}^n b_n(\psi_{2,i})^2 \\ &\leq c_n^2 n^{-1} \sum_g |g|^{-1} \sum_{i,j \in g} |\psi_{2,i} - \psi_{2,j}|_2^2 + n^{-1}\bar{k}_n L_n c_n^2 = 2c_n^2 F(\mathcal{G}_n) + n^{-1}\bar{k}_n L_n c_n^2. \end{aligned}$$

The second inequality uses $|b_n|_{lip} \vee |b_n|_\infty \leq c_n$, and the equality uses $\sum_{i,j \in g} |\psi_{2,i} - \psi_{2,j}|_2^2 = 2|g| \sum_{i \in g} |\psi_{2,i} - \bar{\psi}_{2,g}|_2^2$. The first term is $o_p(1)$ since $c_n^2 F(\mathcal{G}_n) = o_p(1)$, and the second is $o(1)$ since $c_n^2 n^{-1}\bar{k}_n L_n = o(1)$. Then $\text{Var}(\sqrt{n}B_n | \mathcal{F}_n^D) = o_p(1)$, so $B_n = o_p(n^{-1/2})$, again by conditional Markov. The second claim follows by the substitutions $D_{1:n} \rightarrow T_{1:n}$, $T_{1:n} \rightarrow 1$, $\hat{p}_n \rightarrow \hat{q}_n$, $\psi_2 \rightarrow \psi_1$. \square

We now develop the core CLT machinery of the paper. The design propensities

are allowed to be estimated from auxiliary data and the matching parameters \bar{k}_n, L_n are allowed to grow. This form is general enough to accommodate the main result in Theorem 4.2, the extension results in Section 4.1, and the optimal sampling CLT in Section 5.2.

Theorem C.5 (CLT). *Let Assumption C.2 hold, and suppose one of the following: (i) $\bar{k}_n = O(1)$ and $E[a(W)^2] < \infty$, or (ii) $E[|a(W)|^{2+\gamma}] < \infty$ and $\bar{k}_n = o(n^{\gamma/(2(\gamma+2))})$ for some $\gamma > 0$. Let $\mathcal{H}_n^D = \sigma(W_{1:n}, T_{1:n}, \pi_n)$ and $\mathcal{H}_n^T = \sigma(W_{1:n}, \pi_n)$. Then*

$$\begin{aligned} \sqrt{n}E_n[T_i(D_i - \hat{p}_n(\psi_{2,i}))a(W_i)|\mathcal{H}_n^D] &\Rightarrow \mathcal{N}(0, V_a), \quad V_a = E[q(\psi_1)p(\psi_2)(1 - p(\psi_2)) \text{Var}(a|\psi_2)], \\ \sqrt{n}E_n[(T_i - \hat{q}_n(\psi_{1,i}))a(W_i)|\mathcal{H}_n^T] &\Rightarrow \mathcal{N}(0, V_s), \quad V_s = E[q(\psi_1)(1 - q(\psi_1)) \text{Var}(a|\psi_1)]. \end{aligned}$$

Proof. Since the auxiliary randomness ξ_n is independent of the data $W_{1:n}$ and of all design randomness (π_n, η^D, η^T) in Definition C.1, adjoining ξ_n to π_n preserves every independence relation among $(W_{1:n}, \pi_n, \eta^D, \eta^T, T_{1:n})$. In particular, we have $W_{1:n} \perp\!\!\!\perp \pi_n$, $\eta^D \perp\!\!\!\perp (W_{1:n}, \pi_n, T_{1:n})$, and $\eta^T \perp\!\!\!\perp (W_{1:n}, \pi_n)$. We adopt this convention throughout, so ξ_n is a component of π_n , the propensities \hat{q}_n, \hat{p}_n are $\sigma(\psi_{2,1:n}, \pi_n) \subseteq \mathcal{F}_n^D$ -measurable (using $\psi_{1,1:n} = f(\psi_{2,1:n})$ from Assumption C.2(i)), and the σ -algebras $\mathcal{H}_n^D = \sigma(W_{1:n}, T_{1:n}, \pi_n)$, $\mathcal{F}_n^D = \sigma(\psi_{2,1:n}, \pi_n, T_{1:n})$, and $\mathcal{H}_n^T = \sigma(W_{1:n}, \pi_n)$ carry ξ_n .

Note $E[a(W)^2] < \infty$ under either hypothesis, in case (ii) by Jensen's inequality. Let $u_i = a(W_i) - E[a(W)|\psi_{2,i}]$, so $E[u_i|\psi_{2,i}] = 0$ and $E[u_i^2|\psi_{2,i}] = \text{Var}(a|\psi_{2,i})$. Under hypothesis (ii) also $E[|u_i|^{2+\gamma}] < \infty$: by the c_p inequality $|u_i|^{2+\gamma} \leq 2^{1+\gamma}(|a(W_i)|^{2+\gamma} + |E[a(W)|\psi_{2,i}]|^{2+\gamma})$, and $E[|E[a(W)|\psi_{2,i}]|^{2+\gamma}] \leq E[|a(W)|^{2+\gamma}]$ by conditional Jensen and the tower law, giving $E[|u_i|^{2+\gamma}] \leq 2^{2+\gamma}E[|a(W)|^{2+\gamma}] < \infty$. By Lemma C.4 applied to the design $D_{1:n} \sim \text{Loc}(\psi_2, \hat{p}_n(\psi_2))$, $\sqrt{n}E_n[T_i(D_i - \hat{p}_n(\psi_{2,i}))E[a|\psi_{2,i}]] = o_p(1)$, since $E[E[a|\psi_{2,i}]^2] \leq E[a^2] < \infty$ by conditional Jensen and tower law. Hence we have $\sqrt{n}E_n[T_i(D_i - \hat{p}_n(\psi_{2,i}))a(W_i)] = X_n + o_p(1)$ for $X_n \equiv \sqrt{n}E_n[T_i(D_i - \hat{p}_n(\psi_{2,i}))u_i]$.

Order the assignment groups $g(j) \in \mathcal{G}_n$, $j = 1, \dots, M_n$. Since $\hat{p}_n(\psi_2)$ takes at most L_n rational levels $\hat{p}_{n,l} = a_{n,l}/k_{n,l}$ for $l \in [L_n]$ with denominators $k_{n,l} \leq \bar{k}_n$, interior groups at level l have size $k_{n,l}$, with at most one remainder group per level of size $\leq k_{n,l} - 1$. Define the basic terms $z_{j,n} = n^{-1/2} \sum_{i \in g(j)} (D_i - \hat{p}_n(\psi_{2,i}))u_i$, noting $T_i = 1$ for $i \in g(j)$, so $X_n = \sum_{j=1}^{M_n} z_{j,n}$. Our plan is to show $X_n|\mathcal{H}_n^D \Rightarrow \mathcal{N}(0, V_a)$ by justifying the conditions of Proposition D.5, in three steps: (1) the basic terms $(z_{j,n})_{j=1}^{M_n}$ are jointly independent conditional on \mathcal{H}_n^D , (2) the variance process $\Sigma_n \equiv \sum_{j=1}^{M_n} E[z_{j,n}^2|\mathcal{H}_n^D]$ has convergence $\Sigma_n \xrightarrow{p} V_a$, and (3) the conditional Lindeberg condition $\sum_{j=1}^{M_n} E[z_{j,n}^2 \mathbb{1}(|z_{j,n}| > \epsilon)|\mathcal{H}_n^D] = o_p(1)$ holds for each $\epsilon > 0$. We begin with (1).

(1) Note $\mathcal{F}_n^D = \sigma(\psi_{2,1:n}, \pi_n, T_{1:n}) \subseteq \mathcal{H}_n^D$ and $\mathcal{H}_n^D \perp\!\!\!\perp \eta^D$ for assignment randomness η^D of Definition C.1. Observe \widehat{p}_n is \mathcal{F}_n^D -measurable since ξ_n is a component of π_n . Each $z_{j,n}$ has the form $\phi((D_i)_{i \in g(j)}, (\widehat{p}_n(\psi_{2,i}), u_i)_{1:n}, g(j))$ for the deterministic function $\phi(d, (p, u), g) = n^{-1/2} \sum_{i \in g} (d_i - p_i) u_i$, with $(\widehat{p}_n(\psi_{2,i}), u_i)_{1:n}$ being \mathcal{H}_n^D -measurable. Then by Lemma D.11, the $(z_{j,n})_{j=1}^{M_n}$ are jointly independent conditional on \mathcal{H}_n^D .

(2) By Lemma D.13, $E[D_i | \mathcal{H}_n^D] = \widehat{p}_n(\psi_{2,i})$ for $i \in \{T_i = 1\}$, so $E[z_{j,n} | \mathcal{H}_n^D] = 0$. By \mathcal{H}_n^D -measurability of u_r, u_t and $E[D_i | \mathcal{H}_n^D] = \widehat{p}_n(\psi_{2,i})$, we have

$$E[z_{j,n}^2 | \mathcal{H}_n^D] = n^{-1} \sum_{r,t \in g(j)} u_r u_t \text{Cov}(D_r, D_t | \mathcal{H}_n^D).$$

By Lemma D.13, $\text{Var}(D_i | \mathcal{H}_n^D) = \widehat{p}_n(\psi_{2,i})(1 - \widehat{p}_n(\psi_{2,i}))$ for $i \in \{T_i = 1\}$. For $r \neq t$ in an interior $g(j)$ at level l , the covariance $\text{Cov}(D_r, D_t | \mathcal{H}_n^D) = -\widehat{p}_{n,l}(1 - \widehat{p}_{n,l}) / (k_{n,l} - 1)$ with $\widehat{p}_{n,l} = \widehat{p}_n(\psi_{2,i})$ constant within the group. For the remainder group at level l , $(D_i)_{i \in g}$ is conditionally iid Bernoulli($\widehat{p}_{n,l}$) by Definition C.1, so $\text{Cov}(D_r, D_t | \mathcal{H}_n^D) = 0$. Let $l(j)$ denote the assignment propensity level of group $g(j)$. Summing over j , we obtain

$$\Sigma_n = E_n[T_i \widehat{p}_n(\psi_{2,i})(1 - \widehat{p}_n(\psi_{2,i})) u_i^2] - n^{-1} \sum_{j:\text{interior}} \frac{\widehat{p}_{n,l(j)}(1 - \widehat{p}_{n,l(j)})}{k_{n,l(j)} - 1} \sum_{r \neq t \in g(j)} u_r u_t \equiv T_{n2} + T_{n1}.$$

The diagonal term T_{n2} aggregates over all groups using $\sum_j \widehat{p}_{n,l(j)}(1 - \widehat{p}_{n,l(j)}) \sum_{i \in g(j)} u_i^2 = \sum_i T_i \widehat{p}_n(\psi_{2,i})(1 - \widehat{p}_n(\psi_{2,i})) u_i^2$, since \mathcal{G}_n partitions $\{i : T_i = 1\}$ and $\widehat{p}_n(\psi_{2,i})$ is constant within each group. The off-diagonal T_{n1} vanishes on the remainder groups by the iid Bernoulli structure noted above. We will show $T_{n2} \xrightarrow{p} V_a$. To do so, write $\rho_n(\psi_{2,i}) = \widehat{p}_n(\psi_{2,i})(1 - \widehat{p}_n(\psi_{2,i}))$ and $\rho(\psi_2) = p(\psi_2)(1 - p(\psi_2))$, and split $T_{n2} = E_n[T_i(\rho_n - \rho)(\psi_{2,i}) u_i^2] + E_n[T_i \rho(\psi_{2,i}) u_i^2] \equiv T_{n2}^{(1)} + T_{n2}^{(2)}$.

To analyze these terms, we repeatedly use the following fact. If $(b_{i,n})_{i=1}^n$ are uniformly bounded, $\sup_{i,n} |b_{i,n}| \leq M$, with $E_n[b_{i,n}^2] = o_p(1)$, and $(c_i)_{i=1}^n$ are iid with $E[|c_1|] < \infty$, then $E_n[|b_{i,n} c_i|] = o_p(1)$. Indeed, for any $K > 0$, $E_n[|b_{i,n} c_i|] \leq K E_n[|b_{i,n}|] + M E_n[|c_i| \mathbf{1}(|c_i| > K)]$. The first term is $o_p(1)$ since $E_n[|b_{i,n}|] \leq E_n[b_{i,n}^2]^{1/2}$ by Cauchy–Schwarz, the second converges in probability to $M E[|c_1| \mathbf{1}(|c_1| > K)]$ by the iid weak law, and this limit is arbitrarily small for K large.

For $T_{n2}^{(1)}$, the identity $|x(1-x) - y(1-y)| = |x-y||1-x-y| \leq |x-y|$ for $x, y \in [0, 1]$ gives $|\rho_n - \rho| \leq |\widehat{p}_n - p|$ pointwise, hence $E_n[(\rho_n - \rho)^2(\psi_{2,i})] \leq E_n[(\widehat{p}_n - p)^2(\psi_{2,i})] = o_p(1)$ by Assumption C.2. So $|T_{n2}^{(1)}| = |E_n[T_i(\rho_n - \rho)(\psi_{2,i}) u_i^2]| \leq E_n[|\rho_n - \rho|(\psi_{2,i}) u_i^2]$, using $T_i \leq 1$, which is $o_p(1)$ by the fact above with $b_{i,n} = (\rho_n - \rho)(\psi_{2,i})$, bounded by $1/4$, and

$c_i = u_i^2$, iid with $E[u_i^2] \leq \text{Var}(a) < \infty$.

For $T_{n2}^{(2)}$, let $h(W) = \rho(\psi_2)u^2$, which has $E[|\rho(\psi_2)u^2|] \leq E[u^2] < \infty$. We claim that $E_n[T_i h(W_i)] = E[q(\psi_1)h(W)] + o_p(1)$. To see this, we can decompose $E_n[T_i h_i] = E_n[(T_i - \widehat{q}_n(\psi_{1,i}))h_i] + E_n[(\widehat{q}_n - q)(\psi_{1,i})h_i] + E_n[q(\psi_{1,i})h_i]$. The first term is $o_p(1)$ by Lemma D.14 part (2) applied to the sampling design $T_{1:n} \sim \text{Loc}(\psi_1, \widehat{q}_n(\psi_1))$, with the constant sequence $h_n = h$. The second term is bounded in absolute value by $E_n[|\widehat{q}_n - q|(\psi_{1,i})|h_i|]$, which is $o_p(1)$ by the fact above with $b_{i,n} = (\widehat{q}_n - q)(\psi_{1,i})$, bounded by 1, and $c_i = h_i$, iid with $E[|h|] < \infty$. The third term equals $E[q(\psi_1)h(W)] + o_p(1)$ by the iid WLLN, using that $|qh| \leq |h|$. Then since $h = \rho(\psi_2)u^2$, we have $T_{n2}^{(2)} = E_n[T_i \rho(\psi_{2,i})u_i^2] = E[q(\psi_1)p(\psi_2)(1 - p(\psi_2))u^2] + o_p(1)$. This implies $T_{n2} \xrightarrow{P} V_a$, since by tower law and ψ_2 -measurability of $q(\psi_1)$ and since $E[u^2|\psi_2] = \text{Var}(a|\psi_2)$,

$$E[q(\psi_1)p(\psi_2)(1 - p(\psi_2))u^2] = E[q(\psi_1)p(\psi_2)(1 - p(\psi_2)) \text{Var}(a|\psi_2)] = V_a.$$

Next, we apply Lemma D.8 to show $T_{n1} = o_p(1)$, with conditioning σ -algebra \mathcal{F}_n^D . Note that \mathcal{G}_n is \mathcal{F}_n^D measurable, recalling that $\mathcal{F}_n^D = \sigma(\psi_{2,1:n}, \pi_n, T_{1:n})$ for the assignment design (Definition C.1). Define $v_j = \frac{\widehat{p}_{n,l(j)}(1 - \widehat{p}_{n,l(j)})}{k_{n,l(j)} - 1} \sum_{r \neq t \in g(j)} u_r u_t$ for interior $g(j)$ and $v_j = 0$ for the remainder, so $T_{n1} = -n^{-1} \sum_{j=1}^{M_n} v_j$.

Lemma D.8 requires joint conditional independence of $(v_j)_{j=1}^{M_n}$ given \mathcal{F}_n^D . To prove this, we apply Lemma D.16 with $h_i = \psi_{2,i}$ and $\kappa = (\pi_n, T_{1:n})$, so that in the notation of the lemma $\mathcal{F}_n^D = \sigma(\psi_{2,1:n}, \pi_n, T_{1:n}) = \sigma(h_{1:n}, \kappa)$. The conditional independence required is $\kappa \perp\!\!\!\perp W_{1:n}|h_{1:n}$, equivalent to $(\pi_n, T_{1:n}) \perp\!\!\!\perp W_{1:n}|\psi_{2,1:n}$. To see that this holds, chain two reductions. First, $W_{1:n} \perp\!\!\!\perp \pi_n$ unconditionally gives $W_{1:n} \perp\!\!\!\perp \pi_n|\psi_{2,1:n}$ via $(A, B) \perp\!\!\!\perp C \Rightarrow A \perp\!\!\!\perp C|B$ (with $A = W_{1:n}$, $B = \psi_{2,1:n} \in \sigma(W_{1:n})$, $C = \pi_n$). Second, $T_{1:n} \perp\!\!\!\perp W_{1:n} | (\psi_{2,1:n}, \pi_n)$ was established in the argument above. Composing, $W_{1:n} | (\psi_{2,1:n}, \pi_n, T_{1:n}) \stackrel{d}{=} W_{1:n} | (\psi_{2,1:n}, \pi_n) \stackrel{d}{=} W_{1:n} | \psi_{2,1:n}$, which is the required $(\pi_n, T_{1:n}) \perp\!\!\!\perp W_{1:n} | \psi_{2,1:n}$. Hence Lemma D.16 gives $(v_j)_{j=1}^{M_n}$ jointly conditionally independent given \mathcal{F}_n^D . Also, note the partition \mathcal{G}_n is \mathcal{F}_n^D -measurable by construction.

Lemma D.8 also requires the conditional mean-zero property $E[v_j | \mathcal{F}_n^D] = 0$ for each $j = 1, \dots, M_n$. The prefactor $\widehat{p}_{n,l(j)}(1 - \widehat{p}_{n,l(j)}) / (k_{n,l(j)} - 1)$ is \mathcal{F}_n^D -measurable and bounded by $1/4$ since $\widehat{p}_{n,l}(1 - \widehat{p}_{n,l}) \leq 1/4$ and $k_{n,l} - 1 \geq 1$. We claim that $E[u_r u_t | \mathcal{F}_n^D] = 0$ for

$r \neq t$, reducing the conditioning as follows:

$$\begin{aligned} E[u_r u_t | \mathcal{F}_n^D] &= E[u_r u_t | \psi_{2,1:n}, \pi_n, T_{1:n}] = E[u_r u_t | \psi_{2,1:n}, \pi_n] = E[u_r u_t | \psi_{2,1:n}] \\ &= E[u_r u_t | \psi_{2,r}, \psi_{2,t}]. \end{aligned}$$

The first equality is by definition of \mathcal{F}_n^D . For the second equality, by Definition C.1 $T_{1:n}$ is a function of $(\psi_{1,1:n}, \pi_n, \eta^T)$ for sampling randomness $\eta^T \perp\!\!\!\perp (W_{1:n}, \pi_n)$. Since $\psi_{1,1:n} = f(\psi_{2,1:n})$ is $\sigma(\psi_{2,1:n})$ -measurable, $T_{1:n} \perp\!\!\!\perp W_{1:n} | (\psi_{2,1:n}, \pi_n)$. In particular, $(u_r, u_t) \perp\!\!\!\perp T_{1:n} | (\psi_{2,1:n}, \pi_n)$. Recall the basic fact that $(A, B) \perp\!\!\!\perp C \implies A \perp\!\!\!\perp C | B$. The third equality follows from this fact noting that $W_{1:n} \perp\!\!\!\perp \pi_n$, hence $(u_r, u_t) \perp\!\!\!\perp \pi_n | \psi_{2,1:n}$. The fourth equality uses iid sampling: $(u_r, u_t, \psi_{2,r}, \psi_{2,t}) \perp\!\!\!\perp \psi_{2,-(r,t)}$, hence $(u_r, u_t) \perp\!\!\!\perp \psi_{2,-(r,t)} | (\psi_{2,r}, \psi_{2,t})$, again using the basic fact. Then by tower law, we can further reduce

$$E[u_r u_t | \psi_{2,r}, \psi_{2,t}] = E[u_r E[u_t | u_r, \psi_{2,r}, \psi_{2,t}] | \psi_{2,r}, \psi_{2,t}] = E[u_r E[u_t | \psi_{2,t}] | \psi_{2,r}, \psi_{2,t}] = 0.$$

The second equality uses $u_t \perp\!\!\!\perp (u_r, \psi_{2,r}) | \psi_{2,t}$ from iid sampling and the basic fact above. The third uses $E[u_t | \psi_{2,t}] = 0$ by definition. Then we have shown $E[v_j | \mathcal{F}_n^D] = 0$.

Finally, Lemma D.8 requires $n^{-1} \sum_{j=1}^{M_n} E[|v_j| \mathbf{1}(|v_j| > c_n) | \mathcal{F}_n^D] = o_p(1)$ for some sequence c_n with $c_n = \omega(1)$ and $c_n = o(n^{1/2})$. Under assumption (i), set $c_n = n^{1/4}$. Under assumption (ii), set $c_n = a_n \bar{k}_n^{-1+2/\gamma}$, where $a_n \rightarrow \infty$ is a deterministic sequence chosen such that $c_n = o(n^{1/2})$. Such an a_n exists by Lemma D.10 applied to the deterministic null sequence $\bar{k}_n^{-1+2/\gamma} n^{-1/2}$, which is $o(1)$ since the case (ii) hypothesis $\bar{k}_n = o(n^{\gamma/(2(\gamma+2))})$ is equivalent to $\bar{k}_n^{-1+2/\gamma} = o(n^{1/2})$. Then under either assumption, we have a sequence $c_n = \omega(1)$ and $c_n = o(n^{1/2})$.

On interior group $g(j)$ at level l , $\sum_{r \neq t \in g(j)} u_r u_t = (\sum_{r \in g(j)} u_r)^2 - \sum_{r \in g(j)} u_r^2$ and Cauchy-Schwarz give $|v_j| \leq \hat{p}_{n,l}(1 - \hat{p}_{n,l}) \sum_{r \in g(j)} u_r^2 \leq (1/4) \sum_{r \in g(j)} u_r^2$, using the bound $\max_p p(1-p) \leq 1/4$. Recall interior groups have size $|g(j)| \leq \bar{k}_n$. Note also the fact that for any $(a_k)_{k=1}^m$ with $a_k \geq 0$ that $\sum_k a_k \mathbf{1}(\sum_k a_k > c) \leq m \sum_k a_k \mathbf{1}(a_k > c/m)$. Then

$$\begin{aligned} E[|v_j| \mathbf{1}(|v_j| > c_n) | \mathcal{F}_n^D] &\leq (1/4) E\left[\sum_{r \in g(j)} u_r^2 \mathbf{1}\left(\sum_{r \in g(j)} u_r^2 > 4c_n\right) | \mathcal{F}_n^D \right] \\ &\leq (\bar{k}_n/4) \sum_{r \in g(j)} E[u_r^2 \mathbf{1}(u_r^2 > 4c_n/\bar{k}_n) | \mathcal{F}_n^D]. \end{aligned}$$

Summing over j , since interior groups partition a subset of $\{i : T_i = 1\}$ this is

$$\frac{1}{n} \sum_j E[|v_j| \mathbf{1}(|v_j| > c_n) \mid \mathcal{F}_n^D] \leq \frac{\bar{k}_n}{4n} \sum_{i=1}^n T_i E[u_i^2 \mathbf{1}(u_i^2 > 4c_n/\bar{k}_n) \mid \mathcal{F}_n^D].$$

Taking expectation and using $T_i \leq 1$ together with iid sampling and tower law, the right side has expectation bounded by $(\bar{k}_n/4)E[u_i^2 \mathbf{1}(u_i^2 > 4c_n/\bar{k}_n)]$. In case (i), $\bar{k}_n \leq C$ for a constant C , so this is at most $(C/4)E[u_i^2 \mathbf{1}(u_i^2 > 4c_n/C)] \rightarrow 0$ by dominated convergence, since $4c_n/C \rightarrow \infty$ and $E[u_i^2] \leq E[a^2] < \infty$. In case (ii), since $u^2 \mathbf{1}(u^2 > t) \leq |u|^{2+\gamma} t^{-\gamma/2}$ for $t > 0$, the bound is at most $\frac{1}{4^{1+\gamma/2}} E[|u_i|^{2+\gamma}] \bar{k}_n^{-1+\gamma/2} c_n^{-\gamma/2} = \frac{1}{4^{1+\gamma/2}} E[|u_i|^{2+\gamma}] a_n^{-\gamma/2} \rightarrow 0$, using $c_n = a_n \bar{k}_n^{-1+2/\gamma}$. Then by Markov inequality, $n^{-1} \sum_j E[|v_j| \mathbf{1}(|v_j| > c_n) \mid \mathcal{F}_n^D] = o_p(1)$. Lemma D.8 with this c_n , which satisfies $c_n = \omega(1)$ and $c_n = o(n^{1/2})$, gives $T_{n1} = o_p(1)$. Then we have shown the claim $\Sigma_n = T_{n1} + T_{n2} \xrightarrow{P} V_a$.

(3) Finally, we verify the conditional Lindeberg condition $\sum_{j=1}^{M_n} E[z_{j,n}^2 \mathbf{1}(|z_{j,n}| > \epsilon) \mid \mathcal{H}_n^D] = o_p(1)$ holds for each $\epsilon > 0$. Note that the bound $|D_i - \hat{p}_n(\psi_{2,i})| \leq 1$ and Cauchy-Schwarz imply $z_{j,n}^2 \leq \bar{k}_n n^{-1} \sum_{r \in g(j)} u_r^2$, using $|g(j)| \leq \bar{k}_n$. Hence

$$z_{j,n}^2 \mathbf{1}(z_{j,n}^2 > \epsilon^2) \leq \bar{k}_n n^{-1} \sum_{r \in g(j)} u_r^2 \mathbf{1}\left(\sum_{r \in g(j)} u_r^2 > n\epsilon^2/\bar{k}_n\right) \leq \bar{k}_n^2 n^{-1} \sum_{r \in g(j)} u_r^2 \mathbf{1}(u_r^2 > n\epsilon^2/\bar{k}_n^2).$$

The second inequality applies the indicator function fact above to $\sum_{r \in g(j)} u_r^2$ as a sum of $m \leq \bar{k}_n$ non-negative terms. Summing over j , since \mathcal{G}_n partitions $\{i : T_i = 1\}$ we have

$$\sum_{j=1}^{M_n} E[z_{j,n}^2 \mathbf{1}(z_{j,n}^2 > \epsilon^2) \mid \mathcal{H}_n^D] \leq \frac{\bar{k}_n^2}{n} \sum_{i=1}^n T_i E[u_i^2 \mathbf{1}(u_i^2 > n\epsilon^2/\bar{k}_n^2) \mid \mathcal{H}_n^D].$$

Taking expectation of both sides above and using $T_i \leq 1$ together with iid sampling and tower law, the right side has expectation bounded by $\bar{k}_n^2 E[u_i^2 \mathbf{1}(u_i^2 > n\epsilon^2/\bar{k}_n^2)]$.

Under assumption (i), $\bar{k}_n \leq C$ for a constant C , so this is at most $C^2 E[u_i^2 \mathbf{1}(u_i^2 > n\epsilon^2/C^2)] \rightarrow 0$ by dominated convergence, since $n\epsilon^2/C^2 \rightarrow \infty$ and $E[u_i^2] \leq E[a^2] < \infty$. Under assumption (ii), since $u^2 \mathbf{1}(u^2 > t) \leq |u|^{2+\gamma} t^{-\gamma/2}$ for $t > 0$, this is at most $\epsilon^{-\gamma} E[|u_i|^{2+\gamma}] \bar{k}_n^{-2+\gamma} n^{-\gamma/2} \rightarrow 0$, since $\bar{k}_n^{-2+\gamma} = o(n^{\gamma/2})$ by raising the assumed Assumption (ii) rate $\bar{k}_n = o(n^{\gamma/(2(\gamma+2))})$ to the power $2+\gamma$. Then by Markov inequality, we have $\sum_{j=1}^{M_n} E[z_{j,n}^2 \mathbf{1}(|z_{j,n}| > \epsilon) \mid \mathcal{H}_n^D] = o_p(1)$, showing the conditional Lindeberg condition.

