Designing Representative and Balanced Experiments by Local Randomization

Max Cytrynbaum

November 11, 2021

Abstract

This paper studies treatment effect estimation in a novel two-stage model of experimentation. In the first stage, using baseline covariates, the researcher selects units to participate in the experiment from a sample of eligible units. Next, they assign each selected unit to one of two treatment arms. We relate estimator efficiency to representative selection of participants and balanced assignment of treatments. We define a new family of local randomization procedures, which can be used for both selection and assignment. This family nests stratified block randomization and matched pairs, the most commonly used designs in practice in development economics, but also produces many useful new designs, embedding them in a unified framework. When used to select representative units into the experiment, local randomization boosts effective sample size, making estimators behave as if they were estimated using a larger experiment. When used for treatment assignment, local randomization does model-free non-parametric regression adjustment by design. We give novel asymptotically exact inference methods for locally randomized selection and assignment, allowing experimenters to report smaller confidence intervals if they designed a representative experiment. We apply our methods to the setting of two-wave design, where the researcher has access to a pilot study when designing the main experiment. We use local randomization methods to give the first fully efficient solution to this problem.

*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 Suhas Vijaykumar.

†Department of Economics, MIT. Correspondence: mcytryn@mit.edu
1 Introduction

This paper studies experimental designs for estimating the average treatment effect (ATE) in a novel two-stage model of experimentation. In the first stage, the researcher selects units to participate in the experiment from a random sample of eligible units. In the second stage, they assign each selected unit to one of two treatment arms. We focus on estimation efficiency, showing that selecting representative units into the experiment and “locally” balancing treatment assignments gives significant variance reduction. We implement both stages of the design, selection and assignment, using a new family of local randomization procedures. When used to assign treatments, this family nests commonly used designs such as stratified block randomization and matched pairs, but also produces many useful new designs. We show that, by selecting units into the trial using local randomization, estimators of the ATE inherit some of the efficiency of a fictitious larger experiment, without having to actually run this experiment (made precise below).

In practice, experimenters often have (at least) two choices to make when designing an experiment: (1) which units should be selected\(^1\) to participate in the experiment, and (2) how should treatments be assigned among the selected units? For example, in a recent paper Breza et al. (2021) consider the effect of low vs. high intensity of facebook ads discouraging travel on county-level covid-19 infection rates. Given finite resources, designing the experiment requires selecting a small set of counties in which to run ads and collect outcome data. The selected counties are then randomly assigned to either low or high intensity of ads. Similarly, Abaluck et al. (2021) consider the effect of mask promotion on village-level infection rates in Bangladesh. From a sample of 1000 villages, they first randomly select 600 to be included in the experiment. Next, the selected villages are randomly assigned to various interventions that promote mask usage.\(^2\) Can we do better than completely random selection in these examples? This paper defines novel representative selection designs, admitting asymptotically exact inference, which can significantly reduce the sample size required to estimate causal effects.

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

Selecting a representative sample of units into the experiment can significantly reduce estimator variance, while preserving finite-sample unbiasedness. To see why, consider designing an experiment to estimate the effect of distributing surgical masks vs. cloth masks on covid infection rates, simplifying Abaluck et al. (2021). Suppose that average resident age predicts covid infections, with older villages more likely to have higher in-

\(^1\)In the empirical literature, selection is also known as the sampling frame in the empirical literature.

\(^2\)Alternatively, consider a job search assistance program as in Caria et al. (2021). Suppose 5000 job-seekers apply for the program, which includes various employment interventions, but the budget only allows for 1000 participants. Who should we let into the experiment?
fection rates. If selection is completely random, the selected villages may be younger or older than the full sample of eligible villages just by random chance. If we randomly select mostly old villages, our treatment effect estimate will be biased up (ex-post), and conversely it will be biased down if we select villages with only healthy young residents. These chance covariate imbalances increase estimator variance (ex-ante), wasting sample size and experimental resources.

We solve this problem with new designs that finely balance covariates between the selected and not-selected villages, producing a selected sample that is representative of the larger population of eligible villages. There are generally multiple dimensions of heterogeneity along which representativeness can be measured, such as village demographics, baseline covid levels, housing density and so on. Our analysis shows that the largest increase in precision comes from representative selection with respect to covariates that are predictive of treatment effect heterogeneity. The optimal selection design balances the conditional average treatment effect (CATE) itself. During treatment assignment, the second stage of the design, experimenters can further reduce estimator variance by finely balancing covariates between the different treatment arms. We show that the optimal assignment design balances those covariates that are most predictive of experimental outcomes, rather than treatment effect heterogeneity.

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

We give novel algorithms and rate analysis for the combinatorial problem of constructing maximally homogeneous groups of units (Section 7). By group homogeneity, we can think of each group as consisting of units of a certain “covariate type.” Selecting exactly 3 out of 10 units from each group ensures that each group, and therefore each type of unit, is well-represented in the experiment. By contrast, under completely random selection we may not select any units from certain groups, while other groups may be over-represented in the experiment by random chance. Such fluctuations increase estimator variance.

Summarizing the procedure, during selection we solve an optimization problem to match eligible units into homogeneous groups, randomly selecting a fixed proportion of the units in each group to participate in the experiment. During treatment assignment, we repeat the process among the selected units, again forming homogeneous groups of selected units.

---

3Defined formally in Section 2.3 below.
and randomly assigning a fixed proportion of each group to one of two treatment arms. This paper studies efficient treatment effect estimation under the local randomization designs outlined above. We make several methodological contributions to the literature on experimental design.

Our first contribution is to define a family of novel covariate-adaptive designs for treatment assignment. The family of local randomizations contains stratified block randomization and matched pairs, the most commonly used designs in practice in development economics, but also produces many useful new designs. The general version of our method allows varying treatment proportions, extending the principle of matched-pairs designs to arbitrary propensity scores \( p(x) \neq 1/2 \). For example, due to welfare considerations we may want to assign older villages to surgical masks with probability \( 2/3 \), while younger villages instead receive surgical masks with probability \( 1/3 \). Using local randomization in this setting gives a “matched triples” design with varying treatment proportions. Other novel designs are given in the many examples throughout the paper.

Our second contribution concerns selection into the experiment. This paper gives a novel formalization of covariate-adaptive selection of experiment participants, showing how to estimate treatment effects more precisely by constructing a “representative” experiment. We show that optimal selection requires finely balancing the covariates most predictive of treatment effect heterogeneity. Effectively, researchers should ensure that the different “types” of responses to treatment are well-represented among the selected experimental participants. Local randomization defines a large family of covariate-adaptive selection methods, giving a practical implementation of this idea.

In our third contribution, we give a new central limit theorem for inverse propensity weighting (IPW) estimators of the ATE in experiments with locally-randomized selection and treatment assignment. Our results show that local randomization does non-parametric regression adjustment by design, increasing estimator precision without the need to actually specify or estimate a regression model ex-post. When local randomization is used for selection, this non-parametric control attenuates the variance due to treatment effect heterogeneity, to the extent that the covariates used to form homogeneous groups explain this heterogeneity. In particular, we show that the optimal design locally randomizes with respect to the conditional average treatment effect (CATE), minimizing the variance due to selection. For treatment assignment, the optimal design locally randomizes with respect to the balance function, a weighted sum of conditional expectations defined below. The oracle design that locally randomizes with respect to (CATE, balance function) in both stages of the design achieves full efficiency.

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

---

4See Bruhn and McKenzie (2009) for evidence on the practice of experimental design in development economics.

5We consider both the ATE and finite-population fixed-regressor estimands, as in Abadie et al. (2014).
Our fourth contribution gives new methods for asymptotically exact inference on the ATE under locally randomized selection and assignment. In particular, we allow researchers to report smaller confidence intervals if they use the methods in this paper to design a representative experiment. Our construction allows plug-in asymptotically exact inference over the entire class of locally randomized designs defined below. As a consequence, we get new exact variance estimators for matched pairs and classical stratified block randomization. We also explore an interesting connection between our matching-based variance estimators and randomization inference.

In an important application, we study efficient design when the researcher has data from a pilot study. We use local randomization methods to construct the first fully efficient design in this setting, comparing with the recent approaches of Tabord-Meehan (2020) and Bai (2020). To achieve full efficiency, we use pilot data to estimate the optimal propensity score\footnote{For complete randomization, optimal constant treatment proportions are known as the Neyman allocation. We give a local randomization implementation of the conditional Neyman allocation.} for the experiment, which assigns units to each treatment arm in proportion to that arm’s conditional variance.\footnote{One can think of this as a design analogue of (optimal) feasible weighted least squares.} Intuitively, one can think of this as taking more measurements of the thing that is harder to measure. We give a balanced implementation of this propensity score by double stratification, first stratifying units by our pilot estimate of the optimal propensity score, then locally randomizing treatments within these propensity strata. Our analysis shows that this method globally minimizes the selection, assignment, and residual components of the asymptotic variance independently, while previous methods require tradeoffs between these different sources of variance.

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

\subsection{Related Literature}

There is a large literature on experimental design, see e.g. Athey and Imbens (2017) or Rosenberger and Lachin (2016) for surveys. See Bruhn and McKenzie (2009) for evidence on experimental design as used in practice in development economics. The literature may be divided by assumptions about timing of the experiment. Our first set of results assumes the classical timing: (1) all baseline covariates are observed (2) units are selected into the trial and assigned to interventions (3) all outcomes are observed. We also allow additional covariates to be observed after selection: see Example 3.3 below. Alternatively, some papers assume sequential timing, with units arriving one-by-one and treatment decisions made “on the spot”, as in Efron (1971) or Kapelner and Krieger (2014).
results on design with a pilot fit into the *adaptive* design literature, with some outcomes observed during the treatment process. See Hu and Rosenberger (2006) for an overview.

With respect to solution concept, some papers emphasize “robustness”, searching for minimax designs over a class of DGP’s satisfying certain smoothness assumptions. For example, see Kallus (2017) or the discussion and references in Harshaw et al. (2021). Alternatively, one may use a decision theoretic framework to characterize the optimal design, as in Kasy (2016). By contrast, this paper defines a new family of randomization procedures, studying their asymptotic efficiency pointwise in (design, DGP), over a class of DGP’s satisfying certain regularity conditions.

Our selection model is related to the literature on survey sampling, see e.g. Cochran (1977). The designs in this paper are examples of blocking, as in Fisher (1926), Higgins et al. (2015), Fogarty (2018), and Bai (2020). The family we study contains stratified block randomization (with fixed strata) and matched pairs, as analyzed in Bugni et al. (2018) and Bai et al. (2021). Rerandomization, studied in Morgan and Rubin (2012) and Li et al. (2018), is not contained in this family. Follow up work in Cytrynbaum (2021) studies rerandomizing local designs, e.g. giving results for rerandomized SBR and matched pairs, as well as rerandomized versions of all the other designs studied in this paper. The main result shows that local randomization generically dominates rerandomization (giving smaller asymptotic variance).

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

Our first sets of results is most related to Bai (2020), which we build on. Relative to this paper, we study two-stage designs, with both covariate-adaptive selection and treatment assignment. For treatment assignment alone, the designs in Bai (2020) are a special case of our method with (1) $\dim(\psi) = 1$ for the local function defined below (function of covariates we match on) and (2) constant propensity $p = a/k$. Allowing $\dim(\psi) > 1$ is necessary to implement the optimal design for joint selection and assignment, as well as its empirical counterparts. More generally, our method allows balancing with respect to more than just a single covariate.

For $\dim(\psi) > 1$, the algorithm used in Bai that constructs groups of units by sorting their $\psi(X_i)$ values is no longer feasible. We study combinatorial optimization procedures (graph-partitioning) to form groups in higher dimensions, and give new analysis of their statistical rates. Our designs allow propensity $p(x)$ non-constant. This innovation is required for a fully efficient solution to the two-wave design problem, which uses the pilot to estimate, and “locally” implement, an optimally varying propensity score (Neyman allocation). Finally, Bai (2020) allows inference only for the case $p = 1/2$, using the results of Bai et al. (2021) on matched pairs. Our work gives novel, generic inference methods that cover the full family of locally randomized designs, including joint covariate-adaptive selection and assignment.
Our solution to the problem of design with a pilot experiment follows work in Hahn et al. (2011), Tabord-Meehan (2020), and Bai (2020). More broadly, the two-wave setting is an example of response-adaptive design, as in Russo (2016) and Kasy and Sautmann (2021). A detailed comparison with Tabord-Meehan (2020) is given in Section 4.1.1 below. A comparison with the large pilot results of Bai (2020) is given in section 4.2.

The rest of this paper is organized as follows. In section 2 we formally describe our method and preview some of our main results. Section 3 gives our main asymptotic results, describes several useful new designs produced by our framework, and gives the optimal design. Section 4 gives our results on optimal design with a pilot experiment. Section 5 gives our inference methods. Section 6 gives our empirical results. Section 7 collects technical results on algorithms for constructing the local groups, including their asymptotic balancing rates. All proofs are given in the appendix.

2 Motivation and Description of Method

We start with some basic notation. Let $X_i$ denote baseline covariates, observed by the experimenter before choosing the experimental design. Let $Y_i(0), Y_i(1)$ be potential outcomes under interventions $d \in \{0, 1\}$. The data on eligible units $(X_i, Y_i(0), Y_i(1))_{i=1}^n$ is assumed to be drawn iid. Given all observed covariates $(X_i)_{i=1}^n$, the experimenter chooses which units to select into the experiment and how to assign treatments among the selected units. Define the selection variable $T_i \in \{0, 1\}$ with $T_i = 1$ if the unit is selected into the experiment, and $T_i = 0$ otherwise. Let $D_i \in \{0, 1\}$ denote treatment assignment. Since a unit’s outcome is observed if and only if it is selected, we may write

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

Define the conditional average treatment effect (CATE)

$$c(X_i) = E[Y_i(1) - Y_i(0) | X_i]$$

We are interested in the average treatment effect (ATE)

$$ATE = E[Y_i(1) - Y_i(0)] = E[c(X_i)]$$

2.1 Motivating Example

The first stage of the experimental design requires choosing selection variables $(T_i)^n_{i=1}$ using the observed covariates $(X_i)^n_{i=1}$. We begin by examining how selection affects the precision of our estimate of the ATE. To do this, we briefly consider a toy model of an experiment, in which $c(X_i)$ is directly observed for each selected unit $T_i = 1$. Since $E[c(X_i)] = ATE$, we can use a sample-mean estimator

$$\hat{\theta} = E_n[c(X_i) | T_i = 1] = \frac{\sum_{i=1}^n T_i c(X_i)}{\sum_{i=1}^n T_i}$$

---

8We also allow additional covariates to be observed after selection but before treatment assignment, see Example 3.3.
In the example from Abaluck et al. (2021) discussed above, the experimental design required selecting which Bangladeshi villages would participate in the experiment \((T_i = 1)\). The selected villages then receive one of two interventions: distribution of cloth or surgical masks \((D_i = 0, 1)\). Among the \(n\) eligible villages, suppose we select exactly \(qn\) to participate,\(^9\) with selection proportion \(q \in (0, 1)\). For instance, if there are \(n = 500\) villages, but we only have the resources to include 100 in the experiment, then we would set \(q = 1/5\).

\textit{Complete Randomization} - One natural way to draw the selection variables is by complete randomization: picking exactly \(qn\) out of \(n\) units uniformly\(^{10}\) at random. We denote this by \(T_{1:n} \sim \text{CR}(q)\). Under complete randomization, the estimator \(\bar{\theta}\) is unbiased for the ATE. However, if treatment effects are heterogeneous, so that \(c(X)\) varies with \(X\), the estimator may still have large variance.

\subsection{2.1.1 Variance due to Selection}

Using simulated data, Figure 1 shows one draw \(T_{1:n} \sim \text{CR}(q)\) of the selection variables for \(q = 1/5\). By chance, many of the highly educated, high infection rate villages were selected. Suppose, for instance, that these \textit{types} of villages have larger treatment effect on average \((c(X_i)\) large). Then the estimator \(\bar{\theta}\) will \textit{over-estimate} the average treatment effect. The opposite situation is equally likely: we may randomly select many low-education, low infection rate villages to participate, under-estimating the average treatment effect. This creates ex-ante estimator variance due to random selection.

\textbf{Representative Selection} - By contrast, in Figure 1 (b) the distribution of village types \(X_i\) among the selected villages \(X_i \mid T_i = 1\) closely matches the sample of all eligible

\(^9\)For brevity, assume this is an integer for now. We explicitly deal with edge effects in Section 3.

\(^{10}\)Formally, \(P(T_{1:n}) = \binom{n}{qn}^{-1}\) for each \(T_{1:n}\) with \(\sum_i T_i = qn\), and 0 otherwise.
Figure 2: Representative selection: (a) form homogeneous groups of villages (different colors) and (b) randomly select proportion $q$ units to represent each group.

villages $(X_i)_{i=1}^n$. By using representative selection, we force the smaller sample of selected villages to mimic the larger sample of eligible villages, reducing estimator variance.

Local Groups of Villages - How can we construct a representative selection? Consider the following procedure, shown in Figure 2.

1. (Local Groups) Match all eligible villages into homogeneous groups. In particular, for some baseline features $\psi(x)$ of each village chosen by the experimenter, the groups should be local in $\psi(x)$ space. For instance, the units in the lime green group at the bottom left in Figure 2 are locally clustered in $\psi(x) = (x_1, x_2) = (\text{education, infections})$ space. Note that constructing such groups is a hard combinatorial optimization problem. We give new algorithms and statistical analysis for this problem, as discussed in Section 2.3.1 below.

2. (Selection) Choose exactly proportion $q$ units from each group, completely at random, to represent that type of village in the trial. For instance, we have randomly selected one of the five units to represent the lime green group in the experiment, since $q = 1/5$.

By randomizing within local groups instead of globally, Figure 2 shows that we get a significantly more uniform selection of villages, reducing the estimator variance due to selection. Note that one can also think of this procedure as data-adaptive fine stratification.

2.1.2 Variance due to Assignment

Leaving the toy experiment behind, consider the classical setting with observed outcomes $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$ and no selection ($q = 1$). Suppose the experimenter wants proportion $p \in (0, 1)$ units assigned to the intervention $D = 1$. The full version of our method, sketched in Section 2.3.2 below, also allows for non-constant treatment
propensity $p(x)$. For now, assume that $p$ is constant with $p = a/k$. We may use an inverse propensity weighting (IPW) estimator of the ATE, which simplifies to difference-of-means when $p = 1/2$

$$\hat{\theta} = E_n \left[ \frac{D_i Y_i}{p} \right] - E_n \left[ \frac{(1 - D_i) Y_i}{1 - p} \right]$$

If treatment assignments $D_i$ are completely randomized, $\hat{\theta}$ is unbiased ex-ante but may have large variance. The reason for this is that while $E_n[D_i] = p$ globally, the realized propensity can significantly deviate from $p$ in large regions of the covariate space. For example, by random chance the villages in the highest education decile could all be assigned to $D_i = 1$. If education level is positively correlated with mask wearing, this will lead to positive bias ex-post (after the realization of $D_{1:n}$).

To formalize this intuition, let $m_d(X) = E[Y(d)|X]$, with residual $\epsilon^d_i = Y_i - m_d(X_i)$ and define the balance function

$$b(X; p) = m_1(X) \left( \frac{1 - p}{p} \right)^{1/2} + m_0(X) \left( \frac{p}{1 - p} \right)^{1/2} \quad (2.2)$$

We abuse notation and write $b(x) = b(x; p)$ when the propensity is clear. With a few steps of algebra, the IPW estimator can be written as

$$\hat{\theta} = E_n[c(X_i)] + E_n \left[ \frac{(D_i - p)}{\sqrt{p - p^2}} b(X_i) \right] + \text{residuals}_n \quad (2.3)$$

If $D_i = 1$ and $b(X_i)$ is large, we get a large positive fluctuation away from the ideal estimator $E_n[c(X_i)]$ and vice-versa. From this reasoning, we see that $b(x)$ is the exact function that needs to be balanced between treatment arms. In fact, the middle term is a sample correlation between $D_i$ and $b(X_i)$, measuring the confounding between treatments $D_i$ and covariates $X_i$ in our finite sample. We would like a balanced assignment procedure that makes this correlation small (with high probability), reducing estimator variance.

**Balanced Assignment** - To draw a balanced assignment of treatments, we use the same “local group” construction as before. Consider the example in Figure 3 with $qn = 100$ selected villages and treatment proportion $p = 2/3$. After selecting units using steps (1) and (2) above, we use the following procedure for treatment assignment:

(3) (Local Groups) Partition the selected villages $\{i : T_i = 1\}$ into groups that are maximally homogeneous in the baseline features $\psi(X_i)$.

(4) (Assignment) Assign exactly proportion $p$ units in each group to $D_i = 1$ and $(1 - p)$ to $D_i = 0$, completely at random.

Figure 3 shows that the realized propensity is close to $p = 2/3$ in each local region of the covariate space, producing strong (weighted) covariate balance between treatment arms. By contrast, Figure 4 shows selection and assignment by complete randomization, which produces significant imbalances. For instance, almost all of the highly infected villages with medium education levels are assigned $D_i = 1$. 
Figure 3: Balanced assignment: for the villages \( \{ i : T_i = 1 \} \) selected into the experiment, (a) form homogeneous groups (b) assign exactly proportion \( p \) of the villages in each group to treatment.

2.2 Efficient Estimation under Local Randomization

Putting this all together, we propose to use local randomization for both selection into the experiment and assignment of interventions, using the procedure outlined in steps (1)-(4) above. A single realization of this design is shown in Figure 4. We use a modified IPW estimator for treatment effect estimation

\[
\hat{\theta} = E_n \left[ \frac{T_i D_i Y_i}{q p} \right] - E_n \left[ \frac{T_i (1 - D_i) Y_i}{q (1 - p)} \right] \tag{2.4}
\]

Next, we briefly outline the efficiency gains from locally randomized selection and assignment. The claims in this section follow from our asymptotic results, which are discussed in detail in Section 3 below. Let \( \sigma^2_d(X) = \text{Var}(Y(d)|X) \) denote the heteroskedasticity function. Observe that the sample size is \( qn \), the number of units selected into the experiment. Under completely randomized selection and assignment \( \sqrt{qn}(\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V) \)

\[
V = \text{Var}(c(X)) + \text{Var}(b(X)) + E \left[ \frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1 - p} \right] \tag{2.5}
\]

The variance due to random selection is controlled by the CATE \( c(x) \), while the variance from treatment assignment is parameterized by the balance function \( b(x) \).

Our first main result (Theorem 3.17) shows that local randomization dampens both of these sources of variance. In particular, if we use baseline covariate features \( \psi(x) \) to form homogeneous groups, then \( \sqrt{qn}(\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V) \), with smaller asymptotic variance

\[
V = \text{Var}(c(X)) \cdot (1 - (1 - q)R_{c,\psi}^2) + \text{Var}(b(X)) \cdot (1 - R_{b,\psi}^2) + E \left[ \frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1 - p} \right]
\]
Figure 4: Comparison of local vs. complete randomization, used jointly for both selection and treatment assignment.

The increase in precision is controlled by the non-parametric $R^2$ coefficients

\[
R^2_{c,\psi} = \frac{\text{Var}(E[c(X)|\psi(X)])}{\text{Var}(c(X))} \quad R^2_{b,\psi} = \frac{\text{Var}(E[b(X)|\psi(X)])}{\text{Var}(b(X))}
\]  

(2.6)

$R^2_{c,\psi} \in [0, 1]$ gives the proportion of variance in treatment effects explained by $\psi(X)$ (in a non-parametric sense), while $R^2_{b,\psi}$ gives the proportion of variance in $b(X)$, and thus in experimental outcomes, explained by $\psi(X)$.

**Regression Adjustment by Design** - By rearranging, the variance due to treatment assignment may be written

\[
\text{Var}(b(X)) \cdot (1 - R^2_{b,\psi}) = E[(b(X) - E[b(X)|\psi(X)])^2]
\]  

(2.7)

Effectively, $b(X)$ is controlled by the non-parametric regression model $E[b(X)|\psi(X)]$. This shows that local randomization does non-parametric regression adjustment at the design stage, without actually estimating a regression.

**Boosting Experiment Size** - If $\psi(X) = X$ then $R^2_{c,\psi} = 1$. Then the asymptotic variance due to random selection simplifies to

\[
\text{Var}(c(X)) \cdot (1 - (1 - q)R^2_{c,\psi}) = q \text{Var}(c(X))
\]  

(2.8)

In the example above, there were $n = 500$ eligible villages, and we selected $q = 1/5$ of them to be in the experiment. By using local randomization to select villages, the selection component of the variance is dampened from $\text{Var}(c(X)) \rightarrow (1/5)\text{Var}(c(X))$, as if we had run a much larger experiment using all of the eligible villages. More generally, rearranging gives

\[
\text{Var}(c(X)) \cdot (1 - (1 - q)R^2_{c,\psi}) = q \text{Var}(E[c(X)|\psi(X)]) + E[\text{Var}(c(X)|\psi(X))]
\]  

(2.9)
Effectively, we get the large experiment size for the treatment effect heterogeneity predicted by \( \psi(X) \) (first term above). For the heterogeneity not predicted by \( \psi(X) \), we get the smaller true experiment size (second term above).

Choosing the Local Function \( \psi(x) \) - The experimenter has a choice of which baseline covariates \( \psi(x) \) to use when constructing local groups. From Equation 2.6, estimation is more precise as \( \psi(X) \) explains a larger proportion of the variability in \( c(X) \) and \( b(X) \).

In the limit \( \psi(X) = X \), we have \( R^2 = 1 \). However, there are statistical tradeoffs when increasing the dimension of \( \psi(x) \) in finite samples. Including many weakly predictive or irrelevant covariates in \( \psi(x) \) reduces the within-group match quality on the most important covariates, degrading finite sample performance. See Remark 4.6 below and the simulations in Section 6 for more discussion of these issues.

2.3 Formal Definition and Local Group Construction

They key step in the local randomization procedure outlined above was using baseline covariates to partition the experimental units into homogeneous groups. For locally randomized selection, this ensured that the different “types” of units in the eligible sample were well-represented by the selected units. For locally randomized treatment assignment, this ensured strong covariate balance between treatment arms. By homogeneity, we mean that the units in a group have very similar values of \( \psi_i = \psi(X_i) \), a sub-vector or transformation of the baseline covariates. Because of this, units within the same group appear to cluster in a local region of \( \psi(x) \) space, as shown in Figure 2.

Group Homogeneity - Formally, let \( [n] = \bigsqcup \{ i \in g \} \) be a partition of the experimental units into disjoint groups \( g \). For example, \( g_1 = \{1, 4, 7\}, g_2 = \{2, 13, 9\} \) and so on. We require the following homogeneity condition

\[
\sum \sum_{i < j \in g} |\psi_i - \psi_j|^2 = o_p(1) \tag{2.10}
\]

Equation 2.10 requires that the average squared Euclidean distance between features \( \psi_i \) and \( \psi_j \) within a group is going to 0 as \( n \to \infty \). We refer to \( \psi(x) \) as the local function. These are the covariate features that we use to match units into groups.

To make use of within-group homogeneity, we require that units with similar values of \( \psi_i \) have similar experimental outcomes, on average. In particular, the efficiency results in the previous section use the following smoothness condition\(^{11}\)

Assumption 2.1 (Smoothness). The map

\[
\psi \to E[Y(d)|\psi(X) = \psi]
\]

is Lipschitz continuous for \( d = 0, 1 \).

Next, we define local randomization, formalizing steps (1)-(4) in the discussion above.

\(^{11}\)This assumption can be significantly weakened at the cost of extra notation. For instance, we can allow finitely many discontinuities and weaker Hölder smoothness conditions, see Definition 7.1.
Definition 2.2 (Local Randomization). Let \( q = a/k \) with \( \gcd(a,k) = 1 \), and consider the following assumptions

1. (Local Groups) Assume we have a partition of the eligible units \([n] = \bigsqcup_g \{ i \in g \} \) into groups with \( |g| = k \), possibly excepting a single remainder group with \( 0 \leq |g| < k \). Assume that the homogeneity condition 2.10 holds, and that the groups only depend on the data through the function values \( (\psi(X_i))_{i=1}^n \).

2. (Randomization) Independently over all groups with \( |g| = k \), draw selection variables \((T_i)_{i \in g}\) by setting \( T_i = 1 \) for exactly \( a \) out of \( k \) units, completely at random. For units in the remainder group with \( |g| < k \), draw \( T_i \) iid with \( P(T_i = 1) = a/k \).

If the above conditions are satisfied, we say that the design implements \( q \) locally with respect to \( \psi(x) \), denoting

\[
T_{1:n} \sim \text{Loc}_n(\psi, q) \tag{2.11}
\]

For treatment assignment, we repeat the procedure above, replacing the eligible units \([n]\) by the selected units \( \{ i : T_i = 1 \} \) and replacing \( q \) by the treatment propensity \( p \). To be clear that we are assigning treatments only for the selected units, we denote

\[
D_{1:n} \sim \text{Loc}_n(\psi, p | T_{1:n}) \tag{2.12}
\]

The groups in Definition 2.2 may be viewed as fine, data-adaptive strata. They are data-adaptive in the sense that each group \( g \) generally depends on all of the values \( (\psi(X_i))_{i=1}^n \). Groups formed during treatment assignment will also depend on the full vector of selection variables \( T_{1:n} \). By contrast, classical stratified block randomization, as in Bugni et al. (2018), uses large (stratum size growing \( \propto n \)) fixed strata. The case \( q = 1 \) and \( p = 1/2 \) gives a matched pairs design.

2.3.1 Group Construction

The key step in Definition 2.2 is the construction of homogeneous groups satisfying Equation 2.10. Given such groups, the randomization step is easy: we assign binary variables within each group by complete randomization.

Optimal Groups: Maximally homogeneous groups may be obtained by minimizing Equation 2.10 over all feasible partitions of the eligible or selected units.

\[
(g^*) = \arg\min_{(g)} \sum_g \sum_{i,j \in g, i<j} |\psi_i - \psi_j|^2 \tag{2.13}
\]

Proposition 7.6 shows that the optimal groups satisfy condition 2.10 with rate \( O_p(n^{-2/(d+1)}) \) for \( \dim(\psi) = d \). However, computing the optimal groups is a hard combinatorial optimization problem (graph-partitioning).

Block Path Algorithm: The proof of Proposition 7.6 gives a constructive algorithm achieving the same \( O_p(n^{-2/(d+1)}) \) rate, extending and sharpening the rates for a construction of Bai et al. (2021). The procedure partitions the covariate space into blocks of a certain size, preferentially matching units into groups if they lie in the same block. However, the time complexity scales poorly with dimension.
**Greedy Algorithm:** In practice, we find that the following greedy minimization algorithm works well and is computationally efficient. The algorithm constructs groups iteratively. At each time step, we add a unit \( j^* \) to the group currently under construction that minimizes the within-group sum of squares, among the units currently assigned to that group. This process continues until the group size \(|g| = k\), at which point a new random group is initialized. In pseudo-code, we have

**Algorithm 2.3** (Greedy). Set \( I = \{1, \ldots, n\} \) and initialize the first group \( g = \{i\} \) randomly. Update \( I \leftarrow I \setminus \{i\} \). While \( I \neq \emptyset \) repeat the following:

(i) While \(|g| < k\) repeat the following:

(a) Find the optimal unit to add to a group

\[
    j^* = \arg\min_{j \in I} \sum_{i,j \in g} |\psi_i - \psi_j|^2
\]

(b) Update \( g \leftarrow g \cup \{j^*\} \) and \( I \leftarrow I \setminus \{j^*\} \)

(ii) Randomly initialize a new group \( g = \{i\} \)

### 2.3.2 Varying Treatment Proportions

In some situations, experimenters may want treatment assignment proportions \( p(x) \) to vary with baseline covariates. As in the introduction, it could be that the experimenter needs to assign surgical masks \( p = 2/3 \) to villages with more elderly residents, while they can allow \( p = 1/3 \) for younger villages. Allowing non-constant \( p(x) \) is also required for full efficiency, as discussed in Sections 3.2 and 4. We use a **double stratification** approach to extend the definition above: first stratifying units by distinct propensity levels, then locally randomizing within these propensity strata.

**Definition 2.4** (Local Randomization). Suppose that \( p(x) \in (p_j)_j \) for some finite collection of propensities \( p_j = a_j/k_j \) and \( \gcd(a_j, k_j) = 1 \). At the treatment assignment stage, do the following for each distinct propensity level \( p_j \)

(i) (Local Groups) Form homogeneous local groups of size \(|g| = k_j\) among the selected units \( T_i = 1 \) with this propensity level \( p(X_i) = p_j \)

(ii) (Treatment Assignment) Assign exactly \( a_j \) out of \( k_j \) units to treatment in each group, completely at random.

We denote this generalization by \( D_{1:n} \sim \text{Loc}_n(\psi, p(x)) \). See Definition 3.11 for more formal conditions, allowing approximation of arbitrary propensity scores \( p(x) \).

### 2.4 Efficient Design with a Pilot Study

We apply local randomization methods to the problem of design with a pilot study, constructing the first fully efficient design in this setting. Recall that under local randomization \( D_{1:n} \sim \text{Loc}_n(\psi, p) \) we had the variance decomposition

\[
    V = \text{Var}(c(X)) \cdot (1 - (1 - q)R^2_{c,\psi}) + \text{Var}(b(X)) \cdot (1 - R^2_{b,\psi}) + E \left[ \frac{\sigma^2_{1}(X)}{p} + \frac{\sigma^2_{0}(X)}{1 - p} \right]
\]
We seek a design that is fully efficient in the sense that $R^2_{c,\psi} = R^2_{b,\psi} = 1$, and the final residual variance term is also minimized. The first two terms are due to selection and assignment, respectively. The final term comes from the residuals $\epsilon_i^d = Y_i(d) - E[Y_i(d)|X_i]$, the part of outcomes not predicted by baseline covariates. Since $\epsilon_i^d$ is mean-independent of baseline covariates $X_i$, using $X_i$ to form homogeneous groups and randomize “locally” cannot change the residual variance.

To minimize the residual variance, we need to change $p \in (0, 1)$, the proportion of units assigned to each treatment arm.\textsuperscript{12} Intuitively, efficiency requires taking more measurements of the “noisier” treatment arm, which is harder to measure. For example, consider a vaccine trial where vaccine $d = 0$ always has a small positive effect, but the effect of an alternative vaccine $d = 1$ varies widely with unobserved genetic factors. We can reduce estimator variance by allocating more of our sample size towards the harder statistical problem of estimating $E[Y(1)]$. The optimal treatment proportions are given by

$$p^* = \arg\min_{p \in (0, 1)} E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1-p}\right] = \frac{\sqrt{E[\sigma_1^2(X)]}}{\sqrt{E[\sigma_1^2(X)]} + \sqrt{E[\sigma_0^2(X)]}} \quad (2.14)$$

Note that $p^*$ is increasing in the conditional variance $\sigma_d^2(X) = \text{Var}(Y(d)|X)$ of the $d = 1$ outcome.

**Using Pilot Data** - While $p^*$ is not known to the researcher, in settings with a pilot study we can form $\hat{p} \rightarrow p^*$ by estimating the conditional variances above. Locally randomized assignment $D_{1;n} \sim \text{Loc}_{n}(X, \hat{p})$ asymptotically minimizes both the variance due to treatment assignment and the residual variance (over constant propensity designs).

We can use a pilot study to achieve full efficiency by applying this argument conditionally. For intuition, note that if $\sigma_1^2(x)$ and $\sigma_0^2(x)$ were known, we could optimize the treatment proportions at each $X_i = x$, giving a globally optimal propensity score and pilot estimate

$$p^*(x) = \frac{\sigma_1(x)}{\sigma_1(x) + \sigma_0(x)} \quad \hat{p}(x) = \frac{\hat{\sigma}_1(x)}{\hat{\sigma}_1(x) + \hat{\sigma}_0(x)}$$

In Section 4.1, we show that the design $D_{1;n} \sim \text{Loc}_{n}(X, \hat{p}(x))$ is asymptotically fully efficient if $\hat{p}(x)$ is consistent for $p^*(x)$. In Section 4.2, we give an alternative approach that uses the pilot to “estimate what to balance.” This method is also asymptotically fully efficient, and may have better finite sample performance in settings with high-dimensional baseline covariates.

### 2.5 Exact Inference

We introduce new asymptotically exact inference methods for designs with locally randomized selection and assignment.

\textsuperscript{12}Changing $p$ also affects the variance due to assignment, since $b = b(x; p)$ If $R^2_{b,\psi} = 1$, these effects are lower order.
3 Asymptotics and Optimal Designs

This section gives our main asymptotic results. We show a central limit theorem for the IPW estimator in Equation 2.4 above under locally randomized selection and assignment. We show that our framework produces several useful new designs, described in the examples below. Next, we characterize various optimal designs that achieve the asymptotic variance bound. These results give new insights about what experimenters should attempt to balance at the design stage. We begin with the case of constant, rational treatment proportions $p = a/k$. Local randomization with varying treatment proportions $p(x)$, required for full efficiency, is discussed in Section 3.3 below.

For concision, in what follows we denote the balance function $b = b(X; p)$, as defined in Equation 2.2, and similarly $c = c(X)$ for the CATE, when the context is clear. Let $W_{1:n} = (X_i, Y_i(0), Y_i(1))_{i=1}^n$.

3.1 Constant Treatment Proportions

We begin with a central limit theorem, justifying the efficiency claims in Section 2 above. The technical assumptions 8.3 and 8.6 comprise various moment bounds, as well as the Lipschitz condition in Assumption 2.1 above.

**Theorem 3.1 (CLT).** Suppose that conditions 8.3 and 8.6 hold. Assume that the propensity $p = a/k$ is constant with $\gcd(a, k) = 1$. Suppose that selection and assignment are locally randomized:

(i) $T_{1:n} \sim \text{Loc}_n(\psi, q)$ or $T_{1:n} = 1$
(ii) $D_{1:n} \sim \text{Loc}_n(\psi, p \mid T_{1:n})$

Then $\sqrt{n}(\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V(\psi, (q, p)))$

$$V(\psi, (q, p)) = q \text{Var}(c) + (1 - q)E[(c - E[c|\psi])^2] + E[(b - E[b|\psi])^2] + E\left[\frac{\sigma^2_1(X)}{p} + \frac{\sigma^2_0(X)}{1-p}\right]$$

From the second term in the asymptotic variance, we see that locally randomized treatment assignment does non-parametric regression adjustment by design, leaving only the residual variance $E[(b - E[b|\psi])^2] = E[\text{Var}(b|\psi)]$. Similarly, locally randomized selection gives non-parametric control over the variance due to treatment effect heterogeneity. In particular, the first term above “looks like” we ran a larger experiment with sample size $n \geq qn$, up to an additive factor of $E[(c - E[c|\psi])^2] = E[\text{Var}(c|\psi)]$. This extra term is due to the residual treatment effect heterogeneity not explained by $\psi(x)$.

Intuition for key elements of the proof is given in Remark 3.22 below. For the full proof, see section 8.2 of the appendix.

**Example 3.2 (Matched k-tuples).** We can interpret the homogeneous groups of Definition 2.2 as a generalization of matched pairs. In particular, the local randomization algorithms in this paper implement “matched k-tuples” that are tightly matched in the sense of homogeneity Condition 2.10. We use these algorithms to produce homogeneous...
groups during both the selection and treatment assignment stages of the design.

For example, consider a situation with \( n = 700 \) units willing to participate, but there are only resources for \( qn = 300 \) to ultimately be included in the experiment. The researcher must randomly select proportion \( q = 3/7 \) of the eligible units to participate. Suppose that intervention \( D = 1 \) is more expensive, so that only \( p = 1/3 \) of the selected units may be assigned to \( D = 1 \). Following the steps in Definition 2.2, we use local randomization to design an efficient experiment satisfying these constraints. First, Algorithm 2.3 can be used to divide the sample into groups of 7 units that are homogeneous in the baseline covariates \( \psi(x) \). From each group, we select exactly 3 units to participate in the trial, leaving \((3/7) \cdot 700 = 300\) selected units. Next, we again use our algorithm to divide the selected units into homogeneous groups of 3 units, assigning exactly 1 out of 3 in each group to intervention \( D = 1 \). In notation, for the case \( \psi(x) = x \) we have

\[
\begin{align*}
(i) & \quad T_{1:n} \sim \text{Loc}_n(X, 3/7) \\
(ii) & \quad D_{1:n} \sim \text{Loc}_n(X, 1/3 \mid T_{1:n})
\end{align*}
\]

Then by Theorem 3.1 \( \sqrt{qn}(\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V) \)

\[
V = (3/7) \text{Var}(c(X)) + E \left[ \frac{\sigma_1^2(X)}{2/3} + \frac{\sigma_0^2(X)}{1/3} \right]
\]

Because of the tight matching into 7-tuples, the selection component of the asymptotic variance behaves as if we had run the trial with all \( n = 700 \) units. Because of the “matched triples” design during treatment assignment, the assignment variance component is lower order, and doesn’t appear at all in the asymptotic variance. Exact inference for this design is given by plugging \( q = 3/7 \) and \( p = 1/3 \) into the variance estimators in Theorem 5.3.

**Example 3.3** (Selection on Preliminary Covariates). In practice, experimenters may learn additional covariate information between the selection and treatment assignment stages of the design, after committing to select certain units into the experiment. Suppose that \( X_i = (X_{i}^{\text{pre}}, X_{i}^{\text{post}}) \), with only \( X_{i}^{\text{pre}} \) available at selection time. Consider the design

\[
\begin{align*}
(i) & \quad T_{1:n} \sim \text{Loc}_n(X^{\text{pre}}, q) \\
(ii) & \quad D_{1:n} \sim \text{Loc}_n(X, p \mid T_{1:n})
\end{align*}
\]

Then \( \sqrt{qn}(\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V) \)

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

We achieve full efficiency (relative to \( X \)) if and only if \( c(X) = c(X^{\text{pre}}) \). Intuitively, if \( c(X) \) only varies with \( X^{\text{pre}} \), we can use these preliminary covariates to select units into the trial that are “fully representative” of the variation in treatment effects over \( X \).

**Example 3.4** (Complete Randomization). Consider the local function \( \psi(x) = 1 \) for all \( x \). Note that this \( \psi(x) \) provides no information at all about experimental units. In particular, perfect homogeneity in the sense of Condition 2.10 holds for any group, since \( \psi_i = \psi_j = 1 \) for all \( i, j \). By Proposition 8.15 in the appendix, \( \text{Loc}_n(1, p) = \text{CR}(p) \) in distribution,
showing formally that “randomly matching” units into groups of $k$, then treating exactly $a$ out of $k$ in each group is equivalent to complete randomization $CR(a/k)$. From the theorem, if $T_{1:n} \sim \text{Loc}_n(1,q)$ and $D_{1:n} \sim \text{Loc}_n(1,p | T_{1:n})$. Then $\sqrt{qn}(\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0,V)$

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

**Remark 3.5 (Finite Population Estimand).** If the eligible units $[n]$ are equal to the full population, then the thought experiment where these units are randomly sampled from a larger super-population is invalid. For instance, in the Abaluck et al. (2021) example, it may be that all the villages of interest in Bangladesh are eligible to be included in the experiment. In this case, we may consider a finite-population estimand such as $E_n[c(X_i)]$, as in Abadie et al. (2014) or Kolesàr and Armstrong (2021). The proof of Theorem 3.1 shows that

$$\sqrt{qn}(\hat{\theta} - E_n[c(X_i)]) \Rightarrow \mathcal{N}(0,V(\psi, (q,p)) - q \text{Var}(c(X)))$$

Thus, we get the same asymptotic variance as before, minus the extra variability due to sampling eligible units from a larger super-population. Valid inference for this estimand is given by Theorem 5.3 below, exact for the case $\psi(x) = x$.

