Designing Representative and Balanced Experiments by Local Randomization*

Max Cytrynbaum†

November 19, 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

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

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

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

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

[1] We also allow the case where the researcher can select any unit in the entire population, which is known. See Remark 3.5 for discussion.

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

The 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).

Representative selection can significantly reduce estimator variance, while preserving finite-sample unbiasedness. To see why, consider estimating the effect of distributing surgical masks versus cloth masks on covid infection rates, as in Abaluck et al. (2021). Suppose average resident age predicts vulnerability to covid, so that older villages have larger treatment effects from switching to surgical masks from cloth masks. If selection is completely random, the selected villages may be older or younger 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. This illustrates how chance covariate imbalances created during selection can increase estimator variance (ex-ante) when treatment effects are heterogeneous. Representative selection avoids this problem, dampening this source of variance.

This paper implements both representative selection and finely balanced treatment assignment using a new family of *local randomization* algorithms. We introduce the basic principle of local randomization using the example above. Suppose the researcher has a sample of 1000 eligible villages, but logistical constraints only allow 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*[3] in the predictive covariates. For instance, the villages in each group of 10 may be very similar in terms of age, housing density, and baseline infection rates. Randomly select exactly 3 of 10 villages from each group into the experiment, leaving 300 villages. Suppose welfare considerations or logistical constraints require assigning 2/3 of the villages to surgical masks and 1/3 to cloth, as in Abaluck et al. (2021). Then during treatment assignment, again partition the 300 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. Enforcing group homogeneity allows us to 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 type of unit in the larger eligible sample is well-represented in the experiment. Under completely random selection, by contrast, 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 two-stage procedure, during selection we solve an optimization problem to match eligible units into homogeneous groups, randomly selecting a fixed proportion of

[3]Defined formally in Section 2.3 below.

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 and randomly assigning a fixed proportion from each group to one of two treatment arms. By introducing and studying this two-stage procedure, we make several contributions to the literature on experimental design.

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

This paper's second contribution is the selection model. To the best of our knowledge, we give a novel formalization of the principle of covariate-adaptive selection. We show that *efficient selection* requires finely balancing the covariates that are most predictive of treatment effect heterogeneity. Effectively, researchers should ensure that the different "types" of treatment response are well-represented among the selected units. Local randomization methods give an efficient and practical implementation of this idea.

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

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

Our fourth contribution gives new methods for asymptotically *exact* inference on the ATE under locally randomized 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

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

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

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

We also propose a second method that uses the pilot to "estimate what to balance" during both 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 some key predictive covariates outright, in addition to balancing the pilot regressions. Finally, for settings where a pilot experiment is infeasible, we discuss matching on *proxy regressions*, estimated in a related experiment or observational data set. In clinical trials, our results allow exact inference for designs that balance a vector of "risk scores" from previous studies. We also compare our results with the recent approaches of Tabord-Meehan (2020) and Bai (2020) in this setting.

1.1 Related Literature

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

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

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

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

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

Our 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 classical stratified block randomization (SBR) with fixed strata and matched pairs, as analyzed in Bugni et al. (2018) and Bai et al. (2021). Rerandomization, studied in Morgan and Rubin (2012) and Li et al. (2018), is not contained in this family. Follow up work in Cytrynbaum (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 set 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 oracle design for joint selection and assignment, as well as its empirical analogue estimated using pilot data. More generally, our method allows balancing with respect to more than just a single covariate.

For $\dim(\psi) > 1$, the algorithm used in Bai that constructs groups of units by sorting their $\psi(X_i)$ values is no longer feasible. We study combinatorial optimization procedures (graph-partitioning) to form groups in higher dimensions, and give new analysis of their statistical rates. Our designs allow propensity $p(x)$ non-constant. This innovation is required for a fully efficient solution to the two-wave design problem, which uses the pilot to estimate, and "locally" implement, an optimally varying propensity score (Neyman allocation). Finally, Bai (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 8 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[7] 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

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

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)

$$\text{ATE} = E[Y_i(1) - Y_i(0)] = E[c(X_i)]$$

2.1 Motivating Example

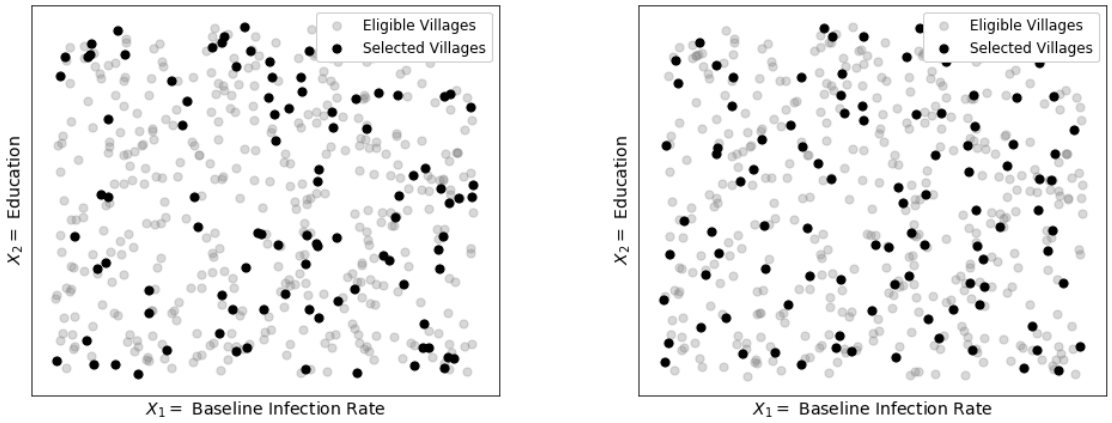
The first stage of the experimental design requires choosing selection variables $(T_i)_{i=1}^n$ using the observed covariates $(X_i)_{i=1}^n$. We begin by examining how selection affects the *precision* of our estimate of the ATE. To do this, we briefly consider a toy model of an experiment, in which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

$$\bar{\theta} = E_n[c(X_i)|T_i = 1] \equiv \frac{\sum_{i=1}^n T_i c(X_i)}{\sum_{i=1}^n T_i} \tag{2.1}$$

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,[8] 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

[7] We also allow additional covariates to be observed after selection but before treatment assignment, see Example 3.3.

[8] For brevity, assume this is an integer for now. We explicitly deal with edge effects in Section 3.



(a) Completely Random (b) Representative

Figure 1: Selecting villages to participate in the experiment using complete randomization versus *representative* selection using the local randomization design defined below.

set $q = 1/5$.

Complete Randomization - One natural way to draw the selection variables is by complete randomization: picking exactly qn out of n units uniformly[9] 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.

2.1.1 Variance due to Selection

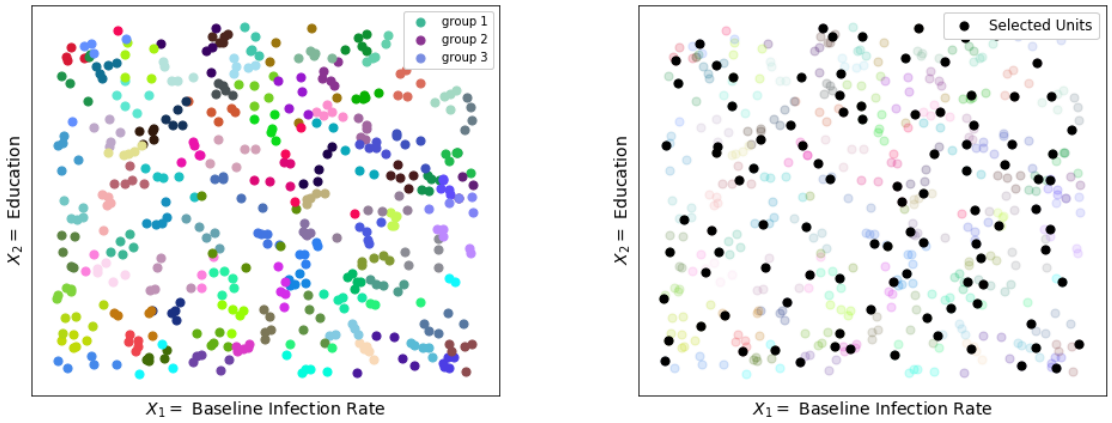
Using simulated data, Figure 1 shows one draw $T_{1:n} \sim \text{CR}(q)$ of the selection variables for $q = 1/5$. By chance, many of the highly educated, high infection rate villages were selected. Suppose, for instance, that these *types* of villages have larger treatment effect on average ($c(X_i)$ large). Then the estimator $\bar{\theta}$ will *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.

Representative Selection - By contrast, in Figure 1 (b) the distribution of village types X_i among the selected villages $X_i \,|\, T_i = 1$ closely matches the sample of all eligible villages $(X_i)_{i=1}^n$. By using representative selection, we force the smaller sample of selected villages to mimick 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

[9]Formally, $P(T_{1:n}) = \binom{n}{qn}^{-1}$ for each $T_{1:n}$ with $\sum_i T_i = qn$, and 0 otherwise.



(a) Local Groups (b) Selection

Figure 2: Representative selection: (a) form homogeneous groups of villages (different colors) and (b) randomly select proportion q units to represent each group.

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 $\psi(x) = (x_1, x_2) =$ (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 $q = 1/5$.

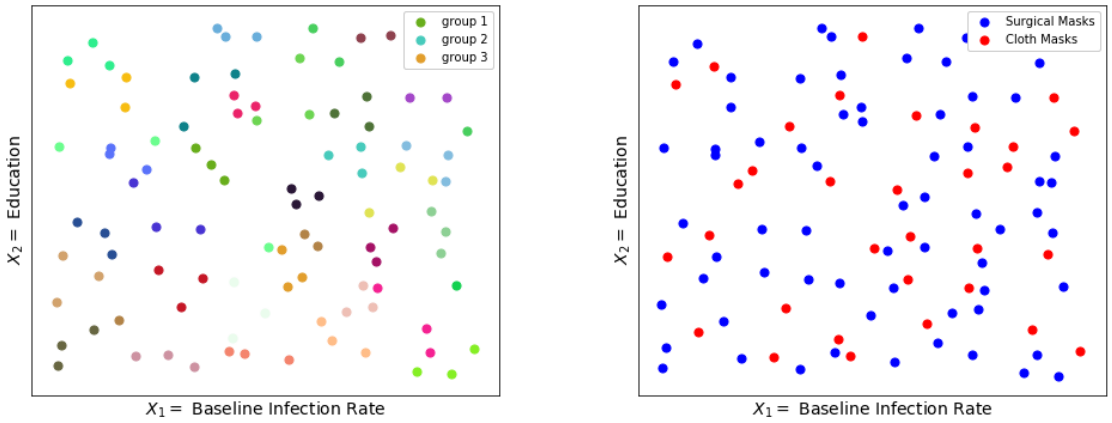
By randomizing within local groups instead of globally, Figure 2 shows that we get a significantly more uniform selection of villages, reducing the estimator variance due to selection. One can also think of this procedure as *data-adaptive* fine stratification.

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 $p = 1/2$

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

If treatment assignments D_i are completely randomized, $\widehat{\theta}$ is unbiased ex-ante but may have large variance. The reason for this is that while $E_n[D_i] = p$ globally, the *realized*



(a) Assignment Groups (b) Treatment Assignments

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.

propensity can significantly deviate from p in large regions of the 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 $\epsilon_i^d = Y_i - m_d(X_i)$ and define the *balance function*

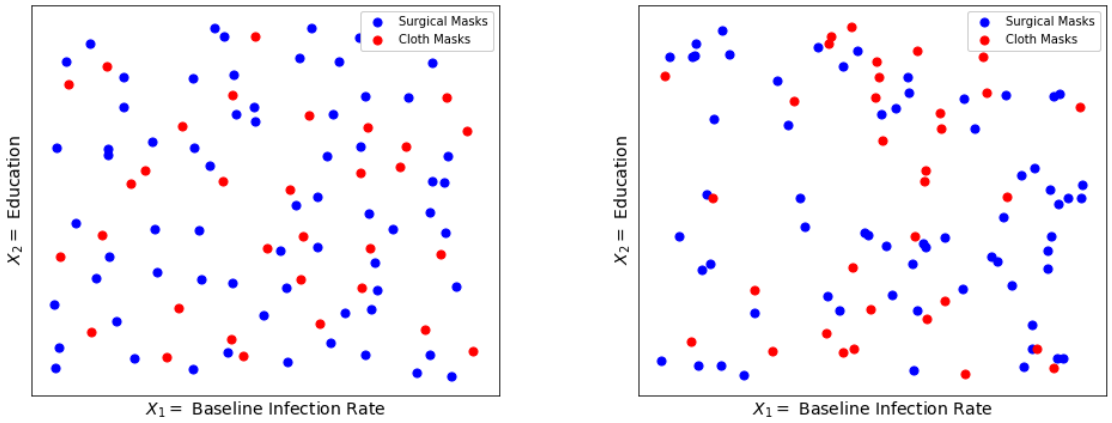
$$b(X; p) = m_1(X) \left(\frac{1-p}{p}\right)^{1/2} + m_0(X) \left(\frac{p}{1-p}\right)^{1/2} \tag{2.2}$$

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

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

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:



(a) Locally Randomized (b) Completely Randomized

Figure 4: Comparison of local vs. complete randomization, used jointly for both selection and 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 $D_i = 1$.

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

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

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 $\sigma_d^2(X) = \text{Var}(Y(d)|X)$ 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 $\sqrt{qn}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$

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

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 $\sqrt{qn}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$, with smaller asymptotic variance

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

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

$$R^2_{c,\psi} = \frac{\text{Var}(E[c(X)|\psi(X)])}{\text{Var}(c(X))} \qquad R^2_{b,\psi} = \frac{\text{Var}(E[b(X)|\psi(X)])}{\text{Var}(b(X))} \tag{2.6}$$

$R^2_{c,\psi} \in [0,1]$ 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

$$\text{Var}(b(X)) \cdot \left(1 - R^2_{b,\psi}\right) = E[(b(X) - E[b(X)|\psi(X)])^2] \tag{2.7}$$

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

$$\text{Var}(c(X)) \cdot \left(1 - (1-q)R^2_{c,\psi}\right) = q\,\text{Var}(c(X)) \tag{2.8}$$

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 $\text{Var}(c(X)) \to (1/5)\,\text{Var}(c(X))$, as if we had run a much *larger experiment* using all of the eligible villages. More generally, rearranging gives

$$\text{Var}(c(X)) \cdot \left(1 - (1-q)R^2_{c,\psi}\right) = q\,\text{Var}(E[c(X)|\psi(X)]) + E[\text{Var}(c(X)|\psi(X))] \tag{2.9}$$

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 4.7 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 $[n] = \bigsqcup_g \{i \in g\}$ 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

$$n^{-1} \sum_g \sum_{\substack{i,j \in g \\ i<j}} |\psi_i - \psi_j|_2^2 = o_p(1) \tag{2.10}$$

Equation 2.10 requires that the average squared Euclidean distance between features ψ_i and ψ_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 ψ_i have similar experimental *outcomes*, on average. In particular, the efficiency results in the previous section use the following smoothness condition[10]

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.

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 $(\psi(X_i))_{i=1}^n$.

[10] 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 8.1.

(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

$$T_{1:n} \sim \mathrm{Loc}_{\mathrm{n}}(\psi, q) \tag{2.11}$$

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

$$D_{1:n} \sim \mathrm{Loc}_{\mathrm{n}}(\psi, p \,|\, T_{1:n}) \tag{2.12}$$

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 $\asymp n$) 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.

$$(g^*) = \operatorname*{argmin}_{(g)} \sum_g \sum_{\substack{i,j \in g \\ i<j}} |\psi_i - \psi_j|_2^2 \tag{2.13}$$

Proposition 8.6 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 8.6 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 construct 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

$$j^* = \underset{j \in I}{\operatorname{argmin}} \sum_{\substack{i,j \in g \\ i<j}} |\psi_i - \psi_j|_2^2$$

(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 $D_{1:n} \sim \mathrm{Loc_n}(\psi, p(x))$. 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 $D_{1:n} \sim \mathrm{Loc_n}(\psi, p)$ we had the variance decomposition

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

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 ϵ_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 treament arm.[11] 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

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

Note that p^* is increasing in the conditional variance $\sigma_d^2(X) = \mathrm{Var}(Y(d)|X)$ 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 $\widehat{p} \to p^*$ by estimating the conditional variances above. Locally randomized assignment $D_{1:n} \sim \mathrm{Loc}_n(X, \widehat{p})$ 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

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

In Section 4.1, we show that the design $D_{1:n} \sim \mathrm{Loc}_n(X, \widehat{p}(x))$ is asymptotically fully efficient if $\widehat{p}(x)$ 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.

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.

[11] 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.

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 9.3 and 9.6 comprise various moment bounds, as well as the Lipschitz condition in Assumption 2.1 above.

Theorem 3.1 (CLT)**.** *Suppose that conditions 9.3 and 9.6 hold. Assume that the propensity $p = a/k$ is constant with $\gcd(a, k) = 1$. Suppose that selection and assignment are locally randomized:*

(i) $T_{1:n} \sim \mathrm{Loc_n}(\psi, q)$ *or* $T_{1:n} = 1$

(ii) $D_{1:n} \sim \mathrm{Loc_n}(\psi, p \,|\, T_{1:n})$

Then $\sqrt{qn}(\widehat{\theta} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V(\psi, (q, p))$

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

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 $E[(b - E[b|\psi])^2] = E[\operatorname{Var}(b|\psi)]$. 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 $E[(c - E[c|\psi])^2] = E[\operatorname{Var}(c|\psi)]$. 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 9.2 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

(i) $T_{1:n} \sim \mathrm{Loc}_{\mathrm{n}}(X, 3/7)$

(ii) $D_{1:n} \sim \mathrm{Loc}_{\mathrm{n}}(X, 1/3 \,|\, T_{1:n})$

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

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

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 $X_i = (X_i^{pre}, X_i^{post})$, with only X_i^{pre} available at selection time. Consider the design

(i) $T_{1:n} \sim \mathrm{Loc}_{\mathrm{n}}(X^{pre}, q)$

(ii) $D_{1:n} \sim \mathrm{Loc}_{\mathrm{n}}(X, p \,|\, T_{1:n})$

Then $\sqrt{qn}(\widehat{\theta} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V)$

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

We achieve full efficiency (relative to X) if and only if $c(X) = c(X^{pre})$. Intuitively, if $c(X)$ only varies with X^{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 9.15 in the appendix, $\mathrm{Loc}_{\mathrm{n}}(1, p) = \mathrm{CR}(p)$ 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 $\mathrm{CR}(a/k)$. From the theorem, if $T_{1:n} \sim \mathrm{Loc}_{\mathrm{n}}(1, q)$ and $D_{1:n} \sim \mathrm{Loc}_{\mathrm{n}}(1, p \,|\, T_{1:n})$. Then $\sqrt{qn}(\widehat{\theta} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V)$

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

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

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

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 9.3 and 9.6 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) $T_{1:n} \sim \mathrm{Loc_n}(\psi^*, q)$ *or* $T_{1:n} = 1$

(ii) $D_{1:n} \sim \mathrm{Loc_n}(\psi^*, p \,|\, T_{1:n})$

Then $\sqrt{qn}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V^*)$

$$V^* = q \operatorname{Var}(c(X)) + E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1-p}\right] \tag{3.1}$$

Under the optimal design, the selection and assignment terms are independently optimized. Note that the term $q \operatorname{Var}(c(X))$ 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 8 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 4.7 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 $\psi(x) = (\tilde{m}_0(x), \tilde{m}_1(x))$ be regressions of *proxy outcomes and treatments* $\tilde{Y}(\tilde{d})$ on X in a previous study (with some shared basline 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 $\tilde{Y}$ may be reading or overall test scores of different students in a previous study, under a related intervention $\tilde{d}$. Such approaches are commonly used in clinical trials, where $\tilde{m}_d(x)$ may be regarded as patient *risk scores*. If we use the design

(i) $T_{1:n} \sim \text{Loc}_n((\tilde{m}_0, \tilde{m}_1), q)$

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

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

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

This variance will be small if the risk scores $(\tilde{m}_0(x), \tilde{m}_1(x))$ 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 $\tilde{m}_d = \widehat{m}_d$ 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 $\|p_n - p\|_\infty = O(r_n^p) = o(1)$ and $p_n(x) \in L_n$ for all x, with rational propensity level set $L_n = \{p_a = q_a/k_a : a = 1, \ldots, |L_n|\}$, 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 $1 \le s \le n-1$, and remainder group $0 \le |g_{a,n}| \le k_a$.

(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 ξ_n. In notation, for some $\xi_n \perp\!\!\!\perp W_{1:n}$ we have $(g_{a,s})_{a,s} \in \sigma(\psi_{1:n}, p_{1:n,n}, \xi_n)$.

(iii) *Homogeneity*: For $\psi_i = \psi(X_i)$

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

(c) (Randomization) Complete randomization of treatment assignments within each local group $(D_i)_{i \in g_{a,s}} \sim \mathrm{CR}(p_a)$, independently over groups $1 \le s \le n-1$. Also, $D_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p_a)$ 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 $D_{1:n} \sim \text{Loc}_n(\psi, p(x))$. For treatment post-selection, we will denote $D_{1:n} \sim \text{Loc}_n(\psi, p(x) \,|\, T_{1:n})$, 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 $D_{1:n} \sim \text{Loc}_n(X, p = \{1/3, 2/3\})$, 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 8 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 $\overline{k}_n = \max_{p_a \in L_n} k_a$ the maximal group size, we impose the technical condition

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

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 $\zeta \perp\!\!\!\perp (X_i, Y_i(0), Y_i(1))_{i=1}^n$ a random element. In our applications, ζ 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 ζ. See the proof of Theorem 9.5 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 ψ 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* ζ, 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 9.3 and 9.7 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 $\zeta \perp\!\!\!\perp W_{1:n}$. Assume that selection and assignment are locally randomized*

(i) $T_{1:n} \sim \mathrm{Loc_n}(\psi', q)$ *or* $T_{1:n} = 1$

(ii) $D_{1:n} \sim \mathrm{Loc_n}(\psi, p(x) \,|\, T_{1:n})$

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

$$\begin{aligned} V\left((\psi', \psi), (q, p)\right) = {} & q\,\mathrm{Var}(c) + (1-q)E[(c - E[c|\psi', \zeta])^2|\zeta] \\ & + E[(b - E[b|\psi, p, \zeta])^2|\zeta] + E\left[\frac{\sigma_1^2(X)}{p(X,\zeta)} + \frac{\sigma_0^2(X)}{1-p(X,\zeta)}\middle|\zeta\right] \end{aligned}$$

Under the same conditions

$$\sqrt{qn}(\widehat{\theta} - E_n[c(X_i)])|\zeta \Rightarrow \mathcal{N}\Big(0, V\left((\psi', \psi), (q, p)\right) - q\,\mathrm{Var}(c(X))\Big)$$

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 9.15 in the appendix, $\mathrm{Loc_n}(S, p(S))$ is equivalent[12] to $\mathrm{SBR}(S, p(S))$. Consider a design with both stratified selection and assignment

[12] $D_{1:n} \sim \mathrm{Loc_n}(S, p(S))$ 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.

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

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

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

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

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 $\mathbb{1}_S$ be a full set of strata indicator functions. Consider the population OLS equations

$$\begin{aligned} b(X) &= \Gamma_b' \mathbb{1}_S + v_b \qquad E[\mathbb{1}_S v_b] = 0 \\ c(X) &= \Gamma_c' \mathbb{1}_S + v_c \qquad E[\mathbb{1}_S v_c] = 0 \end{aligned}$$

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 $\sigma_d^2(x) = \text{Var}(Y(d)|X = x)$ and define

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

The propensity $p^*(x)$ is known as the *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 $p_n(x)$

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

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 9.3 and 9.7 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) $T_{1:n} \sim \mathrm{Loc_n}(\psi^*, q)$ *or* $T_{1:n} = 1$

(ii) $D_{1:n} \sim \mathrm{Loc_n}(\psi^*, p^*(x) \,|\, T_{1:n})$

Then $\sqrt{qn}(\widehat{\theta} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V^*)$

$$V^* = q \operatorname{Var}(c(X)) + \min_{\substack{p \in L_2(X) \\ 0<p<1}} E\left[\frac{\sigma_1^2(X)}{p(X)} + \frac{\sigma_0^2(X)}{1-p(X)}\right] \tag{3.7}$$

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 $\widehat{p}(X_i)$ of the known propensity achieves the semi-parametric variance bound

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

By contrast, if $D_{1:n} \sim \mathrm{Loc_n}(X, p(x))$, then by Theorem 3.17 $\sqrt{n}(\widehat{\theta} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V^*)$. 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 $\sqrt{n} E_n[(D_i - p(X_i)) \mathbb{1}(X_i \in A)] = \Omega_p(1)$, reflecting order $n^{-1/2}$ dispersion of the *sample* treatment proportions away from the target proportions $p(x)$. Unless we update the propensity weights $p(X_i) \to \widehat{p}(X_i)$ to

reflect the realized, in-sample treatment proportions, this dispersion increases the first-order asymptotic variance. By contrast, Lemma 9.4 in the appendix shows that under local randomization

$$\sqrt{n}E_n[(D_i - p_n(X_i))\mathbb{1}(X_i \in A)] = O_p(n^{-1/(d+1)})$$

with $d = \dim(X)$. 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 $D_{1:n} \sim \mathrm{Loc_n}(\psi, p)$ is, using the law of total variance

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

This shows that, at least asymptotically, balancing *any function* of the covariates weakly dominates complete randomization, which achieves the maximal variance $\mathrm{Var}(b(X))$. The proof of Theorem 9.5 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 9.5 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 $D_{1:n} \sim \mathrm{Loc_n}(\psi, p)$ with $p = a/k$. By the fundamental decomposition of the IPW estimator

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

We highlight two key elements of the proof of Theorem 9.5 (from which Theorem 3.17 above is derived), focusing on our conditional CLT for the middle term above

(1) We show that within-group randomization non-parametrically "partials out" the projection $E[b|\psi]$, leaving only the variation not predicted by $\psi(X)$ (to first order). This gives the coupling below

$$\sqrt{n}E_n[(D_i - a/k)b(X_i)] = \sqrt{n}E_n[(D_i - a/k)(b - E[b|\psi])(X_i)] + O_p\left((r_n^{\psi})^{1/2}\right)$$

(2) We show that the leftover residuals $u_i \equiv b_i - E[b|\psi_i]$ behave "as if" randomly assigned to groups, so the effect of data-adaptive group construction and correlated assignment decays asymptotically (for the residuals)

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

For (1), observe that $b_i = E[b|\psi_i] + (b_i - E[b|\psi_i])$. If $\psi \to E[b|\psi]$ is Lipschitz, by Lemma

9.20 the local-group correlation structure of the treatment assignments gives

$$\begin{aligned}\text{Var}\left(\sqrt{n}E_n[(D_i - a/k)E[b|\psi_i]|X_{1:n}, \pi_n\right) &\leq k^{-1}\sum_g \sum_{\substack{i,j\in g \\ i\neq j}} (E[b|\psi_i] - E[b|\psi_j])^2 + O_p(\log(n)/n) \\ &\lesssim_P k^{-1}\sum_g \sum_{\substack{i,j\in g \\ i\neq j}} (\psi_i - \psi_j)^2 = O_p(r_n^\psi) = o_p(1)\end{aligned}$$

using the balancing rate guarantees for groups constructed in section 8. External randomness $\pi_n \perp\!\!\!\perp W_{1:n}$ is needed to break potential ties during group construction.

For (2), since the algorithms in section 8 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 $E[u_i u_j | i, j \in g] = E[E[u_i u_j|\psi_{1:n}, \pi_n]|i, j \in g] = 0$, showing the residuals behave "as if" randomly assigned to groups. The proof of Theorem 9.5 shows that

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

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 9.20, 9.22, and the proof of Theorem 9.5 in the appendix. For concrete rate guarantees on r_n^ψ under weaker smoothness assumptions, see section 8.

4 Efficient Design with a Pilot Experiment

This section proposes new methods that use data from a pilot[13] 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. However, since the residuals $\epsilon_i^d = Y_i(d) - E[Y_i(d)|X_i]$ are mean-independent of the baseline covariates X_i used to form local groups, randomizing within local groups cannot affect the *residual* variance. To further minimize the residual variance, we need to change the design's treatment proportions, assigning more units to the "noisier" treatment arm. Intuitively, one can think of this as taking more measurements of the quantity that is harder to measure. Full efficiency requires implementing the optimal treatment proportions conditionally, in proportion to the conditional standard deviation $\sigma_d(x) = \text{Var}(Y(d)|X = x)^{1/2}$ for $d = 0, 1$. While $\sigma_d(x)$ is unknown, it may be consistently estimated using pilot data.

Thus, we propose to use the pilot experiment to construct a feasible version of the fully optimal design in two steps, using the double stratification approach of Section 3.3. First,

[13]The results from this section also apply with access to observational data from the population of interest, as long as the relevant causal quantites are identified and estimable.

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

$$\widehat{p}(x) = \frac{\widehat{\sigma}_1(x)}{\widehat{\sigma}_1(x) + \widehat{\sigma}_0(x)} \qquad p^*(x) = \frac{\sigma_1(x)}{\sigma_1(x) + \sigma_0(x)} \tag{4.1}$$

Next, we implement these treatment proportions by local randomization within each propensity stratum.