Steps (1)-(3) verify, with conditioning σ -algebra \mathcal{H}_n^D , the conditional independence, variance, and Lindeberg hypotheses of Proposition D.5 for the array $(z_{j,n})_{j=1}^{M_n}$, extended

by $z_{j,n} = 0$ for $M_n < j \leq n$. The proposition gives $E[e^{itX_n} | \mathcal{H}_n^D] = e^{-t^2 V_a / 2} + o_p(1)$, with V_a a constant, hence trivially \mathcal{H}_n^D -measurable; since ξ_n is a component of $\pi_n \subseteq \mathcal{H}_n^D$, this convergence is conditional on the auxiliary randomness ξ_n . Combined with $\sqrt{n}E_n[T_i(D_i - \hat{p}_n(\psi_{2,i}))a(W_i)] = X_n + o_p(1)$, this gives the assignment CLT.

For the sampling CLT, apply the substitution rule of Section C.2: $D_{1:n} \rightarrow T_{1:n}$, $T_{1:n} \rightarrow 1$, $\hat{p}_n \rightarrow \hat{q}_n$, $p \rightarrow q$, $\psi_2 \rightarrow \psi_1$. Under this substitution $\mathcal{H}_n^D = \sigma(W_{1:n}, T_{1:n}, \pi_n)$ becomes $\mathcal{H}_n^T = \sigma(W_{1:n}, \pi_n)$, and every step of the argument above carries through with sampling-design analogs; in particular, under $T_{1:n} \rightarrow 1$ the term T_{n2} becomes the average $E_n[\hat{q}_n(\psi_{1,i})(1 - \hat{q}_n(\psi_{1,i}))u_i^2]$, treated by the same split into $\hat{q}_n - q$, handled by the truncation fact above, and the iid weak law. This yields $\sqrt{n}E_n[(T_i - \hat{q}_n(\psi_{1,i}))a(W_i)] | \mathcal{H}_n^T \Rightarrow \mathcal{N}(0, V_s)$ with V_s obtained from V_a under the same substitution. \square

We now use Theorem C.5 together with the asymptotic independence Lemma D.7 to prove generalizations of Theorem 4.5 and Corollary 4.6 that allow both $q(\psi_1)$ and $p(\psi_2)$ to take finitely many rational levels.

Theorem C.6 (CLT, ATE). *Suppose $T_{1:n} \sim \text{Loc}(\psi_1, q(\psi_1))$ and $D_{1:n} \sim \text{Loc}(\psi_2, p(\psi_2))$, with $\psi_1 = f(\psi_2)$ for some measurable f . Then $\sqrt{n_T}(\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V_{\text{ATE}})$,*

$$V_{\text{ATE}} = E[q(\psi_1)] \left(\text{Var}(\tau) + E\left[\frac{1-q(\psi_1)}{q(\psi_1)} \text{Var}(\tau | \psi_1)\right] + E\left[\frac{\text{Var}(\bar{Y} | \psi_2)}{q(\psi_1)p(\psi_2)(1-p(\psi_2))}\right] \right). \quad (\text{C.3})$$

Proof. This design is a special case of Assumption C.2 obtained by taking $\hat{q}_n \equiv q$ and $\hat{p}_n \equiv p$, so its condition (iii) holds exactly and $\bar{k}_n, L_n = O(1)$. Condition (ii) holds with $\delta > 0$ chosen so the finitely many rational levels of q, p lie in $(\delta, 1 - \delta)$. Case (i) of Theorem C.5 then applies, which requires only $E[a(W)^2] < \infty$ of each integrand below.

Above, we noted the decomposition $\hat{\theta} - \text{ATE} = A_n + B_n + C_n$ for

$$A_n = E_n[\tau_i - \text{ATE}], \quad B_n = E_n[\tau_i(T_i - q(\psi_{1,i}))/q(\psi_{1,i})], \quad C_n = E_n[T_i H_i \bar{Y}_i / q(\psi_{1,i})].$$

Where $H_i = (D_i - p(\psi_{2,i}))/[p(\psi_{2,i})(1 - p(\psi_{2,i}))]$ and $\bar{Y}_i = (1 - p(\psi_{2,i}))Y_i(1) + p(\psi_{2,i})Y_i(0)$. Lemma D.7 requires an increasing chain of σ -algebras $\mathcal{F}_{n,0} \subseteq \mathcal{F}_{n,1} \subseteq \mathcal{F}_{n,2}$. In this context, we set $\mathcal{F}_{n,0} = \sigma(\emptyset)$, $\mathcal{F}_{n,1} = \mathcal{H}_n^T$, and $\mathcal{F}_{n,2} = \mathcal{H}_n^D$. We show conditional CLTs for $\sqrt{n}A_n, \sqrt{n}B_n, \sqrt{n}C_n$ at the three levels of the chain, then combine via the lemma.

First consider $\sqrt{n}A_n$. By Young's inequality, $\text{Var}(\tau) \leq 2E[Y(1)^2 + Y(0)^2] < \infty$, so the iid CLT gives $\sqrt{n}A_n \Rightarrow \mathcal{N}(0, V_A)$ with $V_A = \text{Var}(\tau)$. Since $\mathcal{F}_{n,0}$ is trivial, this is equivalently $\sqrt{n}A_n | \mathcal{F}_{n,0} \Rightarrow \mathcal{N}(0, V_A)$.

Next consider $\sqrt{n}B_n$. Apply the sampling CLT of Theorem C.5 with $a(W) = \tau/q(\psi_1)$, noting $E[a(W)^2] \leq \delta^{-2}E[\tau^2] \leq 2\delta^{-2}E[Y(1)^2 + Y(0)^2] < \infty$ by the propensity bound and Young's inequality. Then by that theorem $\sqrt{n}B_n|\mathcal{F}_{n,1} \Rightarrow \mathcal{N}(0, V_B)$ with

$$V_B = E\left[q(\psi_1)(1 - q(\psi_1)) \text{Var}(\tau/q(\psi_1)|\psi_1)\right] = E\left[\frac{1-q(\psi_1)}{q(\psi_1)} \text{Var}(\tau|\psi_1)\right].$$

Finally consider $\sqrt{n}C_n$. Apply the assignment part of Theorem C.5 with $a(W) = \bar{Y}_i/[q(\psi_{1,i})p(\psi_{2,i})(1 - p(\psi_{2,i}))]$. Note $q(\psi_{1,i})$ and $p(\psi_{2,i})$ are both $\psi_{2,i}$ -measurable, the latter directly and the former since $\psi_1 = f(\psi_2)$. By the propensity bounds, $|a(W)| \leq [\delta^3]^{-1}|\bar{Y}_i|$, so $E[a(W)^2] \leq \delta^{-6}E[\bar{Y}_i^2] \lesssim E[Y(1)^2 + Y(0)^2] < \infty$. Then the result guarantees $\sqrt{n}C_n|\mathcal{F}_{n,2} \Rightarrow \mathcal{N}(0, V_C)$ with

$$V_C = E\left[q(\psi_1)p(\psi_2)(1 - p(\psi_2)) \text{Var}(a|\psi_2)\right] = E\left[\frac{\text{Var}(\bar{Y}|\psi_2)}{q(\psi_1)p(\psi_2)(1-p(\psi_2))}\right].$$

By construction $\sqrt{n}A_n \in \sigma(W_{1:n}) \subseteq \mathcal{F}_{n,1}$ and $\sqrt{n}B_n \in \sigma(W_{1:n}, T_{1:n}) \subseteq \mathcal{F}_{n,2}$, and $\sqrt{n}C_n$ is measurable in $\sigma(W_{1:n}, \pi_n, T_{1:n}, D_{1:n}) \supset \mathcal{F}_{n,2}$. Lemma D.7 applied with $X_{n,k} = \sqrt{n}(A_n, B_n, C_n)$ at $k = 1, 2, 3$ gives that for each $t \in \mathbb{R}$

$$E\left[e^{it\sqrt{n}(A_n+B_n+C_n)}\right] = e^{-t^2(V_A+V_B+V_C)/2} + o_p(1).$$

Finally, $n_T/n = E_n[T_i] \xrightarrow{P} E[q(\psi_1)]$ by Lemma D.14 part (3) applied to the sampling design with $h = 1$, so $(n_T/n)^{1/2} \xrightarrow{P} E[q(\psi_1)]^{1/2}$ by continuous mapping. By Slutsky, $\sqrt{n_T}(\hat{\theta} - \text{ATE}) = (n_T/n)^{1/2}\sqrt{n}(\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, E[q(\psi_1)](V_A + V_B + V_C)) = \mathcal{N}(0, V_{\text{ATE}})$. \square

Corollary C.7 (CLT, SATE). *In the design of Theorem C.6, $\sqrt{n_T}(\hat{\theta} - \text{SATE})|\mathcal{H}_n^T \Rightarrow \mathcal{N}(0, V_{\text{SATE}})$, with*

$$V_{\text{SATE}} = E[q(\psi_1)]\left(E\left[\frac{1-q(\psi_1)}{q(\psi_1)} \text{Var}(\tau|\psi_1)\right] + E\left[\frac{\text{Var}(\bar{Y}|\psi_2)}{q(\psi_1)p(\psi_2)(1-p(\psi_2))}\right]\right). \quad (\text{C.4})$$

In particular, $\sqrt{n_T}(\hat{\theta} - \text{SATE}) \Rightarrow \mathcal{N}(0, V_{\text{SATE}})$ marginally.

Proof. Since $\text{SATE} = E_n[\tau_i] = A_n + \text{ATE}$, $\hat{\theta} - \text{SATE} = (\hat{\theta} - \text{ATE}) - A_n = B_n + C_n$. Using the chain $\mathcal{F}_{n,0} \subseteq \mathcal{F}_{n,1} \subseteq \mathcal{F}_{n,2}$ from the proof of Theorem C.6, that proof establishes the conditional CLTs $\sqrt{n}B_n|\mathcal{F}_{n,1} \Rightarrow \mathcal{N}(0, V_B)$ and $\sqrt{n}C_n|\mathcal{F}_{n,2} \Rightarrow \mathcal{N}(0, V_C)$. Lemma D.7

applied with $X_{n,k} = \sqrt{n}(B_n, C_n)$ at $k = 1, 2$ and base \mathcal{H}_n^T gives

$$\sqrt{n}(B_n + C_n) \mid \mathcal{H}_n^T \Rightarrow \mathcal{N}(0, V_B + V_C).$$

By Lemma D.14 part (3) applied to the sampling design with $h = 1$, $n_T/n \xrightarrow{p} E[q(\psi_1)]$, so $(n_T/n)^{1/2} \xrightarrow{p} E[q(\psi_1)]^{1/2}$ by continuous mapping. Then by Slutsky's theorem $\sqrt{n_T}(\widehat{\theta} - \text{SATE}) = (n_T/n)^{1/2} \cdot \sqrt{n}(B_n + C_n) \mid \mathcal{H}_n^T \Rightarrow \mathcal{N}(0, V_{\text{SATE}})$, with $V_{\text{SATE}} = E[q(\psi_1)](V_B + V_C)$, establishing the conditional CLT. The marginal statement follows by tower and bounded convergence: $E[e^{it\sqrt{n_T}(\widehat{\theta} - \text{SATE})}] = E[E[e^{it\sqrt{n_T}(\widehat{\theta} - \text{SATE})} \mid \mathcal{H}_n^T]] \rightarrow e^{-t^2 V_{\text{SATE}}/2}$. \square

The next lemma shows that when $\psi_1 = \psi_2 = \psi$, the variance V_{ATE} of (C.3) reduces by algebraic manipulation to the variance V of Theorem 4.2.

Lemma C.8 (Variance equivalence). *Suppose $\psi_1 = \psi_2 = \psi$. Then*

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

Proof. Write $p = p(\psi)$, $q = q(\psi)$, $\sigma_d^2 = \sigma_d^2(\psi)$ for brevity. We first show the pointwise identity

$$\frac{\text{Var}(\bar{Y}|\psi)}{p(1-p)} = \frac{\sigma_1^2}{p} + \frac{\sigma_0^2}{1-p} - \text{Var}(\tau|\psi). \quad (\text{C.5})$$

Expanding $\bar{Y} = (1-p)Y(1) + pY(0)$ and $\tau = Y(1) - Y(0)$ gives

$$\begin{aligned} \text{Var}(\bar{Y}|\psi) &= (1-p)^2\sigma_1^2 + p^2\sigma_0^2 + 2p(1-p)\text{Cov}(Y(1), Y(0)|\psi), \\ \text{Var}(\tau|\psi) &= \sigma_1^2 + \sigma_0^2 - 2\text{Cov}(Y(1), Y(0)|\psi). \end{aligned}$$

Summing, we have $\text{Var}(\bar{Y}|\psi) + p(1-p)\text{Var}(\tau|\psi) = (1-p)\sigma_1^2 + p\sigma_0^2$. This follows since the Cov terms cancel and $(1-p)^2 + p(1-p) = 1-p$, $p^2 + p(1-p) = p$. Dividing by $p(1-p)$ gives (C.5). Setting $\psi_1 = \psi_2 = \psi$ in (C.3) and applying (C.5) to the 3rd term,

$$E \left[\frac{\text{Var}(\bar{Y}|\psi)}{qp(1-p)} \right] = E \left[\frac{1}{q} \left(\frac{\sigma_1^2}{p} + \frac{\sigma_0^2}{1-p} \right) \right] - E \left[\frac{\text{Var}(\tau|\psi)}{q} \right].$$

Combining the $\text{Var}(\tau|\psi)$ contribution with the middle term of (C.3),

$$E \left[\frac{1-q}{q} \text{Var}(\tau|\psi) \right] - E \left[\frac{\text{Var}(\tau|\psi)}{q} \right] = -E[\text{Var}(\tau|\psi)].$$

By the law of total variance, $\text{Var}(\tau) - E[\text{Var}(\tau|\psi)] = \text{Var}(E[\tau|\psi]) = \text{Var}(c(\psi))$. Summing the three terms inside the parentheses of (C.3),

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

Multiplying by $E[q]$ gives the claim. \square

Proof of Theorem 4.2. We invoke Theorem C.6 with $\psi_1 = \psi_2 = \psi$, so the constraint $\psi_1 = f(\psi_2)$ holds with f the identity. Theorem C.6 then gives $\sqrt{n_T}(\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V_{\text{ATE}})$ with V_{ATE} as in (C.3). By Lemma C.8, $V_{\text{ATE}} = V$ of (4.2), completing the proof. \square

Proof of Theorem B.2. Let $Z_i = (\psi_i, Y_i(0), Y_i(1), T_i, D_i)$, which are iid under the assumed design, and let $\pi_n \perp\!\!\!\perp Z_{1:n}$ be the randomness generating the fold partition $[n] = I_1 \cup \dots \cup I_K$. By Lemma D.16 with iid variables $Z_{1:n}$, trivial function $h \equiv 1$, random element $\kappa = \pi_n$, and the $\sigma(\pi_n)$ -measurable partition $\{I_1, \dots, I_K\}$, the fold sub-samples $(Z_i)_{i \in I_1}, \dots, (Z_i)_{i \in I_K}$ are jointly conditionally independent given $\sigma(\pi_n)$. In particular, for each fold k , $(Z_i)_{i \in I_k} \perp\!\!\!\perp \mathcal{A}_k | \pi_n$ with $\mathcal{A}_k = \sigma((Z_j)_{j \notin I_k})$.

Fix a fold k and apply Lemma D.22 with $S = I_k$ and $\mathcal{A} = \mathcal{A}_k$. The fold estimator $\hat{m}_d^{(-k)}$ is $\sigma(\mathcal{A}_k, \pi_n)$ -measurable by construction and consistent by hypothesis, so the hypotheses of the lemma hold. Writing $R_i^{(k)}$ for the AIPW summand of Lemma D.22 evaluated at $\hat{m}_d^{(-k)}$, and ϕ_i for the corresponding influence function, $n^{-1} \sum_{i \in I_k} R_i^{(k)} = n^{-1} \sum_{i \in I_k} \phi_i + o_p(n^{-1/2})$. Since $\hat{\theta}_{adj} = E_n[R_i^{(k(i))}] = \sum_{k=1}^K n^{-1} \sum_{i \in I_k} R_i^{(k)}$ and K is fixed, summing over the K folds gives $\hat{\theta}_{adj} = E_n[\phi_i] + o_p(n^{-1/2})$.

Conditional on ψ_i , the residuals ϵ_i^d are mean zero and independent of (T_i, D_i) by Assumption B.1. Hence ϕ_i are iid with $E[\phi_i] = \text{ATE}$. The covariance between the two IPW terms in $\text{Var}(\phi_i)$ vanishes by $D_i(1 - D_i) = 0$, and the residuals' conditional mean zero kills the covariance with $c(\psi_i)$, giving

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

By the Lindeberg–Lévy CLT, $\sqrt{n}(\hat{\theta}_{adj} - \text{ATE}) = \sqrt{n}E_n[\phi_i - \text{ATE}] + o_p(1) \Rightarrow \mathcal{N}(0, \text{Var}(\phi_i))$. Finally $n_T/n = E_n[T_i] \xrightarrow{p} E[q(\psi)]$, so $\sqrt{n_T}(\hat{\theta}_{adj} - \text{ATE}) = (n_T/n)^{1/2} \sqrt{n}(\hat{\theta}_{adj} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$ with $V = E[q(\psi)] \text{Var}(\phi_i)$, the variance of Equation 4.2. \square

C.3 Optimal Sampling Design

Proof of Theorem 5.1. Define $V(q) = E[\sigma_s^2(\psi)/q(\psi)]$. Define the sets

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

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

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

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 $\sup_\psi q^* \leq 1$ we have $E[\sigma_s(\psi)C(\psi)^{1/2}] > 0$. By Holder $E[\sigma_s(\psi)C(\psi)^{1/2}] \leq E[\sigma_s^2(\psi)]^{1/2}E[C(\psi)]^{1/2} < \infty$. Then we have $V(q^*) = B^{-1}E[\sigma_s(\psi)C(\psi)^{1/2}]^2 < \infty$. Also clearly $E[C(\psi)q^*(\psi)] = B$. This shows the claim. Next, suppose that $q^* + t\Delta \in \mathcal{Q}'$ for some $t \in [0, 1]$ and $|\Delta|_\infty < M$. Since $q^* \in \mathcal{Q}'$, by convexity of \mathcal{Q}' , $q^* + t'\Delta \in \mathcal{Q}'$ for all $0 \leq t' \leq t$. Then $dV_{q^*}[\Delta] \equiv \limsup_{t \rightarrow 0^+} t^{-1}(V(q^* + t\Delta) - V(q^*))$ is well-defined. We claim that this limit exists. Under our assumptions $q^*(\psi) > \delta > 0$ pointwise, so that for any ψ and all $t \leq \delta^2/2M < \infty$

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

Since $\|\sigma_s^2\|_1 < \infty$, dominated convergence implies

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

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

Proof of Theorem 5.3. The estimator is $\hat{\theta} = E_n[T_i H_i Y_i / \hat{q}_n(\psi_i)]$, with Horvitz-Thompson weight $H_i = (D_i - p(\psi_i)) / [p(\psi_i)(1 - p(\psi_i))]$, outcome level $\bar{Y}_i = (1 - p(\psi_i))Y_i(1) +$

$p(\psi_i)Y_i(0)$, and individual effect $\tau_i = Y_i(1) - Y_i(0)$. The implemented sampling propensity is the discretized estimate $\widehat{q}_n(\cdot)$, and the assignment propensity p is fixed.

Our plan is to apply the CLT of Theorem C.5 with $\gamma = 2$. To do so, we first verify the key Assumption C.2 with $\psi_1 = \psi_2 = \psi$, sampling propensity \widehat{q}_n , and assignment propensity $\widehat{p}_n \equiv p$.

Part (i) of the assumption holds with f the identity, $\widehat{p}_n \equiv p$ fixed, and ξ_n the pilot randomness of Assumption 5.2(i). To see this, note by Assumption 5.2(i), the variance estimates $\widehat{\sigma}_d^2$ are functions of external pilot data ξ_n independent of the experimental data and design variables, including the auxiliary design randomness (η^D, η^T) of Definition C.1, so $\xi_n \perp\!\!\!\perp (W_{1:n}, \pi_n, \eta^D, \eta^T)$. From the construction in Equation (5.5), $\widehat{q}_n(\cdot)$ is a function of ξ_n and the study covariates $\psi_{1:n}$, hence $\sigma(\xi_n, \psi_{1:n})$ -measurable as required by the assumption. We subsume ξ_n in π_n , so that \widehat{q}_n is $\sigma(\psi_{1:n}, \pi_n)$ -measurable.

Part (ii) of the assumption holds since \widehat{q}_n takes at most L_n rational levels with denominators at most \bar{k}_n and p is a fixed rational level, the bounds $\widehat{q}_n(\psi) \in (\delta, 1]$ and $p \in (\delta, 1 - \delta)$ are Assumption 5.2(ii), and $\bar{k}_n L_n = o(n)$ follows from Assumption 5.2(iv). Part (iii) holds since $E_n[(\widehat{q}_n(\psi_i) - q^*(\psi_i))^2] \xrightarrow{p} 0$ by Lemma D.18, using the discretization rate of Assumption 5.2(iv) and the heteroskedasticity consistency of Assumption 5.2(iii), and $E_n[(\widehat{p}_n(\psi_i) - p(\psi_i))^2] = 0$ identically, so the fixed limit propensities are $q = q^*$ and p . Part (iv) holds since $E[Y(d)^4] < \infty$ gives $E[Y(d)^2] < \infty$, and $E[|\psi|^\alpha] < \infty$ with $\alpha > \dim(\psi) + 1$ is Assumption 5.2(ii). The moment bound $E[|a(W)|^4] < \infty$ is verified for each integrand below, and the rate $\bar{k}_n = o(n^{\gamma/(2(\gamma+2))}) = o(n^{1/4})$ is Assumption 5.2(iv).

Using the pointwise identity $H_i Y_i = \tau_i + H_i \bar{Y}_i$, expand the estimator as in the proof of Theorem C.6, with the implemented propensity \widehat{q}_n in place of q :

$$\begin{aligned} \widehat{\theta} - \text{ATE} &= E_n[\tau_i - \text{ATE}] + E_n[\tau_i(T_i - \widehat{q}_n(\psi_i))/\widehat{q}_n(\psi_i)] + E_n[T_i H_i \bar{Y}_i/\widehat{q}_n(\psi_i)] \\ &\equiv A_n + B_n + C_n. \end{aligned}$$

We apply Lemma D.7 to deduce convergence in distribution of $\sqrt{n}(A_n + B_n + C_n)$ from conditional CLTs for $\sqrt{n}A_n$, $\sqrt{n}B_n$, and $\sqrt{n}C_n$ over the lemma's required increasing chain. In this context, we set $\mathcal{F}_{n,0} = \sigma(\emptyset)$, $\mathcal{F}_{n,1} = \mathcal{H}_n^T$, $\mathcal{F}_{n,2} = \mathcal{H}_n^D$, and $\mathcal{F}_{n,3} = \sigma(W_{1:n}, \pi_n, T_{1:n}, D_{1:n})$.

First consider $\sqrt{n}A_n$. By Young's inequality $\text{Var}(\tau) \leq 2E[Y(1)^2 + Y(0)^2] < \infty$, so the iid CLT gives $\sqrt{n}A_n \Rightarrow \mathcal{N}(0, V_A)$ with $V_A = \text{Var}(\tau)$. Since $\mathcal{F}_{n,0}$ is trivial, this is equivalently $\sqrt{n}A_n | \mathcal{F}_{n,0} \Rightarrow \mathcal{N}(0, V_A)$.

Next consider $\sqrt{n}B_n$. Write $\delta_i = 1/\widehat{q}_n(\psi_i) - 1/q^*(\psi_i)$ and split $B_n = B'_n + B''_n$, where $B'_n = E_n[(T_i - \widehat{q}_n(\psi_i))\tau_i/q^*(\psi_i)]$ and $B''_n = E_n[(T_i - \widehat{q}_n(\psi_i))\tau_i\delta_i]$. For B'_n , apply Theorem C.5 with the fixed integrand $a(W) = \tau/q^*(\psi)$, which satisfies $E[|a(W)|^4] \leq \delta^{-4}E[\tau^4] \lesssim E[Y(1)^4 + Y(0)^4] < \infty$. The theorem gives $\sqrt{n}B'_n|\mathcal{F}_{n,1} \Rightarrow \mathcal{N}(0, V_B)$ with

$$V_B = E[q^*(\psi)(1 - q^*(\psi)) \text{Var}(\tau/q^*(\psi)|\psi)] = E\left[\frac{1-q^*(\psi)}{q^*(\psi)} \text{Var}(\tau|\psi)\right].$$

For B''_n , since $\widehat{q}_n(\psi_i), q^*(\psi_i) \geq \delta$ by Assumption 5.2, we have $|\delta_i| \leq \delta^{-2}|\widehat{q}_n(\psi_i) - q^*(\psi_i)|$, and $|\widehat{q}_n - q^*| \leq 1$. Then using Lemma D.18, we have

$$E_n[\delta_i^4] \leq \delta^{-8}E_n[(\widehat{q}_n(\psi_i) - q^*(\psi_i))^4] \leq \delta^{-8}E_n[(\widehat{q}_n(\psi_i) - q^*(\psi_i))^2] = o_p(1). \quad (\text{C.6})$$

The term $\tau_i\delta_i$ in $B''_n = E_n[(T_i - \widehat{q}_n(\psi_i))\tau_i\delta_i]$ is $\mathcal{F}_{n,1}$ -measurable, and $E[T_i|\mathcal{F}_{n,1}] = \widehat{q}_n(\psi_i)$ by Lemma D.13 applied to the sampling design, so $E[\sqrt{n}B''_n|\mathcal{F}_{n,1}] = 0$. By the variance bound in the proof of Lemma D.14 applied to the sampling design,

$$\text{Var}(\sqrt{n}B''_n|\mathcal{F}_{n,1}) \leq 2E_n[\tau_i^2\delta_i^2] \leq 2E_n[\tau_i^4]^{1/2}E_n[\delta_i^4]^{1/2} = O_p(1)o_p(1) = o_p(1).$$

The second inequality is Cauchy-Schwarz, then $E_n[\tau_i^4] \xrightarrow{p} E[\tau^4] < \infty$ from the WLLN and (C.6), Chebyshev (Lemma D.1) gives $\sqrt{n}B''_n = o_p(1)$, so $\sqrt{n}B_n = \sqrt{n}B'_n + o_p(1)$.

Finally consider $\sqrt{n}C_n$. Split $C_n = C'_n + C''_n$, where

$$\begin{aligned} C'_n &= E_n[T_i(D_i - p(\psi_i))\bar{Y}_i/(q^*(\psi_i)p(\psi_i)(1 - p(\psi_i)))], \\ C''_n &= E_n[T_i(D_i - p(\psi_i))\bar{Y}_i\delta_i/(p(\psi_i)(1 - p(\psi_i)))]. \end{aligned}$$

For C'_n , apply the assignment part of Theorem C.5 with the fixed integrand $a(W) = \bar{Y}/(q^*(\psi)p(\psi)(1 - p(\psi)))$, which satisfies $|a(W)| \leq \delta^{-1}[\delta(1 - \delta)]^{-1}|\bar{Y}|$, hence $E[|a(W)|^4] \lesssim E[\bar{Y}^4] \lesssim E[Y(1)^4 + Y(0)^4] < \infty$. The theorem gives $\sqrt{n}C'_n|\mathcal{F}_{n,2} \Rightarrow \mathcal{N}(0, V_C)$ with

$$V_C = E[q^*(\psi)p(\psi)(1 - p(\psi)) \text{Var}(a|\psi)] = E\left[\frac{\text{Var}(\bar{Y}|\psi)}{q^*(\psi)p(\psi)(1 - p(\psi))}\right].$$

For C''_n , the integrand $\bar{Y}_i\delta_i/(p(\psi_i)(1 - p(\psi_i)))$ is $\mathcal{F}_{n,2}$ -measurable, and $E[D_i|\mathcal{F}_{n,2}] = p(\psi_i)$ by Lemma D.13, so $E[\sqrt{n}C''_n|\mathcal{F}_{n,2}] = 0$. By the variance bound in the proof of Lemma D.14, then $T_i \leq 1$ with $p(\psi_i)(1 - p(\psi_i)) \geq \delta(1 - \delta)$, then Cauchy-Schwarz

with $E_n[\bar{Y}_i^4] = O_p(1)$ from the iid weak law together with (C.6),

$$\begin{aligned}\text{Var}(\sqrt{n}C_n''|\mathcal{F}_{n,2}) &\leq 2E_n[T_i\bar{Y}_i^2\delta_i^2/(p(\psi_i)(1-p(\psi_i)))^2] \\ &\leq 2[\delta(1-\delta)]^{-2}E_n[\bar{Y}_i^4]^{1/2}E_n[\delta_i^4]^{1/2} = o_p(1).\end{aligned}$$

Conditional Chebyshev gives $\sqrt{n}C_n'' = o_p(1)$, so $\sqrt{n}C_n = \sqrt{n}C_n' + o_p(1)$.