**Remark 3.6 (Design as Regression Adjustment).** Theorem 3.1 shows that local randomization does *non-parametric* regression at the design stage, without actually specifying or estimating a regression model ex-post. Similarly, Li et al. (2018) show that rerandomization does *linear* regression by design, up to a small extra variance term. Recent work in Harshaw et al. (2021) gives a novel “Grahm-Schmidt walk” design with MSE bounded by a quantity related to linear ridge regression.

Follow-up work in Cytrynbaum (2021) studies *rerandomizing* local designs by redrawing the complete randomizations within each local group until some global measure of balance is achieved. That paper shows that the variance of rerandomized local designs is controlled by the residuals of a *semi-parametric* regression, with the linear component due to rerandomization. For a fixed regressor $\psi(X)$, non-parametric regression residuals are smaller than OLS residuals. Because of this, locally balanced designs are asymptotically more efficient than globally balanced designs such as rerandomization, or stratified designs with non data-adaptive strata, as in Example 3.18 below.

### 3.2 The Optimal Design

The following corollary of Theorem 3.1 characterizes the *optimal design* for estimating the ATE in a given population, with fixed propensity $p$. The *fully* optimal design requires implementing varying propensity

$$p^*(x) = \frac{\sigma_1(x)}{\sigma_1(x) + \sigma_0(x)}$$

We give this result in Theorem 3.19 below, after discussing how to implement local randomization with non-constant propensity scores $p(x)$ in Section 3.3.

**Theorem 3.7 (Optimal Design).** Suppose that conditions 8.3 and 8.6 hold. Let the function $\psi(x)$ used to form local groups be one of the following
(i) $\psi^*(x) = (b(x;p), c(x))$
(ii) $\psi^*(x) = (m_0(x), m_1(x))$
(iii) $\psi^*(x) = x$

**Suppose that the design is locally randomized:**

(i) $T_{1:n} \sim \text{Loc}_n(\psi^*, q)$ or $T_{1:n} = 1$
(ii) $D_{1:n} \sim \text{Loc}_n(\psi^*, p | T_{1:n})$

Then $\sqrt{\bar{q}_n} (\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V^*)$

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

Under the optimal design, the selection and assignment terms are independently optimized. Note that the term $q \text{Var}(c(X))$ is irreducible.

**Remark 3.8 (What Experimenters Should Balance).** Consider (i) $\psi^*(x) = (b(x;p), c(x))$. Recall that the balance function $b(x)$ describes how outcomes vary with covariates, while $c(x)$ describes how treatment effects vary. This shows that the optimal design reduces both the selection and assignment sources of variance by balancing along two distinct dimensions

(i) (Selection) By forming local groups that are homogeneous in $c(X_i)$, then selecting exactly proportion $q$ units into the experiment ($T_i = 1$), the conditional average treatment effect level $c(x)$ in this group is “well-represented” in the experiment.

(ii) (Assignment) By forming local groups homogeneous in $b(x;p)$, the conditional average outcome level $b(x;p)$ is “well-represented” among the units assigned to each treatment arm.

Since $(b(x;p), c(x))$ are each weighted combinations of $(m_0(x), m_1(x))$, matching on both conditional expectations in (ii) is also sufficient.

**Remark 3.9 (Role of Covariate Dimension).** Note that (i) and (ii) are both infeasible, since none of these objects are observed. By contrast, (iii) $\psi(x) = x$ is feasible, and the theorem shows that this fully attenuates selection and assignment variance, asymptotically. If $x$ is low-dimensional, experimenters should use the designs above with $\psi(x) = x$. However, rates in Section 7 guaranteeing the local balancing condition 3.2 scale poorly with dimension. Intuitively, matching on $\psi(x) = x$ in high dimensions, with many irrelevant components, will likely give poor finite sample performance. In this case, a small subset of the baseline covariates $x$ expected to be most predictive of variation in both treatment effects and outcomes should be selected. See remark 4.6 and section 4.2 for more discussion of this issue.

**Example 3.10 (Matching on Risk Scores).** In view of the optimal designs in Theorem 3.7, consider letting $\psi(x) = (\hat{m}_0(x), \hat{m}_1(x))$ be regressions of proxy outcomes and treatments $\hat{Y}(\hat{d})$ on $X$ in a previous study (with some shared baseline covariates between studies). For example, if $D_i$ is an educational intervention, $X_i$ are student characteristics, and the outcome of interest $Y_i$ is mathematics test scores, the proxy outcome $\hat{Y}$ may be reading or overall test scores of different students in a previous study, under a related intervention $\hat{d}$. Such approaches are commonly used in clinical trials, where $\hat{m}_d(x)$ may be regarded as patient risk scores. If we use the design
(i) \( T_{1:n} \sim \text{Loc}_n((\hat{m}_0, \hat{m}_1), q) \)

(ii) \( D_{1:n} \sim \text{Loc}_n((\hat{m}_0, \hat{m}_1), p \mid T_{1:n}) \)

Then \( \sqrt{mq} (\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V) \)

\[
V = q \text{Var}(c) + (1 - q)E[(c - E[c|\hat{m}_0, \hat{m}_1])^2] + E[(b - E[b|\hat{m}_0, \hat{m}_1])^2] \\
+ E \left[ \frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1 - p} \right]
\]

This variance will be small if the risk scores \((\hat{m}_0(x), \hat{m}_1(x))\) explain variance in treatment effects and outcomes (through \(b(x)\)) well. Theorem 4.8 below shows that, in the case where the proxy study is actually a pilot with the same outcomes and treatments, and \(\hat{m}_d = \hat{m}_d\) are well-specified, consistent regressions, the efficiency bound of Theorem 3.7 is achieved.

### 3.3 Varying Treatment Proportions

This section describes the full version of the method, giving local randomization designs for arbitrarily varying treatment proportions \(p(x)\). We also formally give conditional asymptotics for the situation where \(\psi(x)\) and \(p(x)\) are random, for instance estimated using data from a pilot or proxy study. As outlined in Definition 2.4, we take a double stratification approach:

1. Discretize \(p(x)\) to a coarser, rational-valued propensity \(p_n(x)\)
2. Stratify units by their discretized propensity values \(p_n(X_i)\)
3. Locally randomize with respect to \(\psi(x)\) in each propensity stratum \(\{p_n(X_i) = a/k\}\).

For example, we may have \(p_n(x) \in \{2/5, 2/3\}\). In the stratum \(\{i : p_n(X_i) = 2/5\}\), we form homogeneous groups \(|g| = 5\), randomly assigning \(D_i = 1\) to 2 out of 5 units. The reader comfortable with this informal definition may choose to skip to Section 3.4 below, discussing the fully efficient design. Formally, we require the following conditions

**Definition 3.11** (Local Randomization). Assume the following

(a) Discretized propensities \((p_n)_n\) with approximation rate \(\|p_n - p\|_\infty = O(\frac{r^p_n}{n}) = o(1)\)
and \(p_n(x) \in L_n\) for all \(x\), with rational propensity level set \(L_n = \{p_n = q_a/k_a : a = 1, \ldots, |L_n|\}\), written with \(\gcd(q_a, k_a) = 1\).

(b) (Local groups) Assume we have formed groups \((g_{a,s})_{s=1}^n\) for each propensity level \(p_a \in L_n\), noting that there can be at most \(n\) groups. Assume that the groups satisfy the following conditions

(i) Feasible partition: \([n] = \bigsqcup_{a=1}^{L_n} \bigsqcup_{s=1}^n g_{a,s}\) (disjoint), with propensity level \(p_n(X_i) = p_a\) for all \(i \in g_{a,s}\), full group size \(|g_{a,s}| = k_a\) for \(1 \leq s \leq n - 1\), and remainder group \(0 \leq |g_{a,n}| \leq k_a\).

(ii) Restricted adaptivity: The groups \((g_{a,s})_{a,s}\) only depend on the data through \((\psi(X_i), p_n(X_i))_{i=1}^n\) and external randomness \(\xi_n\). In notation, for some \(\xi_n \perp W_{1:n}\) we have \((g_{a,s})_{a,s} \in \sigma(\psi_{1:n}, p_{1:n}, \xi_n)\).
(iii) Homogeneity: For $\psi_i = \psi(X_i)$

$$n^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} k_a^{-1} \sum_{i,j \in g_{a,s}, i \neq j} (\psi_i - \psi_j)^2 = O_p(r_n^\psi) = o_p(1) \quad (3.2)$$

(c) (Randomization) Complete randomization of treatment assignments within each local group $(D_i)_{i \in g_{a,s}} \sim CR(p_a)$, independently over groups $1 \leq s \leq n - 1$. Also, $D_i \overset{iid}{\sim} Bernoulli(p_a)$ for all units $i$ in the remainder group $s = n$.

We say that such a design implements $p(x)$ locally with respect to $\psi(x)$, denoting $D_{1:n} \sim Loc_n(\psi, p(x))$. For treatment post-selection, we will denote $D_{1:n} \sim Loc_n(\psi, p(x) \mid T_{1:n})$, replacing the set of eligible units $[n]$ in the above definition with the set of selected units $\{i : T_i = 1\}$.

**Example 3.12** (Targeted Matched Triples). Suppose the experimenter expects intervention $D = 1$ to be highly beneficial for units of type A, but less so for type B. In a medical trial, these could be old vs. young patients. Suppose that $X_i$ is a vector that contains unit type, but also other known characteristics of the unit. Welfare or political constraints may require, for example, $p(A) = 2/3 > 1/3 = p(B)$. Concretely, we can let the propensity $p(x)$ vary with unity type, with $p(A) = 2/3$ and $p(B) = 1/3$. For $D_{1:n} \sim Loc_n(X, p = \{1/3, 2/3\})$, the procedure in Definition 2.2 does the following

1. Group type A units into triples $|g| = 3$, matched on the full covariate vector $X_i$.
2. Assign $D = 1$ to exactly 2 out of 3 units in each group
3. Repeat the procedure for the type B units in the pro $\{i : p_n(X_i) = 1/3\}$, with 1 out of 3 treated.

This procedure (i) satisfies the experimenter’s constraint and (ii) gives the efficiency of a “matched-pairs” like design, using the design to control for all of the (potentially continuous) features $X_i$. Exact inference procedures for this design follow from the general formula in section 5, allowing the experimenter to fully take advantage of this efficiency.

**Remark 3.13** (Group Construction Algorithms). Equation 3.2 is the general version of the group homogeneity Condition 2.10 for the case of non-constant propensity scores. We analyze group-construction algorithms in Section 7 below that achieve the rate $r_n^\psi = O(n^{-2/(d+1)})$, with $\dim(\psi) = d$. In practice, we find that the greedy heuristic outlined in Algorithm 2.3 performs well and is fast.

**Remark 3.14** (Propensity Approximation). Our asymptotic framework allows the set of propensity levels $L_n$ to grow with $n$, accommodating asymptotic approximation of arbitrary propensity scores. With $\bar{k}_n = \max_{p_a \in L_n} k_a$ the maximal group size, we impose the technical condition

$$\bar{k}_n \vee |L_n| = O(\sqrt{n}/\log(n)) \quad (3.3)$$

requiring that the maximal group size and number of distinct propensity levels $|L_n|$ don’t grow too fast.

**Remark 3.15** (Balancing Estimated Functions). In Section 4, we study a setting where the local function $\psi(x)$ is estimated using a pilot data set, asking if we can “estimate what
to balance.” To formalize this setting, we write \( \psi = \psi(\cdot, \zeta) \) with \( \zeta \perp (X_i, Y_i(0), Y_i(1))_{i=1}^n \) a random element. In our applications, \( \zeta \) includes objects estimated out-of-sample, e.g. in a pilot or proxy data set. We also allow \( p(\cdot) = p(\cdot, \zeta) \), for instance allowing local randomization using an independent estimate of the optimal propensity score, as in Theorem 4.3. Theorem 3.17 and the inference methods in Section 5 are conditionally valid given the realized value of \( \zeta \). See the proof of Theorem 8.5 in the appendix for more details.

**Example 3.16 (Pilot Study).** Let \( \zeta = (\widehat{m}_d)_{d=0,1} \) be estimates of \( m_d(x) = E[Y(d)|X = x] \), from a pilot experiment. Consider setting

\[
\psi(x, \zeta) = \left( \widehat{m}_1(x) - \widehat{m}_0(x), \widehat{m}_1(x) \left( \frac{1-p}{p} \right)^{1/2} + \widehat{m}_0(x) \left( \frac{p}{1-p} \right)^{1/2} \right)
\]

Notice that this \( \psi \) is a (random) pilot estimate of the optimal design \( \psi^*(x) = (c(x), b(x)) \) from Theorem 3.19. The results below give consistent inference conditional on \( \zeta \), without requiring consistency of the regression estimates. If the regressions \( \widehat{m}_d(x) \) are consistent, Theorem 4.8 in the next section shows selection and assignment that are locally randomized with respect to \( \psi(x, \zeta) \) gives asymptotic efficiency.

Abusing notation, for local randomization with varying propensity score \( p(x) \), discretized to \( p_n(x) \) as above, we use the IPW estimator

\[
\widehat{\theta} = E_n \left[ \frac{T_i D_i Y_i}{q p_n(X_i)} \right] - E_n \left[ \frac{T_i (1-D_i) Y_i}{q (1-p_n(X_i))} \right] \tag{3.4}
\]

We also extend our definition of the balance function

\[
b(x; p(x)) = m_1(x) \left( \frac{1-p(x)}{p(x)} \right)^{1/2} + m_0(x) \left( \frac{p(x)}{1-p(x)} \right)^{1/2} \tag{3.5}
\]

The following theorem gives the most general version of our asymptotic results, extending Theorem 3.1 above.

**Theorem 3.17 (CLT).** Suppose that conditions 8.3 and 8.7 hold. Suppose that local functions \( \psi' = \psi'(\cdot, \zeta) \) and \( \psi = \psi(\cdot, \zeta) \) are increasing in granularity, in the sense that \( \psi' \in \sigma(\psi) \). Suppose that treatment proportions \( p = p(\cdot, \zeta) \) with random element \( \zeta \perp W_{1:n} \). Assume that selection and assignment are locally randomized

(i) \( T_{1:n} \sim \text{Loc}_n(\psi', q) \) or \( T_{1:n} = 1 \)

(ii) \( D_{1:n} \sim \text{Loc}_n(\psi, p(x) | T_{1:n}) \)

Then \( \sqrt{n}(\widehat{\theta} - \text{ATE})|\zeta \Rightarrow \mathcal{N}(0, V((\psi', \psi), (q, p))) \)

\[
V((\psi', \psi), (q, p)) = q \text{Var}(c) + (1 - q) E[(c - E[c|\psi', \zeta])^2 | \zeta] + E[(b - E[b|\psi, p, \zeta])^2 | \zeta] + E \left[ \frac{\sigma^2_0(X)}{p(X, \zeta)} + \frac{\sigma^2_0(X)}{1-p(X, \zeta)} \right] | \zeta
\]

Under the same conditions

\[
\sqrt{n}(\widehat{\theta} - E_n[c(X_i)])|\zeta \Rightarrow \mathcal{N} \left( 0, V((\psi', \psi), (q, p)) - q \text{Var}(c(X)) \right)
\]
Example 3.18 (Jointly Stratified Selection and Assignment). In this example, we show that local randomization subsumes stratified block randomization (SBR) with fixed strata (not data-adaptive). For the case without selection, this design was recently studied using alternative methods in Bugni et al. (2018). Our work gives asymptotics and exact inference for a more general class of designs with jointly stratified selection and assignment. Consider a fixed stratification $\psi(x) = S(x) \in \{1, \ldots, K\}$. For example, if $X$ are village-level infection rates, $S(X) \in \{1, \ldots, 5\}$ could be a discretized “infection-quintile.” By Proposition 8.15 in the appendix, $\text{Loc}_n(S, p(S))$ is equivalent to SBR($S, p(S)$). Consider a design with both stratified selection and assignment

(i) $T_{1:n} \sim \text{Loc}_n(S(x), q)$

(ii) $D_{1:n} \sim \text{Loc}_n(S(x), p(S(x)) | T_{1:n})$

Then by Theorem 3.17, $\sqrt{n}(\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$

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

The variance due to treatment assignment is reduced from $\text{Var}(b(X))$ under complete randomization to $E[(b - E[b|S])^2] = E[\text{Var}(b(X)|S)]$ under SBR. Thus, $b(X)$ is effectively “controlled” by a non-parametric regression with regressor $S(X)$, and similarly for the selection variance term $E[(c - E[c|S])^2]$. Exact inference for this design may be obtained by plugging $q$ and $p(s)$ into the variance estimators in Section 5.

Connection to OLS: We can relate the increase in precision from stratified selection and assignment to linear regression adjustment as follows. Let $I_S$ be a full set of strata indicator functions. Consider the population OLS equations

$$b(X) = \Gamma_b' I_S + v_b \quad E[I_S v_b] = 0$$

$$c(X) = \Gamma_c' I_S + v_c \quad E[I_S v_c] = 0$$

Then the variance expression above can be rewritten in terms of the population OLS residuals

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

This shows that stratified selection and assignment, with large, fixed strata, does linear regression adjustment by design.

3.4 Fully Optimal Designs

In Section 3.2, we gave restricted optimal designs implementing the constant treatment proportions $p = a/k$. In this section, we use the double stratification methods from the previous section to give fully optimal designs, which generically require implementing a non-constant optimal propensity score $p^*(x)$. To state the result, recall

\footnote{\(D_{1:n} \sim \text{Loc}_n(S, p(S))\) may differ from other implementations of SBR due to our explicit handling of remainder groups, when \(p(s) \cdot |\{i : S_i = s\}| \) is not an integer. The theorem shows that these differences are higher order.}

24
\[ \sigma^2_d(x) = \text{Var}(Y(d)|X = x) \]

and define

\[ p^*(x) = \frac{\sigma_1(x)}{\sigma_1(x) + \sigma_0(x)} \]

The propensity \( p^*(x) \) is known as the \textit{Neyman allocation}. Intuitively, \( p^*(x) \) assigns the noisier treatment arm more often (at each \( X_i = x \)), making optimal use of sample size. Such effects are discussed in detail in section 4, where we use local randomization to implement a pilot estimate of \( p^*(x) \). Let \( p_n(x) \) be a discretization of \( p^*(x) \), satisfying the conditions in Definition 3.11. We also require a generalization of Equation 3.5, giving a balance function with the implemented (discretized) propensity weights \( p_n(x) \)

\[
 b_n(x) = m_1(x) \left( \frac{1 - p_n(x)}{p_n(x)} \right)^{1/2} + m_0(x) \left( \frac{p_n(x)}{1 - p_n(x)} \right)^{1/2} \quad (3.6)
\]

Note that \( b_n(x) = b(x,p_n(x)) \).

The next result improves on the restricted optimal designs of Theorem 3.7.

**Theorem 3.19 (Optimal Design).** Suppose that conditions 8.3 and 8.7 hold. Let the function \( \psi(x) \) used to form local groups be one of the following

(i) \( \psi^*(x) = (b_n(x),c(x)) \)

(ii) \( \psi^*(x) = (m_0(x),m_1(x)) \)

(iii) \( \psi^*(x) = x \)

Suppose that the design is locally randomized:

(i) \( T_{1:n} \sim \text{Loc}_n(\psi^*,q) \) or \( T_{1:n} = 1 \)

(ii) \( D_{1:n} \sim \text{Loc}_n(\psi^*,p^*(x)|T_{1:n}) \)

Then \( \sqrt{n}(\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0,V^*) \)

\[
 V^* = q \text{Var}(c(X)) + \min_{\theta < p < 1} E \left[ \frac{\sigma^2_1(X)}{p(X)} + \frac{\sigma^2_0(X)}{1 - p(X)} \right] \quad (3.7)
\]

Under the fully optimal design, the selection, assignment, and residual components of the asymptotic variance are all independently optimized.

### 3.5 Theory Discussion

This section collects various discussions that may be of interest to researchers working on statistical theory for experimental design. In particular, we give intuition for some key elements of the proof of our asymptotic results in Remark 3.22.

**Remark 3.20 (Efficient IPW Estimation).** Suppose there is no selection \( (q = 1) \) and consider simple random sampling, which draws treatments \( D_i \overset{\text{inid}}{\sim} \text{Bernoulli}(p) \), as in Hahn et al. (2011). The results of Hirano et al. (2003) show that in this setting IPW estimation using the true propensity weights \( p(X_i) \) is inefficient, while weighting with a non-parametric estimate \( \hat{p}(X_i) \) of the known propensity achieves the semi-parametric variance
bound

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

By contrast, if \( D_{1:n} \sim \text{Loc}_n(X, p(x)) \), then by Theorem 3.17 \( \sqrt{n}(\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V^*) \). In particular, the IPW estimator with true propensity weights achieves the semi-parametric variance bound, removing the need for an extra step where we non-parametrically “re-estimate” a known quantity.

To understand this improvement, note that under independent treatment assignment, for any “nice” local region \( A \subseteq \mathcal{X} \) we have \( \sqrt{n}E_n[(D_i - p(X_i))1(X_i \in A)] = \Omega_p(1) \), reflecting order \( n^{-1/2} \) dispersion of the sample treatment proportions away from the target proportions \( p(x) \). Unless we update the propensity weights \( p(X_i) \to \hat{p}(X_i) \) to reflect the realized, in-sample treatment proportions, this dispersion increases the first-order asymptotic variance. By contrast, Lemma 8.4 in the appendix shows that under local randomization \( \sqrt{n}E_n[(D_i - p_n(X_i))1(X_i \in A)] = O_p(n^{-1/(d+1)}) \) with \( d = \text{dim}(X) \). Local randomization forces the sample treatment proportions in each fixed region of the covariate space to be equal to the target proportions, up to lower order. Also see Figure 4, which shows an example where in-sample propensity widely diverges from the population propensity under complete randomization.

Remark 3.21 (Robustness vs. Efficiency). By Theorem 3.17, the assignment variance under \( D_{1:n} \sim \text{Loc}_n(\psi, p) \) is, using the law of total variance

\[ E[(b - E[b|\psi, p])^2] = \text{Var}(b(X)) - \text{Var}(E[b|\psi, p]) \leq \text{Var}(b(X)) \]

This shows that, at least asymptotically, balancing any function of the covariates weakly dominates complete randomization, which achieves the maximal variance \( \text{Var}(b(X)) \). The proof of Theorem 8.5 shows that, in finite samples, this variance ordering still holds ex ante, but may fail conditionally on \( X_{1:n} \), e.g. if \( \psi(X) \) predicts outcomes poorly, echoing a similar observation in Imbens (2011) for matched pair designs. The conditional variance calculations in Theorem 8.5 show that such problematic variance components vanish asymptotically at rate \( O_p(n^{-1/2}) \). For details, see the discussion in Remark 3.22 below.

Remark 3.22 (Proof Intuition). In this remark, we give intuition for the proof of our CLT in Theorem 3.1. Consider the special case \( T_{1:n} = 1 \) and \( D_{1:n} \sim \text{Loc}_n(\psi, p) \) with \( p = a/k \). By the fundamental decomposition of the IPW estimator

\[ \hat{\theta} = E_n[c(X_i)] + E_n \left[ \frac{(D_i - a/k)}{\sqrt{a/k - (a/k)^2}} b(X_i) \right] + \text{residuals}_n \]

We highlight two key elements of the proof of Theorem 8.5 (from which Theorem 3.17 above is derived), focusing on our conditional CLT for the middle term above

1. Coupling to non-parametric residuals

\[ \sqrt{n}E_n[(D_i - a/k)b(X_i)] = \sqrt{n}E_n[(D_i - a/k)(b - E[b|\psi])(X_i)] + O_p \left( (r_n^{\psi})^{1/2} \right) \]
(2) Residuals $u_i \equiv b_i - E[b|\psi_i]$ behave “as if” randomly assigned to groups

$$E[u_iu_j|i,j \in g] = E[u_iu_j] = 0$$

For (1), observe that $b_i = E[b|\psi_i] + (b_i - E[b|\psi_i])$. If $\psi \rightarrow E[b|\psi]$ is Lipschitz, by Lemma 8.20 the local-group correlation structure of the treatment assignments gives

$$\text{Var}(\sqrt{n}E_n[(D_i - a/k)u_i|X_{1:n}, \pi_n]) \leq k^{-1} \sum_g \sum_{i,j \in g, i \neq j} (E[b|\psi_i] - E[b|\psi_j])^2 + O_p(\log(n)/n)$$

$$\leq_P k^{-1} \sum_g \sum_{i,j \in g, i \neq j} (\psi_i - \psi_j)^2 = O_p(r_n^\psi) = o_p(1)$$

using the balancing rate guarantees for groups constructed in section 7. External randomness $\pi_n \perp W_{1:n}$ is needed to break potential ties during group construction.

For (2), since the algorithms in section 7 use all values $\psi_{1:n} = (\psi(X_i))_{i=1}^n$ to globally optimize the balance criterion, each group $g$ has stochastic dependence $g = g(\psi_{1:n}, \pi_n)$. Then one argues that $E[u_iu_j|i,j \in g] = E[E[u_iu_j|\psi_{1:n}, \pi_n]|i,j \in g] = 0$, showing the residuals behave “as if” randomly assigned to groups. The proof of Theorem 8.5 shows that

$$\text{Var}(\sqrt{n}E_n[(D_i - a/k)u_i|X_{1:n}, \pi_n]) = E[(b - E[b|\psi])^2] - \frac{a(k-a)}{k^2(k-1)} \sum_g \sum_{i,j \in g, i \neq j} u_iu_j + o_p(1)$$

By this “anti-selection” result for the residuals, the middle term can be shown to be $O_p(n^{-1/2})$. For more details, see Lemmas 8.20, 8.22, and the proof of Theorem 8.5 in the appendix. For concrete rate guarantees on $r_n^\psi$ under weaker smoothness assumptions, see section 7.

## 4 Efficient Design with a Pilot Experiment

This section studies new methods that use data from a pilot\(^{14}\) study in order to design a more efficient main experiment. In particular, we give the first asymptotically fully efficient designs in this setting.

Local randomization methods can be used to minimize the variance due to selection and treatment assignment (Section 3.2). However, further optimizing the residual component of the asymptotic variance requires assigning more units to the “noisier” treatment arm (Section 3.4). Intuitively, one can think of this as taking more measurements of the thing that is harder to measure. The relative variance of the different potential outcomes is unknown, but may be estimated using pilot data. Thus, we propose to use the pilot experiment to construct a feasible version of the optimal design in two steps. First, we

---

\(^{14}\)The results from this section also apply with access to observational data from the population of interest, as long as the relevant causal quantities are identified and estimable.

27
stratify units on a pilot estimate of the optimal treatment proportions

\[
\hat{p}(x) = \frac{\hat{\sigma}_1(x)}{\hat{\sigma}_1(x) + \hat{\sigma}_0(x)} \quad p^\ast(x) = \frac{\sigma_1(x)}{\sigma_1(x) + \sigma_0(x)}
\]  

(4.1)

with \(\sigma_d(x) = \text{Var}(Y(d) | X = x)^{1/2}\). Notice that full efficiency requires implementing the optimal treatment proportions conditionally, at each \(X_i = x\). The optimal propensity \(p^\ast(x)\) is known as the *Neyman allocation*. In the second step, we implement these treatment proportions by local randomization within each propensity stratum. For example, in a propensity stratum \(\{i : \hat{p}(X_i) = 3/5\}\) we partition the units into homogeneous groups of 5, assigning exactly 3 in each group to \(D = 1\). This *double stratification* approach gives a design with (an estimate of) the non-constant optimal treatment proportions \(p^\ast(x)\). By implementing the estimate \(\hat{p}(x)\) using local randomization, we also strongly control the variance due to treatment assignment. In particular, this strategy independently minimizes the selection, assignment, and residual components of the asymptotic variance (Theorem 4.8), giving the first fully efficient design in this setting.

**Remark 4.1** (High Dimensional Covariates). We first propose a method that is *robust* in the sense that it does not require correct specification or estimation of outcome regression models (Section 4.1). However, the finite sample performance of this method may be poor in high-dimensional settings with many weakly predictive covariates. Therefore, we consider a second method that uses the pilot study to “estimate what to balance” (Section 4.2). In particular, locally randomizing with respect to a pilot estimate \(\psi(x) = (\hat{b}_n(x), \hat{c}(x))\) of the optimal design \(\psi^\ast(x) = (b_n(x), c(x))\) given by Theorem 3.19 is asymptotically fully efficient if the pilot regressions are estimated consistently. However, we argue that strong consistency assumptions may be inappropriate for experimental design, leading us to prefer a “robustified” version of the second method with \(\psi(x) = (\hat{b}_n(x), \hat{c}(x), x_{\text{pred}})\), also balancing some predictive covariates *outright*. See the discussion in Remark 4.12.

**Advice for Practice** - The theory developed in this section suggests

(1) For low-dimensional baseline covariate covariates \(X_i\), let selections and treatment assignments be drawn by \(T_{1:n} \sim \text{Loc}_n(X, q)\) and \(D_{1:n} \sim \text{Loc}_n(\hat{X}, \hat{p}(x) | T_{1:n})\), where \(\hat{p}(x)\) is a “working model” of the Neyman allocation estimated using pilot data. For considerations involved in choosing the estimate \(\hat{p}(x)\), see the practical discussion in Remark 4.5 below.

(2) For high-dimensional baseline covariates \(X_i\), let \(\psi(x) = (\hat{b}_n(x), \hat{c}(x), x_{\text{pred}})\), including a pilot estimate of the balance function and CATE, and also including a few of the covariates thought to be most predictive of outcomes outright. Then, let selections and treatment assignments be drawn by \(T_{1:n} \sim \text{Loc}_n(\psi, q)\) and \(D_{1:n} \sim \text{Loc}_n(\psi, \hat{p}(x) | T_{1:n})\).

### 4.1 Robust Solution

This section presents our robust solution to the problem of efficient design with a pilot study. First, we discuss implementation of the Neyman allocation estimate in Equation 4.1 in more detail. Consider a simple implementation of the propensity \(\hat{p}(x)\) using independent sampling

\[D_i^{\text{ind}} \sim \text{Bernoulli}(\hat{p}(X_i))\]
Unfortunately, under this strategy the realized treatment proportions may significantly differ from the pilot estimate \( \hat{p}(x) \) that we wish to implement, due to the \( O(n^{-1/2}) \) dispersion created by independent sampling. This dispersion creates covariate imbalances that need to be removed ex-post in the analysis stage. For instance, this may require well-specified regression adjustment, as in Robinson (1988), or ex-post, non-parametric re-estimation of the (known) propensity score, as in Hirano et al. (2003).

**Double Stratification** - Alternatively, consider assignments \( D_{1:n} \sim \text{Loc}_n(X, \hat{p}(x)) \), using the double stratification approach defined in Section 3.3 above. This gives a fast implementation of \( \hat{p}(x) \), forcing the realized treatment proportions to be exactly \( \hat{p}(x) \), up to lower order dispersion. Using this method, we can implement the estimated optimal treatment proportions \( \hat{p}(x) \), while also strongly controlling the selection and assignment components of the asymptotic variance.

**Remark 4.2** (Large Pilot Asymptotics). In the following theorem, we assume that the pilot sample size \( n_{\text{pilot}} \gg n \), working in large pilot asymptotics as in Hahn et al. (2011). This formally allows for consistent estimation \( \| \hat{\sigma}_d - \sigma_d \|_{2, \mathcal{X}} = o_p(1) \) using the pilot data. While consistency is required for the asymptotic optimality result, we advocate treating \( \hat{\sigma}_d(x) \) as a “working model” of the variance function. Using the pilot to estimate residual variances will generally improve upon an arbitrary propensity score, such as \( p \approx 1/2 \), even if the estimator \( \hat{\sigma}_d(x) \) is misspecified or formed using a small pilot.

**Theorem 4.3.** Suppose \( \| \hat{\sigma}_d - \sigma_d \|_{2, \mathcal{X}} = o_p(1) \) and impose assumptions 8.3 and 8.9. With \( \hat{p}(x) \) the pilot Neyman allocation estimate in Equation 4.1, suppose that

\[
\begin{align*}
(i) & \quad T_{1:n} \sim \text{Loc}_n(X, q) \\
(ii) & \quad D_{1:n} \sim \text{Loc}_n(X, \hat{p}(x) | T_{1:n})
\end{align*}
\]

Then
\[
\sqrt{n} \left( \hat{\theta} - \text{ATE} \right) \Rightarrow \mathcal{N}(0, V^*)
\]

\[
V^* = q \text{ Var}(c(X)) + \min_{0 < p < 1} \left( E \left[ \frac{\sigma^2_0(X)}{p(X)} + \frac{\sigma^2_0(X)}{1 - p(X)} \right] \right)
\]

Theorem 4.3 shows that the selection, assignment, and residual components of the asymptotic variance are independently minimized. In particular, the variance due to treatment assignment is lower order. For the proof of this theorem, see section 8.3 of the appendix.

**Remark 4.4** (Double Stratification). For more intuition, consider the treatment assignment step in detail. The notation \( D_{1:n} \sim \text{Loc}_n(X, \hat{p}(x) | T_{1:n}) \) does the following:

1. Stratify the selected units on \( p_n(X_i) \), values of a discretized\(^{16}\) version of the estimated optimal treatment proportions \( \hat{p}(x) \). This partitions units based on the relative noisiness of potential outcomes \( Y_i(1) \) vs. \( Y_i(0) \) at \( X_i \), forming strata of units with similar optimal treatment probabilities.

\(^{15}\)Formally, for \( A \subseteq \mathcal{X} \) a fixed “nice” region of the covariate space, \( E_n[|D_i - p_n(X_i)|1(X_i \in A)] = O_p((r_n^\alpha/n)^{1/2}) \). See Lemma 8.4 for details.

\(^{16}\)If the discretization has \( |L_n| \) propensity levels with \( \overline{k}_n = \max_{q_n/k_n \in L_n} k_n \), the local balancing results in section 7 require \( |L_n|\overline{k}_n = o(n) \). For the standard discretization \( \{1/k_n, \ldots, k_n - 1/k_n\} \), let \( 1/k_n \asymp \log(n)/\sqrt{n} \).
(2) Within each propensity stratum \( \{i : p_n(X_i) = a/k\} \), match units on baseline covariates \( X_i \) into homogeneous groups of size \( |g| = k \), assigning \( D_i = 1 \) to exactly \( a \) out of \( k \) of them, uniformly at random.

By doing this, we simultaneously (1) implement treatment proportions that optimize the residual variance and (2) do “matched k-tuples” (Example 3.12) within each propensity stratum, producing asymptotically lower order assignment variance.

**Remark 4.5** (Variance Models in Practice). While the asymptotic full efficiency of Theorem 4.3 requires well-specified, consistent \( \hat{\sigma}^2_d(x) \), the asymptotics and inference results in Theorem 3.17 and Theorem 5.3 hold for any, potentially inconsistent, estimate \( \hat{p}(x) \bot \) data of the optimal treatment proportions, including the following:

(a) \( \hat{p} = \frac{\hat{s}_1}{\hat{s}_1 + \hat{s}_0} \) constant with \( \hat{\sigma}^2_d \) estimated from a previous experiment with different, but conceptually related, treatments and outcomes

(b) \( \hat{p} = \frac{\hat{s}_1}{\hat{s}_1 + \hat{s}_0} \) an educated guess of the relative variability of outcomes for \( d = 0, 1 \)

(c) \( \hat{p}(x) = \frac{\hat{s}_1(x)}{\hat{s}_1(x) + \hat{s}_0(x)} \) with \( \hat{\sigma}^2_d(x) \) estimated inconsistently from confounded observational data

In each case, assigning \( D_{1:n} \sim \text{Loc}_n(X, \hat{p}(x) \mid T_{1:n}) \) attempts to reduce estimator variance by (1) taking more samples of the treatment arm with noisier outcomes and (2) implementing these treatment proportions using local randomization methods. Asymptotically exact inference, with no assumptions on the consistency or well-specification of \( \hat{p}(x) \), is given by Theorem 5.3.

**Remark 4.6** (Role of Covariate Dimension). If the algorithms in Section 7 are used to form homogeneous groups and \( D_{1:n} \sim \text{Loc}_n(X, \hat{p}(x) \mid T_{1:n}) \), then Lemma 8.4 shows that the variance due to treatment assignment is asymptotically lower order

\[
\sqrt{n}E_n[(D_i - p_n(X_i))b_n(X_i)] = O_p(n^{-1/(d+1)}) \quad d = \text{dim}(X)
\]

However, if \( X \) is high-dimensional the assignment variance may be non-negligible in finite samples due to the slow convergence rate above. For instance, if many components of \( X \) are irrelevant for predicting outcomes and treatment effect variation, the “signal” of the relevant covariates may be lost when we match on all components of the vector \( X \). Motivated by this, Section 4.2 considers using pilot data to “estimate what to balance,” rather than balancing the whole vector \( X \). See Section 6 for monte carlo evidence showing the deterioration of finite sample performance in high dimensions.

In the remainder of this subsection, we compare our method with the recent “optimal stratification tree” approach of Tabord-Meehan (2020).

### 4.1.1 Comparison with Optimal Stratification Trees

Recall that classical stratified block randomization (SBR) assigns exactly proportion \( p(s) \) units to treatment, completely at random, within the stratum \( \{S(X_i) = s\} \), for some fixed stratification of the covariate space \( S \in \{1, \ldots, K\} \). Tabord-Meehan (2020), henceforth TM, uses pilot data to estimate an efficient stratification and propensity score \((\hat{S}, \hat{p}(s))\) over a set of tree partitions of the covariate space, denoted \( \mathcal{T} \).
Theorem 3.1 of TM\textsuperscript{17} shows that if $D_{1:n} \sim \text{SBR}(\hat{S}, \hat{p}(s))$ then an IPW estimator $\hat{\theta}$ has $
abla \hat{\theta} - \text{ATE} \Rightarrow \mathcal{N}(0, V)$ with asymptotic variance

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

The within-stratum optimal propensity $p^*(X; S)$ and balance function are given by

$$p^*(X; S) = \frac{\text{Var}(Y(1)|S)^{1/2}}{\text{Var}(Y(1)|S)^{1/2} + \text{Var}(Y(0)|S)^{1/2}}$$

$$b(X; S) = m_1(X) \sqrt{\frac{1 - p^*(X; S)}{p^*(X; S)}} + m_0(X) \sqrt{\frac{p^*(X; S)}{1 - p^*(X; S)}}$$

From the minimization problem in the first display, we see that the optimal tree stratification chooses a compromise between:

1. In the first term, minimizing variance due to treatment assignment by choosing strata that predict outcomes well (parameterized by the balance function $b(X; S)$)
2. In the second term, minimizing residual variance by sampling treatment arms proportionally to their relative residual variance, through the induced propensity $p^*(X; S)$

By contrast, for the case with no selection into the experiment ($q = 1$), the local randomization approach of Theorem 4.3 achieves asymptotic variance

$$V^* = \text{Var}(c(X)) + \min_{p \in L_2(X)} \min_{0 < p < 1} E \left[ \frac{\sigma_1^2(X)}{p(X)} + \frac{\sigma_0^2(X)}{1 - p(X)} \right]$$

Both terms are independently optimized, and the variance due to assignment is asymptotically vanishing due to local matching within propensity strata. In particular, $V^* \leq V$.

\textbf{Remark 4.7 (Discussion).} Our strategy exploits the fact that we have stronger control over the variance due to treatment assignment than the residual variance. Note that

1. The variance due to treatment assignment can be made asymptotically lower order by locally randomizing on $\psi(x) = x$.
2. The variance due to residuals $\epsilon^d_i = Y_i(d) - E[Y_i(d)|X_i]$ cannot be affected by matching-like methods, including classical stratification, since the baseline covariates $X_i$ used for stratification are mean-independent of $\epsilon^d_i$. The only parameter we can use to affect the residual variance is the propensity score $p(x)$, through which the proportion of residuals $\epsilon^d_i$ drawn from each arm $d = 0, 1$ may be optimized.

The assignment variance is the easier term: we can ensure that it is lower order by implementing a consistent estimate of the optimal propensity $p^*(x)$ with local randomization methods. By contrast, optimal stratification trees force us to choose a compromise between the easier assignment component and the residual component of the variance, leading to asymptotic sub-optimality.

\textsuperscript{17}For a formal transformation of the SBR variance of Bugni et al. (2018) (cited in TM) into “balance function adjustment” form, see section 8.3 of the appendix.
4.2 Using a Pilot Study to “Estimate what to Balance”

Motivated by settings with high-dimensional covariates, where the robust approach above may have poor finite-sample performance, we consider using the pilot study to “estimate what to balance.” In our discussion of optimal designs in Section 3.4, we showed that locally randomizing with respect to \( \psi^*(x) = (b_n(x), c(x)) \) during both selection and assignment was fully efficient. Remark 3.8 gives a more intuitive discussion of this result. Since the optimal design \( \psi^*(x) \) is unknown to the experimenter, we consider estimating it using data from a pilot study, extending Bai (2020). Under consistency and well-specification of the pilot regressions, Theorem 4.8 shows that this feasible design is asymptotically fully efficient. However, we argue that such formal assumptions are too strong for design theory, giving poor guidance for practice. In particular, Proposition 4.11 shows a sense in which, under such assumptions, all designs are asymptotically equivalent.

**Advice for Practice** - In view of the theory in this section, for design with a pilot study and high-dimensional baseline covariates we advocate a robustified version of “estimating what to balance.” In particular, experimenters can use locally randomized selection and assignment with \( \psi(x) = (\hat{b}_n(x), \hat{c}(x), x_{pred}) \), including pilot estimates of the balance function and CATE, but also including, outright, a small vector of covariates \( x_{pred} \) that are suspected a priori to be good predictors of treatment effect heterogeneity and variation in potential outcomes. Theorem 5.3 gives asymptotically conditionally exact inference, conditional on the pilot data, for the designs in this section.

**Theorem 4.8 (Balancing Pilot Regressions).** Consider pilot regression estimators \( \hat{m}_d \) consistent with rate \( \| \hat{m}_d - m_d \|_2, P = O_p(n^{-\gamma}) \) and \( \| \hat{m}_d \|_\infty < \infty \) a.s. for \( d = 0, 1 \). Require assumption 8.9, and let \( \hat{p}(x) \) be a consistent estimate of the Neyman allocation as in Theorem 4.3, with discretization \( p_n(x) \) satisfying assumption 8.1. Define the estimators

\[
\hat{b}_n(x) = \hat{m}_1(x) \left( \frac{1 - p_n(x)}{p_n(x)} \right)^{1/2} + \hat{m}_0(x) \left( \frac{p_n(x)}{1 - p_n(x)} \right)^{1/2}, \quad \hat{c}(x) = \hat{m}_1(x) - \hat{m}_0(x)
\]

Suppose one of the following

(a) \( \hat{\psi}(x) = (\hat{b}_n, \hat{c})(x) \)

(b) \( \hat{\psi}(x) = (\hat{m}_0, \hat{m}_1)(x) \)

Suppose that the design is given by

(1) \( T_{1:n} \sim \text{Loc}_n(\hat{\psi}, q) \)

(2) \( D_{1:n} \sim \text{Loc}_n(\hat{\psi}, \hat{p}(x) | T_{1:n}) \)

Then \( \sqrt{n}(\hat{\theta} - \text{ATE}) \Rightarrow N(0, \sigma^2) \), with \( \sigma^2 \) the optimal variance of Theorem 3.19.