For example, suppose the pilot estimate of the optimal treatment proportions takes values $\widehat{p}(x) \in \{1/2, 3/5\}$. In the propensity stratum $\{i : \widehat{p}(X_i) = 3/5\}$ we partition the units into homogeneous groups of 5, assigning exactly 3 in each group to $D = 1$, and similarly for the stratum with $\widehat{p}(X_i) = 1/2$. By implementing $\widehat{p}(x)$ by randomization within local groups, we strongly control the variance due to treatment assignment, while also minimizing the residual variance. In particular, our 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)**.** Section 4.1 proposes a *robust* method that does not require correct specification or estimation of outcome regression models However, the finite sample performance of this approach may be poor in high-dimensional settings with many weakly predictive covariates. Therefore, in Section 4.2 we provide an alternative method that uses the pilot study to "estimate what to balance." In particular, we show that locally randomizing with respect to a pilot estimate $\psi(x) = (\widehat{b}_n(x), \widehat{c}(x))$ of the optimal design $\psi^*(x) = (b_n(x), c(x))$ given by Theorem 3.19 is also 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 $\psi(x) = (\widehat{b}_n(x), \widehat{c}(x), x_{pred})$, 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 covariates X_i, let selection variables and treatment assignments be drawn by $T_{1:n} \sim \mathrm{Loc_n}(X, q)$ and $D_{1:n} \sim \mathrm{Loc_n}(X, \widehat{p}(x) \,|\, T_{1:n})$, where $\widehat{p}(x)$ is a "working model" of the Neyman allocation estimated using pilot data. For considerations involved in choosing the estimate $\widehat{p}(x)$, see the practical discussion in Remark 4.5 below.

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

4.1 Robust Solution

This section presents our robust solution. First, we discuss implementation of the Neyman allocation estimate in Equation 4.1 in more detail. Consider a simple implementation of the propensity $\widehat{p}(x)$ using independent sampling

$$D_i \overset{\text{inid}}{\sim} \text{Bernoulli}(\widehat{p}(X_i))$$

Unfortunately, under this strategy the realized treatment proportions may significantly differ from the pilot estimate $\widehat{p}(x)$ 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 $D_{1:n} \sim \text{Loc}_\text{n}(X, \widehat{p}(x))$, using the double stratification approach defined in Section 3.3 above. This gives a *fast implementation* of $\widehat{p}(x)$, forcing the realized treatment proportions to be exactly $\widehat{p}(x)$, up to lower order dispersion.[14] Using this method, we can implement the estimated optimal treatment proportions $\widehat{p}(x)$, 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 $n_{pilot} \asymp n$, working in *large pilot asymptotics* as in Hahn et al. (2011). This formally allows for consistent estimation $\|\widehat{\sigma}_d - \sigma_d\|_{2,\mathcal{X}} = o_p(1)$ using the pilot data.

While consistency is required for the asymptotic optimality result below, in practice we advocate treating $\widehat{\sigma}_d(x)$ as a "working model" of the variance function. Using the pilot data to implement an estimate of the optimal treatment proportions $p^*(x)$ will generally improve upon an arbitrary propensity score, such as $p = 1/2$, even if the estimator $\widehat{\sigma}_d(x)$ is misspecified or formed using a small pilot.

Theorem 4.3. *Suppose* $\|\widehat{\sigma}_d - \sigma_d\|_{2,\mathcal{X}} = o_p(1)$ *and impose assumptions 9.3 and 9.9. With* $\widehat{p}(x)$ *the pilot Neyman allocation estimate in Equation 4.1, suppose that*

(i) $T_{1:n} \sim \text{Loc}_\text{n}(X, q)$

(ii) $D_{1:n} \sim \text{Loc}_\text{n}(X, \widehat{p}(x) \,|\, T_{1:n})$

Then $\sqrt{qn}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V^*)$

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

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 9.3 of the appendix.

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

(1) Stratify the selected units on $p_n(X_i)$, values of a discretized[15] version of the estimated optimal treatment proportions $\widehat{p}(x)$. 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.

[14] Formally, for $A \subseteq \mathcal{X}$ a fixed "nice" region of the covariate space, $E_n[(D_i - p_n(X_i))\mathbb{1}(X_i \in A)] = O_p((r_n^\psi/n)^{1/2})$. See Lemma 9.4 for details

[15] If the discretization has $|L_n|$ propensity levels with $\overline{k}_n = \max_{q_a/k_a \in L_n} k_a$, the local balancing results in section 8 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 $\widehat{\sigma}_d^2(x)$, the asymptotics and inference results in Theorem 3.17 and Theorem 5.3 hold for any, potentially inconsistent, estimate $\widehat{p}(x) \perp\!\!\!\perp$ data of the optimal treatment proportions, including the following:

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

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

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

In each case, assigning $D_{1:n} \sim \mathrm{Loc_n}(X, \widehat{p}(x) \,|\, T_{1:n})$ 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 $\widehat{p}(x)$, is given by Theorem 5.3 in the next section.

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 $(\widehat{S}, \widehat{p}(s))$ over a set of tree partitions of the covariate space, denoted $\mathcal{T}$.

Theorem 3.1 of TM[16] shows that if $D_{1:n} \sim \mathrm{SBR}(\widehat{S}, \widehat{p}(s))$ then an IPW estimator $\widehat{\theta}$ has $\sqrt{n}(\widehat{\theta} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V)$ with asymptotic variance

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

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

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

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

[16]For a formal transformation of the SBR variance of Bugni et al. (2018) (cited in TM) into "balance function adjustment" form, see section 9.3 of the appendix

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

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

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

Remark 4.6 (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 $\epsilon_i^d = Y_i(d) - E[Y_i(d)|X_i]$ *cannot be affected* by matching-like methods, including classical stratification, since the baseline covariates X_i used for stratification are mean-independent of ϵ_i^d. The only parameter we can use to affect the residual variance is the propensity score $p(x)$, through which the proportion of residuals ϵ_i^d 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.

Remark 4.7 (Role of Covariate Dimension)**.** If $D_{1:n} \sim \mathrm{Loc_n}(X, \widehat{p}(x) \,|\, T_{1:n})$, our results show that the variance due to treatment assignment is asymptotically lower order. In particular, the covariate imbalance term in the IPW expansion has

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

See Lemma 9.4 in the appendix for details. 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, the "signal" of the relevant covariates may be lost by matching on all components of the high-dimensional vector X. Motivated by this, Section 4.2 considers using pilot data to "estimate what to balance," rather than balancing the whole vector X. Section 6 gives empirical evidence showing the deterioration of finite sample performance in high dimensions.

4.2 Using a Pilot Study to "Estimate what to Balance"

This section considers an alternative method that uses the pilot study to "estimate what to balance," motivated by concerns about poor finite-sample performance of the robust method in high-dimensions (Remark 4.7). In Section 3.4, we showed that a design that locally randomizes with respect to $\psi^*(x) = (b_n(x), c(x))$ during both selection and assignment is fully efficient.[17] Since the optimal design $\psi^*(x)$ is unknown, we consider estimating it using data from a pilot study, extending Bai (2020). Under well-specification and consistency of the pilot regressions, Theorem 4.8 shows that this feasible design is asymptotically fully efficient.

In spite of this result, 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. This leads us to prefer a "robustified" version of the method, discussed in Remark 4.13.

Advice for Practice - For high-dimensional settings with many weakly predictive covariates, experimenters can draw selection and assignment variables by local randomization with respect to $\psi(x) = (\widehat{b}_n(x), \widehat{c}(x), x_{pred})$, including pilot estimates of the balance function and CATE. They should also include, outright, a small vector of covariates x_{pred} thought to be good predictors of treatment efffect heterogeneity and variation in potential outcomes, if available. Theorem 5.3 in the next section gives asymptotically exact inference, conditional on the pilot data, for the designs in this section.

Theorem 4.8 (Balancing Pilot Regressions)**.** *Consider pilot regression estimators $(\widehat{m}_d)_{d=0,1}$ consistent with rate $\|\widehat{m}_d - m_d\|_{2,P} = O_p(n^{-r_m})$ and $\|\widehat{m}_d\|_\infty < \overline{m} < \infty$ a.s. for $d = 0, 1$. Require assumption 9.9, and let $\widehat{p}(x)$ be a consistent estimate of the Neyman allocation as in Theorem 4.3, with discretization $p_n(x)$ satisfying assumption 9.1. Define the estimators*

$$\widehat{b}_n(x) = \widehat{m}_1(x)\left(\frac{1-p_n(x)}{p_n(x)}\right)^{1/2} + \widehat{m}_0(x)\left(\frac{p_n(x)}{1-p_n(x)}\right)^{1/2} \qquad \widehat{c}(x) = \widehat{m}_1(x) - \widehat{m}_0(x)$$

Suppose one of the following

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

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

Suppose that the design is given by

(1) $T_{1:n} \sim \mathrm{Loc_n}(\widehat{\psi}, q)$

(2) $D_{1:n} \sim \mathrm{Loc_n}(\widehat{\psi}, \widehat{p}(x) \,|\, T_{1:n})$

Then $\sqrt{n}(\widehat{\theta} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V^*)$, with V^* the optimal variance of Theorem 3.19.

Theorem 4.8 achieves full efficiency by *embedding* the pilot regression estimators into the design, as we discuss in the following remark. Note that this result extends Bai (2020). See Remark 4.14 below for a discussion.

[17] See Remark 3.8 above for an intuitive discussion of this result.

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

$$\begin{aligned}\text{assignment}_n &\propto E_n[(D_i - 1/2)b(X_i)] = E_n[(D_i - 1/2)(b - \widehat{b})(X_i)] + E_n[(D_i - 1/2)\widehat{b}(X_i)] \\ &= E_n[(D_i - 1/2)(b - \widehat{b})(X_i)] + O_p(n^{-1})\end{aligned}$$

The final equality follows by matching on $\widehat{b}(x)$ in the design. Up to lower order, we are left with the bias term for a regression-augmented (AIPW) estimator, as in the doubly robust analysis of observational studies (Chernozhukov et al. (2017)). Thus, locally balancing a regression $\widehat{b}(x)$, makes the IPW estimator *behave like* a doubly-robust estimator using the pilot regression $\widehat{b}(x)$ for regression-augmentation.

Design Irrelevance: Motivated by the previous result, we study regression-augmented IPW estimation for locally randomized designs. 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 (Cross-fit AIPW)**.** Let $[n] = \cup_{k=1}^K I_k$ be a random partition. Define the out-of k^{th} fold regression estimate $\widehat{m}_{d,k} \equiv \widehat{m}_d((W_i)_{i \in I_k^c})$. For each k define the within k^{th} fold estimator

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

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

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

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

(2) $D_i \overset{\text{inid}}{\sim} p(X_i)$

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

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

Remark 4.12 (Consistency vs. Robustness)**.** Regression consistency is necessary for point estimation of treatment effects 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 $\text{Loc}_\text{n}(\psi, p(x))$ 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 $\text{Loc}_\text{n}(\psi, p(x))$ under this assumption. Independent sampling amounts to creating covariate imbalances[18] between treatment arms by design, but promising to correct them later with regression adjustment. This is an absurd recommendation for practice. For instance, consider that by Theorem 3.19 assigning $D_{1:n} \sim \text{Loc}_\text{n}(X, p(X))$ 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 $\psi(x) = (\widehat{m}_1(x), \widehat{m}_0(x))$ 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 $\psi(x) = (\widehat{m}_1(x), \widehat{m}_0(x), x_{pred})$, then Theorem 3.17 gives the asymptotic variance components

$$\begin{aligned}\text{selection} &= q\,\text{Var}(c(X)) + (1-q)E[(c(X) - E[c(X)|\widehat{m}_1(X), \widehat{m}_0(X), X_{pred}])^2] \\ \text{assignment} &= E[(b(X) - E[b(X)|\widehat{m}_1(X), \widehat{m}_0(X), X_{pred}])^2]\end{aligned}$$

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 $D_{1:n} \sim \text{Loc}_\text{n}(\widehat{b}, a/k)$. 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.

[18] Such 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.

Group Replicates - As noted in Fisher (1926), exact inference requires replicates of each treatment arm under similar experimental conditions. For $D_{1:n} \sim \mathrm{Loc_n}(\psi, p)$, ideal data would contain exact replicates of the outcomes $(Y_i)_{i \in g}$ in each local group under identical observed experimental conditions (identical (ψ_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 8.

(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

$$\gamma : \{i \in g, D_i = 1\} \longrightarrow \{i \in g', D_i = 1\}$$

Proposition 8.8 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 γ 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) $\gamma^2 = \mathrm{Id}$, $\gamma(i) \neq i$, and $\gamma \in \sigma((g_{a,s})_{a,s}, D_{1:n}, \xi_n)$

(iii) $E_n[(\psi_i - \psi_{\gamma(i)})^2 \,|\, T_i = 1] = O_p(r_n^{\psi})$

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), γ is a well-defined bijective matching, only depending on the groups, treatment values, and external randomness ξ_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 γ be a matching satisfying the conditions of Definition 5.1. With $p_{i,n} = p_n(X_i)$, define the matching estimators

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

Also define the cross-moment estimator and IPW sample moment

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

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 provided by Theorem 5.3 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 9.1, 9.3, and 9.11 are satisfied.*

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

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

   For the fixed regressor case

$$\begin{aligned}&\widehat{v}_{I,2} - \widehat{v}_1\left[p^{-1}\right] - \widehat{v}_0\left[(1-p)^{-1}\right]\\ &= V(\psi,(q,p)) - q\,\text{Var}(c(X)) + qE[\text{Var}(c(X_i)|\psi_i,\zeta)|\zeta] + o_p(1)\\ &\geq V(\psi,(q,p)) - q\,\text{Var}(c(X)) + o_p(1)\end{aligned}$$

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

$$\text{Var}_n\left(\frac{(D_i-p_{i,n})Y_i}{p_{i,n}-p_{i,n}^2}\right) - \widehat{v}_1\left[\frac{1-p}{p}\right] - \widehat{v}_0\left[\frac{p}{1-p}\right] - 2\widehat{v}_{10} = V(\psi,p) + o_p(1)$$

   For the fixed regressor case

$$\widehat{v}_{I,2} - \widehat{v}_1\left[p^{-1}\right] - \widehat{v}_0\left[(1-p)^{-1}\right] \geq V(\psi,p) - \text{Var}(c(X)) + o_p(1)$$

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

$$\begin{aligned}
&\text{Var}_n\left(\frac{(D_i - p_{i,n})Y_i}{p_{i,n} - p_{i,n}^2}\right) - \widehat{v}_1\left[\frac{1-p}{p}\right] - \widehat{v}_0\left[\frac{p}{1-p}\right] \\
&= V((\psi', \psi), (q, p)) + (1-q)\,\text{Var}(E[c(X_i)|\psi_i', \zeta]|\zeta) + o_p(1) \\
&\geq V((\psi', \psi), (q, p)) + o_p(1)
\end{aligned}$$

See section 9.4 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: $D_i \overset{\text{inid}}{\sim} p(X_i)$, also known as simple random sampling (SRS)

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

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

Measuring Homogeneity: To see when $\widehat{v}_1$ will be large, consider the matched products $(Y_i Y_{\gamma(i)})_i$ in the definition. By the Hardy-Littlewood inequality, $\widehat{v}_1$ is maximized[19] when matched outcomes are maximally homogeneous: large outcomes Y_i multiplied by large outcomes $Y_{\gamma(i)}$ and vice-versa. Then we can regard $\widehat{v}_1$ 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 $\widehat{v}_1 + \widehat{v}_0$ 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

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

$\psi'(x) = \psi(x)$. By enforcing this condition, we make the selection and assignment variance components comparable

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

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 $\widehat{v}_1, \widehat{v}_0, \widehat{v}_{10}$ 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 $X_i^{pre} \subsetneq X_i$. 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 $D_{1:n} \sim \mathrm{Loc_n}(\psi, p)$ and $\psi = \psi(\cdot, \zeta)$ a function estimated on proxy data ζ, the proof of Theorem 5.3 shows that

$$\widehat{v}_1[1] = c_1 \cdot R^2(\psi, \zeta) + c_2 + o_p(1) \qquad R^2(\psi, \zeta) = \frac{\mathrm{Var}(E[m_1(X)|\psi(X, \zeta), \zeta]|\zeta)}{\mathrm{Var}(m_1(X))}$$

with constants c_1, c_2 independent[20] 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 ζ. 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 $\theta = \mathrm{ATE}$ in a completely randomized experiment $D_{1:n} \sim \mathrm{CR}(1/2)$. The sample variance $16\,\mathrm{Var}_n\left((D_i - 1/2)Y_i\right)$, while exact if D_i are iid, is *conservative* under complete randomization due correlation of the treatment assignments. Proposition 9.15 shows that $\mathrm{CR}(1/2) = \mathrm{Loc_n}(1, 1/2)$, giving a "random matched pairs" representation of $\mathrm{CR}(1/2)$. 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

$$\widehat{v}_1[1] + \widehat{v}_0[1] = 2E_n\left[(D_i + (1 - D_i))Y_i Y_{\gamma(i)}\right] = 2E_n\left[Y_i Y_{\gamma(i)}\right]$$

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

Then by[21] Theorem 5.3, the following variance estimator is exact

$$\widehat{V} = 16\,\mathrm{Var}_n((D_i - 1/2)Y_i) - 2E_n\left[Y_i Y_{\gamma(i)}\right] - 2E_n\left[Y_i Y_{\phi(i)}\right]$$

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

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

We sample from the following DGP's

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

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

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

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

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

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

[21] Since the matched pairs in $\text{Loc}_\text{n}(\psi = 1, 1/2)$ are formed randomly, it suffices to use the random matchings γ, ϕ to define local groups ex-post.

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 $-(1-q)\,\mathrm{Var}(c(X))$ from local randomization. The boost in effective sample size (ESS) is significant relative to completely random selection. Note that this comes essentially "for free" by being slightly more careful about choosing which units to enroll in the experiment. In Model 3, $c(x) = 0$ identically, so the theory predicts no efficiency gain due to selection. The small variance reduction seen in practice is likely a finite sample effect where balanced first-stage selection facilitates better matches during the second-stage assignment process.

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

Table 1: Effect of Representative Selection

Table 1 suggests our inference procedure is generally slightly conservative in finite samples. This is likely due to the between-group matching used for inference in Section 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 %ΔLength = %ΔSD. In finite samples, our inference procedure gives a significant reduction in CI length, reflecting the increase in estimator precision due to selection.

Our second set of simulations focuses on the choice of $\psi(x)$. Asymptotically, including more baseline covariates always weakly improves estimator efficiency. However, as discussed in Remark 4.7, the within-group matching discrepancies converge slowly in high dimensions, giving balancing rate $r_n^{\psi} = O(n^{-2/(d+1)})$. This suggests preferentially including covariates that are predictive of outcomes and treatment effect heterogeneity (predict functions $b_n(x)$ and $c(x)$), excluding "noise covariates" that contain little additional information. Consider the quadratic model and variable distributions above, but

with $X_i \in \mathbb{R}^{10}$ and

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

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

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

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

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

Table 2: Varying $\psi(x)$

Results - There are significant precision gains from locally randomized selection and assignment, as predicted by Theorem 3.17. 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 $x_1 \dots x_6$ is slightly *worse* than $x_1 \dots x_4$ in finite samples. Recall have the largest weight in the outcome model. Additionally including $x_4 \dots x_6$, which are less predictive of outcomes, degrades match quality for the important covariates $x_1 \dots x_3$, reducing efficiency in finite samples. Including noise covariates x_7, x_8 performs still worse. As in Theorem 3.19, the optimal design $\psi^*(x) = (b_n(x), c(x))$ gives the largest efficiency gains.

7 Conclusion

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

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

A more conceptual extension could study *targeted selection*. For instance, we could acknowledge that the participants that sign up for a trial differ systematically from the target population (site selection bias). It would be interesting to try to use non-constant selection proportions $q(x)$ to "unwind" this bias at the design stage. For instance, we could imagine that the designer is given covariate data or moments from the target population. Can we give practical and statistically efficient designs that reverse the site selection bias?

References

Abadie, A. and G. W. Imbens (2012). A martingale representation for matching estimators. *Journal of the American Statistical Association 107*, 833–843.

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

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

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

Bai, Y. (2020). Optimality of matched-pair designs in randomized controlled trials. Working Paper.

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

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

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

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

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

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

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

Cytrynbaum, M. (2021). Local randomization dominates rerandomization. Working Paper.

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

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

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

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

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

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

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

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

Imbens, G. W. (2011). Experimental design for unit and cluster randomized trials. Working Paper.

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

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

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

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

Kolesàr, M. and T. Armstrong (2021). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica 89*(3), 1141–1177.

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

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

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica 56*(4).

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

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

Tabord-Meehan, M. (2020). Stratification trees for adaptive randomization in randomized controlled trials. Working Paper.

8 Local Balancing Rates

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

Definition 8.1 (Uniformly Piecewise Hölder)**.** Let $\dim(\psi) = 1$. Let ζ_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 $\max_{p_a \in L_n} \text{Card}(B_n(a, \zeta_n)) \leq \bar{B} < \infty$ and $C_n(\zeta_n) < \bar{C} < \infty$, ζ_n-a.s.

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

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

Example 8.2 (Stratified Block Randomization)**.** Recall that by Proposition 9.15, SBR with discrete strata $S(X) \in \{1, \ldots, T\}$ is equivalent to $D_{1:n} \sim \text{Loc}_\text{n}(S, p)$. 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 8.1 with $C_n = 0$, break sets $B_{p,n} = \{1, \ldots, T\}$, 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 8.3 (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 ψ_i value, producing $I_a^{sort} = (i_1, i_2, \ldots, i_{m(a)})$. Ties are broken with data-independent randomness π_n.

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

(iii) Increment $s \to s + 1$ and $I_a \to I_a \setminus g_{a,s}$ 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 8.4 (Balancing I)**.** *Let* $\dim(\psi) = 1$ *and* $\psi(X, \zeta_n) \in [a, b]$ *compact* ζ_n*-a.s. Let* $(g_{a,s})_{a,s}$ *be constructed as in Algorithm 8.3. 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 8.1 with* $\alpha \geq 1/2$*. Then groups* $(g_{a,s})_{a,s}$ *satisfy*

$$n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a^{-1} \sum_{\substack{i,j \in g_{a,s} \\ i \neq j}} (h_{i,n} - h_{j,n})^2 = O(r_n^\psi) = o(1) \tag{8.1}$$

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

For $\dim(\psi) > 1$, Algorithm 8.3 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 $F(s, p) = F(\lfloor s \rfloor, p)\ \forall s \in [1, T]$.

Definition 8.5 (Uniformly Lipschitz)**.** Let $\dim(\psi) \geq 1$. Let ζ_n be a random variable and $(h_n)_{n\geq 1}$ a sequence of deterministic functions. We say $(h_n)_{n\geq 1}$ is *uniformly Lipschitz* if $\forall p_a \in L_n$

$$|h_n(\psi, p_a, \zeta_n) - h_n(\psi', p_a, \zeta_n)| \leq C_n(\zeta_n)|\psi - \psi'|$$

with $C_n < \bar{C} < \infty$, ζ_n-a.s.

Proposition 8.6 (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 8.5. With* $\psi_{i,n} = \psi(X_i, \zeta_n)$, *consider groups* $(g_{a,s})_{a,s}$ *formed by solving the program*

$$\min_{(g_{a,s})_{a,s}} \quad n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a^{-1} \sum_{\substack{i,j\in g_{a,s} \\ i\neq j}} |\psi_{i,n} - \psi_{j,n}|_2^2 \tag{8.2}$$

$$\text{s. t.} \quad p_n(X_i, \zeta_n) = p_a \quad \forall i \in g_{a,s} \quad \forall s = 1, \ldots, n \tag{8.3}$$

Then the balancing rate

$$n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a^{-1} \sum_{\substack{i,j\in g_{a,s} \\ i\neq j}} (h_{i,n} - h_{j,n})^2 = O\left(\left(n/(\overline{k}_n |L_n|) \right)^{-2/(d_\psi+1)} \right)$$

Alternatively, if $(g_{a,s})_{a,s}$ *solve the modified problem*

$$\min_{(g_{a,s})_{a,s}} \quad n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a^{-1} \sum_{\substack{i,j\in g_{a,s} \\ i\neq j}} |\psi_{i,n} - \psi_{j,n}|_2 \tag{8.4}$$

$$\text{s. t.} \quad p_n(X_i, \zeta_n) = p_a \quad \forall i \in g_{a,s} \quad \forall s = 1, \ldots, n \tag{8.5}$$

Then the slower rate is achieved

$$n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a^{-1} \sum_{\substack{i,j\in g_{a,s} \\ i\neq j}} (h_{i,n} - h_{j,n})^2 = O\left(\left(n/(\overline{k}_n |L_n|) \right)^{-1/d_\psi} \right)$$

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

Remark 8.7 (Matched Pair Designs)**.** As above, the case $D_{1:n} \sim \mathrm{Loc}_n(\psi, 1/2)$ for $\psi(X) = X$ is equivalent to classical pairwise matching, as recently analyzed by Bai et al. (2021). In this case, $\overline{k}_n = |L_n| = 2$ 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 8.8 (Balancing III)**.** *Let* $\dim(\psi) = d_\psi$ *and consider groups* $(g_{a,s})_{a,s}$ *with balancing rate* $r_n^\psi = o(1)$. *Define the group representative* $\psi_{a,s} = k_a^{-1} \sum_{i\in g_{a,s}} \psi_i$. *For each* $a = 1, \ldots, |L_n|$, *form between-group matching* $\nu_a : [n] \to [n]$, $\nu_a^2 = \mathrm{Id}$ *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* $\gamma^2 = \mathrm{Id}$ *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 9.3. Then the between-group balancing rate*

$$E_n[T_i(\psi_i - \psi_{\gamma(i)})^2] = O_p\left(\left(n/(\overline{k}_n|L_n|)\right)^{-2/(d_\psi+1)}\right) + O_p(r_n^\psi) \tag{8.6}$$

For the proof of these propositions, see Section 9.1 in the appendix.

9 Proofs

9.1 Proofs - Balancing

9.1.1 Proof of Proposition 8.4

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

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

Using this fact, observe that

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

$$\lesssim n^{-1}\sum_{a=1}^{|L_n|} k_a \sum_{s=1}^{n} \operatorname{diam}(g_{a,s}, \psi(\cdot, \zeta_n)) \mathbb{1}(g_{a,s} \neq \emptyset) \leq n^{-1}\sum_{a=1}^{|L_n|} k_a \cdot (1-0) = O(|L_n|\overline{k}_n n^{-1})$$

For the final equality, for $g_{a,s} \neq \emptyset$ define open intervals $I_{a,s,n} \equiv (\psi^l_{a,s,n}, \psi^u_{a,s,n}) \subseteq [0,1]$. 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

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

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

$$n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a^{-1} \sum_{\substack{i,j \in g_{a,s} \\ i \neq j}} (h_n(\psi(X_i, \zeta_n), p_a, \zeta_n) - h_n(\psi(X_j, \zeta_n), p_a, \zeta_n))^2 \mathbb{1}(E_{a,s})$$

$$\leq n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a^{-1} \sum_{\substack{i,j \in g_{a,s} \\ i \neq j}} C_n(\zeta_n) |\psi_{i,n} - \psi_{j,n}|^{2\alpha} \mathbb{1}(E_{a,s}) = O(|L_n| \overline{k}_n n^{-1})$$

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 $E^c_{a,s}$ occurs. Let $Q_{a,s} \equiv E^c_{a,s} \cap \{\text{diam}(g_{a,s}, \psi(\cdot, \zeta_n)) > 0\}$. 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

$$n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a^{-1} \sum_{\substack{i,j \in g_{a,s} \\ i \neq j}} (h_n(\psi(X_i, \zeta_n), p_a, \zeta_n) - h_n(\psi(X_j, \zeta_n), p_a, \zeta_n))^2 \mathbb{1}(Q_{a,s})$$

$$\leq M_n^2 n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a^{-1} \mathbb{1}(Q_{a,s}) \sum_{\substack{i,j \in g_{a,s} \\ i \neq j}} 1 \leq M_n^2 n^{-1} \sum_{a=1}^{|L_n|} k_a^{-1}(k_a^2 - k_a) \sum_{s=1}^{n} \mathbb{1}(Q_{a,s})$$

To bound the inner sum, define the correspondence

$$\varphi : B_n(a, \zeta_n) \rightrightarrows \{s : \mathbb{1}(Q_{a,s}) = 1\}$$
$$\varphi(b) \longrightarrow \{s : b \in \overline{I}_{a,s}\}$$

where $\overline{I}_{a,s} \equiv [\psi^l_{a,s}, \psi^u_{a,s}]$. 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 $b \in \overline{I}_{a,s} \cap \overline{I}_{a,s'} \cap \overline{I}_{a,s''}$ for *distinct* intervals. Without loss suppose that $\psi^l_{a,s} < \psi^u_{a,s} \leq \psi^l_{a,s'} < \psi^u_{a,s'} \leq \psi^l_{a,s''} < \psi^u_{a,s''}$. 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 $\psi^u_{a,s} \geq b$. Then we have $\psi^l_{a,s''} \geq \psi^u_{a,s'} > \psi^l_{a,s'} \geq \psi^u_{a,s} \geq b$. From the strict inequality, we see that $b \notin \overline{I}_{a,s''}$, which is a contradiction. Clearly, the correspondence φ is a surjection, so we have

$$\sum_{s=1}^{n} \mathbb{1}(Q_{a,s}) = \text{Card}\left(\{s : \mathbb{1}(Q_{a,s}) = 1\}\right) = \text{Card}\left(\bigcup_{b \in B_n(a,\zeta_n)} \varphi(b)\right)$$
$$\leq \sum_{b \in B_n(a,\zeta_n)} \text{Card}(\varphi(b)) \leq 2\,\text{Card}(B_n(a, \zeta_n)) \leq 2\bar{B}$$

Then continuing the display above gives

$$M_n^2 n^{-1} \sum_{a=1}^{|L_n|} k_a^{-1}(k_a^2 - k_a) \sum_{s=1}^{n} \mathbb{1}(Q_{a,s}) \leq M_n^2 \overline{k}_n n^{-1} \sum_{a=1}^{|L_n|} 2 \operatorname{Card}(B_n(a, \zeta_n))$$
$$\leq 2M_n^2 \bar{B} |L_n| \overline{k}_n n^{-1} = O_p(|L_n| \overline{k}_n n^{-1})$$

Putting this all together, we have shown that

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

This completes the proof. □

9.1.2 Proof of Proposition 8.6

Proof. Suppose that $\psi(X, \zeta_n) \in [0,1]^{d_\psi}$, ζ_n-a.s. and define $\psi_{i,n} = \psi(X_i, \zeta_n)$. Fix $m > 1$ an integer. We call $(B_l)_{l=1}^{m^{d_\psi}}$ an *ordered block partition* of $[0,1]^{d_\psi}$ if (1) $\operatorname{diam}(B_l, \|\cdot\|_\infty) \leq 1/m$ 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 $l(i) = \min_{l=1}^{m^{d_\psi}} \{l : \psi_i \in B_l\}$.