By construction $\sqrt{n}A_n \in \sigma(W_{1:n}) \subseteq \mathcal{F}_{n,1}$, $\sqrt{n}B_n' \in \sigma(W_{1:n}, \pi_n, T_{1:n}) = \mathcal{F}_{n,2}$, and $\sqrt{n}C_n' \in \sigma(W_{1:n}, \pi_n, T_{1:n}, D_{1:n}) = \mathcal{F}_{n,3}$. Lemma D.7 applied with $X_{n,k} = \sqrt{n}(A_n, B_n', C_n')$ at $k = 1, 2, 3$ gives, for each $t \in \mathbb{R}$,

$$E[e^{it\sqrt{n}(A_n+B_n'+C_n')}] = e^{-t^2(V_A+V_B+V_C)/2} + o_p(1).$$

Hence $\sqrt{n}(A_n + B_n' + C_n') \Rightarrow \mathcal{N}(0, V_A + V_B + V_C)$, and since $\sqrt{n}(\hat{\theta} - \text{ATE}) = \sqrt{n}(A_n + B_n' + C_n') + o_p(1)$ by the work above, Slutsky gives $\sqrt{n}(\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V_A + V_B + V_C)$.

The variances V_A, V_B, V_C are those of Theorem C.6 evaluated at $\psi_1 = \psi_2 = \psi$ and sampling propensity q^* . The proof of Lemma C.8 establishes a purely algebraic identity for these three terms, giving $V_A + V_B + V_C = \text{Var}(c(\psi)) + E[\sigma_s^2(\psi)/q^*(\psi)]$ with $\sigma_s^2(\psi) = \sigma_1^2(\psi)/p(\psi) + \sigma_0^2(\psi)/(1-p(\psi))$. By Theorem 5.1 and the interior solution condition of Assumption 5.2(ii), q^* minimizes $E[\sigma_s^2(\psi)/q(\psi)]$ over $0 < q \leq 1$ with $E[C(\psi)q(\psi)] = B$. Therefore $\sqrt{n}(\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V^*)$ with

$$V^* = \text{Var}(c(\psi)) + \min_{\substack{0 < q \leq 1 \\ E[C(\psi)q(\psi)] = B}} E[\sigma_s^2(\psi)/q(\psi)].$$

□

C.4 Inference

Lemmas C.10 and D.19 below establish the ATE and SATE parts of Theorem 6.1 for the pairs-of-pairs estimator. The constant- q case is the special case $L = 1$, since (6.2) and (6.5) with $L = 1$ differ by the factor $n_T/(qn) \xrightarrow{p} 1$. We focus on proving consistency for the harder case with $\hat{P}^2 = \hat{P}_\nu^2$. Consistency for $\hat{P}^2 = \hat{P}_N^2$ can be shown similarly.

Below, it is convenient to work with partition $\mathcal{G}_n^\nu = \{g \cup \nu(g) : g \in \mathcal{G}_n\}$ into unions of paired groups. We write $\sum_{g \in \mathcal{G}_n} (\hat{\theta}_g - \hat{\theta}_{\nu(g)})^2 = 2 \sum_{u \in \mathcal{G}_n^\nu} (\hat{\theta}_{g_1(u)} - \hat{\theta}_{g_2(u)})^2$, where $\hat{\theta}_{g_1(u)}$ and $\hat{\theta}_{g_2(u)}$ are the different groups forming union $u \in \mathcal{G}_n^\nu$. The variance estimator of (6.4) is then $\hat{P}^2 = |\mathcal{G}_n|^{-1} \sum_{u \in \mathcal{G}_n^\nu} (\hat{\theta}_{g_1(u)} - \hat{\theta}_{g_2(u)})^2$, and similarly $\hat{P}_l^2 = |\mathcal{G}_{nl}|^{-1} \sum_{u \in \mathcal{G}_{nl}^\nu} (\hat{\theta}_{g_1(u)} - \hat{\theta}_{g_2(u)})^2$

within each stratum. Under sampling-subordinate matching every $g \in \mathcal{G}_n$ and every $u \in \mathcal{G}_n^\nu$ lies entirely in a single propensity stratum $S_{nl} = \{i : q(\psi_i) = q_l\}$, so group propensity q_g is equal to q_l whenever $g \in \mathcal{G}_{nl} := \{g \in \mathcal{G}_n : g \subseteq S_{nl}\}$ or $u \in \mathcal{G}_{nl}^\nu := \{u \in \mathcal{G}_n^\nu : u \subseteq S_{nl}\}$.

For each $l \in [L]$, let $n_l = |S_{nl}|$, and define the within-stratum IPW estimator and the group-level diff-in-means

$$\hat{\theta}_l = \frac{1}{n_l} \sum_{i \in S_{nl}} \frac{T_i(D_i - p)Y_i}{q_l p(1-p)}, \quad \hat{\theta}_g = \frac{1}{a} \sum_{i \in g} D_i Y_i - \frac{1}{k-a} \sum_{i \in g} (1 - D_i) Y_i,$$

so that $\hat{\theta} = \sum_{l=1}^L (n_l/n) \hat{\theta}_l$ is the global IPW estimator from (4.1). Throughout the proofs below, we use the within-group quadratic and across-group cross-product aggregates

$$\hat{v}_1 = \frac{1}{n} \sum_{g \in \mathcal{G}_n} \frac{k^2}{q_g^2} \hat{\theta}_g^2, \quad \hat{v}_2 = \frac{2}{n} \sum_{u \in \mathcal{G}_n^\nu} \frac{k(k - q_u)}{q_u^2} \hat{\theta}_{g_1(u)} \hat{\theta}_{g_2(u)}. \quad (\text{C.7})$$

Two conventions simplify the bookkeeping. First, we identify \mathcal{G}_n with the algorithm's interior groups of full size k , excluding the potential remainder group of size $< k$ in each of the L propensity strata. The IPW estimator $\hat{\theta}$ in (4.1) sums over all sampled units, so the identity $\hat{\theta} = |\mathcal{G}_n|^{-1} \sum_{g \in \mathcal{G}_n} \hat{\theta}_g$ from Section 6 hold up to a remainder group discrepancy of order $L(k-1)/n = o(1)$ that does not affect any consistency statement. Second, we assume $|\mathcal{G}_{nl}|$ is even in the proofs. As noted in the text, in practice the case of an odd number of groups is handled by matching one group twice in \mathcal{G}_{nl}^ν , contributing a single extra cross-product term to \hat{P}_l^2 of order $1/|\mathcal{G}_{nl}|$ that is similarly negligible but tedious to keep track of.

Lemma C.9 (Aggregated form). *The variance estimator \hat{V} of (6.5) satisfies*

$$\hat{V} = \frac{n_T}{n} (\hat{v}_1 - \hat{v}_2 - \hat{\theta}^2). \quad (\text{C.8})$$

Proof of Lemma C.9. Write $\hat{V}_l = \hat{S}_l^2 + ((k - q_l)/q_l) \hat{P}_l^2$ for the within-stratum terms, so $\hat{V} = (n_T/n) \sum_l (n_l/n) [\hat{V}_l + (\hat{\theta}_l - \hat{\theta})^2]$. By the within-stratum definitions of \hat{S}_l^2 and \hat{P}_l^2 as analogues of \hat{S}^2 and \hat{P}^2 from (6.2),

$$\hat{V}_l = \frac{1}{|\mathcal{G}_{nl}|} \sum_{g \in \mathcal{G}_{nl}} (\hat{\theta}_g - \hat{\theta}_l)^2 + \frac{k - q_l}{q_l |\mathcal{G}_{nl}|} \sum_{u \in \mathcal{G}_{nl}^\nu} (\hat{\theta}_{g_1(u)} - \hat{\theta}_{g_2(u)})^2. \quad (\text{C.9})$$

First we record the identity $\widehat{\theta}_l = |\mathcal{G}_{nl}|^{-1} \sum_{g \in \mathcal{G}_{nl}} \widehat{\theta}_g$. To see this, note

$$\begin{aligned} \frac{1}{|\mathcal{G}_{nl}|} \sum_{g \in \mathcal{G}_{nl}} \widehat{\theta}_g &= \frac{1}{|\mathcal{G}_{nl}|} \sum_{g \in \mathcal{G}_{nl}} \left(\frac{1}{a} \sum_{i \in g} D_i Y_i - \frac{1}{k-a} \sum_{i \in g} (1-D_i) Y_i \right) \\ &= \frac{1}{|\mathcal{G}_{nl}|} \left(\frac{1}{a} \sum_{i \in S_{nl}} T_i D_i Y_i - \frac{1}{k-a} \sum_{i \in S_{nl}} T_i (1-D_i) Y_i \right) \\ &= \frac{1}{k|\mathcal{G}_{nl}|p(1-p)} \sum_{i \in S_{nl}} T_i (D_i - p) Y_i = \frac{1}{n_l q_l p(1-p)} \sum_{i \in S_{nl}} T_i (D_i - p) Y_i = \widehat{\theta}_l. \end{aligned}$$

The second equality uses $\sum_{g \in \mathcal{G}_{nl}} \sum_{i \in g} = \sum_{i \in S_{nl}} T_i$, since \mathcal{G}_{nl} partitions $\{i \in S_{nl} : T_i = 1\}$. The third combines fractions using $a = pk$ and the identity $(1-p)D_i - p(1-D_i) = D_i - p$. The fourth uses $|\mathcal{G}_{nl}| = n_l q_l / k$, and the final equality is by definition of $\widehat{\theta}_l$.

Next, we expand squares in (C.9). Using the identity above, the first sum satisfies $\sum_{g \in \mathcal{G}_{nl}} (\widehat{\theta}_g - \widehat{\theta}_l)^2 = \sum_{g \in \mathcal{G}_{nl}} \widehat{\theta}_g^2 - |\mathcal{G}_{nl}| \widehat{\theta}_l^2$. For the second sum, each $g \in \mathcal{G}_{nl}$ appears exactly once across \mathcal{G}_{nl}^ν as either $g_1(u)$ or $g_2(u)$, so $\sum_{u \in \mathcal{G}_{nl}^\nu} (\widehat{\theta}_{g_1(u)} - \widehat{\theta}_{g_2(u)})^2 = \sum_{g \in \mathcal{G}_{nl}} \widehat{\theta}_g^2 - 2 \sum_{u \in \mathcal{G}_{nl}^\nu} \widehat{\theta}_{g_1(u)} \widehat{\theta}_{g_2(u)}$. Substituting into (C.9), we have

$$\begin{aligned} \widehat{V}_l &= \frac{1}{|\mathcal{G}_{nl}|} \left(\sum_{g \in \mathcal{G}_{nl}} \widehat{\theta}_g^2 - |\mathcal{G}_{nl}| \widehat{\theta}_l^2 \right) + \frac{k-q_l}{q_l |\mathcal{G}_{nl}|} \left(\sum_{g \in \mathcal{G}_{nl}} \widehat{\theta}_g^2 - 2 \sum_{u \in \mathcal{G}_{nl}^\nu} \widehat{\theta}_{g_1(u)} \widehat{\theta}_{g_2(u)} \right) \\ &= \left(\frac{1}{|\mathcal{G}_{nl}|} + \frac{k-q_l}{q_l |\mathcal{G}_{nl}|} \right) \sum_{g \in \mathcal{G}_{nl}} \widehat{\theta}_g^2 - \frac{2(k-q_l)}{q_l |\mathcal{G}_{nl}|} \sum_{u \in \mathcal{G}_{nl}^\nu} \widehat{\theta}_{g_1(u)} \widehat{\theta}_{g_2(u)} - \widehat{\theta}_l^2 \\ &= \frac{k^2}{n_l q_l^2} \sum_{g \in \mathcal{G}_{nl}} \widehat{\theta}_g^2 - \frac{2k(k-q_l)}{n_l q_l^2} \sum_{u \in \mathcal{G}_{nl}^\nu} \widehat{\theta}_{g_1(u)} \widehat{\theta}_{g_2(u)} - \widehat{\theta}_l^2. \end{aligned}$$

Next we aggregate across strata. Using the work above, $\sum_{l=1}^L \frac{n_l}{n} \widehat{V}_l$ is

$$\begin{aligned} &= \frac{1}{n} \sum_{l=1}^L \sum_{g \in \mathcal{G}_{nl}} \frac{k^2}{q_l^2} \widehat{\theta}_g^2 - \frac{2}{n} \sum_{l=1}^L \sum_{u \in \mathcal{G}_{nl}^\nu} \frac{k(k-q_l)}{q_l^2} \widehat{\theta}_{g_1(u)} \widehat{\theta}_{g_2(u)} - \sum_{l=1}^L \frac{n_l}{n} \widehat{\theta}_l^2 \\ &= \frac{1}{n} \sum_{g \in \mathcal{G}_n} \frac{k^2}{q_g^2} \widehat{\theta}_g^2 - \frac{2}{n} \sum_{u \in \mathcal{G}_n^\nu} \frac{k(k-q_u)}{q_u^2} \widehat{\theta}_{g_1(u)} \widehat{\theta}_{g_2(u)} - \sum_{l=1}^L \frac{n_l}{n} \widehat{\theta}_l^2 = \widehat{v}_1 - \widehat{v}_2 - \sum_{l=1}^L \frac{n_l}{n} \widehat{\theta}_l^2. \end{aligned}$$

Finally, we handle the cross-stratum correction. This expands as $\sum_l (n_l/n) (\widehat{\theta}_l - \widehat{\theta})^2 = \sum_l (n_l/n) \widehat{\theta}_l^2 - \widehat{\theta}^2$ using $\widehat{\theta} = \sum_l (n_l/n) \widehat{\theta}_l$ and $\sum_l n_l = n$. Adding this to the identity above, the $\sum_l (n_l/n) \widehat{\theta}_l^2$ terms cancel, leaving $\sum_l (n_l/n) [\widehat{V}_l + (\widehat{\theta}_l - \widehat{\theta})^2] = \widehat{v}_1 - \widehat{v}_2 - \widehat{\theta}^2$. Multiplying by n_T/n gives (C.8). \square

Lemma C.10 (Variance Estimator Consistency). *In the setting above, with $\psi_1 = \psi_2 = \psi$, design $T_{1:n} \sim \text{Loc}(\psi, q(\psi))$ and $D_{1:n} \sim \text{Loc}(\psi, p(\psi))$, and sampling-subordinate matching,*

$$\widehat{V} \xrightarrow{p} V. \quad (\text{C.10})$$

Proof of Lemma C.10. Let $\mathcal{H}_n^D = \sigma(W_{1:n}, T_{1:n}, \pi_n)$. Since $\psi_{1:n} \in \sigma(W_{1:n})$, the design filtration $\mathcal{F}_n^D = \sigma(\psi_{1:n}, \pi_n, T_{1:n})$ of Section C.2 satisfies $\mathcal{F}_n^D \subseteq \mathcal{H}_n^D$. Moreover, $\mathcal{H}_n^D \perp\!\!\!\perp \eta$ for the assignment randomness η of Definition C.1, since $\eta \perp\!\!\!\perp (W_{1:n}, \pi_n, T_{1:n})$ by construction. By Definition 2.1 and the construction of \mathcal{G}_n^ν above, the partitions \mathcal{G}_n and \mathcal{G}_n^ν are determined by $\psi_{1:n}$, π_n , and the sampled set $\{i : T_i = 1\}$, hence are \mathcal{F}_n^D -measurable and thus \mathcal{H}_n^D -measurable. By Lemma C.9, $\widehat{V} = (n_T/n)(\widehat{v}_1 - \widehat{v}_2 - \widehat{\theta}^2)$, where

$$\widehat{v}_1 = \frac{1}{n} \sum_{g \in \mathcal{G}_n} \frac{k^2}{q_g^2} \widehat{\theta}_g^2, \quad \widehat{v}_2 = \frac{2}{n} \sum_{u \in \mathcal{G}_n^\nu} \frac{k(k - q_u)}{q_u^2} \widehat{\theta}_{g_1(u)} \widehat{\theta}_{g_2(u)}.$$

First consider \widehat{v}_1 . Define $H_i = (D_i - p)/(p(1-p))$. Note the identity $H_i Y_i = \tau_i + H_i \bar{Y}_i$, where $\bar{Y}_i = (1-p)Y_i(1) + pY_i(0)$. Then $\widehat{\theta}_g = k^{-1} \sum_{i \in g} H_i Y_i = \theta_g + k^{-1} \sum_{i \in g} H_i \bar{Y}_i$, where $\theta_g = k^{-1} \sum_{i \in g} \tau_i$. Since \mathcal{G}_n , $W_{1:n}$ are \mathcal{H}_n^D -measurable and $E[H_i | \mathcal{H}_n^D] = 0$, it follows that $E[\widehat{\theta}_g | \mathcal{H}_n^D] = \theta_g$. Then $E[\widehat{\theta}_g^2 | \mathcal{H}_n^D] = \theta_g^2 + \text{Var}(\widehat{\theta}_g | \mathcal{H}_n^D)$. By Lemma D.13, $\text{Var}(H_i | \mathcal{H}_n^D) = [p(1-p)]^{-1}$ and $\text{Cov}(H_i, H_j | \mathcal{H}_n^D) = -[(k-1)p(1-p)]^{-1}$ for $i \neq j \in g$, so that

$$\begin{aligned} \text{Var}(\widehat{\theta}_g | \mathcal{H}_n^D) &= k^{-2} \text{Var}\left(\sum_{i \in g} H_i \bar{Y}_i \mid \mathcal{H}_n^D\right) = k^{-2} \sum_{i, j \in g} \bar{Y}_i \bar{Y}_j \text{Cov}(H_i, H_j | \mathcal{H}_n^D) \\ &= [k^2 p(1-p)]^{-1} \sum_{i \in g} \bar{Y}_i^2 - [k^2 (k-1)p(1-p)]^{-1} \sum_{i \neq j \in g} \bar{Y}_i \bar{Y}_j. \end{aligned}$$

Substituting $E[\widehat{\theta}_g^2 | \mathcal{H}_n^D] = \theta_g^2 + \text{Var}(\widehat{\theta}_g | \mathcal{H}_n^D)$ into $E[\widehat{v}_1 | \mathcal{H}_n^D]$ gives

$$\begin{aligned} E[\widehat{v}_1 | \mathcal{H}_n^D] &= n^{-1} \sum_{g \in \mathcal{G}_n} \frac{k^2}{q_g^2} \theta_g^2 + n^{-1} \sum_{g \in \mathcal{G}_n} [q_g^2 p(1-p)]^{-1} \sum_{i \in g} \bar{Y}_i^2 \\ &\quad - n^{-1} \sum_{g \in \mathcal{G}_n} [q_g^2 (k-1)p(1-p)]^{-1} \sum_{i \neq j \in g} \bar{Y}_i \bar{Y}_j \equiv A_n + B_n + C_n. \end{aligned}$$

First consider B_n . We have $q_g = q(\psi_i)$ for $i \in g$ by the assumption of sampling-

subordinate matching. Since $\cup_{g \in \mathcal{G}_n} \{i \in g\} = \{i : T_i = 1\}$,

$$B_n = [p(1-p)]^{-1} n^{-1} \sum_{g \in \mathcal{G}_n} \sum_{i \in g} \frac{\bar{Y}_i^2}{q(\psi_i)^2} = [p(1-p)]^{-1} E_n [T_i \bar{Y}_i^2 / q(\psi_i)^2].$$

By Lemma D.14 part (3) applied to the sampling design with $h = \bar{Y}^2 / q(\psi)^2$, $E_n [T_i \bar{Y}_i^2 / q(\psi_i)^2] \xrightarrow{p} E[\bar{Y}^2 / q(\psi)]$, so $B_n \xrightarrow{p} [p(1-p)]^{-1} E[\bar{Y}^2 / q(\psi)]$.

Next consider C_n . Using $q_g = q(\psi_i) = q(\psi_j)$ for $i, j \in g$, we have

$$C_n = -[(k-1)p(1-p)]^{-1} n^{-1} \sum_{g \in \mathcal{G}_n} \sum_{\substack{i, j \in g \\ i \neq j}} \frac{\bar{Y}_i \bar{Y}_j}{q(\psi_i) q(\psi_j)}.$$

By Lemma C.11 at $K = k$ on \mathcal{G}_n with $A_i = B_i = \bar{Y}_i / q(\psi_i)$, the inner double sum converges in probability to $(k-1)E[E[\bar{Y}|\psi]^2 / q(\psi)]$, so $C_n \xrightarrow{p} -[p(1-p)]^{-1} E[E[\bar{Y}|\psi]^2 / q(\psi)]$.

Finally, consider A_n . Expanding $\theta_g^2 = k^{-2} \sum_{i, j \in g} \tau_i \tau_j$ in A_n , the factor of k^2 cancels, and splitting diagonal from off-diagonal,

$$A_n = n^{-1} \sum_{g \in \mathcal{G}_n} \frac{1}{q_g^2} \sum_{i \in g} \tau_i^2 + n^{-1} \sum_{g \in \mathcal{G}_n} \frac{1}{q_g^2} \sum_{\substack{i, j \in g \\ i \neq j}} \tau_i \tau_j \equiv A_n^D + A_n^O.$$

For the diagonal piece, using $q_g = q(\psi_i)$ for $i \in g$, $\cup_{g \in \mathcal{G}_n} g = \{i : T_i = 1\}$, and Lemma D.14 part (3) applied to the sampling design,

$$A_n^D = n^{-1} \sum_{g \in \mathcal{G}_n} \sum_{i \in g} \frac{\tau_i^2}{q(\psi_i)^2} = E_n [T_i \tau_i^2 / q(\psi_i)^2] \xrightarrow{p} E[\tau^2 / q(\psi)].$$

For the off-diagonal piece, using $q_g = q(\psi_i) = q(\psi_j)$ for $i, j \in g$,

$$A_n^O = n^{-1} \sum_{g \in \mathcal{G}_n} \sum_{\substack{i, j \in g \\ i \neq j}} \frac{\tau_i \tau_j}{q(\psi_i) q(\psi_j)}.$$

By Lemma C.11 at $K = k$ on \mathcal{G}_n with $A_i = B_i = \tau_i / q(\psi_i)$, $A_n^O \xrightarrow{p} (k-1)E[c(\psi)^2 / q(\psi)]$, where $c(\psi) = E[\tau|\psi]$. Using $E[\tau^2|\psi] = \sigma_\tau^2(\psi) + c(\psi)^2$ for $\sigma_\tau^2(\psi) = \text{Var}(\tau|\psi)$, we have the limit $A_n \xrightarrow{p} E[\sigma_\tau^2(\psi) / q(\psi)] + kE[c(\psi)^2 / q(\psi)]$. Note also $E[\bar{Y}^2 / q(\psi)] = E[E[\bar{Y}^2|\psi] / q(\psi)]$ and $E[\bar{Y}^2|\psi] = E[\bar{Y}|\psi]^2 + \text{Var}(\bar{Y}|\psi)$, so the other terms

$$B_n + C_n \xrightarrow{p} [p(1-p)]^{-1} (E[\bar{Y}^2 / q(\psi)] - E[E[\bar{Y}|\psi]^2 / q(\psi)]) = [p(1-p)]^{-1} E[\text{Var}(\bar{Y}|\psi) / q(\psi)].$$

Combining with the limit of A_n , we have shown that

$$E[\widehat{v}_1|\mathcal{H}_n^D] \xrightarrow{p} E[\sigma_\tau^2(\psi)/q(\psi)] + kE[c(\psi)^2/q(\psi)] + \frac{E[\text{Var}(\bar{Y}|\psi)/q(\psi)]}{p(1-p)}.$$

Next consider $E[\widehat{v}_2|\mathcal{H}_n^D]$. By Lemma D.11 applied with $u_g = \widehat{\theta}_g$, which is a function of $(D_i)_{i \in g}$ and the \mathcal{H}_n^D -measurable potential outcomes $(Y_i(0), Y_i(1))_{i \in g}$, the group-level estimates $(\widehat{\theta}_g)_{g \in \mathcal{G}_n}$ are jointly independent conditional on \mathcal{H}_n^D ; in particular $\widehat{\theta}_{g_1(u)} \perp\!\!\!\perp \widehat{\theta}_{g_2(u)}|\mathcal{H}_n^D$ since $g_1(u) \neq g_2(u)$. Combined with $E[\widehat{\theta}_g|\mathcal{H}_n^D] = \theta_g$, this gives $E[\widehat{\theta}_{g_1(u)}\widehat{\theta}_{g_2(u)}|\mathcal{H}_n^D] = \theta_{g_1(u)}\theta_{g_2(u)}$. Since \mathcal{G}_n^ν is \mathcal{H}_n^D -measurable, taking conditional expectation through the sum yields

$$E[\widehat{v}_2|\mathcal{H}_n^D] = \frac{2}{n} \sum_{u \in \mathcal{G}_n^\nu} \frac{k(k - q_u)}{q_u^2} \theta_{g_1(u)} \theta_{g_2(u)} = \frac{2}{n} \sum_{u \in \mathcal{G}_n^\nu} \frac{k - q_u}{k q_u^2} \sum_{\substack{i \in g_1(u) \\ j \in g_2(u)}} \tau_i \tau_j.$$

Define $A_i = \sqrt{(k - q(\psi_i))/k} \cdot \tau_i/q(\psi_i)$, which is $\sigma(W_i)$ measurable and has $E[A_i^2] \leq E[\tau_i^2]/\delta^2 < \infty$ using $0 \leq (k - q)/k \leq 1$ and propensity bound $q(\psi) \geq \delta > 0$. By sampling-subordinate matching, $q_u = q(\psi_i) = q(\psi_j)$ for $i, j \in u$, so $A_i A_j = (k - q_u)/(k q_u^2) \tau_i \tau_j$ for all $i, j \in u$. Using $u = g_1(u) \cup g_2(u)$, we have the decomposition

$$\sum_{i \neq j \in u} A_i A_j = \sum_{i \neq j \in g_1(u)} A_i A_j + \sum_{i \neq j \in g_2(u)} A_i A_j + 2 \sum_{\substack{i \in g_1(u) \\ j \in g_2(u)}} A_i A_j.$$

Solving for the cross-group piece and summing over the disjoint unions $u \in \mathcal{G}_n^\nu$,

$$E[\widehat{v}_2|\mathcal{H}_n^D] = \frac{2}{n} \sum_{u \in \mathcal{G}_n^\nu} \sum_{\substack{i \in g_1(u) \\ j \in g_2(u)}} A_i A_j = \frac{1}{n} \sum_{u \in \mathcal{G}_n^\nu} \left[\sum_{i \neq j \in u} A_i A_j - \sum_{i \neq j \in g_1(u)} A_i A_j - \sum_{i \neq j \in g_2(u)} A_i A_j \right].$$

Note that by construction of \mathcal{G}_n^ν , each group $g \in \mathcal{G}_n$ appears exactly once as $g_1(u)$ or $g_2(u)$ for a unique union $u \in \mathcal{G}_n^\nu$, so summing over u gives recovers within-group sums over $g \in \mathcal{G}_n$, $\sum_{u \in \mathcal{G}_n^\nu} \left[\sum_{i \neq j \in g_1(u)} A_i A_j + \sum_{i \neq j \in g_2(u)} A_i A_j \right] = \sum_{g \in \mathcal{G}_n} \sum_{i \neq j \in g} A_i A_j$. Substituting,

$$E[\widehat{v}_2|\mathcal{H}_n^D] = \frac{1}{n} \sum_{u \in \mathcal{G}_n^\nu} \sum_{i \neq j \in u} A_i A_j - \frac{1}{n} \sum_{g \in \mathcal{G}_n} \sum_{i \neq j \in g} A_i A_j. \quad (\text{C.11})$$

By Lemma D.23, \mathcal{G}_n^ν inherits tight matching from \mathcal{G}_n , and by sampling-subordinate matching, $q(\psi_i)$ is constant over $i \in u$ and $i \in g$. Applying Lemma C.11 with $B_i = A_i$

at $K = 2k$ on \mathcal{G}_n^ν and at $K = k$ on \mathcal{G}_n , we obtain

$$\begin{aligned} \frac{1}{n} \sum_{u \in \mathcal{G}_n^\nu} \sum_{i \neq j \in u} A_i A_j &\xrightarrow{p} (2k-1) E[q(\psi_i) E[A_i | \psi_i]^2], \\ \frac{1}{n} \sum_{g \in \mathcal{G}_n} \sum_{i \neq j \in g} A_i A_j &\xrightarrow{p} (k-1) E[q(\psi_i) E[A_i | \psi_i]^2]. \end{aligned}$$

Then $E[\widehat{v}_2 | \mathcal{H}_n^D] \xrightarrow{p} k E[q(\psi_i) E[A_i | \psi_i]^2] = E[(k - q(\psi_i)) c(\psi_i)^2 / q(\psi_i)] = k E[c(\psi_i)^2 / q(\psi_i)] - E[c(\psi_i)^2]$. Finally, by Theorem 4.2 $\widehat{\theta} \xrightarrow{p} \text{ATE}$, so $\widehat{\theta}^2 \xrightarrow{p} \text{ATE}^2$ by continuous mapping.