Theorem 4.8 extends the results in Bai (2020), see Remark 4.14 below for a discussion.

**Remark 4.9 (Embedding Regressions).** Assigning \( D_{1:n} \sim \text{Loc}_n(\hat{\psi}, \hat{p}(x) | T_{1:n}) \) gives full efficiency by **embedding** the regression estimators above into the design. Formally, consider \( D_{1:n} \sim \text{Loc}_n(\hat{b}, 1/2) \). The proof of Theorem 4.8 shows that the variance due to treatment assignment comes from the following term

\[
\text{assignment}_n \propto E_n[(D_i - 1/2)b(X_i)] = E_n[(D_i - 1/2)(b - \hat{b})(X_i)] + E_n[(D_i - 1/2)b(X_i)]
\]

\[
= E_n[(D_i - 1/2)(b - \hat{b})(X_i)] + O_p(n^{-1})
\]
The final equality follows by conditional Chebyshev, using the local group homogeneity condition in Equation 2.10. The remaining term appears in the analysis of regression-augmented (AIPW) estimators in observational studies. From this, we see that locally balancing a regression $\hat{b}(x)$, makes the IPW estimator behave like an AIPW estimator, achieving the semi-parametric efficiency bound for the implemented propensity.

**Design Irrelevance:** Motivated by the previous result, we study the AIPW estimator itself. The next result shows that under (1) regression-augmented estimation (AIPW) and (2) assuming consistent pilot estimation of $m_d(x) = E[Y(d)|X = x]$, all designs are equivalent in first-order asymptotics. Before continuing, we formally define the cross-fit AIPW estimator, as in Chernozhukov et al. (2017)

**Definition 4.10** (AIPW). Let $[n] = \cup_{k=1}^K I_k$ be a random partition. Define the out-of $k^{th}$ fold regression estimate $\hat{m}_{d,k} \equiv \hat{m}_d((W_i)_{i \in I_k})$. For each $k$ define the within $k^{th}$ fold estimator

$$\hat{\theta}_k \equiv n^{-1} \sum_{i \in I_k} \hat{c}_k(X_i) + n^{-1} \sum_{i \in I_k} \left( \frac{D_i(Y_i - \hat{m}_{1,k}(X_i)) - (1 - D_i)(Y_i - \hat{m}_{0,k}(X_i))}{p_n(X_i)} \right)$$

Define the cross-fit AIPW estimator $\hat{\theta} = \sum_k \hat{\theta}_k$.

**Proposition 4.11** (Design Irrelevance). Suppose $(\hat{m}_d)_{d=0,1}$ have $\|\hat{m}_d - m_d\|_{2,P} = o_p(1)$. Suppose that one of the following holds

1. $(D_i)_{i \in I_k} \sim \text{Loc}_n(\psi, p(x)|I_k)$ for $(I_k)_{k=1}^K$ a random partition of $[n]$.
2. $D_i \overset{\text{inid}}{\sim} p(X_i)$

Let $\hat{\theta}$ be a cross-fit AIPW estimator as in Definition 4.10, with weights $p_n = p$ in case (2). Then $\sqrt{n}(\hat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$

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

**Remark 4.12** (Consistency vs. Robustness). Consistent estimation of nuisance parameters is necessary for treatment effect point estimation in observational settings, but unnecessary with experimental data. If we impose this assumption anyway in an experimental context, Proposition 4.11 shows that all designs in the class Loc$_n(\psi, p(x))$ are first-order equivalent to simple random sampling (iid treatments). Then, as a tool for studying design theory, this assumption is self-defeating: by imposing asymptotic regression consistency, first-order asymptotics lose all power to differentiate between designs. Considering this, we argue that such consistency assumptions are too strong to provide guidance for practice in experimental design.

In particular, Proposition 4.11 shows that the naive design $D_i \overset{\text{inid}}{\sim} p(X_i)$ is asymptotically equivalent to the significantly more balanced designs in the class Loc$_n(\psi, p(x))$ under this assumption. Independent sampling amounts to creating covariate imbalances between treatment arms by design, but promising to correct them later with regression adjustment.

---

18Such imbalances are mean-zero (at the known propensity weights), creating variance but not asymptotic bias. For $D_{1:n} \sim \text{SRS}(p)$, the assignment variance is $E[b(X)^2]$. Compare this to $E[(b - E[b|\psi])^2]$ achieved by local randomization.
This is an absurd recommendation for practice. For instance, consider that by Theorem 3.19 assigning $D_{1:n} \sim \text{Loc}_n(X, p(X))$ gives asymptotic full efficiency by design, without the need for ex-post correction of imbalances by regression estimation.

**Remark 4.13** (Robust Design with a Pilot). The full efficiency statement of Theorem 4.8 required assuming pilot regression consistency, which we just argued is too strong for experimental design theory. Alternatively, note that Theorem 3.17 gives the asymptotic variance of local randomization with $\psi(x) = (\hat{m}_1(x), \hat{m}_0(x))$ considered to be a fixed (not changing with $n$) random function. Using this result, we can remain agnostic about the specification of the outcome model. If we also include some predictive covariates $x_{pred}$ outright, setting $\psi(x) = (\hat{m}_1(x), \hat{m}_0(x), x_{pred})$, then Theorem 3.17 gives the asymptotic variance components

\[
\text{selection} = q \text{Var}(c(X)) + (1 - q)E[(c(X) - E[c(X)|\hat{m}_1(X), \hat{m}_0(X), X_{pred}])^2]
\]

\[
\text{assignment} = E[(b(X) - E[b(X)|\hat{m}_1(X), \hat{m}_0(X), X_{pred}])^2]
\]

By additionally including $x_{pred}$, researchers can robustify the design against both (1) misspecification of pilot regressions and (2) pilot estimation error. See also the discussion of matching on regressions estimated in proxy studies in Example 3.10.

**Remark 4.14** (Comparison with Bai (2020)). As mentioned above, Theorem 4.8 extends the large pilot CLT of Bai (2020). In our notation, Theorem 5.1 of Bai gives the case $T_{1:n} = 1$ (no selection) and $D_{1:n} \sim \text{Loc}_n(\hat{b}, a/k)$. See section 5.3 of Bai for an alternative approach to robustness motivated by optimality in a Bayesian framework under the assumption of linear conditional expectations $m_d(x) = \beta'_d x$.

## 5 Inference Methods

This section gives procedures for valid inference on average treatment effects under local randomization. Theorem 5.3 gives asymptotically exact inference for the ATE and fixed regressor estimand $E_n[c(X_i)]$ under a two-stage design where both selection into the experiment and treatment assignment are done by covariate-adaptive local randomization. In particular, we give novel methods allowing experimenters to report smaller uncertainty about treatment effects when selection into the experiment is representative of the heterogeneity in treatment effects. We also illustrate an interesting connection between the matching estimators defined below and classical permutation inference. See Remark 5.4 for a discussion.

**Group Replicates** - As noted in Fisher (1926), exact inference requires replicates of each treatment arm under similar experimental conditions. For $D_{1:n} \sim \text{Loc}_n(\psi, p)$, ideal data would contain exact replicates of the outcomes $(Y_i)_{i \in g}$ in each local group under identical observed experimental conditions (identical $(\psi_i, p_i)$ values). To approximate this ideal situation, we match each group $g$ to a neighboring group in $\psi(x)$, giving approximate replicates of the outcomes in $g$ under similar experimental conditions.

**Matching Construction** - Our approach to finding approximate replicates is based on computing a pairwise matches between the centroids of each group in $\psi(x)$ space. In particular, we do the following...
(1) Form the group centroids \( \bar{\psi}_g = \frac{1}{|g|} \sum_{i \in g} \psi_i \)

(2) For each distinct propensity level \( p_a \), compute a pairwise matching of groups \( g \to g' \) within this propensity level by treating each centroid \( \bar{\psi}_g \) as an experimental unit and using one of the algorithms in Section 7.

(3) For matched groups \( (g, g') \), choose a random matching of treated to treated units, and control to control units. In notation, randomly choose a bijective matching \( \gamma: \{ i \in g, D_i = 1 \} \to \{ i \in g', D_i = 1 \} \)

Proposition 7.8 in the technical material shows that if the local groups \( g \) satisfy our strong homogeneity condition (Equation 3.2), the centroid construction above produces a one-to-one matching \( \gamma \) satisfying the following approximate replicate condition. This is a key condition for the proof of Theorem 5.3 below, and the construction of exact variance estimators.

**Definition 5.1** (Approximate Replicates). The matching function \( \gamma: [n] \to [n] \) gives approximate replicates if

(i) \( D_i = D_{\gamma(i)} \) and \( \gamma: \bigcup_s g_{a,s} \to \bigcup_s g_{a,s} \)

(ii) \( \gamma^2 = \text{Id}, \ gamma(i) \neq i, \) and \( \gamma \in \sigma((g_{a,s})_{a,s}, D_{1:n}, \xi_n) \)

(iii) \( E_n[(\psi_i - \psi_{\gamma(i)})^2 | T_i = 1] = O_p(r^2_n) \)

Part (i) requires that matched units are assigned to the same treatment arm and have the same propensity score \( p_n(X_i) = p_a \). By (ii), \( \gamma \) is a well-defined bijective matching, only depending on the groups, treatment values, and external randomness \( \xi_n \). Condition (iii) is the approximate replicate condition, requiring that matched units have approximately the same experimental conditions (on average).

Next, we use this matching to construct variance estimators. For intuition about the form of these matching estimators and a connection to permutation inference, see the discussion in Remark 5.4 below.

**Definition 5.2** (Variance Estimators). Let \( G \) be a bounded, Lipschitz function. Let \( \gamma \) be a matching satisfying the conditions of Definition 5.1. With \( p_{i,n} = p_n(X_i) \), define the matching estimators

\[
\hat{v}_1[G] \equiv E_n \left[ \frac{T_i D_i}{q p_{i,n}} G(p_{i,n}) Y_i Y_{\gamma(i)} \right] \\
\hat{v}_0[G] \equiv E_n \left[ \frac{T_i (1 - D_i)}{q (1 - p_{i,n})} G(p_{i,n}) Y_i Y_{\gamma(i)} \right]
\]

Also define the cross-moment estimator and IPW sample moment

\[
\hat{v}_{10} \equiv 2n^{-1} \sum_{1 \leq i < j \leq n} \frac{T_i T_j}{q} \mathbb{1}(g(i) = g(j)) \frac{D_i (1 - D_j) Y_i Y_j}{p_{i,n} (1 - p_{j,n})} \\
\hat{v}_{1,2} = E_n \left[ \frac{T_i}{q} \left( \frac{(D_i - p_{i,n}) Y_i}{p_{i,n} - p_{i,n}^2} \right)^2 \right]
\]

The following theorem shows exact or valid (but conservative) inference for various causal estimands and designs by varying the weights \( G(p) \) in the estimators above. The degree of conservativeness in the non-exact cases is also given. We summarize the results as follows:
1. **Exact Inference**

   (a) \( \theta = \text{ATE}, T_{1:n} \sim \text{Loc}_n(\psi, q) \) and \( D_{1:n} \sim \text{Loc}_n(\psi, p(x)|T_{1:n}) \), with \( p(x) = p(\psi) \)
   
   (b) \( \theta = \text{ATE}, T_{1:n} = 1 \) and \( D_{1:n} \sim \text{Loc}_n(\psi, p(x)) \)
   
   (c) \( \theta = E_n[c(X_i)], T_{1:n} \sim \text{Loc}_n(X, q) \) and \( D_{1:n} \sim \text{Loc}_n(X, p(x)|T_{1:n}) \)

2. **Valid Inference**

   (a) \( \theta = \text{ATE}, \) general case \( T_{1:n} \sim \text{Loc}_n(\psi', q) \) and \( D_{1:n} \sim \text{Loc}_n(\psi, p(x)|T_{1:n}) \)

   (b) \( \theta = E_n[c(X_i)], T_{1:n} \sim \text{Loc}_n(\psi, q) \) and \( D_{1:n} \sim \text{Loc}_n(\psi, p(x)|T_{1:n}), p = p(\psi) \)

   (c) \( \theta = E_n[c(X_i)], T_{1:n} = 1 \) and \( D_{1:n} \sim \text{Loc}_n(\psi, p(x)) \)

**Theorem 5.3 (Inference).** Suppose that assumptions 8.1, 8.3, and 8.11 are satisfied.

1. Suppose that \( T_{1:n} \sim \text{Loc}_n(\psi', q) \) and \( D_{1:n} \sim \text{Loc}_n(\psi, p|T_{1:n}) \). If \( \psi' = \psi \) and \( p = p(\psi, \xi) \) then

   \[
   \begin{align*}
   \hat{v}_{1,2} - \hat{v}_{1} \left[ \frac{1 - q\theta}{p} \right] - \hat{v}_{0} \left[ \frac{1 - q(1 - p)}{1 - p} \right] - 2q\hat{v}_{10} - q\hat{\theta}^2 = V(\psi, (q, p)) + o_p(1)
   \end{align*}
   
   For the fixed regressor case

   \[
   \begin{align*}
   \hat{v}_{1,2} - \hat{v}_{1} \left[ p^{-1} \right] - \hat{v}_{0} \left[ (1 - p)^{-1} \right] \\
   = V(\psi, (q, p)) - q \text{Var}(c(X)) + qE[\text{Var}(c(X_i)|\psi, \xi)|\xi] + o_p(1)
   \end{align*}
   \]

   \[
   \geq V(\psi, (q, p)) - q \text{Var}(c(X)) + o_p(1)
   \]

2. Suppose that \( T_{1:n} = 1 \) and \( D_{1:n} \sim \text{Loc}_n(\psi, p) \) then with \( \text{Var}_n \) the sample variance

   \[
   \begin{align*}
   \text{Var}_n \left( \frac{(D_i - p_{i,n})Y_i}{p_{i,n} - p^2_{i,n}} \right) - \hat{v}_{1} \left[ \frac{1 - p}{p} \right] - \hat{v}_{0} \left[ \frac{p}{1 - p} \right] - 2\hat{v}_{10} = V(\psi, p) + o_p(1)
   \end{align*}
   
   For the fixed regressor case

   \[
   \hat{v}_{1,2} - \hat{v}_{1} \left[ p^{-1} \right] - \hat{v}_{0} \left[ (1 - p)^{-1} \right] \geq V(\psi, p) - \text{Var}(c(X)) + o_p(1)
   \]

3. Suppose that \( T_{1:n} \sim \text{Loc}_n(\psi', q) \) and \( D_{1:n} \sim \text{Loc}_n(\psi, p|T_{1:n}) \)

   \[
   \begin{align*}
   \text{Var}_n \left( \frac{(D_i - p_{i,n})Y_i}{p_{i,n} - p^2_{i,n}} \right) - \hat{v}_{1} \left[ \frac{1 - p}{p} \right] - \hat{v}_{0} \left[ \frac{p}{1 - p} \right] \\
   = V((\psi', \psi), (q, p)) + (1 - q) \text{Var}(E[c(X_i)|\psi', \xi]|\xi) + o_p(1) \\
   \geq V((\psi', \psi), (q, p)) + o_p(1)
   \end{align*}
   
See section 8.4 of the appendix for the proof.

**Remark 5.4 (Randomization Inference).** In this remark, we argue that the inference strategy in Theorem 5.3 can be interpreted as an *estimation version* of classical randomization inference. Consider case (2) above with \( T_{1:n} = 1 \) and \( D_{1:n} \sim \text{Loc}_n(\psi, p(x)) \). The appendix shows that, for treatment assignments \( D_{1:n} \sim \text{Loc}_n(\psi, p(x)) \), the first term is
consistent for the asymptotic variance of the least balanced design: \( D_i \sim p(X_i) \), also known as simple random sampling (SRS)

\[
\text{Var}_n \left( \frac{(D_i - p_{i,n})Y_i}{p_{i,n} - Y_i^2} \right) \overset{p}{\to} E \left[ \frac{Y(1)^2}{p(X_i)} \right] + E \left[ \frac{Y(0)^2}{1 - p(X_i)} \right] = \text{Var}(\text{SRS}(p))
\]

Our inference method effectively starts with an estimate of the variance under the least efficient design, then reduces it by subtracting off the correction \( \hat{v}_1 + \hat{v}_0 \). Thus, our estimate of uncertainty about the average treatment effect is small when the correction terms \( \hat{v}_1 \) and \( \hat{v}_0 \) are large.

**Measuring Homogeneity:** To see when \( \hat{v}_1 \) will be large, consider the matched products \((Y_iY_i(0))_i\) in the definition. By the Hardy-Littlewood inequality, \( \hat{v}_1 \) is maximized when matched outcomes are maximally homogeneous: large outcomes \( Y_i \) multiplied by large outcomes \( Y_i(0) \) and vice-versa. Then we can regard \( \hat{v}_1 \) as a measure of the average homogeneity of potential outcomes \( Y_i(1) \) within local groups.

**Permuting Treatments:** The more homogeneous \( Y_i(1) \) and \( Y_i(0) \) are within local groups, the smaller the variance due to treatment assignment. To see this, imagine rerandomizing by permuting the treatment assignment variables \((D_i)_i \in g\) in each group \( g \). If the potential outcome \( Y_i(1) \) is approximately constant over \( i \in g \) (and the same for \( Y_i(0) \)), permuting treatments in this way will have almost no effect on the IPW estimator, since the weighted difference of observed outcomes \( Y_i \) will be approximately unchanged within each group, and thus globally. Summarizing this discussion, on average over the local groups the following are equivalent

1. Correction terms \( \hat{v}_1 + \hat{v}_0 \) are large
2. Potential outcomes \((Y_i(1), Y_i(0))_i \in g\) are homogeneous
3. Estimator variance due to permuting treatments \((D_i)_i \in g\) is small

**Remark 5.5** (Inference Methods for Covariate-Adaptive Selection). Note that the first result requires selection and assignment to be locally randomized using the same function \( \psi'(x) = \psi(x) \). By enforcing this condition, we make the selection and assignment variance components comparable

\[
E[(c - E[c|\psi'])^2] \quad E[(b - E[b|\psi])^2]
\]

In particular, they are both weighted combinations of the residuals \( m_d(x) - E[m_d(x)|\psi(x)] \), \( d = 0, 1 \). This allows us to give exact inference for the two-step process of covariate-adaptive selection and assignment, even though the variance corrections \( \hat{v}_1, \hat{v}_0, \hat{v}_{10} \) are only defined using local groups formed during the treatment assignment stage of the design. Effectively, we require that the “level of covariate balance” during the selection and assignment steps is the same.

This requirement may be undesirable when, as in Example 3.3, extra covariates are gathered about each unit after committing enrollment it in the study \((T_i = 1)\), with

---

19Formally, for the case with weights \( G = 1 \), the optimum is achieved by matching adjacent order statistics within treatment class, \( Y_{(1)}Y_{(2)} + Y_{(3)}Y_{(4)} + \cdots \)
outcome models of the form 

\[ Y_i^{\text{pre}} \subseteq X_i. \]

For this case, conservative inference is given by the third result. It may be possible to use ex-post matching methods, as in Abadie and Imbens (2012), to give exact inference for this case. We leave such extensions to future work.

**Remark 5.6** (Model-free Conditional Inference). Recall that the assignment variance depends on the non-parametric regression residuals \( E[(b - E[b|\psi])^2] \). However, the variance corrections above only use local group structure of the design and realized outcomes \( Y_i \), without explicitly estimating a regression. In particular, these corrections are agnostic to the form of the CATE and balance function, as well as to the true non-parametric regression models \( E[c|\psi], E[b|\psi] \) for these objects. For \( D_{1:n} \sim \text{Loc}_n(\psi, p) \) and \( \psi = \psi(\cdot, \zeta) \) a function estimated on proxy data \( \zeta \), the proof of Theorem 5.3 shows that

\[
\hat{\gamma}_1[1] = c_1 \cdot R^2(\psi, \zeta) + c_2 + o_p(1) \quad R^2(\psi, \zeta) = \frac{\text{Var}(E[m_1(X)|\psi(X, \zeta), \zeta]|\zeta)}{\text{Var}(m_1(X))}
\]

with constants \( c_1, c_2 \) independent of the design. \( R^2(\psi, \zeta) \) measures the non-parametric predictive power of \( \psi(\cdot, \zeta) \) for \( m_1(X) \) at the realized value of \( \zeta \). Thus, conditionally valid inference is achieved by measuring how well \( \psi(\cdot, \zeta) \) (non-parametrically) controls for variation in \( b(X), c(X) \) ex-post, without actually specifying or estimating a regression model.

**Example 5.7** (Complete Randomization). Consider inference for \( \theta = \text{ATE} \) in a completely randomized experiment \( D_{1:n} \sim \text{CR}(1/2) \). The sample variance 16 \( \text{Var}_n((D_i - 1/2)Y_i) \), while exact if \( D_i \) are iid, is conservative under complete randomization due correlation of the treatment assignments. Proposition 8.15 shows that \( \text{CR}(1/2) = \text{Loc}_n(1, 1/2) \), giving a “random matched pairs” representation of \( \text{CR}(1/2) \). Then, consider the following inference strategy: let \( \gamma, \phi : [n] \to [n] \) random, bijective matchings with \( D_i = D_{\gamma(i)} \) and \( D_i = 1 - D_{\phi(i)} \), respectively. In this case, the variance corrections of Theorem 5.3 have

\[
\hat{\gamma}_1[1] + \hat{\theta}_0[1] = 2E_n \left[ (D_i + (1 - D_i))Y_i \gamma(i) \right] = 2E_n \left[ Y_i \gamma(i) \right]
\]

Then by \( \text{Theorem 5.3} \), the following variance estimator is exact

\[
\hat{V} = 16 \text{Var}_n((D_i - 1/2)Y_i) - 2E_n \left[ Y_i \gamma(i) \right] - 2E_n \left[ Y_i \phi(i) \right]
\]

### 6 Empirical Results

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

#### 6.1 Simulations

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

\[
Y_i(0) = X_i'\beta(0) + \sigma_0^2(X_i)\epsilon_i(0) \quad Y_i(1) = X_i'\beta(1) + X_i'AX_i + \sigma_0^2(X_i)\epsilon_i(1)
\]

---

\(^20\) In particular, \( c_1 = \text{Var}(m_1(X)) \) and \( c_2 = E[Y(1)]^2 \)

\(^21\) Since the matched pairs in \( \text{Loc}_n(\psi = 1, 1/2) \) are formed randomly, it suffices to use the random matchings \( \gamma, \phi \) to define local groups ex-post.
We sample from the following DGP’s

**Model 1**: \( \beta(0) = (1,1), \beta(1) = (2,2), A_{12} = A_{21} = 1/2, A_{11} = A_{00} = 0, \) with residual variance \( \sigma_1^2 = \sigma_0^2 = 0.1 \)

**Model 2**: As in (1) but with \( \beta(0) = (1,1), \beta(1) = (2,2), A = 0 \)

**Model 3**: As in (1) but with \( \beta(0) = \beta(1) = (1,1), A = 0 \)

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

**Designs** - In Table 1 we vary \( n \), the number of units eligible for selection, fixing the experiment size \( qn = \sum T_i = 100 \). To highlight the marginal efficiency gains from representative selection, while also using locally randomized treatment assignment, we fix the assignment procedure to be “matched triples” on \( \psi(x) = (x_1, x_2) \) with \( p = 2/3 \), as in example 3.12. In our notation, we let selections \( T_{1:n} \) and treatment assignments \( D_{1:n} \) be given by

1. \( T_{1:n} \sim \text{Loc}_n(\psi, q) \) (selection) with \( n = 100/q \)
2. \( D_{1:n} \sim \text{Loc}_n(\psi, 2/3) \) (treatment assignment)

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

**Results** - The variance reduction increases as the number of eligible units \( n = 100/q \) is larger (\( q \) is smaller). This is expected from Theorem 3.17, which gives a reduction asymptotic variance of \( -(1 - q) \text{Var}(c(X)) \) from local randomization. The boost in effective sample size (ESS) is significant relative to completely random selection. Note that this comes essentially “for free” by being slightly more careful about choosing which units to enroll in the experiment. In Model 3, \( c(x) = 0 \) identically, so the theory predicts no efficiency gain due to selection. The small variance reduction seen in practice is likely a finite sample effect where balanced first-stage selection facilitates better matches during the second-stage assignment process.

Table 1 suggests our inference procedure is generally slightly conservative in finite samples. This is likely due to the between-group matching used for inference in Section 5 giving slightly worse matches than the groups produced by the design itself. The argument in Remark 5.4 shows that worse matches lead to smaller correction terms, and thus larger confidence intervals. For CI length, finite-sample exact inference would have \( %\Delta Length = %\Delta SD \). In finite samples, our inference procedure gives a significant reduction in CI length, reflecting the increase in estimator precision due to selection.

Our second set of simulations focuses on the choice of \( \psi(x) \). Asymptotically, including more baseline covariates always weakly improves estimator efficiency. However, as
Table 1: Effect of Representative Selection

<table>
<thead>
<tr>
<th>Model</th>
<th>$n$</th>
<th>$%\Delta SD$</th>
<th>ESS</th>
<th>Coverage</th>
<th>$%\Delta Length$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>0.0</td>
<td>100</td>
<td>95</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>-17.4</td>
<td>147</td>
<td>97</td>
<td>-10.3</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>-22.8</td>
<td>168</td>
<td>97</td>
<td>-13.6</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>-26.4</td>
<td>185</td>
<td>97</td>
<td>-15.9</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>0.0</td>
<td>100</td>
<td>94</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>-17.7</td>
<td>148</td>
<td>96</td>
<td>-9.5</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>-25.1</td>
<td>178</td>
<td>98</td>
<td>-12.4</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>-28.6</td>
<td>196</td>
<td>97</td>
<td>-14.5</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>0.0</td>
<td>100</td>
<td>96</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>-0.4</td>
<td>101</td>
<td>95</td>
<td>-4.2</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>-5.5</td>
<td>112</td>
<td>96</td>
<td>-4.9</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>-10.6</td>
<td>125</td>
<td>97</td>
<td>-5.6</td>
</tr>
</tbody>
</table>

discussed in Remark 4.6, the within-group matching discrepancies converge slowly in high dimensions, giving balancing rate $r_\psi_n = O(n^{-2/(d+1)})$. This suggests preferentially including covariates that are predictive of outcomes and treatment effect heterogeneity (predict functions $b_n(x)$ and $c(x)$), excluding “noise covariates” that contain little additional information. Consider the quadratic model and variable distributions above, but with $X_i \in \mathbb{R}^{10}$ and

$$\beta(0) = (1, 1, 1, .4, .4, .4, 0, 0, 0, 0) \quad \beta(1) = 2\beta(1) \quad A = (1/20)(I_1 - \text{diag}(I))$$

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

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

In Table 2 below, $\psi^*(x) = (b_n(x), c(x))$, the optimal design for joint selection and assignment given in Section 3.2. The evaluation criteria are the same as above, comparing to the least efficient design with selection and assignment done by complete randomization (CR).

**Results** - There are significant precision gains from locally randomized selection and assignment, as predicted by Theorem 3.17. Theorem 5.3 shows that our inference procedures are asymptotically exact, but we get slight conservativeness in finite samples, likely due to between-group match quality during inference being worse than the within-group matches. Note that, contrary to the asymptotic theory, including $x_1 \ldots x_6$ is slightly
worse than $x_1 \ldots x_4$ in finite samples. Recall have the largest weight in the outcome model. Additionally including $x_4 \ldots x_6$, which are less predictive of outcomes, degrades match quality for the important covariates $x_1 \ldots x_3$, reducing efficiency in finite samples. Including noise covariates $x_7, x_8$ performs still worse. As in Theorem 3.19, the optimal design $\psi^*(x) = (b_n(x), c(x))$ gives the largest efficiency gains.

<table>
<thead>
<tr>
<th>$(n, qn)$</th>
<th>$\psi(x)$</th>
<th>Efficiency</th>
<th>Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(200, 100)$</td>
<td>CR</td>
<td>0.0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>$x_1, x_2$</td>
<td>-28.7</td>
<td>197</td>
</tr>
<tr>
<td></td>
<td>$x_1 \ldots x_4$</td>
<td>-37.4</td>
<td>256</td>
</tr>
<tr>
<td></td>
<td>$x_1 \ldots x_6$</td>
<td>-36.0</td>
<td>244</td>
</tr>
<tr>
<td></td>
<td>$x_1 \ldots x_8$</td>
<td>-33.9</td>
<td>229</td>
</tr>
<tr>
<td></td>
<td>$\psi^*(x)$</td>
<td>-57.1</td>
<td>543</td>
</tr>
<tr>
<td>$(400, 200)$</td>
<td>CR</td>
<td>0.0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>$x_1, x_2$</td>
<td>-27.2</td>
<td>188</td>
</tr>
<tr>
<td></td>
<td>$x_1 \ldots x_4$</td>
<td>-40.4</td>
<td>282</td>
</tr>
<tr>
<td></td>
<td>$x_1 \ldots x_6$</td>
<td>-36.5</td>
<td>248</td>
</tr>
<tr>
<td></td>
<td>$x_1 \ldots x_8$</td>
<td>-33.2</td>
<td>224</td>
</tr>
<tr>
<td></td>
<td>$\psi^*(x)$</td>
<td>-56.7</td>
<td>534</td>
</tr>
</tbody>
</table>

Table 2: Varying $\psi(x)$
References


7 Local Balancing Rates

This section collects technical results guaranteeing the local balancing rate $r_n^\psi = o(1)$.

**Definition 7.1** (Uniformly Piecewise Hölder). Let $\dim(\psi) = 1$. Let $\zeta_n$ be a random variable and $(h_n)_{n \geq 1}$ a sequence of deterministic functions. $(h_n)_{n \geq 1}$ is uniformly piecewise Hölder with bounded breaks if there exists $0 < \alpha \leq 1$

(a) There exists break sets $B_n(a, \zeta_n) \subseteq \mathbb{R}$ and $C_n(\zeta_n)$ such that $\forall p_a \in L_n$

$$|h_n(\psi, p_a, \zeta_n) - h_n(\psi', p_a, \zeta_n)| \leq C_n|\psi - \psi'|^\alpha \quad \forall \psi, \psi': [\psi, \psi'] \cap B_n(a, \zeta_n) = \emptyset$$

with $\max_{p_a \in L_n} \text{Card}(B_n(a, \zeta_n)) \leq B < \infty$ and $C_n(\zeta_n) < C < \infty$, $\zeta_n$-a.s.

(b) $|h_n(\psi, p_a, \zeta_n) - h_n(\psi', p_a, \zeta_n)| \leq \bar{M} < \infty$ for all $\psi, \psi', p_a, \zeta_n$-a.s.

Note that $\alpha = 1, \zeta_n = 1$ corresponds to a piecewise Lipschitz assumption.

**Example 7.2** (Stratified Block Randomization). Recall that by Proposition 8.15, SBR with discrete strata $S(X) \in \{1, \ldots, T\}$ is equivalent to $D_{1,n} \sim \text{Loc}_{\zeta_1}(S, p)$. Let $F(s, p)$ be a piecewise constant version\(^{22}\) of $E[b(X)|S(X) = s, p(S(X)) = p]$ on $[1, T] \times (0, 1)$. Then $F_n = F$ satisfies the conditions of definition 7.1 with $C_n = 0$, break sets $B_{n, p} = \{1, \ldots, T\}$, and $M_n = \max_{s, p} F(s, p) - \min_{s, p} F(s, p)$, and $\alpha = 1$.

Consider a simple rank-statistic based approach for forming the local groups $(g_a, s)_{a, s}$

**Algorithm 7.3** (Rank-Ordering). Require $\dim(\psi) = 1$. Set $I_a = \{i : p_n(X_i, \zeta_n) = p_a\}$ for each $p_a \in L_n$; set $s = 1$ and form index groups $(g_a, s)_{s=1}^n$ as follows:

(i) Sort the indices in $I_a$ by increasing $\psi_i$ value, producing $I_a^{\text{sort}} = (i_1, i_2, \ldots, i_{m(a)})$.

Ties are broken with data-independent randomness $\pi_n$.

(ii) Define $g_a, s = \{i_1, \ldots, i_{k_a}\}$

(iii) Increment $s \rightarrow s + 1$ and $I_a \rightarrow I_a \setminus g_a, s$ and return to step (i)

Under the piecewise Hölder assumption and $\dim(\psi) = 1$, rank ordering is a simple way to guarantee strong-balancing, as the next proposition shows

**Proposition 7.4** (Balancing 1). Let $\dim(\psi) = 1$ and $\psi(X, \zeta_n) \in [a, b]$ compact $\zeta_n$-a.s. Let $(g_a, s)_{a, s}$ be constructed as in Algorithm 7.3. Let $h_{i,n} = h_n(\psi(X_i, \zeta_n), p_n(X_i, \zeta_n), \zeta_n)$ and suppose that $(h_n)_{n \geq 1}$ is uniformly piecewise Hölder as in Definition 7.1 with $\alpha \geq 1/2$. Then groups $(g_a, s)_{a, s}$ satisfy

$$n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^n k_a^{-1} \sum_{i, j \in g_a, s, i \neq j} (h_{i,n} - h_{j,n})^2 = O(r_n^\psi) = o(1) \quad (7.1)$$

In particular, Equation 3.2 holds with rate $r_n^\psi = \bar{F}_n |L_n| n^{-1}$

For $\dim(\psi) > 1$, Algorithm 7.3 is no longer feasible, and we take an explicit optimization approach. We impose a slightly stronger assumption.

\(^{22}\)Since $S(X) \in \{1, \ldots, T\}$ w.p.1, we may define $E[b(X)|S(X) = s, P(S(X)) = p]$ arbitrarily on $\mathbb{R} \setminus [T]$, and still get a version of the conditional expectation. Then take $F(s, p) = F([s], p) \forall s \in [1, T]$. 

44
Then the slower rate is achieved and suppose that
\[ h_n(\psi, p_a, \zeta_n) - h_n(\psi', p_a, \zeta_n) \leq C_n(\zeta_n)|\psi - \psi'| \]
with \( C_n < C < \infty, \zeta_n\text{-a.s.} \)

**Proposition 7.6** (Balancing II). Let \( \dim(\psi) = d_\psi \). Let \( h_{i,n} = h_n(\psi(X_i, \zeta_n), p_n(X_i, \zeta_n), \zeta_n) \) and suppose that \( (h_n)_{n \geq 1} \) is uniformly Lipschitz as in Definition 7.5. With \( \psi_{i,n} = \psi(X_i, \zeta_n) \), consider groups \((g_{a,s})_{a,s}\) formed by solving the program

\[
\min_{(g_{a,s})_{a,s}} n^{-1} \sum_{i=1}^{n} k^{-1}_{a,s} \sum_{i,j \in g_{a,s}, i \neq j} |\psi_{i,n} - \psi_{j,n}|^2 \tag{7.2}
\]

subject to

\[
p_n(X_i, \zeta_n) = p_a \quad \forall i \in g_{a,s} \quad \forall s = 1, \ldots, n \tag{7.3}
\]

Then the balancing rate

\[
n^{-1} \sum_{i=1}^{n} \sum_{s=1}^{L_n} k^{-1}_{a,s} \sum_{i,j \in g_{a,s}, i \neq j} (h_{i,n} - h_{j,n})^2 = O\left(\left(n/(\bar{F}_n|L_n|)\right)^{-2/(d_\psi+1)}\right)
\]

Alternatively, if \((g_{a,s})_{a,s}\) solve the modified program

\[
\min_{(g_{a,s})_{a,s}} n^{-1} \sum_{i=1}^{n} k^{-1}_{a,s} \sum_{i,j \in g_{a,s}, i \neq j} |\psi_{i,n} - \psi_{j,n}|^2 \tag{7.4}
\]

subject to

\[
p_n(X_i, \zeta_n) = p_a \quad \forall i \in g_{a,s} \quad \forall s = 1, \ldots, n \tag{7.5}
\]

Then the slower rate is achieved

\[
n^{-1} \sum_{i=1}^{n} \sum_{s=1}^{L_n} k^{-1}_{a,s} \sum_{i,j \in g_{a,s}, i \neq j} (h_{i,n} - h_{j,n})^2 = O\left(\left(n/(\bar{F}_n|L_n|)\right)^{-1/d_\psi}\right)
\]

In particular, if \( \bar{F}_n|L_n|n^{-1} = o(1) \) then \((g_{a,s})_{a,s}\) satisfy Equation 3.2.

**Remark 7.7** (Matched Pair Designs). As above, the case \( D_{1:n} \sim \text{Loc}_n(\psi, 1/2) \) for \( \psi(X) = X \) is equivalent to classical pairwise matching, as recently analyzed by Bai et al. (2021). In this case, \( \bar{F}_n = |L_n| = 2 \) for all \( n \geq 1 \). Plugging in to the slow rate in the proposition, we get \( n^{-1/d} \), recovering the rate of that paper. The (new) fast rate in this case is \( n^{-2/(d+1)} \), which requires instead minimizing the sum of squares above.

The following additional result on finding a well-balanced between-group matching is needed for variance estimation in Theorem 5.3 below.

**Proposition 7.8** (Balancing III). Let \( \dim(\psi) = d_\psi \) and consider groups \((g_{a,s})_{a,s}\) with balancing rate \( r_\psi^a = o(1) \). Define the group representative \( \psi_{a,s} = k^{-1}_{a,s} \sum_{i \in g_{a,s}} \psi_i \). For each \( a = 1, \ldots, |L_n| \), form between-group matching \( \nu_{a} : [n] \to [n] \), \( \nu_{a}^2 = \text{Id} \) as the solution to

\[
\min_{\nu_{a}} \sum_{s=1}^{n} \|\psi_{a,s} - \psi_{a,\nu_{a}(s)}\|^2.
\]

Consider \( \gamma : [n] \to [n] \) with \( \gamma^2 = \text{Id} \) any matching of
units subordinate to the optimal group matching. Specifically, require $\gamma(i) \in g_{a,\nu_a(s)}$ for each $i \in g_{a,s}$. Let selection variables $T_{1:n}$ satisfy assumption 8.3. Then the between-group balancing rate

$$E_n[T_i(\psi_i - \psi_{\gamma(i)})^2] = O_p\left(\frac{n}{(k_n|L_n|)^{-2/(d_\psi+1)}}\right) + O_p(r_n^{\psi})$$

(7.6)

For the proof of these propositions, see section 8.1 in the appendix.
8 Proofs

8.1 Proofs - Balancing

8.1.1 Proof of Proposition 7.4

Proof. Consider Hölder coefficient $1/2 \leq \alpha \leq 1$. Let $\psi_{i,n} = \psi(X_i, \zeta_n)$, let $\psi^d_{a,s,n} \equiv \min_{i \in g_{a,s}} \psi_{i,n}$ and $\psi^u_{a,s,n} \equiv \max_{i \in g_{a,s}} \psi_{i,n}$. Define $\text{diam}(g_{a,s}, \psi(\cdot, \zeta_n)) \equiv \psi^u_{a,s,n} - \psi^d_{a,s,n}$ if $g_{a,s} \neq \emptyset$, and 0 otherwise. Then since $x^{2\alpha} = x^{2(\alpha-1)/2} \leq x$ for all $0 \leq x \leq 1$, observe that

$$\text{diam}(g_{a,s}, \psi(\cdot, \zeta_n))^{2\alpha} = (b - a)^{2\alpha} \left( \frac{\text{diam}(g_{a,s}, \psi(\cdot, \zeta_n))}{b - a} \right)^{2\alpha} \leq (b - a)^{2\alpha - 1} \text{diam}(g_{a,s}, \psi(\cdot, \zeta_n)) \lesssim \text{diam}(g_{a,s}, \psi(\cdot, \zeta_n))$$

Using this fact, observe that

$$n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_{a}^{-1} \sum_{i,j \in g_{a,s}, i \neq j} |\psi_{i,n} - \psi_{j,n}|^{2\alpha} \leq n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_{a}^{-2} - k_{a}^{-1} \text{diam}(g_{a,s}, \psi(\cdot, \zeta_n))^{2\alpha} \mathbb{1}(g_{a,s} \neq \emptyset)$$

$$\lesssim n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} k_{a} \sum_{s=1}^{n} \text{diam}(g_{a,s}, \psi(\cdot, \zeta_n)) \mathbb{1}(g_{a,s} \neq \emptyset) \leq n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} k_{a} \cdot (1 - 0) = O(\lfloor L_n \rfloor |\kappa_n| n^{-1})$$

For the final equality, for $g_{a,s} \neq \emptyset$ define open intervals $I_{a,s,n} \equiv (\psi^d_{a,s,n}, \psi^u_{a,s,n}) \subseteq [0, 1]$. By (i) uniqueness of group membership, and (ii) the ordered construction of the groups within each level $a$, the intervals $(I_{a,s,n})_{s=1}^{n}$ are mutually disjoint for fixed $a$. Then we have

$$\sum_{s=1}^{n} \text{diam}(g_{a,s}, \psi(\cdot, \zeta_n)) = \sum_{s=1}^{n} \mathcal{L}(I_{a,s,n}) = \mathcal{L} \left( \bigcup_{s=1}^{n} \bigcup_{g_{a,s} \neq \emptyset} I_{a,s,n} \right) \leq \mathcal{L}([0, 1])$$

Define $E_{a,s} \equiv \{[\psi^d_{a,s,n}, \psi^u_{a,s,n}] \cap B_n(a, \zeta_n) = \emptyset\}$ for each $a, s$. (1) First, we analyze balance in the the groups where $E_{a,s}$ occurs

$$n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_{a}^{-1} \sum_{i,j \in g_{a,s}, i \neq j} (h_{n}(\chi_{X_i, \zeta_n}, p_{a}, \zeta_n) - h_{n}(\chi_{X_j, \zeta_n}, p_{a}, \zeta_n))^{2} \mathbb{1}(E_{a,s})$$

$$\leq n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_{a}^{-1} \sum_{i,j \in g_{a,s}, i \neq j} C_n(\zeta_n) |\psi_{i,n} - \psi_{j,n}|^{2\alpha} \mathbb{1}(E_{a,s}) = O(\lfloor L_n \rfloor |\kappa_n| n^{-1})$$

The first inequality is by the piecewise Lipschitz assumption, noting that $[\psi_{i,n}, \psi_{j,n}] \cap B_{a,n} = \emptyset \ \forall i, j \in g_{a,s}$ if $E_{a,s}$ occurs. The second equality drops the indicator $\mathbb{1}(E_{a,s})$ and uses the fact proved above.