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 π_n). This process results in at most one partially filled group (with less than k_a units), say $g_{a,s'}$. Increment $l \to l+1$, adding units to the group until $\operatorname{Card}(g_{a,s'}) = k_a$. 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*

$$l(i) \leq l(j) \quad \forall i \in g_{a,s}, j \in g_{a,s'} \quad s < s' \quad a = 1, \ldots, |L_n| \tag{9.1}$$

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

$$n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a^{-1} \sum_{\substack{i,j \in g_{a,s} \\ i \neq j}} |\psi_{i,n} - \psi_{j,n}|_2^2 = n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} k_a^{-1} \sum_{t=1}^{k_a^2 - k_a} |\psi_{i_{a,s,t},n} - \psi_{j_{a,s,t},n}|_2^2 \mathbb{1}(g_{a,s} \neq \emptyset)$$

(1) Suppose $E_{a,s,t}$ occurs. Define $d_{a,s,t} \equiv |\psi_{i_{a,s,t},n} - \psi_{j_{a,s,t},n}|_2 \leq \max_{l=1}^{m^{d_\psi}} \operatorname{diam}(B_l, |\cdot|_2) \leq$

$\sqrt{d_\psi}/m$ on this event. Then we may bound

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

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

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

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

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

Using this calculation, we have

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

The second inequality follows by the ordering property in equation 9.1 above, since for each $t=1,\ldots,k_a^2-k_a$, the intervals $([l(i_{a,s,t}),l(j_{a,s,t})])_{s=1}^n$ are non-overlapping, and there are at most m^{d_ψ} blocks. Summarizing the above work, we have shown that

$$n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}k_a^{-1}\sum_{\substack{i,j\in g_{a,s}\\ i\neq j}}|\psi_{i,n}-\psi_{j,n}|_2=n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}k_a^{-1}\sum_{t=1}^{k_a^2-k_a}d_{a,s,t}\leq\frac{\sqrt{d}_\psi}{m}+2\sqrt{d_\psi}|L_n|\overline{k}_n n^{-1}m^{d_\psi-1}$$

Note the following fact: for $x,y\in[0,1]^{d_\psi}$, $|x-y|_2^2=d_\psi\left(\frac{|x-y|_2}{\sqrt{d_\psi}}\right)^2\leq\sqrt{d}_\psi|x-y|_2$, 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

$$n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}k_a^{-1}\sum_{\substack{i,j\in g_{a,s}\\ i\neq j}}|\psi_{i,n}-\psi_{j,n}|_2^2 \leq n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}k_a^{-1}\sum_{t=1}^{k_a^2-k_a} d_{a,s,t}^2\mathbb{1}(E_{a,s,t})+\sqrt{d_\psi}d_{a,s,t}\mathbb{1}(E_{a,s,t}^c)$$

$$\leq \frac{d_\psi}{m^2}+2d_\psi|L_n|\overline{k}_n n^{-1}m^{d_\psi-1}$$

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

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

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

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

By construction, $(g_{a,s})_{a,s}$ satisfies the constraints in the programs 8.2 and 8.4. Then any groups $(g^*_{a,s})_{a,s}$ that attain the minima in 8.2 and 8.4 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 8.4, then

$$n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}k_a^{-1}\sum_{\substack{i,j\in g^*_{a,s}\\ i\neq j}}(h_n(\psi(X_i,\zeta_n),p_n(X_i,\zeta_n),\zeta_n)-h_n(\psi(X_j,\zeta_n),p_n(X_i,\zeta_n),\zeta_n))^2$$

$$= n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}k_a^{-1}\sum_{\substack{i,j\in g^*_{a,s}\\ i\neq j}}(h_n(\psi(X_i,\zeta_n),p_a,\zeta_n)-h_n(\psi(X_j,\zeta_n),p_a,\zeta_n))^2$$

$$\leq C_n(\zeta_n)n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}k_a^{-1}\sum_{\substack{i,j\in g^*_{a,s}\\ i\neq j}}|\psi_{i,n}-\psi_{j,n}|_2^2 \leq \sqrt{d_\psi}\bar{C}n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}k_a^{-1}\sum_{\substack{i,j\in g^*_{a,s}\\ i\neq j}}|\psi_{i,n}-\psi_{j,n}|_2$$

$$\leq \sqrt{d_\psi}\bar{C}n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}k_a^{-1}\sum_{\substack{i,j\in g_{a,s}\\ i\neq j}}|\psi_{i,n}-\psi_{j,n}|_2 = O\left(\left(n/(|L_n|\overline{k}_n)\right)^{-1/d_\psi}\right)$$

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. □

Proof of Proposition 8.8. 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

$$\begin{aligned}(\psi_i - \psi_{\gamma(i)})^2 &= (\psi_i - \psi_{a,s} + \psi_{a,s} - \psi_{a,\nu_a(s)} + \psi_{a,\nu_a(s)} - \psi_{\gamma(i)})^2 \\ &\leq 3(\psi_i - \psi_{a,s})^2 + 3(\psi_{a,s} - \psi_{a,\nu_a(s)})^2 + 3(\psi_{a,\nu_a(s)} - \psi_{\gamma(i)})^2\end{aligned}$$

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

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

by Jensen again. Since by construction $\gamma(i) \in g_{a,\nu_a(s)}$, the same reasoning applies to the third term above, giving

$$n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{i\in g_{a,s}}(\psi_i - \psi_{a,s})^2 + (\psi_{a,\nu_a(s)} - \psi_{\gamma(i)})^2 \leq n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}k_a^{-1}\sum_{\substack{i,j\in g_{a,s}\\ i\neq j}}(\psi_i - \psi_j)^2 = O(r_n^{\psi})$$

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 $n_a \equiv \sum_{s=1}^{n}\mathbb{1}(g_{a,s} \neq \emptyset)$. For $m_a \geq 1$, the same ordered block construction as in Proposition 8.6 can be used to construct a matching ν_a with

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

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

$$\begin{aligned}&n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{i\in g_{a,s}}(\psi_{a,s} - \psi_{a,\nu_a(s)})^2 \asymp n^{-1}\sum_{a=1}^{|L_n|}k_a n_a^{\frac{d_\psi-1}{d_\psi+1}} = n^{-1}\sum_{a=1}^{|L_n|}k_a^{\frac{2}{d_\psi+1}}(k_a n_a)^{\frac{d_\psi-1}{d_\psi+1}} \\ &\leq (\overline{k}_n^{\frac{2}{d_\psi+1}} n^{-1}|L_n|)\cdot |L_n|^{-1}\sum_{a=1}^{|L_n|}(k_a n_a)^{\frac{d_\psi-1}{d_\psi+1}} \leq \overline{k}_n^{\frac{2}{d_\psi+1}} n^{-1}|L_n|\left(|L_n|^{-1}\sum_{a=1}^{|L_n|}k_a n_a\right)^{\frac{d_\psi-1}{d_\psi+1}} \\ &\leq \overline{k}_n^{\frac{2}{d_\psi+1}} n^{-1}|L_n|(|L_n|^{-1}n)^{\frac{d_\psi-1}{d_\psi+1}} = ((|L_n|\overline{k}_n)^{-1}n)^{\frac{-2}{d_\psi+1}}\end{aligned}$$

Then the matching solving the program in Proposition 8.8 achieves this rate by optimality. This finishes the proof. □

9.2 Proofs - Asymptotics

9.2.1 Assumptions

Assumption 9.1 (Propensity Approximation)**.** *There exist deterministic functions p_n, $\widehat{p}_n$, p, and ψ with dependence*

$$p_{i,n} = p_n(X_i, \zeta_n) \qquad \widehat{p}_{i,n} = \widehat{p}_n(X_i, \zeta_n) \qquad p_i = p(X_i, \zeta_n) \qquad \psi_i = \psi(X_i, \zeta_n) \tag{9.2}$$

These satisfy

(a) $p_{i,n}, p_i, \widehat{p}_{i,n} \in (\delta, 1-\delta)$ ζ_n*-a.s.*

(b) $\|p_n(\cdot, \zeta_n) - \widehat{p}_n(\cdot, \zeta_n)\|_\infty \leq r_n^p = o(1)$, ζ_n*-a.s.*

(c) $(E[(\widehat{p}_{i,n} - p_i)^2 | \zeta_n])^{1/2} = O_p(\bar{r}_n^p)$

(d) $\mathrm{Im}(p_n) \subseteq L_n$ *for* $n \geq 1$

Assumption 9.2 (CLT)**.** *With variables defined as in assumption 9.1, for $t \in [T]$, consider $(f_{t,n})_{n\geq 1}$ deterministic functions. Define*

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

Let $\zeta_n \perp\!\!\!\perp (W_{1:n}, \pi_n, \tau)$ and suppose that

(a) $E[f_{t,n}(X, \zeta_n)^2 | \zeta_n] = O_p(1)$ *and* $E[u_{i,t,n}^4 | \zeta_n] = O_p(1)$ *and* $E[\sigma_d^2(X)^2] < \infty$

(b) $\exists\ u_{i,t} \in L_2(X_i, \zeta_n)$ *with* $E[(u_{i,t,n} - u_{i,t})^2 | \zeta_n] = o_p(1)$ *and* $E[u_{i,t}^2 u_{i,t'}^2 | \zeta_n] = O_p(1)$ $\forall t, t' \in [T]$

(c) $E[(p_i - p_i^2) u_{i,t} u_{i,t'} | \zeta_n] \xrightarrow{p} \Sigma_{tt'}$ *for* $t, t' \in [T]$

(d) There exist $c > 0$ and polynomial $h(\cdot)$ such that ζ_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) $\overline{k}_n \vee |L_n| = o\left(\left(\frac{\log n}{n}\right)^{1/2}\right)$

Assumption 9.3 (Selection)**.** *With variables defined as in assumption 9.1, consider $T_{1:n}, D_{1:n} \in \{0, 1\}$. Let $D_{1:n} \sim \mathrm{Loc_n}(\psi, p)$ with local group structure $(g_{a,s})_{a,s}$ and suppose*

(a) $T_{1:n} \sim \mathrm{Loc_n}(\psi', q)$ *and* $q = a'/k' \in (\delta, 1-\delta)$, $\gcd(a', k') = 1$

(b) $\psi'_{1:n} \in \sigma(\psi_{1:n})$*, so that* $T_{1:n} \in \sigma(\psi_{1:n}, \pi_n^t, \tau^t)$

(c) $\{i : T_i = 1\} = \bigsqcup_{a,s} g_{a,s}$ *with* $(g_{a,s})_{a,s} \in \sigma(T_{1:n}, \psi_{1:n}, p_{1:n,n}, \pi_n^d) \subseteq \sigma(\psi_{1:n}, p_{1:n,n}, \pi_n^d, \pi_n^t, \tau^t)$

9.2.2 Coupling and General CLT

Lemma 9.4 (Coupling). *Suppose assumption 9.1 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 $h_n(\psi, p_a, \zeta_n)$ satisfy the smoothness conditions in Definition 8.1 or 8.5. Define $h_{i,n} = h(\psi_i, p_{i,n}, \zeta_n)$. Let $D_{1:n} \sim \text{Loc}_n(\psi, p)$ and either $T_{1:n} = 1$ or $T_{1:n} \sim \text{Loc}_n(\psi', q)$ and satisfies assumption 9.3. If $E[\max_{i=1}^n h_{i,n}] = O(\log n)$

$$\sqrt{n}E_n[T_i(D_i - p_{i,n})f_n(X_i, \zeta_n)] = \sqrt{n}E_n[T_i(D_i - p_{i,n})u_{i,n}] + O_p\left(\left(r_n^\psi \vee \frac{\overline{k}_n|L_n|\log n}{n}\right)^{1/2}\right)$$

Proof. Let $\mathcal{F}_{x,n} = \sigma(X_{1:n}, \pi_n^d, \pi_n^t, \tau^t, \zeta_n)$. Then by assumption 9.3, $(g_{a,s})_{a,s}, (h_{i,n})_{i=1}^n \in \mathcal{F}_{x,n}$, and $\mathcal{F}_{x,n} \perp\!\!\!\perp \tau^d$. Then by Lemma 9.20 and Proposition 8.4 or Proposition 8.6 (depending on the smoothness condition)

$$\text{Var}\left(\sqrt{n}E_n[T_i(D_i - p_{i,n})h_{i,n}]|\mathcal{F}_{x,n}\right) \le n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n} k_a^{-1}\sum_{\substack{i,j\in g_{a,s}\\ i\neq j}}(h_{i,n} - h_{j,n})^2 + n^{-1}\overline{k}_n|L_n|\max_{i=1}^n h_{i,n}$$
$$\le O_p(r_n^\psi) + n^{-1}\overline{k}_n|L_n|\max_{i=1}^n h_{i,n} = O_p\left(r_n^\psi \vee \frac{\overline{k}_n|L_n|\log n}{n}\right)$$

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 9.20, so the claim follows from conditional Chebyshev (Lemma 9.16). □

Theorem 9.5 (CLT). *Impose assumptions 9.1 and 9.2. Require that either assumption 9.3 holds, or $T_{1:n} = 1$ and $q = 1$. Let $\mathcal{F}_{x,n} = \sigma(X_{1:n}, \pi_n^d, \pi_n^t, \tau^t, \zeta_n)$. Then weak convergence holds as in definition 9.13*

$$\sqrt{n}\left(E_n[T_i(D_i - p_{i,n})u_{i,t,n}]\right)_{t=1}^T \Big|\mathcal{F}_{x,n} \Rightarrow \mathcal{N}(0, \Sigma)$$

with variance matrix

$$\Sigma_{tt'} = qE[(p_i - p_i^2)u_{i,t}u_{i,t'}|\zeta_n] + o_p(1) \qquad t, t' \in [T]$$

Let $\mathcal{F}_{(x,d),n} = \sigma(X_{1:n}, \pi_n^d, \pi_n^t, \tau^d, \tau^t, \zeta_n)$. We have

$$\sqrt{n}E_n\left[\frac{T_iD_i\epsilon_i^1}{p_{i,n}} + \frac{T_i(1-D_i)\epsilon_i^0}{1-p_{i,n}}\right]\Big|\mathcal{F}_{(x,d),n} \Rightarrow \mathcal{N}\left(0,\ \operatorname*{plim}_{n\to\infty} qE\left[\frac{\sigma_1^2(X)}{p(X,\zeta_n)} + \frac{\sigma_0^2(X)}{1-p(X,\zeta_n)}\Big|\zeta_n\right]\right)$$

Proof of Theorem 9.5. The proof is an application of the martingale CLT of Proposition 9.14. Consider the following filtration structure

$$\mathcal{F}_{0,n} = \{\Omega_n, \emptyset\}$$
$$\mathcal{F}_{(a-1)n+s,n} = \sigma(X_{1:n}, \pi_n^d, \pi_n^t, \tau^t, \zeta_n, \tau_{a',1:n}^d : 1 \le a' \le a-1, \tau_{a,1:s}) \quad 1 \le s \le n; \quad 1 \le a \le |L_n|$$
$$\mathcal{F}_{|L_n|n+j,n} = \sigma(X_{1:n}, \pi_n^d, \pi_n^t, \tau^t, \zeta_n, \tau^d, \epsilon_{1:j}^d : d = 0, 1) \quad 1 \le j \le n$$

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 $\overline{u}_{i,n} = \sum_{t=1}^T \lambda_t u_{i,t,n}$. Define the martingale difference increments

$$Z_{(a-1)n+s,n} = \sum_{i \in g_{a,s}} (D_i - p_a)\overline{u}_{i,n} \qquad 1 \leq s \leq n; \quad 1 \leq a \leq |L_n|$$

$$Z_{|L_n|n+i,n} = \frac{T_i D_i \epsilon_i^1}{p_{i,n}} - \frac{T_i(1-D_i)\epsilon_i^0}{1-p_{i,n}} \quad 1 \leq i \leq n$$

We focus on the conditional statements, from which marginal convergence will follow. To apply Proposition 9.14, we need to check (1) martingale difference condition (2) variance process limit (and appropriate measurability) and (3) conditional Lindberg. We check each in turn.

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

$$(A, B) \perp\!\!\!\perp C \implies A \perp\!\!\!\perp C|B \tag{9.3}$$

(1) Let $1 \leq s \leq n$ and $1 \leq a \leq |L_n|$. By the representation in Lemma 9.19.(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

$$\begin{aligned}
&E[D_i \mathbb{1}(i \in g_{a,s})|\mathcal{F}_{(a-1)n+(s-1),n}] \\
&= E[G_{i,n}(\psi_{1:n}, p_{1:n,n}, \pi^d_n, \pi^t_n, \tau^t, \tau^d_{a,s})|X_{1:n}, \pi^d_n, \pi^t_n, \tau^t, \zeta_n, \tau_{a',1:n} : 1 \leq a' \leq a-1, \tau_{a,1:(s-1)}] \\
&= E[G_{i,n}(\psi_{1:n}, p_{1:n,n}, \pi^d_n, \pi^t_n, \tau^t, \tau^d_{a,s})|X_{1:n}, \pi^d_n, \pi^t_n, \tau^t, \zeta_n] \\
&= E[D_i \mathbb{1}(i \in g_{a,s})|X_{1:n}, \pi^d_n, \pi^t_n, \tau^t, \zeta_n] = E[D_i \mathbb{1}(i \in g_{a,s})|\mathcal{F}_{x,n}] = p_a \mathbb{1}(i \in g_{a,s})
\end{aligned}$$

The final equality follows by Lemma 9.19.(ii) since $\mathcal{F}_{x,n} \perp\!\!\!\perp \tau^d$ and $(g_{a,s})_{a,s} \in \mathcal{F}_{x,n}$ by assumptions 9.3 and 9.1. Using $\overline{u}_{i,n} \in \sigma(X_{1:n}, \zeta_n)$ and this calculation, $E[Z_{(a-1)n+(s-1),n}|\mathcal{F}_{(a-1)n+(s-1),n}]$ is equal to

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

(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 $h : \mathbb{R}^2 \to \mathbb{R}$ be a fixed function and note that

$$\begin{aligned}
E[h(\epsilon^1_j, \epsilon^0_j)|\mathcal{F}_{n|L_n|+(j-1),n}] &= E[h(\epsilon^1_j, \epsilon^0_j)|X_{1:n}, \pi^d_n, \pi^t_n, \tau^t, \zeta_n, \tau^d, \epsilon^{0,1}_{1:(j-1)}] \\
&= E[h(\epsilon^1_j, \epsilon^0_j)|X_{1:n}, \epsilon^{0,1}_{1:(j-1)}] = E[h(\epsilon^1_j, \epsilon^0_j)|X_j]
\end{aligned}$$

This follows by making repeated use of 9.3, 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 $(Z_{j,n})_{j=1}^{n(|L_n|+1)}$ 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 $\mathbb{1}(i \in g_{a,s}), u_{i,n} \in \mathcal{F}_{x,n} \subseteq \mathcal{F}_{(a-1)n+(s-1),n}$. Similarly to the argument in (1) above, we have $E[Z^2_{(a-1)n+s,n}|\mathcal{F}_{(a-1)n+(s-1),n}] = E[Z^2_{(a-1)n+s,n}|\mathcal{F}_{x,n}]$. Since $(g_{a,s})_{a,s} \in \mathcal{F}_{x,n}$ and $\mathcal{F}_{x,n} \perp\!\!\!\perp \tau^d$, Lemma 9.19.(iii) applies, giving covariances

$$n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n} E[Z^2_{(a-1)n+s,n}|\mathcal{F}_{(a-1)n+(s-1),n}] = n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\frac{q_a(k_a-q_a)}{k_a^2}\sum_{i\in g_{a,s}}^{n}\overline{u}^2_{i,n}$$
$$- n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n-1}\frac{q_a(k_a-q_a)}{k_a^2(k_a-1)}\sum_{\substack{i,j\in g_{a,s}\\ i\neq j}}^{n}\overline{u}_{i,n}\overline{u}_{j,n} \equiv A_n + B_n$$

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

$$p_{i,n} - p_{i,n}^2 - (p_i - p_i^2) = (p_{i,n} - p_i)[1 - (p_i + p_{i,n})] = (p_{i,n} - \widehat{p}_{i,n} + \widehat{p}_{i,n} - p_i)[1 - (p_i + p_{i,n})]$$

Then we have

$$A_n = n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{i\in g_{a,s}}^{n} p_{i,n}(1-p_{i,n})\overline{u}^2_{i,n} = E_n[T_i p_{i,n}(1-p_{i,n})\overline{u}^2_{i,n}]$$
$$= E_n[T_i p_i(1-p_i)\overline{u}^2_{i,n}] + E_n[T_i(p_{i,n} - \widehat{p}_{i,n} + \widehat{p}_{i,n} - p_i)[1-(p_i+p_{i,n})]\overline{u}^2_{i,n}]$$

Consider the second term. First note that

$$E[|E_n[T_i(p_{i,n} - \widehat{p}_{i,n})[1-(p_i+p_{i,n})]\overline{u}^2_{i,n}]||\zeta_n] \leq E[|p_n - \widehat{p}_n|_\infty \overline{u}^2_{i,n}|\zeta_n]$$
$$\leq r_n^p \cdot T\sum_{t\in[T]}\lambda_t^2 E[u^2_{i,t,n}|\zeta_n] \leq r_n^p \cdot T\sum_{t\in[T]}\lambda_t^2 E[f_{t,n}(X_i,\zeta_n)^2|\zeta_n] = O(r_n^p)$$

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

$$E[|E_n[T_i(\widehat{p}_{i,n} - p_i)[1-(p_i+p_{i,n})]\overline{u}^2_{i,n}]||\zeta_n] \leq E[|\widehat{p}_{i,n} - p_i|\overline{u}^2_{i,n}|\zeta_n]$$
$$\leq E[(\widehat{p}_{i,n} - p_i)^2|\zeta_n]^{1/2} E[\overline{u}^4_{i,n}|\zeta_n]^{1/2} = o_p(\bar{r}_n^p)O_p(1)$$

Summarizing, we have shown that $A_n = E_n[T_i p_i(1-p_i)\overline{u}^2_{i,n}] + O_p(r_n^p) + O_p(\bar{r}_n^p)$.

Define the population residual sum $\overline{u}_i = \sum_{t=1}^T \lambda_t u_{i,t}$. Claim that

$$E_n[T_i p_i(1-p_i)(\overline{u}^2_{i,n} - \overline{u}_i^2)] = O_p\left(\max_{t\in[T]} E[(u_{i,t,n} - u_{i,t})^2|\zeta_n]^{1/2}\right)$$

Note that we have

$$E\left[|E_n[T_i p_i(1-p_i)(\overline{u}_{i,n}^2 - \overline{u}_i^2)]||\zeta_n\right] \leq E[|p_i(1-p_i)(\overline{u}_{i,n}^2 - \overline{u}_i^2)||\zeta_n] \leq (1/4)E[|(\overline{u}_{i,n}^2 - \overline{u}_i^2)||\zeta_n]$$
$$\lesssim E[(\overline{u}_{i,n} - \overline{u}_i)^2|\zeta_n]^{1/2} E[(\overline{u}_{i,n} + \overline{u}_i)^2|\zeta_n]^{1/2} \leq E[(\overline{u}_{i,n} - \overline{u}_i)^2]^{1/2} \left(E[\overline{u}_{i,n}^2|\zeta_n]^{1/2} + E[\overline{u}_i^2|\zeta_n]^{1/2}\right)$$
$$\leq E[(\overline{u}_{i,n} - \overline{u}_i)^2|\zeta_n]^{1/2} O_p(1) \lesssim \sum_{t=1}^{T} |\lambda_t| E[(u_{i,t,n} - u_{i,t})^2|\zeta_n]^{1/2}$$

The third inequality uses $a^2 - b^2 = (a-b)(a+b)$ and Cauchy-Schwarz. For the fifth inequality, observe that

$$E[\overline{u}_{i,n}^2|\zeta_n]^{1/2} \leq \sum_{t\in[T]} |\lambda_t| E[u_{i,t,n}^2|\zeta_n]^{1/2} \leq \sum_{t\in[T]} |\lambda_t| E[f_{t,n}(X_i,\zeta_n)^2|\zeta_n]^{1/2} = O_p(1)$$
$$E[\overline{u}_i^2|\zeta_n]^{1/2} \leq \sum_{t\in[T]} |\lambda_t| E[u_{i,t}^2|\zeta_n]^{1/2} = O_p(1)$$

The second inequality follows by properties of projection. The third inequality is triangle inequality. The claim then follows from conditional Markov inequality (Lemma 9.16). Above work has shown that

$$A_n = E_n[T_i p_i(1-p_i)\overline{u}_i^2] + O_p(r_n^p) + O_p(\bar{r}_n^p) + O_p\left(\max_{t\in[T]} E[(u_{i,t,n} - u_{i,t})^2|\zeta_n]^{1/2}\right)$$
$$= \sum_{t,t'=1}^{T} \lambda_t \lambda_{t'} E_n[T_i p_i(1-p_i) u_{i,t} u_{i,t'}] + o_p(1)$$
$$= \sum_{t,t'=1}^{T} \lambda_t \lambda_{t'} E_n[(T_i - q) p_i(1-p_i) u_{i,t} u_{i,t'}] + q \sum_{t,t'=1}^{T} \lambda_t \lambda_{t'} E_n[p_i(1-p_i) u_{i,t} u_{i,t'}] + o_p(1)$$

If $T_{1:n} = 1$ and $q = 1$, the first term is identically 0. Otherwise, by assumption 9.3 we have $T_{1:n} \sim \text{Loc}_n(\psi', q)$. Then Lemma 9.20 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 $E[u_{i,t}^2 u_{i,t'}^2|\zeta_n] < \infty$, which we assume.

For the second term, note that $E[E_n[p_i(1-p_i)u_{i,t}u_{i,t'}]|\zeta_n] = E[p_i(1-p_i)u_{i,t}u_{i,t'}|\zeta_n]$ and

$$\text{Var}\left(E_n[p_i(1-p_i)u_{i,t}u_{i,t'}]|\zeta_n\right) = n^{-2} \sum_{i\neq j}^{n} \text{Cov}\left(p_i(1-p_i)u_{i,t}u_{i,t'}, p_j(1-p_j)u_{j,t}u_{j,t'}|\zeta_n\right)$$
$$+ n^{-1} E_n\left[\text{Var}(p_i(1-p_i)u_{i,t}u_{i,t'}|\zeta_n)\right] \leq (4n)^{-1} E[u_{i,t}^2 u_{i,t'}^2|\zeta_n] = O_p(n^{-1})$$

For the inequality, note that $p_i(1-p_i)u_{i,t}u_{i,t'} \perp\!\!\!\perp p_j(1-p_j)u_{j,t}u_{j,t'}|\zeta_n$ 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 9.16 and summarizing the above work, we have shown that

$$A_n = q \sum_{t,t'=1}^{T} \lambda_t \lambda_{t'} E[(p - p^2)(X_i) u_{i,t} u_{i,t'}|\zeta_n] + O_p(n^{-1/2}) + o_p(1)$$

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 9.3. Applying Lemma 9.22 on residual selection with functions $F_{i,n}(X_i,\zeta_n) = \sum_{t\in[T]} \lambda_t f_{t,n}(X_i,\zeta_n)$ and random element $\xi_n = (\pi_n^d, \pi_n^t, \tau^t)$ gives the following moments for $i,j,l,f \in [n]$ distinct indices

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

We apply these facts, computing

$$E[B_n|\mathcal{F}_{(\psi,p_n),n}] = E\left[n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n-1}\frac{q_a(k_a-q_a)}{k_a^2(k_a-1)}\sum_{i\neq j}^{n}\mathbb{1}(i,j\in g_{a,s})\overline{u}_{i,n}\overline{u}_{j,n}\middle|\mathcal{F}_{(\psi,p_n),n}\right]$$
$$= n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n-1}\frac{q_a(k_a-q_a)}{k_a^2(k_a-1)}\sum_{i\neq j}^{n}\mathbb{1}(i,j\in g_{a,s})E[\overline{u}_{i,n}\overline{u}_{j,n}|\mathcal{F}_{(\psi,p_n),n}] = 0$$

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

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

Case 1: $\{i,j\}\cap\{l,f\} = \emptyset$ Then $E[\overline{u}_{i,n}\overline{u}_{j,n}\overline{u}_{l,n}\overline{u}_{f,n}|\mathcal{F}_{(\psi,p_n),n}] = 0$. 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 $E[\overline{u}_{i,n}^2\overline{u}_{j,n}\overline{u}_{f,n}|\mathcal{F}_{(\psi,p_n),n}] = 0$.
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

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

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

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

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

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 $k_a w_a^2 = \frac{k_a}{k_a-1}p_a(1-p_a) \leq$

$2 \cdot (1/4)$ for $k_a \geq 2$. For the final equality, observe that

$$E[n^{-1}E_n[T_i E[\overline{u}_{i,n}^2|\psi_i, p_{i,n}, \zeta_n]^2]|\zeta_n] \leq E[E[\overline{u}_{i,n}^2|\psi_i, p_{i,n}, \zeta_n]^2|\zeta_n] \leq E[\overline{u}_{i,n}^4|\zeta_n]$$
$$\leq T^3 \sum_{t\in[T]} E[u_{i,t,n}^4|\zeta_n] = O_p(1)$$

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

$$E[Z_{|L_n|n+i,n}^2|\mathcal{F}_{|L_n|n+(i-1),n}] = E\left[\frac{T_i D_i(\epsilon_i^1)^2}{p_{i,n}^2} + \frac{T_i(1-D_i)(\epsilon_i^0)^2}{(1-p_{i,n})^2}\middle|\mathcal{F}_{|L_n|n+(i-1),n}\right]$$
$$= \frac{T_i D_i E[(\epsilon_i^1)^2|X_i]}{p_{i,n}^2} + \frac{T_i(1-D_i)E[(\epsilon_i^0)^2|X_i]}{(1-p_{i,n})^2} = \frac{T_i D_i \sigma_1^2(X_i)}{p_{i,n}^2} + \frac{T_i(1-D_i)\sigma_0^2(X_i)}{(1-p_{i,n})^2}$$

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

$$n^{-1}\sum_{i=1}^{n} E[Z_{|L_n|n+i,n}^2|\mathcal{F}_{|L_n|n+(i-1),n}] = n^{-1}\sum_{i=1}^{n} \frac{T_i D_i \sigma_1^2(X_i)}{p_{i,n}^2} + \frac{T_i(1-D_i)\sigma_0^2(X_i)}{(1-p_{i,n})^2}$$
$$= E_n\left[T_i D_i\left(\frac{\sigma_1^2(X_i)}{p_{i,n}^2} - \frac{\sigma_0^2(X_i)}{(1-p_{i,n})^2}\right)\right] + E_n\left[\frac{T_i\sigma_0^2(X_i)}{(1-p_{i,n})^2}\right]$$
$$= E_n\left[T_i(D_i - p_{i,n})\left(\frac{\sigma_1^2(X_i)}{p_{i,n}^2} - \frac{\sigma_0^2(X_i)}{(1-p_{i,n})^2}\right)\right] + E_n\left[T_i\left(\frac{\sigma_1^2(X_i)}{p_{i,n}} + \frac{\sigma_0^2(X_i)}{(1-p_{i,n})}\right)\right]$$
$$= E_n\left[T_i(D_i - p_{i,n})\left(\frac{\sigma_1^2(X_i)}{p_{i,n}^2} - \frac{\sigma_0^2(X_i)}{(1-p_{i,n})^2}\right)\right] + E_n\left[(T_i - q)\left(\frac{\sigma_1^2(X_i)}{p_{i,n}} + \frac{\sigma_0^2(X_i)}{(1-p_{i,n})}\right)\right]$$
$$+ qE_n\left[\left(\frac{\sigma_1^2(X_i)}{p_{i,n}} + \frac{\sigma_0^2(X_i)}{(1-p_{i,n})}\right)\right] \equiv A_n + B_n + C_n$$

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

Consider A_n. Let $v_n(X_i) \equiv \frac{\sigma_1^2(X_i)}{p_{i,n}^2} - \frac{\sigma_0^2(X_i)}{(1-p_{i,n})^2}$. By Lemma 9.20 we have $A_n = O_p(n^{-1/2})$ if $\sup_{n\geq 1} E[v_n(X_i)^2] < \infty$. To see this, note that

$$E[v_n(X_i)^2] \leq E\left[2\frac{\sigma_1^2(X_i)^2}{p_{i,n}^4} + 2\frac{\sigma_0^2(X_i)^2}{(1-p_{i,n})^4}\right] \leq 2\delta^{-4}E[\sigma_1^2(X)^2 + \sigma_0^2(X)^2] < \infty$$

The first inequality is Young's, then using propensity score bound and assumption. If $T_{1:n} = 1$ and $q = 1$, then $B_n = 0$. Otherwise, by assumption 9.3 we have $T_{1:n} \sim \text{Loc}_n(\psi', q)$, then $B_n = O_p(n^{-1/2})$ by Lemma 9.20 and an identical argument.