Putting things together, recall that $\widehat{V} = (n_T/n)(\widehat{v}_1 - \widehat{v}_2 - \widehat{\theta}^2)$. By Lemma D.24, we have $\widehat{v}_1 - E[\widehat{v}_1 | \mathcal{H}_n^D] = o_p(1)$ and $\widehat{v}_2 - E[\widehat{v}_2 | \mathcal{H}_n^D] = o_p(1)$. Moreover, from the limits above

$$E[\widehat{v}_1 - \widehat{v}_2 | \mathcal{H}_n^D] - \widehat{\theta}^2 \xrightarrow{p} E[\sigma_\tau^2(\psi) / q(\psi)] + \frac{E[\text{Var}(\bar{Y} | \psi) / q(\psi)]}{p(1-p)} + E[c(\psi)^2] - \text{ATE}^2.$$

We have $E[c(\psi)^2] - \text{ATE}^2 = \text{Var}(c(\psi))$ and $\sigma_\tau^2(\psi) = \sigma_1^2(\psi) + \sigma_0^2(\psi) - 2 \text{Cov}(Y(1), Y(0) | \psi)$ and $\text{Var}(\bar{Y} | \psi) = (1-p)^2 \sigma_1^2(\psi) + p^2 \sigma_0^2(\psi) + 2p(1-p) \text{Cov}(Y(1), Y(0) | \psi)$. Then we have

$$\sigma_\tau^2(\psi) + \frac{\text{Var}(\bar{Y} | \psi)}{p(1-p)} = \frac{\sigma_1^2(\psi)}{p} + \frac{\sigma_0^2(\psi)}{1-p}.$$

Dividing through by $q(\psi)$, taking expectations, and combining with the $\text{Var}(c(\psi))$ term gives

$$\widehat{v}_1 - \widehat{v}_2 - \widehat{\theta}^2 \xrightarrow{p} \text{Var}(c(\psi)) + E \left[\frac{\sigma_1^2(\psi)}{pq(\psi)} + \frac{\sigma_0^2(\psi)}{(1-p)q(\psi)} \right] = V/E[q(\psi)], \quad (\text{C.12})$$

with the second equality by (4.2). Finally, $n_T/n = E_n[T_i] \xrightarrow{p} E[q(\psi)]$ by Lemma D.14 part (3) applied to the sampling design with $h = 1$, so by continuous mapping and Slutsky,

$$\widehat{V} = (n_T/n)(\widehat{v}_1 - \widehat{v}_2 - \widehat{\theta}^2) \xrightarrow{p} E[q(\psi)] \cdot V/E[q(\psi)] = V.$$

This finishes the proof. \square

The next lemma is stated for a generic partition \mathcal{P}_n , which we will set to either \mathcal{G}_n or \mathcal{G}_n^ν depending on the context in which the lemma is applied.

Lemma C.11 (Bilinear Form). *Let \mathcal{P}_n be a \mathcal{F}_n^D -measurable partition of $\{i : T_i = 1\}$ into groups of size K with $n^{-1} \sum_{g \in \mathcal{P}_n} \sum_{i \in g} |\psi_i - \bar{\psi}_g|_2^2 = o_p(1)$, for the standing filtration*

$\mathcal{F}_n^D = \sigma(\psi_{1:n}, \pi_n, T_{1:n})$ of Section C.2. Let $A_i, B_i \in \sigma(W_i)$ with $E[A^2 + B^2] < \infty$. Then

$$\frac{1}{n} \sum_{g \in \mathcal{P}_n} \sum_{\substack{i, j \in g \\ i \neq j}} A_i B_j \xrightarrow{p} (K-1) E[q(\psi_i) \cdot E[A_i | \psi_i] E[B_i | \psi_i]]. \quad (\text{C.13})$$

Proof of Lemma C.11. Let $A_i = v_a(\psi_i) + \epsilon_i^a$ with $E[\epsilon_i^a | \psi_i] = 0$, similarly $B_i = v_b(\psi_i) + \epsilon_i^b$.

$$\frac{1}{n} \sum_{g \in \mathcal{P}_n} \sum_{\substack{i, j \in g \\ i \neq j}} A_i B_j = \frac{1}{n} \sum_{g \in \mathcal{P}_n} \sum_{\substack{i, j \in g \\ i \neq j}} (v_a(\psi_i) + \epsilon_i^a)(v_b(\psi_j) + \epsilon_j^b) \equiv A_n + B_n + C_n + R_n,$$

where A_n collects the $v_a(\psi_i)v_b(\psi_j)$ terms, R_n collects the $\epsilon_i^a \epsilon_j^b$ terms, and B_n, C_n collect the cross terms $v_a(\psi_i)\epsilon_j^b$ and $\epsilon_i^a v_b(\psi_j)$ respectively. First consider A_n . Denote $v_{ia} = v_a(\psi_i)$, $v_{ja} = v_a(\psi_j)$, and v_{ib}, v_{jb} for b . Re-indexing $\sum_{i \neq j \in g} v_{ia} v_{jb} = \sum_{i < j \in g} (v_{ia} v_{jb} + v_{ja} v_{ib})$. By the identity $a_i b_j + b_i a_j = -(a_i - a_j)(b_i - b_j) + a_i b_i + a_j b_j$, this is

$$\begin{aligned} A_n &= \frac{1}{n} \sum_{g \in \mathcal{P}_n} \sum_{i < j \in g} (v_{ia} v_{jb} + v_{ja} v_{ib}) \\ &= \frac{1}{n} \sum_{g \in \mathcal{P}_n} \sum_{i < j \in g} (v_{ia} v_{ib} + v_{ja} v_{jb}) - \frac{1}{n} \sum_{g \in \mathcal{P}_n} \sum_{i < j \in g} (v_{ia} - v_{ja})(v_{ib} - v_{jb}) \equiv A_n^1 + A_n^2. \end{aligned}$$

First consider A_n^2 . Note the identity $\sum_{i, j \in g} |\psi_i - \psi_j|_2^2 = 2K \sum_{i \in g} |\psi_i - \bar{\psi}_g|_2^2$, so the tight matching hypothesis implies $X_n \equiv n^{-1} \sum_{g \in \mathcal{P}_n} \sum_{i, j \in g} |\psi_i - \psi_j|_2^2 = o_p(1)$. By Lemma D.10, we can choose $\mu_n \rightarrow \infty$ with $\mu_n X_n \xrightarrow{p} 0$. Then for $\lambda_n = \mu_n^{1/2}$, we have $\lambda_n \rightarrow \infty$ and $\lambda_n^2 X_n \xrightarrow{p} 0$. By Lemma D.6(ii), there exist sequences $(z_n^a)_{n \geq 1}$ and $(z_n^b)_{n \geq 1}$ with $|z_n^a|_{lip} \vee |z_n^b|_{lip} \leq \lambda_n$, $|z_n^a - v_a|_{2, \psi} = o(1)$, and $|z_n^b - v_b|_{2, \psi} = o(1)$. Denote $z_{in}^a = z_n^a(\psi_i)$, $z_{in}^b = z_n^b(\psi_i)$. We can expand relative to the Lipschitz approximations using $v_{ia} - v_{ja} = (v_{ia} - z_{in}^a) + (z_{in}^a - z_{jn}^a) + (z_{jn}^a - v_{ja})$ and similarly for b . Then $|xy| \leq x^2 + y^2$ and $(p_1 + p_2 + p_3)^2 \leq 3(p_1^2 + p_2^2 + p_3^2)$ give

$$|(v_{ia} - v_{ja})(v_{ib} - v_{jb})| \lesssim \sum_{c \in \{a, b\}} (|v_{ic} - z_{in}^c|^2 + |z_{in}^c - z_{jn}^c|^2 + |z_{jn}^c - v_{jc}|^2).$$

Using this to bound $|A_n^2|$ termwise, we have

$$\begin{aligned} |A_n^2| &\leq n^{-1} \sum_g \sum_{i < j \in g} \sum_{c \in \{a, b\}} (|v_{ic} - z_{in}^c|^2 + |z_{in}^c - z_{jn}^c|^2 + |z_{jn}^c - v_{jc}|^2) \\ &\lesssim \sum_{c \in \{a, b\}} [K E_n[T_i |v_{ic} - z_{in}^c|^2] + \lambda_n^2 X_n] = o_p(1). \end{aligned}$$

The first inequality is the pointwise bound above. The second equality follows by counting $\sum_{i < j \in g} |v_{ic} - z_{in}^c|^2 \leq (K-1) \sum_{i \in g} |v_{ic} - z_{in}^c|^2$ and similarly for the third term. Also, we use the Lipschitz bound $|z_{in}^c - z_{jn}^c|^2 \leq \lambda_n^2 |\psi_i - \psi_j|_2^2$, and $\sum_g \sum_{i \in g} f_i = \sum_i T_i f_i$ since \mathcal{P}_n partitions $\{T_i = 1\}$. For the final equality, $E_n[T_i |v_{ic} - z_{in}^c|^2] \leq E_n[|v_{ic} - z_{in}^c|^2] = o_p(1)$ by Markov, since $E[|v_{ic} - z_{in}^c|^2] = |z_n^c - v_c|_{2, \psi}^2 = o(1)$. For the second term, $\lambda_n^2 X_n \xrightarrow{p} 0$ by our choice of λ_n above. Then we have shown that $A_n^2 = o_p(1)$.

Next consider A_n^1 . Since $\sum_{i < j \in g} (v_{ia} v_{ib} + v_{ja} v_{jb}) = (K-1) \sum_{i \in g} v_{ia} v_{ib}$ for $|g| = K$,

$$A_n^1 = \frac{K-1}{n} \sum_{g \in \mathcal{P}_n} \sum_{i \in g} v_{ia} v_{ib} = (K-1) E_n[T_i v_{ia} v_{ib}] \xrightarrow{p} (K-1) E[q_i v_{ia} v_{ib}].$$

The middle equality since \mathcal{P}_n partitions $\{i : T_i = 1\} \subseteq [n]$. The convergence follows from Lemma D.14 part (3) applied to the sampling design with $h = v_a v_b$, since $E[|q_i v_{ia} v_{ib}|] \leq E[v_{ia}^2]^{1/2} E[v_{ib}^2]^{1/2} \lesssim E[A_i^2]^{1/2} E[B_i^2]^{1/2} < \infty$. Combining, $A_n \xrightarrow{p} (K-1) E[q(\psi_i) \cdot v_a(\psi_i) v_b(\psi_i)] = (K-1) E[q(\psi_i) \cdot E[A_i | \psi_i] E[B_i | \psi_i]]$.

Next claim that $R_n = o_p(1)$. We verify the conditions of Lemma D.8 with the filtration \mathcal{F}_n^D for the terms

$$u_g \equiv \sum_{\substack{i, j \in g \\ i \neq j}} \epsilon_i^a \epsilon_j^b.$$

We apply Lemma D.16 with $h_i = \psi_i$ and $\kappa = (\pi_n, T_{1:n})$, so that $\sigma(h_{1:n}, \kappa) = \sigma(\psi_{1:n}, \pi_n, T_{1:n}) = \mathcal{F}_n^D$. The required hypothesis $\kappa \perp\!\!\!\perp W_{1:n} | h_{1:n}$ is the statement $(\pi_n, T_{1:n}) \perp\!\!\!\perp W_{1:n} | \psi_{1:n}$, which we verify by chaining two reductions. First, $W_{1:n} \perp\!\!\!\perp \pi_n$ unconditionally gives $W_{1:n} \perp\!\!\!\perp \pi_n | \psi_{1:n}$ via $(A, B) \perp\!\!\!\perp C \Rightarrow A \perp\!\!\!\perp C | B$ with $A = W_{1:n}$, $B = \psi_{1:n} \in \sigma(W_{1:n})$, $C = \pi_n$. Second, by Definition C.1 applied to the sampling design, $T_{1:n}$ is a function of $(\psi_{1:n}, \pi_n, \eta^T)$ for sampling randomness $\eta^T \perp\!\!\!\perp (W_{1:n}, \pi_n)$, so $T_{1:n} \perp\!\!\!\perp W_{1:n} | (\psi_{1:n}, \pi_n)$. Composing, $W_{1:n} | (\psi_{1:n}, \pi_n, T_{1:n}) \stackrel{d}{=} W_{1:n} | (\psi_{1:n}, \pi_n) \stackrel{d}{=} W_{1:n} | \psi_{1:n}$, which is the required $(\pi_n, T_{1:n}) \perp\!\!\!\perp W_{1:n} | \psi_{1:n}$. Since \mathcal{P}_n is \mathcal{F}_n^D -measurable by hypothesis, Lemma D.16 gives $(u_g)_{g \in \mathcal{P}_n}$ jointly conditionally independent given \mathcal{F}_n^D . Note the key fact that $(A, B) \perp\!\!\!\perp$

$C \Rightarrow A \perp\!\!\!\perp C|B$. We first reduce the conditioning:

$$E[\epsilon_i^a \epsilon_j^b | \psi_{1:n}, \pi_n, T_{1:n}] = E[\epsilon_i^a \epsilon_j^b | \psi_{1:n}, \pi_n] = E[\epsilon_i^a \epsilon_j^b | \psi_{1:n}] = E[\epsilon_i^a \epsilon_j^b | \psi_i, \psi_j]. \quad (\text{C.14})$$

For the first equality, $T_{1:n} \perp\!\!\!\perp W_{1:n} | (\psi_{1:n}, \pi_n)$ as established above, hence $(\epsilon_i^a, \epsilon_j^b) \perp\!\!\!\perp T_{1:n} | (\psi_{1:n}, \pi_n)$ by the conditional independence fact above, since $\epsilon_i^a, \epsilon_j^b \in \sigma(W_{1:n})$. For the second equality, by design $W_{1:n} \perp\!\!\!\perp \pi_n$, hence $(\epsilon_i^a, \epsilon_j^b) \perp\!\!\!\perp \pi_n | \psi_{1:n}$ by the conditional independence fact above. The third equality drops $\psi_{-(i,j)}$: by iid sampling, $(\epsilon_i^a, \epsilon_j^b, \psi_i, \psi_j) \perp\!\!\!\perp \psi_{-(i,j)}$, hence $(\epsilon_i^a, \epsilon_j^b) \perp\!\!\!\perp \psi_{-(i,j)} | (\psi_i, \psi_j)$ by the fact above. Then note

$$E[\epsilon_i^a \epsilon_j^b | \psi_i, \psi_j] = E[\epsilon_j^b E[\epsilon_i^a | \epsilon_j^b, \psi_i, \psi_j] | \psi_i, \psi_j] = E[\epsilon_j^b E[\epsilon_i^a | \psi_i] | \psi_i, \psi_j] = 0.$$

The first equality is by tower law. The second uses $\epsilon_i^a \perp\!\!\!\perp (\epsilon_j^b, \psi_j) | \psi_i$, which holds by the fact above and because iid sampling gives $(\epsilon_i^a, \psi_i) \perp\!\!\!\perp (\epsilon_j^b, \psi_j)$. The third is $E[\epsilon_i^a | \psi_i] = 0$ by definition of the residual. Hence $E[u_g | \mathcal{F}_n^D] = 0$ for each $g \in \mathcal{P}_n$.

Next, to invoke Lemma D.8, we need to show $\frac{1}{n} \sum_g E[|u_g| \mathbf{1}(|u_g| > c_n) | \mathcal{F}_n^D] = o_p(1)$ for suitable $c_n \rightarrow \infty$. Note for any $a, b \geq 0$, $ab \mathbf{1}(ab > c) \leq a^2 \mathbf{1}(a^2 > c) + b^2 \mathbf{1}(b^2 > c)$ and for any $a_k \geq 0$, $k = 1, \dots, m$, we have

$$\sum_k a_k \mathbf{1}(\sum_k a_k > c) \leq m \sum_k a_k \mathbf{1}(a_k > c/m). \quad (\text{C.15})$$

By triangle and Young's inequalities,

$$|u_g| \leq \sum_{i \neq j \in g} |\epsilon_i^a| |\epsilon_j^b| \leq \frac{1}{2} \sum_{i \neq j \in g} ((\epsilon_i^a)^2 + (\epsilon_j^b)^2) = \frac{K-1}{2} \sum_{i \in g, c \in \{a, b\}} (\epsilon_i^c)^2.$$

Then for any $c_n \rightarrow \infty$,

$$\begin{aligned} E[|u_g| \mathbf{1}(|u_g| > c_n) | \mathcal{F}_n^D] &\leq \frac{K-1}{2} E\left[\sum_{i \in g, c} (\epsilon_i^c)^2 \mathbf{1}\left(\frac{K-1}{2} \sum_{i \in g, c} (\epsilon_i^c)^2 > c_n\right) \mid \mathcal{F}_n^D\right] \\ &\leq K(K-1) \sum_{i \in g, c \in \{a, b\}} E[(\epsilon_i^c)^2 \mathbf{1}((\epsilon_i^c)^2 > c_n [K(K-1)]^{-1}) \mid \mathcal{F}_n^D] \\ &= K(K-1) \sum_{i \in g, c \in \{a, b\}} E[(\epsilon_i^c)^2 \mathbf{1}((\epsilon_i^c)^2 > c_n [K(K-1)]^{-1}) \mid \psi_i]. \end{aligned}$$

The second inequality applies (C.15) to $\sum_{i \in g, c} (\epsilon_i^c)^2$ as a sum of $m = 2K$ non-negative terms and threshold $c = 2c_n / (K-1)$, giving per-term threshold $c/m = c_n [K(K-1)]^{-1}$.

The third reduces conditioning to ψ_i by the same argument as in the conditional-mean calculation. Then we can bound $\frac{1}{n} \sum_g E[|u_g| \mathbf{1}(|u_g| > c_n) | \mathcal{F}_n^D]$ above by

$$\begin{aligned} & \frac{K(K-1)}{n} \sum_{c \in \{a,b\}} \sum_g \sum_{i \in g} E[(\epsilon_i^c)^2 \mathbf{1}((\epsilon_i^c)^2 > c_n [K(K-1)]^{-1}) | \psi_i] \\ &= \frac{K(K-1)}{n} \sum_{c \in \{a,b\}} \sum_{i=1}^n T_i E[(\epsilon_i^c)^2 \mathbf{1}((\epsilon_i^c)^2 > c_n [K(K-1)]^{-1}) | \psi_i] \\ &\leq K(K-1) \sum_{c \in \{a,b\}} E_n [E[(\epsilon_i^c)^2 \mathbf{1}((\epsilon_i^c)^2 > c_n [K(K-1)]^{-1}) | \psi_i]]. \end{aligned}$$

The first expression follows from the bound on $E[|u_g| \mathbf{1}(|u_g| > c_n) | \mathcal{F}_n^D]$ developed above. The equality uses that \mathcal{P}_n partitions $\{T_i = 1\}$, so $\sum_g \sum_{i \in g} f_i = \sum_{i=1}^n T_i f_i$. The second inequality bounds $T_i \leq 1$ and rewrites the sum as E_n . Taking expectations, by iid sampling and tower law we obtain the bound

$$E \frac{1}{n} \sum_g E[|u_g| \mathbf{1}(|u_g| > c_n) | \mathcal{F}_n^D] \leq K(K-1) \sum_{c \in \{a,b\}} E[(\epsilon_i^c)^2 \mathbf{1}((\epsilon_i^c)^2 > c_n [K(K-1)]^{-1})].$$

The RHS converges to 0 as $c_n \rightarrow \infty$ by dominated convergence since $E[(\epsilon_i^c)^2] \leq \text{Var}(C_i) < \infty$ for $C \in \{A, B\}$. Then by Markov $\frac{1}{n} \sum_g E[|u_g| \mathbf{1}(|u_g| > c_n) | \mathcal{F}_n^D] = o_p(1)$. Choosing $c_n = n^{1/4}$ so that $c_n = \omega(1)$ and $c_n = o(n^{1/2})$, Lemma D.8 gives $R_n = o_p(1)$.

The terms $B_n, C_n = o_p(1)$ by an analogous argument. For example, for B_n let $u_g^B = \sum_{i \neq j \in g} v_{ia} \epsilon_j^b$. The conditional mean $E[v_{ia} \epsilon_j^b | \mathcal{F}_n^D] = 0$ since $v_{ia} \in \mathcal{F}_n^D$ and $E[\epsilon_j^b | \mathcal{F}_n^D] = E[\epsilon_j^b | \psi_j] = 0$ by a similar conditioning reduction argument as above. The Lindberg condition is verified by the same indicator-function argument with $(\epsilon_i^a)^2$ replaced by v_{ia}^2 , using $E[v_{ia}^2] = E[E[A_i | \psi_i]^2] \leq E[A_i^2] < \infty$ by Jensen's inequality.

Combining our work above, we have $A_n + B_n + C_n + R_n \xrightarrow{p} (K-1) E[q(\psi_i) \cdot E[A_i | \psi_i] E[B_i | \psi_i]]$, completing the proof. \square

Proof of Theorem 6.1. The ATE part $\widehat{V} \xrightarrow{p} V$ is Lemma C.10. The SATE part $\widehat{V}_{\text{SATE}} \xrightarrow{p} V_{\text{SATE}} + E[q(\psi)] E[\text{Var}(\tau | \psi)]$ is Lemma D.19, with $\sigma_\tau^2(\psi) = \text{Var}(\tau | \psi)$. In the constant- q case $L = 1$, the cross-stratum correction in (6.5) vanishes ($\widehat{\theta}_l = \widehat{\theta}$) and the resulting expression is $(n_T / (qn))$ times (6.2), with $n_T / (qn) \xrightarrow{p} 1$ by Lemma D.14 part (3) and Slutsky, so the limits agree. \square

Secondary Online Appendix to “Optimal Stratification of Survey Experiments”

Max Cytrynbaum

This secondary online appendix is not intended for publication.

D.1 Lemmas

Lemma D.1 (Conditional Convergence). *Let $(\mathcal{A}_n)_{n \geq 1}$ and $(A_n)_{n \geq 1}$ a sequence of σ -algebras and RV's. Define conditional convergence*

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

Then the following results hold

- (i) $A_n = o_p(1) \iff A_n = o_{p, \mathcal{A}_n}(1)$ and $A_n = O_p(1) \iff A_n = O_{p, \mathcal{A}_n}(1)$
- (ii) $E[|A_n| | \mathcal{A}_n] = o_p(1)/O_p(1) \implies A_n = o_p(1)/O_p(1)$
- (iii) $\text{Var}(A_n | \mathcal{A}_n) = o_p(c_n^2)/O_p(c_n^2) \implies A_n - E[A_n | \mathcal{A}_n] = o_p(c_n)/O_p(c_n)$.
- (iv) If $(A_n)_{n \geq 1}$ has $A_n \leq \bar{A} < \infty$ \mathcal{A}_n -a.s. $\forall n$ and $A_n = o_p(1) \implies E[|A_n| | \mathcal{A}_n] = o_p(1)$

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

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

If $A_n = o_p(1)$, then $E[P(|A_n| > \epsilon | \mathcal{A}_n)] = o(1)$, so $P(|A_n| > \epsilon | \mathcal{A}_n) = o_p(1)$ by Markov inequality. Conversely, if $P(|A_n| > \epsilon | \mathcal{A}_n) = o_p(1)$, then $E[P(|A_n| > \epsilon | \mathcal{A}_n)] = o(1)$ since $(P(|A_n| > \epsilon | \mathcal{A}_n))_{n \geq 1}$ 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 $\epsilon > 0$

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

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

Definition D.2 (Conditional Weak Convergence). For random variables $A_n, A \in \mathbb{R}^d$ and σ -algebras $(\mathcal{F}_n)_n, \mathcal{F}$, let μ_n and μ denote the laws of A_n given \mathcal{F}_n and of A given \mathcal{F} , which exist as regular conditional distributions (e.g. [Klenke \(2020\)](#), Theorem 8.29). Let $d_{BL}(\mu, \nu) = \sup_f \left| \int f d\mu - \int f d\nu \right|$ denote the bounded Lipschitz metric on $\mathcal{P}(\mathbb{R}^d)$, where the supremum is over functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ with $\|f\|_\infty + |f|_{lip} \leq 1$. Define

$$A_n | \mathcal{F}_n \Rightarrow A | \mathcal{F} \iff d_{BL}(\mu_n, \mu) = o_p(1).$$

Theorem D.3 (Conditional Lévy Continuity). For $A_n, A \in \mathbb{R}^d$, convergence holds $A_n | \mathcal{F}_n \Rightarrow A | \mathcal{F}$ if and only if $E[e^{it'A_n} | \mathcal{F}_n] = E[e^{it'A} | \mathcal{F}] + o_p(1)$ for each $t \in \mathbb{R}^d$.

Proof. We show the result for $d = 1$. The general case is similar. The laws μ_n and μ exist as regular conditional distributions ([Klenke \(2020\)](#), Theorem 8.29). In particular, $\omega \mapsto \mu_n^\omega(B)$ is \mathcal{F}_n -measurable for each Borel $B \subseteq \mathbb{R}$, and likewise for μ^ω . Equivalently, $\omega \mapsto \mu_n^\omega$ and $\omega \mapsto \mu^\omega$ are measurable into $\mathcal{P}(\mathbb{R})$ equipped with the σ -algebra generated by the evaluation maps $\mu \mapsto \mu(B)$, $B \in \mathcal{B}(\mathbb{R})$. This evaluation σ -algebra coincides with the Borel σ -algebra of the weak topology on $\mathcal{P}(\mathbb{R})$ ([Klenke \(2020\)](#), Exercise 24.1.2), which is metrized by d_{BL} . Since d_{BL} is continuous on $\mathcal{P}(\mathbb{R}) \times \mathcal{P}(\mathbb{R})$, composing the joint map $\omega \mapsto (\mu_n^\omega, \mu^\omega)$ with d_{BL} yields a real-valued random variable $\omega \mapsto d_{BL}(\mu_n^\omega, \mu^\omega)$.