(2) Next, we analyze the terms where $E_{a,s}^c$ occurs. Let $Q_{a,s} \equiv E_{a,s}^c \cap \{\text{diam}(g_{a,s}, \psi(\cdot, \zeta_n)) > \}$
0}. Note that on the event \( \{ \text{diam}(g_{a,s}, \psi(\cdot, \zeta_n)) = 0 \} \), we have \( \psi_{i,n} = \psi_{j,n} \ \forall i, j \in g_{a,s} \), so that \( h_n(\psi(X_i, \zeta_n), p_a, \zeta_n) - h_n(\psi(X_j, \zeta_n), p_a, \zeta_n) = 0 \). Then it suffices to consider the terms

\[
\sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_a^{-1} \sum_{i,j \in g_{a,s}, i \neq j} (h_n(\psi(X_i, \zeta_n), p_a, \zeta_n) - h_n(\psi(X_j, \zeta_n), p_a, \zeta_n))^2 I(Q_{a,s})
\]

\[
\leq M_n^2 n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_a^{-1} I(Q_{a,s}) \sum_{i,j \in g_{a,s}, i \neq j} 1 \leq M_n^2 n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} k_a^{-1} (k_a^2 - k_a) \sum_{s=1}^{n} I(Q_{a,s})
\]

To bound the inner sum, define the correspondence

\[
\varphi : B_n(a, \zeta_n) \mapsto \{s : I(Q_{a,s}) = 1\}
\]

\[
\varphi(b) \mapsto \{s : b \in T_{a,s}\}
\]

where \( T_{a,s} \equiv [\psi_{a,s}^d, \psi_{a,s}^u] \). This correspondence assigns each breakpoint to the set of group intervals it belongs to. We claim that \( \text{Card}(\varphi(b)) \leq 2 \) for all \( b \in B_n(a, \zeta_n) \). Suppose not, then we have \( b \in T_{a,s} \cap T_{a,s'} \cap T_{a,s''} \) for distinct intervals. Without loss suppose that \( \psi_{a,s}^d < \psi_{a,s}^u \leq \psi_{a,s'}^u < \psi_{a,s'}^l < \psi_{a,s''}^u \). The weak inequality follows from the grouping algorithm, and strict inequality since all indices are in \( \{s : \text{diam}(g_{a,s}, \psi(\cdot, \zeta_n)) > 0\} \). By assumption \( \psi_{a,s}^u \geq b \). Then we have \( \psi_{a,s'}^u \geq \psi_{a,s'}^l \geq \psi_{a,s}^u \geq b \). From the strict inequality, we see that \( b \not\in T_{a,s''} \), which is a contradiction. Clearly, the correspondence \( \varphi \) is a surjection, so we have

\[
\sum_{s=1}^{n} I(Q_{a,s}) = \text{Card}(\{s : I(Q_{a,s}) = 1\}) = \text{Card}\left( \bigcup_{b \in B_n(a, \zeta_n)} \varphi(b) \right)
\]

\[
\leq \sum_{b \in B_n(a, \zeta_n)} \text{Card}(\varphi(b)) \leq 2 \text{Card}(B_n(a, \zeta_n)) \leq 2 \tilde{B}
\]

Then continuing the display above gives

\[
M_n^2 n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} k_a^{-1} (k_a^2 - k_a) \sum_{s=1}^{n} I(Q_{a,s}) \leq M_n^2 \bar{k}_n n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} 2 \text{Card}(B_n(a, \zeta_n))
\]

\[
\leq 2 M_n^2 \tilde{B} |L_n| \bar{k}_n n^{-1} = O_p(\|L_n| \bar{k}_n n^{-1})
\]

Putting this all together, we have shown that

\[
\sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_a^{-1} \sum_{i,j \in g_{a,s}, i \neq j} (h_n(\psi(X_i, \zeta_n), p_a, \zeta_n) - h_n(\psi(X_j, \zeta_n), p_a, \zeta_n))^2 = O_p(\|L_n| \bar{k}_n n^{-1})
\]

This completes the proof.  

\[\square\]
8.1.2 Proof of Proposition 7.6

Proof. Suppose that $\psi(X, \zeta_n) \in [0, 1]^{d_\psi}$, $\zeta_n$-a.s. and define $\psi_{i,n} = \psi(X_i, \zeta_n)$. Fix $m > 1$ an integer. We call $(B_l)_{l=1}^{\infty}$ an ordered block partition of $[0, 1]^{d_\psi}$ if (1) $\text{diam}(B_l, \| \cdot \|_\infty) \leq 1/m$ and (2) for all $x \in B_l$ and $y \in B_{l+1}$ there exists $j^*(l)$ such that $|x_i - y_i| \leq 1/m$ for all $i \in [d_\psi]$ $i \neq j^*(l)$ and $|x_j^*(l) - y_j^*(l)| \leq 2/m$. Intuitively, the $(B_l)_l$ form a contiguous, exhaustive path through $[0, 1]^{d_\psi}$. See Bai et al. (2021) for an explicit construction. Abusing notation, define $l(i) = \min_{n=1}^{\infty} \{l : \psi_i \in B_l\}$.

Algorithm - Fix $a$ and form groups $(g_{a,s})_{s=1}^{n}$ by induction as follows. For all units with $l(i) = 1$ and $p_n(X_i, \zeta_n) = p_a$, form groups arbitrarily (possibly using external randomness $\pi_n$). This process results in at most one partially filled group (with less than $k_a$ units), say $g_{a,s'}$. Increment $l \rightarrow l + 1$, adding units to the group until $\text{Card}(g_{a,s'}) = k_a$. Suppose that $g_{a,s'}$ is completed with a unit from $B_{l'}$. Then repeat the process above starting with the next group $g_{a,s'} + 1$ and the units in block $B_{l'}$, excluding all units (with $p_n(X_i, \zeta_n) = p_a$) that are already in a group. Since there are $n < \infty$ units, this process terminates. Repeat this for each $a = 1, \ldots, |L_n|$. 

By construction, this creates groups $(g_{a,s})_{s=1}^{n}$ with the ordering property

$$l(i) \leq l(j) \quad \forall i \in g_{a,s}, j \in g_{a,s'} \quad s < s' \quad a = 1, \ldots, |L_n| \quad (8.1)$$

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

$$n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_a^{-1} \sum_{i,j \in g_{a,s}} \left| \psi_{i,n} - \psi_{j,n} \right|^2 = n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_a^{-1} \sum_{t=1}^{k_a^2 - k_a} \left| \psi_{i_{a,s,t},n} - \psi_{j_{a,s,t},n} \right|^2 1(g_{a,s} \neq \emptyset)$$

(1) Suppose $E_{a,s,t}$ occurs. Define $d_{a,s,t} \equiv \left| \psi_{i_{a,s,t},n} - \psi_{j_{a,s,t},n} \right| \leq \max_{t=1}^{k_a^2} \text{diam}(B_t, \| \cdot \|_2) \leq \sqrt{d_\psi/m}$ on this event. Then we may bound

$$n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_a^{-1} \sum_{t=1}^{k_a^2 - k_a} d_{a,s,t}^2 1(E_{a,s,t}) 1(g_{a,s} \neq \emptyset) \leq \frac{d_\psi}{nm^2} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_a^{-1} \sum_{t=1}^{k_a^2 - k_a} 1(g_{a,s} \neq \emptyset)$$

$$\leq \frac{d_\psi}{nm^2} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_a 1(g_{a,s} \neq \emptyset) = \frac{d_\psi}{m^2}$$

The final equality since the final double sum exactly counts the number of units in the sample (by group) $\sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_a 1(g_{a,s} \neq \emptyset) = \sum_i T_i \leq n$. An identical calculation shows that the un-squared distances

$$n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_a^{-1} \sum_{t=1}^{k_a^2 - k_a} d_{a,s,t} 1(E_{a,s,t}) \leq \frac{\sqrt{d_\psi}}{m}$$

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

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

Using this calculation, we have

\[
n^{-1} \sum_{a=1}^{L_n} \sum_{s=1}^{n} k_a^{-1} d_{a,s,t} \mathbb{I}(E_{a,s,t}^c) \leq \sum_{a=1}^{L_n} \sum_{s=1}^{n} \frac{2 \sqrt{d_\psi}}{mnk_a} \sum_{t=1}^{n} [l(i_{a,s,t}) - l(j_{a,s,t})] \mathbb{I}(E_{a,s,t}^c)
\]

\[
\leq \frac{2 \sqrt{d_\psi}}{mnk_a} \sum_{a=1}^{L_n} \sum_{t=1}^{n} m^{d_\psi} \leq \frac{2 \sqrt{d_\psi}}{mn} \leq 2 \sqrt{d_\psi} |L_n| k_{n} n^{-1} m^{d_\psi - 1}
\]

The second inequality follows by the ordering property in equation 8.1 above, since for each \( t = 1, \ldots, k_a^2 - k_a \), the intervals \( [l(i_{a,s,t}), l(j_{a,s,t})] \) are non-overlapping, and there are at most \( m^{d_\psi} \) blocks. Summarizing the above work, we have shown that

\[
n^{-1} \sum_{a=1}^{L_n} \sum_{s=1}^{n} k_a^{-1} \sum_{i,j \in g_{a,s}, i \neq j} |\psi_{i,n} - \psi_{j,n}| \leq n^{-1} \sum_{a=1}^{L_n} \sum_{s=1}^{n} k_a^{-1} \sum_{t=1}^{n} d_{a,s,t} \leq \frac{\sqrt{d_\psi}}{m} + 2 \sqrt{d_\psi} |L_n| k_{n} n^{-1} m^{d_\psi - 1}
\]

Note the following fact: for \( x, y \in [0, 1]^{d_\psi} \), \( |x - y|_2 \leq \sqrt{d_\psi} \left( \frac{|x - y|_2}{\sqrt{d_\psi}} \right)^2 \leq \sqrt{d_\psi} |x - y|_2 \), using \( \frac{|x - y|_2}{\sqrt{d_\psi}} \leq 1 \) and \( c^2 \leq c \) for \( 0 \leq c \leq 1 \). In particular, we have \( d_{a,s,t}^2 \leq \sqrt{d_\psi} \cdot d_{a,s,t} \). Then we have

\[
n^{-1} \sum_{a=1}^{L_n} \sum_{s=1}^{n} k_a^{-1} \sum_{i,j \in g_{a,s}, i \neq j} |\psi_{i,n} - \psi_{j,n}| \leq n^{-1} \sum_{a=1}^{L_n} \sum_{s=1}^{n} k_a^{-1} \sum_{t=1}^{n} d_{a,s,t} \mathbb{I}(E_{a,s,t}) + \sqrt{d_\psi} d_{a,s,t} \mathbb{I}(E_{a,s,t}^c)
\]

\[
\leq \frac{d_\psi}{m^2} + 2d_\psi |L_n| k_{n} n^{-1} m^{d_\psi - 1}
\]

Let \( (g_{a,s})_{a,s} \) be constructed as above, with \( m \asymp \left(n/(|L_n| \bar{k}_n)\right)^{1/d_\psi} \). By the first “unsquared” bound

\[
n^{-1} \sum_{a=1}^{L_n} \sum_{s=1}^{n} k_a^{-1} \sum_{i,j \in g_{a,s}, i \neq j} |\psi_{i,n} - \psi_{j,n}| = O \left( \left(n/(|L_n| \bar{k}_n)\right)^{-1/d_\psi} \right)
\]

Let \( (g_{a,s})_{a,s} \) be constructed as in the algorithm above, with \( m \asymp \left(n/(|L_n| \bar{k}_n)\right)^{1/(d_\psi + 1)} \).
Then by the second bound

\[ n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_a^{-1} \sum_{\substack{i, j \in g_{a,s} \setminus j \neq s}} |\psi_{i,n} - \psi_{j,n}|_2^2 = O \left( \left( n/(|L_n|\bar{K}_n) \right)^{-2/(d_\psi+1)} \right) \]

By construction, \((g_{a,s})_{a,s}\) satisfies the constraints in the programs 7.2 and 7.4. Then any groups \((g_{a,s}^*)_{a,s}\) that attain the minima in 7.2 and 7.4 achieve the rates above by optimality. The conclusion the follows from the Lipschitz condition. For instance, for the slower rate, let \((g_{a,s}^*)_{a,s}\) optimal for program 7.4, then

\[ n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_a^{-1} \sum_{\substack{i, j \in g_{a,s}^* \setminus j \neq s}} (h_n(\psi(X_i, \zeta_n), p_n(X_i, \zeta_n), \zeta_n) - h_n(\psi(X_j, \zeta_n), p_n(X_j, \zeta_n), \zeta_n))^2 \]

\[ = n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_a^{-1} \sum_{\substack{i, j \in g_{a,s}^* \setminus j \neq s}} (h_n(\psi(X_i, \zeta_n), p_a, \zeta_n) - h_n(\psi(X_j, \zeta_n), p_a, \zeta_n))^2 \]

\[ \leq C_n(\zeta_n) n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_a^{-1} \sum_{\substack{i, j \in g_{a,s}^* \setminus j \neq s}} |\psi_{i,n} - \psi_{j,n}|_2^2 \leq \sqrt{d_\psi} C n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_a^{-1} \sum_{\substack{i, j \in g_{a,s}^* \setminus j \neq s}} |\psi_{i,n} - \psi_{j,n}|_2 \]

\[ \leq \sqrt{d_\psi} C n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} k_a^{-1} \sum_{\substack{i, j \in g_{a,s}^* \setminus j \neq s}} |\psi_{i,n} - \psi_{j,n}|_2 = O \left( \left( n/(|L_n|\bar{K}_n) \right)^{-1/d_\psi} \right) \]

The first equality is by constraint satisfaction. The first inequality by the Lipschitz assumption. The second inequality uses the norm comparison fact above. The final inequality is by constrained optimality of \((g_{a,s}^*)_{a,s}\), and since \((g_{a,s})_{a,s}\) satisfy the propensity constraints by construction. The fast rate case is similar, just omitting the norm comparison step.

\[ \square \]

**Proof of Proposition 7.8.** We suppress the \(n\) notation, letting \(\psi_i = \psi_{i,n}\). First, note that under our matching construction, for \(i \in g_{a,s}\) by comparing with group representatives and an application of Jensen we have

\[ (\psi_i - \psi_{\gamma(i)})^2 = (\psi_i - \psi_{a,s} + \psi_{a,s} - \psi_{a,\nu_0(s)} + \psi_{a,\nu_0(s)} - \psi_{\gamma(i)})^2 \leq 3(\psi_i - \psi_{a,s})^2 + 3(\psi_{a,s} - \psi_{a,\nu_0(s)})^2 + 3(\psi_{a,\nu_0(s)} - \psi_{\gamma(i)})^2 \]

Consider the first term. We have \(\psi_i - \psi_{a,s} = \psi_i - k_a^{-1} \sum_{j \in g_{a,s}} \psi_j = k_a^{-1} \sum_{j \in g_{a,s}, j \neq i} (\psi_i - \psi_j)\).

\[ (\psi_i - \psi_{a,s})^2 \leq \left( \frac{k_a - 1}{k_a} \right)^2 (k_a - 1)^{-1} \sum_{j \in g_{a,s}, j \neq i} (\psi_i - \psi_j)^2 \leq k_a^{-1} \sum_{j \in g_{a,s}, j \neq i} (\psi_i - \psi_j)^2 \]

by Jensen again. Since by construction \(\gamma(i) \in g_{a,\nu_0(s)}\), the same reasoning applies to the
third term above, giving
\[
n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{i \in g_{a,s}} (\psi_i - \psi_{a,s})^2 + (\psi_{a,\nu_a(s)} - \psi_{\gamma(i)})^2 \leq n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{i,j \in g_{a,s}, i \neq j} (\psi_i - \psi_j)^2 = O(r_n^\psi)
\]

Next, consider the middle term \(\sum_{s=1}^{n} \sum_{i \in g_{a,s}} (\psi_{a,s} - \psi_{a,\nu_a(s)})^2 = k_a \sum_{s=1}^{n} (\psi_{a,s} - \psi_{a,\nu_a(s)})^2\).

Let \(n_a \equiv \sum_{s=1}^{n} 1(g_{a,s} \neq \emptyset)\). For \(m_a \geq 1\), the same ordered block construction as in Proposition 7.6 can be used to construct a matching \(\nu_a\) with

\[
k_a n_a \sum_{s=1}^{n} (\psi_{a,s} - \psi_{a,\nu_a(s)})^2 \asymp \frac{k_a n_a}{m_a^2} + k_a m_a^{d_\psi - 1} \asymp k_a n_a^{\frac{d_\psi - 1}{2}}
\]

letting \(m_a \asymp n_a^{1/(d_\psi + 1)}\). Accumulating over levels using \(\sum_{a=1}^{\lfloor L_n \rfloor} n_a k_a = n\) and Jensen gives

\[
n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{i \in g_{a,s}} (\psi_{a,s} - \psi_{a,\nu_a(s)})^2 \asymp n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} k_a n_a^{\frac{d_\psi - 1}{2}} = n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} k_a^{\frac{2}{\psi + 1}} (k_a n_a)^{\frac{d_\psi - 1}{2}}
\]

\[
\leq \left(\frac{2}{k_n^{\frac{2}{\psi + 1}}} n^{-1} \lfloor L_n \rfloor \right) \cdot \lfloor L_n \rfloor^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} (k_a n_a)^{\frac{d_\psi - 1}{2}} \leq \frac{2}{k_n^{\frac{2}{\psi + 1}}} n^{-1} \lfloor L_n \rfloor \left( \lfloor L_n \rfloor^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} k_a n_a \right)^{\frac{d_\psi - 1}{2}}
\]

\[
\leq \frac{2}{k_n^{\frac{2}{\psi + 1}}} n^{-1} \lfloor L_n \rfloor (\lfloor L_n \rfloor^{-1} n)^{\frac{d_\psi - 1}{2}} = \left( \lfloor L_n \rfloor^{-1} n \right)^{\frac{d_\psi - 2}{\psi + 1}}
\]

Then the matching solving the program in Proposition 7.8 achieves this rate by optimality. This finishes the proof. \(\square\)
8.2 Proofs - Asymptotics

8.2.1 Assumptions

**Assumption 8.1** (Propensity Approximation). There exist deterministic functions \( p_n, \hat{p}_n, p, \) and \( \psi \) with dependence

\[
p_{i,n} = p_n(X_i, \zeta_n) \quad \hat{p}_{i,n} = \hat{p}_n(X_i, \zeta_n) \quad p_i = p(X_i, \zeta_n) \quad \psi_i = \psi(X_i, \zeta_n)
\]  

(8.2)

These satisfy

(a) \( p_{i,n}, p_i, \hat{p}_{i,n} \in (\delta, 1 - \delta) \zeta_n\)-a.s.
(b) \( \|p_n(\cdot, \zeta_n) - \hat{p}_n(\cdot, \zeta_n)\|_\infty \leq r_n = o(1), \zeta_n\)-a.s.
(c) \( E[(\hat{p}_{i,n} - p_i)^2|\zeta_n]^{1/2} = O_p(r_n^2) \)
(d) \( \text{Im}(p_n) \subseteq L_n \) for \( n \geq 1 \)

**Assumption 8.2** (CLT). With variables defined as in assumption 8.1, for \( t \in [T] \), consider \( (f_{i,n})_{n \geq 1} \) deterministic functions. Define

\[
f_{i,t,n} = f_{t,n}(X_i, \zeta_n) \quad u_{i,t,n} = f_{i,n}(X_i, \zeta_n) - E[f_{i,n}(X_i, \zeta_n)|\psi(X_i, \zeta_n), p_n(X_i, \zeta_n), \zeta_n]
\]

Let \( \zeta_n \perp \perp (W_{i:n}, \pi_n, \tau) \) and suppose that

(a) \( E[f_{i,n}(X, \zeta_n)^2|\zeta_n] = O_p(1) \) and \( E[u_{i,t,n}^4|\zeta_n] = O_p(1) \) and \( E[\sigma_4^2(X)^2] < \infty \)
(b) \( \exists u_{i,t} \in L_2(X, \zeta_n) \) with \( E[(u_{i,t,n} - u_{i,t})^2|\zeta_n] = o_p(1) \) and \( E[u_{i,t,n}^2u_{i,t'}^2|\zeta_n] = O_p(1) \) \( \forall t, t' \in [T] \)
(c) \( E[(p_i - p_i^2)u_{i,t,n}u_{i,t'}|\zeta_n] \xrightarrow{d} \Sigma_{tt'} \) for \( t, t' \in [T] \)
(d) There exist \( c > 0 \) and polynomial \( h(\cdot) \) such that \( \zeta_n\)-a.s.

\[
E[u_{i,t,n}^2 \mathbb{1}(u_{i,t,n}^2 > z)|\zeta_n] \leq h(z) \exp(-cz)
\]

for all \( n \geq 1, z \geq 0 \)
(e) \( \bar{F}_n \lor |L_n| = o\left(\left(\log\frac{n}{n}\right)^{1/2}\right) \)

**Assumption 8.3** (Selection). With variables defined as in assumption 8.1, consider \( T_{1:n}, D_{1:n} \in \{0, 1\} \). Let \( D_{1:n} \sim \text{Loc}_n(\psi, p) \) with local group structure \( (g_{a,s})_{a,s} \) and suppose

(a) \( T_{1:n} \sim \text{Loc}_n(\psi', q) \) and \( q = a'/k' \in (\delta, 1 - \delta), \gcd(a', k') = 1 \)
(b) \( \psi_{1:n} \in \sigma(\psi_{1:n}), \) so that \( T_{1:n} \in \sigma(\psi_{1:n}, \pi^t_n, \tau) \)
(c) \( \{i : T_i = 1\} = \bigcup_{a,s} g_{a,s} \text{ with } (g_{a,s})_{a,s} \in \sigma(T_{1:n}, \psi_{1:n}, p_{1:n,n}, \pi^d_{n}) \subseteq \sigma(\psi_{1:n}, p_{1:n,n}, \pi^d_{n}, \pi^t_{n}, \tau) \)
### 8.2.2 Coupling and General CLT

**Lemma 8.4 (Coupling).** Suppose assumption 8.1 holds. Consider a function \( f_n(X_i, \zeta_n) \) with
\[
  h_n(\psi, p_a, \zeta) \equiv E[f_n(X_i, \zeta_n) | \psi(X_i, \zeta_n) = \psi, p_n(X_i, \zeta_n) = p_a, \zeta_n = \zeta] \\
  u_{i,n} \equiv f_n(X_i, \zeta_n) - h_n(\psi(X_i, \zeta_n), p_n(X_i, \zeta_n), \zeta_n) \equiv f_i,n - h_{i,n}
\]
Let \( h_n(\psi, p_a, \zeta) \) satisfy the smoothness conditions in Definition 8.1. Theorem 8.5 holds as in definition 8.13. Let \( T_{1:n} \sim \text{Loc}(\psi, p) \) and either \( T_{1:n} = 1 \) or \( T_{1:n} \sim \text{Loc}(\psi', q) \) and satisfies assumption 8.3. If \( E[\max_{i=1}^n h_{i,n}] = O(\log n) \)
\[
\sqrt{n} E_n[T_i(D_i - p_i,n)]f_n(X_i, \zeta_n)] = \sqrt{n} E_n[T_i(D_i - p_i,n)u_{i,n}] + O_p \left( r_n^\psi \sqrt{\frac{\mathbb{E}n[L_n]}{n}} \log n \right)^{1/2}
\]

**Proof.** Let \( \mathcal{F}_{x,n} = \sigma(X_{1:n}, \pi_n^d, \pi_n^t, \tau_n^d, \tau_n^t, \zeta_n) \). Then by assumption 8.3, \( (g_{a,s})_{a,s}, (h_{i,n})_{i=1}^n \in \mathcal{F}_{x,n} \), and \( \mathcal{F}_{x,n} \perp \tau^d \). Then by Lemma 8.20 and Proposition 7.4 or Proposition 7.6 (depending on the smoothness condition)
\[
\text{Var} \left( \sqrt{n} E_n[T_i(D_i - p_i)n]h_{i,n} \right| \mathcal{F}_{x,n}) \leq n^{-1} \sum_{a=1}^{L_n} \sum_{s=1}^n k_a^{-1} \sum_{i,j \in g_{a,s} \atop i \neq j} (h_{i,n} - h_{j,n})^2 + n^{-1} \mathbb{E}n[L_n] \max_{i=1}^n h_{i,n}
\]
\[
\leq O_p(r_n^\psi) + n^{-1} \mathbb{E}n[L_n] \max_{i=1}^n h_{i,n} = O_p \left( r_n^\psi \sqrt{\frac{\mathbb{E}n[L_n]}{n}} \log n \right)^{1/2}
\]
The final equality is by Markov. Also we have \( E[\sqrt{n} E_n[T_i(D_i - p_i,n)h_{i,n}] \mathcal{F}_{x,n}] = 0 \) by Lemma 8.20, so the claim follows from conditional Chebyshev (Lemma 8.16).

**Theorem 8.5 (CLT).** Impose assumptions 8.1 and 8.2. Require that either assumption 8.3 holds, or \( T_{1:n} = 1 \) and \( q = 1 \). Let \( \mathcal{F}_{x,n} = \sigma(X_{1:n}, \pi_n^d, \pi_n^t, \tau_n^d, \tau_n^t, \zeta_n) \). Then weak convergence holds as in definition 8.13
\[
\sqrt{n} \left( E_n[T_i(D_i - p_i,n)u_{i,t,n}] \right)_{t=1}^T \ | \mathcal{F}_{x,n} \Rightarrow \mathcal{N}(0, \Sigma)
\]
with variance matrix
\[
\Sigma_{t,t'} = q E[(p_i - p_i^2)u_{i,t}u_{i,t'} \zeta_n] + o_p(1) \quad t, t' \in [T]
\]
Let \( \mathcal{F}_{x,d,n} = \sigma(X_{1:n}, \pi_n^d, \pi_n^t, \tau_n^d, \tau_n^t, \zeta_n) \). We have
\[
\sqrt{n} E_n \left[ \frac{T_i D_i \epsilon_i^1}{p_i,n} + \frac{T_i(1 - D_i)\epsilon_i^0}{1 - p_i,n} \right] \ | \mathcal{F}_{x,d,n} \Rightarrow \mathcal{N} \left( 0, \text{plim}_{n \to \infty} q E \left[ \frac{\sigma_1^2(X)}{p(X, \zeta_n)} + \frac{\sigma_0^2(X)}{1 - p(X, \zeta_n)} \right] \zeta_n \right)
\]

**Proof of Theorem 8.5.** The proof is an application of the martingale CLT of Proposition 8.14. Consider the following filtration structure
\[
\mathcal{F}_{0,n} = \{ \Omega_n, \emptyset \}
\]
\[
\mathcal{F}_{(a-1)n+s,n} = \sigma(X_{1:n}, \pi_n^d, \pi_n^t, \tau_n^d, \tau_n^t, \zeta_n, \tau_{a,1:s} : 1 \leq d' \leq a - 1, \tau_{0,1:s}) \quad 1 \leq s \leq n; \quad 1 \leq a \leq |L_n|
\]
\[
\mathcal{F}_{L_n|n+j,n} = \sigma(X_{1:n}, \pi_n^d, \pi_n^t, \tau_n^d, \tau_n^t, \epsilon_{1,j} : d = 0, 1) \quad 1 \leq j \leq n
\]

54
We will apply the Cramer-Wold device to show the first claim. To that end, let \((\lambda_t)_{t=1}^T \in \mathbb{R}^T\) and define weighted residuals \(\pi_{i,n} = \sum_{t=1}^T \lambda_t u_{i,t}n\). Define the martingale difference increments

\[
Z_{(a−1)n+s,n} = \sum_{i \in g_{a,s}} (D_i - p_a)\pi_{i,n} \quad 1 \leq s \leq n; \quad 1 \leq a \leq |L_n|
\]

\[
Z_{|L_n|n+i,n} = \frac{T_iD_i\epsilon^0_i}{p_i,n} - \frac{T_i(1 - D_i)\epsilon^0_i}{1 - p_i,n} \quad 1 \leq i \leq n
\]

We focus on the conditional statements, from which marginal convergence will follow. To apply Proposition 8.14, we need to check (1) martingale difference condition (2) variance increments

**Martingale Condition** - First, we claim that \((Z_{n,L_n})_s=1^{n|L_n|+n}\) defines an MDS. In what follows, we will make frequent use of the crucial fact that

\[(A, B) \perp \perp C \implies A \perp \perp C|B \quad (8.3)\]

(1) Let \(1 \leq s \leq n\) and \(1 \leq a \leq |L_n|\). By the representation in Lemma 8.19.(i)

\[
D_i\mathbb{1}(i \in g_{a,s}) = \sum_{\ell=1}^{k_a} \mathbb{1}(i = g_{a,s,\ell})\tau^d_{a,s,\ell} = G_i,n(g_{a,s},\tau^d_{a,s}) = G_i,n(\psi_{1:n},p_{1:n,n},\pi^d_n,\pi^t_n,\tau^t,\tau^d_{a,s})
\]

Then using the fact above

\[
E[D_i\mathbb{1}(i \in g_{a,s})|\mathcal{F}_{(a−1)n+(s−1),n}] = E[G_i,n(\psi_{1:n},p_{1:n,n},\pi^d_n,\pi^t_n,\tau^t,\tau^d_{a,s})|X_{1:n},\pi^d_n,\pi^t_n,\tau^t,\tau^d_{a,s},\pi^t_{a',1:n},1 \leq a' \leq a-1, \tau_{a,1:(s-1)}]
\]

\[
= E[G_i,n(\psi_{1:n},p_{1:n,n},\pi^d_n,\pi^t_n,\tau^t,\tau^d_{a,s})|X_{1:n},\pi^d_n,\pi^t_n,\tau^t,\pi^t_{a,s},\tau^t_{a,s},\pi^t_{a',1:n},1 \leq a' \leq a-1, \tau_{a,1:(s-1)}]
\]

\[
= E[D_i\mathbb{1}(i \in g_{a,s})|X_{1:n},\pi^d_n,\pi^t_n,\tau^t,\pi^t_{a,s},\tau^t_{a,s},\pi^t_{a',1:n},1 \leq a' \leq a-1, \tau_{a,1:(s-1)}] = E[D_i\mathbb{1}(i \in g_{a,s})|\mathcal{F}_{x,n}] = p_a \mathbb{1}(i \in g_{a,s})
\]

The final equality follows by Lemma 8.19.(ii) since \(\mathcal{F}_{x,n} \perp \perp \tau^d_{a,s}\) and \((g_{a,s})_{a,s} \in \mathcal{F}_{x,n}\) by assumptions 8.3 and 8.1. Using \(\pi_{i,n} \in \sigma(X_{1:n},\zeta_n)\) and this calculation, \(E[Z_{(a−1)n+(s−1),n}|\mathcal{F}_{(a−1)n+(s−1),n}]\) is equal to

\[
E \left[ \sum_{i=1}^n (D_i - p_a)\mathbb{1}(i \in g_{a,s})\pi_{i,n}|X_{1:n},\pi^d_n,\pi^t_n,\tau^t,\tau^d_{a,s},\tau_{a',1:n},1 \leq a' \leq a-1, \tau_{a,1:(s-1)} \right] = 0
\]

(2) By assumption, \(p_{i,n} \in \sigma(X_{1:n},\zeta_n) \subseteq \mathcal{F}_{n|L_n|+j,n}\) for \(j \geq 1\). By the representation in the previous section, we see that \(D_j \in \mathcal{F}_{n|L_n|+j,n}\) for all \(1 \leq j \leq n\). Let \(\lambda: \mathbb{R}^2 \rightarrow \mathbb{R}\) be a fixed function and note that

\[
E[\lambda(e^1_j, e^0_j)|\mathcal{F}_{n|L_n|+(j−1),n}] = E[\lambda(e^1_j, e^0_j)|X_{1:n},\pi^d_n,\pi^t_n,\tau^t,\zeta_n,\tau^d,\pi_{a,n},\tau^t_{a,n}]
\]

\[
= E[\lambda(e^1_j, e^0_j)|X_{1:n},\pi^d_{0,1:n},\tau^t_{0,1:n},\tau^d_{0,1:n}] = E[\lambda(e^1_j, e^0_j)|X_{j}]
\]

This follows by making repeated use of 8.3, independent sampling, and that the external variables \((\pi^d_n,\pi^t_n,\tau^t,\zeta_n,\tau^d) \perp \perp (X_{1:n},Y_{1:n}(1),Y_{1:n}(0))\). In particular, applying this shows that \(E[\epsilon^d_j|\mathcal{F}_{n|L_n|+(j−1),n}] = E[\epsilon^d_j|X_{j}] = 0\) for \(d = 0, 1\). Then we see that \(E[Z_{n|L_n|+j,n}|\mathcal{F}_{n|L_n|+(j−1),n}] = \)
0. This shows that \((Z_j, n^{(l_n)^{j=1}})\) is an MDS. Next, we characterize the variance process.

**Variance Process** - We will characterize the probability limit of each variance process, and show the relevant measurability condition.

(1) For \(1 \leq s \leq n\) and \(1 \leq a \leq |L_n|\), we have \(\mathbf{1}(i \in g_{as}), u_{i,n} \in F_{x,n} \subseteq F_{(a-1)n+(s-1),n}\). Similarly to the argument in (1) above, we have \(E[Z^2_{(a-1)n+s,n}|F_{(a-1)n+(s-1),n}] = E[Z^2_{(a-1)n+s,n}|F_{x,n}]\). Since \((g_{as})_{a,s} \in F_{x,n}\) and \(F_{x,n} \perp \tau^d\), Lemma 8.19.(iii) applies, giving covariances

\[
n^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} E[Z^2_{(a-1)n+s,n}|F_{(a-1)n+(s-1),n}] = n^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \frac{q_a(k_a - q_a)}{k_a^2} \sum_{i \in g_{as}} \pi_{i,n}^2
\]

\[
- n^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \frac{q_a(k_a - q_a)}{k_a^2(k_a - 1)} \sum_{i,j \in g_{as}, i \neq j} \tau_{i,n} \tau_{j,n} = A_n + B_n
\]

Clearly, these are both \(F_{x,n}\)-measurable, so we have verified the required measurability condition. We first consider the diagonal terms \(A_n\). Note that by algebra

\[
p_i - p_i^2 = \left(1 - (p_i + p_i)|1 - (p_i + p_i)|\right] = \left(p_i - \hat{p}_i, n + \hat{p}_i, n - p_i\right)[1 - (p_i + p_i)]
\]

Then we have

\[
A_n = n^{-1} \sum_{a=1}^{[L_n]} \sum_{i=1}^{n} p_{i,n}(1 - p_{i,n})^2 \pi_{i,n}^2 = E_n[T_i p_{i,n}(1 - p_{i,n}) \pi_{i,n}^2]
\]

\[
= E_n[T_i p_{i,n}(1 - p_{i,n}) \pi_{i,n}^2] + E_n[T_i (p_{i,n} - \hat{p}_{i,n} + \hat{p}_{i,n} - p_i)[1 - (p_i + p_i)] \pi_{i,n}^2]
\]

Consider the second term. First note that

\[
E[|E_n[T_i (p_{i,n} - \hat{p}_{i,n})[1 - (p_i + p_i)] \pi_{i,n}^2]|\zeta_n] \leq E[|\hat{p}_{i,n} - p_{i,n}| \pi_{i,n}^2|\zeta_n] \leq r_n p \cdot T \sum_{t \in [T]} \lambda_t^2 E[u_{i,t,n}] \leq r_n p \cdot T \sum_{t \in [T]} \lambda_t^2 E[f_{i,n}(X_i, \zeta_n)2] \zeta_n = O(r_n^p)
\]

The first inequality since \(T_i[1 - (p_i + p_i)] \leq 1\) and since \((\pi_{i,n})_{n=1}^{n}\) are identically distributed, conditional on \(\zeta_n\). The second inequality is by Jensen, and the final inequality by properties of projection. Similarly, we have

\[
E[|E_n[T_i (p_{i,n} - p_i)[1 - (p_i + p_i)] \pi_{i,n}^2]|\zeta_n] \leq E[|\hat{p}_{i,n} - p_{i,n}| \pi_{i,n}^2|\zeta_n] \leq E[(\hat{p}_{i,n} - p_i)^2|\zeta_n]^{1/2} E[\pi_{i,n}^2|\zeta_n]^{1/2} = o_p(r_n^p) O_p(1)
\]

Summarizing, we have shown that \(A_n = E_n[T_i p_{i,n} - p_{i,n}) \pi_{i,n}^2] + O_p(r_n^p) + O_p(r_n^p)\).

Define the population residual sum \(\pi_i = \sum_{t=1}^T \lambda_t u_{i,t}\). Claim that

\[
E_n[T_i p_{i,n} - p_{i,n}) (\pi_{i,n}^2 - \pi_{i}^2)] = O_p \left(\max_{t \in [T]} E[(u_{i,t,n} - u_{i,t})^2|\zeta_n]^{1/2} \right)
\]

56
Note that we have
\[
E\left[\left|E_n[T_ip_i(1 - p_i)(\pi^2_{i,n} - \pi^2_i)]\right|\right] \leq E\left[|p_i(1 - p_i)(\pi^2_{i,n} - \pi^2_i)|\right] \leq (1/4)E\left[|\pi^2_{i,n} - \pi^2_i|\right],
\]
\[
\leq E[(\pi_{i,n} - \pi_i)^2|\zeta_n|^{1/2}E[(\pi_{i,n} + \pi_i)^2|\zeta_n|^{1/2} \leq E[(\pi_{i,n} - \pi_i)^2]^{1/2}\left(E[\pi^2_{i,n}|\zeta_n|^{1/2} + E[\pi^2_i|\zeta_n|^{1/2}\right)]
\]
\[
\leq E[(\pi_{i,n} - \pi_i)^2|\zeta_n|^{1/2}O_p(1) \leq \sum_{i=1}^{T} |\lambda_i|E[((u_{i,t,n} - u_{i,t})^2)|\zeta_n|^{1/2}
\]
The third inequality uses \(a^2 - b^2 = (a - b)(a + b)\) and Cauchy-Schwarz. For the fifth inequality, observe that
\[
E[\pi^2_{i,n}|\zeta_n|^{1/2} \leq \sum_{i=1}^{T} |\lambda_i|E[u_{i,t}|\zeta_n|^{1/2} = O_p(1)
\]
\[
E[\pi^2_i|\zeta_n|^{1/2} \leq \sum_{i=1}^{T} |\lambda_i|E[u_{i,t}|\zeta_n|^{1/2} = O_p(1)
\]
The second inequality follows by properties of projection. The third inequality is triangle inequality. The claim then follows from conditional Markov inequality (Lemma 8.16).

Above work has shown that
\[
A_n = E_n[T_ip_i(1 - p_i)\pi^2_i] + O_p(r^p_n) + O_p(r^p_n) + O_p\left(\max_{t \in [T]} E[((u_{i,t,n} - u_{i,t})^2)|\zeta_n|^{1/2}\right)
\]
\[
= \sum_{i,t'} \lambda_i \lambda_{i'} E_n[T_ip_i(1 - p_i)u_{i,t}u_{i,t'}] + o_p(1)
\]
\[
= \sum_{i,t'} \lambda_i \lambda_{i'} E_n[(T_i - q)p_i(1 - p_i)u_{i,t}u_{i,t'}] + q \sum_{i,t'} \lambda_i \lambda_{i'} E_n[p_i(1 - p_i)u_{i,t}u_{i,t'}] + o_p(1)
\]
If \(T_{1:n} = 1\) and \(q = 1\), the first term is identically 0. Otherwise, by assumption 8.3 we have \(T_{1:n} \sim \text{Locn}(\psi', q)\). Then Lemma 8.20 shows that \(E_n[(T_i - q)p_i(1 - p_i)u_{i,t}u_{i,t'}] = O_p(n^{-1/2})\) for each \(t \in [T]\) if \(E[u^2_{i,t}u^2_{i,t'}|\zeta_n] < \infty\), which we assume.

For the second term, note that \(E[\pi_n(p_i(1 - p_i)u_{i,t}u_{i,t'})] = E[p_i(1 - p_i)u_{i,t}u_{i,t'}|\zeta_n]\) and
\[
\text{Var}(E_n[p_i(1 - p_i)u_{i,t}u_{i,t'}]|\zeta_n) = n^{-2} \sum_{i \neq j} \text{Cov} \left(p_i(1 - p_i)u_{i,t}u_{i,t'}, p_j(1 - p_j)u_{j,t}u_{j,t'}|\zeta_n\right)
\]
\[
+ n^{-1} E_n[\text{Var}(p_i(1 - p_i)u_{i,t}u_{i,t'})|\zeta_n] \leq (4n)^{-1} E[u^2_{i,t}u^2_{i,t'}|\zeta_n] = O_p(n^{-1})
\]
For the inequality, note that \(p_i(1 - p_i)u_{i,t}u_{i,t'} \perp p_j(1 - p_j)u_{j,t}u_{j,t'}|\zeta_n\) for \(i \neq j\) by our assumptions, so the off-diagonal terms vanish. Then, we bound the variance by a second moment. By conditional Chebyshev in Lemma 8.16 and summarizing the above work, we have shown that
\[
A_n = q \sum_{i,t'} \lambda_i \lambda_{i'} E[(p - p^2)(X_i)u_{i,t}u_{i,t'}|\zeta_n] + O_p(n^{-1/2}) + o_p(1)
\]
Next, consider the off-diagonal terms \(B_n\). Let \(\mathcal{F}_{(\psi, p_n), n} = \sigma(\psi_{1:n}, p_{1:n:n}, \pi^n_d, \pi^n_t, \tau^t, \zeta_n)\).
Note that \((g_{a,s})_{a,s} \in \mathcal{F}(\psi,p_n),n\) by assumption 8.3. Applying Lemma 8.22 on residual selection with functions \(f_{i,n}(X_i, \zeta_n) = \sum_{t \in [T]} \lambda_t f_{i,n}(X_i, \zeta_n)\) and random element \(\xi_n = (\pi_n, \pi_n', s')\) gives the following moments for \(i, j, l, f \in [n]\) distinct indices

\[
E[\overline{u}_{i,n} \overline{u}_{j,n} | \mathcal{F}(\psi,p_n),n] = 0 \quad E[\overline{u}_{i,n} \overline{a}_{j,n} \overline{u}_{l,n} \overline{a}_{f,n} | \mathcal{F}(\psi,p_n),n] = 0 \quad E[\overline{u}_{i,n}^2 \overline{u}_{j,n} \overline{a}_{f,n} | \mathcal{F}(\psi,p_n),n] = 0
\]

We apply these facts, computing

\[
E[B_n | \mathcal{F}(\psi,p_n),n] = E \left[ n^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n-1} q_a(k_a-q_a) \sum_{i \neq j}^n 1(i, j \in g_{a,s}) \overline{u}_{i,n} \overline{u}_{j,n} | \mathcal{F}(\psi,p_n),n \right] = 0
\]

Define \(w_a \equiv \frac{q_a(k_a-q_a)}{k_a^2(k_a-1)}\). Then the conditional variance \(E[B_n^2 | \mathcal{F}(\psi,p_n),n]\) is given by

\[
n^{-2} \sum_{a, a'=1}^{[L_n]} \sum_{s, s'=1}^{n-1} w_a w_{a'} \sum_{i \neq j}^n 1(i, j \in g_{a,s}) 1(l, f \in g_{a',s'}) E[\overline{u}_{i,n} \overline{u}_{j,n} \overline{u}_{l,n} \overline{a}_{f,n} | \mathcal{F}(\psi,p_n),n]
\]

Case 1: \(\{i, j\} \cup \{l, f\} = \emptyset\) then \(E[\overline{u}_{i,n} \overline{u}_{j,n} \overline{a}_{f,n} | \mathcal{F}(\psi,p_n),n] = 0\). By uniqueness of group identity, we are in this case if \((a, s) \neq (a', s')\).