Next consider C_n. We can calculate

$$\begin{aligned}\left|E_n\left[\sigma_1^2(X_i)\left(\frac{1}{p_{i,n}}-\frac{1}{p_i}\right)\right]\right| &= \left|E_n\left[\sigma_1^2(X_i)\frac{p_i-p_{i,n}}{p_i p_{i,n}}\right]\right| \leq \delta^{-2}E_n\left[\sigma_1^2(X_i)|p_i-\widehat{p}_{i,n}+\widehat{p}_{i,n}-p_{i,n}|\right] \\ &\lesssim E_n\left[\sigma_1^2(X_i)|p_i-\widehat{p}_{i,n}|\right]+r_n^p E_n\left[\sigma_1^2(X_i)\right] \\ &= E_n\left[\sigma_1^2(X_i)|p_i-\widehat{p}_{i,n}|\right]+O_p(r_n^p)\end{aligned}$$

Now taking expectations, note that

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

Then by Lemma 9.16, we have $E_n\left[\sigma_1^2(X_i)|p_i-\widehat{p}_{i,n}|\right] = O_p(\bar{r}_n^p)$. The analysis for the term involving $\sigma_0^2(X_i)$ is identical. Then by WLLN, we have shown that

$$\begin{aligned}C_n &= qE_n\left[\frac{\sigma_1^2(X_i)}{p_i}+\frac{\sigma_0^2(X_i)}{(1-p_i)}\right]+O_p(r_n^p\vee\bar{r}_n^p) \\ &= qE\left[\frac{\sigma_1^2(X_i)}{p_i}+\frac{\sigma_0^2(X_i)}{(1-p_i)}\middle|\zeta_n\right]+O_p(n^{-1/2})+o_p(1)\end{aligned}$$

Summarizing the work above, the conditional variance process is

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

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

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

$$\begin{aligned}&n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^n E\left[Z_{(a-1)n+s,n}^2\mathbb{1}(|Z_{(a-1)n+s,n}|>\sqrt{n}\epsilon)|\mathcal{F}_{x,n}\right] \\ &= n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^n E\left[\left(\sum_{i\in g_{a,s}}(d_i-p_a)\overline{u}_{i,n}\right)^2\mathbb{1}\left(\left(\sum_{i\in g_{a,s}}(d_i-p_a)\overline{u}_{i,n}\right)^2>n\epsilon^2\right)\middle|\mathcal{F}_{x,n}\right] \\ &\leq n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^n |g_{a,s}|E\left[\sum_{i\in g_{a,s}}(d_i-p_a)^2\overline{u}_{i,n}^2\mathbb{1}\left(|g_{a,s}|\sum_{i\in g_{a,s}}(d_i-p_a)^2\overline{u}_{i,n}^2>n\epsilon^2\right)\middle|\mathcal{F}_{x,n}\right]\end{aligned}$$

Continuing, this is

$$\begin{aligned}&\leq \overline{k}_n n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^n E\left[\sum_{i\in g_{a,s}}\overline{u}_{i,n}^2\mathbb{1}\left(\sum_{i\in g_{a,s}}\overline{u}_{i,n}^2>n\epsilon^2\overline{k}_n^{-1}\right)\middle|\mathcal{F}_{x,n}\right] \\ &\leq T\overline{k}_n n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^n E\left[\sum_{i\in g_{a,s}}\sum_{t\in[T]}u_{i,t,n}^2\mathbb{1}\left(\sum_{i\in g_{a,s}}\sum_{t\in[T]}u_{i,t,n}^2>n\epsilon^2(\overline{k}_n T)^{-1}\right)\middle|\mathcal{F}_{x,n}\right]\end{aligned}$$

The first and third inequalities follow because $\left(\sum_{i\in I} a_i\right)^2 \le |I| \sum_{i\in I} a_i^2$ for any scalar array $(a_i)_{i\in I}$ by Jensen inequality. The second inequality uses $|d_i - p_a| \le \max(p_a, 1-p_a) \le 1$ and $|g_{a,s}| \le \overline{k}_n$ by definition. Now note that for any *positive* scalar array $(a_i)_{i\in I}$ we have

$$\sum_{i\in I} a_i \cdot \mathbb{1}\left(\sum_{i\in I} a_i > c\right) \le |I| \sum_{i\in I} a_i \cdot \mathbb{1}\left(a_i > c/|I|\right) \tag{9.4}$$

To see this, note that in the case $a_j = \max_{i\in I} a_i$, the LHS above is bounded by $|I|a_j \mathbb{1}(|I|a_j > c)$. 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

$$\begin{aligned}
&T\overline{k}_n n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} E\left[\sum_{i\in g_{a,s}} \sum_{t\in[T]} u_{i,t,n}^2 \mathbb{1}\left(\sum_{i\in g_{a,s}} \sum_{t\in[T]} u_{i,t,n}^2 > n\epsilon^2(\overline{k}_n T)^{-1}\right) \middle| \mathcal{F}_{x,n}\right] \\
&\le T\overline{k}_n n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} E\left[\overline{k}_n T \sum_{i\in g_{a,s}} \sum_{t\in[T]} u_{i,t,n}^2 \mathbb{1}\left(u_{i,t,n}^2 > n\epsilon^2(\overline{k}_n T)^{-2}\right) \middle| \mathcal{F}_{x,n}\right] \\
&= (T\overline{k}_n)^2 n^{-1} \sum_{t\in[T]} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} \sum_{i\in g_{a,s}} E\left[u_{i,t,n}^2 \mathbb{1}\left(u_{i,t,n}^2 > n\epsilon^2(\overline{k}_n T)^{-2}\right) \middle| \mathcal{F}_{x,n}\right]
\end{aligned}$$

Continuing using $\{i : T_i = 1\} = \bigsqcup_{a,s} g_{a,s}$ from assumption 9.3

$$\begin{aligned}
&= (T\overline{k}_n)^2 \sum_{t\in[T]} E_n\left[T_i E\left[u_{i,t,n}^2 \mathbb{1}\left(u_{i,t,n}^2 > n\epsilon^2(\overline{k}_n T)^{-2}\right) \middle| \mathcal{F}_{x,n}\right]\right] \\
&\le (T\overline{k}_n)^2 \sum_{t\in[T]} E_n\left[E\left[u_{i,t,n}^2 \mathbb{1}\left(u_{i,t,n}^2 > n\epsilon^2(\overline{k}_n T)^{-2}\right) \middle| \mathcal{F}_{x,n}\right]\right]
\end{aligned}$$

The third equality follows since $g_{a,s} \in \mathcal{F}_{x,n}$ for all a, s. Taking an expectation of the final upper bound gives, using linearity and tower law gives

$$\begin{aligned}
&(T\overline{k}_n)^2 \sum_{t\in[T]} E\left[u_{i,t,n}^2 \mathbb{1}\left(u_{i,t,n}^2 > n\epsilon^2(\overline{k}_n T)^{-2}\right) | \zeta_n\right] \lesssim (\overline{k}_n)^2 h(n\epsilon^2(\overline{k}_n T)^{-2}) \exp\left(-cn\epsilon^2(\overline{k}_n T)^{-2}\right) \\
&\lesssim (\overline{k}_n)^2 \exp\left(-(1/2)cn\epsilon^2(\overline{k}_n T)^{-2}\right) = \exp\left(2\log(\overline{k}_n)\left(\frac{-(1/2)cn\epsilon^2}{2\log(\overline{k}_n)(\overline{k}_n T)^2} + 1\right)\right) = o(1)
\end{aligned}$$

The first inequality is by assumption. The second inequality follows using that f is polynomial, so that $f(s_n)\exp(-(1/2)cs_n) = o(1)$ for any $s_n \to \infty$, and $n(\overline{k}_n)^{-2} \to \infty$ by assumption. The last line follows since $n^{-1}\log(\overline{k}_n)(\overline{k}_n)^2 = o(1)$ by assumption. By Markov inequality and the chain of upper bounds, we have shown that

$$n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} E\left[Z_{(a-1)n+s,n}^2 \mathbb{1}(|Z_{(a-1)n+s,n}| > \sqrt{n}\epsilon) | \mathcal{F}_{x,n}\right] = o_p(1)$$

(2) Let $1 \leq i \leq n$ and consider the bound

$$Z^2_{n|L_n|+i,n} = T_i \left(\frac{D_i \epsilon_i^1}{p_{i,n}} - \frac{(1-D_i)\epsilon_i^0}{1-p_{i,n}} \right)^2 = \frac{T_i D_i (\epsilon_i^1)^2}{p_{i,n}^2} + \frac{T_i (1-D_i)(\epsilon_i^0)^2}{(1-p_{i,n})^2}$$
$$\leq \max(p_{i,n}^{-1}, (1-p_{i,n})^{-1})^2 \sum_{d=0,1} (\epsilon_i^d)^2 \leq \delta^{-2} \sum_{d=0,1} (\epsilon_i^d)^2$$

Then we have

$$\begin{aligned}
& n^{-1} \sum_{i=1}^n E\left[Z^2_{|L_n|n+i,n} \mathbb{1}(|Z_{|L_n|n+i,n}| > \sqrt{n}\epsilon) | \mathcal{F}_{(x,d),n} \right] \\
& \leq n^{-1} \sum_{i=1}^n E\left[\delta^{-2} \sum_{d=0,1} (\epsilon_i^d)^2 \mathbb{1}\left(\delta^{-2} \sum_{d=0,1} (\epsilon_i^d)^2 > n\epsilon^2 \right) \middle| X_{1:n}, \pi_n^d, \pi_n^t, \tau^d, \tau^t, \zeta_n \right] \\
& = n^{-1} \sum_{i=1}^n E\left[\delta^{-2} \sum_{d=0,1} (\epsilon_i^d)^2 \mathbb{1}\left(\delta^{-2} \sum_{d=0,1} (\epsilon_i^d)^2 > n\epsilon^2 \right) \middle| X_{1:n} \right] \\
& = n^{-1} \sum_{i=1}^n E\left[\delta^{-2} \sum_{d=0,1} (\epsilon_i^d)^2 \mathbb{1}\left(\delta^{-2} \sum_{d=0,1} (\epsilon_i^d)^2 > n\epsilon^2 \right) \middle| X_i \right]
\end{aligned}$$

Consider the expectation of the final term, which is

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

The inequality holds for any sequence $M(n) > \delta^2 n\epsilon^2$. The RHS is $o(1)$ as $n \to \infty$ 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 $n^{-1} \sum_{i=1}^n E\left[Z^2_{|L_n|n+i,n} \mathbb{1}(|Z_{|L_n|n+i,n}| > \sqrt{n}\epsilon) | \mathcal{F}_{(x,d),n} \right] = o_p(1)$ by Markov inequality.

The MDS condition, variance limits and measurability, and Lindberg conditions in probability (conditional on $\mathcal{F}_{x,n}$ and $\mathcal{F}_{(x,d),n}$ respectively) are sufficient to invoke Proposition 9.14, which proves the first two claims.

□

Assumption 9.6 (CLT I)**.** *Suppose $p = a/k$. In this context, define*

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

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

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

for all $n \geq 1$, $z \geq 0$

(c) $F^d(\psi) = E[m_d(X_i)|\psi_i = \psi]$ *satisfies the smoothness conditions in either Definition 8.1 or 8.5. Also,* $E[\max_{i=1}^n |F_i^d|] = O(\log n)$.

Assumption 9.7 (CLT II)**.** *Require Assumption 9.1 with* $\zeta_n = \zeta$ *for all* $n \geq 1$ *and* $\widehat{p}_{i,n} = p_i = p(X_i, \zeta)$. *Assume* $\psi_i' = \psi'(X_i, \zeta) \in \sigma(\psi_i)$ *for* $\psi_i = \psi(X_i, \zeta)$. *In this context, define the following*

$$\begin{aligned} F_{i,n}^d &= E[m_d(X_i)|p_{i,n}, \psi_i, \zeta] & u_{i,n}^d &= m_d(X_i) - F_{i,n}^d \\ F_i^d &= E[m_d(X_i)|p_i, \psi_i, \zeta] & u_i^d &= m_d(X_i) - F_i^d \\ \tilde{F}_i^d &= E[m_d(X_i)|\psi_i', \zeta] & \tilde{u}_i^d &= m_d(X_i) - \tilde{F}_i^d \end{aligned}$$

For $d = 0, 1$, *require the following conditions*

(a) $E[m_d(X)^2] < \infty$ *and* $E[\sigma_d^2(X)^2] < \infty$

(b) $E[(u_{i,n}^d - u_i^d)^2|\zeta] = o_p(1)$.

(c) $E[(u_{i,n}^d)^4|\zeta] < \infty$, $E[(u_i^d)^4|\zeta] < \infty$, *and* $E[(\tilde{u}_i^d)^4|\zeta] < \infty$, ζ*-a.s.*

(d) There exist $c > 0$ *and polynomial* $h(\cdot)$ *such that* ζ*-a.s.*

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

for all $n \geq 1$, $z \geq 0$

(e) $F_n^d(\psi, p_a, \zeta)$ *and* $\tilde{F}^d(\psi', \zeta)$ *satisfy the smoothness conditions in either Definition 8.1 or 8.5. Also,* $E[\max_{i=1}^n |F_{i,n}^d|] = O(\log n)$ *and* $E[\max_{i=1}^n |\tilde{F}_i^d|] = O(\log n)$

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

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 9.8 (Normalization)**.** *Under the assumptions of Theorem 3.17*

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

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

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

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

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

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 9.5 define

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

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

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

We start by verifying Assumption 9.2 for these variables, justifying application of the general CLT in Theorem 9.5. Observe that

$$E[f^2_{1,n}|\zeta] \le 2\delta^{-2}E[m_1^2(X_i)+m_0^2(X_i)|\zeta] \lesssim E[m_1^2(X_i)+m_0^2(X_i)] < \infty$$
$$E[u^4_{i,n}|\zeta] \le 8\delta^{-4}E[(u^1_{i,n})^4+(u^0_{i,n})^4|\zeta] < \infty$$

In both lines, the first inequality follows by Young's inequality and our propensity bound. This shows assumption 9.2.(a).

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

$$\begin{aligned}\left|\frac{u^1_i}{p_i} - \frac{u^1_{i,n}}{p_{i,n}}\right| &\le \delta^{-2}\left|u^1_i p_{i,n} - u^1_{i,n}p_i\right| \lesssim |p_{i,n}-p_i||u^1_i| + |u^1_i - u^1_{i,n}| \\ &\le r^p_n|u^1_i| + |p_i||u^1_i - u^1_{i,n}|\end{aligned}$$

The first inequality by our propensity bound. The second inequality by telescoping. The final inequality follows since $|\widehat{p}_n - p_n|_\infty \le r^p_n$ by assumption 9.1 and since we let $p = \widehat{p}_n$ known. Using this gives

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

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 9.2.(b). Define the target assignment variance

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

Then assumption 9.2.(c) is trivially satisfied since $\zeta_n = \zeta\ \forall n \ge 1$ is constant. By Young's

inequality and propensity bound, $u_{i,n}^2 \leq 2\delta^{-2}((u_{i,n}^1)^2 + (u_{i,n}^0)^2)$. Then we have

$$\begin{aligned} E[(u_{i,n})^2 \mathbb{1}((u_{i,n})^2 > z)|\zeta] &\leq E[2\delta^{-2}((u_{i,n}^1)^2 + (u_{i,n}^0)^2)\mathbb{1}(2\delta^{-2}((u_{i,n}^1)^2 + (u_{i,n}^0)^2) > z)|\zeta] \\ &\lesssim E[((u_{i,n}^1)^2 + (u_{i,n}^0)^2)\mathbb{1}(((u_{i,n}^1)^2 + (u_{i,n}^0)^2) > (1/2)\delta^2 z)|\zeta] \\ &\leq 2\sum_{d=0,1} E[(u_{i,n}^d)^2 \mathbb{1}((u_{i,n}^d)^2) > (1/4)\delta^2 z)|\zeta] \end{aligned}$$

The final equality follows by fact 9.4 used in the general CLT proof. Then by our assumption on the tails of $u_{i,n}^d$, we see that assumption 9.2.(d) is satisfied, and assumption 9.2.(e) is the same. This completes the verification of assumptions needed for Theorem 9.5.

Let $\mathcal{F}_{x,n} = \sigma(X_{1:n}, \pi_n^d, \pi_n^t, \tau^t, \zeta)$ and $\mathcal{F}_{(x,d),n} = \sigma(X_{1:n}, \pi_n^d, \pi_n^t, \tau^t, \tau^d, \zeta)$. Invoking Theorem 9.5 gives weak convergence

$$\sqrt{n}E_n[T_i(D_i - p_{i,n})u_{i,n}]\big|\mathcal{F}_{x,n} \Rightarrow \mathcal{N}(0, q\Sigma_{11})$$

as well as

$$\sqrt{n}E_n\left[\frac{T_i D_i \epsilon_i^1}{p_{i,n}} + \frac{T_i(1-D_i)\epsilon_i^0}{1-p_{i,n}}\right]\Bigg|\mathcal{F}_{(x,d),n} \Rightarrow \mathcal{N}\left(0, qE\left[\frac{\sigma_1^2(X)}{p(X,\zeta)} + \frac{\sigma_0^2(X)}{1-p(X,\zeta)}\Bigg|\zeta\right]\right)$$

(2) *Selection Term*: Next, we will apply Theorem 9.5 again using only the selection variables $T_{1:n} \sim \text{Loc}_\text{n}(\psi', q)$ and different residuals. Define $\tilde{f}_{1,n}(X_i) = c(X_i) = m_1(X_i) - m_0(X_i)$ and

$$\tilde{u}_i \equiv \tilde{u}_{i,1} \equiv m_1(X_i) - m_0(X_i) - E[m_1(X_i) - m_0(X_i)|\psi_i', \zeta] \equiv \tilde{u}_i^1 - \tilde{u}_i^0$$

This has the same form as $f_{1,n}, u_{i,1,n}$ defined above, with $\tilde{u}_i$ now a linear combination of $(\tilde{u}_i^1, \tilde{u}_i^0)$ with weights $(1, -1)$ instead of $(p_{i,n}^{-1}, (1-p_{i,n})^{-1})$. Then an identical argument shows that assumption 9.2 is satisfied for this choice of $\tilde{f}_{1,n}(X_i, \zeta)$ and $\tilde{u}_i$. Then applying Theorem 9.5 with $\mathcal{F}_{x,n}^T = \sigma(X_{1:n}, \pi_n^t, \zeta)$ and switching roles $T_{1:n} \to 1$ and $D_{1:n} \to T_{1:n}$ gives weak convergence

$$\sqrt{n}E_n[(T_i - q)(\tilde{u}_i^1 - \tilde{u}_i^0)]\big|\mathcal{F}_{x,n}^T \Rightarrow \mathcal{N}(0, \Sigma^t)$$

with variance

$$\Sigma^t = E[(q - q^2)(\tilde{u}_i^1 - \tilde{u}_i^0)^2|\zeta] = q(1-q)E[(c_i - E[c_i|\psi_i', \zeta])^2|\zeta]$$

Residual Coupling: In the notation of coupling Lemma 9.4, the projection of $f_{1,n}$ is

$$\begin{aligned} h_n(\psi, p_a, \zeta') &\equiv E\left[f_{1,n}(X_i, \zeta)|\psi_i = \psi, p_{i,n} = p_a, \zeta = \zeta'\right] \\ &= E\left[\frac{b_n(X_i, \zeta)}{\sqrt{p_{i,n} - p_{i,n}^2}}\Bigg|\psi_i = \psi, p_{i,n} = p_a, \zeta = \zeta'\right] = \frac{F_n^1(\psi, p_a, \zeta')}{p_a} + \frac{F_n^0(\psi, p_a, \zeta')}{1-p_a} \end{aligned}$$

Then F_n satisfies the smoothness conditions in Definition 8.1 or 8.5 if F_n^d does for $d = 0, 1$. Similarly $\tilde{F}(\psi', \zeta') \equiv E[c(X_i)|\psi_i' = \psi', \zeta = \zeta'] = F^1(\psi', \zeta') - F^0(\psi', \zeta')$ satisfies the

smoothness conditions. Under our assumptions (including the maximal inequalities for $\tilde{F}_i^d$ and $F_{i,n}^d$), Lemma 9.4 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

$$\begin{aligned} q\widehat{\theta} &= qE_n\left[\frac{T_i(D_i - p_{i,n})Y_i}{q(p_{i,n} - p_{i,n}^2)}\right] = E_n\left[T_i c(X_i)\right] + E_n\left[\frac{T_i(D_i - p_{i,n})}{\sqrt{p_{i,n} - p_{i,n}^2}}b_{i,n}\right] \\ &+ E_n\left[\frac{T_i D_i \epsilon_i^1}{p_{i,n}} - \frac{T_i(1 - D_i)\epsilon_i^0}{1 - p_{i,n}}\right] = qE_n\left[c(X_i)\right] + E_n\left[(T_i - q)c(X_i)\right] \\ &+ E_n\left[\frac{T_i(D_i - p_{i,n})}{\sqrt{p_{i,n} - p_{i,n}^2}}b_{i,n}\right] + E_n\left[\frac{T_i D_i \epsilon_i^1}{p_{i,n}} - \frac{T_i(1 - D_i)\epsilon_i^0}{1 - p_{i,n}}\right] \end{aligned}$$

Using the work above gives the coupling

$$\begin{aligned} q\sqrt{n}(\widehat{\theta} - \text{ATE}) &= q\sqrt{n}E_n\left[c(X_i) - \text{ATE}\right] + \sqrt{n}E_n\left[(T_i - q)\tilde{u}_i\right] \\ &+ \sqrt{n}E_n\left[T_i(D_i - p_{i,n})u_{i,n}\right] + \sqrt{n}E_n\left[\frac{T_i D_i \epsilon_i^1}{p_{i,n}} - \frac{T_i(1 - D_i)\epsilon_i^0}{1 - p_{i,n}}\right] + o_p(1) \\ &\equiv A_n + B_n + C_n + D_n + o_p(1) \end{aligned}$$

Note that $A_n \Rightarrow \mathcal{N}(0, q^2 \operatorname{Var}(c(X))$ 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, $E[e^{itC_n}|\mathcal{F}_{x,n}] \xrightarrow{p} \phi_c(t) = e^{-\frac{1}{2}q\Sigma_{11}t^2}$. Then for any $t \in \mathbb{R}$ we have

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

Then note that $e^{it(A_n + B_n + C_n)}\left(E[e^{itC_n}|\mathcal{F}_{(x,d),n}] - \phi_d(t)\right) = o_p(1)$ and is totally bounded. Then by Lemma 9.16, $E\left[e^{it(A_n + B_n + C_n)}\left(E[e^{itD_n}|\mathcal{F}_{(x,d),n}] - \phi_d(t)\right)|\zeta\right] = o_p(1)$. Then we have

$$\begin{aligned} E[e^{itq\sqrt{n}(\widehat{\theta} - \text{ATE})}|\zeta] &= E[e^{it(A_n + B_n + C_n)}\phi_d(t)|\zeta] + o_p(1) = \phi_d(t)E[e^{it(A_n + B_n)}E[e^{itC_n}|\mathcal{F}_{x,n}]|\zeta] + o_p(1) \\ &= \phi_d(t)\phi_c(t)E[e^{itA_n}E[e^{itB_n}|\mathcal{F}_{x,n}^T]|\zeta] + o_p(1) = \phi_d(t)\phi_c(t)\phi_b(t)E[e^{itA_n}|\zeta] + o_p(1) \\ &= \phi_d(t)\phi_c(t)\phi_b(t)E[e^{itA_n}] + o_p(1) = \phi_a(t)\phi_b(t)\phi_c(t)\phi_d(t) + o_p(1) \end{aligned}$$

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 $W_{1:n} \perp\!\!\!\perp \zeta$, 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 $\sqrt{qn}(\widehat{\theta} - \text{ATE})|\zeta \Rightarrow \mathcal{N}(0, V(\psi, p))$ with the claimed limit.