Suppose first that $d_{BL}(\mu_n, \mu) \xrightarrow{p} 0$. Let $\varphi_n(t) = E[e^{itA_n} | \mathcal{F}_n]$ and $\varphi(t) = E[e^{itA} | \mathcal{F}]$ be the characteristic functions of μ_n and μ . Fix $t \in \mathbb{R}$. The functions $x \mapsto \cos(tx)$ and $x \mapsto \sin(tx)$ are bounded by 1 and Lipschitz with constant $|t|$, so each lies in the d_{BL} unit ball after scaling by $(1 + |t|)^{-1}$. By definition of d_{BL} as a supremum over the bounded-Lipschitz unit ball, $|\varphi_n(t) - \varphi(t)| \leq 2(1 + |t|) d_{BL}(\mu_n, \mu) = o_p(1)$. Since t was arbitrary, $\varphi_n(t) - \varphi(t) = o_p(1)$ for every $t \in \mathbb{R}$, which is [Definition D.2](#).

Suppose now that $\varphi_n(t) - \varphi(t) = o_p(1)$ for every $t \in \mathbb{R}$. By Urysohn's subsequence principle (e.g. [Klenke \(2020\)](#), Corollary 6.13), random variables $Y_n \xrightarrow{p} Y$ if and only if every subsequence (Y_{n_k}) contains a further subsequence $(Y_{n_{k_j}})$ with $Y_{n_{k_j}} \rightarrow Y$ almost surely. Applying this principle to $Y_n = d_{BL}(\mu_n, \mu)$ and $Y = 0$, it suffices to show that any subsequence (n_k) admits a further subsequence (n_{k_j}) along which $d_{BL}(\mu_{n_{k_j}}, \mu) \rightarrow 0$ almost surely.

For each fixed $t \in \mathbb{R}$, $|\varphi_n(t) - \varphi(t)| \leq 2$ and $\varphi_n(t) - \varphi(t) = o_p(1)$, so $E|\varphi_n(t) - \varphi(t)| = o(1)$ by bounded convergence. For any $K > 0$, Tonelli and dominated convergence give

$$E \int_0^1 |\varphi_n(t/K) - \varphi(t/K)| dt = \int_0^1 E|\varphi_n(t/K) - \varphi(t/K)| dt = o(1).$$

Markov's inequality then implies for each rational $K > 0$,

$$\int_0^1 |\varphi_n(t/K) - \varphi(t/K)| dt = o_p(1). \quad (\text{D.1})$$

Fix the subsequence (n_k) . We have countably many $o_p(1)$ statements along this subsequence: $\varphi_n(t) - \varphi(t) \xrightarrow{p} 0$ for each $t \in \mathbb{Q}$, and (D.1) for each rational $K > 0$. Applying Urysohn's principle again to convert each in-probability convergence to almost-sure convergence on a sub-subsequence, then using a standard diagonal argument across this countable family, we obtain a further subsequence (n_{k_j}) and a probability-one event Ω_0 on which simultaneously $\varphi_{n_{k_j}}(t, \omega) \rightarrow \varphi(t, \omega)$ for all $t \in \mathbb{Q}$ and

$$\int_0^1 |\varphi_{n_{k_j}}(t/K, \omega) - \varphi(t/K, \omega)| dt \rightarrow 0 \quad \text{for all rational } K > 0.$$

Fix $\omega \in \Omega_0$. The tightness estimate inside the proof of Lévy's continuity theorem (Klenke (2020), proof of Theorem 15.23) gives, with $\alpha = 1 - \sin(1) > 0$, for each $K > 0$ and $\text{Re}(z)$ denoting the real component of a complex number z ,

$$\mu_{n_{k_j}}^\omega([-K, K]^c) \leq \alpha^{-1} \int_0^1 (1 - \text{Re} \varphi_{n_{k_j}}(t/K, \omega)) dt.$$

Using the pointwise bound $1 - \text{Re} \varphi_{n_{k_j}}(t/K, \omega) \leq 1 - \text{Re} \varphi(t/K, \omega) + |\varphi_{n_{k_j}}(t/K, \omega) - \varphi(t/K, \omega)|$ in the integral, the second integral term vanishes as $j \rightarrow \infty$ for each rational $K > 0$ on Ω_0 by the second simultaneous fact above, leaving

$$\limsup_j \mu_{n_{k_j}}^\omega([-K, K]^c) \leq \alpha^{-1} \int_0^1 (1 - \text{Re} \varphi(t/K, \omega)) dt \equiv T(K, \omega).$$

The function $\varphi(\cdot, \omega)$ is the characteristic function of the probability measure μ^ω , hence continuous at 0 with $\varphi(0, \omega) = 1$, so $T(K, \omega) \rightarrow 0$ as $K \rightarrow \infty$ by dominated convergence. For any $\epsilon > 0$ we can therefore choose rational $K > 0$ with $T(K, \omega) < \epsilon$, giving $\limsup_j \mu_{n_{k_j}}^\omega([-K, K]^c) \leq \epsilon$. Hence the sequence $(\mu_{n_{k_j}}^\omega)_j$ is tight for each $\omega \in \Omega_0$.

For each $t \in \mathbb{Q}$, $\int e^{itx} d\mu_{n_{k_j}}^\omega(x) = \varphi_{n_{k_j}}(t, \omega) \rightarrow \varphi(t, \omega) = \int e^{itx} d\mu^\omega(x)$ on Ω_0 by the first simultaneous fact above. The complex exponentials $\{x \mapsto e^{itx} : t \in \mathbb{Q}\}$ form a separating family for finite measures on \mathbb{R} , since two such measures whose integrals against e^{itx} agree for every $t \in \mathbb{Q}$ have characteristic functions agreeing on \mathbb{Q} , hence everywhere on \mathbb{R} by continuity of characteristic functions, hence are equal by uniqueness.

By tightness of $(\mu_{n_{k_j}}^\omega)_j$ and [Klenke \(2020\)](#) Theorem 13.34, $\mu_{n_{k_j}}^\omega \Rightarrow \mu^\omega$ as $j \rightarrow \infty$ for each $\omega \in \Omega_0$. Since d_{BL} metrizes weak convergence on $\mathcal{P}(\mathbb{R})$, this gives $d_{BL}(\mu_{n_{k_j}}^\omega, \mu^\omega) \rightarrow 0$ for each $\omega \in \Omega_0$, i.e. almost surely. This finishes the proof. \square

We require a slight modification of the martingale difference CLT in [Billingsley \(1995\)](#), allowing the weak limit to be a mixture of normals.

Proposition D.4 (MDS-CLT). *Consider probability spaces $(\Omega_n, \mathcal{A}_n, P_n)$ 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 $n \rightarrow \infty$. Make the following definitions*

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

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

$$L_n^\epsilon = \sum_{k=1}^n E[Y_{k,n}^2 \mathbf{1}(|Y_{k,n}| \geq \epsilon) | \mathcal{F}_{0,n}] = o_p(1) \quad (\text{D.2})$$

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 \(1995\)](#).

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

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

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

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 $I_{k,n} = \mathbf{1}(|Y_{k,n}| > \epsilon)$. 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 $Z_n = o_p(1)$. \square

The next proposition specializes Proposition D.4 to triangular arrays that are jointly

independent conditional on a single σ -algebra, which may itself vary with n . This is the form we use in Section C.2.

Proposition D.5 (Conditional CLT). *For each n , let $\mathcal{F}_{0,n}$ be a σ -algebra and $(Y_{k,n})_{k=1}^n$ random variables that are jointly independent conditional on $\mathcal{F}_{0,n}$, with $E[Y_{k,n}|\mathcal{F}_{0,n}] = 0$ and $E[Y_{k,n}^2|\mathcal{F}_{0,n}] < \infty$. Let $S_n = \sum_{k=1}^n Y_{k,n}$ and $\Sigma_n = \sum_{k=1}^n E[Y_{k,n}^2|\mathcal{F}_{0,n}]$, and suppose $\Sigma_n = \sigma^2 + o_p(1)$ for some $\sigma^2 \in \mathcal{F}_{0,n}$. Suppose the conditional Lindeberg condition holds: for each $\epsilon > 0$, $\sum_{k=1}^n E[Y_{k,n}^2 \mathbf{1}(|Y_{k,n}| \geq \epsilon)|\mathcal{F}_{0,n}] = o_p(1)$. Then $E[e^{itS_n}|\mathcal{F}_{0,n}] = e^{-\frac{1}{2}t^2\sigma^2} + o_p(1)$ for each $t \in \mathbb{R}$.*

Proof. Define the increasing filtration $\mathcal{F}_{k,n} = \sigma(\mathcal{F}_{0,n}, Y_{1,n}, \dots, Y_{k,n})$ for $k \geq 1$, to which $(Y_{k,n})_{k=1}^n$ is adapted. By joint conditional independence, $Y_{k,n} \perp\!\!\!\perp (Y_{1,n}, \dots, Y_{k-1,n})|\mathcal{F}_{0,n}$. Hence $E[Y_{k,n}|\mathcal{F}_{k-1,n}] = E[Y_{k,n}|\mathcal{F}_{0,n}] = 0$ and $\sigma_{k,n}^2 = E[Y_{k,n}^2|\mathcal{F}_{k-1,n}] = E[Y_{k,n}^2|\mathcal{F}_{0,n}] \in \mathcal{F}_{0,n}$. So $(Y_{k,n}, \mathcal{F}_{k,n})$ is a martingale difference sequence with $\mathcal{F}_{0,n}$ -measurable conditional variances, and $\Sigma_n = \sum_{k=1}^n \sigma_{k,n}^2$. The variance and Lindeberg hypotheses of Proposition D.4 hold by assumption. The proposition then gives $E[e^{itS_n}|\mathcal{F}_{0,n}] = e^{-\frac{1}{2}t^2\sigma^2} + o_p(1)$. \square

Lemma D.6 (Lipschitz Approximation). *Let $Z \in \mathbb{R}^d$ be a random variable. Define $\mathcal{L} = \{g(Z) \in L_2(Z) : |g|_{lip} \vee |g|_\infty < \infty\}$.*

(i) \mathcal{L} is dense in $L_2(Z)$.

(ii) For any $f \in L_2(Z)$ and any sequence $\lambda_n \rightarrow \infty$, there exists a sequence $(z_n)_{n \geq 1}$ in \mathcal{L} with $|z_n|_{lip} \leq \lambda_n$ and $|z_n - f|_{2,Z} \rightarrow 0$.

Proof. For (i), let $\mathbf{1}(Z \in A)$ 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 $z \rightarrow d(z, A)$ 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 $f_n(z) \rightarrow \mathbf{1}(z \in A)$ pointwise as $n \rightarrow \infty$. Then by dominated convergence $\int (f_n(z) - \mathbf{1}(z \in A))^2 dP(z) \rightarrow 0$, 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 $\sum_k a_k \mathbf{1}(Z \in A_k)$ with $|a_k| < \infty$ for all k . If $g_{nk}(z) = (1 + nd(z, A_k))^{-1}$ the same argument shows that $\sum_k a_k g_{nk}(Z) - \sum_k a_k \mathbf{1}(Z \in A_k) \rightarrow 0$ 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.

For (ii), by (i) there exists a sequence $(\tilde{z}_k)_{k \geq 1}$ in \mathcal{L} with $M_k := |\tilde{z}_k|_{lip} < \infty$ and $|\tilde{z}_k - f|_{2,Z} \rightarrow 0$. Define $k(n) = \max\{k \geq 1 : M_k \leq \lambda_n\}$, taking $k(n) = 0$ and $\tilde{z}_0 \equiv 0$ if no such k exists. Since $\lambda_n \rightarrow \infty$, for any fixed $K \geq 1$ we have $\lambda_n \geq M_K$ eventually, hence $k(n) \geq K$ eventually and so $k(n) \rightarrow \infty$. Setting $z_n := \tilde{z}_{k(n)}$ gives $|z_n|_{lip} \leq \lambda_n$ by construction and $|z_n - f|_{2,Z} \rightarrow 0$ since $k(n) \rightarrow \infty$. \square

Lemma D.7 (Asymptotic Independence). *Consider a probability space (Ω, \mathcal{A}, P) with σ -algebras $\mathcal{F}_{n,k} \subseteq \mathcal{F}_{n,k+1} \subseteq \mathcal{A}$ 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 $X_{n,k} | \mathcal{F}_{n,k-1} \Rightarrow L_k | \mathcal{F}_{n,0}$, as in Definition D.2. Then $(X_{n,1}, \dots, X_{n,m}) | \mathcal{F}_{n,0} \Rightarrow (L_1, \dots, L_m) | \mathcal{F}_{n,0}$, with jointly independent limit.*

Proof. By Levy continuity, it suffices to show $E[e^{it'(X_{1,n}, \dots, X_{m,n})} | \mathcal{F}_{n,0}] \rightarrow \prod_{k=1}^m E[e^{it_k L_k} | \mathcal{F}_{n,0}]$ for all $t \in \mathbb{R}^m$. We work by induction on k . By assumption, $E[e^{itX_{n,1}} | \mathcal{F}_{n,0}] \xrightarrow{p} E[e^{itL_1} | \mathcal{F}_{n,0}]$ for all $t \in \mathbb{R}$. Assume by induction that the conclusion holds for $1 \leq k \leq k' \leq m$. Then

$$\begin{aligned} E[e^{i \sum_{k=1}^{k'+1} t_k X_{n,k}} | \mathcal{F}_{n,0}] &= E[e^{i \sum_{k=1}^{k'} t_k X_{n,k}} E[e^{it_{k'+1} X_{n,k'+1}} | \mathcal{F}_{n,k'}] | \mathcal{F}_{n,0}] \\ &= E[e^{i \sum_{k=1}^{k'} t_k X_{n,k}} (E[e^{it_{k'+1} X_{n,k'+1}} | \mathcal{F}_{n,k'}] - E[e^{it_{k'+1} L_{k'+1}} | \mathcal{F}_{n,0}]) | \mathcal{F}_{n,0}] \\ &\quad + E[e^{it_{k'+1} L_{k'+1}} | \mathcal{F}_{n,0}] E[e^{i \sum_{k=1}^{k'} t_k X_{n,k}} | \mathcal{F}_{n,0}] = \prod_{k=1}^{k'+1} E[e^{it_k L_k} | \mathcal{F}_{n,0}] + o_p(1) \end{aligned}$$

The first equality is by tower law and our measurability and increasing σ -algebra assumption. For the final equality, note $|e^{i \sum_{k=1}^{k'} t_k X_{n,k}} (E[e^{it_{k'+1} X_{n,k'+1}} | \mathcal{F}_{n,k'}] - E[e^{it_{k'+1} L_{k'+1}} | \mathcal{F}_{n,0}])| \leq |E[e^{it_{k'+1} X_{n,k'+1}} | \mathcal{F}_{n,k'}] - E[e^{it_{k'+1} L_{k'+1}} | \mathcal{F}_{n,0}]| \xrightarrow{p} 0$. 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. \square

Lemma D.8 (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 $(\mathcal{A}_n)_{n \geq 1}$ be σ -algebras such that \mathcal{G}_n is \mathcal{A}_n -measurable, $E[u_g | \mathcal{A}_n] = 0$, and for all $g \neq g' \in \mathcal{G}_n$ $u_g \perp\!\!\!\perp u_{g'} | \mathcal{A}_n$. If $n^{-1} \sum_{g \in \mathcal{G}_n} E[|u_g| \mathbb{1}(|u_g| > c_n) | \mathcal{A}_n] \xrightarrow{p} 0$ 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 $\bar{\mu}_{sn} = E[u_s \mathbb{1}(|u_s| \leq c_n) | \mathcal{A}_n]$. 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 $E[|T_{n1}||\mathcal{A}_n] \leq (1/n) \sum_{s=1}^n E[|u_s| \mathbf{1}(|u_s| > c_n)|\mathcal{A}_n] = o_p(1)$ by assumption. Then $T_{n1} = o_p(1)$ by conditional Markov (Lemma D.1). Next consider T_{n2} . Note that by definition $E[\bar{u}_{sn} - \bar{\mu}_{sn}|\mathcal{A}_n] = 0$ for each $1 \leq s \leq n$. Note that for $s \neq s'$ $\text{Cov}(\bar{u}_{sn}, \bar{u}_{s'n}|\mathcal{A}_n) = 0$ by the conditional independence assumption. Then $\text{Var}(T_{n2}|\mathcal{A}_n) = n^{-2} \sum_{s=1}^n \text{Var}(\bar{u}_{sn}|\mathcal{A}_n) \leq n^{-2} \sum_{s=1}^n E[\bar{u}_{sn}^2|\mathcal{A}_n] \leq n^{-1} c_n^2 = o(1)$, so that $T_{n2} = o_p(1)$ by conditional Chebyshev. Finally, since $E[u_s|\mathcal{A}_n] = 0$, we have $\bar{\mu}_{sn} = -E[u_s \mathbf{1}(|u_s| > c_n)|\mathcal{A}_n]$. Then $E[|T_{n3}||\mathcal{A}_n] \leq (1/n) \sum_{s=1}^n E[|u_s| \mathbf{1}(|u_s| > c_n)|\mathcal{A}_n] = o_p(1)$, so that $T_{n3} = o_p(1)$ by conditional Markov as before. This finishes the proof. \square

Lemma D.9. *Suppose $E[|X|^\alpha] < \infty$ for $\alpha > 0$. Then $\max_{i=1}^n |X_i| = o_p(n^{1/\alpha})$.*

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

Lemma D.10. *If $X_n \xrightarrow{p} 0$, there is a deterministic sequence $\lambda_n \rightarrow \infty$ with $\lambda_n X_n \xrightarrow{p} 0$.*

Proof. For each $k \geq 1$, choose N_k strictly increasing such that $P(|X_n| > 2^{-k}) < 2^{-k}$ for all $n \geq N_k$. Set $\lambda_n = 2^{k/2}$ for $N_k \leq n < N_{k+1}$, and $\lambda_n = 1$ for $n < N_1$. Then $\lambda_n \rightarrow \infty$ since $N_k \rightarrow \infty$. For $\epsilon > 0$ and $n \in [N_k, N_{k+1})$ with $k \geq 2 \log_2(1/\epsilon)$, $P(|\lambda_n X_n| > \epsilon) = P(|X_n| > \epsilon \cdot 2^{-k/2}) \leq P(|X_n| > 2^{-k}) < 2^{-k} \rightarrow 0$. \square

In the following lemmas, \mathcal{A}_n denotes a generic σ -algebra with $\mathcal{F}_n^D \subseteq \mathcal{A}_n$ and $\mathcal{A}_n \perp\!\!\!\perp \eta$. In our applications, we take \mathcal{A}_n to be either \mathcal{F}_n^D or \mathcal{H}_n^D depending on the context.

Lemma D.11 (Group Aggregate Independence). *Let $D_{1:n} \sim \text{Loc}(\psi, p(\psi))$ with associated partition \mathcal{G}_n and randomness η from Definition C.1. Let \mathcal{A}_n be a σ -algebra such that $\mathcal{F}_n^D \subseteq \mathcal{A}_n$ and $\mathcal{A}_n \perp\!\!\!\perp \eta$. Suppose $u_g = \phi((D_i)_{i \in g}, X, g)$ for some \mathcal{A}_n -measurable variable X and deterministic measurable function $\phi(\cdot)$. Then $(u_g)_{g \in \mathcal{G}_n}$ are jointly independent conditional on \mathcal{A}_n .*

Proof. By Equation (C.1), $(D_i)_{i \in g} = f(g, \eta_{s(g)})$ for the truncation function $f(\cdot)$ and the \mathcal{F}_n^D -measurable injection $s : \mathcal{G}_n \rightarrow [L] \times [n+1]$ of Definition C.1. Substituting, $u_g = \phi(f(g, \eta_{s(g)}), X, g) \equiv h(g, X, \eta_{s(g)})$ for the composition $h(g, X, \eta) = \phi(f(g, \eta), X, g)$. By construction, \mathcal{G}_n and $I = \{i : T_i = 1\}$ are \mathcal{F}_n^D measurable, thus \mathcal{A}_n measurable. The claim now follows from Lemma D.12, applied with finite index set $J = [L] \times [n+1]$, family $(\eta_{l,j})_{(l,j) \in J}$, injection $s(\cdot)$, and fixed function $h(\cdot)$ as above. \square

Lemma D.12 (Independence). *Let \mathcal{G}_n be a partition of $I \subseteq [n]$ with \mathcal{G}_n, I both \mathcal{A}_n -measurable. Let $(\eta_t)_{t \in J}$ be jointly independent random variables with J a finite set and $(\eta_t)_{t \in J} \perp\!\!\!\perp \mathcal{A}_n$, and let $s : \mathcal{G}_n \rightarrow J$ an \mathcal{A}_n -measurable injection assigning $g \in \mathcal{G}_n$ to $\eta_{s(g)}$. For each $g \in \mathcal{G}_n$, let $Z_g = h(g, X, \eta_{s(g)})$ for an \mathcal{A}_n -measurable random vector X and deterministic $h(\cdot)$. Then $(Z_g)_{g \in \mathcal{G}_n}$ are jointly independent conditional on \mathcal{A}_n .*

Proof. Let A be an \mathcal{A}_n -measurable event. For each fixed partition \mathcal{G} of I and each injection $r : \mathcal{G} \rightarrow J$, define $A(r, \mathcal{G}) \equiv A \cap \{\mathcal{G}_n = \mathcal{G}\} \cap \{s(\cdot) = r(\cdot)\}$ and observe that $A(r, \mathcal{G}) \in \mathcal{A}_n$. Observe that the finite family $\{A(r, \mathcal{G})\}_{(r, \mathcal{G})}$ partitions A , that is $A = \cup_{(r, \mathcal{G})} A(r, \mathcal{G})$. Then for any collection of bounded test functions $(\phi_g)_{g \subseteq I}$, we calculate

$$\begin{aligned} E[\mathbb{1}_A \prod_{g \in \mathcal{G}_n} \phi_g(Z_g)] &= E[\mathbb{1}_A \prod_{g \in \mathcal{G}_n} \phi_g(h(g, X, \eta_{s(g)}))] \\ &= \sum_{(r, \mathcal{G})} E[\mathbb{1}_{A(r, \mathcal{G})} \prod_{g \in \mathcal{G}} \phi_g(h(g, X, \eta_{r(g)}))] = \sum_{(r, \mathcal{G})} E[\mathbb{1}_{A(r, \mathcal{G})} E[\prod_{g \in \mathcal{G}} \phi_g(h(g, X, \eta_{r(g)})) | \mathcal{A}_n]] \end{aligned}$$

The first equality is by definition of Z_g . The second since $A = \cup_{(r, \mathcal{G})} A(r, \mathcal{G})$ and $\mathcal{G}_n = \mathcal{G}$, $s = r$ on $A(r, \mathcal{G})$. The third is by tower law and \mathcal{A}_n -measurability of $A(r, \mathcal{G})$. Continuing,

$$\begin{aligned} &= \sum_{(r, \mathcal{G})} E[\mathbb{1}_{A(r, \mathcal{G})} \prod_{g \in \mathcal{G}} E[\phi_g(h(g, X, \eta_{r(g)})) | \mathcal{A}_n]] \\ &= \sum_{(r, \mathcal{G})} E[\mathbb{1}_{A(r, \mathcal{G})} \prod_{g \in \mathcal{G}_n} E[\phi_g(Z_g) | \mathcal{A}_n]] = E[\mathbb{1}_A \prod_{g \in \mathcal{G}_n} E[\phi_g(Z_g) | \mathcal{A}_n]]. \end{aligned}$$

The first equality is by conditional independence of $(\eta_{r(g)})_{g \in \mathcal{G}}$ given \mathcal{A}_n and since X is \mathcal{A}_n -measurable and $h(\cdot)$ is fixed. Conditional independence holds since indices $\{r(g)\}_{g \in \mathcal{G}}$ are distinct by injectivity of $r(\cdot)$ and $(\eta_t)_{t \in J} \perp\!\!\!\perp \mathcal{A}_n$ and are jointly independent. The second equality uses $Z_g = h(g, X, \eta_{r(g)})$ and $\mathcal{G}_n = \mathcal{G}$ on $A(r, \mathcal{G})$. The final equality recombines the partition $A = \cup_{(r, \mathcal{G})} A(r, \mathcal{G})$. Then we have shown $E[\mathbb{1}_A \prod_{g \in \mathcal{G}_n} \phi_g(Z_g)] = E[\mathbb{1}_A \prod_{g \in \mathcal{G}_n} E[\phi_g(Z_g) | \mathcal{A}_n]]$ for any \mathcal{A}_n -measurable event A . By definition of conditional expectation, this implies $E[\prod_{g \in \mathcal{G}_n} \phi_g(Z_g) | \mathcal{A}_n] = \prod_{g \in \mathcal{G}_n} E[\phi_g(Z_g) | \mathcal{A}_n]$. Since the test function collection $(\phi_g)_{g \subseteq I}$ was arbitrary, this shows the claim. \square

Lemma D.13 (Design Properties). *Let $D_{1:n} \sim \text{Loc}(\psi, p(\psi))$ with partition \mathcal{G}_n , group propensity p_g for $g \in \mathcal{G}_n$, and randomness η from Definition C.1. Let \mathcal{A}_n be a σ -algebra such that $\mathcal{F}_n^D \subseteq \mathcal{A}_n$ and $\mathcal{A}_n \perp\!\!\!\perp \eta$. For $i \in g$, $E[D_i | \mathcal{A}_n] = p_g$ and $\text{Var}(D_i | \mathcal{A}_n) = p_g(1-p_g)$.*

For $i \neq j$,

$$\text{Cov}(D_i, D_j | \mathcal{A}_n) = \begin{cases} -\frac{p_g(1-p_g)}{|g|-1} & i, j \in g \text{ interior} \\ 0 & \text{otherwise.} \end{cases}$$

Proof. By Definition C.1 applied to $D_{1:n}$, conditional on \mathcal{A}_n the within-group vector $(D_i)_{i \in g}$ is distributed as $\text{CR}(p_g)$ for interior groups g and as iid Bernoulli(p_g) of length $|g|$ for the remainder. For $i \in g$, D_i is a single coordinate of $(D_i)_{i \in g}$, marginally Bernoulli(p_g) in both cases. Hence $E[D_i | \mathcal{A}_n] = p_g$ and $\text{Var}(D_i | \mathcal{A}_n) = p_g(1 - p_g)$.

For distinct i, j with associated groups $g(i) \neq g(j)$, Lemma D.11 applied with $u_g = (D_i)_{i \in g}$ implies the within-group vectors are jointly independent conditional on \mathcal{A}_n , so $\text{Cov}(D_i, D_j | \mathcal{A}_n) = 0$. For $i \neq j$ in the same remainder group, the iid Bernoulli structure of $(D_i)_{i \in g}$ conditional on \mathcal{A}_n gives $\text{Cov}(D_i, D_j | \mathcal{A}_n) = 0$. For $i \neq j$ in the same interior group g with $|g| = k$ and $p_g = a/k$, (D_i, D_j) has the joint marginal of $\text{CR}(a/k)$ on two distinct coordinates conditional on \mathcal{A}_n :

$$\begin{aligned} \text{Cov}(D_i, D_j | \mathcal{A}_n) &= P(D_i = D_j = 1 | \mathcal{A}_n) - p_g^2 = \binom{k}{a}^{-1} \binom{k-2}{a-2} - p_g^2 \\ &= \frac{a(a-1)}{k(k-1)} - \frac{a^2}{k^2} = -\frac{(a/k)(1-a/k)}{k-1} = -\frac{p_g(1-p_g)}{k-1}. \end{aligned}$$

□

Lemma D.14 (Stratified WLLN). *Let $D_{1:n} \sim \text{Loc}(\psi, p(\psi))$. Let \mathcal{A}_n be a σ -algebra such that $\mathcal{F}_n^D \subseteq \mathcal{A}_n$ and $\mathcal{A}_n \perp \eta$ in Definition C.1. Let $(h_n(W_i))_{i=1}^n$ be \mathcal{A}_n -measurable.*

1. *If $\sup_{n \geq 1} E[h_n(W)^2] < \infty$, then $E_n[T_i(D_i - p(\psi_i))h_n(W_i)] = O_p(n^{-1/2})$.*
2. *If $(h_n(W))_{n \geq 1}$ is uniformly integrable, then $E_n[T_i(D_i - p(\psi_i))h_n(W_i)] = o_p(1)$.*
3. *Let h be a fixed measurable function with $E[|h(W)|] < \infty$. Then $E_n[T_i h(W_i)] = E[q(\psi_i)h(W_i)] + o_p(1)$ and $E_n[T_i D_i h(W_i)] = E[q(\psi_i)p(\psi_i)h(W_i)] + o_p(1)$.*

Proof. Denote $a_i = T_i(D_i - p(\psi_i))h_n(W_i)$. Note $T_i, p(\psi_i), h_n(W_i) \in \mathcal{A}_n$ by $\mathcal{F}_n^D \subseteq \mathcal{A}_n$ and the hypothesis on h_n . We first show $E[E_n[a_i] | \mathcal{A}_n] = 0$. For i with $T_i = 1$, by the partition property i lies in some group $g(i) \in \mathcal{G}_n$ with $p_g = p(\psi_i)$, and Lemma D.13 gives $E[D_i | \mathcal{A}_n] = p(\psi_i)$. For i with $T_i = 0$ we have $a_i = 0$. In either case, we have $E[a_i | \mathcal{A}_n] = T_i h_n(W_i)(E[D_i | \mathcal{A}_n] - p(\psi_i)) = 0$. and summing gives $E[E_n[a_i] | \mathcal{A}_n] = 0$.