Case 2: \(i \in \{l, f\}\) or \(j \in \{l, f\}\), exclusive. By symmetry, without loss assume \(i = l\), then the inner conditional expectation is \(E[\overline{u}_{i,n} \overline{u}_{j,n} \overline{u}_{l,n} | \mathcal{F}(\psi,p_n),n] = 0\).

Case 3: \(\{i, j\} = \{l, f\}\). By symmetry, the terms associated with \((i, j) = (l, f)\) and \((i, j) = (f, l)\) are the same. Summarizing these observations, the remaining terms are

\[
E[B_n^2 | \mathcal{F}(\psi,p_n),n] = 4n^{-2} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n-1} w_a^2 \sum_{i, j \in g_{a,s}}^n E[\overline{u}_{i,n}^2 \overline{u}_{j,n}^2 | \mathcal{F}(\psi,p_n),n]
\]

\[
\leq 4n^{-2} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} w_a^2 \sum_{i, j \in g_{a,s}}^n E[\overline{u}_{i,n}^2 | \psi_1, p_{i,n}, \zeta_n] E[\overline{u}_{j,n}^2 | \psi_1, p_{j,n}, \zeta_n]
\]

\[
\leq 2n^{-2} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} w_a^2 \left( \sum_{i, j \in g_{a,s}}^n E[\overline{u}_{i,n}^2 | \psi_1, p_{i,n}, \zeta_n] \right)^2 \leq 2n^{-2} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} w_a^2 k_a \sum_{i, j \in g_{a,s}}^n E[\overline{u}_{i,n}^2 | \psi_1, p_{i,n}, \zeta_n]^2
\]

\[
\leq n^{-2} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \sum_{i \in g_{a,s}}^n E[\overline{u}_{i,n}^2 | \psi_1, p_{i,n}, \zeta_n]^2 = n^{-1} E_n[T_l E[\overline{u}_{i,n}^2 | \psi_1, p_{i,n}, \zeta_n]^2] = O_p(n^{-1})
\]

The first inequality follows by adding in the \(s = n\) term (positive) and the moments above. The second inequality follows by adding and subtracting the diagonal terms. The third inequality is by Jensen. The fourth inequality follows since \(k_a w_a^2 = \frac{k_a}{k_a - 1} p_a (1 - p_a) \leq\)
\[ 2 \cdot (1/4) \text{ for } k_a \geq 2. \text{ For the final equality, observe that} \]
\[ E[n^{-1} E_n [T_i E[\pi_{i,n}^2 | \psi, p_{i,n}, \zeta_n]^2 | \zeta_n]] \leq E[E[\pi_{i,n}^2 | \psi, p_{i,n}, \zeta_n]^2 | \zeta_n] \leq E[\pi_{i,n}^2 | \zeta_n] \leq T^2 \sum_{t \in [T]} E[u_{t,i,n}^4 | \zeta_n] = O_p(1) \]

The first inequality is by conditional Jensen and tower law. The second inequality again by Jensen. Putting this together, we have
\[ A_n + B_n = q \sum_{t,t'=1}^T \lambda_t \lambda_{t'} E[(p - p^2)(X_i)u_{i,t}u_{i,t'} | \zeta_n] + o_p(1) \]

(2) Let \( 1 \leq i \leq n \) and note that since \( D_i(1 - D_i) = 0 \) we have
\[ E[Z_{i,n+i,n}^2 | F_{i,n} | n+(i-1),n] = E \left[ \frac{T_i D_i (\epsilon_i^1)^2}{p_{i,n}^2} + \frac{T_i (1 - D_i)(\epsilon_i^0)^2}{(1 - p_{i,n})^2} \right] \]
\[ = \frac{T_i D_i E[(\epsilon_i^1)^2 | X_i]}{p_{i,n}} + \frac{T_i (1 - D_i) E[(\epsilon_i^0)^2 | X_i]}{(1 - p_{i,n})^2} = \frac{T_i D_i \sigma_i^2(X_i)}{p_{i,n}^2} + \frac{T_i (1 - D_i) \sigma_i^2(X_i)}{(1 - p_{i,n})^2} \]

The second equality uses the fact from our discussion when proving the MDS property (2) above with \( h(x, y) = x^2 \) and \( h(x, y) = y^2 \). Then the sum of conditional variances is
\[ n^{-1} \sum_{i=1}^n E[Z_{i,n+i,n}^2 | F_{i,n} | n+(i-1),n] = n^{-1} \sum_{i=1}^n \left[ \frac{T_i D_i \sigma_i^2(X_i)}{p_{i,n}^2} + \frac{T_i (1 - D_i) \sigma_i^2(X_i)}{(1 - p_{i,n})^2} \right] \]
\[ \equiv E_n \left[ T_i D_i \left( \frac{\sigma_i^2(X_i)}{p_{i,n}^2} - \frac{\sigma_i^2(X_i)}{(1 - p_{i,n})^2} \right) \right] + E_n \left[ T_i \sigma_i^2(X_i) \right] \]
\[ = E_n \left[ T_i(D_i - p_{i,n}) \left( \frac{\sigma_i^2(X_i)}{p_{i,n}^2} - \frac{\sigma_i^2(X_i)}{(1 - p_{i,n})^2} \right) \right] + E_n \left[ T_i \left( \frac{\sigma_i^2(X_i)}{p_{i,n}} + \frac{\sigma_i^2(X_i)}{(1 - p_{i,n})} \right) \right] \]
\[ = E_n \left[ T_i(D_i - p_{i,n}) \left( \frac{\sigma_i^2(X_i)}{p_{i,n}^2} - \frac{\sigma_i^2(X_i)}{(1 - p_{i,n})^2} \right) \right] + E_n \left[ (T_i - q) \left( \frac{\sigma_i^2(X_i)}{p_{i,n}} + \frac{\sigma_i^2(X_i)}{(1 - p_{i,n})} \right) \right] \]
\[ + qE_n \left[ \left( \frac{\sigma_i^2(X_i)}{p_{i,n}} + \frac{\sigma_i^2(X_i)}{(1 - p_{i,n})} \right) \right] \equiv A_n + B_n + C_n \]

From the first expression, we see that this process is \( F_{x,d,n} \)-measurable. The first three equalities follow by algebra.

Consider \( A_n \). Let \( v_n(X_i) = \frac{\sigma_i^2(X_i)}{p_{i,n}^2} - \frac{\sigma_i^2(X_i)}{(1 - p_{i,n})^2} \). By Lemma 8.20 we have \( A_n = O_p(n^{-1/2}) \) if \( \sup_n E[v_n(X_i)^2] < \infty \). To see this, note that
\[ E[v_n(X_i)^2] \leq E \left[ \frac{2\sigma_i^2(X_i)^2}{p_{i,n}^2} + \frac{2\sigma_i^2(X_i)^2}{(1 - p_{i,n})^2} \right] \leq 2\delta^{-4} E[\sigma_i^2(X)^2 + \sigma_0^2(X)^2] < \infty \]

The first inequality is Young’s, then using propensity score bound and assumption. If \( T_1 = 1 \) and \( q = 1 \), then \( B_n = 0 \). Otherwise, by assumption 8.3 we have \( T_1 \sim \text{Loc}_{\psi}(\psi', q) \), then \( B_n = O_p(n^{-1/2}) \) by Lemma 8.20 and an identical argument.
Next consider \( C_n \). We can calculate

\[
\left| E_n \left[ \sigma_1^2(X_i) \left( \frac{1}{p_{i,n}} - \frac{1}{p_i} \right) \right] \right| = \left| E_n \left[ \sigma_1^2(X_i) \frac{p_i - p_{i,n}}{p_i p_{i,n}} \right] \right| \leq \delta^{-2} E_n \left[ \sigma_1^2(X_i) | p_i - \hat{p}_{i,n} + \hat{p}_{i,n} - p_{i,n} | \right] \\
\leq E_n \left[ \sigma_1^2(X_i) | p_i - \hat{p}_{i,n} | \right] + r_n^p E_n \left[ \sigma_1^2(X_i) \right] \\
= E_n \left[ \sigma_1^2(X_i) | p_i - \hat{p}_{i,n} | \right] + O_p(r_n^p)
\]

Now taking expectations, note that

\[
E[ E_n[\sigma_1^2(X_i)|p_i - \hat{p}_{i,n}|||\zeta_n] = E[\sigma_1^2(X_i)|p_i - \hat{p}_{i,n}|||\zeta_n] \leq (E[\sigma_1^2(X_i)^2])^{1/2}(E[(p_i - \hat{p}_{i,n})^2]|\zeta_n)^{1/2}
\]

Then by Lemma 8.16, we have \( E_n[\sigma_1^2(X_i)|p_i - \hat{p}_{i,n}|||\zeta_n] = O_p(\bar{r}_n^p) \). The analysis for the term involving \( \sigma_0^2(X_i) \) is identical. Then by WLLN, we have shown that

\[
C_n = q E_n \left[ \frac{\sigma_1^2(X_i)}{p_i} + \frac{\sigma_0^2(X_i)}{(1 - p_i)} \right] + O_p(r_n^p + \bar{r}_n^p)
\]

Summarizing the work above, the conditional variance process is

\[
n^{-1} \sum_{i=1}^{n} E[Z_{L_n|n+i,n}^2 | F_{L_n|n+i-1,n}] = qE \left[ \frac{\sigma_1^2(X_i)}{p_i} + \frac{\sigma_0^2(X_i)}{(1 - p_i)} \right] + O_p(n^{-1/2}) + O_p(r_n^p) + O_p(\bar{r}_n^p)
\]

**Lindberg Condition** - We show the Lindberg condition in probability, conditional on the appropriate set.

(1) For \( 1 \leq s \leq n \) and \( 1 \leq a \leq \lfloor L_n \rfloor \), \( Z_{(a-1)n+s,n} = \sum_{i \in g_{a,s}} (d_i - p_a)\bar{u}_{i,n} \). Now note that

\[
n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} E \left[ Z_{(a-1)n+s,n}^2 \mathbb{1}(|Z_{(a-1)n+s,n}| > \sqrt{n}e) \right] \left| F_{x,n} \right|
\]

\[
= n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} E \left[ \left( \sum_{i \in g_{a,s}} (d_i - p_a)\bar{u}_{i,n} \right)^2 \mathbb{1} \left( \sum_{i \in g_{a,s}} (d_i - p_a)\bar{u}_{i,n} \right)^2 > ne^2 \right] \left| F_{x,n} \right|
\]

\[
\leq n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} |g_{a,s}| E \left[ \sum_{i \in g_{a,s}} (d_i - p_a)^2\bar{u}_{i,n}^2 \mathbb{1} \left( \sum_{i \in g_{a,s}} (d_i - p_a)^2\bar{u}_{i,n}^2 > ne^2 \right) \left| F_{x,n} \right| 
\]

Continuing, this is

\[
\leq \bar{K}_n n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} E \left[ \sum_{i \in g_{a,s}} \bar{u}_{i,n}^2 \mathbb{1} \left( \sum_{i \in g_{a,s}} \bar{u}_{i,n}^2 > ne^2 \bar{K}_n^{-1} \right) \left| F_{x,n} \right| 
\]

\[
\leq T\bar{K}_n n^{-1} \sum_{a=1}^{\lfloor L_n \rfloor} \sum_{s=1}^{n} E \left[ \sum_{i \in g_{a,s}} \sum_{t \in [T]} u_{i,t,n}^2 \mathbb{1} \left( \sum_{i \in g_{a,s}} \sum_{t \in [T]} u_{i,t,n}^2 > ne^2 (\bar{K}_n T)^{-1} \right) \left| F_{x,n} \right| 
\]

60
The first and third inequalities follow because \( (\sum_{i \in I} a_i)^2 \leq |I| \sum_{i \in I} a_i^2 \) for any scalar array \((a_i)_{i \in I}\) by Jensen inequality. The second inequality uses \(|d_i - p_n| \leq \max(p_n, 1 - p_n) \leq 1\) and \(|g_{a,s}| \leq \bar{k}_n\) by definition. Now note that for any positive scalar array \((a_i)_{i \in I}\) we have
\[
\sum_{i \in I} a_i \cdot 1 \left( \sum_{i \in I} a_i > c \right) \leq |I| \sum_{i \in I} a_i \cdot 1 \left( a_i > c / |I| \right) \tag{8.4}
\]
To see this, note that in the case \(a_j = \max_{i \in I} a_i\), the LHS above is bounded by \(|I| a_j 1 (|I| a_j > c)\). Summing this bound over \(j \in I\) accounts for all the cases, giving a uniform upper bound. Using this in the final expression above yields
\[
T\bar{k}_n n^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} E \left[ \sum_{i \in g_{a,s}} \sum_{t \in [T]} u_{i,t,n}^2 I \left( \sum_{i \in g_{a,s}} \sum_{t \in [T]} u_{i,t,n}^2 > n e^2 (\bar{k}_n T)^{-1} \right) \left| F_{x,n} \right. \right] \\
\leq T\bar{k}_n n^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} E \left[ \bar{k}_n T \sum_{i \in g_{a,s}} \sum_{t \in [T]} u_{i,t,n}^2 I \left( u_{i,t,n}^2 > n e^2 (\bar{k}_n T)^{-2} \right) \left| F_{x,n} \right. \right] \\
= (T\bar{k}_n)^2 n^{-1} \sum_{t \in [T]} \sum_{a=1}^{[L_n]} E_n \left[ T_t E \left[ u_{i,t,n}^2 I \left( u_{i,t,n}^2 > n e^2 (\bar{k}_n T)^{-2} \right) \left| F_{x,n} \right. \right] \right] \\
\leq (T\bar{k}_n)^2 \sum_{t \in [T]} E_n \left[ u_{i,t,n}^2 I \left( u_{i,t,n}^2 > n e^2 (\bar{k}_n T)^{-2} \right) \left| F_{x,n} \right. \right]
\]
The third equality follows since \(g_{a,s} \in F_{x,n}\) for all \(a, s\). Taking an expectation of the final upper bound gives, using linearity and tower law gives
\[
(T\bar{k}_n)^2 \sum_{t \in [T]} E \left[ u_{i,t,n}^2 I \left( u_{i,t,n}^2 > n e^2 (\bar{k}_n T)^{-2} \right) \left| g_o \right. \right] \leq (\bar{k}_n)^2 h(n e^2 (\bar{k}_n T)^{-2}) \exp \left( -c e^2 (\bar{k}_n T)^{-2} \right) \\
\leq (\bar{k}_n)^2 \exp \left( -1/2 c e^2 (\bar{k}_n T)^{-2} \right) = \exp \left( 2 \log(\bar{k}_n) \left( -\frac{(1/2)c e^2}{2 \log(\bar{k}_n) (\bar{k}_n T)^2} + 1 \right) \right) = o(1)
\]
The first inequality is by assumption. The second inequality follows using that \(f\) is polynomial, so that \(f(s_n) \exp(-1/2 c s_n) = o(1)\) for any \(s_n \to \infty\), and \(n(\bar{k}_n)^{-2} \to \infty\) by assumption. The last line follows since \(n^{-1} \log(\bar{k}_n) (\bar{k}_n)^2 = o(1)\) by assumption. By Markov inequality and the chain of upper bounds, we have shown that
\[
n^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} E \left[ Z_{(a-1)n+s}^2 \right] (|Z_{(a-1)n+s}^2| > \sqrt{ne} |F_{x,n}|) = o_p(1)
\]
61
(2) Let \(1 \leq i \leq n\) and consider the bound

\[
Z_{n|L_{n+i},n}^2 = T_i \left( \frac{D_i \epsilon_i^1 - (1 - D_i) \epsilon_i^0}{p_{i,n} - (1 - p_{i,n})} \right)^2 = \frac{T_i D_i (\epsilon_i^1)^2}{p_{i,n}^2} + \frac{T_i (1 - D_i) (\epsilon_i^0)^2}{(1 - p_{i,n})^2}
\]

\[
\leq \max(p_{i,n}^{-1}, (1 - p_{i,n})^{-1}) \sum_{d=0,1} (\epsilon_i^d)^2 \leq \delta^2 \sum_{d=0,1} (\epsilon_i^d)^2
\]

Then we have

\[
n^{-1} \sum_{i=1}^n E \left[ Z_{|L_{n+i},n}^2 \mathbb{1}(\{|Z_{|L_{n+i},n}| > \sqrt{n} \epsilon\}|F_{(x,d),n}) \right]
\]

\[
\leq n^{-1} \sum_{i=1}^n E \left[ \delta^{-2} \sum_{d=0,1} (\epsilon_i^d)^2 \mathbb{1}(\delta^{-2} \sum_{d=0,1} (\epsilon_i^d)^2 > n \epsilon^2) \right] \left| X_{1,n}, \alpha_i^d, \beta_i^d, \gamma_i^t, \tau_i^t, \zeta_n \right|
\]

\[
= n^{-1} \sum_{i=1}^n E \left[ \delta^{-2} \sum_{d=0,1} (\epsilon_i^d)^2 \mathbb{1}(\delta^{-2} \sum_{d=0,1} (\epsilon_i^d)^2 > n \epsilon^2) \right] \left| X_{1,n} \right|
\]

\[
= n^{-1} \sum_{i=1}^n E \left[ \delta^{-2} \sum_{d=0,1} (\epsilon_i^d)^2 \mathbb{1}(\delta^{-2} \sum_{d=0,1} (\epsilon_i^d)^2 > n \epsilon^2) \right] \left| X_i \right|
\]

Consider the expectation of the final term, which is

\[
\delta^{-2} n^{-1} \sum_{i=1}^n E \left[ \sum_{d=0,1} (\epsilon_i^d)^2 \mathbb{1}(\sum_{d=0,1} (\epsilon_i^d)^2 > \delta^2 n \epsilon^2) \right] \leq \delta^{-2} E \left[ \sum_{d=0,1} (\epsilon_i^d)^2 \mathbb{1}(\sum_{d=0,1} (\epsilon_i^d)^2 > M(n)) \right]
\]

The inequality holds for any sequence \(M(n) > \delta^2 n \epsilon^2\). The RHS is \(o(1)\) as \(n \rightarrow \infty\) since

\[
E \left[ \sum_{d=0,1} (\epsilon_i^d)^2 \right] = \sum_{d=0,1} E[\sigma_d^2(X)] < \infty
\]

by law of total variance. Then \(n^{-1} \sum_{i=1}^n E \left[ Z_{|L_{n+i},n}^2 \mathbb{1}(\{|Z_{|L_{n+i},n}| > \sqrt{n} \epsilon\}|F_{(x,d),n}) \right] = o_p(1)\) by Markov inequality.

The MDS condition, variance limits and measurability, and Lindberg conditions in probability (conditional on \(F_{x,n}\) and \(F_{(x,d),n}\) respectively) are sufficient to invoke Proposition 8.14, which proves the first two claims.

\[\Box\]

**Assumption 8.6 (CLT 1).** Suppose \(p = a/k\). In this context, define

\[
F_i^d = E[m_d(X_i)|\psi_i] \quad \quad u_i^d = m_d(X_i) - F_i^d
\]

For \(d = 0, 1\), require the following conditions

(a) \(E[m_d(X)^2] < \infty\), \(E[\sigma_d^2(X)^2] < \infty\), and \(E[(u_i^d)^4] < \infty\)

(b) There exist \(c > 0\) and polynomial \(h(\cdot)\) such that

\[
E[(u_i^d)^2 \mathbb{1}((u_i^d)^2 > z)] \leq h(z) \exp(-cz)
\]

62
For all \( n \geq 1, z \geq 0 \)

(c) \( F^d(\psi) = E[m_d(X_i)|\psi = \psi] \) satisfies the smoothness conditions in either Definition 7.1 or 7.5. Also, \( E[\max_{i=1}^n |F^d_i|] = O(\log n) \).

**Assumption 8.7** (CLT II). Require Assumption 8.1 with \( \zeta_n = \zeta \) for all \( n \geq 1 \) and \( \hat{p}_{i,n} = p_i = p(X_i, \zeta) \). Assume \( \psi'_i = \psi'(X_i, \zeta) \in \sigma(\psi_i) \) for \( \psi_i = \psi(X_i, \zeta) \). In this context, define the following

\[
\begin{align*}
F^d_{i,n} &= E[m_d(X_i)|p_{i,n}, \psi_i, \zeta] \\
F^d_i &= E[m_d(X_i)|p_i, \psi_i, \zeta] \\
\tilde{F}^d_i &= E[m_d(X_i)|\psi'_i, \zeta]
\end{align*}
\]

For \( d = 0, 1 \), require the following conditions

(a) \( E[m_d(X)^2] < \infty \) and \( E[\sigma^2_d(X)^2] < \infty \)

(b) \( E[(u^d_{i,n} - u^d_i)^2|\zeta] = o_p(1) \).

(c) \( E[(u^d_{i,n})^4|\zeta] < \infty, E[(u^d_i)^4|\zeta] < \infty, \) and \( E[(\tilde{u}^d_i)^4|\zeta] < \infty, \zeta\text{-a.s.} \)

(d) There exist \( c > 0 \) and polynomial \( h(\cdot) \) such that \( \zeta\text{-a.s.} \)

\[
E[(u^d_{i,n})^2 \mathbb{1}((u^d_{i,n})^2 > z)|\zeta] \vee E[(\tilde{u}^d_i)^2 \mathbb{1}((\tilde{u}^d_i)^2 > z)|\zeta] \leq h(z) \exp(-cz)
\]

for all \( n \geq 1, z \geq 0 \)

(e) \( F^d_n(\psi, p_0, \zeta) \) and \( \tilde{F}^d_n(\psi', \zeta) \) satisfy the smoothness conditions in either Definition 7.1 or 7.5. Also, \( E[\max_{i=1}^n |F^d_{i,n}|] = O(\log n) \) and \( E[\max_{i=1}^n |\tilde{F}^d_i|] = O(\log n) \)

(f) \( \bar{k}_n \lor |L_n| = o \left( \left( \frac{\log n}{n} \right)^{1/2} \right) \)

In the following easy lemma, we justify focusing on the IPW estimator with population normalization \( q \) in this section, rather than the (random) realized propensity \( E_n[T_i]^{-1} \).

**Lemma 8.8** (Normalization). Under the assumptions of Theorem 3.17

\[
E_n \left[ \frac{T_i(D_i - p_n(X_i))Y_i}{q(p_n - p_n^2)(X_i)} \right] - E_n \left[ \frac{(D_i - p_n(X_i))Y_i}{(p_n - p_n^2)(X_i)} \right] = 1 = O_p(n^{-1})
\]

**Proof.** Suppose that \( q = a/k \). First note that by Definition 2.2, we have w.p.1

\[
\sum_{i=1}^n T_i = \sum_{s=1}^{n-1} \sum_{i \in g_s} T_i + \sum_{i \in g_n} T_i \in [\lfloor n/k \rfloor a, \lfloor n/k \rfloor a + k] \subseteq [qn - a, qn + k]
\]

Then \( E_n[T_i] \in [q - a/n, q + k/n] \). In particular, this shows that \( E_n[T_i]^{-1} = O_p(1) \) and \( E_n[T_i] - q = O_p(n^{-1}) \). By Theorem 3.17, we have

\[
E_n \left[ \frac{T_i(D_i - p_n(X_i))Y_i}{(p_n - p_n^2)(X_i)} \right] = \hat{\theta}/q = \text{ATE} / q + O_p(n^{-1/2}) = O_p(1)
\]

Then the difference above can be written

\[
\frac{1}{qE_n[T_i]} E_n \left[ \frac{T_i(D_i - p_n(X_i))Y_i}{(p_n - p_n^2)(X_i)} \right] (E_n[T_i] - q) = O_p(1)O_p(n^{-1}) = O_p(n^{-1})
\]
This finishes the claim.

Proof of Theorem 3.17. (1) Assignment Term: With $u^d_{i,n}$ and $u^d_i$ as above, in the notation of Theorem 8.5 define

$$f_{i,n}(X_i, \zeta) = \frac{b_n(X_i, \zeta)}{\sqrt{p_{i,n} - p_i^2}} = \frac{m_1(X_i)}{p_{i,n} - p_i} + \frac{m_0(X_i)}{1 - p_{i,n}}$$

$$u_{i,1,n} \equiv u_{i,n} = \frac{b_n(X_i, \zeta)}{\sqrt{p_{i,n} - p_i^2}} - E \left[ \frac{b_n(X_i, \zeta)}{\sqrt{p_{i,n} - p_i^2}} \mid \psi_i, p_{i,n}, \zeta \right] = \frac{u^1_{i,n}}{p_{i,n}} + \frac{u^0_{i,n}}{1 - p_{i,n}}$$

$$u_{i,1} \equiv u_i = \frac{b(X_i, \zeta)}{\sqrt{p_i - p_i^2}} - E \left[ \frac{b(X_i, \zeta)}{\sqrt{p_i - p_i^2}} \mid \psi_i, p_i, \zeta \right] = \frac{u^1_i}{p_i} + \frac{u^0_i}{1 - p_i}$$

We start by verifying Assumption 8.2 for these variables, justifying application of the general CLT in Theorem 8.5. Observe that

$$E[f_{i,n}^2(\zeta)] \leq 2\delta^{-2}E[m^2(X_i) + m^2_0(X_i)] \zeta \lesssim E[m^2(X_i) + m^2_0(X_i)] < \infty$$

$$E[u^4_{i,n} \zeta] \leq 8\delta^{-4}E[(u^1_{i,n})^4 + (u^0_{i,n})^4] \zeta < \infty$$

In both lines, the first inequality follows by Young’s inequality and our propensity bound. This shows assumption 8.2.(a).

Next, we claim that $E[(u_{i,n} - u_i)^2 | \zeta] = o_p(1)$.

$$\left| \frac{u^1_i}{p_i} - \frac{u^1_{i,n}}{p_{i,n}} \right| \leq \delta^{-2} \left| u^1_{i,n} p_{i,n} - u^1_i p_i \right| \lesssim |p_{i,n} - p_i| |u^1_i| + |u^1_i - u^1_{i,n}|$$

$$\leq r_p^p |u^1_i| + |p_i| |u^1_i - u^1_{i,n}|$$

The first inequality by our propensity bound. The second inequality by telescoping. The final inequality follows since $|\tilde{p}_n - p_n| \leq r_p^p$ by assumption 8.1 and since we let $p = \tilde{p}_n$ known. Using this gives

$$E \left[ \left( \frac{u^1_i}{p_i} - \frac{u^1_{i,n}}{p_{i,n}} \right)^2 \mid \zeta \right] \lesssim (r_p^p)^2 E[(u^1_i)^2 | \zeta] + E[|u^1_i - u^1_{i,n}|^2 | \zeta] = o_p(1)$$

The proof for $d = 0$ is identical. This suffices to show $E[(u_{i,n} - u_i)^2 | \zeta] = o_p(1)$, completing the proof of assumption 8.2.(b). Define the target assignment variance

$$\Sigma_{11} = E[(p_i - p_i^2)u^2_i | \zeta] = E \left[ (p_i - p_i^2) \left( \frac{b_i - E[b_i | \psi_i, p_i, \zeta]}{\sqrt{p_i - p_i^2}} \right)^2 \right] \zeta = E \left[ (b_i - E[b_i | \psi_i, p_i, \zeta])^2 | \zeta \right]$$

Then assumption 8.2.(c) is trivially satisfied since $\zeta_n = \zeta \forall n \geq 1$ is constant. By Young’s
inequality and propensity bound, \( u_{1,n}^2 \leq 2\delta^{-2}((u_{1,n}^1)^2 + (u_{1,n}^0)^2) \). Then we have
\[
\begin{align*}
E[(u_{i,n}^2) \mathbb{1}((u_{i,n}^2) > z)|\zeta] &\leq E[2\delta^{-2}((u_{i,n}^1)^2 + (u_{i,n}^0)^2) \mathbb{1}((u_{i,n}^1)^2 + (u_{i,n}^0)^2) > z)|\zeta] \\
&\leq E[((u_{i,n}^1)^2 + (u_{i,n}^0)^2) \mathbb{1}((u_{i,n}^1)^2 + (u_{i,n}^0)^2) > (1/2)\delta^2 z)|\zeta] \\
&\leq 2 \sum_{d=0,1} E[((u_{i,n}^d)^2) \mathbb{1}((u_{i,n}^d)^2) > (1/4)\delta^2 z)|\zeta]
\end{align*}
\]

The final equality follows by fact 8.4 used in the general CLT proof. Then by our assumption on the tails of \( u_{1,n}^d \), we see that assumption 8.2.(d) is satisfied, and assumption 8.2.(e) is the same. This completes the verification of assumptions needed for Theorem 8.5.

Let \( F_{x,n} = \sigma(X_{1,n}, \pi_{x,n}^d, \pi_{x,n}^t, \tau, \zeta) \) and \( F_{(x,d),n} = \sigma(X_{1,n}, \pi_{x,n}^d, \pi_{x,n}^t, \tau, \tau, \zeta) \). Invoking Theorem 8.5 gives weak convergence
\[
\sqrt{n}E_n[T_i(D_i - p_{i,n})u_{i,n}]|F_{x,n} \Rightarrow \mathcal{N}(0, q\Sigma_{11})
\]
as well as
\[
\sqrt{n}E_n \left[ T_iD_i\ell^1_{i,n} + \frac{T_i(1-D_i)\ell^0_{i,n}}{1-p_{i,n}} \right]|F_{(x,d),n} \Rightarrow \mathcal{N} \left( 0, qE \left[ \frac{\sigma^2(X)}{p(X,\zeta)} + \frac{\sigma^2(X)}{1-p(X,\zeta)} \right] \zeta \right)
\]

(2) Selection Term: Next, we will apply Theorem 8.5 again using only the selection variables \( T_{1,n} \sim \text{Loc}_a(\psi', q) \) and different residuals. Define \( f_{1,n}(X_i) = c(X_i) = m_1(X_i) - m_0(X_i) \) and
\[
\tilde{u}_i \equiv \tilde{u}_{i,1} \equiv m_1(X_i) - m_0(X_i) - E[m_1(X_i) - m_0(X_i)|\psi', \zeta] \equiv \tilde{u}_1 - \tilde{u}_0
\]
This has the same form as \( f_{1,n}, u_{i,1,n} \) defined above, with \( \tilde{u}_i \) now a linear combination of \((\tilde{u}_1, \tilde{u}_0)\) with weights \((1, -1)\) instead of \((p_{i,n}, (1-p_{i,n})^{-1})\). Then an identical argument shows that assumption 8.2 is satisfied for this choice of \( \tilde{f}_{1,n}(X_i, \zeta) \) and \( \tilde{u}_i \). Then applying Theorem 8.5 with \( F'_{x,n} = \sigma(X_{1,n}, \pi_{x,n}^l, \zeta) \) and switching roles \( T_{1,n} \to 1 \) and \( D_{1,n} \to T_{1,n} \) gives weak convergence
\[
\sqrt{n}E_n[(T_i - q)(\tilde{u}_i - \tilde{u}_0)]|F'_{x,n} \Rightarrow \mathcal{N}(0, \Sigma')
\]
with variance
\[
\Sigma' = E[(q - q'^2)(\tilde{u}_i - \tilde{u}_0)^2|\zeta] = q(1-q)E[(c_i - E[c_i|\psi', \zeta])^2|\zeta]
\]

Residual Coupling: In the notation of coupling Lemma 8.4, the projection of \( f_{1,n} \) is
\[
h_n(\psi, p_a, \zeta') \equiv E[f_{1,n}(X_i, \zeta)|\psi = \psi, p_{i,n} = p_a, \zeta = \zeta']
\]
\[
= E \left[ \frac{b_n(X_i, \zeta)}{p_{i,n} - p_{i,n}^2} \right] \psi = \psi, p_{i,n} = p_a, \zeta = \zeta'] = \frac{F_1^3(\psi, p_a, \zeta')}{p_a} + \frac{F_0^3(\psi, p_a, \zeta')}{1-p_a}
\]

Then \( F_n \) satisfies the smoothness conditions in Definition 7.1 or 7.5 if \( F_n^d \) does for \( d = 0, 1 \). Similarly \( \tilde{F}(\psi', \zeta') \equiv E[e(X_i)|\psi' = \psi', \zeta = \zeta'] = F^1(\psi', \zeta') - F^0(\psi', \zeta') \) satisfies the
smoothness conditions. Under our assumptions (including the maximal inequalities for $\tilde{F}_i^d$ and $F_{i,n}^d$), Lemma 8.4 gives the couplings

$$\sqrt{n}E_n \left[ \frac{T_i(D_i - p_{i,n})}{\sqrt{p_{i,n} - p_{i,n}^2}} b_{i,n} \right] = \sqrt{n}E_n \left[ T_i(D_i - p_{i,n})u_{i,n} \right] + o_p(1)$$

$$\sqrt{n}E_n \left[ (T_i - q)c(X_i) \right] = \sqrt{n}E_n \left[ (T_i - q)\tilde{u}_i \right] + o_p(1)$$

To finish the proof, we apply the weak convergence and couplings above to the form of the IPW estimator. Consider the decomposition

$$q\hat{\theta} = qE_n \left[ \frac{T_i(D_i - p_{i,n})}{q(p_{i,n} - p_{i,n}^2)} Y_i \right] = E_n \left[ T_i c(X_i) \right] + E_n \left[ \frac{T_i(D_i - p_{i,n})}{\sqrt{p_{i,n} - p_{i,n}^2}} b_{i,n} \right]$$

$$qE_n \left[ \frac{T_i D_i c_i^1}{p_{i,n} - 1 - p_{i,n}} \right] = qE_n \left[ c(X_i) \right] + E_n \left[ (T_i - q)c(X_i) \right]$$

Using the work above gives the coupling

$$q\sqrt{n}(\hat{\theta} - \text{ATE}) = q\sqrt{n}E_n \left[ c(X_i) - \text{ATE} \right] + \sqrt{n}E_n \left[ (T_i - q)\tilde{u}_i \right]$$

$$+ \sqrt{n}E_n \left[ T_i(D_i - p_{i,n})u_{i,n} \right] + \sqrt{n}E_n \left[ \frac{T_i D_i c_i^1}{p_{i,n} - 1 - p_{i,n}} \right] + o_p(1)$$

$$\equiv A_n + B_n + C_n + D_n + o_p(1)$$

Note that $A_n \Rightarrow \mathcal{N}(0, q^2 \text{Var}(c(X)))$ by vanilla CLT. Let $\phi_a, \phi_b, \phi_c, \phi_d$ be the limiting characteristic functions for each of these four terms given by the weak convergence results above. For instance, $E[e^{it\epsilon_n} | \mathcal{F}_{x,n}] \overset{P}{\to} \phi_c(t) = e^{-\frac{t^2}{2}\text{Var}(\epsilon_n)}$. Then for any $t \in \mathbb{R}$ we have

$$E[e^{itq\hat{\theta}} | \mathcal{F}_{x,n}] \overset{P}{\to} \phi_c(t) = e^{-\frac{t^2}{2}\text{Var}(\epsilon_n)}$$

Then note that $e^{itA_n + B_n + C_n} (E[e^{itC_n} | \mathcal{F}_{x,d,n}] - \phi_d(t)) = o_p(1)$ and is totally bounded. Then by Lemma 8.16, $E[e^{itA_n + B_n + C_n} (E[e^{itD_n} | \mathcal{F}_{x,d,n}] - \phi_d(t)]) = o_p(1)$. Then we have

$$E[e^{itq\sqrt{n}(\hat{\theta} - \text{ATE})}] = E[e^{it(A_n + B_n + C_n + D_n)}] = E[e^{itA_n + B_n} E[e^{itC_n} | \mathcal{F}_{x,n}] | \zeta] + o_p(1)$$

The second equality follows since $B_n \in \mathcal{F}_{x,n}$. The fourth equality follows since $A_n \in \mathcal{F}_{x,n}^T$. The third equality by identical reasoning as above, using that $E[e^{itB_n} | \mathcal{F}_{x,n}] - \phi_b(t) = o_p(1)$ and is totally bounded. The fourth equality since $\mathcal{W}_{1,n} \perp \zeta$, then the final equality is by vanilla CLT. Then by adding variances and continuous mapping applied to $x \to q^{-1/2}x$, we have $\sqrt{q}(\hat{\theta} - \text{ATE}) | \zeta \Rightarrow \mathcal{N}(0, V(\psi, p))$ with the claimed limit.
\textbf{Fixed Regressors}: Consider that by our estimator decomposition

\[ q\sqrt{n}(\hat{\theta} - E_n[c(X_i)]) = q\sqrt{n}(\hat{\theta} - \text{ATE}) - q\sqrt{n}(E_n[c(X_i)] - \text{ATE}) \]

\[ = B_n + C_n + D_n + o_p(1) \]

The claim then follows by work above, setting \( A_n = 0 \), so \( \phi_n(t) = 1 \). This finishes the proof. \( \square \)

\section{8.3 Proofs - Design with a Pilot}

\textbf{Assumption 8.9}. Consider \( \hat{\sigma}_d^2(x) \) estimated out-of-sample. The following hold for \( d = 0,1 \)

(a) \( E[m_d(X)^2] < \infty \) and \( E[\sigma_d^2(X)^2] < \infty \) a.s.

(b) \( \|\hat{\sigma}_d - \sigma_d\|_{2,X} = o_p(1) \) or \( \|\hat{\sigma}_d^2 - \sigma_d^2\|_{2,X} = o_p(1) \)

(c) \( \min(\sigma_d^2(X), \hat{\sigma}_d^2(X)) \geq \sigma > 0 \) and \( \sigma_d^2(X) \leq B, (X, \zeta_n) \)-a.s.

(d) \( m_d(x) \) satisfies either Definition 7.1 or 7.5 and \( E[\max_{i=1}^n m_d(X_i)] = O(\log n) \)

\textbf{Proof of Theorem 4.3}. Define \( \zeta_n = (\hat{p}_d)_{d=0,1} \), so \( \zeta_n \perp (W_{i:n}, \pi_d^i, \pi_a^i, \tau_d, \tau^i) \). In the notation of coupling Lemma 8.4, the projection of \( f_{1,n}(X_i, \zeta_n) = \frac{b_n(X_i, \zeta)}{\sqrt{p_{i,n} - p_{i,n}^2}} \) is given by

\[ h_n(x, p_a, \zeta) \equiv E[f_{1,n}(X_i, \zeta)|X_i = x, p_{i,n} = p_a, \zeta_n = \zeta] \]

\[ = \frac{b_n(X_i, \zeta)}{\sqrt{p_{i,n} - p_{i,n}^2}}|X_i = x, p_{i,n} = p_a, \zeta_n = \zeta] = \frac{m_1(x)}{p_a} + \frac{m_0(x)}{1 - p_a} = h_n(x, p_a) \]

By the propensity bound \( p_a^{-1} \lor (1 - p_a)^{-1} < \delta^{-1} \), we see \( h_n \) satisfies the smoothness conditions in Definition 7.1 or 7.5 if \( m_d(x) \) does for \( d = 0,1 \), which we assume. Similarly, we have

\[ \tilde{h}_n(x, \zeta) \equiv E[c(X_i)|X_i = x, \zeta_n = \zeta] = m_1(x) - m_0(x) = \tilde{h}(x) \]

Then \( \tilde{h}(x) \) also satisfies one of the smoothness conditions. Note that

\[ u_{i,n} = \frac{b_n(X_i, \zeta)}{\sqrt{p_{i,n} - p_{i,n}^2}} - h_n(X_i, p_{i,n}, \zeta_n) = \frac{b_n(X_i, \zeta)}{\sqrt{p_{i,n} - p_{i,n}^2}} - \frac{b_n(X_i, \zeta)}{\sqrt{p_{i,n} - p_{i,n}^2}} = 0 \]

\[ u_i \equiv c(X_i) - \tilde{h}_n(X_i, \zeta_n) = c(X_i) - c(X_i) = 0 \]

Then Lemma 8.4 implies that

\[ \sqrt{n}E_n[(T_i - q)c(X_i)] = o_p(1) \quad \sqrt{n}E_n[T_i(D_i - p_{i,n})b_{i,n}] = o_p(1) \]

Next, we verify conditions in assumptions 8.1, 8.2 of the general CLT Theorem 8.5. Aside from \( E[m_d(X)^2] < \infty \), the remaining assumptions were only required for the conditional CLT of \( \sqrt{n}E_n[T_i(D_i - p_{i,n})u_{i,n}] = 0 \) in this case. Then we only need verify that \( E[(\hat{p}_{i,n} -} \]
Then \( p_i^2 | \zeta_n |^{1/2} = o_p(1) \). A calculation shows that
\[
\hat{p}(x) - p^*(x) = \frac{\hat{\sigma}_1(x)}{\sigma_1(x) + \sigma_0(x)} - \frac{\sigma_1(x)}{\sigma_1(x) + \sigma_0(x)} = \frac{(\hat{\sigma}_1 - \sigma_1)(x)\sigma_0(x) + \sigma_1(x)(\sigma_0 - \hat{\sigma}_0)(x)}{(\hat{\sigma}_1 + \hat{\sigma}_0)(x)(\sigma_1 + \sigma_0)(x)}
\]
Then almost surely
\[
(\hat{p}(X_i) - p^*(X_i))^2 \leq \frac{1}{8\sigma^4} \sum_{d=0,1} (\hat{\sigma}_d - \sigma_d)^2(X_i)
\]
For the case \( \|\hat{\sigma}_d - \sigma_d\|_{2,\chi} = o_p(1) \), the claim follows immediately. Otherwise, observe that
\[
\sum_{d=0,1} (\hat{\sigma}_d - \sigma_d)^2(X_i) = \sum_{d=0,1} (\hat{\sigma}_d^2 - \sigma_d^2)(X_i) \leq \sum_{d=0,1} (\hat{\sigma}_d - \sigma_d)^2(X_i)
\]
This shows that \( \|\hat{\sigma}_d^2 - \sigma_d^2\|_{2,\chi} = o_p(1) \) is also sufficient. Let \( F(x,d), n = \sigma(X_1:n, \pi_1, \pi_2, \tau_1, \tau_2, \zeta_1) \).

By Theorem 8.5, we have
\[
\sqrt{nE_n} \left[ \frac{T_iD_i\epsilon_i^1}{p_i,n} + \frac{T_i(1 - D_i)\epsilon_i^0}{1 - p_i,n} \right] \Rightarrow N \left( 0, \frac{\sigma^2(X)}{p^*(X)} + \frac{\sigma_0^2(X)}{1 - p^*(X)} \right)
\]
Summarizing using the coupling above, we have shown that
\[
q\sqrt{n}(\hat{\theta} - \theta) = q\sqrt{nE_n} [c(X_i) - \theta] + \sqrt{nE_n} \left[ \frac{T_iD_i\epsilon_i^1}{p_i,n} - \frac{T_i(1 - D_i)\epsilon_i^0}{1 - p_i,n} \right] + o_p(1)
\]
The conclusion follows by repeating the argument at the end of the proof of Theorem 3.17.