Fixed Regressors: Consider that by our estimator decomposition

$$\begin{aligned} q\sqrt{n}(\widehat{\theta} - E_n[c(X_i)]) &= q\sqrt{n}(\widehat{\theta} - \text{ATE}) - q\sqrt{n}(E_n[c(X_i)] - \text{ATE}) \\ &= B_n + C_n + D_n + o_p(1) \end{aligned}$$

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

9.3 Proofs - Design with a Pilot

Assumption 9.9. *Consider $\widehat{\sigma}_d^2(x)$ 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) $\|\widehat{\sigma}_d - \sigma_d\|_{2,\mathcal{X}} = o_p(1)$ or $\|\widehat{\sigma}_d^2 - \sigma_d^2\|_{2,\mathcal{X}} = o_p(1)$

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

(d) $m_d(x)$ satisfies either Definition 8.1 or 8.5 and $E[\max_{i=1}^n m_d(X_i)] = O(\log n)$

Proof of Theorem 4.3. Define $\zeta_n = (\widehat{p}_d)_{d=0,1}$, so $\zeta_n \perp\!\!\!\perp (W_{1:n}, \pi_n^d, \pi_n^t, \tau^d, \tau^t)$. In the notation of coupling Lemma 9.4, 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

$$\begin{aligned} h_n(x, p_a, \zeta) &\equiv E\left[f_{1,n}(X_i, \zeta) | X_i = x, p_{i,n} = p_a, \zeta_n = \zeta\right] \\ &= E\left[\frac{b_n(X_i, \zeta)}{\sqrt{p_{i,n} - p_{i,n}^2}} \middle| X_i = x, p_{i,n} = p_a, \zeta_n = \zeta\right] = \frac{m_1(x)}{p_a} + \frac{m_0(x)}{1 - p_a} = h_n(x, p_a) \end{aligned}$$

By the propensity bound $p_a^{-1} \vee (1 - p_a)^{-1} < \delta^{-1}$, we see h_n satisfies the smoothness conditions in Definition 8.1 or 8.5 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

$$\begin{aligned} u_{i,n} &\equiv \frac{b_n(X_i, \zeta)}{\sqrt{p_{i,n} - p_{i,n}^2}} - h_n(X_i, p_{i,n}, \zeta_n) = \frac{b_n(X_i, \zeta)}{\sqrt{p_{i,n} - p_{i,n}^2}} - \frac{b_n(X_i, \zeta)}{\sqrt{p_{i,n} - p_{i,n}^2}} = 0 \\ u_i &\equiv c(X_i) - \tilde{h}_n(X_i, \zeta_n) = c(X_i) - c(X_i) = 0 \end{aligned}$$

Then Lemma 9.4 implies that

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

Next, we verify conditions in assumptions 9.1, 9.2 of the general CLT Theorem 9.5. 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 $(E[(\widehat{p}_{i,n} -$

$p_i)^2|\zeta_n])^{1/2} = o_p(1)$. A calculation shows that

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

Then almost surely

$$\begin{aligned}(\widehat{p}(X_i)-p^*(X_i))^2 &\le \frac{(\widehat{\sigma}_1-\sigma_1)^2(X_i)\sigma_0^2(X_i)+\sigma_1^2(X_i)(\sigma_0-\widehat{\sigma}_0)^2(X_i)}{8\underline{\sigma}^4} \\ &\le \frac{B}{8\underline{\sigma}^4}\sum_{d=0,1}(\widehat{\sigma}_d-\sigma_d)^2(X_i)\end{aligned}$$

For the case $\|\widehat{\sigma}_d-\sigma_d\|_{2,\mathcal{X}} = o_p(1)$, the claim follows immediately. Otherwise, observe that

$$\sum_{d=0,1}(\widehat{\sigma}_d-\sigma_d)^2(X_i) = \sum_{d=0,1}\frac{(\widehat{\sigma}_d^2-\sigma_d^2)^2(X_i)}{(\widehat{\sigma}_d+\sigma_d)^2(X_i)} \lesssim \sum_{d=0,1}(\widehat{\sigma}_d^2-\sigma_d^2)^2(X_i)$$

This shows that $\|\widehat{\sigma}_d^2-\sigma_d^2\|_{2,\mathcal{X}} = o_p(1)$ is also sufficient. Let $\mathcal{F}_{(x,d),n} = \sigma(X_{1:n}, \pi_n^d, \pi_n^t, \tau^d, \tau^t, \zeta_n)$. By Theorem 9.5, we have

$$\sqrt{n}E_n\left[\frac{T_iD_i\epsilon_i^1}{p_{i,n}}+\frac{T_i(1-D_i)\epsilon_i^0}{1-p_{i,n}}\right]\Big|\mathcal{F}_{(x,d),n} \Rightarrow \mathcal{N}\left(0,\ qE\left[\frac{\sigma_1^2(X)}{p^*(X)}+\frac{\sigma_0^2(X)}{1-p^*(X)}\right]\right)$$

Summarizing using the coupling above, we have shown that

$$q\sqrt{n}(\widehat{\theta}-\theta) = q\sqrt{n}E_n\left[c(X_i)-\theta\right]+\sqrt{n}E_n\left[\frac{T_iD_i\epsilon_i^1}{p_{i,n}}-\frac{T_i(1-D_i)\epsilon_i^0}{1-p_{i,n}}\right]+o_p(1)$$

The conclusion follows by repeating the argument at the end of the proof of Theorem 3.17. □

Proposition 9.10 (Consistent Regressions)**.** *Require the assumptions and definitions in Theorem 4.8. Let $p(x)$ be a fixed, known propensity score. If $T_{1:n} \sim \mathrm{Loc_n}(\widehat{\psi}, q)$ and $D_{1:n} \sim \mathrm{Loc}(\widehat{\psi}, p\,|\,T_{1:n})$ then $\sqrt{n}(\widehat{\theta}-\mathrm{ATE}) \Rightarrow \mathcal{N}(0,V)$*

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

Proof of Proposition 9.10 and Theorem 4.8. First, we consider coupling arguments. Define $\zeta_n = (\widehat{m}_d)_d$. In the notation of Lemma 9.4, let $f_n(X_i,\zeta_n) = \widehat{c}(X_i)$. Let $\psi = (\psi_1,\psi_2) \in \mathbb{R}^2$.

$$\begin{aligned}h_n(\psi,p_a,\zeta) &= E[f_n(X_i,\zeta_n)|\psi(X_i,\zeta_n)=\psi, p_{i,n}=p_a, \zeta_n=\zeta] \\ &= E\left[\widehat{c}(X_i)|(\widehat{c}(X_i),\widehat{b}_n(X_i))=(\psi_1,\psi_2), p_{i,n}=p_a,\zeta_n=\zeta\right] = \psi_1\end{aligned}$$

Then h_n is a Lipschitz function of $\psi=(\psi_1,\psi_2)$. In particular, in the notation of Lemma 9.4 $h_{i,n} = \widehat{c}(X_i)$, so $u_{i,n} = f_n(X_i,\zeta_n) - h_{i,n} = 0$. Note that

$$\max_{i=1}^n h_{i,n}^2 = \max_{i=1}^n \widehat{c}(X_i)^2 \le 2\max_d \|\widehat{m}_d\|_\infty^2 \le 2\overline{m}^2$$

Then $E[\max_{i=1}^n h_{i,n}^2] = O(1)$, so the relevant maximal inequality is satisfied. Then Lemma 9.4 implies

$$\sqrt{n}E_n[(T_i - q)\widehat{c}(X_i)] = \sqrt{n}E_n[(T_i - q)u_{i,n}] + o_p(1) = o_p(1)$$

Similarly, let $f_n'(X_i, \zeta_n) = \frac{\widehat{b}_n(X_i)}{\sqrt{p_{i,n} - p_{i,n}^2}}$ and define projection

$$h_n'(\psi, p_a, \zeta) = E\left[\frac{\widehat{b}_n(X_i)}{\sqrt{p_{i,n} - p_{i,n}^2}}\middle| (\widehat{c}(X_i), \widehat{b}_n(X_i)) = (\psi_1, \psi_2), p_{i,n} = p_a, \zeta_n = \zeta\right] = \frac{\psi_2}{\sqrt{p_a - p_a^2}}$$

Since $p_a \vee 1 - p_a \leq \delta^{-1}$ by assumption, h_n' is Lipschitz in (ψ_1, ψ_2). Again, $f_n'(X_i, \zeta_n) - h_{i,n}' = 0$ and $E[\max_{i=1}^n (h_{i,n}')^2] = O(1)$ by similar arguments. Then Lemma 9.4 implies

$$\sqrt{n}E_n\left[T_i(D_i - p_{i,n})\frac{\widehat{b}_n(X_i)}{\sqrt{p_{i,n} - p_{i,n}^2}}\right] = o_p(1)$$

Then by telescoping, the estimator expansion

$$\begin{aligned}
&q\sqrt{n}(\widehat{\theta} - \text{ATE}) = q\sqrt{n}E_n\left[c(X_i) - \text{ATE}\right] + \sqrt{n}E_n\left[(T_i - q)(c(X_i) - \widehat{c}(X_i) + \widehat{c}(X_i))\right] \\
&+ \sqrt{n}E_n\left[T_i(D_i - p_{i,n})\frac{b_{i,n} - \widehat{b}_n(X_i) + \widehat{b}_n(X_i)}{\sqrt{p_{i,n} - p_{i,n}^2}}\right] + \sqrt{n}E_n\left[\frac{T_i D_i \epsilon_i^1}{p_{i,n}} - \frac{T_i(1 - D_i)\epsilon_i^0}{1 - p_{i,n}}\right] \\
&= q\sqrt{n}E_n\left[c(X_i) - \text{ATE}\right] + \sqrt{n}E_n\left[(T_i - q)(c(X_i) - \widehat{c}(X_i))\right] \\
&+ \sqrt{n}E_n\left[T_i(D_i - p_{i,n})\frac{b_{i,n} - \widehat{b}_n(X_i)}{\sqrt{p_{i,n} - p_{i,n}^2}}\right] + \sqrt{n}E_n\left[\frac{T_i D_i \epsilon_i^1}{p_{i,n}} - \frac{T_i(1 - D_i)\epsilon_i^0}{1 - p_{i,n}}\right] + o_p(1) \\
&\equiv A_n + B_n + C_n + D_n
\end{aligned}$$

Define $\mathcal{F}_{x,n} = \sigma(X_{1:n}, \pi_n^t, \tau^t, \zeta_n)$. Then $(g_{a,s})_{a,s} \in \sigma(\widehat{\psi}_{1:n}, p_{1:n,n}, \pi_n^t, \tau^t) \subseteq \mathcal{F}_{x,n}$, and for all $i \in [n]$, $\frac{b_{i,n} - \widehat{b}_n(X_i)}{\sqrt{p_{i,n} - p_{i,n}^2}} \in \mathcal{F}_{x,n}$. Also $\mathcal{F}_{x,n} \perp\!\!\!\perp \tau^d$ by randomization. Then Lemma 9.20 applies, showing $E[C_n|\mathcal{F}_{x,n}] = 0$ and conditional variance

$$\begin{aligned}
\text{Var}(C_n|\mathcal{F}_{x,n}) &\leq 2E_n\left[\frac{(b_{i,n} - \widehat{b}_n(X_i))^2}{p_{i,n} - p_{i,n}^2}\right] = 2E_n\left[\left(\frac{m_1(X_i) - \widehat{m}_1(X_i)}{p_{i,n}} + \frac{m_0(X_i) - \widehat{m}_0(X_i)}{1 - p_{i,n}}\right)^2\right] \\
&\leq 4\delta^{-2}\sum_{d=0,1} E_n[(m_d(X_i) - \widehat{m}_d(X_i))^2] = O_p(n^{-2r_m})
\end{aligned}$$

The last line by conditional Markov, since $\zeta_n \perp\!\!\!\perp W_{1:n}$ we have $E[E_n[(m_d(X_i) - \widehat{m}_d(X_i))^2]|\zeta_n] = \|\widehat{m}_d - m_d\|_{2,P}^2 = O_p(n^{-2r_m})$. Then by conditional Chebyshev (Lemma 9.16), $C_n = O_p(n^{-r_m})$. An identical argument, using out-of-sample condition $\zeta_n \perp\!\!\!\perp W_{1:n}$, estimator rates, and Lemma 9.20 shows $B_n = O_p(n^{-r_m})$. Summarizing the work above

$$q\sqrt{n}(\widehat{\theta} - \theta) = q\sqrt{n}E_n\left[c(X_i) - \theta\right] + \sqrt{n}E_n\left[\frac{T_i D_i \epsilon_i^1}{p_{i,n}} - \frac{T_i(1 - D_i)\epsilon_i^0}{1 - p_{i,n}}\right] + o_p(1)$$

This is identical to the final coupling in the proof of Theorem 4.3, which shows the conclusion for Theorem 4.8. For Proposition 9.10, the required condition $(E[(\widehat{p}_{i,n} - p_i)^2|\zeta_n])^{1/2} = o_p(1)$ is trivial, since we may take $\widehat{p} = p$ (known propensity). □

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

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

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

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

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

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

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 ξ_n be external randomness forming the partition: $(I_k)_{k=1}^K = F_n(\xi_n)$ for deterministic F_n. Define the operator $E_{n,k}[Z_i] = n^{-1}\sum_{i\in I_k} Z_i$. A standard expansion of the AIPW estimator (algebra) gives, for the k^{th} fold estimator

$$\tilde{\theta}_k = E_{n,k}[c(X_i)] + E_{n,k}\left[(D_i - p_n(X_i))\left(\frac{(m_1 - \widehat{m}_{1,k})(X_i)}{p_n(X_i)} + \frac{(m_0 - \widehat{m}_{0,k})(X_i)}{1-p_n(X_i)}\right)\right]$$
$$+ E_{n,k}\left[\left(\frac{D_i\epsilon_i^1}{p_n(X_i)} - \frac{(1-D_i)\epsilon_i^0}{1-p_n(X_i)}\right)\right] \equiv A_{n,k} + B_{n,k} + C_{n,k}$$

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 τ_k^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 $\mathcal{F}_{n,k} = \sigma((Y_i, D_i, X_i)_{i\in I_k^c}, X_{1:n}, \pi_n^d, \xi_n)$. We check the conditions to apply Lemma 9.20. By cross-fitting of the design, $\{i : \mathbb{1}(i \in I_k) = 1\} = \bigsqcup_{a,s} g_{a,s}^k$. Observe that $(g_{a,s}^k)_{a,s} \in \sigma((X_i)_{i\in I_k}, \pi_n^d(k), \xi_n) \subseteq \mathcal{F}_{n,k}$ and, in the notation of Lemma 9.20

$$h_n(W_i) \equiv \left(\frac{(m_1 - \widehat{m}_{1,k})(X_i)}{p_n(X_i)} + \frac{(m_0 - \widehat{m}_{0,k})(X_i)}{1-p_n(X_i)}\right) \in \mathcal{F}_{n,k}$$

By Lemma 9.19.(i) and cross-fitting of the design.

$$(D_i)_{i\in I_k^c} \in \sigma\left((g_{a,s}^{k'})_{a,s}, \tau_{k'}^d : k' \neq k\right) \in \sigma((X_i)_{i\in I_{k'}}, \pi_n^d(k'), \xi_n, \tau_{k'}^d : k' \neq k) \tag{9.5}$$

Then we see that

$$\mathcal{F}_{n,k} \subseteq \sigma((W_i)_{i\in I_k^c}, (D_i)_{i\in I_k^c}, X_{1:n}, \pi_n^d, \xi_n) \subseteq \sigma((W_i)_{i\in I_k^c}, X_{1:n}, \pi_n^d, \xi_n, \tau_{k'}^d : k' \neq k) \perp\!\!\!\perp \tau_k^d$$

Then we may apply Lemma 9.20 with $S_i = \mathbb{1}(i \in I_k)$, which gives $E[B_{n,k}|\mathcal{F}_{n,k}] = 0$ and

$$\begin{aligned}\mathrm{Var}(\sqrt{n}B_{n,k}|\mathcal{F}_{n,k}) &\leq 2E_n\left[\mathbb{1}(i \in I_k)\left(\frac{(m_1 - \widehat{m}_{1,k})(X_i)}{p_n(X_i)} + \frac{(m_0 - \widehat{m}_{0,k})(X_i)}{1 - p_n(X_i)}\right)^2\right] \\ &\leq 4\delta^{-2}E_n\left[\mathbb{1}(i \in I_k)\sum_{d=0,1}(m_d - \widehat{m}_{d,k})(X_i)^2\right]\end{aligned}$$

Now, taking a conditional expectation

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

To see the equality, note that $\mathbb{1}(i \in I_k) \in \sigma(\xi_n)$ and by tower law and Equation 9.5 above

$$\begin{aligned}&E\left[\mathbb{1}(i \in I_k)(m_d - \widehat{m}_{d,k})(X_i)^2|(Y_i, D_i, X_i)_{i\in I_k^c}, \xi_n\right] \\ =&E\left[E\left[\mathbb{1}(i \in I_k)(m_d - \widehat{m}_{d,k})(X_i)^2|(W_i)_{i\in I_k^c}, (\pi_n^d(k'), \tau_{k'}^d) : k' \neq k, \xi_n\right]|(Y_i, D_i, X_i)_{i\in I_k^c}, \xi_n\right] \\ =&\mathbb{1}(i \in I_k)\|m_d - \widehat{m}_{d,k}\|_{2,P}^2\end{aligned}$$

Then by conditional Markov and conditional Chebyshev (Lemma 9.16), 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

$$\begin{aligned}\sqrt{n}(\tilde{\theta} - \mathrm{ATE}) = \sqrt{n}\left(\sum_{k=1}^K \tilde{\theta}_k - \mathrm{ATE}\right) &= \sqrt{n}E_n[c(X_i) - \mathrm{ATE}] + \sqrt{n}\sum_{k=1}^K B_{n,k} \\ &+ \sqrt{n}E_n\left[\left(\frac{D_i\epsilon_i^1}{p_n(X_i)} - \frac{(1 - D_i)\epsilon_i^0}{1 - p_n(X_i)}\right)\right] \\ &= \sqrt{n}E_n[c(X_i) - \mathrm{ATE}] + \sqrt{n}E_n\left[\left(\frac{D_i\epsilon_i^1}{p_n(X_i)} - \frac{(1 - D_i)\epsilon_i^0}{1 - p_n(X_i)}\right)\right] + o_p(1)\end{aligned}$$

Similar reasoning as in the proof of 3.17 may now be used to show the claimed CLT. The case $D_i \overset{\text{inid}}{\sim} p(X_i)$ follows from Chernozhukov et al. (2017), Theorem 5.1. □

9.4 Proofs - Inference

9.4.1 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 $\sqrt{n}(\widehat{\theta} - \text{ATE})$ is

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

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

$$\begin{aligned}\operatorname{Var}(c(X_i)) &= E[(c(X_i) - \text{ATE})^2] = E[(c(X_i) - \text{ATE})^2|\zeta] = E[(c(X_i) - E[c(X_i)|\zeta])^2|\zeta] \\ &= \operatorname{Var}(c(X_i)|\zeta) = E[((F_i^1 - F_i^0) + (u_i^1 - u_i^0))^2|\zeta] - \text{ATE}^2 \\ &= E[(F_i^1 - F_i^0)^2|\zeta] + E[(u_i^1 - u_i^0))^2|\zeta] - \text{ATE}^2\end{aligned}$$

Similarly we have

$$E[(c(X_i) - E[c(X_i)|\psi_i, \zeta])^2|\zeta] = E[(u_i^1 - u_i^0))^2|\zeta]$$

Then the first two terms can be written

$$q \operatorname{Var}(c(X_i)) + (1-q)E[(c(X_i) - E[c(X_i)|\psi_i, \zeta])^2|\zeta] = qE[(F_i^1 - F_i^0)^2|\zeta] + E[(u_i^1 - u_i^0))^2|\zeta] - q\,\text{ATE}^2$$

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

$$\begin{aligned}E[b(X_i)|\psi_i, p(\psi_i, \zeta), \zeta] &= E\left[m_1(X_i)\sqrt{\frac{1-p_i}{p_i}} + m_0(X_i)\sqrt{\frac{p_i}{1-p_i}}\,\middle|\,\psi_i, \zeta\right] \\ &= F_i^1\sqrt{\frac{1-p_i}{p_i}} + F_i^0\sqrt{\frac{p_i}{1-p_i}}\end{aligned}$$

Using this, the middle term becomes

$$E[(b(X_i) - E[b(X_i)|\psi_i, p(\psi_i, \zeta), \zeta])^2|\zeta] = E\left[\left(u_i^1\sqrt{\frac{1-p_i}{p_i}} + u_i^0\sqrt{\frac{p_i}{1-p_i}}\right)^2\middle|\zeta\right]$$
$$= E\left[(u_i^1)^2\frac{1-p_i}{p_i} + (u_i^0)^2\frac{p_i}{1-p_i} + 2u_i^1u_i^0\middle|\zeta\right]$$
$$= E\left[(u_i^1)^2\frac{1}{p_i} + (u_i^0)^2\frac{1}{1-p_i}\middle|\zeta\right] - E[(u_i^1 - u_i^0)^2|\zeta]$$

Then adding all four terms gives

$$\begin{aligned}
&qE[(F_i^1 - F_i^0)^2|\zeta] + E[(u_i^1 - u_i^0))^2|\zeta] - q\,\mathrm{ATE}^2 + E\left[\frac{(u_i^1)^2}{p_i} + \frac{(u_i^0)^2}{1-p_i}\middle|\zeta\right] \\
&-E[(u_i^1 - u_i^0)^2|\zeta] + E\left[\frac{\sigma_1^2(X_i)}{p_i} + \frac{\sigma_0^2(X_i)}{1-p_i}\middle|\zeta\right] \\
&= E\left[\frac{(u_i^1)^2 + \sigma_1^2(X_i)}{p_i} + \frac{(u_i^0)^2 + \sigma_0^2(X_i)}{1-p_i}\middle|\zeta\right] + qE[(F_i^1 - F_i^0)^2|\zeta] - q\,\mathrm{ATE}^2
\end{aligned}$$

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

$$\begin{aligned}
E\left[\frac{Y_i(1)^2}{p_i}\middle|\zeta\right] &= E\left[\frac{(m_1(X_i) + \epsilon_i^1)^2}{p_i}\middle|\zeta\right] = E\left[\frac{(F_i^1 + u_i^1 + \epsilon_i^1)^2}{p_i}\middle|\zeta\right] \\
&= E\left[\frac{(F_i^1)^2 + (u_i^1)^2 + (\epsilon_i^1)^2}{p_i}\middle|\zeta\right] = E\left[\frac{(F_i^1)^2 + (u_i^1)^2 + \sigma_1^2(X_i)}{p_i}\middle|\zeta\right]
\end{aligned}$$

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

$$\begin{aligned}
&E[p_i^{-1}u_i^1F_i^1|\zeta] = E[p_i^{-1}E[u_i^1F_i^1|\psi_i,\zeta]\zeta] = E[p_i^{-1}F_i^1E[u_i^1|\psi_i,\zeta]\zeta] = 0 \\
&E[p_i^{-1}\epsilon_i^1F_i^1|\zeta] = E[p_i^{-1}F_i^1E[\epsilon_i^1|X_i,\zeta]|\zeta] = E[p_i^{-1}F_i^1E[\epsilon_i^1|X_i]|\zeta] = 0 \\
&E[p_i^{-1}\epsilon_i^1u_i^1|\zeta] = E[p_i^{-1}u_i^1E[\epsilon_i^1|X_i,\zeta]|\zeta] = E[p_i^{-1}u_i^1E[\epsilon_i^1|X_i]|\zeta] = 0 \\
&E[p_i^{-1}(\epsilon_i^1)^2|\zeta] = E[p_i^{-1}E[(\epsilon_i^1)^2|X_i,\zeta]|\zeta] = E[p_i^{-1}E[(\epsilon_i^1)^2|X_i]|\zeta] = E[p_i^{-1}\sigma_1^2(X_i)|\zeta]
\end{aligned}$$

Similarly for $d = 0$

$$E\left[\frac{Y_i(0)^2}{1-p_i}\middle|\zeta\right] = E\left[\frac{(F_i^0)^2 + (u_i^0)^2 + \sigma_0^2(X_i)}{1-p_i}\middle|\zeta\right]$$

Comparing the calculations above, we see that

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

Now observe that

$$\begin{aligned}
&E\left[\frac{(F_i^1)^2}{p_i}\middle|\zeta\right] + E\left[\frac{(F_i^0)^2}{1-p_i}\middle|\zeta\right] - qE\left[(F_i^1 - F_i^0)^2|\zeta\right] \\
&= E\left[\frac{(F_i^1)^2(1-qp_i)}{p_i}\middle|\zeta\right] + E\left[\frac{(F_i^0)^2(1-q(1-p_i))}{1-p_i}\middle|\zeta\right] + 2qE[F_i^1F_i^0|\zeta]
\end{aligned}$$

Summarizing, we have shown that

$$\begin{aligned}
V(\psi,(q,p)) = E\left[\frac{Y_i(1)^2}{p_i}\middle|\zeta\right] + E\left[\frac{Y_i(0)^2}{1-p_i}\middle|\zeta\right] - E\left[\frac{(F_i^1)^2(1-qp_i)}{p_i}\middle|\zeta\right] \\
- E\left[\frac{(F_i^0)^2(1-q(1-p_i))}{1-p_i}\middle|\zeta\right] - 2qE[F_i^1F_i^0|\zeta] - q\,\mathrm{ATE}^2
\end{aligned}$$

Then by Lemma 9.12 we have

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

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

$$\begin{aligned}
&(1-q)E[(u_i^1-u_i^0))^2|\zeta] + E\left[(u_i^1)^2\frac{1}{p_i} + (u_i^0)^2\frac{1}{1-p_i}\middle|\zeta\right] - E[(u_i^1-u_i^0)^2|\zeta] \\
&+ E\left[\frac{\sigma_1^2(X_i)}{p_i} + \frac{\sigma_0^2(X_i)}{1-p_i}\middle|\zeta\right] = E\left[\frac{(u_i^1)^2+\sigma_1^2(X_i)}{p_i} + \frac{(u_i^0)^2+\sigma_0^2(X_i)}{1-p_i}\middle|\zeta\right] - qE[(u_i^1-u_i^0)^2|\zeta] \\
&\leq E\left[\frac{(u_i^1)^2+\sigma_1^2(X_i)}{p_i} + \frac{(u_i^0)^2+\sigma_0^2(X_i)}{1-p_i}\middle|\zeta\right]
\end{aligned}$$

Then comparing with the calculations above.

$$\begin{aligned}
V(\psi,p) - q\operatorname{Var}(c(X)) &\leq E\left[\frac{Y_i(1)^2}{p_i}\middle|\zeta\right] + E\left[\frac{Y_i(0)^2}{1-p_i}\middle|\zeta\right] - E\left[\frac{(F_i^1)^2}{p_i}\middle|\zeta\right] - E\left[\frac{(F_i^0)^2}{1-p_i}\middle|\zeta\right] \\
&= \widehat{v}_{I,2} - \widehat{v}_1\left[p^{-1}\right] - \widehat{v}_0\left[(1-p)^{-1}\right] + o_p(1)
\end{aligned}$$

The final equality follows by Lemma 9.12. Conservativeness is given by the residual variance

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

This completes the proof of claim (1).

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

$$\begin{aligned}
F_i^d &= E[m_d(X_i)|\psi_i, p(X_i,\zeta),\zeta] \\
u_i^d &= m_d(X_i) - E[m_d(X_i)|\psi_i, p(X_i,\zeta),\zeta]
\end{aligned}$$

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

$$\begin{aligned}
V(\psi,p) &= \operatorname{Var}(c(X_i)) + E[(b-E[b|\psi_i,p_i,\zeta])^2|\zeta] + E\left[\frac{\sigma_1^2(X_i)}{p_i} + \frac{\sigma_0^2(X_i)}{1-p_i}\middle|\zeta\right] \\
&= E[(F_i^1-F_i^0)^2|\zeta] + E[(u_i^1-u_i^0)^2|\zeta] - \mathrm{ATE}^2 \\
&\quad + E\left[(u_i^1)^2\frac{1}{p_i} + (u_i^0)^2\frac{1}{1-p_i}\middle|\zeta\right] - E[(u_i^1-u_i^0)^2|\zeta] + E\left[\frac{\sigma_1^2(X_i)}{p_i} + \frac{\sigma_0^2(X_i)}{1-p_i}\middle|\zeta\right] \\
&= E\left[\frac{(u_i^1)^2+\sigma_1^2(X_i)}{p_i} + \frac{(u_i^0)^2+\sigma_0^2(X_i)}{1-p_i}\middle|\zeta\right] + E[(F_i^1-F_i^0)^2|\zeta] - \mathrm{ATE}^2 \\
&= E\left[\frac{Y_i(1)^2}{p_i}\middle|\zeta\right] + E\left[\frac{Y_i(0)^2}{1-p_i}\middle|\zeta\right] - \mathrm{ATE}^2 \\
&\quad - E\left[\frac{(F_i^1)^2}{p_i}\middle|\zeta\right] - E\left[\frac{(F_i^0)^2}{1-p_i}\middle|\zeta\right] + E[(F_i^1-F_i^0)^2|\zeta]
\end{aligned}$$

Note that the last line is

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

Then by Lemma 9.12

$$V(\psi, p) = \text{Var}_n\left(\frac{(D_i - p_{i,n})Y_i}{p_{i,n} - p_{i,n}^2}\right) - \widehat{v}_1\left[\frac{1-p}{p}\right] - \widehat{v}_0\left[\frac{p}{1-p}\right] - 2\widehat{v}_{10} + o_p(1)$$

This completes the proof of the second claim. For the fixed regressors case, by the work above

$$\begin{aligned} V(\psi, p) - \text{Var}(c(X_i)) &= E\left[\frac{(u_i^1)^2 + \sigma_1^2(X_i)}{p_i} + \frac{(u_i^0)^2 + \sigma_0^2(X_i)}{1-p_i}\middle|\zeta\right] - E[(u_i^1 - u_i^0)^2|\zeta] \\ &\leq E\left[\frac{(u_i^1)^2 + \sigma_1^2(X_i)}{p_i} + \frac{(u_i^0)^2 + \sigma_0^2(X_i)}{1-p_i}\middle|\zeta\right] \\ &= E\left[\frac{Y_i(1)^2}{p_i}\middle|\zeta\right] + E\left[\frac{Y_i(0)^2}{1-p_i}\middle|\zeta\right] - E\left[\frac{(F_i^1)^2}{p_i}\middle|\zeta\right] - E\left[\frac{(F_i^0)^2}{1-p_i}\middle|\zeta\right] \end{aligned}$$

Claim (3) follows from (2) and Lemma 9.12, noting that for the general case $\psi_i' \in \sigma(\psi_i)$

$$\begin{aligned} &\text{Var}(c(X_i)) - q\,\text{Var}(c(X_i)) + (1-q)E[(c - E[c|\psi_i', \zeta])^2|\zeta] \\ &= (1-q)\Big[\text{Var}(c(X_i)) - E[\text{Var}(c(X_i)|\psi_i', \zeta)|\zeta]\Big] = (1-q)\,\text{Var}(E[c(X_i)|\psi_i', \zeta]|\zeta) \end{aligned}$$

This finishes the proof. □

Assumption 9.11 (Inference)**.** *Require assumption 9.1 with $\zeta_n = \zeta$ for $n \geq 1$ and $\widehat{p} = p$ and assumption 9.3. Define the projection*

$$F_n^d(\psi, p_a, \zeta') = E[m_d(X_i)|\psi(X_i, \zeta) = \psi, p_n(X_i, \zeta) = p_a, \zeta = \zeta']$$

Also define $F_{i,n}^d = F_n^d(\psi_{i,n}, p_{i,n}, \zeta)$. Similarly, define $F_i^d = E[m_d(X_i)|\psi(X_i, \zeta), p(X_i, \zeta), \zeta]$, $u_{i,n}^d = m_d(X_i) - F_{i,n}^d$, and $u_i^d = m_d(X_i) - F_i^d$. Define $\sigma_{u,i,n}^2 = E[(u_{i,n})^2|\psi_i, p_{i,n}, \zeta]$ for $u_{i,n} = u_{i,n}^d$ and $d = 0, 1$. Require that

(1) $E[Y(d)^4] < \infty$ and $E[(F_i^d)^4|\zeta] < \bar{F}$ and $E[(F_{i,n}^d)^4|\zeta] < \bar{F} < \infty$, ζ-a.s. for $d = 0, 1$

(2) $\sigma_{u^d,i,n}^2 < \overline{\sigma}^2$ and $\sigma_d^2(X_i) \leq \overline{\sigma}^2 < \infty$ for $d = 0, 1$

(3) $F_n^d(\psi, p_a, \zeta)$ satisfies Hölder conditions in either Definition 8.1 or 8.5 for $d = 0, 1$.

(4) $E[(F_{i,n}^d - F_i^d)^2|\zeta]^{1/2} = O_p(r_n^F)$ for $d = 0, 1$.