Next, we bound $\text{Var}(E_n[a_i] | \mathcal{A}_n) = n^{-2} \sum_{i,j} \text{Cov}(a_i, a_j | \mathcal{A}_n)$. By \mathcal{A}_n -measurability of the prefactors, $\text{Cov}(a_i, a_j | \mathcal{A}_n) = T_i T_j h_n(W_i) h_n(W_j) \text{Cov}(D_i, D_j | \mathcal{A}_n)$. Lemma D.13 gives $\text{Var}(D_i | \mathcal{A}_n) = p_g(1 - p_g) \leq 1/4$ for $i \in g$, $|\text{Cov}(D_i, D_j | \mathcal{A}_n)| = p_g(1 - p_g) / (|g| - 1) \leq |g|^{-1}$

for distinct i, j in an interior g . To see the inequality, note $p_g(1 - p_g) \leq 1/4 \leq (k - 1)/k$ for $k = |g| \geq 2$. Also $\text{Cov}(D_i, D_j | \mathcal{A}_n) = 0$ for all other pairs i, j by Lemma D.13. Then

$$\begin{aligned} \text{Var}(E_n[a_i] | \mathcal{A}_n) &\leq n^{-2} \sum_i T_i h_n(W_i)^2 + n^{-2} \sum_{g \text{ interior}} |g|^{-1} \sum_{i \neq j \in g} T_i T_j |h_n(W_i)| |h_n(W_j)| \\ &\leq n^{-1} E_n[T_i h_n(W_i)^2] + n^{-2} \sum_{g \text{ interior}} |g|^{-1} \left(\sum_{i \in g} |h_n(W_i)| \right)^2 \\ &\leq n^{-1} E_n[T_i h_n(W_i)^2] + n^{-2} \sum_{g \text{ interior}} \sum_{i \in g} h_n(W_i)^2 \leq 2n^{-1} E_n[T_i h_n(W_i)^2]. \end{aligned}$$

The first inequality uses $\text{Var}(D_i | \mathcal{A}_n) \leq 1$ and $|\text{Cov}(D_i, D_j | \mathcal{A}_n)| \leq |g|^{-1}$, together with the triangle inequality. The second uses $0 \leq T_i \leq 1$ to drop the indicators and extends $\sum_{i \neq j \in g}$ to $\sum_{i, j \in g}$. The third uses Cauchy-Schwarz: $(\sum_{i \in g} |h_n(W_i)|)^2 \leq |g| \sum_{i \in g} h_n(W_i)^2$.

Suppose $\sup_{n \geq 1} E[h_n(W)^2] < \infty$. By the law of total variance and work above we have $E[(E_n[a_i])^2] = E[\text{Var}(E_n[a_i] | \mathcal{A}_n)] \leq 2n^{-1} E[h_n(W)^2] \leq 2n^{-1} \sup_n E[h_n(W)^2]$. The final quantity is $O(n^{-1})$, so Markov inequality gives $E_n[a_i] = o_p(n^{-1/2})$.

Suppose instead $(h_n(W))_{n \geq 1}$ is UI. Let $\bar{h}_n(W_i) = h_n(W_i) \mathbf{1}(|h_n(W_i)| \leq n^{1/4})$. Decompose $E_n[a_i] = A_n + B_n$ where $A_n = E_n[T_i(D_i - p(\psi_i))(h_n(W_i) - \bar{h}_n(W_i))]$ and $B_n = E_n[T_i(D_i - p(\psi_i))\bar{h}_n(W_i)]$. For A_n , we have the bound $|A_n| \leq E_n[|h_n(W_i) - \bar{h}_n(W_i)|] = E_n[|h_n(W_i)| \mathbf{1}(|h_n(W_i)| > n^{1/4})]$ using $|T_i(D_i - p(\psi_i))| \leq 1$. By uniform integrability, $E[|h_n(W)| \mathbf{1}(|h_n(W)| > n^{1/4})] \rightarrow 0$, so $E[|A_n|] \rightarrow 0$ and $A_n = o_p(1)$ by Markov. For B_n , we have $E[B_n | \mathcal{A}_n] = 0$ by the work above and the variance bound applied with $\bar{h}_n(W_i)$ in place of $h_n(W_i)$ gives $\text{Var}(B_n | \mathcal{A}_n) \leq 2n^{-1} E_n[\bar{h}_n(W_i)^2] \leq 2n^{-1/2} = o(1)$ since $|\bar{h}_n(W_i)| \leq n^{1/4}$. Then $B_n = o_p(1)$ by Chebyshev (Lemma D.1). Then $E_n[a_i] = o_p(1)$.

For (3), decompose $E_n[T_i h(W_i)] = E_n[(T_i - q(\psi_i))h(W_i)] + E_n[q(\psi_i)h(W_i)]$. The first term is $o_p(1)$ by (2) applied to the sampling design, via the substitution $D_{1:n} \rightarrow T_{1:n}$ and $T_{1:n} \rightarrow 1$. Note the constant sequence $h_n = h$ is uniformly integrable since $E[|h(W)|] < \infty$. The second term equals $E[q(\psi_i)h(W_i)] + o_p(1)$ by vanilla WLLN for iid data, using $|qh| \leq |h|$ integrable. This gives the first identity. For the second identity, decompose $E_n[T_i D_i h(W_i)] = E_n[T_i(D_i - p(\psi_i))h(W_i)] + E_n[T_i p(\psi_i)h(W_i)]$. The first term is $o_p(1)$ by (2) applied to the assignment design with constant sequence $h_n = h$ uniformly integrable since $E[|h(W)|] < \infty$. The second term equals $E[q(\psi_i)p(\psi_i)h(W_i)] + o_p(1)$ by the first identity applied with ph in place of h , integrable since $|ph| \leq |h|$. \square

Lemma D.15 (Variance Identity). *Let $D_{1:n} \sim \text{Loc}(\psi, p(\psi))$ with partition \mathcal{G}_n . Let \mathcal{A}_n be a σ -algebra such that $\mathcal{F}_n^D \subseteq \mathcal{A}_n$ and $\mathcal{A}_n \perp\!\!\!\perp \eta$ in Definition C.1, and let $(h_n(W_i))_{i=1}^n$*

be \mathcal{A}_n -measurable. Then $\text{Var}(\sqrt{n}E_n[T_i(D_i - p(\psi_i))h_n(W_i)] \mid \mathcal{A}_n)$ is bounded above by

$$n^{-1} \sum_{g \in \mathcal{G}_n} |g|^{-1} \sum_{i,j \in g} (h_n(W_i) - h_n(W_j))^2 + n^{-1} \bar{k}_n L_n \cdot \max_{i=1}^n h_n(W_i)^2. \quad (\text{D.3})$$

Proof. Define $a_i = T_i(D_i - p(\psi_i))h_n(W_i)$. As in the proof of Lemma D.14, we have $\text{Cov}(a_i, a_j \mid \mathcal{A}_n) = T_i T_j h_n(W_i) h_n(W_j) \text{Cov}(D_i, D_j \mid \mathcal{A}_n)$. By Lemma D.13, the covariance $\text{Cov}(D_i, D_j \mid \mathcal{A}_n) = 0$ unless i, j lie in the same group, so

$$\begin{aligned} \text{Var}(\sqrt{n}E_n[a_i] \mid \mathcal{A}_n) &= n^{-1} \sum_{i,j} T_i T_j h_n(W_i) h_n(W_j) \text{Cov}(D_i, D_j \mid \mathcal{A}_n) \\ &= n^{-1} \sum_{g \in \mathcal{G}_n} \sum_{i,j \in g} h_n(W_i) h_n(W_j) \text{Cov}(D_i, D_j \mid \mathcal{A}_n). \end{aligned}$$

The second equality uses $T_i = T_j = 1$ for all $i, j \in g$ since \mathcal{G}_n partitions $\{i : T_i = 1\}$. For interior g with $|g| = k$ and $p_g = a/k$, Lemma D.13 gives $\text{Cov}(D_i, D_i \mid \mathcal{A}_n) = a(k - a)/k^2$ and $\text{Cov}(D_i, D_j \mid \mathcal{A}_n) = -a(k - a)/[k^2(k - 1)]$ for $i \neq j$. Then

$$\begin{aligned} \sum_{i,j \in g} h_n(W_i) h_n(W_j) \text{Cov}(D_i, D_j \mid \mathcal{A}_n) &= \frac{a(k - a)}{k^2} \sum_{i \in g} h_n(W_i)^2 - \frac{a(k - a)}{k^2(k - 1)} \sum_{i \neq j \in g} h_n(W_i) h_n(W_j) \\ &= \frac{a(k - a)}{k^2(k - 1)} \left[(k - 1) \sum_i h_n(W_i)^2 - \sum_{i \neq j} h_n(W_i) h_n(W_j) \right]. \end{aligned}$$

Using $\sum_{i \neq j \in g} h_n(W_i) h_n(W_j) = (\sum_i h_n(W_i))^2 - \sum_i h_n(W_i)^2$, the bracket equals $k \sum_i h_n(W_i)^2 - (\sum_i h_n(W_i))^2$, which equals $\frac{1}{2} \sum_{i,j \in g} (h_n(W_i) - h_n(W_j))^2$ by the identity $\sum_{i,j} (h_n(W_i) - h_n(W_j))^2 = 2[|g| \sum h_n(W_i)^2 - (\sum h_n(W_i))^2]$. Hence

$$\begin{aligned} \sum_{i,j \in g} h_n(W_i) h_n(W_j) \text{Cov}(D_i, D_j \mid \mathcal{A}_n) &= \frac{a(k - a)}{2k^2(k - 1)} \sum_{i,j \in g} (h_n(W_i) - h_n(W_j))^2 \\ &\leq \frac{1}{|g|} \sum_{i,j \in g} (h_n(W_i) - h_n(W_j))^2. \end{aligned}$$

Note we used $a(k - a)/[2k^2(k - 1)] \leq 1/k$ via $a(k - a) \leq k^2/4$ and $k(k - 1) \geq k^2/2$ for $k \geq 2$. For remainder groups g , $(D_i)_{i \in g}$ has independent components conditional on \mathcal{A}_n , so $\text{Cov}(D_i, D_j \mid \mathcal{A}_n) = 0$ for $i \neq j$ and $\text{Var}(D_i \mid \mathcal{A}_n) = p_g(1 - p_g) \leq 1$. Hence $\sum_{i,j \in g} h_n(W_i) h_n(W_j) \text{Cov}(D_i, D_j \mid \mathcal{A}_n) = \sum_{i \in g} h_n(W_i)^2 \text{Var}(D_i \mid \mathcal{A}_n) \leq \sum_{i \in g} h_n(W_i)^2$. Ag-

gregating over groups, $\text{Var}(\sqrt{n}E_n[a_i]|\mathcal{A}_n)$ is bounded above by

$$n^{-1} \sum_{g \text{ interior}} |g|^{-1} \sum_{i,j \in g} (h_n(W_i) - h_n(W_j))^2 + n^{-1} \sum_{g \text{ remainder}} \sum_{i \in g} h_n(W_i)^2.$$

The second term is bounded by $\sum_{g \text{ remainder}} \sum_{i \in g} h_n(W_i)^2 \leq \max_{i=1}^n h_n(W_i)^2 \cdot \sum_l |g_{l,\text{rem}}| \leq \bar{k}_n L_n \cdot \max_{i=1}^n h_n(W_i)^2$, since there are at most L_n remainder groups each of size at most \bar{k}_n . Extending the interior sum to include remainder groups gives the claimed bound. \square

Lemma D.16 (Random Partitions). *Let $W_{1:n}$ be iid random variables, $h_i = h(W_i)$ for a measurable function h , and κ a random element with $\kappa \perp\!\!\!\perp W_{1:n}|h_{1:n}$. Define $\mathcal{A}_n = \sigma(h_{1:n}, \kappa)$, and let \mathcal{G}_n be an \mathcal{A}_n -measurable partition of a subset of $[n]$. For each $g \in \mathcal{G}_n$, let $Z_g = F(g, (W_i)_{i \in g})$ for a deterministic measurable function F . Then $(Z_g)_{g \in \mathcal{G}_n}$ are jointly conditionally independent given \mathcal{A}_n .*

Proof. Since $\kappa \perp\!\!\!\perp W_{1:n}|h_{1:n}$, we have $W_{1:n}|h_{1:n}, \kappa \stackrel{d}{=} W_{1:n}|h_{1:n}$. Moreover, by iid sampling of $W_{1:n}$ and $h_i = h(W_i)$, this factorizes as $W_{1:n}|h_{1:n} \stackrel{d}{=} \prod_{i=1}^n (W_i|h_i)$. Hence $(W_i)_{i=1}^n$ are jointly conditionally independent given \mathcal{A}_n , with conditional marginal $W_i|h_i$.

Let $A \in \mathcal{A}_n$ and $(\phi_g)_{g \subseteq [n]}$ a finite collection of bounded measurable test functions. For each partition \mathcal{G} of a subset of $[n]$, define $A(\mathcal{G}) \equiv A \cap \{\mathcal{G}_n = \mathcal{G}\} \in \mathcal{A}_n$ by assumption, so the finite family $\{A(\mathcal{G})\}_{\mathcal{G}}$ partitions $A = \cup_{\mathcal{G}} A(\mathcal{G})$. Then

$$\begin{aligned} E[\mathbf{1}_A \Pi_g \phi_g(Z_g)] &= \sum_{\mathcal{G}} E[\mathbf{1}_{A(\mathcal{G})} \Pi_{g \in \mathcal{G}} \phi_g(F(g, (W_i)_{i \in g})))] \\ &= \sum_{\mathcal{G}} E[\mathbf{1}_{A(\mathcal{G})} E[\Pi_{g \in \mathcal{G}} \phi_g(F(g, (W_i)_{i \in g})) | \mathcal{A}_n]] \\ &= \sum_{\mathcal{G}} E[\mathbf{1}_{A(\mathcal{G})} \Pi_{g \in \mathcal{G}} E[\phi_g(F(g, (W_i)_{i \in g})) | \mathcal{A}_n]] \\ &= \sum_{\mathcal{G}} E[\mathbf{1}_{A(\mathcal{G})} \Pi_{g \in \mathcal{G}_n} E[\phi_g(Z_g) | \mathcal{A}_n]] = E[\mathbf{1}_A \Pi_g E[\phi_g(Z_g) | \mathcal{A}_n]]. \end{aligned}$$

The first equality decomposes A along $\mathcal{G}_n = \mathcal{G}$ and uses $Z_g = F(g, (W_i)_{i \in g})$ with g ranging in \mathcal{G} on $A(\mathcal{G})$. The second equality is by tower law and \mathcal{A}_n -measurability of $A(\mathcal{G})$. The key third equality uses the joint conditional independence of $(W_i)_{i=1}^n$ given \mathcal{A}_n established above and the fact that \mathcal{G} is a disjoint partition. The fourth equality uses $\mathcal{G}_n = \mathcal{G}$ and $Z_g = F(g, (W_i)_{i \in g})$ on $A(\mathcal{G})$. The fifth re-aggregates the partition $A = \cup_{\mathcal{G}} A(\mathcal{G})$. Since $A \in \mathcal{A}_n$ and (ϕ_g) were arbitrary, this gives $E[\Pi_g \phi_g(Z_g) | \mathcal{A}_n] = \Pi_g E[\phi_g(Z_g) | \mathcal{A}_n]$, hence joint conditional independence. \square

Theorem 4.2 describes the limiting variance of $\widehat{\theta}$. The next result complements it by quantifying the finite-sample variance, in the simplest setting of representative sampling $q \equiv 1$ and constant assignment propensity p . It makes precise the rate at which $n \text{Var}(\widehat{\theta})$ approaches the asymptotic variance V , and isolates the gap as a nonnegative term governed by the matching objective $F(\mathcal{G}_n)$.

Proposition D.17 (Finite-Sample Variance). *Suppose $q(\psi) = 1$ and $p(\psi) = p = a/k$ is constant, with $D_{1:n} \sim \text{Loc}(\psi, p)$ and assignment partition \mathcal{G}_n . Suppose also that $|Y(d)| \leq M$ almost surely and that $\psi \mapsto E[Y(d)|\psi]$ is B -Lipschitz, for $d \in \{0, 1\}$. Then the estimator $\widehat{\theta}$ of (4.1) satisfies*

$$n \text{Var}(\widehat{\theta}) = V + \Delta_n, \quad 0 \leq \Delta_n \leq \frac{B^2}{2p^2(1-p)^2} E[F(\mathcal{G}_n)] + \frac{(k-1)M^2}{np(1-p)}. \quad (\text{D.4})$$

Proof. Write $v = p(1-p)$ and let $b(\psi) = E[\bar{Y}|\psi]$ for the outcome level $\bar{Y} = (1-p)Y(1) + pY(0)$. Since $q = 1$ we have $T_i = 1$ and $n_T = n$. Since $T_{1:n}$ is deterministic, \mathcal{H}_n^D reduces to $\sigma(W_{1:n}, \pi_n)$ and \mathcal{F}_n^D reduces to $\sigma(\psi_{1:n}, \pi_n)$, with $\mathcal{F}_n^D \subseteq \mathcal{H}_n^D$ and $\mathcal{H}_n^D \perp\!\!\!\perp \eta^D$. The partition \mathcal{G}_n is \mathcal{F}_n^D -measurable by Definition 2.1. The outcomes are bounded, so $\widehat{\theta}$, τ , and \bar{Y} are bounded and every variance below is finite.

We first reduce to a conditional variance. Recall from the start of Section C.2 the decomposition $\widehat{\theta} - \text{ATE} = A_n + B_n + C_n$. With $q = 1$ the term $B_n = E_n[\tau_i(T_i - q)/q]$ vanishes. The remaining terms are $A_n = E_n[\tau_i - \text{ATE}]$ and $C_n = E_n[H_i \bar{Y}_i]$, with $H_i = (D_i - p)/v$, so $\widehat{\theta} - \text{ATE} = A_n + C_n$. The term A_n is $\sigma(W_{1:n})$ -measurable. By Lemma D.13, $E[D_i|\mathcal{H}_n^D] = p$, hence $E[C_n|\mathcal{H}_n^D] = 0$. Since $\sigma(W_{1:n}) \subseteq \mathcal{H}_n^D$ this gives $E[C_n] = 0$ and $\text{Cov}(A_n, C_n) = E[A_n E[C_n|\mathcal{H}_n^D]] = 0$. Therefore $n \text{Var}(\widehat{\theta}) = n \text{Var}(A_n) + n \text{Var}(C_n)$. The summands of A_n are iid and mean zero, so $n \text{Var}(A_n) = \text{Var}(\tau)$. The conditional mean $E[C_n|\mathcal{H}_n^D] = 0$ and the law of total variance give $\text{Var}(C_n) = E[\text{Var}(C_n|\mathcal{H}_n^D)]$.

Next we record the exact conditional variance. For a \mathcal{H}_n^D -measurable array $h = (h_i)_{i=1}^n$ define $S_n(h) = n^{-1/2} \sum_{i=1}^n (D_i - p)h_i$ with variance $\Gamma_n(h) = \text{Var}(S_n(h)|\mathcal{H}_n^D)$. Then $\sqrt{n}C_n = v^{-1}S_n(\bar{Y})$, so $n \text{Var}(C_n) = v^{-2}E[\Gamma_n(\bar{Y})]$. By Lemma D.13, $\text{Var}(D_i|\mathcal{H}_n^D) = v$, and $\text{Cov}(D_i, D_j|\mathcal{H}_n^D) = -v/(k-1)$ for $i \neq j$ in a common interior group and is zero for all other pairs. Since $E[S_n(h)|\mathcal{H}_n^D] = 0$, expanding the square gives

$$\Gamma_n(h) = n^{-1} \sum_{i,j} h_i h_j \text{Cov}(D_i, D_j|\mathcal{H}_n^D) = \frac{v}{n} \sum_{i=1}^n h_i^2 - \frac{v}{(k-1)n} \sum_{g \text{ interior}} \sum_{i \neq j \in g} h_i h_j. \quad (\text{D.5})$$

Centering within each interior group, with $\bar{h}_g = |g|^{-1} \sum_{i \in g} h_i$ and remainder group r ,

this is

$$\Gamma_n(h) = \frac{vk}{(k-1)n} \sum_{g \text{ interior}} \sum_{i \in g} (h_i - \bar{h}_g)^2 + \frac{v}{n} \sum_{i \in r} h_i^2. \quad (\text{D.6})$$

We now compute the term $E[\Gamma_n(\bar{Y})]$ in $n \text{Var}(C_n) = v^{-2}E[\Gamma_n(\bar{Y})]$ above using the decomposition in (D.5). Consider the off-diagonal sum first. The partition \mathcal{G}_n is \mathcal{F}_n^D -measurable, so the indicator $\chi_{ij} = \mathbb{1}(i, j \text{ in a common interior group})$ is \mathcal{F}_n^D -measurable, and the off-diagonal sum in (D.5) can be written $\sum_{i \neq j} \chi_{ij} h_i h_j$. By the same conditioning reductions used in the proof of Theorem C.5, for $i \neq j$ we have

$$E[\bar{Y}_i \bar{Y}_j | \mathcal{F}_n^D] = E[\bar{Y}_i \bar{Y}_j | \psi_{1:n}, \pi_n] = E[\bar{Y}_i | \psi_i] E[\bar{Y}_j | \psi_j] = b(\psi_i) b(\psi_j).$$

Since each χ_{ij} is \mathcal{F}_n^D -measurable, the tower law gives

$$E\left[\sum_{i \neq j} \chi_{ij} \bar{Y}_i \bar{Y}_j\right] = E\left[\sum_{i \neq j} \chi_{ij} E[\bar{Y}_i \bar{Y}_j | \mathcal{F}_n^D]\right] = E\left[\sum_{i \neq j} \chi_{ij} b(\psi_i) b(\psi_j)\right].$$

So the off-diagonal sum in (D.5) has the same expectation for $h = \bar{Y}$ as for $h = b$.

Consider the diagonal terms next. The diagonal sum in (D.5) is $vn^{-1} \sum_{i=1}^n h_i^2$, which we evaluate for $h = \bar{Y}$ and for $h = b$. By definition of conditional variance, $E[\bar{Y}_i^2 | \psi_i] = b(\psi_i)^2 + \text{Var}(\bar{Y} | \psi_i)$, so $E[\bar{Y}_i^2] = E[b(\psi)^2] + E[\text{Var}(\bar{Y} | \psi)]$, while $E[b(\psi_i)^2] = E[b(\psi)^2]$. Hence the diagonal sum for $h = \bar{Y}$ exceeds the diagonal sum for $h = b$ in expectation by

$$vn^{-1} \sum_{i=1}^n (E[\bar{Y}_i^2] - E[b(\psi_i)^2]) = vE[\text{Var}(\bar{Y} | \psi)].$$

Subtracting the instances of (D.5) for $h = \bar{Y}$ and $h = b$ and taking expectations, the off-diagonal contributions cancel and the diagonal contributions give

$$E[\Gamma_n(\bar{Y})] = E[\Gamma_n(b)] + v E[\text{Var}(\bar{Y} | \psi)]. \quad (\text{D.7})$$

It remains to bound the term $E[\Gamma_n(b)]$. The function $b = (1-p)E[Y(1)|\psi] + pE[Y(0)|\psi]$ is a convex combination of B -Lipschitz functions, hence B -Lipschitz. For any group g ,

$$\sum_{i \in g} (b(\psi_i) - \bar{b}_g)^2 = \frac{1}{2|g|} \sum_{i, j \in g} (b(\psi_i) - b(\psi_j))^2 \leq \frac{B^2}{2|g|} \sum_{i, j \in g} |\psi_i - \psi_j|_2^2 = B^2 \sum_{i \in g} |\psi_i - \bar{\psi}_g|_2^2.$$

Summing over interior groups and using $F(\mathcal{G}_n) = n^{-1} \sum_g \sum_{i \in g} |\psi_i - \bar{\psi}_g|_2^2$ of (2.1),

$$\frac{vk}{(k-1)n} \sum_{g \text{ interior}} \sum_{i \in g} (b(\psi_i) - \bar{b}_g)^2 \leq \frac{vk}{k-1} B^2 F(\mathcal{G}_n) \leq \frac{1}{2} B^2 F(\mathcal{G}_n).$$

The last step uses $v \leq 1/4$ and $k/(k-1) \leq 2$ for $k \geq 2$. The remainder group has $|r| \leq k-1$, and $|b| \leq M$ since $b(\psi) = E[\bar{Y}|\psi]$ and $|\bar{Y}| \leq M$. So the remainder term in (D.6) satisfies $vn^{-1} \sum_{i \in r} b(\psi_i)^2 \leq v(k-1)M^2/n$. Combining the two bounds with (D.6),

$$0 \leq \Gamma_n(b) \leq \frac{1}{2} B^2 F(\mathcal{G}_n) + v(k-1)M^2/n. \quad (\text{D.8})$$

Finally we assemble the pieces. Combining $n \text{Var}(C_n) = v^{-2} E[\Gamma_n(\bar{Y})]$ with (D.7),

$$n \text{Var}(\hat{\theta}) = \text{Var}(\tau) + v^{-2} E[\Gamma_n(\bar{Y})] = \text{Var}(\tau) + v^{-1} E[\text{Var}(\bar{Y}|\psi)] + v^{-2} E[\Gamma_n(b)].$$

The pointwise identity (C.5) gives $v^{-1} \text{Var}(\bar{Y}|\psi) = \sigma_1^2(\psi)/p + \sigma_0^2(\psi)/(1-p) - \text{Var}(\tau|\psi)$. The law of total variance gives $\text{Var}(\tau) = \text{Var}(c(\psi)) + E[\text{Var}(\tau|\psi)]$. Adding these,

$$\text{Var}(\tau) + v^{-1} E[\text{Var}(\bar{Y}|\psi)] = \text{Var}(c(\psi)) + E[\sigma_1^2(\psi)/p + \sigma_0^2(\psi)/(1-p)].$$

This equals V by (4.3) with $q = 1$. Set $\Delta_n = v^{-2} E[\Gamma_n(b)]$. Then $n \text{Var}(\hat{\theta}) = V + \Delta_n$, with $\Delta_n \geq 0$ by (D.6). The bound (D.8) gives $\Delta_n \leq v^{-2} (\frac{1}{2} B^2 E[F(\mathcal{G}_n)] + v(k-1)M^2/n)$, which is the inequality in (D.4). \square

Proof of Theorem 4.7. Fix $P \in \mathcal{P}$. The design has $q \equiv 1$ and constant $p = a/k$, and under P the potential outcomes satisfy $|Y(d)| \leq M$ almost surely with $E[Y(d)|\psi]$ Lipschitz of constant at most B , for $d \in \{0, 1\}$. So Proposition D.17 applies under P , giving $n \text{Var}_P(\hat{\theta}) = V(P) + \Delta_n(P)$, where $V(P)$ is the variance (4.3) under P and, with $v = p(1-p)$,

$$0 \leq \Delta_n(P) \leq \frac{B^2}{2v^2} E_P[F(\mathcal{G}_n)] + \frac{(k-1)M^2}{vn}.$$

Hence $|n \text{Var}_P(\hat{\theta}) - V(P)| = \Delta_n(P)$. The moment bound $E_P[|\psi|_2^\alpha] \leq K$ in the definition of \mathcal{P} lets us apply Corollary A.4 under P , with $\bar{k}_n = k$ and $L_n = 1$, giving

$$E_P[F(\mathcal{G}_n)] \leq C(\alpha, d) K^{2/\alpha} n^{2/\alpha - 2/(\dim(\psi)+1)},$$

where $C(\alpha, d)$ depends only on α and d and, in particular, not on P . Taking the

supremum over $P \in \mathcal{P}$,

$$\sup_{P \in \mathcal{P}} |n \operatorname{Var}_P(\hat{\theta}) - V(P)| \leq \frac{B^2 C(\alpha, d) K^{2/\alpha}}{2v^2} n^{2/\alpha-2/(\dim(\psi)+1)} + \frac{(k-1)M^2}{vn}.$$

Since $v = p(1-p)$, k , α , and K are fixed, this is $\lesssim B^2 n^{2/\alpha-2/(\dim(\psi)+1)} + M^2/n$. \square

Lemma D.18 (Propensity Convergence). *Suppose $|\hat{\sigma}_d^2 - \sigma_d^2|_{2,\psi}^2 = o_p(1)$ for $d = 0, 1$, with $\inf_\psi \sigma_d^2(\psi) > 0$ and $p(\psi) \in (\delta, 1 - \delta)$, and discretization $E_n[(\hat{q}_n(\psi_i) - \hat{q}(\psi_i))^2] = o_p(1)$. Then $E_n[(\hat{q}_n(\psi_i) - q^*(\psi_i))^2] = o_p(1)$.*

Proof. By the triangle inequality, $E_n[(\hat{q}_n - q^*)^2] \leq 2E_n[(\hat{q}_n - \hat{q})^2] + 2E_n[(\hat{q} - q^*)^2]$, with the first term $o_p(1)$ by hypothesis. It remains to bound $E_n[(\hat{q} - q^*)^2]$. Recall that

$$\hat{q}(\psi) = B \cdot \frac{\hat{\sigma}_s(\psi) C(\psi)^{-1/2}}{E_n[\hat{\sigma}_s(\psi_i) C(\psi_i)^{1/2}]}, \quad q^*(\psi) = B \cdot \frac{\sigma_s(\psi) C(\psi)^{-1/2}}{E[\sigma_s(\psi) C(\psi)^{1/2}]}.$$

Our approach is to first establish $|\hat{\sigma}_s - \sigma_s|_{2,\psi}^2 = o_p(1)$, then use it to control both the pointwise numerator difference $\hat{\sigma}_s - \sigma_s$ and the gap between the sample and population denominator normalizations.