**Proposition 8.10 (Consistent Regressions).** **Require the assumptions and definitions in Theorem 4.8.** Let \( p(x) \) be a fixed, known propensity score. If \( T_1:n \sim \text{Loc}_n(\psi, q) \) and \( D_1:n \sim \text{Loc}(\hat{\psi}, q | T_1:n) \) then \( \sqrt{n}(\hat{\theta} - \text{ATE}) \Rightarrow N(0, V) \)

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

**Proof of Proposition 8.10 and Theorem 4.8.** First, we consider coupling arguments. Define \( \zeta_n = (\hat{m}_d)_d \). In the notation of Lemma 8.4, let \( f_n(X_i, \zeta_n) = \hat{c}(X_i) \). Let \( \psi = (\psi_1, \psi_2) \in \mathbb{R}^2 \)
\[
h_n(\psi, p_a, \zeta) = E[f_n(X_i, \zeta_n) | \psi(X_i, \zeta_n) = \psi, p_i,n = p_a, \zeta_n = \zeta] = \psi_1
\]
Then \( h_n \) is a Lipschitz function of \( \psi = (\psi_1, \psi_2) \). In particular, in the notation of Lemma 8.4 \( h_{i,n} = \hat{c}(X_i) \), so \( u_{i,n} = f_n(X_i, \zeta_n) - h_{i,n} = 0 \). Note that
\[
\max_{i=1}^n h_{i,n}^2 = \max_{i=1}^n \hat{c}(X_i)^2 \leq 2 \max_d \|\hat{m}_d\|_{\infty} \leq 2m^2
\]
Then $E[\max_{i=1}^{n} h_{i,n}^2] = O(1)$, so the relevant maximal inequality is satisfied. Then Lemma 8.4 implies
\[
\sqrt{n} E_n[(T_i - q)\hat{c}(X_i)] = \sqrt{n} E_n[(T_i - q)u_{i,n}] + o_p(1) = o_p(1)
\]
Similarly, let $f_n'(X_i, \zeta_n) = \frac{\hat{b}_n(X_i)}{p_{i,n} - \hat{p}_{i,n}}$ and define projection
\[
h'_n(\psi, p_a, \zeta) = E \left[ \frac{\hat{b}_n(X_i)}{\sqrt{p_{i,n} - \hat{p}_{i,n}^2}} \left( \hat{c}(X_i), \hat{b}_n(X_i) \right) = (\psi_1, \psi_2), p_{i,n} = p_a, \zeta_n = \zeta \right] = E \left[ \frac{\hat{b}_n(X_i)}{\sqrt{p_{i,n} - \hat{p}_{i,n}^2}} \right] = o_p(1)
\]
Since $p_a \sqrt{1 - p_a} \leq \delta^{-1}$ by assumption, $h'_n$ is Lipschitz in $(\psi_1, \psi_2)$. Again, $f_n'(X_i, \zeta_n) - h'_n = 0$ and $E[\max_{i=1}^{n} (h'_{i,n})^2] = O(1)$ by similar arguments. Then Lemma 8.4 implies
\[
\sqrt{n} E_n \left[ T_i(D_i - p_{i,n}) \frac{\hat{b}_n(X_i)}{\sqrt{p_{i,n} - \hat{p}_{i,n}^2}} \right] = o_p(1)
\]
Then by telescoping, the estimator expansion
\[
q\sqrt{n}(\hat{\theta} - \text{ATE}) = q\sqrt{n} E_n \left[ c(X_i) - \text{ATE} \right] + \sqrt{n} E_n \left[ (T_i - q)(c(X_i) - \hat{c}(X_i) + \hat{c}(X_i)) \right] + \sqrt{n} E_n \left[ \frac{T_i D_i \epsilon_i}{p_{i,n}} - \frac{T_i (1 - D_i) \epsilon_i^0}{1 - p_{i,n}} \right] + o_p(1)
\]
Define $\mathcal{F}_{x,n} = \sigma(X_{1:n}, \pi_1^n, \tau^t, \zeta_n)$. Then $(g_{a,s})_{a,s} \in \sigma(\hat{\psi}_{1:n}, p_{1:n,n}, \pi_1^n, \tau^t) \subseteq \mathcal{F}_{x,n}$, and for all $i \in [n], \frac{b_{i,n} - \hat{b}_{i,n}}{\sqrt{p_{i,n} - \hat{p}_{i,n}^2}} \in \mathcal{F}_{x,n}$. Also $\mathcal{F}_{x,n} \perp \tau^d$ by randomization. Then Lemma 8.20 applies, showing $E[C_n | \mathcal{F}_{x,n}] = 0$ and conditional variance
\[
\text{Var}(C_n | \mathcal{F}_{x,n}) \leq 2E_n \left[ \frac{(b_{i,n} - \hat{b}_{i,n}(X_i))^2}{p_{i,n} - \hat{p}_{i,n}^2} \right] = 2E_n \left[ \left( \frac{m_1(X_i) - \hat{m}_1(X_i)}{p_{i,n}} + \frac{m_0(X_i) - \hat{m}_0(X_i)}{1 - p_{i,n}} \right)^2 \right] \leq 4\delta^{-2} \sum_{d=0,1} E_n[(m_d(X_i) - \hat{m}_d(X_i))^2] = O_p(n^{-2\delta_m})
\]
The last line by conditional Markov, since $\zeta_n \perp W_{1:n}$ we have $E[E_n[(m_d(X_i) - \hat{m}_d(X_i))^2] | \zeta_n] = \|\hat{m}_d - m_d\|_{2,P} = O_p(n^{-2\delta_m})$. Then by conditional Chebyshev (Lemma 8.16), $C_n = O_p(n^{-\delta_m})$. An identical argument, using out-of-sample condition $\zeta_n \perp W_{1:n}$, estimator rates, and Lemma 8.20 shows $B_n = O_p(n^{-\delta_m})$. Summarizing the work above
\[
q\sqrt{n}(\hat{\theta} - \theta) = q\sqrt{n} E_n \left[ c(X_i) - \theta \right] + \sqrt{n} E_n \left[ \frac{T_i D_i \epsilon_i}{p_{i,n}} - \frac{T_i (1 - D_i) \epsilon_i^0}{1 - p_{i,n}} \right] + o_p(1)
\]
This is identical to the final coupling in the proof of Theorem 4.3, which shows the conclusion for Theorem 4.8. For Proposition 8.10, the required condition \( (E[|\hat{p}_n - p_i|^2)^2] \) is trivial, since we may take \( \hat{p} = p \) (known propensity).

**Calculations for Section 4.1.1.** Define \( c'(X) = E[Y(1) - Y(0)|S] = E[c(X)|S] \) and \( v_0^2(S) = \text{Var}(Y(d)|S) \). By the law of total variance

\[
\text{Var}(c(X)) = \text{Var}(E[c(X)|S]) + E[\text{Var}(c(X)|S)] = \text{Var}(c'(S)) + E[\text{Var}(c(X)|S)]
\]

\[
\text{Var}(Y(d)|S) = E[\text{Var}(Y(d)|X)|S] + E[\text{Var}(Y(d)|X)|S] = \text{Var}(m_d(X)|S) + E[\sigma_d^2(X)|S]
\]

We can use these to rewrite the asymptotic variance in TM (2020) as

\[
\text{Var}(c'(S)) + E\left[ \frac{v_1^2(S)}{p(S)} + \frac{v_0^2(S)}{1 - p(S)} \right] = \text{Var}(c(X)) - E[\text{Var}(c(X)|S)] + E\left[ \frac{\sigma_1^2(S)}{p(S)} + \frac{\sigma_0^2(S)}{1 - p(S)} \right] + E\left[ \frac{1}{p(S)} \text{Var}(m_1(X)|S) + \frac{1}{1 - p(S)} \text{Var}(m_0(X)|S) \right]
\]

Since \( c(X) = m_1(X) - m_0(X) \), adding the 2nd and 4th terms gives

\[
E\left[ \frac{1 - p(S)}{p(S)} \text{Var}(m_1(X)|S) + \frac{1}{p(S)} \text{Var}(m_0(X)|S) + 2 \text{Cov}(m_1(X), m_0(X)|S) \right] = E\left[ \text{Var}(b(X; S)|S) \right] = E[(b(X; S) - E[b(X; S)|S])^2]
\]

Alternatively, this follows from Theorem 3.17 with \( \psi(x) = S(x) \) and \( p = p(S(x)) \).

**Proof of Proposition 4.11.** First, consider the case \( (D_i)_{i \in I_k} \sim \text{Loc}_n(\psi, p|I_k) \). Let \( \xi_n \) be external randomness forming the partition: \( (I_k)_{k=1}^K \sim F_n(\xi_n) \) for deterministic \( F_n \). Define the operator \( E_{n,k}[Z] = n^{-1} \sum_{i \in I_k} Z \). A standard expansion of the AIPW estimator (algebra) gives, for the \( k \)th fold estimator

\[
\hat{\theta}_k = E_{n,k}[c(X_i)] + E_{n,k}\left[ \frac{(D_i - p_n(X_i))}{p_n(X_i)} \left( \frac{m_1 - \hat{m}_{1,k}(X_i)}{p_n(X_i)} + \frac{m_0 - \hat{m}_{0,k}(X_i)}{1 - p_n(X_i)} \right) \right]
\]

\[
+ E_{n,k}\left[ \left( \frac{D_i e_i^1}{p_n(X_i)} - \frac{(1 - D_i) e_i^0}{1 - p_n(X_i)} \right) \right] = A_{n,k} + B_{n,k} + C_{n,k}
\]

Consider \( B_{n,k} \). Let \( (g_{a,s}^k)_{a,s} \) denote the groups for the \( k \)th design \( (D_i)_{i \in I_k} \sim \text{Loc}_n(\psi, p|I_k) \). Let \( \pi_n^d(k) \) and \( \pi_n^d \) denote tie-breaking randomness and within-group randomization variables for the \( k \)th design and \( \pi_n^d = (\pi_n^d(k))_k \). Define \( F_{n,k} = \sigma((Y_i, D_i, X_i)_{i \in I_k}, X_{1:n}, \pi_n^d, \xi_n) \). We check the conditions to apply Lemma 8.20. By cross-fitting of the design, \( \{i : 1 < i \in I_k \} = I_{a,s} g_{a,s}^k \). Observe that \( (g_{a,s}^k)_{a,s} \in \sigma((X_i)_{i \in I_k}, \pi_n^d(k), \xi_n) \subseteq F_{n,k} \) and, in the notation of Lemma 8.20

\[
h_n(W_i) = \left( \frac{m_1 - \hat{m}_{1,k}(X_i)}{p_n(X_i)} + \frac{m_0 - \hat{m}_{0,k}(X_i)}{1 - p_n(X_i)} \right) \in F_{n,k}
\]

70
By Lemma 8.19.(i) and cross-fitting of the design.

\[(D_i)_{i \in I_k} \in \sigma\left(\left(\phi^i_{a,s}, \tau^d_k : k' \neq k\right) \in \sigma((X_i)_{i \in I_k}, \pi^d_n(k'), \xi, \tau^d_k : k' \neq k)\right) \quad (8.5)\]

Then we see that

\[F_{n,k} \subseteq \sigma((W_i)_{i \in I_k}, (D_i)_{i \in I_k}, X_{1:n}, \pi^d_n, \xi_n) \subseteq \sigma((W_i)_{i \in I_k}, X_{1:n}, \pi^d_n, \xi, \tau^d_k : k' \neq k) \perp \tau^d_k\]

Then we may apply Lemma 8.20 with \(S_i = 1(i \in I_k)\), which gives \(E[B_{n,k}|F_{n,k}] = 0\) and

\[
\text{Var}(\sqrt{n}B_{n,k}|F_{n,k}) \leq 2E_n \left[1(i \in I_k) \left(\frac{(m_1 - \hat{m}_{1,k})(X_i)}{p_n(X_i)} + \frac{(m_0 - \hat{m}_{0,k})(X_i)}{1 - p_n(X_i)}\right)^2\right]
\leq 4\delta^{-2}E_n \left[1(i \in I_k) \sum_{d=0,1} (m_d - \hat{m}_{d,k})(X_i)^2\right]
\]

Now, taking a conditional expectation

\[E \left[E_n \left[1(i \in I_k) \sum_{d=0,1} (m_d - \hat{m}_{d,k})(X_i)^2\right]|(Y_i, D_i, X_i)_{i \in I_k}, \xi_n\right] = (|I_k|/n) \sum_{d=0,1} \|m_d - \hat{m}_{d,k}\|_2^2\]

To see the equality, note that \(1(i \in I_k) \in \sigma(\xi_n)\) and by tower law and Equation 8.5 above

\[E \left[1(i \in I_k)(m_d - \hat{m}_{d,k})(X_i)^2|(Y_i, D_i, X_i)_{i \in I_k}, \xi_n\right] = E \left[E \left[1(i \in I_k)(m_d - \hat{m}_{d,k})(X_i)^2|(W_i)_{i \in I_k}, (\pi^d_n(k'), \tau^d_k : k' \neq k, \xi_n)\right|(Y_i, D_i, X_i)_{i \in I_k}, \xi_n\right] = 1(i \in I_k) \|m_d - \hat{m}_{d,k}\|_2^2\]

Then by conditional Markov and conditional Chebyshev (Lemma 8.16), we have shown that \(\sqrt{n}B_{n,k} = o_p(1)\), so that \(\sqrt{n} \sum_{k=1}^{K} B_{n,k} = o_p(1)\). Then returning to the decomposition above

\[\sqrt{n}(\hat{\theta} - \text{ATE}) = \sqrt{n} \left(\sum_{k=1}^{K} \hat{\theta}_k - \text{ATE}\right) = \sqrt{n}E_n[c(X_i) - \text{ATE}] + \sqrt{n} \sum_{k=1}^{K} B_{n,k}
+ \sqrt{n}E_n \left[\left(D_i \epsilon^1_i \frac{1 - D_i \epsilon^0_i}{p_n(X_i)} - \left(\left(D_i \epsilon^1_i \frac{1 - D_i \epsilon^0_i}{p_n(X_i)} - \left(\left(D_i \epsilon^1_i \frac{1 - D_i \epsilon^0_i}{p_n(X_i)} - (1 - D_i) \epsilon^0_i\right)\right)\right)\right)\right]
= \sqrt{n}E_n[c(X_i) - \text{ATE}] + \sqrt{n}E_n \left[\left(D_i \epsilon^1_i \frac{1 - D_i \epsilon^0_i}{p_n(X_i)} - \left(\left(D_i \epsilon^1_i \frac{1 - D_i \epsilon^0_i}{p_n(X_i)} - (1 - D_i) \epsilon^0_i\right)\right)\right] + o_p(1)

Similar reasoning as in the proof of 3.17 may now be used to show the claimed CLT. The case \(D_i \sim p(X_i)\) follows from Chernozhukov et al. (2017), Theorem 5.1.
8.4 Proofs - Inference

8.4.1 Proof of Theorem 5.3

Proof. First, we show the claim in (1). Suppose \( p = p(\psi, \zeta) \) and \( \psi' = \psi \). In this case

\[
F_i^d = E[m_d(X_i)|\psi_i, p(\psi_i, \zeta), \zeta] = E[m_d(X_i)|\psi_i, \zeta]
\]

\[
u_i^d = m_d(X_i) - E[m_d(X_i)|\psi_i, \zeta]
\]

By Theorem 3.17, in this case the limiting variance of \( \sqrt{n}(\hat{\theta} - \text{ATE}) \) is

\[
V(\psi, (q, p)) = q \text{Var}(c(X)) + (1-q)E[(c - E[c|\psi, \zeta])^2|\zeta]
\]

\[
+ E[(b - E[b|\psi, p, \zeta])^2|\zeta] + E \left[ \frac{\sigma_1^2(X)}{p(X, \zeta)} + \frac{\sigma_0^2(X)}{1-p(X, \zeta)} \right] \zeta
\]

Since \( \zeta \bot X_i \) we have \( E[c(X_i)|\zeta] = E[c(X_i)] = \text{ATE} \) and

\[
\text{Var}(c(X_i)) = E[(c(X_i) - \text{ATE})^2] = E[(c(X_i) - \text{ATE})^2|\zeta] = E[(c(X_i) - E[c(X_i)|\zeta])^2|\zeta]
\]

\[
= \text{Var}(c(X_i))|\zeta] = E[(F_1 - F_0) + (u_1 - u_0)]^2|\zeta - \text{ATE}^2
\]

\[
= E[(F_1 - F_0)^2|\zeta] + E[(u_1 - u_0)]^2|\zeta - \text{ATE}^2
\]

Similarly we have

\[
E[(c(X_i) - E[c(X_i)|\psi_i, \zeta])^2|\zeta] = E[(u_1 - u_0)^2|\zeta]
\]

Then the first two terms can be written

\[
q \text{Var}(c(X_i)) + (1-q)E[(c(X_i) - E[c(X_i)|\psi_i, \zeta])^2|\zeta] = qE[(F_1 - F_0)^2|\zeta] + E[(u_1 - u_0)^2|\zeta - q \text{ATE}^2
\]

Under our assumptions, with \( p_i = p(\psi_i, \zeta) \) we have

\[
E[b(X_i)|\psi_i, p(\psi_i, \zeta), \zeta] = E \left[ m_1(X_i) \sqrt{\frac{1-p_i}{p_i}} + m_0(X_i) \sqrt{\frac{p_i}{1-p_i}} \right] \psi_i, \zeta
\]

\[
= F_1^i \sqrt{\frac{1-p_i}{p_i}} + F_0^i \sqrt{\frac{p_i}{1-p_i}}
\]

Using this, the middle term becomes

\[
E[(b(X_i) - E[b(X_i)|\psi_i, p(\psi_i, \zeta), \zeta])^2|\zeta] = E \left[ \left( u_1^i \sqrt{\frac{1-p_i}{p_i}} + u_0^i \sqrt{\frac{p_i}{1-p_i}} \right)^2 \zeta \right]
\]

\[
= E \left[ (u_1^i)^2 \frac{1-p_i}{p_i} + (u_0^i)^2 \frac{p_i}{1-p_i} + 2u_1^iu_0^i \right] \zeta
\]

\[
= E \left[ (u_1^i)^2 \frac{1}{p_i} + (u_0^i)^2 \frac{1}{1-p_i} \right] \zeta - E[(u_1^i - u_0^i)^2|\zeta]
\]

72
Similarly for

Then adding all four terms gives

\[
qE[(F_i^1 - F_i^0)^2 | \zeta] + E[(u_i^1 - u_i^0)^2 | \zeta] - q \text{ATE}^2 + E \left[ \frac{(u_i^1)^2}{p_i} + \frac{(u_i^0)^2}{1 - p_i} \right] \zeta \\
= E \left[ \frac{(u_i^1)^2 + \sigma_i^2(X_i)}{p_i} \right] + E \left[ \frac{(u_i^0)^2 + \sigma_i^0(X_i)}{1 - p_i} \right] \zeta + qE[(F_i^1 - F_i^0)^2 | \zeta] - q \text{ATE}^2
\]

Next, we write this in terms of estimable quantities. Note that

\[
E \left[ \frac{Y_i(1)^2}{p_i} \right] = E \left[ \frac{(m_1(X_i) + \epsilon_i^1)^2}{p_i} \right] = E \left[ \frac{(F_i^1 + u_i^1 + \epsilon_i^1)^2}{p_i} \right] = E \left[ \frac{(F_i^1)^2 + (u_i^1)^2 + (\epsilon_i^1)^2}{p_i} \right]
\]

To see the last two equalities, note that since \( p_i = p(\psi, \zeta) \)

\[
E[p_i^{-1} u_i F_i^1 | \zeta] = E[p_i^{-1} E[u_i F_i^1 | \psi_i, \zeta] | \zeta] = E[p_i^{-1} F_i^1 E[u_i | \psi_i, \zeta] | \zeta] = 0 \\
E[p_i^{-1} \epsilon_i^1 F_i^1 | \zeta] = E[p_i^{-1} F_i^1 E[\epsilon_i^1 | X_i, \zeta] | \zeta] = E[p_i^{-1} F_i^1 E[\epsilon_i^1 | X_i] | \zeta] = 0 \\
E[p_i^{-1} \epsilon_i^1 u_i | \zeta] = E[p_i^{-1} u_i E[\epsilon_i^1 | X_i, \zeta] | \zeta] = E[p_i^{-1} u_i E[\epsilon_i^1 | X_i] | \zeta] = 0 \\
E[p_i^{-1} (\epsilon_i^1)^2 | \zeta] = E[p_i^{-1} E[(\epsilon_i^1)^2 | X_i, \zeta] | \zeta] = E[p_i^{-1} E[(\epsilon_i^1)^2 | X_i] | \zeta] = E[p_i^{-1} \sigma_i^2(X_i) | \zeta]
\]

Similarly for \( d = 0 \)

\[
E \left[ \frac{Y_i(0)^2}{1 - p_i} \right] = E \left[ \frac{(F_i^0)^2 + (u_i^0)^2 + \sigma_i^0(X_i)}{1 - p_i} \right] \zeta
\]

Comparing the calculations above, we see that

\[
V(\psi, (q, p)) = E \left[ \frac{Y_i(1)^2}{p_i} \right] + E \left[ \frac{Y_i(0)^2}{1 - p_i} \right] \zeta - E \left[ \frac{(F_i^1)^2}{p_i} \right] \zeta - E \left[ \frac{(F_i^0)^2}{1 - p_i} \right] \zeta + qE[(F_i^1 - F_i^0)^2 | \zeta] - q \text{ATE}^2
\]

Now observe that

\[
E \left[ \frac{(F_i^1)^2}{p_i} \right] + E \left[ \frac{(F_i^0)^2}{1 - p_i} \right] \zeta - qE[(F_i^1 - F_i^0)^2 | \zeta] \\
= E \left[ \frac{(F_i^1)^2(1 - q p_i)}{p_i} \right] + E \left[ \frac{(F_i^0)^2 (1 - q (1 - p_i))}{1 - p_i} \right] \zeta + 2qE[F_i^1 F_i^0 | \zeta]
\]

Summarizing, we have shown that

\[
V(\psi, (q, p)) = E \left[ \frac{Y_i(1)^2}{p_i} \right] + E \left[ \frac{Y_i(0)^2}{1 - p_i} \right] \zeta - E \left[ \frac{(F_i^1)^2(1 - q p_i)}{p_i} \right] \zeta \\
- E \left[ \frac{(F_i^0)^2 (1 - q (1 - p_i))}{1 - p_i} \right] \zeta - 2qE[F_i^1 F_i^0 | \zeta] - q \text{ATE}^2
\]
Then by Lemma 8.12 we have

\[ V(\psi, (q, p)) = \hat{v}_{1,2} - \hat{v}_1 \left[ \frac{1 - qp}{p} \right] - \hat{v}_0 \left[ \frac{1 - q(1 - p)}{1 - p} \right] - 2q\hat{v}_{10} - q\theta^2 + o_p(1) \]

Next, consider the limiting variance \( V(\psi, p) - q \text{Var}(c(X)) \) for the fixed regressor case. By the work above, this is

\[ (1 - q)E[(u_i^1 - u_i^0)^2|\zeta] + E \left[ \frac{(u_i^1)^2}{p_i} + \frac{(u_i^0)^2}{1 - p_i} \right] - E[(u_i^1 - u_i^0)^2|\zeta] + E \left[ \frac{\sigma_i^2(X_i^1)}{p_i} + \frac{\sigma_0^2(X_i)}{1 - p_i} \right] - E \left[ \frac{\sigma_i^2(X_i)}{p_i} + \frac{\sigma_0^2(X_i)}{1 - p_i} \right] \leq E \left[ \frac{(u_i^1)^2}{p_i} + \frac{(u_i^0)^2}{1 - p_i} \right] \]

Then comparing with the calculations above.

\[ V(\psi, p) - q \text{Var}(c(X)) \leq E \left[ \frac{Y_i(1)^2}{p_i} \right] + E \left[ \frac{Y_i(0)^2}{1 - p_i} \right] - E \left[ \frac{(F_i^1)^2}{p_i} \right] - E \left[ \frac{(F_i^0)^2}{1 - p_i} \right] = \hat{v}_{1,2} - \hat{v}_1 \left[ p^{-1} \right] - \hat{v}_0 \left[ (1 - p)^{-1} \right] + o_p(1) \]

The final equality follows by Lemma 8.12. Conservativeness is given by the residual variance

\[ qE[(u_i^1 - u_i^0)^2|\zeta] = qE[(c - E[c|\psi, \zeta])^2|\zeta] = qE[\text{Var}(c(X_i)|\psi, \zeta)|\zeta] \]

This completes the proof of claim (1).

Next, we show claim (2). In this general case we have \( p_i = p(X_i, \zeta) \)

\[ F_i^d = E[m_d(X_i)|\psi, p(X_i, \zeta), \zeta] \]
\[ u_i^d = m_d(X_i) - E[m_d(X_i)|\psi, p(X_i, \zeta), \zeta] \]

Identical calculations, substituting the new definitions of \( F_i^d, u_i^d \), show the same identities above for each of the variance components. Then we have

\[ V(\psi, p) = \text{Var}(c(X_i)) + E[(b - E[b|\psi, p_i, \zeta])^2|\zeta] + E \left[ \frac{\sigma_i^2(X_i)}{p_i} + \frac{\sigma_0^2(X_i)}{1 - p_i} \right] \]
\[ = E[(F_i^1 - F_i^0)^2|\zeta] + E[(u_i^1 - u_i^0)^2|\zeta] - ATE^2 \]
\[ + E \left[ \frac{(u_i^1)^2}{p_i} + \frac{(u_i^0)^2}{1 - p_i} \right] - E[(u_i^1 - u_i^0)^2|\zeta] + E \left[ \frac{\sigma_i^2(X_i)}{p_i} + \frac{\sigma_0^2(X_i)}{1 - p_i} \right] \]
\[ = E \left[ \frac{Y_i(1)^2}{p_i} \right] + E \left[ \frac{Y_i(0)^2}{1 - p_i} \right] - ATE^2 \]
\[ - E \left[ \frac{F_i(1)^2}{p_i} \right] - E \left[ \frac{F_i(0)^2}{1 - p_i} \right] + E[(F_i^1 - F_i^0)^2|\zeta] \]
Note that the last line is
\[
E \left[ \frac{(F_i^1)^2}{p_i} \right] + E \left[ \frac{(F_i^0)^2}{1-p_i} \right] - E[(F_i^1 - F_i^0)^2] \]
\[
= E \left[ \frac{(F_i^1)^2 - p_i}{p_i} \right] + E \left[ \frac{(F_i^0)^2 - p_i}{1-p_i} \right] + 2E[F_i^1 F_i^0 | \zeta]
\]

Then by Lemma 8.12
\[
V(\psi, p) = \text{Var}_n \left( \frac{(D_i - p_{i,n}) Y_i}{p_{i,n} - p_{i,n}^2} \right) - \hat{\nu}_1 \left[ \frac{1-p_i}{p_i} \right] - \hat{\nu}_0 \left[ \frac{p_i}{1-p_i} \right] - 2\hat{\nu}_{10} + o_p(1)
\]

This completes the proof of the second claim. For the fixed regressors case, by the work above
\[
V(\psi, p) - \text{Var}(c(X_i)) = E \left[ \frac{(u_i^1)^2 + \sigma^2(X_i)}{p_i} + \frac{(u_i^0)^2 + \sigma^2(X_i)}{1-p_i} \right] - E[(u_i^1 - u_i^0)^2 | \zeta]
\]
\[
\leq E \left[ \frac{(u_i^1)^2 + \sigma^2(X_i)}{p_i} + \frac{(u_i^0)^2 + \sigma^2(X_i)}{1-p_i} \right] - E \left[ \frac{(F_i^1)^2}{p_i} \right] - E \left[ \frac{(F_i^0)^2}{1-p_i} \right]
\]

Claim (3) follows from (2) and Lemma 8.12, noting that for the general case \( \psi'_i \in \sigma(\psi) \)
\[
\text{Var}(c(X_i)) - q \text{Var}(c(X_i)) + (1-q)E[(c - E[c|\psi'_i, \zeta])^2 | \zeta]
\]
\[
= (1-q) \left[ \text{Var}(c(X_i)) - E[\text{Var}(c(X_i)|\psi'_i, \zeta)] | \zeta \right] = (1-q) \text{Var}(E[c(X_i)|\psi'_i, \zeta]) | \zeta)
\]

This finishes the proof.

\( \square \)

**Assumption 8.11 (Inference).** Require assumption 8.1 with \( \zeta_n = \zeta \) for \( n \geq 1 \) and \( \hat{p} = p \) and assumption 8.3. Define the projection
\[
F_n^d(\psi, p, \zeta') = E[m_d(X_i) | \psi(X_i, \zeta) = \psi, p_n(X_i, \zeta) = p_n, \zeta = \zeta']
\]

Also define \( F_{i,n}^d = F_n^d(\psi_{i,n}, p_{i,n}, \zeta) \). Similarly, define \( F_i^d = E[m_d(X_i) | \psi(X_i, \zeta), p(X_i, \zeta, \zeta), \zeta_n = m_d(X_i) - F_i^d, \) and \( u_i^d = m_d(X_i) - F_i^d \). Define \( \sigma^2_{u,i,n} = E[(u_i,n)^2 | \psi_i, p_i, \zeta] \) for \( u_i,n = u_i^d \) and \( d = 0, 1 \). Require that
\( (1) \) \( E[Y(d)^4] < \infty \) and \( E[(F_i^d)^4 | \zeta] < \tilde{F} \) and \( E[(F_i^d)^4 | \zeta] < \tilde{F} < \infty, \) \( \zeta \)-a.s. for \( d = 0, 1 \)
\( (2) \) \( \sigma^2_{u,i,n} < \tilde{\sigma}^2 \) and \( \sigma^2(\zeta) \leq \tilde{\sigma}^2 < \infty \) for \( d = 0, 1 \)
\( (3) \) \( F_n^d(\psi, p, \zeta) \) satisfies Hölder conditions in either Definition 7.1 or 7.5 for \( d = 0, 1 \).
\( (4) \) \( E[(F_i^d - F_i^d)^2]^{1/2} = O_p(r_F^d) \) for \( d = 0, 1 \).
\( (5) \) \( \gamma \) is an admissible matching as in Definition 5.1
\( (6) \) \( \overline{F}_n \lor L_n = o(n) \) and \( r_n^\psi \lor r_n^F \lor r_n^p = o(1) \)
Lemma 8.12. Suppose that assumptions 8.1, 8.3, and 8.11 hold. Then the IPW second moment (for the selected units)

\[ \hat{v}_{1,2} = E_n \left[ \frac{T_i}{q} \left( \frac{(D_i - p_{i,n})Y_i}{p_{i,n} - p_{i,n}^2} \right)^2 \right] = E \left[ \frac{Y_i(1)^2}{p_i} \right] + E \left[ \frac{Y_i(0)^2}{1 - p_i} \right] + O_p(n^{-1/2}) + O_p(r_n^p) \]

Let G be any bounded, Lipschitz function. Then the matching terms satisfy

\[ \hat{v}_1[G] = E_n \left[ \frac{T_i}{q} G(p_{i,n}) Y_i Y_{\gamma(i)} \right] = E \left[ G(p_i)(F_i^1)^2 | \xi \right] + O_p(r_{1,n}) \]

\[ \hat{v}_0[G] = E_n \left[ \frac{T_i}{q} (1 - D_i)^2 G(p_{i,n}) Y_i Y_{\gamma(i)} \right] = E \left[ G(p_i)(F_i^n)^2 | \xi \right] + O_p(r_{0,n}) \]

\[ \hat{v}_{10} = 2n^{-1} \sum_{1 \leq i < j \leq n} \frac{T_i T_j}{q} \frac{1(g(i) = g(j))}{\text{Card}(g)} \frac{D_i(1 - D_j)Y_i Y_j}{p_{i,n}(1 - p_{j,n})} = E[F_i^1 F_i^n | \xi] + O_p(r_{10,n}) \]

With convergence rates

\[ r_{1,n} \vee r_{0,n} = O(n^{-1/2}) + O(r_n^p) + O(r_n^F) + O(r_n^\psi) \]

\[ r_{10,n} = O(n^{-1/2}) + O(r_n^F) + O(r_n^\psi) + O((\bar{K}_n/n)^{1/2}) + O(\sqrt{E_n}/n) \]

Proof. First, we show consistency for \( \hat{\theta} \). From Theorem 8.5, we have \( \hat{\theta} = \text{ATE} + O_p(n^{-1/2}) \), so \( (\hat{\theta})^2 = \text{ATE}^2 + O_p(n^{-1/2}) \). Since \( D_i(1 - D_i) = 0 \), the first term is

\[ E_n \left[ \frac{T_i}{q} \left( \frac{(D_i - p_{i,n})Y_i}{p_{i,n} - p_{i,n}^2} \right)^2 \right] = E_n \left[ \frac{T_i}{q} \left( \frac{D_i Y_i(1)}{p_{i,n}^2} - \frac{(1 - D_i)Y_i(0)}{1 - p_{i,n}} \right)^2 \right] \]

Consider the \( d = 1 \) term

\[ E_n \left[ \frac{T_i}{q} \frac{D_i Y_i(1)^2}{p_{i,n}^2} \right] = E_n \left[ \frac{T_i}{q} \frac{(D_i - p_{i,n})Y_i(1)^2}{p_{i,n}^2} \right] + E_n \left[ \frac{T_i}{q} \frac{Y_i(1)^2}{p_{i,n}^2} \right] \]

\[ = E_n \left[ \frac{T_i}{q} \frac{(D_i - p_{i,n})Y_i(1)^2}{p_{i,n}^2} \right] + E_n \left[ \frac{T_i}{q} \frac{Y_i(1)^2}{p_{i,n}^2} \right] \equiv A_n + B_n + C_n \]

We check the conditions of Lemma 8.20.(4). Let \( \mathcal{F}_n = \sigma(W_{1:n}, \zeta, \pi_{n}^l, \pi_{n}^d, \tau^t) \). Then by Assumption 8.3 and 8.1, \( (g_{a,s})_{a,s} \in \mathcal{F}_n \) and \( \mathcal{F}_n \perp \tau^d \). Moreover, note that

\[ \sup_{n \geq 1} E \left[ \frac{Y_i(1)^4}{(p_{i,n} q)^2} \right] \leq \sup_{n \geq 1} E \left[ \frac{Y_i(1)^4}{p_{i,n} q^2} \right] \leq \delta^{-6} E[Y(1)^4] < \infty \]

Then Lemma 8.20.(4) shows that \( A_n, B_n = O_p(n^{-1/2}) \), applying first with \( S_{1:n} = T_{1:n} \), then with \( S_n = 1 \). For the third term, \( E[Y(1)^4] < \infty \) by assumption. Note that \( G(p) \equiv p^{-1} \) is bounded and Lipschitz on \( [\delta^{-1}, 1] \). Then Lemma 8.18 applied with \( H(W) = Y(1) \) and
Consider that for any $G_i$ we have shown that identical. This finishes the proof of the claim for $A_i$. Consider 

Matching Corrections - First note that $Y_i(d) = m_d(X_i) + \epsilon_i^d = F_{i,n}^d + u_{i,n}^d + \epsilon_i^d$. Then for $i \neq j$ and $d, d' \in \{0, 1\}$, we have 

$$Y_i(d)Y_j(d') = (F_{i,n}^d + u_{i,n}^d + \epsilon_i^d)(F_{j,n}^{d'} + u_{j,n}^{d'} + \epsilon_j^{d'}) = F_{i,n}^dF_{j,n}^{d'} + F_{i,n}^{d'}u_{i,n}^d + \epsilon_i^d$$

(8.6) 

$$+ F_{i,n}^d(u_{j,n}^{d'} + \epsilon_j^{d'}) + (u_{i,n}^d + \epsilon_i^d)(u_{j,n}^{d'} + \epsilon_j^{d'}) \equiv F_{i,n}^dF_{j,n}^{d'} + R_{i,j}^{d,d'}$$

(8.7) 

Define $g_{a,s}^1 = g_{a,s} \cap \{i : D_i = 1\}$. Using this decomposition, and noting $\{i : T_i = 1\} = \bigsqcup_{s} g_{a,s}$ by Assumption 8.3, we have 

$$\hat{\nu}_1 = E_n \left[ T_iD_iG(p_{i,n})Y_iY_{\gamma(i)} \right] = n^{-1} \sum_{a=1}^{\lceil L_n \rceil} \sum_{s=1}^n \sum_{i \in g_{a,s}^1} (qp_a)^{-1}G(p_a) Y_i(1)Y_{\gamma(i)}(1)$$

$$= n^{-1} \sum_{a=1}^{\lceil L_n \rceil} \sum_{s=1}^n \sum_{i \in g_{a,s}^1} (qp_a)^{-1}G(p_a)(F_{i,n}^1F_{\gamma(i),n} + R_{\gamma(i)}^{i,1}) \equiv A_n + B_n$$

Consider that for any $i, j \in [n]$ and $d \in \{0, 1\}$ 

$$F_{i,n}^dF_{j,n}^d = F_{i,n}^dF_{j,n}^d - 1/2((F_{i,n}^d)^2 + (F_{j,n}^d)^2) + 1/2((F_{i,n}^d)^2 + (F_{j,n}^d)^2)$$

$$= (1/2)((F_{i,n}^d)^2 + (F_{j,n}^d)^2 - (F_{i,n}^d - F_{j,n}^d)^2)$$

Then the first part of the numerator above becomes 

$$n^{-1} \sum_{a=1}^{\lceil L_n \rceil} \sum_{s=1}^n \sum_{i \in g_{a,s}^1} (qp_a)^{-1}G(p_a)F_{i,n}^1F_{\gamma(i),n}^1 = (2n)^{-1} \sum_{a=1}^{\lceil L_n \rceil} \sum_{s=1}^n \sum_{i \in g_{a,s}^1} (qp_a)^{-1}G(p_a)[(F_{i,n}^1)^2 + (F_{\gamma(i),n}^1)^2]$$

$$+ (2n)^{-1} \sum_{a=1}^{\lceil L_n \rceil} \sum_{s=1}^n \sum_{i \in g_{a,s}^1} (qp_a)^{-1}G(p_a)(F_{i,n}^1^2 - F_{\gamma(i),n}^1)^2 \equiv A_n^1 + A_n^2$$

Consider $A_n^1$. Since $p_{i,n} = p_{\gamma(i),n} = p_a$ for $i \in g_{a,s}$ by matching admissibility, this is 

$$A_n^1 = (2n)^{-1} \sum_{a=1}^{\lceil L_n \rceil} \sum_{s=1}^n \sum_{i \in g_{a,s}^1} (qp_{i,n})^{-1}G(p_{i,n})(F_{i,n}^1)^2$$

The equality follows since $\gamma$ is bijective and $D_i = D_{\gamma(i)}$, so the second term sums over
each treated unit exactly twice. Continuing, the expression above is

\[ n^{-1} \sum_{i=1}^{n} T_i(D_i - p_{i,n})(q p_{i,n})^{-1} G(p_{i,n})(F_{i,n}^1)^2 + n^{-1} \sum_{i=1}^{n} \frac{T_i}{q} G(p_{i,n})(F_{i,n}^1)^2 \]

\[ n^{-1} \sum_{i=1}^{n} T_i(D_i - p_{i,n})(q p_{i,n})^{-1} G(p_{i,n})(F_{i,n}^1)^2 + n^{-1} \sum_{i=1}^{n} \frac{T_i - q}{q} G(p_{i,n})(F_{i,n}^1)^2 \]

\[ + n^{-1} \sum_{i=1}^{n} G(p_{i,n})(F_{i,n}^1)^2 \]

As in the argument for \( \hat{W}_I \), let \( \mathcal{F}_n = \sigma(W_{1:n}, \zeta_n, \pi_n, \tau^d, \pi_n, \tau^d) \). Note that \( \mathcal{F}_n \perp \tau^d, (g_{a,s})_{a,s} \in \mathcal{F}_n \), and \( (q p_{i,n})^{-1} G(p_{i,n})(F_{i,n}^1)^2 \in \mathcal{F}_n \). Moreover, note that

\[ \sup_{n \geq 1} E \left[ \left( (q p_{i,n})^{-1} G(p_{i,n})(F_{i,n}^1)^2 \right)^2 \right] \leq \delta^{-4} \sup_{n \geq 1} E \left[ (F_{i,n}^1)^4 \right] = \delta^{-4} \sup_{n \geq 1} E \left[ (F_{i,n}^1)^4 | \zeta_n \right] < \infty \]

Then by Lemma 8.20.(4), the first term is \( O_p(n^{-1/2}) \). Similarly, the second term is \( O_p(n^{-1/2}) \) by Lemma 8.20.(4), applied with selection variables \( S_{1:n} = 1 \). For the third term, by boundedness we have \( |E_n[G(p_{i,n})(F_{i,n}^1)^2 - (F_{i,n}^1)^2]| \lesssim E_n[|F_{i,n}^1 - F_{i,n}^1||F_{i,n}^1 + F_{i,n}^1|] \).

Taking a conditional expectation

\[ E[E_n[|F_{i,n}^1 - F_{i,n}^1||F_{i,n}^1 + F_{i,n}^1| | \zeta]] \lesssim E[(F_{i,n}^1)^2 | \zeta]^{1/2} (E[(F_{i,n}^1)^2 | \zeta]^{1/2} + E[(F_{i,n}^1)^2 | \zeta]^{1/2}) \]

\[ \leq O_p(r_n^F)E[m_1(X)^2]^{1/2} = O_p(r_n^F) \]

The first inequality by Cauchy-Schwarz and triangle inequality. The second inequality is by assumption and properties of projection. Finally, note that by applying Lemma 8.18

\[ E_n[G(p_{i,n})(F_{i,n}^1)^2] = E[G(p_{i,n})(F_{i,n}^1)^2 | \zeta] + O_p(n^{-1/2}) + O_p(r_n^p) \]

The lemma applies since \( E[(F_{i,n}^1)^4 | \zeta] < \infty, \zeta \)-a.s. by assumption. Putting this all together, we have shown that \( A_{1n}^1 = E[G(p_{i,n})(F_{i,n}^1)^2 | \zeta] + O_p(n^{-1/2}) + O_p(r_n^p) + O_p(r_n^F) \).

Next, consider \( A_{2n}^1 \). By the Lipschitz assumption 7.5, on an almost sure set given by the definition, \( |A_{2n}^1| \) is bounded by

\[ (2n)^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \sum_{i \in \vartheta_{a,s}^1} (q p_{a})^{-1} |G(p_{a})| (F_{i,n}^1 - F_{i,n}^1)_{(\gamma(i),n)}^2 \lesssim n^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \sum_{i \in \vartheta_{a,s}^1} (\psi_i - \psi_{\gamma(i)})^2 \]

\[ \leq n^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} (\psi_i - \psi_{\gamma(i)})^2 = E_n[T_i(\psi_i - \psi_{\gamma(i)})^2] = O_p(r_n^\psi) \]

The final equality follows by matching admissibility as in Definition 5.1.
Next, we bound the residual terms above

\[ B_n = n^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \sum_{i \in g_{a,s}^1} (q_{pa})^{-1} G(p_a) R_{i,\gamma(i)}^{1,1} \]

\[ = n^{-1} \sum_{a=1}^{[L_n]} (q_{pa})^{-1} G(p_a) \sum_{s=1}^{n} \sum_{i,j=1}^{n} D_i(\mathbb{I}(i \in g_{a,s}) \mathbb{I}(j = \gamma(i)) R_{i,j}^{1,1} \]

By definition the matching can be written as a deterministic function

\[ \gamma = \gamma((g_{a,s})_{a,s}; \psi; 1:n, P_{1:n}, n, D_{1:n}, T_{1:n}) = \gamma(\psi; 1:n, P_{1:n}, \pi^t_n, \pi^d_n, \tau^t, \tau^d, \zeta) \]

By assumption 8.3 and Lemma 8.19.(i). Let \( F_{(\psi, p_n, \tau^d), n} = \sigma(\psi; 1:n, P_{1:n}, \pi^t_n, \pi^d_n, \tau^t, \tau^d, \zeta) \). Again by Lemma 8.19.(i), \( D_i \in F_{(\psi, p_n, \tau^d), n} \) for all \( i \), and clearly \( F_{d_n} \in F_{(\psi, p_n, \tau^d), n} \).