(5) γ is an admissible matching as in Definition 5.1

(6) $\overline{k}_n \vee L_n = o(n)$ and $r_n^\psi \vee r_n^F \vee r_n^p = o(1)$

Lemma 9.12. *Suppose that assumptions 9.1, 9.3, and 9.11 hold. Then the IPW second moment (for the selected units)*

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

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

$$\widehat{v}_1[G] \equiv E_n\left[\frac{T_i D_i}{q p_{i,n}} G(p_{i,n}) Y_i Y_{\gamma(i)}\right] = E\left[G(p_i)(F_i^1)^2|\zeta\right] + O_p(r_{1,n})$$
$$\widehat{v}_0[G] \equiv E_n\left[\frac{T_i(1-D_i)}{q(1-p_{i,n})} G(p_{i,n}) Y_i Y_{\gamma(i)}\right] = E\left[G(p_i)(F_i^0)^2|\zeta\right] + O_p(r_{0,n})$$
$$\widehat{v}_{10} \equiv 2n^{-1} \sum_{1\le i<j\le n} \frac{T_i T_j}{q} \frac{\mathbb{1}(g(i)=g(j))}{\mathrm{Card}(g)} \frac{D_i(1-D_j)Y_i Y_j}{p_{i,n}(1-p_{j,n})} = E[F_i^1 F_i^0|\zeta] + O_p(r_{10,n})$$

With convergence rates

$$r_{1,n} \vee r_{0,n} = O(n^{-1/2}) + O(r_n^p) + O(r_n^F) + O(r_n^\psi)$$
$$r_{10,n} = O(n^{-1/2}) + O(r_n^F) + O(r_n^\psi) + O((\overline{k}_n/n)^{1/2}) + O\left(\sqrt{|L_n|}/n\right)$$

Proof. First, we show consistency for $\widehat{v}_I$. From Theorem 9.5, we have $\widehat{\theta} = \mathrm{ATE} + O_p(n^{-1/2})$, so $(\widehat{\theta})^2 = \mathrm{ATE}^2 + O_p(n^{-1/2})$. Since $D_i(1-D_i) = 0$, the first term is

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

Consider the $d = 1$ term

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

We check the conditions of Lemma 9.20.(4). Let $\mathcal{F}_n = \sigma(W_{1:n}, \zeta_n, \pi_n^t, \pi_n^d, \tau^t)$. Then by Assumption 9.3 and 9.1, $(g_{a,s})_{a,s}$, $\frac{Y_i(1)^2}{p_{i,n}^2 q}$, $\frac{Y_i(1)^2}{p_{i,n} q} \in \mathcal{F}_n$ and $\mathcal{F}_n \perp\!\!\!\perp \tau^d$. Moreover, note that

$$\sup_{n\ge1} E\left[\frac{Y_i(1)^4}{(p_{i,n}q)^2}\right] \le \sup_{n\ge1} E\left[\frac{Y_i(1)^4}{p_{i,n}^4 q^2}\right] \le \delta^{-6} E[Y(1)^4] < \infty$$

Then Lemma 9.20.(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 9.18 applied with $H(W) = Y(1)$ and

$G(p) = p^{-1}$ shows

$$E_n\left[\frac{Y_i(1)^2}{p_{i,n}}\right] = E\left[\frac{Y_i(1)^2}{p_i}\middle|\zeta\right] + O_p(r_n^p) + O_p(n^{-1/2})$$

We have shown that $E_n\left[\frac{T_i}{q}\frac{D_i Y_i(1)^2}{p_{i,n}^2}\right] = E\left[\frac{Y_i(1)^2}{p_i}\right] + O_p(n^{-1/2}) + O_p(r_n^p)$. The case $d = 0$ is identical. This finishes the proof of the claim for $\widehat{v}_I$.

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

$$Y_i(d)Y_j(d') = (F_{i,n}^d + u_{i,n}^d + \epsilon_i^d)(F_{j,n}^{d'} + u_{j,n}^{d'} + \epsilon_j^{d'}) = F_{i,n}^d F_{j,n}^{d'} + F_{j,n}^{d'}(u_{i,n}^d + \epsilon_i^d) \tag{9.6}$$

$$+ F_{i,n}^d(u_{j,n}^{d'} + \epsilon_j^{d'}) + (u_{i,n}^d + \epsilon_i^d)(u_{j,n}^{d'} + \epsilon_j^{d'}) \equiv F_{i,n}^d F_{j,n}^{d'} + R_{i,j}^{d,d'} \tag{9.7}$$

Define $g_{a,s}^1 = g_{a,s} \cap \{i : D_i = 1\}$. Using this decomposition, and noting $\{i : T_i = 1\} = \bigsqcup_{a,s} g_{a,s}$ by Assumption 9.3, we have

$$\widehat{v}_1 = E_n\left[\frac{T_i D_i}{q p_{i,n}} G(p_{i,n}) Y_i Y_{\gamma(i)}\right] = n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{i \in g_{a,s}^1}(qp_a)^{-1}G(p_a)Y_i(1)Y_{\gamma(i)}(1)$$

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

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

$$\begin{aligned} F_{i,n}^d F_{j,n}^d &= F_{i,n}^d F_{j,n}^d - 1/2((F_{i,n}^d)^2 + (F_{j,n}^d)^2) + 1/2((F_{i,n}^d)^2 + (F_{j,n}^d)^2) \\ &= (1/2)[(F_{i,n}^d)^2 + (F_{j,n}^d)^2 - (F_{i,n}^d - F_{j,n}^d)^2] \end{aligned}$$

Then the first part of the numerator above becomes

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

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

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

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

$$= n^{-1}\sum_{i=1}^{n} T_i D_i (qp_{i,n})^{-1}G(p_{i,n})(F_{i,n}^1)^2$$

The equality follows since γ is bijective and $D_i = D_{\gamma(i)}$, so the second term sums over

each treated unit exactly twice. Continuing, the expression above is

$$
\begin{aligned}
&= n^{-1}\sum_{i=1}^{n} T_i(D_i - p_{i,n})(qp_{i,n})^{-1}G(p_{i,n})(F^1_{i,n})^2 + n^{-1}\sum_{i=1}^{n}\frac{T_i}{q}G(p_{i,n})(F^1_{i,n})^2 \\
&= n^{-1}\sum_{i=1}^{n} T_i(D_i - p_{i,n})(qp_{i,n})^{-1}G(p_{i,n})(F^1_{i,n})^2 + n^{-1}\sum_{i=1}^{n}\frac{T_i - q}{q}G(p_{i,n})(F^1_{i,n})^2 \\
&\quad + n^{-1}\sum_{i=1}^{n} G(p_{i,n})(F^1_{i,n})^2
\end{aligned}
$$

As in the argument for $\widehat{v}_I$, let $\mathcal{F}_n = \sigma(W_{1:n}, \zeta_n, \pi^t_n, \pi^d_n, \tau^t)$. Note that $\mathcal{F}_n \perp\!\!\!\perp \tau^d$, $(g_{a,s})_{a,s} \in \mathcal{F}_n$, and $(qp_{i,n})^{-1}G(p_{i,n})(F^1_{i,n})^2 \in \mathcal{F}_n$. Moreover, note that

$$
\sup_{n\geq 1} E\left[\left((qp_{i,n})^{-1}G(p_{i,n})(F^1_{i,n})^2\right)^2\right] \lesssim \delta^{-4}\sup_{n\geq 1} E\left[(F^1_{i,n})^4\right] = \delta^{-4}\sup_{n\geq 1} E\left[E[(F^1_{i,n})^4|\zeta_n]\right] < \infty
$$

Then by Lemma 9.20.(4), the first term is $O_p(n^{-1/2})$. Similarly, the second term is $O_p(n^{-1/2})$ by Lemma 9.20.(4), applied with selection variables $S_{1:n} = 1$. For the third term, by boundedness we have $|E_n[G(p_{i,n})(F^1_{i,n})^2 - (F^1_i)^2]| \lesssim E_n[|F^1_{i,n} - F^1_i||F^1_{i,n} + F^1_i|]$. Taking a conditional expectation

$$
\begin{aligned}
E[E_n[|F^1_{i,n} - F^1_i||F^1_{i,n} + F^1_i|]|\zeta] &\lesssim E[(F^1_{i,n} - F^1_i)^2|\zeta]^{1/2}\left(E[(F^1_{i,n})^2|\zeta]^{1/2} + E[(F^1_i)^2|\zeta]^{1/2}\right) \\
&\leq O_p(r^F_n)E[m_1(X)^2]^{1/2} = O_p(r^F_n)
\end{aligned}
$$

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 9.18

$$
E_n[G(p_{i,n})(F^1_i)^2] = E[G(p_i)(F^1_i)^2|\zeta] + O_p(n^{-1/2}) + O_p(r^p_n)
$$

The lemma applies since $E[(F^1_i)^4|\zeta] < \infty$, ζ-a.s. by assumption. Putting this all together, we have shown that $A^1_n = E[G(p_i)(F^1_i)^2|\zeta] + O_p(n^{-1/2}) + O_p(r^p_n) + O_p(r^F_n)$.

Next, consider A^2_n. By the Lipschitz assumption 8.5, on an almost sure set given by the definition, $|A^2_n|$ is bounded by

$$
\begin{aligned}
&(2n)^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{i\in g^1_{a,s}}(qp_a)^{-1}|G(p_a)|(F^1_{i,n} - F^1_{\gamma(i),n})^2 \lesssim n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{i\in g^1_{a,s}}(\psi_i - \psi_{\gamma(i)})^2 \\
&\leq n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{i\in g_{a,s}}(\psi_i - \psi_{\gamma(i)})^2 = E_n[T_i(\psi_i - \psi_{\gamma(i)})^2] = O_p(r^\psi_n)
\end{aligned}
$$

The final equality follows by matching admissibility as in Definition 5.1.

Next, we bound the residual terms above

$$\begin{aligned} B_n &= n^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} \sum_{i \in g^1_{a,s}} (qp_a)^{-1} G(p_a) R^{1,1}_{i,\gamma(i)} \\ &= n^{-1} \sum_{a=1}^{|L_n|} (qp_a)^{-1} G(p_a) \sum_{s=1}^{n} \sum_{i,j=1}^{n} D_i \mathbb{1}(i \in g_{a,s}) \mathbb{1}(j = \gamma(i)) R^{1,1}_{i,j} \end{aligned}$$

By definition the matching can be written as a deterministic function

$$\gamma = \gamma((g_{a,s})_{a,s}, \psi_{1:n}, p_{1:n,n}, D_{1:n}, T_{1:n}) = \gamma(\psi_{1:n}, p_{1:n,n}, \pi^t_n, \pi^d_n, \tau^t, \tau^d)$$

By assumption 9.3 and Lemma 9.19.(i). Let $\mathcal{F}_{(\psi,p_n,\tau^d),n} = \sigma(\psi_{1:n}, p_{1:n,n}, \pi^t_n, \pi^d_n, \tau^t, \tau^d, \zeta)$. Again by Lemma 9.19.(i), $D_i \in \mathcal{F}_{(\psi,p_n,\tau^d),n}$ for all i, and clearly $F^d_{i,n} \in \mathcal{F}_{(\psi,p_n,\tau^d),n}$. In the sequel, we will make use of the following conditional moments. Momentarily suppressing the 1-superscript notation, applying Lemma 9.22 with $\mathcal{F}_n = \mathcal{F}_{(\psi,p_n,\tau^d),n}$, $\xi_n = (\pi^d_n, \pi^t_n, \tau^t, \tau^d)$, and $\zeta_n = \zeta$ gives

$$\begin{aligned} 0 &= E[u_{i,n}|\mathcal{F}_{(\psi,p_n,\tau^d),n}] = E[\epsilon_i|\mathcal{F}_{(\psi,p_n,\tau^d),n}] = E[u_{i,n}u_{j,n}|\mathcal{F}_{(\psi,p_n,\tau^d),n}] = E[u_{i,n}\epsilon_j|\mathcal{F}_{(\psi,p_n,\tau^d),n}] \\ &= E[\epsilon_i\epsilon_j|\mathcal{F}_{(\psi,p_n,\tau^d),n}] = E[u_{i,n}u_{j,n}\epsilon_l|\mathcal{F}_{(\psi,p_n,\tau^d),n}] = E[u_{i,n}\epsilon_j\epsilon_l|\mathcal{F}_{(\psi,p_n,\tau^d),n}] = E[u_{i,n}u_{j,n}u_{l,n}|\mathcal{F}_{(\psi,p_n,\tau^d),n}] \\ &= E[\epsilon_i\epsilon_j\epsilon_l|\mathcal{F}_{(\psi,p_n,\tau^d),n}] = E[\epsilon_i\epsilon_j\epsilon_l\epsilon_f|\mathcal{F}_{(\psi,p_n,\tau^d),n}] = E[\epsilon_i\epsilon_j\epsilon_l u_{f,n}|\mathcal{F}_{(\psi,p_n,\tau^d),n}] \\ &= E[\epsilon_i\epsilon_j u_{l,n}u_{f,n}|\mathcal{F}_{(\psi,p_n,\tau^d),n}] = E[\epsilon_i u_{j,n}u_{l,n}u_{f,n}|\mathcal{F}_{(\psi,p_n,\tau^d),n}] = E[u_{i,n}u_{j,n}u_{l,n}u_{f,n}|\mathcal{F}_{(\psi,p_n,\tau^d),n}] \end{aligned}$$

Recall that

$$R^{1,1}_{i,j} = F^1_{j,n}(u^1_{i,n} + \epsilon^1_i) + F^1_{i,n}(u^1_{j,n} + \epsilon^1_j) + (u^1_{i,n} + \epsilon^1_i)(u^1_{j,n} + \epsilon^1_j)$$

By the moments above, $E[R^{1,1}_{i,j}|\mathcal{F}_{(\psi,p_n,\tau^d),n}] = 0$ for $1 \leq i \neq j \leq n$. Since $D_i\mathbb{1}(i \in g_{a,2s})\mathbb{1}(j = \gamma(i)) \in \mathcal{F}_{(\psi,p_n,\tau^d),n}$ and $\mathbb{1}(j = \gamma(i)) = 0$ for $j = i$ by definition, we have showns that $E[B_n|\mathcal{F}_{(\psi,p_n,\tau^d),n}] = 0$. Next, consider the conditional variance $E[B^2_n|\mathcal{F}_{(\psi,p_n,\tau^d),n}]$, given by

$$\begin{aligned} E[B^2_n|\mathcal{F}_{(\psi,p_n,\tau^d),n}] &= n^{-2} \sum_{a,a'=1}^{|L_n|} \frac{G(p_a)G(p_{a'})}{q^2 p_a p_{a'}} \sum_{s,s'=1}^{n} \sum_{i \neq j=1}^{n} \sum_{l \neq f=1}^{n} C^{a,a',s,s'}_{i,j,l,f} E[R^{1,1}_{i,j} R^{1,1}_{l,f}|\mathcal{F}_{(\psi,p_n,\tau^d),n}] \\ C^{a,a',s,s'}_{i,j,l,f} &\equiv D_i D_l \mathbb{1}(i \in g_{a,s})\mathbb{1}(l \in g_{a',s'})\mathbb{1}(j = \gamma(i))\mathbb{1}(f = \gamma(l)) \end{aligned}$$

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 $\mathbb{1}(j = \gamma(i))\mathbb{1}(i = \gamma(l)) = 0$, since each unit is matched uniquely. The other subcases follow by permuting labels.

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

$$\begin{aligned} R^{1,1}_{i,j} R^{1,1}_{l,f} &= F^1_{f,n}(u^1_{l,n} + \epsilon^1_l)[F^1_{j,n}(u^1_{i,n} + \epsilon^1_i) + F^1_{i,n}(u^1_{j,n} + \epsilon^1_j) + (u^1_{i,n} + \epsilon^1_i)(u^1_{j,n} + \epsilon^1_j)] \\ &+ F^1_{l,n}(u^1_{f,n} + \epsilon^1_f)[F^1_{j,n}(u^1_{i,n} + \epsilon^1_i) + F^1_{i,n}(u^1_{j,n} + \epsilon^1_j) + (u^1_{i,n} + \epsilon^1_i)(u^1_{j,n} + \epsilon^1_j)] \\ &+ (u^1_{l,n} + \epsilon^1_l)(u^1_{f,n} + \epsilon^1_f)[F^1_{j,n}(u^1_{i,n} + \epsilon^1_i) + F^1_{i,n}(u^1_{j,n} + \epsilon^1_j) + (u^1_{i,n} + \epsilon^1_i)(u^1_{j,n} + \epsilon^1_j)] \end{aligned}$$

Let i, j, l, f jointly distinct indices. Then since $F^1_{i,n} \in \mathcal{F}_{(\psi,p_n,\tau^d),n}$ for all i, we see that in

this case $E[R^{1,1}_{i,j}R^{1,1}_{l,f}|\mathcal{F}_{(\psi,p_n,\tau^d),n}] = 0$ 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 $C^{a,a',s,s'}_{i,j,j,i} = 0$ for all i,j. Then we only need consider terms of the form

$$\begin{aligned}(R^{1,1}_{i,j})^2 = (F^1_{j,n})^2(u^1_{i,n}+\epsilon^1_i)^2 + (F^1_{i,n})^2(u^1_{j,n}+\epsilon^1_j)^2 + (u^1_{i,n}+\epsilon^1_i)^2(u^1_{j,n}+\epsilon^1_j)^2 \\ + F^1_{j,n}F^1_{i,n}(u^1_{i,n}+\epsilon^1_i)(u^1_{j,n}+\epsilon^1_j) + F^1_{j,n}(u^1_{i,n}+\epsilon^1_i)^2(u^1_{j,n}+\epsilon^1_j) \\ + F^1_{i,n}(u^1_{j,n}+\epsilon^1_j)^2(u^1_{i,n}+\epsilon^1_i)\end{aligned}$$

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

$$0 = E[u_{i,n}\epsilon_i|\mathcal{F}_{(\psi,p_n,\tau^d),n}] = E[u_{i,n}\epsilon_i u_{j,n}\epsilon_j|\mathcal{F}_{(\psi,p_n,\tau^d),n}] = E[(u^1_{i,n}+\epsilon^1_i)(u^1_{j,n}+\epsilon^1_j)^2|\mathcal{F}_{(\psi,p_n,\tau^d),n}]$$

again by Lemma 9.22. Then all of the cross terms are conditionally mean zero. This leaves only diagonal terms of the form $E[(R^{1,1}_{i,j})^2|\mathcal{F}_{(\psi,p_n,\tau^d),n}]$, which can be written

$$\begin{aligned}&E\left[(F^1_{j,n})^2((u^1_{i,n})^2+(\epsilon^1_i)^2) + (F^1_{i,n})^2((u^1_{j,n})^2+(\epsilon^1_j)^2)|\mathcal{F}_{(\psi,p_n,\tau^d),n}\right] \\ &+ E\left[(u^1_{i,n}u^1_{j,n})^2 + (u^1_{i,n}\epsilon^1_j)^2 + (u^1_{j,n}\epsilon^1_i)^2 + (\epsilon^1_i\epsilon^1_j)^2|\mathcal{F}_{(\psi,p_n,\tau^d),n}\right] \\ &= (F^1_{j,n})^2(\sigma^2_{u^1,n}(X_i) + E[\sigma^2_1(X_i)|\psi,p_{i,n},\zeta]) + (F^1_{i,n})^2(\sigma^2_{u^1,n}(X_j) + E[\sigma^2_1(X_i)|\psi,p_{i,n},\zeta]) \\ &+ \sigma^2_{u^1,n}(X_i)\sigma^2_{u^1,n}(X_j) + \sigma^2_{u^1,n}(X_i)E[\sigma^2_1(X_j)|\psi,p_{i,n},\zeta] + \sigma^2_{u^1,n}(X_j)E[\sigma^2_1(X_i)|\psi,p_{i,n},\zeta] \\ &+ E[\sigma^2_1(X_i)|\psi,p_{i,n},\zeta]E[\sigma^2_1(X_j)|\psi,p_{i,n},\zeta] \leq 2\overline{\sigma}^2[(F^1_{j,n})^2 + (F^1_{i,n})^2] + 4(\overline{\sigma}^2)^2\end{aligned}$$

The second equality uses Lemma 9.22, 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 $E[B_n^2|\mathcal{F}_{(\psi,p_n,\tau^d),n}]$ is

$$\begin{aligned}&n^{-2}\sum_{a=1}^{|L_n|}\frac{G(p_a)^2}{q^2p_a^2}\sum_{s=1}^{n}\sum_{i\neq j=1}^{n} D_i\mathbb{1}(i \in g_{a,s})\mathbb{1}(j=\gamma(i))E[(R^{1,1}_{i,j})^2|\mathcal{F}_{(\psi,p_n,\tau^d),n}] \\ &\lesssim n^{-2}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{i\neq j=1}^{n} D_i\mathbb{1}(i \in g_{a,s})\mathbb{1}(j=\gamma(i))\left(2\overline{\sigma}^2[(F^1_{j,n})^2+(F^1_{i,n})^2] + 4(\overline{\sigma}^2)^2\right) \\ &= n^{-2}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{i\neq j=1}^{n} \mathbb{1}(i \in g_{a,s})\mathbb{1}(j=\gamma(i))\left(2\overline{\sigma}^2[D_j(F^1_{j,n})^2+D_i(F^1_{i,n})^2] + 4D_i(\overline{\sigma}^2)^2\right)\end{aligned}$$

The last equality since $D_i\mathbb{1}(j=\gamma(i)) = D_j$. Finally, note that this is

$$\begin{aligned}&n^{-2}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{i\in g_{a,s}}^{n}\left[2\overline{\sigma}^2[D_{\gamma(i)}(F^1_{\gamma(i),n})^2 + D_i(F^1_{i,n})^2] + 4D_i(\overline{\sigma}^2)^2\right] \\ &\lesssim n^{-2}\sum_{i=1}^{n} T_iD_i(F^1_{i,n})^2 + n^{-2}\cdot(qn) = O_p(n^{-1})\end{aligned}$$

The final equality follows since $E_n[T_iD_i(F^1_{i,n})^2] = O_p(1)$ by the calculations above for the term A^1_n. Then by conditional Chebyshev (Lemma 9.16), $B_n = O_p(n^{-1/2})$. Putting this

all together, we have shown that

$$\widehat{v}_1 = A_n^1 + A_n^2 + B_n = E[G(p_i)(F_i^1)^2|\zeta] + O_p(n^{-1/2}) + O_p(r_n^p) + O_p(r_n^F) + O_p(r_n^\psi)$$

By symmetry of our assumptions with respect to permutation of the labels $d \in \{0,1\}$, the same result holds for the estimator $\widehat{v}_0$.

To conclude, we analyze the cross-moment estimator, expanding using Equation 9.6

$$\begin{aligned}
\widehat{v}_{10} &= 2n^{-1} \sum_{1\le i<j\le n} \frac{T_i T_j}{q} \frac{\mathbb{1}(g(i)=g(j))}{\mathrm{Card}(g)} \frac{D_i(1-D_j)Y_iY_j}{p_n(X_i)(1-p_n)(X_j)} \\
&= (qn)^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} \frac{1}{k_a} \frac{k_a^2}{q_a(k_a-q_a)} \sum_{i\in g_{a,s}^1} \sum_{j\in g_{a,s}^0} Y_i(1)Y_j(0) \\
&= (qn)^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} \frac{1}{k_a-1} \frac{k_a(k_a-1)}{q_a(k_a-q_a)} \sum_{i\in g_{a,s}^1} \sum_{j\in g_{a,s}^0} (F_{i,n}^1 F_{j,n}^0 + R_{i,j}^{1,0}) \equiv S_n + T_n
\end{aligned}$$

First consider S_n. Let $\mathcal{F}_{(\psi,p_n),n} = \sigma(\psi_{1:n}, p_{1:n,n}, \pi_n^t, \pi_n^d, \tau^t, \zeta)$, then we have $(g_{a,s})_{a,s}, F_{i,n}^d \in \mathcal{F}_{(\psi,p_n),n}$ and $\mathcal{F}_{(\psi,p_n),n} \perp\!\!\!\perp \tau^d$, so by Lemma 9.19

$$\begin{aligned}
E[S_n|\mathcal{F}_{(\psi,p_n),n}] &= (qn)^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} \frac{1}{k_a-1} \frac{k_a(k_a-1)}{q_a(k_a-q_a)} \sum_{i,j\in g_{a,s}} E[D_i(1-D_j)|\mathcal{F}_{(\psi,p_n),n}] F_{i,n}^1 F_{j,n}^0 \\
&= (qn)^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} \frac{1}{k_a-1} \frac{k_a(k_a-1)}{q_a(k_a-q_a)} \sum_{\substack{i,j\in g_{a,s} \\ i\neq j}} \frac{q_a(k_a-q_a)}{k_a(k_a-1)} F_{i,n}^1 F_{j,n}^0 \\
&= (qn)^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} \frac{1}{k_a-1} \sum_{\substack{i,j\in g_{a,s} \\ i<j}} F_{i,n}^1 F_{j,n}^0 + F_{j,n}^1 F_{i,n}^0 \\
&= (qn)^{-1} \sum_{a=1}^{|L_n|} \sum_{s=1}^{n} \frac{1}{k_a-1} \sum_{\substack{i,j\in g_{a,s} \\ i<j}} F_{i,n}^1 F_{i,n}^0 + F_{j,n}^1 F_{j,n}^0 + (F_{i,n}^1 - F_{j,n}^1)(F_{j,n}^0 - F_{i,n}^0) \\
&\equiv S_n^1 + S_n^2
\end{aligned}$$

First consider S_n^1. Note that the final sum is

$$\begin{aligned}
&\sum_{\substack{i,j\in g_{a,s} \\ i<j}} F_{i,n}^1 F_{i,n}^0 + \sum_{\substack{i,j\in g_{a,s} \\ i<j}} F_{j,n}^1 F_{j,n}^0 = \sum_{i=1}^{n} F_{i,n}^1 F_{i,n}^0 \mathbb{1}(i\in g_{a,s}) \sum_{j>i} \mathbb{1}(j\in g_{a,s}) \\
+ &\sum_{j=1}^{n} F_{j,n}^1 F_{j,n}^0 \mathbb{1}(j\in g_{a,s}) \sum_{i<j} \mathbb{1}(i\in g_{a,s}) = \sum_{i=1}^{n} F_{i,n}^1 F_{i,n}^0 \mathbb{1}(i\in g_{a,s})(k_a-1)
\end{aligned}$$

Then we have $S_n^1 = (qn)^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{i\in g_{a,s}}^{n} F_{i,n}^1 F_{i,n}^0 = q^{-1}E_n[T_i F_{i,n}^1 F_{i,n}^0]$. Next we write

$$\begin{aligned}&E_n[T_i F_{i,n}^1 F_{i,n}^0] = E_n[T_i(F_{i,n}^1 - F_i^1 + F_i^1)(F_{i,n}^0 - F_i^0 + F_i^0)] \\ &= E_n[T_i F_i^1 F_i^0] + E_n[T_i(F_{i,n}^1 - F_i^1)(F_{i,n}^0 - F_i^0)] + E_n[T_i(F_{i,n}^1 - F_i^1)F_i^0] + E_n[T_i F_i^1(F_{i,n}^0 - F_i^0)]\end{aligned}$$

The first term has

$$\begin{aligned}E[|E_n[T_i(F_{i,n}^1 - F_i^1)(F_{i,n}^0 - F_i^0)]||\zeta] &\leq E[|F_{i,n}^1 - F_i^1||F_{i,n}^0 - F_i^0||\zeta] \\ &\leq E[|F_{i,n}^1 - F_i^1|^2|\zeta]^{1/2}E[|F_{i,n}^0 - F_i^0|^2|\zeta]^{1/2} = O_p((r_n^F)^2)\end{aligned}$$

Also $E[|E_n[T_i(F_{i,n}^1 - F_i^1)F_i^0]||\zeta] \leq E[|F_{i,n}^1 - F_i^1|^2|\zeta]^{1/2}E[|F_i^1|^2|\zeta]^{1/2} = O_p(r_n^F)$. Then by conditional Markov, the last three terms above are $O_p(r_n^F)$. Finally, note that

$$E_n[T_i F_i^1 F_i^0] = E_n[(T_i - q)F_i^1 F_i^0] + qE_n[F_i^1 F_i^0] = qE[F_i^1 F_i^0|\zeta] + O_p(n^{-1/2})$$

For the final equality, note that $E_n[(T_i - q)F_i^1 F_i^0]$ by Lemma 9.20 since by assumption

$$E[(F_i^1)^2(F_i^0)^2|\zeta] \lesssim E[(F_i^1)^4|\zeta] + E[(F_i^0)^4|\zeta] < \infty$$

The conclusion then follows conditional Markov manipulations applied to $qE_n[F_i^1 F_i^0]$. We have shown that $S_n^1 = E[F_i^1 F_i^0|\zeta] + O_p(n^{-1/2}) + O_p(r_n^F)$.