Write $\sigma_s^2(\psi) = \sigma_1^2(\psi)/p(\psi) + \sigma_0^2(\psi)/(1-p(\psi))$ and $\sigma_s = \sqrt{\sigma_s^2}$, with $\hat{\sigma}_s^2, \hat{\sigma}_s$ defined analogously from $\hat{\sigma}_d^2$. By hypothesis $\sigma_d^2(\psi) \geq c_l > 0$ pointwise and $p(\psi), 1-p(\psi) \leq 1-\delta$, so $\sigma_s^2(\psi) \geq c_l/(1-\delta) =: \tilde{c} > 0$, hence $\hat{\sigma}_s + \sigma_s \geq \sigma_s \geq \sqrt{\tilde{c}}$. Using the identity $(\hat{\sigma}_s - \sigma_s)^2 = (\hat{\sigma}_s^2 - \sigma_s^2)^2/(\hat{\sigma}_s + \sigma_s)^2$ and the pointwise bound $(\hat{\sigma}_s^2 - \sigma_s^2)^2 \leq 2\delta^{-2}[(\hat{\sigma}_1^2 - \sigma_1^2)^2 + (\hat{\sigma}_0^2 - \sigma_0^2)^2]$ (using $p(\psi), 1-p(\psi) \geq \delta$),

$$|\hat{\sigma}_s - \sigma_s|_{2,\psi}^2 \lesssim \sum_{d=0,1} |\hat{\sigma}_d^2 - \sigma_d^2|_{2,\psi}^2 = o_p(1). \quad (\text{D.9})$$

Conditional on ξ_n , $(\psi_i)_{i=1}^n$ are iid and $\hat{\sigma}_s$ is fixed, so $E[E_n[(\hat{\sigma}_s(\psi_i) - \sigma_s(\psi_i))^2]|\xi_n] = |\hat{\sigma}_s - \sigma_s|_{2,\psi}^2 = o_p(1)$ by (D.9). By conditional Markov (Lemma D.1), $E_n[(\hat{\sigma}_s(\psi_i) - \sigma_s(\psi_i))^2] = o_p(1)$. Also $E[\sigma_s^2(\psi)] \leq \delta^{-1}E[\sigma_1^2 + \sigma_0^2] \leq 2\delta^{-1}E[Y(1)^2 + Y(0)^2] < \infty$ by Young's inequality, $p(\psi), 1-p(\psi) \geq \delta$, and Assumption 5.2(ii).

Define denominators $\mu = E[\sigma_s(\psi)C(\psi)^{1/2}]$ and $\mu_n = E_n[\hat{\sigma}_s(\psi_i)C(\psi_i)^{1/2}]$. Cauchy-Schwarz gives $\mu \leq E[\sigma_s^2]^{1/2}E[C]^{1/2} < \infty$, and pointwise $\sigma_s C^{1/2} \geq \tilde{c}^{1/2}C_l^{1/2}$ gives $\mu \geq \tilde{c}^{1/2}C_l^{1/2} > 0$. Conditional on ξ_n the units $(\psi_i)_{i=1}^n$ are iid and $\hat{\sigma}_s$ is a fixed function, so conditional Chebyshev gives $E_n[\hat{\sigma}_s C^{1/2}] = E[\hat{\sigma}_s C^{1/2}|\xi_n] + O_p(n^{-1/2})$. By Cauchy-Schwarz, $(E[\hat{\sigma}_s C^{1/2}|\xi_n] - E[\sigma_s C^{1/2}])^2 \leq C_u \cdot E_\psi[(\hat{\sigma}_s - \sigma_s)^2|\xi_n] = o_p(1)$ using (D.9).

Combining and applying Young's inequality,

$$(\mu_n - \mu)^2 = O_p(n^{-1}) + o_p(1) = o_p(1). \quad (\text{D.10})$$

Then $\mu_n \geq \mu/2$ on an event of probability tending to one. Finally, decompose

$$\widehat{q}(\psi) - q^*(\psi) = \frac{BC(\psi)^{-1/2}}{\mu_n \mu} [\mu (\widehat{\sigma}_s(\psi) - \sigma_s(\psi)) - \sigma_s(\psi) (\mu_n - \mu)].$$

On the event $\{\mu_n \geq \mu/2\}$, using $C^{-1} \leq C_l^{-1}$ and $\mu \geq \tilde{c}^{1/2} C_l^{1/2}$,

$$E_n[(\widehat{q} - q^*)^2] \lesssim E_n[(\widehat{\sigma}_s(\psi_i) - \sigma_s(\psi_i))^2] + (\mu_n - \mu)^2 \cdot E_n[\sigma_s(\psi_i)^2].$$

The first term is $o_p(1) = o_p(1)$ by work above; the second is $o_p(1) \cdot O_p(1) = o_p(1)$ using (D.10) and $E_n[\sigma_s^2] \xrightarrow{p} E[\sigma_s^2] < \infty$ by the iid weak law. \square

We turn to consistency of the SATE variance estimator $\widehat{V}_{\text{SATE}}$ of (6.5).

Lemma D.19 (Conservative SATE Inference). *In the setting of Lemma C.10,*

$$\widehat{V}_{\text{SATE}} \xrightarrow{p} V_{\text{SATE}} + E[q(\psi)]E[\sigma_\tau^2(\psi)]. \quad (\text{D.11})$$

Proof of Lemma D.19. Let $\mathcal{H}_n^D = \sigma(W_{1:n}, T_{1:n}, \pi_n)$ as in the proof of Lemma C.10. The within-stratum bracket in (6.5) is $(k/q_l)\widehat{P}_l^2 = (k/(q_l|\mathcal{G}_{nl}|)) \sum_{u \in \mathcal{G}_{nl}^\nu} (\widehat{\theta}_{g_1(u)} - \widehat{\theta}_{g_2(u)})^2$. By sampling-subordinate matching every union $u \in \mathcal{G}_n^\nu$ lies in a single propensity stratum with $q_u = q_l$, and $|\mathcal{G}_{nl}| = n_l q_l / k$, so $k/(q_l|\mathcal{G}_{nl}|) = k^2/(n_l q_l^2)$. Summing the within-stratum bracket over l gives the aggregated form $\widehat{V}_{\text{SATE}} = (n_T/n)\widehat{Q}$ with

$$\widehat{Q} = \frac{1}{n} \sum_{u \in \mathcal{G}_n^\nu} \frac{k^2}{q_u^2} (\widehat{\theta}_{g_1(u)} - \widehat{\theta}_{g_2(u)})^2.$$

Expand the square. Each $g \in \mathcal{G}_n$ appears exactly once across the unions with $q_u = q_g$, so the group-square terms collect into \widehat{v}_1 of (C.7). Comparing the resulting cross-product coefficient $2k^2/q_u^2$ with that of \widehat{v}_2 in (C.7) gives $2k^2/q_u^2 - 2k(k - q_u)/q_u^2 = 2k/q_u$, hence

$$\widehat{Q} = \widehat{v}_1 - \widehat{v}_2 - T_n, \quad T_n = \frac{2k}{n} \sum_{u \in \mathcal{G}_n^\nu} \frac{1}{q_u} \widehat{\theta}_{g_1(u)} \widehat{\theta}_{g_2(u)}.$$

By (C.12) and $\widehat{\theta} \xrightarrow{p} \text{ATE}$, we have $\widehat{v}_1 - \widehat{v}_2 \xrightarrow{p} V/E[q(\psi)] + \text{ATE}^2$.

Next consider T_n . Each summand of T_n is a union-level cross product with bounded \mathcal{H}_n^D -measurable weight $2k/q_u \leq 2k/\delta$, so the argument of Lemma D.24 gives $T_n - E[T_n|\mathcal{H}_n^D] = o_p(1)$. By Lemma D.11, $\widehat{\theta}_{g_1(u)} \perp\!\!\!\perp \widehat{\theta}_{g_2(u)}|\mathcal{H}_n^D$, and $E[\widehat{\theta}_g|\mathcal{H}_n^D] = \theta_g$, so $E[T_n|\mathcal{H}_n^D] = \frac{2k}{n} \sum_u q_u^{-1} \theta_{g_1(u)} \theta_{g_2(u)}$. Define $A_i = \tau_i/q(\psi_i)^{1/2}$, which is $\sigma(W_i)$ -measurable with $E[A_i^2] \leq E[\tau_i^2]/\delta < \infty$. By sampling-subordinate matching $q_u = q(\psi_i) = q(\psi_j)$ for $i, j \in u$, so $A_i A_j = \tau_i \tau_j / q_u$, and $\theta_g = k^{-1} \sum_{i \in g} \tau_i$ gives $q_u^{-1} \theta_{g_1(u)} \theta_{g_2(u)} = k^{-2} \sum_{i \in g_1(u), j \in g_2(u)} A_i A_j$. Using $u = g_1(u) \cup g_2(u)$ and that each $g \in \mathcal{G}_n$ appears once across the unions,

$$E[T_n|\mathcal{H}_n^D] = \frac{2}{nk} \sum_{u \in \mathcal{G}_n^\nu} \sum_{\substack{i \in g_1(u) \\ j \in g_2(u)}} A_i A_j = \frac{1}{nk} \left(\sum_{u \in \mathcal{G}_n^\nu} \sum_{i \neq j \in u} A_i A_j - \sum_{g \in \mathcal{G}_n} \sum_{i \neq j \in g} A_i A_j \right).$$

By Lemma D.23 the unions \mathcal{G}_n^ν inherit tight matching from \mathcal{G}_n , so Lemma C.11 with $B_i = A_i$ at $K = 2k$ on \mathcal{G}_n^ν and at $K = k$ on \mathcal{G}_n gives the two sums the limits $(2k-1)E[c(\psi)^2]$ and $(k-1)E[c(\psi)^2]$, using $E[A_i|\psi_i] = c(\psi_i)/q(\psi_i)^{1/2}$ and hence $E[q(\psi_i)E[A_i|\psi_i]^2] = E[c(\psi_i)^2]$. Subtracting and dividing by k gives $E[T_n|\mathcal{H}_n^D] \xrightarrow{p} E[c(\psi)^2]$, hence $T_n \xrightarrow{p} E[c(\psi)^2]$.

Combining, $\widehat{Q} \xrightarrow{p} V/E[q(\psi)] + \text{ATE}^2 - E[c(\psi)^2] = V/E[q(\psi)] - \text{Var}(c(\psi))$, using $E[c(\psi)^2] - \text{ATE}^2 = \text{Var}(c(\psi))$. By the variance formula (4.2), $V/E[q(\psi)] - \text{Var}(c(\psi)) = E[q(\psi)^{-1}(\sigma_1^2(\psi)/p + \sigma_0^2(\psi)/(1-p))]$. Since $n_T/n \xrightarrow{p} E[q(\psi)]$ by Lemma D.14 part (3), Slutsky gives

$$\widehat{V}_{\text{SATE}} = (n_T/n)\widehat{Q} \xrightarrow{p} E[q(\psi)]E[q(\psi)^{-1}(\sigma_1^2(\psi)/p + \sigma_0^2(\psi)/(1-p))].$$

It remains to identify this limit. With $\psi_1 = \psi_2 = \psi$ and p constant, $\text{Var}(\tau|\psi) = \sigma_\tau^2(\psi)$ and the identity (C.5) give the pointwise equality $\frac{1-q(\psi)}{q(\psi)}\sigma_\tau^2(\psi) + \frac{\text{Var}(\bar{Y}|\psi)}{q(\psi)p(1-p)} = q(\psi)^{-1}(\sigma_1^2(\psi)/p + \sigma_0^2(\psi)/(1-p)) - \sigma_\tau^2(\psi)$. Substituting into the variance of Corollary C.7 gives $V_{\text{SATE}} = E[q(\psi)](E[q(\psi)^{-1}(\sigma_1^2(\psi)/p + \sigma_0^2(\psi)/(1-p))] - E[\sigma_\tau^2(\psi)])$. Hence $\widehat{V}_{\text{SATE}} \xrightarrow{p} V_{\text{SATE}} + E[q(\psi)]E[\sigma_\tau^2(\psi)]$. \square

The within-group estimator (6.3) admits an exact, design-based characterization of its conditional bias. We state it for $q = 1$ without stratified sampling, where $\mathcal{H}_n^D = \sigma(W_{1:n}, \pi_n)$ carries the data and the matching randomness, $E[\widehat{\theta}|\mathcal{H}_n^D] = \text{SATE}$, and only the assignment $D_{1:n}$ is random.

Lemma D.20 (Within-group Variance Estimator). *Let $q = 1$, $p = a/k$, and $\min(a, k - a) \geq 2$. Then for $\bar{S}_\tau^2 = \frac{1}{|\mathcal{G}_n|} \sum_{g \in \mathcal{G}_n} \frac{1}{k-1} \sum_{i \in g} (\tau_i - \theta_g)^2$, the within-group estimator has*

$$E[\widehat{V}_{\text{SATE}}|\mathcal{H}_n^D] = n \text{Var}(\widehat{\theta}|\mathcal{H}_n^D) + \bar{S}_\tau^2.$$

The bias \bar{S}_τ^2 obeys the bound $0 \leq \bar{S}_\tau^2 \leq \frac{k}{k-1} S_\tau^2 \leq 2S_\tau^2$ for $S_\tau^2 = \frac{1}{n-1} \sum_{i=1}^n (\tau_i - \text{SATE})^2$.

Proof. Write $H_i = (D_i - p)/(p(1-p))$ and $\bar{Y}_i = (1-p)Y_i(1) + pY_i(0)$, so that $H_i Y_i = \tau_i + H_i \bar{Y}_i$ and, since $q = 1$, $\hat{\theta}_g = k^{-1} \sum_{i \in g} H_i Y_i = \theta_g + k^{-1} \sum_{i \in g} H_i \bar{Y}_i$ with $\hat{\theta} = |\mathcal{G}_n|^{-1} \sum_g \hat{\theta}_g$.

Unbiasedness of \hat{P}_N^2 . For an interior group g , treated units reveal $D_i Y_i = D_i Y_i(1)$, so $(a-1)s_{1,g}^2 = \frac{a-1}{a} \sum_{i \in g} D_i Y_i(1)^2 - \frac{1}{a} \sum_{i \neq j \in g} D_i D_j Y_i(1) Y_j(1)$. By Lemma D.13 with $\mathcal{A}_n = \mathcal{H}_n^D$, $E[D_i | \mathcal{H}_n^D] = a/k$ and $E[D_i D_j | \mathcal{H}_n^D] = a(a-1)/(k(k-1))$ for $i \neq j \in g$, and substituting $\sum_{i \neq j} Y_i(1) Y_j(1) = k^2 \mu_g(1)^2 - \sum_i Y_i(1)^2$ with $\mu_g(d) = k^{-1} \sum_{i \in g} Y_i(d)$,

$$E[s_{1,g}^2 | \mathcal{H}_n^D] = \frac{1}{k-1} \sum_{i \in g} (Y_i(1) - \mu_g(1))^2 =: S_{1,g}^2,$$

Likewise, $E[s_{0,g}^2 | \mathcal{H}_n^D] = S_{0,g}^2$. Hence $E[\hat{V}_{\text{SATE}} | \mathcal{H}_n^D] = k E[\hat{P}_N^2 | \mathcal{H}_n^D] = k \bar{P}_N^2$ for $\bar{P}_N^2 = |\mathcal{G}_n|^{-1} \sum_g (S_{1,g}^2/a + S_{0,g}^2/(k-a))$.

Design variance. Since $E[H_i | \mathcal{H}_n^D] = 0$, $E[\hat{\theta} | \mathcal{H}_n^D] = |\mathcal{G}_n|^{-1} \sum_g \theta_g = \text{SATE}$. The groups are conditionally independent given \mathcal{H}_n^D , and Lemma D.13 gives $\text{Var}(H_i | \mathcal{H}_n^D) = [p(1-p)]^{-1}$ and $\text{Cov}(H_i, H_j | \mathcal{H}_n^D) = -[(k-1)p(1-p)]^{-1}$ for $i \neq j \in g$, so for variance $S_{\bar{Y},g}^2 = \frac{1}{k-1} \sum_{i \in g} (\bar{Y}_i - \frac{1}{k} \sum_{j \in g} \bar{Y}_j)^2$, $\text{Var}(\hat{\theta}_g | \mathcal{H}_n^D) = k^{-2} \sum_{i,j \in g} \bar{Y}_i \bar{Y}_j \text{Cov}(H_i, H_j | \mathcal{H}_n^D) = \frac{S_{\bar{Y},g}^2}{k p(1-p)}$. The within-group identity $\bar{Y}_i - k^{-1} \sum_j \bar{Y}_j = (1-p)(Y_i(1) - \mu_g(1)) + p(Y_i(0) - \mu_g(0))$, combined with $S_{\tau,g}^2 = S_{1,g}^2 + S_{0,g}^2 - 2S_{10,g}^2$ for within-group covariance $S_{10,g}^2$, expands to

$$\frac{S_{\bar{Y},g}^2}{p(1-p)} = \frac{S_{1,g}^2}{p} + \frac{S_{0,g}^2}{1-p} - S_{\tau,g}^2 = k \left(\frac{S_{1,g}^2}{a} + \frac{S_{0,g}^2}{k-a} \right) - S_{\tau,g}^2,$$

where $S_{\tau,g}^2 = (k-1)^{-1} \sum_{i \in g} (\tau_i - \theta_g)^2$. Using $\text{Var}(\hat{\theta} | \mathcal{H}_n^D) = |\mathcal{G}_n|^{-2} \sum_g \text{Var}(\hat{\theta}_g | \mathcal{H}_n^D)$ and $n = k|\mathcal{G}_n|$,

$$n \text{Var}(\hat{\theta} | \mathcal{H}_n^D) = \frac{1}{|\mathcal{G}_n|} \sum_g \frac{S_{\bar{Y},g}^2}{p(1-p)} = k \bar{P}_N^2 - \bar{S}_\tau^2.$$

Combined with the previous display, $E[\hat{V}_{\text{SATE}} | \mathcal{H}_n^D] = k \bar{P}_N^2 = n \text{Var}(\hat{\theta} | \mathcal{H}_n^D) + \bar{S}_\tau^2$.

Bound. The finite-population decomposition $\sum_{i=1}^n (\tau_i - \text{SATE})^2 = \sum_g \sum_{i \in g} (\tau_i - \theta_g)^2 + k \sum_g (\theta_g - \text{SATE})^2$ and nonnegativity of the second term give $(k-1)|\mathcal{G}_n| \bar{S}_\tau^2 \leq (n-1)S_\tau^2$, hence $\bar{S}_\tau^2 \leq \frac{n-1}{(k-1)|\mathcal{G}_n|} S_\tau^2 < \frac{k}{k-1} S_\tau^2 \leq 2S_\tau^2$ for $k \geq 2$. \square

Under sampling-subordinate matching, \mathcal{G}_n^D is constructed by running the matching algorithm with assignment group size k separately in each sampling propensity stratum $S_l \cap \{i : T_i = 1\}$, $l = 1, \dots, L_n^T$, where L_n^T counts sampling, not assignment, propensity

levels. The next lemma verifies tight matching $F(\mathcal{G}_n^D) = o_p(1)$ in this setting, allowing L_n^T to grow.

Lemma D.21 (Subordinate Tight Matching). *Under sampling-subordinate matching with k constant, suppose $E[|\psi_2|^{\alpha_2}] < \infty$ for some $\alpha_2 > d_2 + 1$ and $L_n^T = o(n^{1-(d_2+1)/\alpha_2})$. Then $F(\mathcal{G}_n^D) = o_p(1)$.*

Proof. Write $\mathcal{G}_n^D = \cup_{l=1}^{L_n^T} \mathcal{G}_n^{D,l}$, where $\mathcal{G}_n^{D,l}$ is the algorithm's output on the index set $S_l \cap \{i : T_i = 1\}$ with one propensity level and group size k . The per-stratum bound from the proof of Theorem A.2 gives $F(\mathcal{G}_n^{D,l}) \leq M_n^2 \cdot n^{-1} C_{d_2} k^{2/(d_2+1)} (n_l^D)^{(d_2-1)/(d_2+1)}$, where $n_l^D = |S_l \cap \{i : T_i = 1\}|$. Summing over l and applying Jensen to $x \mapsto x^{(d_2-1)/(d_2+1)}$ with $\sum_l n_l^D \leq n$, as in the proof of Theorem A.2,

$$F(\mathcal{G}_n^D) \leq M_n^2 \cdot C_{d_2} k^{2/(d_2+1)} (L_n^T/n)^{2/(d_2+1)}.$$

Since $k = O(1)$ and $M_n^2 = o_p(n^{2/\alpha_2})$ by Lemma A.9, this is $o_p(n^{2/\alpha_2} (L_n^T/n)^{2/(d_2+1)}) = o_p(1)$ by the rate hypothesis $L_n^T = o(n^{1-(d_2+1)/\alpha_2})$. \square

The next lemma couples a generic AIPW summand to its influence-function linearization on a subsample $S \subseteq [n]$. Applied to Theorem B.2 below, the variable π_n is auxiliary randomness used to split the data into folds, $S = I_k$ is one such random fold, and \mathcal{A} is the σ -algebra generated by the out-of-fold data $(Z_j)_{j \notin S}$ used to fit \widehat{m}_d . The required independence $(Z_i)_{i \in S} \perp\!\!\!\perp \mathcal{A} | \pi_n$ then says that the units in the fold S are independent of the training set conditional on the fold partition randomness.

Lemma D.22 (AIPW Linearization). *Let $T_i \stackrel{\text{iid}}{\sim} \text{Bernoulli}(q_i)$ and $D_i \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p_i)$, and suppose Assumption B.1 holds. Let $Z_i = (\psi_i, Y_i(0), Y_i(1), T_i, D_i)$, let $\pi_n \perp\!\!\!\perp Z_{1:n}$, and let $S \subseteq [n]$ be $\sigma(\pi_n)$ -measurable. Let \mathcal{A} be a σ -algebra with $(Z_i)_{i \in S} \perp\!\!\!\perp \mathcal{A} | \pi_n$, and let $\widehat{m}_0, \widehat{m}_1$ be regression estimators that are $\sigma(\mathcal{A}, \pi_n)$ -measurable with $|\widehat{m}_d - m_d|_{2,\psi} = o_p(1)$ for $d \in \{0, 1\}$. Define the AIPW summand and the influence function*

$$R_i = \widehat{m}_1(\psi_i) - \widehat{m}_0(\psi_i) + \frac{T_i D_i (Y_i - \widehat{m}_1(\psi_i))}{q_i p_i} - \frac{T_i (1 - D_i) (Y_i - \widehat{m}_0(\psi_i))}{q_i (1 - p_i)},$$

$$\phi_i = c(\psi_i) + \frac{T_i D_i \epsilon_i^1}{q_i p_i} - \frac{T_i (1 - D_i) \epsilon_i^0}{q_i (1 - p_i)}, \quad \epsilon_i^d = Y_i(d) - m_d(\psi_i).$$

Then $n^{-1} \sum_{i \in S} R_i = n^{-1} \sum_{i \in S} \phi_i + o_p(n^{-1/2})$.

Proof. Write $\widehat{c} = \widehat{m}_1 - \widehat{m}_0$ for the estimated CATE and define $\widehat{b}(\psi) = (1 - p(\psi))\widehat{m}_1(\psi) + p(\psi)\widehat{m}_0(\psi)$ estimating the conditional mean $b(\psi) = E[\bar{Y}|\psi]$ for outcome level $\bar{Y} = (1 - p(\psi))Y(1) + p(\psi)Y(0)$. Since $p(\psi), 1 - p(\psi) \in (0, 1)$ pointwise, we have bounds $|\widehat{c} - c|_{2,\psi} \leq \sum_d |\widehat{m}_d - m_d|_{2,\psi} = o_p(1)$ and $|\widehat{b} - b|_{2,\psi} \leq \sum_d |\widehat{m}_d - m_d|_{2,\psi} = o_p(1)$ by Assumption B.1. We first rearrange the summand R_i . Substituting $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$ and $Y_i(d) = m_d(\psi_i) + \epsilon_i^d$ into the inverse propensity terms gives

$$\begin{aligned} \frac{T_i D_i Y_i}{q_i p_i} - \frac{T_i (1 - D_i) Y_i}{q_i (1 - p_i)} &= \frac{T_i}{q_i} \left[\frac{D_i m_1(\psi_i)}{p_i} - \frac{(1 - D_i) m_0(\psi_i)}{1 - p_i} \right] + \frac{T_i D_i \epsilon_i^1}{q_i p_i} - \frac{T_i (1 - D_i) \epsilon_i^0}{q_i (1 - p_i)} \\ &= \frac{T_i}{q_i} c(\psi_i) + \frac{T_i (D_i - p_i)}{q_i p_i (1 - p_i)} b(\psi_i) + \frac{T_i D_i \epsilon_i^1}{q_i p_i} - \frac{T_i (1 - D_i) \epsilon_i^0}{q_i (1 - p_i)}. \end{aligned}$$

The second equality applies the algebraic identity $D_i \alpha / p_i - (1 - D_i) \beta / (1 - p_i) = (\alpha - \beta) + (D_i - p_i)(\alpha / p_i + \beta / (1 - p_i))$ with $\alpha = m_1(\psi_i)$ and $\beta = m_0(\psi_i)$, together with $b(\psi_i) / (p_i (1 - p_i)) = m_1(\psi_i) / p_i + m_0(\psi_i) / (1 - p_i)$. Applying the identity with $\alpha = \widehat{m}_1(\psi_i)$ and $\beta = \widehat{m}_0(\psi_i)$ to the regression terms of R_i ,

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

Adding the displays, using $\frac{T_i}{q_i} c - \frac{T_i - q_i}{q_i} \widehat{c} = c + \frac{T_i - q_i}{q_i} (c - \widehat{c})$, gives $R_i = \phi_i + A_i + B_i$, where

$$A_i = \frac{T_i - q_i}{q_i} (c - \widehat{c})(\psi_i), \quad B_i = \frac{T_i (D_i - p_i)}{q_i p_i (1 - p_i)} (b - \widehat{b})(\psi_i).$$

It remains to show $\sqrt{nn^{-1}} \sum_{i \in S} A_i = o_p(1)$ and $\sqrt{nn^{-1}} \sum_{i \in S} B_i = o_p(1)$.