In the sequel, we will make use of the following conditional moments. Momentarily suppressing the 1-superscript notation, applying Lemma 8.22 with \( F_n = F_{(\psi, p_n, \tau^d), n} \), \( \xi_n = (\pi^t_n, \pi^d_n, \tau^t, \tau^d) \), and \( \zeta_n = \zeta \) gives

\[ 0 = E[u_{i,n} F_{(\psi, p_n, \tau^d), n}] = E[\xi F_{(\psi, p_n, \tau^d), n}] = E[u_{i,n} u_{j,n} F_{(\psi, p_n, \tau^d), n}] = E[u_{i,n} u_{j,n} F_{(\psi, p_n, \tau^d), n}] \]

By the moments above, \( E[R_{i,j}^{1,1} F_{(\psi, p_n, \tau^d), n}] = 0 \) for \( 1 \leq i \neq j \leq n \). Since \( D_i(\mathbb{I}(i \in g_{a,2s}) \mathbb{I}(j = \gamma(i)) \in F_{(\psi, p_n, \tau^d), n} \) and \( \mathbb{I}(j = \gamma(i)) \) for \( j = i \) by definition, we have shown that \( E[B_n F_{(\psi, p_n, \tau^d), n}] = 0 \). Next, consider the conditional variance \( E[B_n^2 F_{(\psi, p_n, \tau^d), n}] \), given by

\[ E[B_n^2 F_{(\psi, p_n, \tau^d), n}] = n^{-2} \sum_{a,d=1}^{[L_n]} G(p_{ad}) C_{i,j,l,f}^{a,a',s,s'} \sum_{i,j=1}^{n} \sum_{i,j=1}^{n} C_{i,j,l,f}^{a,a',s,s'} E[R_{i,j}^{1,1} R_{i,j}^{1,1} F_{(\psi, p_n, \tau^d), n}] \]

Case 1: \( f \in \{i,j\} \) or \( l \in \{i,j\} \) exclusive. By definition of our matching, e.g. for the case \( \{f = i, l \neq j\} \) we have \( \mathbb{I}(j = \gamma(i)) \mathbb{I}(i = \gamma(l)) = 0 \), since each unit is matched uniquely. The other subcases follow by permuting labels.

Case 2: \( \{i,j\} \cap \{f,l\} = \emptyset \). Then we have

\[ R_{i,j}^{1,1} R_{i,f}^{1,1} = \sum_{i,j} u_{i,n} (u_{i,n} + \epsilon_i) [F_{i,j,n}(u_{i,n} + \epsilon_i) + F_{i,j,n}(u_{j,n} + \epsilon_j) + (u_{i,n} + \epsilon_i)(u_{j,n} + \epsilon_j)] \]

Let \( i, j, l, f \) jointly distinct indices. Then since \( F_{i,n} \in F_{(\psi, p_n, \tau^d), n} \) for all \( i \), we see that in
this case \( E[R_{i,j}^{1,1} R_{i,j}^{1,1} | \mathcal{F}(p_{\nu}, r^d), n] = 0 \) by the moments above.

Case 3: \( \{i, j\} = \{f, l\} \). Note that if \( \mathbb{1}(i \in g_{a,s}) \mathbb{1}(j = \gamma(i)) = 1 \) then \( \mathbb{1}(j \in g_{a,s}) \mathbb{1}(i = \gamma(j)) = 0 \), since \( j \) can’t both be in an even and odd group, so \( C_{i,j,i,j}^{a,a', s,s'} = 0 \) for all \( i, j \).

Then we only need consider terms of the form

\[
\begin{align*}
(R_{i,j}^{1,1})^2 &= (F_{j,n}^{1})^2(u_{i,n} + \epsilon_i^j)^2 + (F_{i,n}^{1})^2(u_{j,n} + \epsilon_j^i)^2 + (u_{i,n} + \epsilon_i^j)(u_{j,n} + \epsilon_j^i) \\
&+ F_{j,n}^{1} F_{i,n}^{1} (u_{i,n} + \epsilon_i^j)(u_{j,n} + \epsilon_j^i) + F_{j,n}^{1} (u_{i,n} + \epsilon_i^j)(u_{j,n} + \epsilon_j^i) \\
&+ F_{i,n}^{1} (u_{j,n} + \epsilon_j^i)(u_{i,n} + \epsilon_i^j) 
\end{align*}
\]

In addition to the moments above, note also that for any \( i \neq j \)

\[
0 = E[u_{i,n} \epsilon_i | \mathcal{F}(p_{\nu}, r^d), n] = E[u_{i,n} \epsilon_i u_{j,n} \epsilon_j | \mathcal{F}(p_{\nu}, r^d), n] = E[(u_{i,n} + \epsilon_i^j)(u_{j,n} + \epsilon_j^i)^2 | \mathcal{F}(p_{\nu}, r^d), n]
\]

again by Lemma 8.22. Then all of the cross terms are conditionally mean zero. This leaves only diagonal terms of the form \( E[(R_{i,i}^{1,1})^2 | \mathcal{F}(p_{\nu}, r^d), n] \), which can be written

\[
\begin{align*}
&= E[(F_{j,n}^{1})^2((u_{i,n}^2) + (\epsilon_i^j)^2) + (F_{j,n}^{1})^2((u_{j,n}^2) + (\epsilon_j^i)^2)] \\
&+ E[(u_{i,n}^2u_{j,n}^2) + (u_{i,n}^2\epsilon_i^j)^2 + (u_{j,n}^2\epsilon_j^i)^2] + (\epsilon_i^j\epsilon_j^i)^2 | \mathcal{F}(p_{\nu}, r^d), n]
\end{align*}
\]

The second equality uses Lemma 8.22, and the final inequality is by assumption. Also, note that \( \mathbb{1}(i \in g_{a,s}) \mathbb{1}(i \in g_{a', s'}) = 0 \) for \( (a, s) \neq (a', s') \) by uniqueness of group membership. This leaves only the diagonal terms, so that \( E[B_n^2 | \mathcal{F}(p_{\nu}, r^d), n] \) is

\[
n^{-2} \sum_{a=1}^{|L_a|} \frac{G(p_a)^2}{q^2 p_a^2} \sum_{s=1}^n \sum_{i \neq j=1}^n D_i \mathbb{1}(i \in g_{a,s}) \mathbb{1}(j = \gamma(i)) E[(R_{i,j}^{1,1})^2 | \mathcal{F}(p_{\nu}, r^d), n]
\]

\[
\lesssim n^{-2} \sum_{a=1}^{|L_a|} \sum_{s=1}^n \sum_{i \neq j=1}^n D_i \mathbb{1}(i \in g_{a,s}) \mathbb{1}(j = \gamma(i)) \left(2\sigma^2[(F_{j,n}^{1})^2 + (F_{i,n}^{1})^2 + 4(\sigma^2)^2] \right)
\]

\[
= n^{-2} \sum_{a=1}^{|L_a|} \sum_{s=1}^n \sum_{i \neq j=1}^n \mathbb{1}(i \in g_{a,s}) \mathbb{1}(j = \gamma(i)) \left(2\sigma^2[D_j(F_{j,n}^{1})^2 + D_i(F_{i,n}^{1})^2 + 4D_i(\sigma^2)^2] \right)
\]

The last equality since \( D_i \mathbb{1}(j = \gamma(i)) = D_j \). Finally, note that this is

\[
n^{-2} \sum_{a=1}^{|L_a|} \sum_{s=1}^n \sum_{i \in g_{a,s}} \left[2\sigma^2[D_j(\gamma(i))F_{j,n}^{1})^2 + D_i(F_{i,n}^{1})^2 + 4D_i(\sigma^2)^2] \right]
\]

\[
\lesssim n^{-2} \sum_{i=1}^n T_i D_i(F_{i,n}^{1})^2 + n^{-2} \cdot (gn) = O_p(n^{-1})
\]

The final equality follows since \( E_n[T_i D_i(F_{i,n}^{1})^2] = O_p(1) \) by the calculations above for the term \( A_n^1 \). Then by conditional Chebyshev (Lemma 8.16), \( B_n = O_p(n^{-1/2}) \). Putting this
all together, we have shown that
\[ \hat{\nu}_1 = A_n^1 + A_n^2 + B_n = E[G(p_i)(F_i^n)^2] + O_p(n^{-1/2}) + O_p(r_n^p) + O_p(r_n^F) + O_p(r_n^\psi) \]
By symmetry of our assumptions with respect to permutation of the labels \( d \in \{0, 1\} \), the same result holds for the estimator \( \hat{\nu}_0 \).

To conclude, we analyze the cross-moment estimator, expanding using Equation 8.6
\[
\hat{\nu}_{10} = 2n^{-1} \sum_{1 \leq i < j \leq n} \frac{T_i T_j 1(g(i) = g(j))}{q} \frac{D_i(1 - D_j)Y_i Y_j}{\text{Card}(g)} \frac{p_n(X_i)(1 - p_n)(X_j)}{p_n(X_i) + 1 - p_n(X_j)}
\]
\[
= (qn)^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^n \frac{k_a^2}{k_a - 1} \frac{k_a(k_a - 1)}{q_a(k_a - q_a)} \sum_{i,j \in g_{a,s}, j \in g_{a,s}} Y_i(1)Y_j(0)
\]
\[
= (qn)^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^n \frac{k_a}{k_a - 1} \frac{k_a(k_a - 1)}{q_a(k_a - q_a)} \sum_{i,j \in g_{a,s}, j \in g_{a,s}} (F_{i,n} F_{j,n} + F_{i,j}^1) \equiv S_n + T_n
\]
First consider \( S_n \). Let \( F(x,y),(n) = \sigma(\psi_1,n, p_1,n, \pi_1,n, \pi_2,n, \tau_1, \zeta) \), then we have \((g_{a,s})_a,s, F_{i,n}^d \in F(x,y),(n)\) and \( F(x,y),(n) \perp \tau_d \), so by Lemma 8.19
\[
E[S_n|F(x,y),(n)] = (qn)^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^n \frac{k_a}{k_a - 1} \frac{k_a(k_a - 1)}{q_a(k_a - q_a)} \sum_{i,j \in g_{a,s}} E[D_i(1 - D_j)|F(x,y),(n)]F_{i,n} F_{j,n}
\]
\[
= (qn)^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^n \frac{k_a}{k_a - 1} \frac{k_a(k_a - 1)}{q_a(k_a - q_a)} \sum_{i,j \in g_{a,s}} \frac{q_a(k_a - q_a)}{k_a(k_a - 1)} F_{i,n} F_{j,n}
\]
\[
= (qn)^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^n \frac{k_a}{k_a - 1} \sum_{i \in g_{a,s}} \sum_{j \in g_{a,s}} F_{i,n} F_{j,n} + F_{i,n}^1 F_{j,n}
\]
\[
= (qn)^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^n \frac{k_a}{k_a - 1} \sum_{i \in g_{a,s}} \sum_{j \in g_{a,s}} F_{i,n} F_{j,n} + F_{i,n}^1 F_{j,n} + (F_{i,n} - F_{i,n}^1)(F_{j,n}^1 - F_{i,n}^0)
\]
\[
\equiv S_n^1 + S_n^2
\]
First consider \( S_n^1 \). Note that the final sum is
\[
\sum_{i,j \in g_{a,s}} F_{i,n} F_{j,n}^0 + \sum_{i,j \in g_{a,s}} F_{i,n}^1 F_{j,n}^0 = \sum_{i=1}^n F_{i,n} F_{i,n}^0 1(i \in g_{a,s}) \sum_{j \in g_{a,s}} 1(j \in g_{a,s})
\]
\[
+ \sum_{j=1}^n F_{j,n} F_{j,n}^0 1(j \in g_{a,s}) \sum_{i \in g_{a,s}} 1(i \in g_{a,s}) = \sum_{i=1}^n F_{i,n} F_{i,n}^0 1(i \in g_{a,s})(k_a - 1)
\]
81
Then we have $S^1_n = (qn)^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^n \sum_{i \in g_{a,s}} F^1_{i,n} F^0_{i,n} = q^{-1} E_n[T_i F^1_{i,n} F^0_{i,n}]$. Next we write

$$E_n[T_i F^1_{i,n} F^0_{i,n}] = E_n[T_i (F^1_{i,n} - F^1_i + F^1_i) (F^0_{i,n} - F^0_i + F^0_i)]$$

$$= E_n[T_i F^1_{i,n} F^0_i] + E_n[T_i (F^1_{i,n} - F^1_i) (F^0_{i,n} - F^0_i)] + E_n[T_i (F^1_i - F^1_i) F^0_{i,n}] + E_n[T_i F^1_{i,n} (F^0_{i,n} - F^0_i)]$$

The first term has

$$E[[E_n[T_i (F^1_{i,n} - F^1_i) (F^0_{i,n} - F^0_i)]]|\xi] \leq E[[F^1_{i,n} - F^1_i]\|F^0_{i,n} - F^0_i]|\xi]$$

$$\leq E[[F^1_{i,n} - F^1_i]^2|\xi]^{1/2} E[[F^0_{i,n} - F^0_i]^2|\xi]^{1/2} = O_p((r^F_n)^2)$$

Also $E[[E_n[T_i (F^1_{i,n} - F^1_i) F^0_i)]|\xi] \leq E[[F^1_{i,n} - F^1_i]^2|\xi]^{1/2} E[[F^0_i|^2|\xi]^{1/2} = O_p(r^F_n)$. Then by conditional Markov, the last three terms above are $O_p(r^F_n)$. Finally, note that

$$E_n[T_i F^1_i F^0_i] = E_n[(T_i - q) F^1_i F^0_i] + q E_n[F^1_i F^0_i] = q E[F^1_i F^0_i|\xi] + O_p(n^{-1/2})$$

For the final equality, note that $E_n[(T_i - q) F^1_i F^0_i]$ by Lemma 8.20 since by assumption

$$E[(F^1_i)^2(F^0_i)^2|\xi] \leq E[(F^1_i)^4|\xi] + E[(F^0_i)^4|\xi] < \infty$$

The conclusion then follows conditional Markov manipulations applied to $q E_n[F^1_i F^0_i]$. We have shown that $S^1_n = E[F^1_i F^0_i|\xi] + O_p(n^{-1/2}) + O_p(r^F_n)$. Next, consider the differences in $S^2_n$ above. By the Lipschitz assumption, on an almost sure set given by definition 7.5 we have

$$|S^2_n| \leq (qn)^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^n \frac{1}{k_a - 1} \sum_{i,j \in g_{a,s}} |F^1_{i,n} - F^1_{j,n}||F^0_{j,n} - F^0_{i,n}|$$

$$\leq n^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^n \frac{1}{k_a - 1} \sum_{i,j \in g_{a,s}, i < j} (\psi_i - \psi_j)^2 \leq n^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^n \frac{1}{k_a - 1} \sum_{i,j \in g_{a,s}, i \neq j} (\psi_i - \psi_j)^2 = O_p(r^\psi_n)$$

Putting this all together, we have shown that

$$E[S_n|F(\psi,p_n),n] = E[F^1_i F^0_i|\xi] + O_p(n^{-1/2}) + O_p(r^F_n) + O_p(r^\psi_n)$$

Next, we will show the coupling $S_n - E[S_n|F(\psi,p_n),n] = O_p((k_n/n)^{1/2})$. As previously, define the higher-order inverse propensity weight $w_{a,n} \equiv E[D_i(1 - D_j)|F(\psi,p_n),n]^{-1}1(i,j \in g_{a,s}) = k_a(k_a - 1) q_a(k_a - q_a)$. By work above, we have

$$S_n - E[S_n|F(\psi,p_n),n] = (qn)^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^n \frac{1}{k_a - 1} \sum_{i,j \in g_{a,s}, i \neq j} F^1_{i,n} F^0_{j,n} (D_i(1 - D_j)w_{a,n} - 1)$$

82
Therefore, Var$(S_n|\mathcal{F}_{(\psi,pn),n})$ is given by the conditional expectation w.r.t. $\mathcal{F}_{(\psi,pn),n}$ of

$$\sum_{a,a'=1}^{L_n} \sum_{s,s'=1}^{n} \frac{(qn)^{-2}}{(k_a - 1)^2} \sum_{i,j \in g_{a,s}, l,f \in g_{a',s'}} F^1_{i,n} F^1_{l,n} F^0_{f,n} F^0_{j,n} (D_i(1 - D_j)w_{a,n} - 1)(D_l(1 - D_f)w_{a',n} - 1)$$

By the weight definition $E[D_i(1 - D_j)w_{a,n} I(i, j \in g_{a,s})|\mathcal{F}_{(\psi,pn),n}] = 1$ for all $i \neq j$. Then taking a conditional expectation, the above simplifies to

$$\sum_{a=1}^{L_n} \sum_{s=1}^{n} \frac{(qn)^{-2}}{(k_a - 1)^2} \sum_{i,j \in g_{a,s}, l \neq f} F^1_{i,n} F^1_{l,n} F^0_{f,n} F^0_{j,n} \Gamma_{i,j,l,f}^{a,n}$$

$$\Gamma_{i,j,l,f}^{a,n} \equiv E[D_iD_l(1 - D_j)(1 - D_f)|\mathcal{F}_{(\psi,pn),n}]w_{a,n}^2 - 1$$

Case 1: $\{i, j\} \cap \{l, f\} = \emptyset$. Note that in this case we have $k_a \geq 4$. By the representation in Lemma 8.19.(i), we have

$$|\Gamma_{i,j,l,f}^{a,n}| = \left| \frac{(k_a - 4)}{(k_a)} \left( \frac{1}{(k_a - 2)} \right) - 1 \right| \leq 5$$

Using triangle inequality, the corresponding terms of the sum are bounded by

$$\lesssim \sum_{a=1}^{L_n} \sum_{s=1}^{n} \frac{(qn)^{-2}}{(k_a - 1)^2} \sum_{i,j,l,f \in g_{a,s}} \left| F^1_{i,n} F^1_{l,n} F^0_{f,n} F^0_{j,n} \Gamma_{i,j,l,f}^{a,n} \right|$$

$$\lesssim \sum_{a=1}^{L_n} \sum_{s=1}^{n} \frac{5(qn)^{-2}}{(k_a - 1)^2} \sum_{i,j \in g_{a,s}, l \neq f} \left( \sum_{i} \left| F^1_{i,n} \right|^4 + \sum_{i} \left| F^0_{f,n} \right|^4 \right)$$

$$\lesssim 7\tilde{K}_n(qn)^{-1} \left( E_n[F^1_{i,n}] + E_n[F^0_{f,n}] \right)$$

$$\lesssim \tilde{K}_n(qn)^{-1} \left( E_n[F^1_{i,n}]^4 + E_n[F^0_{f,n}]^4 \right) = O_p(\tilde{K}_n^{-1})$$

The first inequality is by Lemma 8.17, and the second noting $\max_{x \geq 4} \frac{x^2}{x^2 - 1} \leq 4/3$ and $k_a \leq \tilde{K}_n$. The third inequality by assumption 8.3, and the final equality follows by conditional Markov, since $E[|F_{i,n}|^4|\xi] < \infty$ by assumption.

Case 2: $i \in \{l, f\}$ or $j \in \{l, f\}$. Then we must have $k_a \geq 3$.

Subcase (a): $i = l$, then we have

$$|\Gamma_{i,j,i,f}^{a,n}| = \left| E[D_i(1 - D_j)(1 - D_f)|\mathcal{F}_{(\psi,pn),n}]w_{a,n}^2 - 1 \right| = \left| \frac{(k_a - 3)}{(k_a)} \left( \frac{1}{(k_a - 2)} \right) - 1 \right|$$

Using algebra, one can show that $|\Gamma_{i,j,i,f}^{a,n}| \leq 2\frac{k_a}{k_a} + 1 \leq 2\delta^{-1} + 1$. Again applying triangle
inequality, these terms in the sum are bounded by

$$\sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \frac{(qn)^{-2}}{(k_a - 1)^2} \sum_{i,j,f \in \mathcal{g}_{s,a}, i,j,f \text{ distinct}} |F_{i,n}^1||F_{f,n}^0||F_{j,n}^0||\Gamma_{i,j,i,f}^a|$$

$$\leq \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \frac{(2\delta^{-1} + 1)(qn)^{-2}}{(k_a - 1)^2} \sum_{i,j,f \in \mathcal{g}_{s,a}, i,j,f \text{ distinct}} |F_{i,n}^1|^2|F_{f,n}^0|F_{j,n}^0|$$

$$\leq \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \frac{(2\delta^{-1} + 1)(qn)^{-2}}{(k_a - 1)^2} \left( \sum_{i \in \mathcal{g}_{s,a}} |F_{i,n}^1|^4 + \sum_{i \in \mathcal{g}_{s,a}} |F_{i,n}^0|^4 \right)$$

$$\lesssim n^{-1}(E_n[|F_{i,n}^1|^4] + E_n[|F_{i,n}^0|^4]) = O_p(n^{-1})$$

Subcase (b): $j = f$, then we have

$$|\Gamma_{i,j,i,j}^a| = |E[D_i D_l (1 - D_j) | F_{(\psi,p,n)}]|w_{a,n}^2 - 1| = \left| (k_a - 3) \left( \frac{k_a}{q_a - 2} \right)^{-1} w_{a,n}^2 - 1 \right|$$

$$\leq \frac{(q_a - 1)k_a}{(k_a - 2)q_a} + 1 \leq 2 \frac{k_a}{k_a - q_a} + 1 \leq 2 \delta^{-1} + 1$$

Then the relevant terms in the sum are bounded by

$$\sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \frac{(2\delta^{-1} + 1)(qn)^{-2}}{(k_a - 1)^2} \sum_{i,j \in \mathcal{g}_{s,a}, i,j \text{ distinct}} |F_{i,n}^1||F_{i,n}^1||F_{f,n}^0|^2 = O_p(n^{-1})$$

by the same argument as the previous subcase, permuting labels \{0, 1\}.

Subcase (c): $i = f$, then $\Gamma_{i,i,i,f}^a = E[D_i D_l (1 - D_j) (1 - D_i) | F_{(\psi,p,n)}]w_{a,n}^2 - 1 = -1$. Then the relevant terms in the sum are bounded by

$$\sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \frac{(qn)^{-2}}{(k_a - 1)^2} \sum_{i,j,f \in \mathcal{g}_{s,a}, i,j,f \text{ distinct}} |F_{i,n}^1||F_{i,n}^0||F_{j,n}^0|^2$$

$$\leq \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \frac{(qn)^{-2}}{(k_a - 1)^2} \left( k_a^3 + k_a \right) \sum_{i \in \mathcal{g}_{s,a}} |F_{i,n}^1|^4 + |F_{i,n}^0|^4$$

$$\lesssim \overline{k}_n n^{-1}(E_n[|F_{i,n}^1|^4] + E_n[|F_{i,n}^0|^4]) = O_p(\overline{k}_n n^{-1})$$

The first inequality is by Lemma 8.17, and the second uses $k_a \leq \overline{k}_n$ and $\max_{k \geq 3} \frac{k(k+1)}{(k-1)^2} = 3$.

Subcase (d): $j = l$, then $\Gamma_{i,j,i,l}^a = E[D_i D_l (1 - D_j)(1 - D_l) | F_{(\psi,p,n)}]w_{a,n}^2 - 1 = -1$. Then the relevant terms in the sum are bounded by

$$\sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \frac{(qn)^{-2}}{(k_a - 1)^2} \sum_{i,j,f \in \mathcal{g}_{s,a}, i,j,f \text{ distinct}} |F_{i,n}^1||F_{j,n}^0||F_{j,n}^0||F_{f,n}^0| = O_p(\overline{k}_n n^{-1})$$

84
since this is the same expression as the previous case.

Case 3: \( \{i, j\} = \{l, f\} \).
Subcase (a): \( (i, j) = (l, f) \), then we have
\[
|\Gamma_{i,j,l,f}^{a,n}| = |E[D_i(1 - D_j)]|F_{(\psi, p_n), n}|w_{a,n}^2 - 1| = |w_{a,n}^{-1}w_{a,n}^2 - 1| \leq w_{a,n} + 1 \leq \delta^{-2} + 1
\]
Then the relevant terms in the sum are bounded by
\[
\sum_{a=1}^{L_n} \sum_{s=1}^{n} \frac{(q_n)^{-2}(\delta^{-2} + 1)}{(k_a - 1)^2} \sum_{i,j \in g_{a,s}} |F_{i,n}^1|^2 |F_{j,n}^0|^2 \\
\leq \sum_{a=1}^{L_n} \sum_{s=1}^{n} \frac{(q_n)^{-2}(\delta^{-2} + 1)k_a}{(k_a - 1)^2} \sum_{i \in g_{a,s}} |F_{i,n}^1|^4 + |F_{j,n}^0|^4 \\
\lesssim n^{-1}(E_n[|F_{i,n}^1|^4] + E_n[|F_{j,n}^0|^4]) = O_p(n^{-1})
\]
The first inequality is by Lemma 8.17, then using \( \max_{k \geq 2} \frac{k}{(k-1)^2} = 2 \).
Subcase (b): \( (i, j) = (l, f) \), then \( \Gamma_{i,j,l,f}^{a,n} = E[D_iD_j(1 - D_j)(1 - D_i)]|F_{(\psi, p_n), n}|w_{a,n}^2 - 1 = -1 \).
Then the relevant terms are bounded by
\[
\sum_{a=1}^{L_n} \sum_{s=1}^{n} \frac{(q_n)^{-2}}{(k_a - 1)^2} \sum_{i,j \in g_{a,s}} |F_{i,n}^1||F_{i,n}^0||F_{j,n}^1||F_{j,n}^0| \\
\leq \sum_{a=1}^{L_n} \sum_{s=1}^{n} \frac{n^{-2}k_a}{(k_a - 1)^2} \sum_{i \in g_{a,s}} |F_{i,n}^1|^4 + |F_{j,n}^0|^4 \\
\leq 2n^{-1}(E_n[|F_{i,n}^1|^4] + E_n[|F_{j,n}^0|^4]) = O_p(n^{-1})
\]
By Lemma 8.17 again. From the work above and triangle inequality, we see that
\[
\text{Var}(S_n|F_{(\psi, p_n), n}) \leq \sum_{a=1}^{L_n} \sum_{s=1}^{n} \frac{n^{-2}}{(k_a - 1)^2} \sum_{i,j \in g_{a,s}, l,f \in g_{a,s}} |F_{i,n}^1|F_{l,n}^0F_{j,n}^0\Gamma_{i,j,l,f}^{a,n}| \\
= O_p(n^{-1}) + O_p(\bar{k}_n n^{-1})
\]
Then by Lemma 8.16, we have the coupling \( S_n - E[S_n|F_{(\psi, p_n), n}] = O_p((\bar{k}_n/n)^{1/2}) \).
Together with the result above, we have shown that
\[
S_n = E[F_i^1F_0^0|\zeta] + O_p(n^{-1/2}) + O_p(r_n^\psi) + O_p(r_n^\zeta) + O_p((\bar{k}_n/n)^{1/2})
\]
Next, we consider the residual terms \( T_n \).
\[
T_n = (q_n)^{-1} \sum_{a=1}^{L_n} \frac{w_{a,n}}{k_a} \sum_{s=1}^{n} \sum_{i \neq j=1} D_i(1 - D_j)1(i \in g_{a,s})1(j \in g_{a,s})R_{i,j}^{1,0} \\
R_{i,j}^{1,0} = F_{j,n}^0(u_{i,n}^0 + \epsilon_i^0) + F_{i,n}^1(u_{j,n}^0 + \epsilon_j^0) + (u_{i,n}^1 + \epsilon_i^1)(u_{j,n}^0 + \epsilon_j^0)
\]
Again, let \( F_{(\psi, p_n, \tau_n^d), n} = \sigma(\psi_{1:n}, p_{1:n}, \tau_n^d, \tau_n^d, \zeta) \). By the residual moments analyzed above, we have \( E[R_{i,j}^{1,0}|F_{(\psi, p_n, \tau_n^d), n}] = 0 \) for \( i \neq j \). Since \( D_i(1 - D_j)1(i \in g_{a,s})1(j \in g_{a,s}) \in F_{(\psi, p_n, \tau_n^d), n} \), we have \( E[T_n|F_{(\psi, p_n, \tau_n^d), n}] = 0 \). Next, we consider the conditional variance
$E[T_n^2|F_{(\psi,p_{\tau},\tau^{\prime}),n}]$. This can be written as

$$(gn)^{-2} \sum_{a,a'=1}^{\lfloor L_n \rfloor} \frac{\sum_{a'=1}^{\lfloor L_n \rfloor} w_{a,n} w_{a',n}}{(k_{a} - 1)(k_{a'} - 1)} \sum_{n} \sum_{n} \sum_{n} k_{a,i,j,f}^{a',s,s'} E[R_{i,j}^{0} R_{i,f}^{0} | F_{(\psi,p_{\tau},\tau^{\prime}),n}]$$

$R_{i,j,f}^{0} \equiv D_i D_j (1 - D_j)(1 - D_f) 1(i \in g_{a,s}) 1(j \in g_{a,s}) 1(l \in g_{a',s'}) 1(f \in g_{a',s'})$

Case 1: $\{i,j\} \cap \{l,f\} = \emptyset$. Then we have $E[R_{i,j}^{0} R_{i,f}^{0} | F_{(\psi,p_{\tau},\tau^{\prime}),n}] = 0$ by the work in our analysis of $\hat{\psi}_1$, noting that we did not rely on the superscript values. If $(a,s) \neq (a',s')$, then $\{i,j\} \cap \{l,f\} = \emptyset$ on the event $\{i,j \in g_{a,s}\} \cap \{l,f \in g_{a',s'}\}$, so all terms with $(a,s) \neq (a',s')$ are 0.

Case 2: $i \in \{l,f\}$ or $j \in \{l,f\}$, exclusive. If $i = f$ then $D_i D_j (1 - D_j)(1 - D_f) = D_i D_j (1 - D_j)(1 - D_f) = 0$ and similarly if $j = l$. This leaves the subcases $i = l$ or $j = f$, exclusive. Suppose $i = l$, then

$R_{i,j}^{0} R_{i,f}^{0} = F_{i,j}^{0} (u_{i,n} + \epsilon_i) [F_{i,j}^{0} (u_{i,n} + \epsilon_i) + F_{i,n}^{1} (u_{j,n} + \epsilon_j) + (u_{i,n} + \epsilon_i) (u_{j,n} + \epsilon_j)]$

Again applying Lemma 8.22, only the first cross moments remains, leaving

$|E[R_{i,j}^{0} R_{i,f}^{0} | F_{(\psi,p_{\tau},\tau^{\prime}),n}]| = |E[F_{i,j}^{0} F_{i,n}^{0} (u_{i,n} + \epsilon_i)]^2 F_{(\psi,p_{\tau},\tau^{\prime}),n}]|$

$|E[F_{i,j}^{0} F_{i,n}^{0} (\sigma^{2}_{a^1,\tau}(X_i) + E[\sigma^{2}_{a^1,\tau}(X_i)]|\psi, p_{n})| \leq 2|F_{i,j}^{0} F_{i,n}^{0} \sigma^2|$}

Suppose $j = f$. Then reasoning similarly the only non-zero cross-moment is

$|E[R_{i,j}^{0} R_{i,f}^{0} | F_{(\psi,p_{\tau},\tau^{\prime}),n}]| = |E[F_{i,n}^{1} F_{i,n}^{0} (u_{j,n} + \epsilon_j)]^2 | F_{(\psi,p_{\tau},\tau^{\prime}),n}]| \leq 2|F_{i,n}^{1} F_{i,n}^{0} \sigma^2|$}

Case 3: $\{i,j\} = \{l,f\}$. Note that $D_i D_j (1 - D_i)(1 - D_j) = 0$, so we need only consider the case where $i = l$ and $j = f$. We have

$(R_{i,j}^{0})^2 = (F_{i,j}^{0} (u_{i,n} + \epsilon_i) + F_{i,n}^{1} (u_{j,n} + \epsilon_j) + (u_{i,n} + \epsilon_i) (u_{j,n} + \epsilon_j))^2$

As in our analysis of $\hat{\psi}_1$, using Lemma 8.22, all the cross terms are conditionally mean zero. Also

$E[u_{i,n} \epsilon_i ] \cdot F_{(\psi,p_{\tau},\tau^{\prime}),n}] = E[u_{i,n} \epsilon_i (u_{j,n} + \epsilon_j)]^2 | F_{(\psi,p_{\tau},\tau^{\prime}),n}] = 0$

Using these observations and application of Lemma 8.22 to compute higher order moments as in the previous analysis, $E[(R_{i,j}^{0})^2 | F_{(\psi,p_{\tau},\tau^{\prime}),n}]$ simplifies as

$(F_{i,j}^{0})^2 (\sigma^{2}_{a^1,\tau}(X_i) + E[\sigma^{2}_{a^1,\tau}(X_i)]|\psi, p_{n}) + (F_{i,n}^{1})^2 (\sigma^{2}_{a^2,\tau,n}(X_j) + E[\sigma^{2}_{a^2,\tau,n}(X_j)]|\psi, p_{n})$

$+ \sigma^{2}_{a^1,\tau,n}(X_i) \sigma^{2}_{a^2,\tau,n}(X_j) + E[\sigma^{2}_{a^1,\tau,n}(X_i)|\psi, p_{n}) + \sigma^{2}_{a^2,\tau,n}(X_j)] + E[\sigma^{2}_{a^1,\tau,n}(X_i)|\psi, p_{n}) \sigma^{2}_{a^2,\tau,n}(X_j)]$

$+ E[\sigma^{2}_{a^1,\tau,n}(X_i)|\psi, p_{n}) \sigma^{2}_{a^2,\tau,n}(X_j)] \leq 2(F_{i,j}^{0})^2 \sigma^2 + 2(F_{i,n}^{1})^2 \sigma^2 + 4(\sigma^2)^2$
From the discussion above, we can simplify

\[
E[|T_n^2|;\mathcal{F}_{(\psi,p_n,\tau^d)}] = (qn)^{-2}\sum_{a=1}^{[L_n]} w_{a,n} \sum_{s=1}^{n} \sum_{i \neq j=1}^{n} \kappa_{a,s,s}^{i,j} E[(R_{i,j}^{1,0})^2;\mathcal{F}_{(\psi,p_n,\tau^d)}] \\
+ (qn)^{-2}\sum_{a=1}^{[L_n]} w_{a,n} \sum_{s=1}^{n} \sum_{i \neq j=1}^{n} \kappa_{a,s,s}^{i,j} E[R_{i,j}^{1,0} R_{i,j}^{1,0};\mathcal{F}_{(\psi,p_n,\tau^d)}] \\
+ (qn)^{-2}\sum_{a=1}^{[L_n]} w_{a,n} \sum_{s=1}^{n} \sum_{i \neq j=1}^{n} \kappa_{a,s,s}^{i,j} E[R_{i,j}^{1,0} R_{i,j}^{1,0};\mathcal{F}_{(\psi,p_n,\tau^d)}] \\
= T_{n,1} + T_{n,2} + T_{n,3}
\]

Note the weight bound

\[
w_{a,n} = \frac{k_a^2}{q_a^2(k_a - q_a)^2} \leq \frac{k_a^4}{q_a^2(k_a - q_a)^2} = \frac{1}{p_a(1-p_a)^2} \leq \delta^{-4}
\]

Then we have

\[
T_{n,1} = (qn)^{-2}\sum_{a=1}^{[L_n]} w_{a,n} \sum_{s=1}^{n} \sum_{i \neq j=1}^{n} D_i (1 - D_j) \mathbb{I}(i,j \in g_{a,s}) E[(R_{i,j}^{1,0})^2;\mathcal{F}_{(\psi,p_n,\tau^d)}] \\
\leq \delta^{-4}(qn)^{-2}\sum_{a=1}^{[L_n]} \frac{1}{(k_a - 1)^2} \sum_{s=1}^{n} \sum_{i \neq j=1}^{n} D_i (1 - D_j) \mathbb{I}(i,j \in g_{a,s}) [2(F_{j,n}^0)^2 \sigma^2 + 2(F_{i,n}^1)^2 \sigma^2 + 4(\sigma^2)^2]
\]

The first term in the brackets is bounded by

\[
\leq n^{-2} \sum_{a=1}^{[L_n]} \frac{1}{(k_a - 1)^2} \sum_{s=1}^{n} \sum_{j=1}^{n} (1 - D_j) \mathbb{I}(j \in g_{a,s})(F_{j,n}^0)^2 \sum_{i=1}^{n} D_i \mathbb{I}(i \in g_{a,s})
\]

\[
\leq n^{-2} \sum_{a=1}^{[L_n]} \frac{1}{(k_a - 1)^2} \sum_{s=1}^{n} \sum_{j=1}^{n} (1 - D_j) \mathbb{I}(j \in g_{a,s})(F_{j,n}^0)^2 q_a \mathbb{I}(g_{a,s} \neq \emptyset)
\]

\[
= n^{-2} \sum_{a=1}^{[L_n]} \frac{q_a}{(k_a - 1)^2} \sum_{s=1}^{n} \sum_{j=1}^{n} (1 - D_j) \mathbb{I}(j \in g_{a,s})(F_{j,n}^0)^2
\]

\[
\leq n^{-2} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \sum_{j=1}^{n} (1 - D_j) \mathbb{I}(j \in g_{a,s})(F_{j,n}^0)^2 = n^{-1} E_n[T_i(1 - D_j)(F_{j,n}^0)^2] = O_p(n^{-1})
\]

Identical reasoning can be used to show that

\[
n^{-2} \sum_{a=1}^{[L_n]} \frac{1}{(k_a - 1)^2} \sum_{s=1}^{n} \sum_{i \neq j=1}^{n} D_i (1 - D_j) \mathbb{I}(i,j \in g_{a,s}) 2(F_{i,n}^1)^2 \sigma^2 = O_p(n^{-1})
\]
Then the last term in the bracket gives

\[ 4\delta^{-4}n^{-2} \sum_{a=1}^{[L_n]} \sum_{i \neq j=1}^{n} \frac{D_i(1 - D_j)1(i, j \in g_{a,s})(\bar{\sigma}^2)^2}{(k_a - 1)^2} \]

\[ \leq n^{-2} \sum_{a=1}^{[L_n]} \frac{q_a(k_a - q_a)}{(k_a - 1)^2} \left[ \left( \sum_{s=1}^{n} 1(g_{a,s} \neq \emptyset) \right) + 1 \right] \]

\[ \leq n^{-2} \sum_{a=1}^{[L_n]} \left( \sum_{s=1}^{n} 1(g_{a,s} \neq \emptyset) \right) + 1 \]

\[ \leq 4\delta^{-4}(\bar{\sigma}^2)^2n^{-2}[n + |L_n|] \]

\[ = O_p(n^{-1}) + O_p([L_n]n^{-2}) \]

Summarizing, we have shown that \( \bar{T}_{n,1} = O_p(n^{-1}) \). Next, consider \( \bar{T}_{n,2} \). By triangle inequality

\[ |T_{n,2}| \leq (qn)^{-2} \sum_{a=1}^{[L_n]} \frac{w_{a,n}^2}{(k_a - 1)^2} \sum_{s=1}^{n} \sum_{i \neq j=1}^{n} \sum_{f \notin \{i,j\}}^{n} \kappa_{i,j,i,f}^{a,a,s,s} |E[R_{i,j}^{1,0} R_{i,f}^{1,0} | \mathcal{F}(\psi,p_a),n]| \]

\[ \leq n^{-2} \sum_{a=1}^{[L_n]} \frac{w_{a,n}^2}{(k_a - 1)^2} \sum_{s=1}^{n} \sum_{i \neq j=1}^{n} \sum_{f \notin \{i,j\}}^{n} D_i(1 - D_j)(1 - D_j)1(i, j, f \in g_{a,s})|F_{f,n}^0 F_{j,n}^0| \]

Then \( E[|\bar{T}_{n,2}|^2 | \mathcal{F}(\psi,p_a),n] \) is bounded above by

\[ \sum_{a=1}^{[L_n]} \frac{n^{-2}w_{a,n}^2}{(k_a - 1)^2} \sum_{s=1}^{n} \sum_{i \neq j=1}^{n} \sum_{f \notin \{i,j\}}^{n} E[D_i(1 - D_j)(1 - D_j)| \mathcal{F}(\psi,p_a),n]|1(i, j, f \in g_{a,s})|F_{f,n}^0 F_{j,n}^0| \]

\[ = \sum_{a=1}^{[L_n]} \frac{n^{-2}w_{a,n}^2}{(k_a - 1)^2} \sum_{s=1}^{n} \sum_{i \neq j=1}^{n} \sum_{f \notin \{i,j\}}^{n} \frac{q_a(k_a - q_a)(k_a - q_a - 1)}{k_a(k_a - 1)(k_a - 2)} 1(i, j, f \in g_{a,s})|F_{f,n}^0 F_{j,n}^0| \]

We can simplify the \((a, n)\) weight as

\[ \frac{w_{a,n}^2 q_a(k_a - q_a)(k_a - q_a - 1)}{k_a(k_a - 1)^3(k_a - 2)} = \frac{k_a^2(k_a - 1)^2(k_a - q_a)(k_a - q_a - 1)}{k_a(k_a - 1)^3(k_a - 2)q_a^2(k_a - q_a)^2} \]

\[ = \frac{k_a(k_a - q_a - 1)}{(k_a - 1)(k_a - 2)q_a^2(k_a - q_a)} \]
The inner summand is
\[
\sum_{s=1}^{n} \sum_{i \neq j=1}^{n} \sum_{f \notin \{i,j\}} 1(i, j, f \in g_{a,s}) |F_{f,n}^0, F_{j,n}^0| = \sum_{s=1}^{n} \sum_{j \neq f} 1(j, f \in g_{a,s}) |F_{f,n}^0, F_{j,n}^0| \sum_{i \notin \{j,f\}} 1(i \in g_{a,s})
\]
\[
= \sum_{s=1}^{n} (k_a - 2) \sum_{j \neq f} 1(j, f \in g_{a,s}) |F_{f,n}^0, F_{j,n}^0| \leq \sum_{s=1}^{n} (k_a - 2) \left( \sum_{j \in g_{a,s}} |F_{j,n}^0| \right)^2
\]
\[
= \sum_{s=1}^{n} (k_a - 2) k_a^2 \left( k_a^{-1} \sum_{j \in g_{a,s}} |F_{j,n}^0| \right)^2 \leq \sum_{s=1}^{n} (k_a - 2) k_a^2 k_a^{-1} \sum_{j \in g_{a,s}} (F_{j,n}^0)^2
\]

The first inequality is by adding and subtracting the diagonal terms, and the second is by Jensen. Using the two preceding displays, we have
\[
E[T_{n,2}^2 | F_{(\psi,p_n),n}] \leq \sum_{a=1}^{L_n} n^{-2} k_a^2 (k_a - 2)(k_a - q_a - 1) \sum_{s=1}^{n} \sum_{j \in g_{a,s}} (F_{j,n}^0)^2
\]
\[
\leq n^{-2} \sum_{a=1}^{L_n} 2k_a^2 \sum_{s=1}^{n} \sum_{j \in g_{a,s}} (F_{j,n}^0)^2 \leq \delta^{-1} n^{-2} \sum_{a=1}^{L_n} \sum_{s=1}^{n} \sum_{j \in g_{a,s}} (F_{j,n}^0)^2
\]
\[
\leq n^{-1} E_n[T_{i}(F_{j,n}^0)^2] = O_p(n^{-1})
\]

The first inequality simplifies, using \(\max_{k \geq 2} k_a / k_a - 1 = 2\) and \(\max_{k \geq 2} k_a / k_a - 1 \leq 1\). The second inequality uses \(k_a / q_a = \frac{1}{p_a} \leq \delta^{-1}\) and \(q_a - 1 \leq 1\). By Lemma 8.16, \(T_{n,2} = O_p(n^{-1})\). A very similar argument shows that \(T_{n,3} = O_p(n^{-1})\). Putting this together, we have shown that \(E[T_{n}^2 | F_{(\psi,p_n),n}] = T_{n,1} + T_{n,2} + T_{n,3} = O_p(n^{-1})\), so that \(T_n = O_p(n^{-1/2})\) by Lemma 8.16.