Next, consider the differences in S_n^2 above. By the Lipschitz assumption, on an almost sure set given by definition 8.5 we have

$$\begin{aligned}|S_n^2| &\leq (qn)^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\frac{1}{k_a - 1}\sum_{\substack{i,j\in g_{a,s}\\ i<j}}|F_{i,n}^1 - F_{j,n}^1||F_{j,n}^0 - F_{i,n}^0| \\ &\lesssim n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\frac{1}{k_a - 1}\sum_{\substack{i,j\in g_{a,s}\\ i<j}}(\psi_i - \psi_j)^2 \lesssim n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}k_a^{-1}\sum_{\substack{i,j\in g_{a,s}\\ i\neq j}}(\psi_i - \psi_j)^2 = O_p(r_n^\psi)\end{aligned}$$

Putting this all together, we have shown that

$$E[S_n|\mathcal{F}_{(\psi,p_n),n}] = E[F_i^1 F_i^0|\zeta] + O_p(n^{-1/2}) + O_p(r_n^F) + O_p(r_n^\psi)$$

Next, we will show the coupling $S_n - E[S_n|\mathcal{F}_{(\psi,p_n),n}] = O_p((\overline{k}_n/n)^{1/2})$. As previously, define the higher-order inverse propensity weight $w_{a,n} \equiv E[D_i(1-D_j)|\mathcal{F}_{(\psi,p_n),n}]^{-1}\mathbb{1}(i,j \in g_{a,s}) = \frac{k_a(k_a-1)}{q_a(k_a-q_a)}$. By work above, we have

$$S_n - E[S_n|\mathcal{F}_{(\psi,p_n),n}] = (qn)^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\frac{1}{k_a - 1}\sum_{\substack{i,j\in g_{a,s}\\ i\neq j}}F_{i,n}^1 F_{j,n}^0(D_i(1-D_j)w_{a,n} - 1)$$

Therefore, $\mathrm{Var}(S_n|\mathcal{F}_{(\psi,p_n),n})$ is given by the conditional expectation w.r.t. $\mathcal{F}_{(\psi,p_n),n}$ of

$$\sum_{a,a'=1}^{|L_n|}\sum_{s,s'=1}^{n}\frac{(qn)^{-2}}{(k_a-1)^2}\sum_{\substack{i,j\in g_{a,s}\\ i\neq j}}\sum_{\substack{l,f\in g_{a',s'}\\ l\neq f}}F^1_{i,n}F^1_{l,n}F^0_{f,n}F^0_{j,n}(D_i(1-D_j)w_{a,n}-1)(D_l(1-D_f)w_{a',n}-1)$$

By the weight definition $E[D_i(1-D_j)w_{a,n}\mathbb{1}(i,j\in g_{a,s})|\mathcal{F}_{(\psi,p_n),n}]=1$ for all $i\neq j$. Then taking a conditional expectation, the above simplifies to

$$\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\frac{(qn)^{-2}}{(k_a-1)^2}\sum_{\substack{i,j\in g_{a,s}\\ i\neq j}}\sum_{\substack{l,f\in g_{a,s}\\ l\neq f}}F^1_{i,n}F^1_{l,n}F^0_{f,n}F^0_{j,n}\Gamma^{a,n}_{i,j,l,f}$$
$$\Gamma^{a,n}_{i,j,l,f}\equiv E[D_iD_l(1-D_j)(1-D_f)|\mathcal{F}_{(\psi,p_n),n}]w^2_{a,n}-1$$

Case 1: $\{i,j\}\cap\{l,f\}=\emptyset$. Note that in this case we have $k_a\geq 4$. By the representation in Lemma 9.19.(i), we have

$$|\Gamma^{a,n}_{i,j,l,f}|=\left|\binom{k_a-4}{q_a-2}\binom{k_a}{q_a}^{-1}w^2_{a,n}-1\right|=\left|\frac{(q_a-1)(k_a-q_a-1)}{(k_a-2)(k_a-3)}-1\right|\leq 5$$

Using triangle inequality, the corresponding terms of the sum are bounded by

$$\begin{aligned}
&\lesssim\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\frac{(qn)^{-2}}{(k_a-1)^2}\sum_{\substack{i,j,l,f\in g_{a,s}\\ i,j,l,f\text{ distinct}}}|F^1_{i,n}F^1_{l,n}F^0_{f,n}F^0_{j,n}||\Gamma^{a,n}_{i,j,l,f}|\\
&\leq\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\frac{5(qn)^{-2}}{(k_a-1)^2}k_a^3\left(\sum_{i\in g_{a,s}}|F^1_{i,n}|^4+\sum_{i\in g_{a,s}}|F^0_{i,n}|^4\right)\\
&\leq 7\overline{k}_n(qn)^{-1}\left(E_n[T_i|F^1_{i,n}|^4]+E_n[T_i|F^0_{i,n}|^4]\right)\\
&\lesssim\overline{k}_n(qn)^{-1}\left(E_n[|F^1_{i,n}|^4]+E_n[|F^0_{i,n}|^4]\right)=O_p(\overline{k}_nn^{-1})
\end{aligned}$$

The first inequality is by Lemma 9.17, and the second noting $\max_{x\geq 4}\frac{x}{x-1}\leq 4/3$ and $k_a\leq\overline{k}_n$. The third inequality by assumption 9.3, and the final equality follows by conditional Markov, since $E[(F_{i,n})^4|\zeta]<\infty$ by assumption.

Case 2: $i\in\{l,f\}$ or $j\in\{l,f\}$, exclusive. Then we must have $k_a\geq 3$.
Subcase (a): $i=l$, then we have

$$|\Gamma^{a,n}_{i,j,i,f}|=|E[D_i(1-D_j)(1-D_f)|\mathcal{F}_{(\psi,p_n),n}]w^2_{a,n}-1|=\left|\binom{k_a-3}{q_a-1}\binom{k_a}{q_a}^{-1}w^2_{a,n}-1\right|$$

Using algebra, one can show that $|\Gamma^{a,n}_{i,j,i,f}|\leq 2\frac{k_a}{q_a}+1\leq 2\delta^{-1}+1$. Again applying triangle

inequality, these terms in the sum are bounded by

$$\begin{aligned}
&\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\frac{(qn)^{-2}}{(k_a-1)^2}\sum_{\substack{i,j,f\in g_{a,s}\\ i,j,f \text{ distinct}}}|F^1_{i,n}|^2|F^0_{f,n}F^0_{j,n}||\Gamma^{a,n}_{i,j,i,f}|\\
\leq&\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\frac{(2\delta^{-1}+1)(qn)^{-2}}{(k_a-1)^2}\sum_{\substack{i,j,f\in g_{a,s}\\ i,j,f \text{ distinct}}}|F^1_{i,n}|^2|F^0_{f,n}F^0_{j,n}|\\
\leq&\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\frac{(2\delta^{-1}+1)(qn)^{-2}}{(k_a-1)^2}k_a^2\left(\sum_{i\in g_{a,s}}|F^1_{i,n}|^4+\sum_{i\in g_{a,s}}|F^0_{i,n}|^4\right)\\
\lesssim& n^{-1}(E_n[|F^1_{i,n}|^4]+E_n[|F^0_{i,n}|^4])=O_p(n^{-1})
\end{aligned}$$

Subcase (b): $j=f$, then we have

$$\begin{aligned}
|\Gamma^{a,n}_{i,j,l,j}|=|E[D_iD_l(1-D_j)|\mathcal{F}_{(\psi,p_n),n}]w^2_{a,n}-1|&=\left|\binom{k_a-3}{q_a-2}\binom{k_a}{q_a}^{-1}w^2_{a,n}-1\right|\\
\leq\frac{(q_a-1)k_a(k_a-1)}{(k_a-2)q_a(k_a-q_a)}+1\leq 2\frac{k_a}{k_a-q_a}+1&\leq 2\delta^{-1}+1
\end{aligned}$$

Then the relevant terms in the sum are bounded by

$$\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\frac{(2\delta^{-1}+1)(qn)^{-2}}{(k_a-1)^2}\sum_{\substack{i,l,j\in g_{a,s}\\ i,l,j \text{ distinct}}}|F^1_{i,n}||F^1_{l,n}||F^0_{j,n}|^2=O_p(n^{-1})$$

by the same argument as the previous subcase, permuting labels $\{0,1\}$.
Subcase (c): $i=f$, then $\Gamma^{a,n}_{i,j,l,f}=E[D_iD_l(1-D_j)(1-D_i)|\mathcal{F}_{(\psi,p_n),n}]w^2_{a,n}-1=-1$. Then the relevant terms in the sum are bounded by

$$\begin{aligned}
&\sum_{a=3}^{|L_n|}\sum_{s=1}^{n}\frac{(qn)^{-2}}{(k_a-1)^2}\sum_{\substack{i,j,l\in g_{a,s}\\ i,j,l \text{ distinct}}}|F^1_{i,n}||F^0_{i,n}||F^1_{l,n}||F^0_{j,n}|\\
\leq&\sum_{a=3}^{|L_n|}\sum_{s=1}^{n}\frac{(qn)^{-2}}{(k_a-1)^2}(k_a^3+k_a)\sum_{i\in g_{a,s}}|F^1_{i,n}|^4+|F^0_{i,n}|^4\\
\lesssim&\overline{k}_nn^{-1}(E_n[|F^1_{i,n}|^4]+E_n[|F^0_{i,n}|^4])=O_p(\overline{k}_nn^{-1})
\end{aligned}$$

The first inequality is by Lemma 9.17, 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 $\Gamma^{a,n}_{i,j,l,f}=E[D_iD_j(1-D_j)(1-D_f)|\mathcal{F}_{(\psi,p_n),n}]w^2_{a,n}-1=-1$ Then the relevant terms in the sum are bounded by

$$\sum_{a=3}^{|L_n|}\sum_{s=1}^{n}\frac{(qn)^{-2}}{(k_a-1)^2}\sum_{\substack{i,j,f\in g_{a,s}\\ i,j,f \text{ distinct}}}|F^1_{i,n}||F^0_{j,n}||F^1_{j,n}||F^0_{f,n}|=O_p(\overline{k}_nn^{-1})$$

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

$$|\Gamma^{a,n}_{i,j,i,j}| = |E[D_i(1-D_j)|\mathcal{F}_{(\psi,p_n),n}]w^2_{a,n} - 1| = \left|w^{-1}_{a,n}w^2_{a,n} - 1\right| \leq w_{a,n} + 1 \leq \delta^{-2} + 1$$

Then the relevant terms in the sum are bounded by

$$\begin{aligned}
&\sum_{a=1}^{|L_n|}\sum_{s=1}^{n} \frac{(qn)^{-2}(\delta^{-2}+1)}{(k_a-1)^2} \sum_{\substack{i,j\in g_{a,s}\\ i,j \text{ distinct}}} |F^1_{i,n}|^2|F^0_{j,n}|^2 \\
\leq &\sum_{a=1}^{|L_n|}\sum_{s=1}^{n} \frac{(qn)^{-2}(\delta^{-2}+1)k_a}{(k_a-1)^2} \sum_{i\in g_{a,s}} |F^1_{i,n}|^4 + |F^0_{j,n}|^4 \\
\lesssim &\, n^{-1}(E_n[|F^1_{i,n}|^4] + E_n[|F^0_{i,n}|^4]) = O_p(n^{-1})
\end{aligned}$$

The first inequality is by Lemma 9.17, then using $\max_{k\geq 2}\frac{k}{(k-1)^2} = 2$.
Subcase (b): $(i,j) = (f,l)$, then $\Gamma^{a,n}_{i,j,l,f} = E[D_iD_j(1-D_j)(1-D_i)|\mathcal{F}_{(\psi,p_n),n}]w^2_{a,n} - 1 = -1$.
Then the relevant terms are bounded by

$$\begin{aligned}
&\sum_{a=1}^{|L_n|}\sum_{s=1}^{n} \frac{(qn)^{-2}}{(k_a-1)^2} \sum_{\substack{i,j\in g_{a,s}\\ i,j \text{ distinct}}} |F^1_{i,n}||F^0_{i,n}||F^0_{j,n}||F^1_{j,n}| \leq \sum_{a=1}^{|L_n|}\sum_{s=1}^{n} \frac{n^{-2}k_a}{(k_a-1)^2} \sum_{i\in g_{a,s}} |F^1_{i,n}|^4 + |F^0_{j,n}|^4 \\
\leq &2n^{-1}(E_n[|F^1_{i,n}|^4] + E_n[|F^0_{i,n}|^4]) = O_p(n^{-1})
\end{aligned}$$

By Lemma 9.17 again. From the work above and triangle inequality, we see that

$$\begin{aligned}
\mathrm{Var}(S_n|\mathcal{F}_{(\psi,p_n),n}) &\leq \sum_{a=1}^{|L_n|}\sum_{s=1}^{n} \frac{n^{-2}}{(k_a-1)^2} \sum_{\substack{i,j\in g_{a,s}\\ i\neq j}} \sum_{\substack{l,f\in g_{a,s}\\ l\neq f}} |F^1_{i,n}F^1_{l,n}F^0_{f,n}F^0_{j,n}||\Gamma^{a,n}_{i,j,l,f}| \\
&= O_p(n^{-1}) + O_p(\overline{k}_n n^{-1})
\end{aligned}$$

Then by Lemma 9.16, we have the coupling $S_n - E[S_n|\mathcal{F}_{(\psi,p_n),n}] = O_p((\overline{k}_n/n)^{1/2})$.
Together with the result above, we have shown that

$$S_n = E[F^1_iF^0_i|\zeta] + O_p(n^{-1/2}) + O_p(r^F_n) + O_p(r^\psi_n) + O_p((\overline{k}_n/n)^{1/2})$$

Next, we consider the residual terms T_n.

$$\begin{aligned}
T_n &= (qn)^{-1}\sum_{a=1}^{|L_n|} \frac{w_{a,n}}{k_a} \sum_{s=1}^{n}\sum_{i\neq j=1}^{n} D_i(1-D_j)\mathbb{1}(i\in g_{a,s})\mathbb{1}(j\in g_{a,s})R^{1,0}_{i,j} \\
R^{1,0}_{i,j} &= F^0_{j,n}(u^1_{i,n}+\epsilon^1_i) + F^1_{i,n}(u^0_{j,n}+\epsilon^0_j) + (u^1_{i,n}+\epsilon^1_i)(u^0_{j,n}+\epsilon^0_j)
\end{aligned}$$

Again, let $\mathcal{F}_{(\psi,p_n,\tau^d),n} = \sigma(\psi_{1:n}, p_{1:n,n}, \pi^d_n, \pi^t_n, \tau^t, \tau^d, \zeta)$. By the residual moments analyzed above, we have $E[R^{1,0}_{i,j}|\mathcal{F}_{(\psi,p_n,\tau^d),n}] = 0$ for $i \neq j$. Since $D_i(1-D_j)\mathbb{1}(i\in g_{a,s})\mathbb{1}(j\in g_{a,s}) \in \mathcal{F}_{(\psi,p_n,\tau^d),n}$, we have $E[T_n|\mathcal{F}_{(\psi,p_n,\tau^d),n}] = 0$. Next, we consider the conditional variance

$E[T_n^2|\mathcal{F}_{(\psi,p_n,\tau^d),n}]$. This can be written as

$$(qn)^{-2}\sum_{a,a'=1}^{|L_n|}\frac{w_{a,n}w_{a',n}}{(k_a-1)(k_{a',n}-1)}\sum_{s,s'=1}^{n}\sum_{i\neq j=1}^{n}\sum_{l\neq f=1}^{n}\kappa_{i,j,l,f}^{a,a',s,s'}E[R_{i,j}^{1,0}R_{l,f}^{1,0}|\mathcal{F}_{(\psi,p_n,\tau^d),n}]$$
$$\kappa_{i,j,l,f}^{a,a',s,s'}\equiv D_iD_l(1-D_j)(1-D_f)\mathbb{1}(i\in g_{a,s})\mathbb{1}(j\in g_{a,s})\mathbb{1}(l\in g_{a',s'})\mathbb{1}(f\in g_{a',s'})$$

Case 1: $\{i,j\}\cap\{l,f\}=\emptyset$. Then we have $E[R_{i,j}^{1,0}R_{l,f}^{1,0}|\mathcal{F}_{(\psi,p_n,\tau^d),n}]=0$ by the work in our analysis of $\widehat{v}_1$, 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 $D_iD_l(1-D_j)(1-D_f)=D_iD_l(1-D_j)(1-D_i)=0$ and similarly if $j=l$. This leaves the subcases $i=l$ or $j=f$, exclusive. Suppose $i=l$, then

$$\begin{aligned}R_{i,j}^{1,0}R_{i,f}^{1,0}&=F_{f,n}^0(u_{i,n}^1+\epsilon_i^1)[F_{j,n}^0(u_{i,n}^1+\epsilon_i^1)+F_{i,n}^1(u_{j,n}^0+\epsilon_j^0)+(u_{i,n}^1+\epsilon_i^1)(u_{j,n}^0+\epsilon_j^0)]\\&+F_{i,n}^1(u_{f,n}^0+\epsilon_f^0)[F_{j,n}^0(u_{i,n}^1+\epsilon_i^1)+F_{i,n}^1(u_{j,n}^0+\epsilon_j^0)+(u_{i,n}^1+\epsilon_i^1)(u_{j,n}^0+\epsilon_j^0)]\\&+(u_{i,n}^1+\epsilon_i^1)(u_{f,n}^0+\epsilon_f^0)[F_{j,n}^0(u_{i,n}^1+\epsilon_i^1)+F_{i,n}^1(u_{j,n}^0+\epsilon_j^0)+(u_{i,n}^1+\epsilon_i^1)(u_{j,n}^0+\epsilon_j^0)]\end{aligned}$$

Again applying Lemma 9.22, only the first cross moments remains, leaving

$$\begin{aligned}|E[R_{i,j}^{1,0}R_{i,f}^{1,0}|\mathcal{F}_{(\psi,p_n,\tau^d),n}]|&=|E[F_{f,n}^0F_{j,n}^0(u_{i,n}^1+\epsilon_i^1)^2|\mathcal{F}_{(\psi,p_n,\tau^d),n}]|\\&=|F_{f,n}^0F_{j,n}^0|(\sigma_{u^1,n}^2(X_i)+E[\sigma_1^2(X_i)|\psi_i,p_{i,n}])\leq 2|F_{f,n}^0F_{j,n}^0|\overline{\sigma}^2\end{aligned}$$

Suppose $j=f$. Then reasoning similarly the only non-zero cross-moment is

$$|E[R_{i,j}^{1,0}R_{l,j}^{1,0}|\mathcal{F}_{(\psi,p_n,\tau^d),n}]|=|E[F_{l,n}^1F_{i,n}^1(u_{j,n}^0+\epsilon_j^0)^2|\mathcal{F}_{(\psi,p_n,\tau^d),n}]|\leq 2|F_{l,n}^1F_{i,n}^1|\overline{\sigma}^2$$

Case 3: $\{i,j\}=\{l,f\}$. Note that $D_iD_j(1-D_i)(1-D_j)=0$, so we need only consider the case where $i=l$ and $j=f$. We have

$$\begin{aligned}(R_{i,j}^{1,0})^2&=(F_{j,n}^0(u_{i,n}^1+\epsilon_i^1)+F_{i,n}^1(u_{j,n}^0+\epsilon_j^0)+(u_{i,n}^1+\epsilon_i^1)(u_{j,n}^0+\epsilon_j^0))^2\\&=(F_{j,n}^0)^2(u_{i,n}^1+\epsilon_i^1)^2+(F_{i,n}^1)^2(u_{j,n}^0+\epsilon_j^0)^2+(u_{i,n}^1+\epsilon_i^1)^2(u_{j,n}^0+\epsilon_j^0)^2\\&+2\left(F_{j,n}^0F_{i,n}^1(u_{i,n}^1+\epsilon_i^1)(u_{j,n}^0+\epsilon_j^0)+F_{j,n}^0(u_{i,n}^1+\epsilon_i^1)^2(u_{j,n}^0+\epsilon_j^0)+F_{i,n}^1(u_{j,n}^0+\epsilon_j^0)^2(u_{i,n}^1+\epsilon_i^1)\right)\end{aligned}$$

As in our analysis of $\widehat{v}_1$, using Lemma 9.22, all the cross terms are conditionally mean zero. Also

$$E[u_{i,n}^1\epsilon_i^1|\mathcal{F}_{(\psi,p_n,\tau^d),n}]=E[u_{i,n}^1\epsilon_i^1(u_{j,n}^0+\epsilon_j^0)^2|\mathcal{F}_{(\psi,p_n,\tau^d),n}]=0$$

Using these observations and application of Lemma 9.22 to compute higher order moments as in the previous analysis, $E[(R_{i,j}^{1,0})^2|\mathcal{F}_{(\psi,p_n,\tau^d),n}]$ simplifies as

$$\begin{aligned}&(F_{j,n}^0)^2(\sigma_{u^1,n}^2(X_i)+E[\sigma_1^2(X_i)|\psi_i,p_{i,n},\zeta])+(F_{i,n}^1)^2(\sigma_{u^0,n}^2(X_j)+E[\sigma_0^2(X_j)|\psi_j,p_{j,n},\zeta])\\&+\sigma_{u^1,n}^2(X_i)\sigma_{u^0,n}^2(X_j)+\sigma_{u^1,n}^2(X_i)E[\sigma_0^2(X_j)|\psi_j,p_{j,n},\zeta]+E[\sigma_1^2(X_i)|\psi_i,p_{i,n},\zeta]\sigma_{u^0,n}^2(X_j)\\&+E[\sigma_1^2(X_i)|\psi_i,p_{i,n},\zeta]E[\sigma_0^2(X_j)|\psi_j,p_{j,n},\zeta]\leq 2(F_{j,n}^0)^2\overline{\sigma}^2+2(F_{i,n}^1)^2\overline{\sigma}^2+4(\overline{\sigma}^2)^2\end{aligned}$$

From the discussion above, we can simplify

$$\begin{aligned}
E[T_n^2|\mathcal{F}_{(\psi,p_n,\tau^d),n}] &= (qn)^{-2}\sum_{a=1}^{|L_n|}\frac{w_{a,n}^2}{(k_a-1)^2}\sum_{s=1}^{n}\sum_{i\neq j=1}^{n}\kappa_{i,j,i,j}^{a,a,s,s}E[(R_{i,j}^{1,0})^2|\mathcal{F}_{(\psi,p_n,\tau^d),n}] \\
&+ (qn)^{-2}\sum_{a=1}^{|L_n|}\frac{w_{a,n}^2}{(k_a-1)^2}\sum_{s=1}^{n}\sum_{i\neq j=1}^{n}\sum_{f\notin\{i,j\}}\kappa_{i,j,i,f}^{a,a,s,s}E[R_{i,j}^{1,0}R_{i,f}^{1,0}|\mathcal{F}_{(\psi,p_n,\tau^d),n}] \\
&+ (qn)^{-2}\sum_{a=1}^{|L_n|}\frac{w_{a,n}^2}{(k_a-1)^2}\sum_{s=1}^{n}\sum_{i\neq j=1}^{n}\sum_{l\notin\{i,j\}}\kappa_{i,j,l,j}^{a,a,s,s}E[R_{i,j}^{1,0}R_{l,j}^{1,0}|\mathcal{F}_{(\psi,p_n,\tau^d),n}] \\
&\equiv \overline{T}_{n,1}+\overline{T}_{n,2}+\overline{T}_{n,3}
\end{aligned}$$

Note the weight bound

$$w_{a,n}^2 = \frac{k_a^2(k_a-1)^2}{q_a^2(k_a-q_a)^2} \leq \frac{k_a^4}{q_a^2(k_a-q_a)^2} = \frac{1}{p_a^2(1-p_a)^2} \leq \delta^{-4}$$

Then we have

$$\begin{aligned}
\overline{T}_{n,1} &= (qn)^{-2}\sum_{a=1}^{|L_n|}\frac{w_{a,n}^2}{(k_a-1)^2}\sum_{s=1}^{n}\sum_{i\neq j=1}^{n}D_i(1-D_j)\mathbb{1}(i,j\in g_{a,s})E[(R_{i,j}^{1,0})^2|\mathcal{F}_{(\psi,p_n,\tau^d),n}] \\
&\leq \delta^{-4}(qn)^{-2}\sum_{a=1}^{|L_n|}\frac{1}{(k_a-1)^2}\sum_{s=1}^{n}\sum_{i\neq j=1}^{n}D_i(1-D_j)\mathbb{1}(i,j\in g_{a,s})\left[2(F_{j,n}^0)^2\overline{\sigma}^2+2(F_{i,n}^1)^2\overline{\sigma}^2+4(\overline{\sigma}^2)^2\right]
\end{aligned}$$

The first term in the brackets is bounded by

$$\begin{aligned}
&\lesssim n^{-2}\sum_{a=1}^{|L_n|}\frac{1}{(k_a-1)^2}\sum_{s=1}^{n}\sum_{j=1}^{n}(1-D_j)\mathbb{1}(j\in g_{a,s})(F_{j,n}^0)^2\sum_{\substack{i=1\\i\neq j}}^{n}D_i\mathbb{1}(i\in g_{a,s}) \\
&\leq n^{-2}\sum_{a=1}^{|L_n|}\frac{1}{(k_a-1)^2}\sum_{s=1}^{n}\sum_{j=1}^{n}(1-D_j)\mathbb{1}(j\in g_{a,s})(F_{j,n}^0)^2q_a\mathbb{1}(g_{a,s}\neq\emptyset) \\
&= n^{-2}\sum_{a=1}^{|L_n|}\frac{q_a}{(k_a-1)^2}\sum_{s=1}^{n}\sum_{j=1}^{n}(1-D_j)\mathbb{1}(j\in g_{a,s})(F_{j,n}^0)^2 \\
&\leq n^{-2}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{j=1}^{n}(1-D_j)\mathbb{1}(j\in g_{a,s})(F_{j,n}^0)^2 = n^{-1}E_n[T_i(1-D_j)(F_{j,n}^0)^2] = O_p(n^{-1})
\end{aligned}$$

Identical reasoning can be used to show that

$$n^{-2}\sum_{a=1}^{|L_n|}\frac{1}{(k_a-1)^2}\sum_{s=1}^{n}\sum_{i\neq j=1}^{n}D_i(1-D_j)\mathbb{1}(i,j\in g_{a,s})2(F_{i,n}^1)^2\overline{\sigma}^2 = O_p(n^{-1})$$

The last term in the bracket gives

$$\begin{aligned}
&4\delta^{-4}n^{-2}\sum_{a=1}^{|L_n|}\frac{1}{(k_a-1)^2}\sum_{s=1}^{n}\sum_{i\neq j=1}^{n}D_i(1-D_j)\mathbb{1}(i,j\in g_{a,s})(\overline{\sigma}^2)^2\\
\lesssim\ & n^{-2}\sum_{a=1}^{|L_n|}\frac{q_a(k_a-q_a)}{(k_a-1)^2}\left[\left(\sum_{s=1}^{n}\mathbb{1}(g_{a,s}\neq\emptyset)\right)+1\right]\\
\leq\ & n^{-2}\sum_{a=1}^{|L_n|}\left[\left(\sum_{s=1}^{n}\mathbb{1}(g_{a,s}\neq\emptyset)\right)+1\right]\leq 4\delta^{-4}(\overline{\sigma}^2)^2n^{-2}[n+|L_n|]\\
=\ & O_p(n^{-1})+O_p(|L_n|n^{-2})
\end{aligned}$$

Summarizing, we have shown that $\overline{T}_{n,1}=O_p(n^{-1})$. Next, consider $\overline{T}_{n,2}$. By triangle inequality

$$\begin{aligned}
|\overline{T}_{n,2}|&\leq (qn)^{-2}\sum_{a=1}^{|L_n|}\frac{w_{a,n}^2}{(k_a-1)^2}\sum_{s=1}^{n}\sum_{i\neq j=1}^{n}\sum_{f\notin\{i,j\}}\kappa_{i,j,i,f}^{a,a,s,s}|E[R_{i,j}^{1,0}R_{i,f}^{1,0}|\mathcal{F}_{(\psi,p_n,\tau^d),n}]|\\
&\lesssim n^{-2}\sum_{a=1}^{|L_n|}\frac{w_{a,n}^2}{(k_a-1)^2}\sum_{s=1}^{n}\sum_{i\neq j=1}^{n}\sum_{f\notin\{i,j\}}D_i(1-D_j)(1-D_f)\mathbb{1}(i,j,f\in g_{a,s})|F_{f,n}^0F_{j,n}^0|
\end{aligned}$$

Then $E[|\overline{T}_{n,2}||\mathcal{F}_{(\psi,p_n),n}]$ is bounded above by

$$\begin{aligned}
&\sum_{a=1}^{|L_n|}\frac{n^{-2}w_{a,n}^2}{(k_a-1)^2}\sum_{s=1}^{n}\sum_{i\neq j=1}^{n}\sum_{f\notin\{i,j\}}E[D_i(1-D_j)(1-D_f)|\mathcal{F}_{(\psi,p_n),n}]\mathbb{1}(i,j,f\in g_{a,s})|F_{f,n}^0F_{j,n}^0|\\
=&\sum_{a=1}^{|L_n|}\frac{n^{-2}w_{a,n}^2}{(k_a-1)^2}\sum_{s=1}^{n}\sum_{i\neq j=1}^{n}\sum_{f\notin\{i,j\}}\frac{q_a(k_a-q_a)(k_a-q_a-1)}{k_a(k_a-1)(k_a-2)}\mathbb{1}(i,j,f\in g_{a,s})|F_{f,n}^0F_{j,n}^0|
\end{aligned}$$

We can simplify the (a,n) weight as

$$\begin{aligned}
&\frac{w_{a,n}^2q_a(k_a-q_a)(k_a-q_a-1)}{k_a(k_a-1)^3(k_a-2)}=\frac{k_a^2(k_a-1)^2(k_a-q_a)(k_a-q_a-1)}{k_a(k_a-1)^3(k_a-2)q_a^2(k_a-q_a)^2}\\
=&\frac{k_a(k_a-q_a-1)}{(k_a-1)(k_a-2)q_a^2(k_a-q_a)}
\end{aligned}$$

The inner summand is

$$\sum_{s=1}^{n}\sum_{i\neq j=1}^{n}\sum_{f\notin\{i,j\}}\mathbb{1}(i,j,f\in g_{a,s})|F^0_{f,n}F^0_{j,n}| = \sum_{s=1}^{n}\sum_{j\neq f}^{n}\mathbb{1}(j,f\in g_{a,s})|F^0_{f,n}F^0_{j,n}|\sum_{i\notin\{j,f\}}\mathbb{1}(i\in g_{a,s})$$
$$=\sum_{s=1}^{n}(k_a-2)\sum_{j\neq f}^{n}\mathbb{1}(j,f\in g_{a,s})|F^0_{f,n}F^0_{j,n}| \leq \sum_{s=1}^{n}(k_a-2)\left(\sum_{j\in g_{a,s}}|F^0_{j,n}|\right)^2$$
$$=\sum_{s=1}^{n}(k_a-2)k_a^2\left(k_a^{-1}\sum_{j\in g_{a,s}}|F^0_{j,n}|\right)^2 \leq \sum_{s=1}^{n}(k_a-2)k_a^2k_a^{-1}\sum_{j\in g_{a,s}}(F^0_{j,n})^2$$

The first inequality is by adding and subtracting the diagonal terms, and the second is by Jensen. Using the two preceding displays, we have

$$\begin{aligned}E[|\overline{T}_{n,2}||\mathcal{F}_{(\psi,p_n),n}] &\lesssim \sum_{a=1}^{|L_n|}\frac{n^{-2}k_a^2(k_a-2)(k_a-q_a-1)}{(k_a-1)(k_a-2)q_a^2(k_a-q_a)}\sum_{s=1}^{n}\sum_{j\in g_{a,s}}(F^0_{j,n})^2\\ &\leq n^{-2}\sum_{a=1}^{|L_n|}\frac{2k_a}{q_a^2}\sum_{s=1}^{n}\sum_{j\in g_{a,s}}(F^0_{j,n})^2 \leq \delta^{-1}n^{-2}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{j\in g_{a,s}}(F^0_{j,n})^2\\ &\lesssim n^{-1}E_n[T_i(F^0_{j,n})^2] = O_p(n^{-1})\end{aligned}$$

The first inequality simplifies, using $\max_{k\geq 2}\frac{k}{k-1}=2$ and $\max_{k\geq 2}\frac{k-1}{k}\leq 1$. The second inequality uses $k_a/q_a=\frac{1}{p_a}\leq\delta^{-1}$ and $q_a^{-1}\leq 1$. By Lemma 9.16, $\overline{T}_{n,2}=O_p(n^{-1})$. A very similar argument shows that $\overline{T}_{n,3}=O_p(n^{-1})$. Putting this together, we have shown that $E[T_n^2|\mathcal{F}_{(\psi,p_n,\tau^d),n}]=\overline{T}_{n,1}+\overline{T}_{n,2}+\overline{T}_{n,3}=O_p(n^{-1})$, so that $T_n=O_p(n^{-1/2})$ by Lemma 9.16.