Consider $T_n = n^{-1} \sum_{i \in S} A_i$. Write $\Delta = c - \widehat{c}$, which is $\sigma(\mathcal{A}, \pi_n)$ -measurable by assumption. Since $Z_{1:n}$ are iid and $\pi_n \perp\!\!\!\perp Z_{1:n}$, the $Z_{1:n}$ remain iid given π_n . Since $(Z_i)_{i \in S} \perp\!\!\!\perp \mathcal{A} | \pi_n$, the variables $(Z_i)_{i \in S}$ are, conditionally on $\sigma(\mathcal{A}, \pi_n)$, also iid with their marginal law. For each $i \in S$, $E[A_i | \mathcal{A}, \pi_n] = 0$ because Δ is $\sigma(\mathcal{A}, \pi_n)$ -measurable and $E[T_i - q_i | \psi_i] = 0$, so $E[T_n | \mathcal{A}, \pi_n] = 0$. Then

$$\begin{aligned} \text{Var}(\sqrt{n} T_n | \mathcal{A}, \pi_n) &= n^{-1} \sum_{i \in S} \text{Var}(A_i | \mathcal{A}, \pi_n) \\ &= n^{-1} \sum_{i \in S} E \left[\frac{1 - q_i}{q_i} \Delta(\psi_i)^2 | \mathcal{A}, \pi_n \right] \leq \delta^{-1} |\Delta|_{2,\psi}^2 = o_p(1). \end{aligned}$$

The first equality follows by conditional independence of $(A_i)_{i \in S}$ given $\sigma(\mathcal{A}, \pi_n)$. The second equality follows since Δ is $\sigma(\mathcal{A}, \pi_n)$ -measurable and $E[(T_i - q_i)^2 | \psi_i] = q_i(1 - q_i)$. The inequality uses $|S| \leq n$ and $q_i > \delta$. Then by the conditional Markov inequality (Lemma D.1), $\sqrt{n}T_n = o_p(1)$. The argument for $n^{-1} \sum_{i \in S} B_i$ is identical, conditioning in addition on $(T_i)_{i \in S}$ and using $E[D_i - p_i | \psi_i, T_i] = 0$ together with $p_i \in (\delta, 1 - \delta)$. This gives $\sqrt{nn}^{-1} \sum_{i \in S} (A_i + B_i) = o_p(1)$, which proves the claim. \square

Lemma D.23 (Matching for Unions). *Let \mathcal{G}_n satisfy $n^{-1} \sum_{g \in \mathcal{G}_n} \sum_{i \in g} |\psi_i - \bar{\psi}_g|_2^2 = o_p(1)$ and $E[|\psi|_2^\alpha] < \infty$ for some $\alpha > \dim(\psi) + 1$. Suppose \mathcal{G}_n^ν matches group centroids $(\bar{\psi}_g)_{g \in \mathcal{G}_n}$ into pairs using either the algorithm of Section 3 or Derigs (1988). Then*

$$n^{-1} \sum_{u \in \mathcal{G}_n^\nu} \sum_{i \in u} |\psi_i - \bar{\psi}_u|_2^2 = o_p(1).$$

Proof of Lemma D.23. In what follows, let $k = \max_{g \in \mathcal{G}_n} |g|$. Note we have the identity $\sum_{i, j \in g} |\psi_i - \psi_j|_2^2 = 2|g| \sum_{i \in g} |\psi_i - \bar{\psi}_g|_2^2$. Each $u \in \mathcal{G}_n^\nu$ has the form $g \cup \nu(g)$ for $g \in \mathcal{G}_n$. Indexing by g double-counts each union, so

$$n^{-1} \sum_{u \in \mathcal{G}_n^\nu} \sum_{i \in u} |\psi_i - \bar{\psi}_u|_2^2 \leq \frac{1}{2n} \sum_{u \in \mathcal{G}_n^\nu} \sum_{i, j \in u} |\psi_i - \psi_j|_2^2 = \frac{1}{4n} \sum_{g \in \mathcal{G}_n} \sum_{i, j \in g \cup \nu(g)} |\psi_i - \psi_j|_2^2.$$

Splitting the inner sum into within-half and cross-half pairs and using the bijection $g \mapsto \nu(g)$ to combine the two within-half contributions,

$$\frac{1}{n} \sum_{u \in \mathcal{G}_n^\nu} \sum_{i, j \in u} |\psi_i - \psi_j|_2^2 = \frac{1}{n} \sum_{g \in \mathcal{G}_n} \sum_{i, j \in g} |\psi_i - \psi_j|_2^2 + \frac{1}{n} \sum_{g \in \mathcal{G}_n} \sum_{\substack{i \in g \\ j \in \nu(g)}} |\psi_i - \psi_j|_2^2.$$

The first term is $o_p(1)$ by the tight matching assumption for \mathcal{G}_n . For the cross term, by Jensen's inequality $|\psi_i - \psi_j|_2^2 \leq 3|\psi_i - \bar{\psi}_g|_2^2 + 3|\bar{\psi}_g - \bar{\psi}_{\nu(g)}|_2^2 + 3|\bar{\psi}_{\nu(g)} - \psi_j|_2^2$. Then

$$\begin{aligned} \frac{1}{n} \sum_{g \in \mathcal{G}_n} \sum_{\substack{i \in g \\ j \in \nu(g)}} |\psi_i - \psi_j|_2^2 &\leq \frac{3k}{n} \sum_{g \in \mathcal{G}_n} \sum_{i \in g} |\psi_i - \bar{\psi}_g|_2^2 + \frac{3k^2}{n} \sum_{g \in \mathcal{G}_n} |\bar{\psi}_g - \bar{\psi}_{\nu(g)}|_2^2 \\ &\quad + \frac{3k}{n} \sum_{g \in \mathcal{G}_n} \sum_{j \in \nu(g)} |\psi_j - \bar{\psi}_{\nu(g)}|_2^2. \end{aligned}$$

The first and third terms are $o_p(1)$ by the tight matching assumption for \mathcal{G}_n . It remains to handle the second term. Since ν partitions the centroids $\{\bar{\psi}_g\}_{g \in \mathcal{G}_n}$ into matched pairs

using either the algorithm in Section 3 or the optimal pairs, Theorem A.3 guarantees that $n^{-1} \sum_{g \in \mathcal{G}_n} |\bar{\psi}_g - \bar{\psi}_{\nu(g)}|_2^2 \leq M_n^2 \cdot O(n^{-2/(\dim(\psi)+1)})$ for $M_n^2 \equiv \max_{g \in \mathcal{G}_n} |\bar{\psi}_g|_2^2$. For M_n^2 , by Jensen $|\bar{\psi}_g|_2 \leq |g|^{-1} \sum_{i \in g} |\psi_i|_2 \leq \max_{i \in g} |\psi_i|_2$, so $M_n \leq \max_{i \in [n]} |\psi_i|_2 = o_p(n^{1/\alpha})$ by Lemma D.9 under $E[|\psi|_2^\alpha] < \infty$. Hence $M_n^2 = o_p(n^{2/\alpha})$, and the second term is $o_p(n^{2/\alpha-2/(\dim(\psi)+1)}) = o_p(1)$ since $\alpha > \dim(\psi) + 1$. \square

Lemma D.24 (Coupling for Inference). *Under the conditions of Lemma C.10, we have $\hat{v}_1 - E[\hat{v}_1 | \mathcal{H}_n^D] = o_p(1)$ and $\hat{v}_2 - E[\hat{v}_2 | \mathcal{H}_n^D] = o_p(1)$ for $\mathcal{H}_n^D = \sigma(W_{1:n}, T_{1:n}, \pi_n)$.*

Proof of Lemma D.24. First consider \hat{v}_1 . Set $u_g = (k^2/q_g^2)(\hat{\theta}_g^2 - E[\hat{\theta}_g^2 | \mathcal{H}_n^D])$, so that $\hat{v}_1 - E[\hat{v}_1 | \mathcal{H}_n^D] = n^{-1} \sum_{g \in \mathcal{G}_n} u_g$. We verify the conditions of Lemma D.8 with the filtration \mathcal{H}_n^D . The partition \mathcal{G}_n is \mathcal{H}_n^D -measurable, and $E[u_g | \mathcal{H}_n^D] = 0$ by construction. As noted in the proof of Lemma C.10, $\mathcal{F}_n^D \subseteq \mathcal{H}_n^D$ and $\mathcal{H}_n^D \perp\!\!\!\perp \eta$ for the assignment randomness η of Definition C.1. Since $\hat{\theta}_g$ is a function of $(D_i)_{i \in g}$ and the \mathcal{H}_n^D -measurable potential outcomes $(Y_i(0), Y_i(1))_{i \in g}$, the term $u_g = (k^2/q_g^2)(\hat{\theta}_g^2 - E[\hat{\theta}_g^2 | \mathcal{H}_n^D])$ has the form $\phi((D_i)_{i \in g}, X, g)$ for a deterministic function ϕ and the \mathcal{H}_n^D -measurable variable $X = (W_{1:n}, (E[\hat{\theta}_{g'}^2 | \mathcal{H}_n^D])_{g' \in \mathcal{G}_n})$. By Lemma D.11, the $(u_g)_{g \in \mathcal{G}_n}$ are jointly independent conditional on \mathcal{H}_n^D ; in particular $u_g \perp\!\!\!\perp u_{g'} | \mathcal{H}_n^D$ for $g \neq g'$.

For $i \in g$, $T_i = 1$ and $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$, so $|\hat{\theta}_g| = |a^{-1} \sum_{i \in g} D_i Y_i(1) - (k - a)^{-1} \sum_{i \in g} (1 - D_i) Y_i(0)| \leq \max_{i \in g} |Y_i(1)| + \max_{i \in g} |Y_i(0)|$. Squaring and using $(p + q)^2 \leq 2(p^2 + q^2)$ together with $\max_{i \in g} Y_i(d)^2 \leq \sum_{i \in g} Y_i(d)^2$ gives $\hat{\theta}_g^2 \leq 2 \sum_{i \in g} (Y_i(1)^2 + Y_i(0)^2)$. The right side is \mathcal{H}_n^D -measurable, so the bound is preserved under conditional expectation: $E[\hat{\theta}_g^2 | \mathcal{H}_n^D] \leq 2 \sum_{i \in g} (Y_i(1)^2 + Y_i(0)^2)$. Hence $|u_g| \leq \tilde{u}_g := (4k^2/q_g^2) \sum_{i \in g} (Y_i(1)^2 + Y_i(0)^2)$, which is \mathcal{H}_n^D -measurable, and since $\mathbf{1}(|u_g| > c_n) \leq \mathbf{1}(\tilde{u}_g > c_n)$, we have $E[|u_g| \mathbf{1}(|u_g| > c_n) | \mathcal{H}_n^D] \leq \tilde{u}_g \mathbf{1}(\tilde{u}_g > c_n)$.

Note that for positive constants $(a_l)_{l=1}^m$, $\sum_l a_l \mathbf{1}(\sum_l a_l > c) \leq m \sum_l a_l \mathbf{1}(a_l > c/m)$. Viewing \tilde{u}_g as a sum of $2k$ terms indexed by $(i, d) \in g \times \{0, 1\}$, each equal to $4k^2 Y_i(d)^2 / q_g^2$, and applying the inequality with $m = 2k$, we get

$$\tilde{u}_g \mathbf{1}(\tilde{u}_g > c_n) \leq (8k^3/q_g^2) \sum_{i \in g, d} Y_i(d)^2 \mathbf{1}(Y_i(d)^2/q_g^2 > c_n/(8k^3)).$$

Since \mathcal{G}_n partitions $\{i : T_i = 1\}$ and $q_g = q(\psi_i)$ for $i \in g$, setting $Z_i(d) = Y_i(d)^2/q(\psi_i)^2$,

$$\frac{1}{n} \sum_{g \in \mathcal{G}_n} \tilde{u}_g \mathbf{1}(\tilde{u}_g > c_n) \leq \frac{8k^3}{n} \sum_{i=1}^n T_i \sum_{d \in \{0,1\}} Z_i(d) \mathbf{1}(Z_i(d) > c_n/(8k^3)). \quad (\text{D.12})$$

By assumption $q(\psi_i) \geq \delta > 0$, so that $E[Z_i(d)] \leq E[Y_i(d)^2]/\delta^2 < \infty$. Then by dominated convergence $E[Z_i(d)\mathbf{1}(Z_i(d) > c_n/(8k^3))] \rightarrow 0$ as $c_n \rightarrow \infty$. Using $T_i \leq 1$ and iid sampling of W_i , the right side of (D.12) has expectation at most $8k^3 \sum_d E[Z_i(d)\mathbf{1}(Z_i(d) > c_n/(8k^3))] \rightarrow 0$, hence is $o_p(1)$ by Markov for any $c_n \rightarrow \infty$. Chaining through the conditional truncation bound $E[|u_g|\mathbf{1}(|u_g| > c_n)|\mathcal{H}_n^D] \leq \tilde{u}_g\mathbf{1}(\tilde{u}_g > c_n)$ established above,

$$\frac{1}{n} \sum_{g \in \mathcal{G}_n} E[|u_g|\mathbf{1}(|u_g| > c_n) | \mathcal{H}_n^D] \leq \frac{1}{n} \sum_{g \in \mathcal{G}_n} \tilde{u}_g\mathbf{1}(\tilde{u}_g > c_n) = o_p(1).$$

Choosing $c_n = n^{1/4}$, Lemma D.8 gives $n^{-1} \sum_{g \in \mathcal{G}_n} u_g = o_p(1)$, i.e., $\hat{v}_1 - E[\hat{v}_1|\mathcal{H}_n^D] = o_p(1)$. The argument for \hat{v}_2 is similar, with $u_u = (2k(k-q_u)/q_u^2)(\hat{\theta}_{g_1(u)}\hat{\theta}_{g_2(u)} - E[\hat{\theta}_{g_1(u)}\hat{\theta}_{g_2(u)}|\mathcal{H}_n^D])$ for $u \in \mathcal{G}_n^\nu$ in place of u_g . Each u_u is a function of $((D_i)_{i \in g})_{g \in \{g_1(u), g_2(u)\}}$ and \mathcal{H}_n^D -measurable quantities, and since distinct unions are built from disjoint groups of \mathcal{G}_n , the joint conditional independence of $((D_i)_{i \in g})_{g \in \mathcal{G}_n}$ given \mathcal{H}_n^D from Lemma D.11 yields $(u_u)_{u \in \mathcal{G}_n^\nu}$ jointly conditionally independent given \mathcal{H}_n^D . The conditional truncation bound is verified as for \hat{v}_1 , using $|\hat{\theta}_{g_1(u)}\hat{\theta}_{g_2(u)}| \leq \frac{1}{2}(\hat{\theta}_{g_1(u)}^2 + \hat{\theta}_{g_2(u)}^2)$. \square

Rounding Procedure. Next, we discuss a specific rounding procedure that can be used to satisfy the assumptions on $\hat{q}_n(\psi)$ for the sampling design $T_{1:n} \sim \text{Loc}(\psi, \hat{q}_n(\psi))$. For an estimator $\hat{q}(\psi)$ with $\delta \leq \hat{q}(\psi) \leq 1$, define the discretization $\hat{q}_n(\psi) = \lfloor k_n \hat{q}(\psi) \rfloor / k_n$, where $\lfloor x \rfloor$ denotes the nearest integer to x . Then $\hat{q}_n(\psi)$ takes values in the rational set $\{a/k_n : a = 0, 1, \dots, k_n\}$, with at most $L_n \leq k_n + 1$ distinct propensity levels and common group size $\bar{k}_n = \max_{l \in [L_n]} k_l = k_n$. The following lemma shows that this discretization scheme satisfies part (iv) of Assumption 5.2.

Lemma D.25 (Discretization by Rounding). *Under the discretization scheme above, $E_n[(\hat{q}_n(\psi_i) - \hat{q}(\psi_i))^2] \leq 1/(4k_n^2)$ and $\bar{k}_n L_n = O(k_n^2)$. In particular, if $k_n \rightarrow \infty$ and $k_n = o(n^{(1-(\dim(\psi)+1)/\alpha)/2} \wedge n^{1/4})$, then part (iv) of Assumption 5.2 holds.*

Proof. By construction $\hat{q}_n(\psi) = a(\psi)/k_n$ with $a(\psi) = \lfloor k_n \hat{q}(\psi) \rfloor \in \{0, 1, \dots, k_n\}$, since $\hat{q}(\psi) \in [\delta, 1] \subseteq [0, 1]$. Hence $L_n \leq k_n + 1$ and $\bar{k}_n = k_n$. Rounding to the nearest integer satisfies $|\lfloor x \rfloor - x| \leq 1/2$, so $\sup_\psi |\hat{q}_n(\psi) - \hat{q}(\psi)| \leq 1/(2k_n)$, and squaring gives $E_n[(\hat{q}_n(\psi_i) - \hat{q}(\psi_i))^2] \leq 1/(4k_n^2)$. Thus $\bar{k}_n L_n \leq k_n(k_n + 1) = O(k_n^2)$. For the final claim, $k_n = o(n^{(1-(\dim(\psi)+1)/\alpha)/2})$ gives $\bar{k}_n L_n = o(n^{1-(\dim(\psi)+1)/\alpha})$, while $\bar{k}_n = k_n = o(n^{1/4})$, so part (iv) of Assumption 5.2 holds. \square

D.2 Targeting the SATE

Recall the motivating example in Section 2, where [Chioda et al. \(2026\)](#) sample $n_T \approx 4,402$ Ugandan job applicants into an experiment from a pool of $n = 7,431$ applicants to a particular job posting. The superpopulation interpretation embeds these n applicants into a larger distribution P of Ugandan job-seekers from which they are imagined to be drawn iid, with $\text{ATE} = E_P[Y(1) - Y(0)]$.

However, the sampling process into the eligible pool is not controlled by the researcher, and it's unclear exactly who the superpopulation P represents in this case. Because of this, the researcher may instead wish to target $\text{SATE} = E_n[Y_i(1) - Y_i(0)]$ in the realized eligible population that we physically sample from.

What is the interpretation of the superpopulation model in this case? To explain this, we note that our work in [Corollary C.7](#) in fact implies a stronger conditional weak convergence statement $\sqrt{n_T}(\hat{\theta} - \text{SATE})|W_{1:n} \Rightarrow \mathcal{N}(0, V_{\text{SATE}})$ in the sense that

$$\sup_{x \in \mathbb{R}} |P(\sqrt{n_T}(\hat{\theta} - \text{SATE}) \leq x \mid W_{1:n}) - \Phi(x/\sqrt{V_{\text{SATE}}})| \xrightarrow{P} 0.$$

This is a quasi design-based result, since asymptotic normality is delivered by design randomness $T_{1:n} \sim \text{Loc}(\psi_1, q)$ and $D_{1:n} \sim \text{Loc}(\psi_2, p)$ alone. The iid model for the data further constrains the form of the limiting variance, showing that it converges in probability to the simple analytical expression presented in the text.

Extension to a fully design-based model is beyond the scope of this paper. See [Bai et al. \(2026\)](#) for a more robust purely design-based model of finely stratified assignment and a related theoretical discussion.

D.3 Empirical Application Details

This section provides descriptions of each paper and implementation details for our empirical application in [Section 7](#). The stratification variables ψ are used in their raw scale at the loader stage; the matching algorithm in [Section 3](#) applies per-coordinate normalization internally, and the imputation regressions ([Appendix B.3](#)) apply standardization where the chosen regression family requires it.

- (1) [Abebe et al. \(2021\)](#) estimates the effect of an application incentive on the ability of applicants for clerical employment in Ethiopia. We let Y be the authors' index of cognitive ability, D be the application incentive, and $\psi(X)$ be gender, age, work

experience in years, self-reported gpa and previous wage, and indicators for being born in Addis Ababa, speaking Amharic, and studying engineering.

- (2) [Banerjee et al. \(2025\)](#) estimates the effect of various strategies to promote child-vaccination on the number of children completing the full vaccination sequence. We let Y_i be the number of children receiving the measles shot over the full trial period in village i and D_i whether the village received the SMS reminder intervention. We let ψ include village population, the proportion of individuals in the baseline survey of that village who were in a “scheduled caste”, a “backward class”, who received nursery education or less, baseline proportions of vaccine completions in two age cohorts, and the proportion who had a vaccine card.
- (3) [Baysan \(2022\)](#) estimates the effect of political information campaigns about concentration of executive power in Turkey on voter polarization. Data is at the ballot box level, while assignment to information campaigns is at the neighborhood level. We let Y_i be the vote share of “No” votes, averaged over ballot boxes in a neighborhood, in the 2017 referendum. We let D be whether the village was exposed to any information campaign, and ψ include the village-level average vote share for the CHP in the 2015 election, a measure of turnout, and the number of ballots collected.
- (4) [Casey et al. \(2021\)](#) estimates the effect of increased information given to political parties about voter preferences over candidates on whether the most voter-preferred candidate was selected to run by the party. We let D be assignment to the treatment package at the party-constituency level, Y whether the most preferred candidate was selected, and let ψ include competitiveness of the race, candidate professional qualifications index, and candidate public service motivation index.
- (5) [DellaVigna et al. \(2022\)](#) estimates the effect of employer gifts and other “social preference” related interventions on worker productivity. We study the first experiment in the paper and let $D = 1$ if the worker received either a positive gift or an in-kind gift and $D = 0$ for no gift or a negative gift. We let Y be worker output in the last working period. We let ψ be total productivity during the first 8 periods of the trial (excluding the final gift period), gender, and age.
- (6) [Domurat et al. \(2021\)](#) estimates the effect of informational interventions about Covered California insurance policies on takeup of insurance. We let $D = 0$ for the control group and $D = 1$ if assigned to any of the letter campaigns in arms 3,

4, or 5. Y is an indicator for insurance takeup. ψ includes a measure of household income, mean age, household subsidy size, and an indicator for being Latino.

- (7) [Finkelstein et al. \(2012\)](#) reports the effect of winning the 2008 Oregon Medicaid lottery on various health and public service utilization outcomes. We use data from wave one, restricting to single person households. We estimate an ITT effect with Y the number of emergency department (ED) visits in the post-period and D an indicator for winning the lottery. We let ψ include gender, age, any visits to the ED in the pre-period, number of visits in the pre-period, total SNAP benefits in the pre-period, and indicators for ever being on SNAP or having a chronic condition.
- (8) [Hussam et al. \(2022\)](#) estimates the value of employment on measures of psychosocial wellbeing. Treatment assignment is at the block level in the refugee camp, with a sample of five individuals chosen in each block. We aggregate by taking the mean of outcomes and covariates in each block. We let $D = 0$ if the block was randomized to cash only and $D = 1$ if it was randomized to employment. Y is the endline mental health index, and ψ includes the baseline mental health index, average gender (in $[0, 1]$), proportion who had a family member killed, and a measure of sociability.
- (9) [Lowe \(2021\)](#) estimates the effect of collaborative and adversarial intergroup contact on cross-caste friendships using randomization to different teams in a cricket league in India. We let Y be number of other caste friends at endline, D be whether the person is assigned to a mixed caste team, and ψ include number of other caste friends at baseline, age, and a measure of cricket ability.

D.4 Empirical Results for the SATE

This section presents our empirical results for SATE estimation and inference, using the variance estimator $\widehat{V}_{\text{SATE}}$ from Section 6 on the same nine DGPs and seven designs as in Section 7. Table 2 reports the standard deviation of $\widehat{\theta} - \text{SATE}$, percent change in CI length relative to complete randomization, empirical coverage of SATE, and the percent change in confidence interval length relative to the corresponding ATE confidence interval \widehat{C} in Table 1.

Coverage of SATE is close to nominal across all 63 (paper, design) cells, in line with the asymptotically valid inference guarantee from Theorem 6.1. The percent change in confidence interval length relative to the ATE confidence interval is uni-

| | Design, Paper | Abe. | Ban. | Bay. | Cas. | Del. | Dom. | Fin. | Hus. | Low. |
|----------------------|---------------|------|------|------|------|------|------|------|------|------|
| % Δ SD | CR | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | CR, Loc | -6 | -45 | -49 | -2 | -41 | 1 | -9 | -3 | -12 |
| | Loc | -7 | -42 | -49 | 3 | -42 | 2 | -7 | -9 | -9 |
| | Hom. | -16 | -43 | -49 | -8 | -41 | -4 | -7 | -14 | -12 |
| | Pilot S | -14 | -51 | -46 | -8 | -43 | 3 | -22 | -15 | -15 |
| | Pilot L | -18 | -52 | -47 | -6 | -43 | -2 | -24 | -10 | -19 |
| | Obs | -16 | -54 | -48 | -3 | -44 | -5 | -25 | -13 | -18 |
| % Δ CI | CR | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | CR, Loc | -6 | -43 | -46 | -1 | -38 | 0 | -8 | -6 | -10 |
| | Loc | -6 | -41 | -46 | 0 | -38 | 0 | -8 | -6 | -9 |
| | Hom. | -13 | -41 | -46 | -6 | -40 | -3 | -9 | -9 | -9 |
| | Pilot S | -13 | -49 | -42 | -7 | -39 | 0 | -21 | -10 | -16 |
| | Pilot L | -13 | -50 | -44 | -6 | -40 | -3 | -21 | -10 | -18 |
| | Obs | -14 | -51 | -44 | -6 | -41 | -2 | -21 | -10 | -17 |
| Cover | CR | 0.96 | 0.95 | 0.94 | 0.96 | 0.95 | 0.97 | 0.97 | 0.95 | 0.96 |
| | CR, Loc | 0.96 | 0.96 | 0.95 | 0.97 | 0.96 | 0.96 | 0.97 | 0.94 | 0.97 |
| | Loc | 0.97 | 0.96 | 0.95 | 0.95 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 |
| | Hom. | 0.96 | 0.96 | 0.95 | 0.96 | 0.95 | 0.96 | 0.96 | 0.96 | 0.96 |
| | Pilot S | 0.96 | 0.96 | 0.95 | 0.96 | 0.97 | 0.96 | 0.96 | 0.96 | 0.96 |
| | Pilot L | 0.97 | 0.96 | 0.95 | 0.95 | 0.97 | 0.96 | 0.97 | 0.95 | 0.97 |
| | Obs | 0.96 | 0.97 | 0.96 | 0.95 | 0.96 | 0.96 | 0.97 | 0.95 | 0.96 |
| % Δ CI vs ATE | CR | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | CR, Loc | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 0 |
| | Loc | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 0 |
| | Hom. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 0 |
| | Pilot S | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 0 |
| | Pilot L | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 0 |
| | Obs | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 0 |
| n_T | 1451 | 903 | 550 | 91 | 446 | 1000 | 1903 | 116 | 770 | |
| $\dim(\psi)$ | 8 | 7 | 3 | 3 | 3 | 4 | 7 | 4 | 3 | |

Table 2: Empirical Results (SATE).

formly non-positive. By Corollary 4.6, targeting the SATE lowers the asymptotic variance to $V_{\text{SATE}} = V - q \text{Var}(\tau)$. The realized reduction is smaller, since inference on the SATE is conservative. By Theorem 6.1, the estimators satisfy $\widehat{V}_{\text{SATE}} \xrightarrow{p} V_{\text{SATE}} + E[q(\psi)]E[\text{Var}(\tau|\psi)]$ and $\widehat{V} \xrightarrow{p} V$, so their probability limits differ by $q \text{Var}(c(\psi))$ using the identity $\text{Var}(\tau) = \text{Var}(c(\psi)) + E[\text{Var}(\tau|\psi)]$. The gap is largest in papers with the most heterogeneity in $c(\psi)$, where researchers stand to gain the most precision from targeting the SATE over the ATE.