Summarizing, we have shown that
\[
\hat{v}_{10} = S_n + T_n = E[S_n | F_{(\psi,p_n),n}] + (S_n - E[S_n | F_{(\psi,p_n),n}]) + T_n
\]
\[
= E[F_i^0 F_i^0 | \zeta] + O_p(n^{-1/2}) + O_p(r_n^k) + O_p(r_n^\psi) + O_p((T_n/n)^{1/2}) + O_p\left(\sqrt{|L_n|}/n\right)
\]

This completes the proof of the lemma. \(\square\)

8.5 Propositions

**Definition 8.13** (Conditional Weak Convergence). For random variables \(A_n \in \mathbb{R}^d\) and \(\sigma\)-algebras \((\mathcal{F}_n)_n\), define conditional weak convergence
\[
A_n | \mathcal{F}_n \Rightarrow A \quad \iff \quad E[e^{it A_n} | \mathcal{F}_n] = E[e^{it A}] + o_p(1) \quad \forall t \in \mathbb{R}^d
\]

We require a slight modification of the martingale difference CLT in Billingsley.

**Proposition 8.14** (MDS-CLT). Consider probability spaces \((\Omega_n, \mathcal{G}_n, P_n)\) each equipped with filtration \((\mathcal{F}_{k,n})_{k \geq 0}\). Suppose \((Y_{k,n})_{k=1}^n\) is adapted to \((\mathcal{F}_{k,n})_{k \geq 0}\) and has \(E[Y_{k,n} | \mathcal{F}_{k-1,n}] = \)
0 for all \( k \geq 1 \) with \( r_n \to \infty \). Make the following definitions

\[
S_{k,n} = \sum_{j=1}^{k} Y_{k,n} \quad \Sigma_{k,n} = \sum_{j=1}^{k} \sigma_{k,n}^2
\]

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

\[
\sum_{k=1}^{n} E[Y_{k,n}^2 I(|Y_{k,n}| \geq \epsilon)|\mathcal{F}_{0,n}] = o_p(1) \tag{8.8}
\]

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

Proof. We may compute

\[
E\left[e^{i\epsilon S_n} - e^{-\frac{1}{2} \epsilon^2 \sigma^2}|\mathcal{F}_{0,n}\right] = E[e^{i\epsilon S_n}(1 - e^{\frac{1}{2} \epsilon^2 \Sigma_n} e^{-\frac{1}{2} \epsilon^2 \sigma^2})|\mathcal{F}_{0,n}] + E[e^{-\frac{1}{2} \epsilon^2 \sigma^2}(e^{\frac{1}{2} \epsilon^2 \Sigma_n} e^{i\epsilon S_n} - 1)|\mathcal{F}_{0,n}]
\]

For the first term, by conditional Jensen inequality

\[
|E[e^{i\epsilon S_n}(1 - e^{\frac{1}{2} \epsilon^2 \Sigma_n} e^{-\frac{1}{2} \epsilon^2 \sigma^2})|\mathcal{F}_{0,n}]| \leq E||(1 - e^{\frac{1}{2} \epsilon^2 \Sigma_n} e^{-\frac{1}{2} \epsilon^2 \sigma^2})||\mathcal{F}_{0,n}|
\]

The first equality since \( \Sigma_n \in \mathcal{F}_{0,n} \). Since \( \Sigma_n = \sigma^2 + o_p(1) \), the second equality follows by continuous mapping. Following the telescoping argument in Billingsley (henceforth BL), the second term is

\[
|e^{-\frac{1}{2} \epsilon^2 \sigma^2} E[(e^{\frac{1}{2} \epsilon^2 \Sigma_n} e^{i\epsilon S_n} - 1)|\mathcal{F}_{0,n}]| = e^{-\frac{1}{2} \epsilon^2 \sigma^2} E[(e^{\frac{1}{2} \epsilon^2 \Sigma_n} e^{i\epsilon S_n} - 1)|\mathcal{F}_{0,n}]|
\]

The second to last inequality follows 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 \( 1 \leq k \leq n \). Define \( Z_n = \sum_{k=1}^{n} E[|E[(e^{i\epsilon Y_{k,n}} - e^{\frac{1}{2} \epsilon^2 \sigma_{k,n}})^2]|\mathcal{F}_{k-1,n}||\mathcal{F}_{0,n}] \), and note that this is positive. For the final inequality, let \( c > \sigma^2 \) and note that

\[
(e^{\frac{1}{2} \epsilon^2 \Sigma_n} - e^{\frac{1}{2} \epsilon^2 c}) Z_n \leq (e^{\frac{1}{2} \epsilon^2 \Sigma_n} - e^{\frac{1}{2} \epsilon^2 c}) Z_n I(\Sigma_n \geq c)
\]

The first inequality since \( Z_n \geq 0 \). The second is by (conditional) Jensen inequality. For the final equality

\[
P\left(2n(e^{\frac{1}{2} \epsilon^2 \Sigma_n} - e^{\frac{1}{2} \epsilon^2 c}) I(\Sigma_n \geq c) > \epsilon\right) \leq P(\Sigma_n \geq c) = o_p(1) \]
Fix $\epsilon > 0$ and let $I_{k,n} = \mathbb{1}(Y_{k,n} > \epsilon)$. The same argument as in BL shows that for $K_t \geq t^2 \lor |t|^3$

$$|E[(e^{tY_{k,n}} - e^{-\frac{1}{2}t^2\sigma^2_{k,n}})|\mathcal{F}_{k-1,n}]| \leq K_t(\sigma^4_{k,n} + Y^2_{k,n}I_{k,n} + \epsilon \sigma^2_{k,n})$$

so that the sum above is

$$\sum_{k=1}^{r_n} E[|E[(e^{tY_{k,n}} - e^{-\frac{1}{2}t^2\sigma^2_{k,n}})|\mathcal{F}_{k-1,n}]||\mathcal{F}_{0,n}] \leq K_t \sum_{k=1}^{r_n} E[\sigma^4_{k,n} + Y^2_{k,n}I_{k,n} + \epsilon \sigma^2_{k,n}|\mathcal{F}_{0,n}]$$

$$= K_t E \left[ \sum_{k=1}^{r_n} \sigma^4_{k,n} + Y^2_{k,n}I_{k,n} + \epsilon \sigma^2_{k,n}|\mathcal{F}_{0,n} \right] \leq K_t E \left[ \sum_{k=1}^{r_n} (\max \sigma^2_{k,n} + \epsilon) + \sum_{k=1}^{r_n} Y^2_{k,n}I_{k,n}|\mathcal{F}_{0,n} \right]$$

$$\leq K_t E \left[ \sum_{n} (\epsilon^2 + \sum_{k=1}^{r_n} E[Y^2_{k,n}I_{k,n}|\mathcal{F}_{k-1,n}] + \epsilon) + \sum_{k=1}^{r_n} Y^2_{k,n}I_{k,n}|\mathcal{F}_{0,n} \right]$$

$$= K_t \sum_{k=1}^{r_n} (\epsilon^2 + \epsilon) + K_t (\sum_{n} + 1) \sum_{k=1}^{r_n} E[Y^2_{k,n}I_{k,n}|\mathcal{F}_{0,n}] \leq K_t c(\epsilon^2 + \epsilon) + K_t(c + 1) o_p(1) + o_p(1)$$

The first inequality by the fact from BL in second to last display. The first equality is by Tonelli’s theorem. The second inequality uses the definition of $\Sigma_n$. The third inequality follows as in BL by taking max$_k$ on both sides of

$$\sigma^2_{k,n} = E[Y^2_{k,n}|\mathcal{F}_{k-1,n}] \leq E[\epsilon^2 + Y^2_{k,n}I_{k,n}|\mathcal{F}_{k-1,n}] \leq \epsilon^2 + \sum_{k=1}^{n} E[Y^2_{k,n}I_{k,n}|\mathcal{F}_{k-1,n}]$$

The second to last equality follows by tower law and because $\Sigma_n \in \mathcal{F}_{0,n}$ as assumed above. The final inequality because (1) $\Sigma_n = c + \sum_n - c \leq c + (\sum_n - c) + \mathbb{1}(\sum_n \geq c) = c + o_p(1)$ as shown before and (2) from the assumed Lindberg condition in probability. Then we have shown that $E \left[ e^{it\Sigma_n} - e^{-\frac{1}{2}t^2\sigma^2} | \mathcal{F}_{0,n} \right] = o_p(1)$. \hfill \square

**Proposition 8.15 (SBR Equivalence).** Suppose $n = \ell k$ for $\ell \in \mathbb{N}$. Then $\text{Loc}_n(1, a/k) = \text{CR}(a/k)$ for any $1 \leq a \leq k - 1$

Let $\tau \in S_n$ and $g(\tau) = 1/|\tau|$. For $1 \leq t \leq \ell$, let $g_t = \{(t - 1)k + 1, \ldots, tk\}$ and the pre-image $\tau^{-1}(g_t) = \{\tau^{-1}(i) : i \in g_t\}$. Then for $D_{1:n} \sim \text{Loc}_n(1, a/k)$ by definition

$$P(D_{1:n} = d_{1:n}) = \sum_{\tau \in S_n} P(D_{1:n} = d_{1:n}|\tau) g(\tau) = \sum_{\tau \in S_n} \prod_{t=1}^{\ell} P(D_{r^{-1}(g_t)} = d_{r^{-1}(g_t)}|\tau) g(\tau)$$

$$= \sum_{\tau \in S_n} \prod_{t=1}^{\ell} \sum_{\xi_t \in \{0,1\}^k} P(D_{r^{-1}(g_t)} = d_{r^{-1}(g_t)}|\xi_t, \tau) g(\xi_t|\tau) g(\tau)$$

$$= \sum_{\tau \in S_n} \prod_{t=1}^{\ell} \sum_{\xi_t \in \{0,1\}^k} P(D_{r^{-1}(g_t)} = d_{r^{-1}(g_t)}|\xi_t, \tau) g(\xi_t) g(\tau)$$

Where by abuse of notation we also denote $g(\xi_t) = (\xi_t)^{-1} \mathbb{1}(\#\{i : \xi_{it} = 1\} = a)$. The last equality by definition of the design since the index permutation is independent of the within-group randomizations. The inner probability is in $\{0,1\}$, since the design is totally
Putting this together, we see that the second equality follows since (\xi_t)_t \perp \tau. 

Consider \( d, d' \in \{0, 1\} \) s.t. \#\{i : d_i = 1\} = a\ell \) and likewise for \( d' \). Clearly there exists \( \pi \in \mathcal{S}_n \) s.t. \( d_{\pi(i)} = d'_i \) for all \( 1 \leq i \leq n \). Then we have

\[
P(D_{1:n} = d') = P((D_i) = (d'_i)) = P((D_i) = (d_{\pi(i)})) = P((D_{\pi^{-1}(i)} = (d_i))
\]

Now note that

\[
(D_i) = H((\xi_t)_t, \tau) \overset{d}{=} H((\xi_t)_t, \pi \circ \tau) = (D_{\pi^{-1}(i)})
\]

The second equality follows since \( \tau \sim \text{Unif}(\mathcal{S}_n) \), so \( \pi \circ \tau \overset{d}{=} \tau \) and because \( (\xi_t)_t \perp \tau \). Putting this together, we see that \( P(D_{1:n} = d') = P(D_{1:n} = d) \). Since \( d, d' \) were arbitrary

\[
\text{Loc}_n(1, a/k) = \text{Unif}({d_{1:n} : E_n[d_i] = a/k}) = \text{CR}(a/k)
\]

### 8.6 Lemmas

**Lemma 8.16** (Conditional Convergence). Let \((\mathcal{G}_n)_{n \geq 1}\) and \((A_n)_{n \geq 1}\) a sequence of \( \sigma \)-algebras and RV’s. Define conditional convergence

\[
A_n = o_{p, \mathcal{G}_n}(1) \iff P(|A_n| > \epsilon | \mathcal{G}_n) = o_p(1) \quad \forall \epsilon > 0
\]

\[
A_n = O_{p, \mathcal{G}_n}(1) \iff P(|A_n| > s_n | \mathcal{G}_n) = o_p(1) \quad \forall s_n \to \infty
\]

Then the following results hold

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

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

(iii) \( \text{Var}(A_n | \mathcal{G}_n) = o_p(c_n^2)/O_p(c_n^2) \implies A_n - E[A_n | \mathcal{G}_n] = o_p(c_n)/O_p(c_n) \) for all positive \( c_n \)

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

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

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

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

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

The equality is by (i) and our assumption. Since \( \epsilon > 0 \) was arbitrary \( E[|A_n| | \mathcal{G}_n] = o_p(1) \). \(\square\)
Lemma 8.17 (Product Diagonalization). Consider positive scalar arrays \((x_i)_{i=1}^k, (y_i)_{i=1}^k\). Then the following inequalities are valid

\[
\begin{align*}
\sum_{i,j,l,f=1 \atop \text{i,j,l,f distinct}}^k x_i x_l y_j y_f &\leq k^3 \sum_{i=1}^k x_i^4 + k^3 \sum_{j=1}^k y_j^4, \\
\sum_{i,j,f=1 \atop \text{i,j,f distinct}}^k x_i y_j y_f &\leq k^2 \sum_{i=1}^k x_i^4 + k^2 \sum_{j=1}^k y_j^4
\end{align*}
\]

\[
\begin{align*}
\sum_{i,j=1 \atop \text{i,j distinct}}^k x_i y_i x_j y_j &\leq k \sum_{i=1}^k x_i^4 + k \sum_{j=1}^k y_j^4 + k^3 \sum_{i=1}^k x_i^4 + k^3 \sum_{j=1}^k y_j^4 \\
\sum_{i,j=1 \atop \text{i,j distinct}}^k x_i^2 y_j^2 &\leq k \sum_{i=1}^k x_i^4 + k \sum_{j=1}^k y_j^4 \\
\sum_{i,j=1 \atop \text{i,j distinct}}^k x_i y_j x_j y_j &\leq k \sum_{i=1}^k x_i^4 + k \sum_{j=1}^k y_j^4
\end{align*}
\]

Proof. Let \(a \geq 1\), so that \(\varphi(x) = x^a\) is convex on \(\mathbb{R}_{\geq 0}\). Note that by Jensen’s inequality

\[
\left( \sum_{i=1}^k x_i \right)^a = k^a \left( \sum_{i=1}^k x_i \right)^a \leq k^a k^{-1} \sum_{i=1}^k x_i^a = \sum_{i=1}^k x_i^a
\]

For the first statement

\[
\sum_{i,j,l,f=1 \atop \text{i,j,l,f distinct}}^k x_i x_l y_j y_f \leq \left( \sum_{i,j=1}^k x_i y_j \right)^2 \leq \left( \sum_{i=1}^k x_i \right)^2 \left( \sum_{j=1}^k y_j \right)^2 \leq \left( \sum_{i=1}^k x_i \right)^4 + \left( \sum_{j=1}^k y_j \right)^4 \leq k^3 \sum_{i=1}^k x_i^4 + k^3 \sum_{j=1}^k y_j^4
\]

The first inequality by positivity, the second by Young’s inequality, and the third by the fact above. Next we have

\[
\sum_{i,j,f=1 \atop \text{i,j,f distinct}}^k x_i^2 y_j y_f \leq \left( \sum_{i=1}^k x_i^2 \right) \left( \sum_{j=1}^k y_j \right)^2 \leq k \left( \sum_{i=1}^k x_i^2 \right) \left( \sum_{j=1}^k y_j^2 \right) \leq k \left( \sum_{i=1}^k x_i^2 \right)^2 + k \left( \sum_{j=1}^k y_j^2 \right)^2 \leq k^2 \sum_{i=1}^k x_i^4 + k^2 \sum_{j=1}^k y_j^4
\]

The inequalities follow by positivity and inspection, the fact above, Young’s inequality, and the fact above again, respectively. Note that by AM-GM inequality and Jensen, for
\(a, b, c, d \geq 0\) we have \(abcd \leq \left(\frac{1}{4}(a + b + c + d)\right)^4 \leq \frac{1}{4}(a^4 + b^4 + c^4 + d^4)\). Then we have
\[
\sum_{i,j=1 \atop i,j \text{ distinct}}^{k} x_i y_i x_j y_j \leq \left(\sum_{i=1}^{k} x_i y_i\right) \left(\sum_{i=1}^{k} x_i\right) \left(\sum_{j=1}^{k} y_j\right)
\]
\[
\leq \left(\sum_{i=1}^{k} x_i^2\right)^{1/2} \left(\sum_{i=1}^{k} y_i^2\right)^{1/2} \left(\sum_{l=1}^{k} x_l\right) \left(\sum_{j=1}^{k} y_j\right) \leq \|x\|_2 \|y\|_2 \left(\sum_{i=1}^{k} x_i\right) \left(\sum_{j=1}^{k} y_j\right)
\]
\[
\leq \|x\|^2_2 + \|y\|^2_2 + \left(\sum_{l=1}^{k} x_l\right)^4 + \left(\sum_{j=1}^{k} y_j\right)^4 \leq k \sum_{i=1}^{k} x_i^4 + k \sum_{i=1}^{k} y_i^4 + k^3 \sum_{i=1}^{k} x_i^4 + k^3 \sum_{j=1}^{k} y_j^4
\]
Next we have
\[
\sum_{i,j=1 \atop i,j \text{ distinct}}^{k} x_i^2 y_j^2 \leq \left(\sum_{i=1}^{k} x_i^2\right) \left(\sum_{j=1}^{k} y_j^2\right) \leq \left(\sum_{i=1}^{k} x_i^2\right)^2 + \left(\sum_{j=1}^{k} y_j^2\right)^2 \leq k \sum_{i=1}^{k} x_i^4 + k \sum_{j=1}^{k} y_j^4
\]
Next we have
\[
\sum_{i,j=1 \atop i,j \text{ distinct}}^{k} x_i y_i x_j y_j \leq \left(\sum_{i=1}^{k} x_i y_i\right)^2 \leq \|x\|^2 \|y\|^2 \leq \|x\|^4 + \|y\|^4 \leq k \sum_{i=1}^{k} x_i^4 + k \sum_{j=1}^{k} y_j^4
\]
\[
\square
\]

**Lemma 8.18** (Discretization Properties). Suppose assumption 8.1 holds with \(\zeta_n = \zeta\) for \(n \geq 1\), \(\zeta \perp W_{1,n}\), and \(\hat{p}_{i,n} = p_{i,n} = p(X_i, \zeta)\). Suppose \(G\) is bounded, Lipschitz and \(H\) a function with \(E[H(W_i, \zeta)|G(p_{i,n})]\) \(\leq \infty, \zeta\text{-a.s.}\) Then we have
\[
E_n[H(W_i, \zeta)G(p_{i,n})] = E[H(W_i, \zeta)G(p_i)] + O_p(r_n^p) + O_p(n^{-1/2})
\]
In particular, this holds for any \(C^1(K)\) function on \(K\) compact.

**Proof.** Let \(H_i \equiv H(W_i, \zeta)\). We may calculate
\[
|E_n[H_i G(p_{i,n})] - E_n[H_i G(p_i)]| \leq E_n \left[|H_i| |G(p_{i,n}) - G(p_i)|\right] \leq BE_n \left[|H_i| |p_{i,n} - p_i|\right] \leq |p_n - p| \infty E_n \left[|H_i|\right] = O_p(r_n^p)O_p(1)
\]
Now note that \(E[E_n[H(W_i, \zeta)G(p_i)]|\zeta] = E[H(W_i, \zeta)G(p_i)]\) and
\[
\text{Var} \left(E_n[H(W_i, \zeta)G(p(X_i, \zeta))]\right) = n^{-2} \sum_{i,j} \text{Cov}(H_i G(p_{i,n}), H_j G(p_{j,n}))\zeta)
\]
\[
= n^{-1} E_n \left[\text{Var}(H_i G(p_{i,n})|\zeta)\right] \leq n^{-1} E[H_i^2|\zeta] = O_p(n^{-1})
\]
The second equality by Lemma 8.21 with \(r_i(W_i) = 1\) and \(\xi_n = 1\). The first inequality by a second moment bound of the variance and boundedness of \(G\). By conditional Chebyshev (Lemma 8.16), this shows the claim.

\[
\square
\]

**Lemma 8.19** (Design Properties). Let \(D_{1:n} \sim \text{Gr}_n((g_{a,s})_{a,s}, n)\). Let \(\tau^d = (\tau^d_{a,s})_{a,s}\) jointly independent with \((\tau^d_{a,s})_{a,s}\) \(\sim \text{CR}(q_a/k_a)\) for \(1 \leq s \leq n - 1\) and \((\tau^d_{a,s})_{a,s}\) \(\sim \text{SRS}(q_a/k_a)\).
Let \( (F_n)_{n \geq 0} \) a sequence of \( \sigma \)-algebras with \( \sigma((g_{a,s})_{a,s}) \subseteq F_n \) and \( F_n \perp \perp \tau^d \). Then the following hold

(i) For all \( 1 \leq i \leq n \), there are deterministic \( G_i \) such that

\[
D_i \mathbb{1}(i \in g_{a,s}) = \sum_{\ell=1}^{k_a} \mathbb{1}(i = g_{a,s,\ell}) \tau^d_{a,s,\ell} \equiv G_i(g_{a,s}, \tau^d_{a,s})
\]

(ii) For each \( i \in [n] \) we have

\[
E[D_i \mathbb{1}(i \in g_{a,s}) | F_n] = \mathbb{1}(i \in g_{a,s}) \cdot p_a
\]

In particular, \( E[D_i \mathbb{1}(i \in g_{a,s}) | F_n] = \mathbb{1}(i \in g_{a,s}) \cdot p_a \).

(iii) For \( 1 \leq i \leq n \) and \( 1 \leq s \leq n \)

\[
\text{Var}(D_i | F_n) \mathbb{1}(i \in g_{a,s}) \leq \frac{1}{4} \mathbb{1}(i \in g_{a,s})
\]

For \( 1 \leq i, j \leq n \) distinct indices and \( 1 \leq s \leq n - 1 \)

\[
E[D_i D_j | F_n] \mathbb{1}(i, j \in g_{a,s}) = \frac{q_a(q_a - 1)}{k_a(k_a - 1)} \mathbb{1}(i, j \in g_{a,s})
\]

\[
\text{Cov}(D_i, D_j | F_n) \mathbb{1}(i, j \in g_{a,s}) = -\frac{q_a(k_a - q_a)}{k_a^2(k_a - 1)} \mathbb{1}(i, j \in g_{a,s})
\]

For \( 1 \leq i, j \leq n \) distinct indices and \( s = n \)

\[
\text{Cov}(D_i, D_j | F_n) \mathbb{1}(i, j \in g_{a,s}) = 0
\]

For \( 1 \leq i, j \leq n \) distinct indices

\[
\text{Cov}(D_i, D_j | F_n) \mathbb{1}(g(i) \neq g(j)) = 0
\]

In particular we have

\[
\text{Var}(D_i | F_n) \mathbb{1}(i \in g_{a,s}) \leq \frac{1}{4} \mathbb{1}(i \in g_{a,s})
\]

\[
| \text{Cov}(D_i, D_j | F_n) \mathbb{1}(i, j \in g_{a,s}) | \leq k_a^{-1} \mathbb{1}(i, j \in g_{a,s}) \mathbb{1}(s \neq n)
\]

**Proof.** For the first statement, note that

\[
D_i \mathbb{1}(i \in g_{a,s}) = \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \sum_{\ell=1}^{k_a} \mathbb{1}(i = g_{a,s,\ell}) \tau^d_{a,s,\ell} \mathbb{1}(i \in g_{a,s}) = \sum_{\ell=1}^{k_a} \mathbb{1}(i = g_{a,s,\ell}) \tau^d_{a,s,\ell} \equiv G_i(g_{a,s}, \tau^d_{a,s})
\]
For the second, note that under (i) and the assumed conditions
\[
E[D_i|\mathcal{F}_n] = \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \sum_{\ell=1}^{k_a} E[1(i = g_{a,s,\ell})\tau_{a,s,\ell}^d|\mathcal{F}_n] = \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \sum_{\ell=1}^{k_a} 1(i = g_{a,s,\ell})E[\tau_{a,s,\ell}^d|\mathcal{F}_n]
\]
\[
= \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \sum_{\ell=1}^{k_a} 1(i = g_{a,s,\ell})E[\tau_{a,s,\ell}^d]
= \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \sum_{\ell=1}^{k_a} 1(i = g_{a,s,\ell})p_a
\]
The second equality by the first containment. The third equality is by independence. Then note that \(\sum_{\ell=1}^{k_a} 1(i = g_{a,s,\ell}) = 1(i \in g_{a,s})\).

For (iii), by the decomposition above and \(\sigma\)-algebra assumption, for \(1 \leq j \neq i \leq n\) we have
\[
\text{Cov}(D_i, D_j|\mathcal{F}_n) = \sum_{a,a'=1}^{[L_n]} \sum_{s,s'=1}^{n} \sum_{\ell=1}^{k_a} \sum_{\ell'=1}^{k_{a'}} 1(i = g_{a,s,\ell})1(j = g_{a',s',\ell'}) \text{Cov}(\tau_{a,s,\ell}^d, \tau_{a',s',\ell'}^d|\mathcal{F}_n)
\]
By \(\sigma\)-algebra independence and joint independence of groupwise randomizations
\[
\text{Cov}(\tau_{a,s,\ell}^d, \tau_{a',s',\ell'}^d|\mathcal{F}_n) = \text{Cov}(\tau_{a,s,\ell}^d, \tau_{a',s',\ell'}^d)
= \begin{cases} 
0 & (a, s) \neq (a', s') \\
p_a - p_a^2 & (a, s, \ell) = (a', s', \ell') \\
-\frac{q_a(k_a-q_a)}{k_a(k_a-1)} & (a, s) = (a', s'); \ell \neq \ell' \quad 1 \leq s \leq n \\
0 & (a, s) = (a', s'); \ell \neq \ell' \quad s = n
\end{cases}
\]
The third line follows since by definition of \(\text{CR}(q_a/k_a)\), for \((a, s) = (a', s')\) we have
\[
\text{Cov}(\tau_{a,s,\ell}^d, \tau_{a',s',\ell'}^d) = P(\tau_{a,s,\ell}^d = \tau_{a',s',\ell'}^d = 1) - (q_a/k_a)^2 = (q_a/k_a)^{-1} \left(\frac{k_a}{q_a} - 2\right) - (q_a/k_a)^2
= \frac{q_a(k_a-q_a)}{k_a(k_a-1)} - \frac{q_a(k_a-q_a)}{k_a(k_a-1)}^2
\]
The bounds follow by inspection.

\[\square\]

**Lemma 8.20** (Stochastic Balance). Let \((\mathcal{F}_n)_{n \geq 1}\) such that \(\sigma((g_{a,s})_{a,s}, (h_n(W_i))_{i=1}^{n}) \subseteq \mathcal{F}_n\) for a sequence of functions \((h_n)_{n \geq 1}\). Let \(D_{1:n} \sim \text{Loc}_n(\psi,p)\) and require \(\mathcal{F}_n \perp \perp \tau^d\). Let \(S_{1:n} \in \{0,1\}^n\) such that \(i : S_i = 1 = \bigcup_{a,s} g_{a,s}\). The following hold

1. \(E_n[S_i(D_i - p_n(X_i))h_n(W_i)] = n^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \sum_{i \in g_{a,s}} (D_i - p_{i,n})h_n(W_i)\)
2. \(E[E_n[S_i(D_i - p_n(X_i))h_n(W_i)]|\mathcal{F}_n] = 0\)
3. \(\text{Var}(E_n[S_i(D_i - p_n(X_i))h_n(W_i)]|\mathcal{F}_n) \leq 2n^{-1}E_n[S_ih_n(W_i)^2] \leq 2n^{-1}E_n[h_n(W_i)^2]\)
4. If \(\sup_{a\geq 1} E[h_n(W_i)^2] < \infty\) and \(\exists (\mathcal{F}_n)_{n \geq 1}\) satisfying the conditions above, then \(E_n[S_i(D_i - p_n(X_i))h_n(W_i)] = O_p(n^{-1/2})\)
Finally, (5) we have a representation by within-group differences

\[
\text{Var}(E_n[S_i(D_i - p_{i,n})(W_i)])h_n(W_i)|\mathcal{F}_n) \leq n^{-2} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \frac{k^{-1}_{a}}{i,j \in g_{a,s} \cap \mathbb{M}^{2}} (h_n(W_i) - h_n(W_j))^2
\]

\[+ n^{-2}K_n|L_n| \cdot \max_{i=1}^{n} h_n(W_i)^2 \]

**Proof.** For (1), by assumption \(\{i : S_i = 1\} = \bigcup_{a,s} g_{a,s}\), so we have

\[
E_n[S_i(D_i - p_{i,n})h_n(W_i)] = n^{-1} \sum_{i : S_i = 1} (D_i - p_{i,n})h_n(W_i) = n^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \sum_{i \in g_{a,s}} (D_i - p_{a})h_n(W_i)
\]

\[= n^{-1} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \sum_{i=1}^{n} (D_i - p_{a})h_n(W_i) \mathbb{1}(i \in g_{a,s})\]

For (2), by Lemma 8.19.(ii) and our measurability assumptions

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

Then \(E[E_n[(D_i - p_{i,n})h_n(W_i)]|\mathcal{F}_n] = 0\). Using the expansion, \(\text{Var}(E_n[(D_i - p_{i,n})h_n(W_i)]|\mathcal{F}_n)\) is

\[
= n^{-2} \sum_{a,a'} \sum_{s,s'=1}^{n} \sum_{i,j} \text{Cov}((D_i - p_{a})h_n(W_i) \mathbb{1}(i \in g_{a,s}), (D_j - p_{a'})h_n(W_j) \mathbb{1}(j \in g_{a',s'})|\mathcal{F}_n)
\]

\[= n^{-2} \sum_{a,a'} \sum_{s,s'=1}^{n} \sum_{i,j} h_n(W_i)h_n(W_j) \mathbb{1}(i \in g_{a,s}) \mathbb{1}(j \in g_{a',s'}) \text{Cov}(D_i, D_j|\mathcal{F}_n)
\]

\[= n^{-2} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \sum_{i,j} h_n(W_i)h_n(W_j) \mathbb{1}(i, j \in g_{a,s}) \text{Cov}(D_i, D_j|\mathcal{F}_n)
\]

The final equality follows from Lemma 8.19.(iii). By triangle inequality and the covariance bound in Lemma 8.19.(iii), this is bounded above by

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

\[\leq n^{-1} \text{E}_n[h_n(W_i)^2] + \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \sum_{i \neq j} h_n(W_i)|h_n(W_j)| \mathbb{1}(i, j \in g_{a,s})k^{-1}_{a}
\]

\[\leq n^{-1} \text{E}_n[h_n(W_i)^2] + n^{-2} \sum_{a=1}^{[L_n]} \sum_{s=1}^{n} \left( \sum_{i=1}^{n} h_n(W_i) \mathbb{1}(i \in g_{a,s}) \right)^2 k^{-1}_{a}
\]

97
Continuing the calculation

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

The first inequality by removing the \( \mathbb{1}(s \neq n) \) indicator. The second inequality by adding and subtracting diagonal terms. The third inequality is Jensen’s. This completes the proof of (3). Claim (4) follows by applying (2), (3) with \((\mathcal{F}_n)_{n \geq 1}\) any sequence satisfying the conditions, followed by conditional Markov inequality (Lemma 8.16).

For the final identity (5), note that from Lemma 8.19.(iii) for \( s \neq n \)

\[
\sum_{i,j \in g_{a,s}, i < j} h_n(W_i) h_n(W_j) \text{Cov}(D_i,D_j|\mathcal{F}_n) = \frac{q_a (k_a - q_a)}{k_a^2} \sum_{i \in g_{a,s}} h_n(W_i)^2 + \frac{q_a (k_a - q_a)}{k_a^2 (k_a - 1)} \sum_{i,j \in g_{a,s}, i < j} (-2) h_n(W_i) h_n(W_j)
\]

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

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

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

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

The second term is

\[
n^{-2} \sum_{a=1}^{[L_n]} \sum_{i=1}^{n} h_n(W_i)^2 \mathbb{1}(i \in g_{a,n}) \leq n^{-2} \max_{i=1}^{n} h_n(W_i)^2 \sum_{a=1}^{[L_n]} \sum_{i=1}^{n} \mathbb{1}(i \in g_{a,n}) \\
\leq n^{-2} L_n \max_{i=1}^{n} h_n(W_i)^2
\]

98
This finishes the proof.

**Lemma 8.21** (Conditional Moments). Write \( W_i = (X_i, Y_i(1), Y_i(0)) \). Let \((h_i)_{i=1}^n\) and \((r_i)_{i=1}^n\) a collection of deterministic functions. Consider random elements \((\xi_n, \zeta_n)\) such that \((W_{i:n}, \xi_n, \zeta_n)\) are jointly independent. Let \(\mathcal{F}_n = \sigma((r_i(W_i, \zeta_n))_{i=1}^n, \xi_n, \zeta_n)\). Then for any \(I \subseteq [n]\)

\[
E \left[ \prod_{i \in I} h_i(W_i, \zeta_n) | \mathcal{F}_n \right] = \prod_{i \in I} E [h_i(W_i, \zeta_n)]
\]

In particular, if \(\exists j \in I\) such that \(E[h_j(W_j, \zeta_n)]r_j(W_j, \zeta_n, \zeta_n) = 0\) then

\[
E \left[ \prod_{i \in I} h_i(W_i, \zeta_n) | \mathcal{F}_n \right] = 0
\]

**Proof.** First, note that since the elements of \(W_{1:n}\) are jointly independent, and \(W_{1:n} \perp \zeta_n\), one can show that \((W_i, \zeta_n)_{i=1}^n\) are jointly independent given \(\zeta_n\). The proof heavily relies on the following fact: \((A, B) \perp C \implies A \perp C | B\). We have

\[
E \left[ \prod_{i \in I} h_i(W_i, \zeta_n) | \mathcal{F}_n \right] = E \left[ \prod_{i \in I} h_i(W_i, \zeta_n) \right] (r_i(W_i, \zeta_n))_{i=1}^n, \zeta_n, \xi_n
\]

\[
= E \left[ \prod_{i \in I} h_i(W_i, \zeta_n) \right] (r_i(W_i, \zeta_n))_{i=1}^n, \zeta_n
\]

\[
= E \left[ \prod_{i \in I \setminus \{j\}} h_i(W_i, \zeta_n) E \left[ h_j(W_j, \zeta_n) \prod_{i \in I \setminus \{j\}} h_i(W_i, \zeta_n), (r_i(W_i, \zeta_n))_{i \in I \setminus \{j\}}, \zeta_n \right] \right] (r_i(W_i, \zeta_n))_{i \in I}, \zeta_n
\]

The second equality is by \(\xi_n \perp (W_{1:n}, \zeta_n)\) and the fact above. The third equality is by independent sampling the fact above with \((A, B) = ((h_i(W_i, \zeta_n))_{i \in I}, (r_i(W_i, \zeta_n))_{i \in I})\) and \(C = (r_i(W_i, \zeta_n))_{i \in I\setminus\{j\}}\), applied under the conditional measure given \(\zeta_n\). The final equality is by tower law. The inner expectation is

\[
E \left[ h_j(W_j, \zeta_n) \prod_{i \in I \setminus \{j\}} h_i(W_i, \zeta_n), (r_i(W_i, \zeta_n))_{i \in I \setminus \{j\}}, \zeta_n \right]
\]

\[
= E \left[ h_j(W_j, \zeta_n) \prod_{i \in I \setminus \{j\}} h_i(W_i, \zeta_n), (r_i(W_i, \zeta_n))_{i \in I \setminus \{j\}}, r_j(W_j, \zeta_n), \zeta_n \right]
\]

\[
= E [h_j(W_j, \zeta_n)]r_j(W_j, \zeta_n, \zeta_n)
\]

The final equality uses the fact above with \(C = \left(\prod_{i \in I \setminus \{j\}} h_i(W_i, \zeta_n), (r_i(W_i, \zeta_n))_{i \in I \setminus \{j\}}\right)\) and \((A, B) = (h_j(W_j, \zeta_n), r_j(W_j, \zeta_n))\), again applied conditionally on \(\zeta_n\). Returning to
the display above, we have

\[ E \left[ \prod_{i \in I} h_i(W_i, \zeta_n) | \mathcal{F}_n \right] = E \left[ h_j(W_j, \zeta_n) | r_j(W_j, \zeta_n), \zeta_n \right] E \left[ \prod_{i \in I \setminus \{j\}} h_i(W_i, \zeta_n) \right] (r_i(W_i, \zeta_n))_{i \in I \setminus \{j\}, \zeta_n} \]

\[ = E \left[ h_j(W_j, \zeta_n) | r_j(W_j, \zeta_n), \zeta_n \right] E \left[ \prod_{i \in I \setminus \{j\}} h_i(W_i, \zeta_n) \right] (r_i(W_i, \zeta_n))_{i \in I \setminus \{j\}, \zeta_n} \]

The last equality uses \((A, B) = \left( \prod_{i \in I \setminus \{j\}} h_i(W_i, \zeta_n), (r_i(W_i, \zeta_n))_{i \in I \setminus \{j\}} \right)\) and \(C = r_j(W_j, \zeta_n)\).

This expression has the same form as in the first display, so the conclusion follows by induction. \(\square\)

**Lemma 8.22 (Residual Selection).** Let \(\mathcal{F}_n = \sigma(\psi_{1:n}, p_{1:n}, \zeta_n, \xi_n)\), with \(p_{1:n} = p_n(X_1, \zeta_n)\), \(\psi_i = \psi(X_i, \zeta_n)\). Assume \((W_{1:n}, \xi_n, \zeta_n)\) are jointly independent. Suppose that for \((F_{i,n})_{i=1}^n\) deterministic functions

\[ u_{i,n} = F_{i,n}(X_i, \zeta_n) - E[F_{i,n}(X_i, \zeta_n) | \psi_i, p_{i,n}, \zeta_n] \quad \sigma_{i,n}^2 = E[u_{i,n}^2 | \psi_i, p_{i,n}, \zeta_n] \]

Let \(i, j, l, f \in [n]\) distinct indices. The following moments hold for arbitrary mixtures of superscripts \(\epsilon_i \in \{\epsilon_1, \epsilon_0\}\), which we suppress.

\[ E[u_{i,n}|\mathcal{F}_n] = E[\epsilon_i | \mathcal{F}_n] = E[u_{i,n} \epsilon_i | \mathcal{F}_n] = E[u_{i,n} u_{j,n} | \mathcal{F}_n] = E[\epsilon_i \epsilon_j | \mathcal{F}_n] = E[\epsilon_i u_{n} \epsilon_j | \mathcal{F}_n] = E[\epsilon_i u_{n} u_{f,n} | \mathcal{F}_n] = E[u_{i,n}^2 | \psi_i, p_{i,n}, \zeta_n] = 0 \]

The higher order moments

\[ E[u_{i,n}^2 | \mathcal{F}_n] = \sigma_{i,n}^2 \]
\[ E[(u_{i,n} u_{j,n})^2 | \mathcal{F}_n] = \sigma_{i,n}^2 \sigma_{j,n} \]
\[ E[\epsilon_i^2 | \mathcal{F}_n] = E[\sigma^2(X_i) | \psi_i, p_{i,n}, \zeta_n] \]
\[ E[(u_{i,n} \epsilon_i)^2 | \mathcal{F}_n] = \sigma_{i,n}^2 \sigma_{j,n} \]
\[ E[(\epsilon_i \epsilon_j)^2 | \mathcal{F}_n] = E[\sigma^2(X_i) | \psi_i, p_{i,n}, \zeta_n] \]

**Proof.** We start with the first three inequalities. By applying Lemma 8.21 with \(h_i(W_i, \zeta_n) = u_{i,n}, r_i(W_i, \zeta_n) = (\psi(X_i, \zeta_n), p_n(X_i, \zeta_n))\) and \(I = [n]\), we have \(E[u_{i,n} | \mathcal{F}_n] = E[u_{i,n} | \psi_i, p_{i,n}, \zeta_n] = 0\). Similarly, we have

\[ E[\epsilon_i | \mathcal{F}_n] = E[\epsilon_i | \psi_i, p_{i,n}, \zeta_n] = E[E[\epsilon_i | X_i, \zeta_n] | \psi_i, p_{i,n}, \zeta_n] = E[E[\epsilon_i | X_i] | \psi_i, p_{i,n}, \zeta_n] = 0 \]

Also, \(E[u_{i,n} \epsilon_i | \mathcal{F}_n] = 0\) because

\[ E[u_{i,n} \epsilon_i | \psi_i, p_{i,n}, \zeta_n] = E[E[u_{i,n} \epsilon_i | X_i, \zeta_n] | \psi_i, p_{i,n}, \zeta_n] = E[u_{i,n} E[\epsilon_i | X_i] | \psi_i, p_{i,n}, \zeta_n] = 0 \]

The claims in the first display follow from these moments by applying Lemma 8.21 with \(r_i(W_i, \zeta_n) = (\psi_i, p_{i,n})\) and varying the index set \(I\) and functions \((h_i)_{i=1}^n\). For instance, for the second to last statement \(I = \{i, j\}\), \(h_i(W_i, \zeta_n) = u_{i,n} + \epsilon_i\) and \(h_j(W_j, \zeta_n) = (u_{j,n} + \epsilon_j)^2\).
The other equalities are similar.

For the second display, by Lemma 8.21 we have $E[u_{i,n}^2 | \mathcal{F}_n] = E[u_{i,n}^2 | \psi_i, p_{i,n}, \zeta_n] = \sigma_{u,i,n}^2$ and

\[
E[\epsilon_i^2 | \mathcal{F}_n] = E[\epsilon_i^2 | \psi_i, p_{i,n}, \zeta_n] = E[E[\epsilon_i^2 | X_i, \zeta_n] | \psi_i, p_{i,n}, \zeta_n] = E[E[\epsilon_i^2 | X_i] | \psi_i, p_{i,n}, \zeta_n] = E[\sigma_i^2(X_i) | \psi_i, p_{i,n}, \zeta_n]
\]

The first line by applying Lemma 8.21, then tower law and independence. The other equalities in the display now again follow by applying the Lemma 8.21 with $r_i(W_i, \zeta_n) = (\psi_i, p_{i,n})$ and varying the index set $I$ and functions $(h_i)_{i=1}^n$. For instance, for the last equality let $I = \{i, j\}$, $h_i(W_i, \zeta_n) = \epsilon_i^2$. \qed