Summarizing, we have shown that

$$\begin{aligned}\widehat{v}_{10} &= S_n+T_n = E[S_n|\mathcal{F}_{(\psi,p_n),n}]+(S_n-E[S_n|\mathcal{F}_{(\psi,p_n),n}])+T_n\\ &= E[F_i^1F_i^0|\zeta]+O_p(n^{-1/2})+O_p(r_n^F)+O_p(r_n^\psi)+O_p((\overline{k}_n/n)^{1/2})+O_p\left(\sqrt{|L_n|}/n\right)\end{aligned}$$

This completes the proof of the lemma. □

9.5 Propositions

Definition 9.13 (Conditional Weak Convergence)**.** For random variables $A_n\in\mathbb{R}^d$ and σ-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 9.14 (MDS-CLT)**.** *Consider probability spaces $(\Omega_n,\mathcal{G}_n,P_n)$ each equipped with filtration $(\mathcal{F}_{k,n})_{k\geq 0}$. Suppose $(Y_{k,n})_{k=1}^{r_n}$ 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*

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

Denote $S_n \equiv S_{r_n,n}$ *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* $\epsilon > 0$

$$\sum_{k=1}^{n} E[Y^2_{k,n}\mathbb{1}(|Y_{k,n}| \geq \epsilon)|\mathcal{F}_{0,n}] = o_p(1) \tag{9.8}$$

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

Proof. We may compute

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

For the first term, by conditional Jensen inequality

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

The first equality since $\Sigma_n \in \mathcal{F}_{0,n}$. 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

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

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 $Z_n = \sum_{k=1}^{n} E[|E[(e^{itY_{k,n}} - e^{-\frac{1}{2}t^2\sigma^2_{k,n}})|\mathcal{F}_{k-1,n}]||\mathcal{F}_{0,n}]$, and note that this is positive. For the final inequality, let $c > \sigma^2$ and note that

$$\begin{aligned}(e^{\frac{1}{2}t^2\Sigma_n} - e^{\frac{1}{2}t^2c})Z_n &\leq (e^{\frac{1}{2}t^2\Sigma_n} - e^{\frac{1}{2}t^2c})Z_n\mathbb{1}(\Sigma_n \geq c) \\ &\leq 2n(e^{\frac{1}{2}t^2\Sigma_n} - e^{\frac{1}{2}t^2c})\mathbb{1}(\Sigma_n \geq c) = o_p(1)\end{aligned}$$

The first inequality since $Z_n \geq 0$. The second is by (conditional) Jensen inequality. For the final equality

$$P\left(2n(e^{\frac{1}{2}t^2\Sigma_n} - e^{\frac{1}{2}t^2c})\mathbb{1}(\Sigma_n \geq c) > \epsilon\right) \leq P\left(\Sigma_n \geq c\right) = o_p(1)$$

Fix $\epsilon > 0$ and let $I_{k,n} = \mathbb{1}(|Y_{k,n}| > \epsilon)$. The same argument as in BL shows that for $K_t \geq t^2 \vee |t|^3$

$$|E[(e^{itY_{k,n}} - e^{-\frac{1}{2}t^2\sigma_{k,n}^2})|\mathcal{F}_{k-1,n}]| \leq K_t(\sigma_{k,n}^4 + Y_{k,n}^2 I_{k,n} + \epsilon\sigma_{k,n}^2)$$

so that the sum above is

$$\begin{aligned}
&\sum_{k=1}^{r_n} E[|E[(e^{itY_{k,n}} - e^{-\frac{1}{2}t^2\sigma_{k,n}^2})|\mathcal{F}_{k-1,n}]||\mathcal{F}_{0,n}] \leq K_t \sum_{k=1}^{r_n} E[\sigma_{k,n}^4 + Y_{k,n}^2 I_{k,n} + \epsilon\sigma_{k,n}^2|\mathcal{F}_{0,n}] \\
=&K_t E\left[\sum_{k=1}^{r_n} \sigma_{k,n}^4 + Y_{k,n}^2 I_{k,n} + \epsilon\sigma_{k,n}^2|\mathcal{F}_{0,n}\right] \leq K_t E\left[\Sigma_n(\max_k \sigma_{k,n}^2 + \epsilon) + \sum_{k=1}^{r_n} Y_{k,n}^2 I_{k,n}|\mathcal{F}_{0,n}\right] \\
\leq&K_t E\left[\Sigma_n\left(\epsilon^2 + \sum_{k=1}^{r_n} E[Y_{k,n}^2 I_{k,n}|\mathcal{F}_{k-1,n}] + \epsilon\right) + \sum_{k=1}^{r_n} Y_{k,n}^2 I_{k,n}|\mathcal{F}_{0,n}\right] \\
=&K_t\Sigma_n(\epsilon^2 + \epsilon) + K_t(\Sigma_n + 1)\sum_{k=1}^{r_n} E[Y_{k,n}^2 I_{k,n}|\mathcal{F}_{0,n}] \leq K_t c(\epsilon^2 + \epsilon) + K_t(c+1)o_p(1) + o_p(1)
\end{aligned}$$

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 Σ_n. The third inequality follows as in BL by taking $\max_k$ on both sides of

$$\sigma_{k,n}^2 = E[Y_{k,n}^2|\mathcal{F}_{k-1,n}] \leq E[\epsilon^2 + Y_{k,n}^2 I_{k,n}|\mathcal{F}_{k-1,n}] \leq \epsilon^2 + \sum_{k=1}^{n} E[Y_{k,n}^2 I_{k,n}|\mathcal{F}_{k-1,n}]$$

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) $\Sigma_n = c + \Sigma_n - c \leq c + (\Sigma_n - c)_+\mathbb{1}(\Sigma_n \geq c) = c + o_p(1)$ as shown before and (2) from the assumed Lindberg condition in probability. Then we have shown that $E\left[e^{itS_n} - e^{-\frac{1}{2}t^2\sigma^2}|\mathcal{F}_{0,n}\right] = o_p(1)$. □

Proposition 9.15 (SBR Equivalence)**.** *Suppose $n = \ell k$ for $\ell \in \mathbb{N}$. Then* $\mathrm{Loc_n}(1, a/k) = \mathrm{CR}(a/k)$ *for any* $1 \leq a \leq k-1$

Let $\tau \in \mathcal{S}_n$ and $g(\tau) = 1/n!$. 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 $D_{1:n} \sim \mathrm{Loc_n}(1, a/k)$ by definition

$$\begin{aligned}
P(D_{1:n} = d_{1:n}) &= \sum_{\tau\in\mathcal{S}_n} P(D_{1:n} = d_{1:n}|\tau)g(\tau) = \sum_{\tau\in\mathcal{S}_n}\prod_{t=1}^{\ell} P(D_{\tau^{-1}(g_t)} = d_{\tau^{-1}(g_t)}|\tau)g(\tau) \\
&= \sum_{\tau\in\mathcal{S}_n}\prod_{t=1}^{\ell}\sum_{\xi_t\in\{0,1\}^k} P(D_{\tau^{-1}(g_t)} = d_{\tau^{-1}(g_t)}|\xi_t, \tau)g(\xi_t|\tau)g(\tau) \\
&= \sum_{\tau\in\mathcal{S}_n}\prod_{t=1}^{\ell}\sum_{\xi_t\in\{0,1\}^k} P(D_{\tau^{-1}(g_t)} = d_{\tau^{-1}(g_t)}|\xi_t, \tau)g(\xi_t)g(\tau)
\end{aligned}$$

Where by abuse of notation we also denote $g(\xi_t) = \binom{k}{a}^{-1}\mathbb{1}(\#\{i : \xi_{it} = 1\} = a)$. 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

determined by $((\xi_t)_t, \tau)$. In particular, there exists H deterministic s.t. $D_{1:n} = H((\xi_t)_t, \tau)$ with $(\xi_t)_t \perp\!\!\!\perp \tau$.

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

$$(D_i) = H((\xi_t)_t, \tau) \stackrel{d}{=} H((\xi_t)_t, \pi \circ \tau) = (D_{\pi^{-1}(i)})$$

The second equality follows since $\tau \sim \text{Unif}(\mathcal{S}_n)$, so $\pi \circ \tau \stackrel{d}{=} \tau$ and because $(\xi_t)_t \perp\!\!\!\perp \tau$. Putting this together, we see that $P(D_{1:n} = d') = P(D_{1:n} = d)$. Since d, d' were arbitrary

$$\text{Loc}_\text{n}(1, a/k) = \text{Unif}(\{d_{1:n} : E_n[d_i] = a/k\}) = \text{CR}(a/k)$$

9.6 Lemmas

Lemma 9.16 (Conditional Convergence)**.** *Let $(\mathcal{G}_n)_{n\geq 1}$ and $(A_n)_{n\geq 1}$ a sequence of σ-algebras and RV's. Define conditional convergence*

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

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* $(c_n)_n$

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

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

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

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

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

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

Lemma 9.17 (Product Diagonalization)**.** *Consider positive scalar arrays* $(x_i)_{i=1}^k$, $(y_i)_{i=1}^k$. *Then the following inequalities are valid*

$$\sum_{\substack{i,j,l,f=1\\ i,j,l,f \text{ distinct}}}^{k} x_i x_l y_j y_f \leq k^3 \sum_{i=1}^{k} x_i^4 + k^3 \sum_{j=1}^{k} y_j^4; \qquad \sum_{\substack{i,j,f=1\\ i,j,f \text{ distinct}}}^{k} x_i^2 y_j y_f \leq k^2 \sum_{i=1}^{k} x_i^4 + k^2 \sum_{j=1}^{k} y_j^4$$

$$\sum_{\substack{i,j,l=1\\ i,j,l \text{ distinct}}}^{k} x_i y_i x_l y_j \leq k \sum_{i=1}^{k} x_i^4 + k \sum_{i=1}^{k} y_i^4 + k^3 \sum_{i=1}^{k} x_l^4 + k^3 \sum_{j=1}^{k} y_j^4$$

$$\sum_{\substack{i,j=1\\ i,j \text{ distinct}}}^{k} x_i^2 y_j^2 \leq k \sum_{i=1}^{k} x_i^4 + k \sum_{j=1}^{k} y_j^4; \qquad \sum_{\substack{i,j=1\\ i,j \text{ distinct}}}^{k} x_i y_i x_j y_j \leq k \sum_{i=1}^{k} x_i^4 + k \sum_{j=1}^{k} y_j^4$$

Proof. Let $a \geq 1$, so that $\varphi(x) = x^a$ is convex on $\mathbb{R}_{\geq 0}$. Note that by Jensen's inequality

$$\left(\sum_{i=1}^{k} x_i \right)^a = k^a \left(k^{-1} \sum_{i=1}^{k} x_i \right)^a \leq k^a k^{-1} \sum_{i=1}^{k} x_i^a = k^{a-1} \sum_{i=1}^{k} x_i^a$$

For the first statement

$$\begin{aligned} \sum_{\substack{i,j,l,f=1\\ i,j,l,f \text{ distinct}}}^{k} x_i x_l y_j y_f &\leq \left(\sum_{i,j=1}^{k} x_i y_j \right)^2 = \left(\left(\sum_{i=1}^{k} x_i \right) \left(\sum_{j=1}^{k} y_j \right) \right)^2 \leq \left(\sum_{i=1}^{k} x_i \right)^4 + \left(\sum_{j=1}^{k} y_j \right)^4 \\ &\leq k^3 \sum_{i=1}^{k} x_i^4 + k^3 \sum_{j=1}^{k} y_j^4 \end{aligned}$$

The first inequality by positivity, the second by Young's inequality, and the third by the fact above. Next we have

$$\begin{aligned} \sum_{\substack{i,j,f=1\\ i,j,f \text{ distinct}}}^{k} x_i^2 y_j y_f &\leq \left(\sum_{i=1}^{k} x_i^2 \right) \left(\sum_{j=1}^{k} y_j \right)^2 \leq k \left(\sum_{i=1}^{k} x_i^2 \right) \left(\sum_{j=1}^{k} y_j^2 \right) \\ &\leq k \left(\sum_{i=1}^{k} x_i^2 \right)^2 + k \left(\sum_{j=1}^{k} y_j^2 \right)^2 \leq k^2 \sum_{i=1}^{k} x_i^4 + k^2 \sum_{j=1}^{k} y_j^4 \end{aligned}$$

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 $abcd \leq (\frac{1}{4}(a+b+c+d))^4 \leq \frac{1}{4}(a^4+b^4+c^4+d^4)$. Then we have

$$\sum_{\substack{i,j,l=1 \\ i,j,l \text{ distinct}}}^{k} x_i y_i x_l y_j \leq \left(\sum_{i=1}^{k} x_i y_i\right)\left(\sum_{l=1}^{k} x_l\right)\left(\sum_{j=1}^{k} y_j\right)$$

$$\leq \left(\sum_{i=1}^{k} x_i^2\right)^{1/2}\left(\sum_{i=1}^{k} y_i^2\right)^{1/2}\left(\sum_{l=1}^{k} x_l\right)\left(\sum_{j=1}^{k} y_j\right) \leq \|x\|_2\|y\|_2\left(\sum_{l=1}^{k} x_l\right)\left(\sum_{j=1}^{k} y_j\right)$$

$$\leq \|x\|_2^4 + \|y\|_2^4 + \left(\sum_{l=1}^{k} x_l\right)^4 + \left(\sum_{j=1}^{k} y_j\right)^4 \leq k\sum_{i=1}^{k} x_i^4 + k\sum_{i=1}^{k} y_i^4 + k^3\sum_{i=1}^{k} x_l^4 + k^3\sum_{j=1}^{k} y_j^4$$

Next we have

$$\sum_{\substack{i,j=1 \\ i,j \text{ distinct}}}^{k} x_i^2 y_j^2 \leq \left(\sum_{i=1}^{k} x_i^2\right)\left(\sum_{j=1}^{k} y_j^2\right) \leq \left(\sum_{i=1}^{k} x_i^2\right)^2 + \left(\sum_{j=1}^{k} y_j^2\right)^2 \leq k\sum_{i=1}^{k} x_i^4 + k\sum_{j=1}^{k} y_j^4$$

Next we have

$$\sum_{\substack{i,j=1 \\ i,j \text{ distinct}}}^{k} x_i y_i x_j y_j \leq \left(\sum_{i=1}^{k} x_i y_i\right)^2 \leq \|x\|^2\|y\|^2 \leq \|x\|^4 + \|y\|^4 \leq k\sum_{i=1}^{k} x_i^4 + k\sum_{j=1}^{k} y_j^4$$

□

Lemma 9.18 (Discretization Properties)**.** *Suppose assumption 9.1 holds with $\zeta_n = \zeta$ for $n \geq 1$, $\zeta \perp\!\!\!\perp W_{1:n}$, and $\widehat{p}_{i,n} = p_{i,n} = p(X_i, \zeta)$. Suppose G is bounded, Lipschitz and H a function with $E[H(W_i,\zeta)^2|\zeta] < \infty$, ζ-a.s. Then we have*

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

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

$$\begin{aligned}|E_n[H_iG(p_{i,n})] - E_n[H_iG(p_i)]| &\leq E_n\left[|H_i||G(p_{i,n}) - G(p_i)|\right] \leq BE_n\left[|H_i||p_{i,n} - p_i|\right] \\ &\lesssim |p_n - p|_\infty E_n[|H_i|] = O_p(r_n^p)O_p(1)\end{aligned}$$

Now note that $E[E_n[H(W_i,\zeta)G(p_i)]|\zeta] = E[H(W_i,\zeta)G(p_i)|\zeta]$ and

$$\begin{aligned}&\mathrm{Var}\left(E_n[H(W_i,\zeta)G(p(X_i,\zeta))]|\zeta\right) = n^{-2}\sum_{i,j}\mathrm{Cov}(H_iG(p_{i,n}), H_{j,n}G(p_{j,n})|\zeta) \\ &= n^{-1}E_n\left[\mathrm{Var}(H_iG(p_{i,n})|\zeta)\right] \lesssim n^{-1}E[H_i^2|\zeta] = O_p(n^{-1})\end{aligned}$$

The second equality by Lemma 9.21 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 9.16), this shows the claim. □

Lemma 9.19 (Design Properties)**.** *Let $D_{1:n} \sim \mathrm{Gr}_n((g_{a,s})_{a,s}, L_n)$. Let $\tau^d = (\tau^d_{a,s})_{a,s}$ jointly independent with $(\tau^d_{a,s,\ell})_{\ell=1}^{k_a} \sim \mathrm{CR}(q_a/k_a)$ for $1 \leq s \leq n-1$ and $(\tau^d_{a,n,\ell})_{\ell=1}^{k_a} \sim \mathrm{SRS}(q_a/k_a)$.*

Let $(\mathcal{F}_n)_{n\geq 0}$ a sequence of σ-algebras with $\sigma((g_{a,s})_{a,s}) \subseteq \mathcal{F}_n$ and $\mathcal{F}_n \perp\!\!\!\perp \tau^d$. 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

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

In particular, $E[D_i \mathbb{1}(i \in g_{a,s})|\mathcal{F}_n] = \mathbb{1}(i \in g_{a,s}) \cdot p_a$.

(iii) For $1 \leq i \leq n$ and $1 \leq s \leq n$

$$\mathrm{Var}(D_i|\mathcal{F}_n)\mathbb{1}(i \in g_{a,s}) = p_a(1-p_a)\mathbb{1}(i \in g_{a,s})$$

For $1 \leq i, j \leq n$ distinct indices and $1 \leq s \leq n-1$

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

For $1 \leq i, j \leq n$ distinct indices and $s = n$

$$\mathrm{Cov}(D_i, D_j|\mathcal{F}_n)\mathbb{1}(i,j \in g_{a,s}) = 0$$

For $1 \leq i, j \leq n$ distinct indices

$$\mathrm{Cov}(D_i, D_j|\mathcal{F}_n)\mathbb{1}(g(i) \neq g(j)) = 0$$

In particular we have

$$\mathrm{Var}(D_i|\mathcal{F}_n)\mathbb{1}(i \in g_{a,s})| \leq (1/4)\mathbb{1}(i \in g_{a,s})$$
$$|\mathrm{Cov}(D_i, D_j|\mathcal{F}_n)\mathbb{1}(i,j \in g_{a,s})| \leq k_a^{-1}\mathbb{1}(i,j \in g_{a,s})\mathbb{1}(s \neq n)$$

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

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

The second equality by the first containment. The third equality is by independence. Then note that $\sum_{\ell=1}^{k_a} \mathbb{1}(i = g_{a,s,\ell}) = \mathbb{1}(i \in g_{a,s})$.

For (iii), by the decomposition above and σ-algebra assumption, for $1 \leq j \neq i \leq n$ we have

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

By σ-algebra independence and joint independence of groupwise randomizations

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

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

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

The bounds follow by inspection. □

Lemma 9.20 (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 $D_{1:n} \sim \mathrm{Loc_n}(\psi, p)$ and require $\mathcal{F}_n \perp\!\!\!\perp \tau^d$. Let $S_{1:n} \in \{0,1\}^n$ such that $\{i : S_i = 1\} = \bigsqcup_{a,s} g_{a,s}$. 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) $\mathrm{Var}(E_n[S_i(D_i - p_n(X_i)))h_n(W_i)]|\mathcal{F}_n) \leq 2n^{-1}E_n[S_i h_n(W_i)^2] \leq 2n^{-1}E_n[h_n(W_i)^2]$

(4) If $\sup_{n\geq 1} E[h_n(W)^2] < \infty$ 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

$$\mathrm{Var}(E_n[S_i(D_i - p_n(X_i)))h_n(W_i)]|\mathcal{F}_n) \leq n^{-2}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n} k_a^{-1} \sum_{\substack{i,j\in g_{a,s}\\ i\neq j}} (h_n(W_i) - h_n(W_j))^2$$
$$+ n^{-2}\overline{k}_n|L_n| \cdot \max_{i=1}^{n} h_n(W_i)^2$$

Proof. For (1), by assumption $\{i : S_i = 1\} = \bigsqcup_{a,s} g_{a,s}$, so we have

$$E_n[S_i(D_i - p_{i,n})h_n(W_i)] = n^{-1}\sum_{i:S_i=1}(D_i - p_{i,n})h_n(W_i) = n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{i\in g_{a,s}}(D_i - p_{i,n})h_n(W_i)$$
$$= n^{-1}\sum_{a=1}^{|L_n|}\sum_{s=1}^{n}\sum_{i=1}^{n}(D_i - p_a)h_n(W_i)\mathbb{1}(i \in g_{a,s})$$

For (2), by Lemma 9.19.(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, $\mathrm{Var}(E_n[(D_i - p_{i,n})h_n(W_i)]|\mathcal{F}_n)$ is

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

The final equality follows from Lemma 9.19.(iii). By triangle inequality and the covariance bound in Lemma 9.19.(iii), this is bounded above by

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

Continuing the calculation

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

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 $(\mathcal{F}_n)_{n\ge1}$ any sequence satisfying the conditions, followed by conditional Markov inequality (Lemma 9.16).

For the final identity (5), note that from Lemma 9.19.(iii) for $s \neq n$

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

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

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

Substituting in the first display above, the diagonal terms cancel. For the claimed constant, note that $\max_{p\in(0,1)}p(1-p) \le 1/4$ and $\max_{k\ge2}\frac{k}{k-1} \le 2$, so $\frac{q_a(k_a-q_a)}{k_a^2(k_a-1)} \le k_a^{-1}$. Aggregating over (a,s) gives

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

The second term is

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

This finishes the proof. □

Lemma 9.21 (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 (ξ_n, ζ_n) such that $(W_{1:n}, \xi_n, \zeta_n)$ are jointly independent. Let $\mathcal{F}_n = \sigma((r_i(W_i, \zeta_n))_{i=1}^n, \xi_n, \zeta_n)$. Then for any $I \subseteq [n]$*

$$E\left[\prod_{i\in I} h_i(W_i, \zeta_n)|\mathcal{F}_n\right] = \prod_{i\in I} E\left[h_i(W_i, \zeta_n)|r_i(W_i, \zeta_n), \zeta_n\right]$$

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 $W_{1:n} \perp\!\!\!\perp \zeta_n$, one can show that $((W_i, \zeta_n))_{i=1}^n$ are jointly independent given ζ_n. The proof heavily relies on the following fact: $(A, B) \perp\!\!\!\perp C \implies A \perp\!\!\!\perp C|B$. We have

$$\begin{aligned}
&E\left[\prod_{i\in I} h_i(W_i, \zeta_n)|\mathcal{F}_n\right] = E\left[\prod_{i\in I} h_i(W_i, \zeta_n)\middle|(r_i(W_i, \zeta_n))_{i=1}^n, \zeta_n, \xi_n\right] \\
&= E\left[\prod_{i\in I} h_i(W_i, \zeta_n)\middle|(r_i(W_i, \zeta_n))_{i=1}^n, \zeta_n\right] = E\left[\prod_{i\in I} h_i(W_i, \zeta_n)\middle|(r_i(W_i, \zeta_n))_{i\in I}, \zeta_n\right] \\
&= E\left[\prod_{i\in I\setminus\{j\}} h_i(W_i, \zeta_n) E\left[h_j(W_j, \zeta_n)\middle|\prod_{i\in I\setminus\{j\}} h_i(W_i, \zeta_n), (r_i(W_i, \zeta_n))_{i\in I}, \zeta_n\right]\middle|(r_i(W_i, \zeta_n))_{i\in I}, \zeta_n\right]
\end{aligned}$$

The second equality is by $\xi_n \perp\!\!\!\perp (W_{1:n}, \zeta_n)$ 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 $C = (r_i(W_i, \zeta_n))_{i\in I^c}$, applied under the conditional measure given ζ_n. The final equality is by tower law. The inner expectation is

$$\begin{aligned}
&E\left[h_j(W_j, \zeta_n)\middle|\prod_{i\in I\setminus\{j\}} h_i(W_i, \zeta_n), (r_i(W_i, \zeta_n))_{i\in I}, \zeta_n\right] \\
&= E\left[h_j(W_j, \zeta_n)\middle|\prod_{i\in I\setminus\{j\}} h_i(W_i, \zeta_n), (r_i(W_i, \zeta_n))_{i\in I\setminus\{j\}}, r_j(W_j, \zeta_n), \zeta_n\right] \\
&= E\left[h_j(W_j, \zeta_n)|r_j(W_j, \zeta_n), \zeta_n\right]
\end{aligned}$$

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 ζ_n. Returning to

the display above, we have

$$E\left[\prod_{i\in I} h_i(W_i,\zeta_n)|\mathcal{F}_n\right] = E\left[h_j(W_j,\zeta_n)|r_j(W_j,\zeta_n),\zeta_n\right] E\left[\prod_{i\in I\setminus\{j\}} h_i(W_i,\zeta_n)\middle|(r_i(W_i,\zeta_n))_{i\in I},\zeta_n\right]$$
$$= E\left[h_j(W_j,\zeta_n)|r_j(W_j,\zeta_n),\zeta_n\right] E\left[\prod_{i\in I\setminus\{j\}} h_i(W_i,\zeta_n)\middle|(r_i(W_i,\zeta_n))_{i\in I\setminus\{j\}},\zeta_n\right]$$

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. □

Lemma 9.22 (Residual Selection)**.** *Let $\mathcal{F}_n = \sigma(\psi_{1:n}, p_{1:n,n}, \zeta_n, \xi_n)$, with $p_{i,n} = p_n(X_i,\zeta_n)$, $\psi_i = \psi(X_i,\zeta_n)$. Assume $(W_{1:n},\xi_n,\zeta_n)$ are jointly independent. Suppose that for $(F_{i,n})_{i=1}^n$ deterministic functions*

$$u_{i,n} = F_{i,n}(X_i,\zeta_n) - E[F_{i,n}(X_i,\zeta_n)|\psi_i,p_{i,n},\zeta_n] \qquad \sigma^2_{u,i,n} = E[u^2_{i,n}|\psi_i,p_{i,n},\zeta_n]$$

Let $i,j,l,f \in [n]$ distinct indices. The following moments hold for arbitrary mixtures of superscripts $\epsilon_i \in \{\epsilon_i^1,\epsilon_i^0\}$, which we suppress.

$$\begin{aligned}
&E[u_{i,n}|\mathcal{F}_n] = E[\epsilon_i|\mathcal{F}_n] = E[u_{i,n}\epsilon_i|\mathcal{F}_n] = E[u_{i,n}u_{j,n}|\mathcal{F}_n] = E[u_{i,n}\epsilon_j|\mathcal{F}_n] = E[\epsilon_i\epsilon_j|\mathcal{F}_n]\\
=&E[u_{i,n}u_{j,n}\epsilon_l|\mathcal{F}_n] = E[u_{i,n}\epsilon_j\epsilon_l|\mathcal{F}_n] = E[u_{i,n}u_{j,n}u_{l,n}|\mathcal{F}_n] = E[\epsilon_i\epsilon_j\epsilon_l|\mathcal{F}_n] = E[\epsilon_i\epsilon_j\epsilon_l\epsilon_f|\mathcal{F}_n]\\
=&E[\epsilon_i\epsilon_j\epsilon_l u_{f,n}|\mathcal{F}_n] = E[\epsilon_i\epsilon_j u_{l,n}u_{f,n}|\mathcal{F}_n] = E[\epsilon_i u_{j,n}u_{l,n}u_{f,n}|\mathcal{F}_n] = E[u_{i,n}u_{j,n}u_{l,n}u_{f,n}|\mathcal{F}_n]\\
=&E[u_{i,n}\epsilon_i u_{j,n}\epsilon_j|\mathcal{F}_n] = E[(u_{i,n}+\epsilon_i)(u_{j,n}+\epsilon_j)^2|\mathcal{F}_n] = 0
\end{aligned}$$

The higher order moments

$$\begin{aligned}
E[u^2_{i,n}|\mathcal{F}_n] &= \sigma^2_{u,i,n}\\
E[(u_{i,n}u_{j,n})^2|\mathcal{F}_n] &= \sigma^2_{u,i,n}\sigma^2_{u,j,n}\\
E[\epsilon_i^2|\mathcal{F}_n] &= E[\sigma^2(X_i)|\psi_i,p_{i,n},\zeta_n]\\
E[(u_{i,n}\epsilon_j)^2|\mathcal{F}_n] &= \sigma^2_{u,n}(X_i)E[\sigma^2(X_j)|\psi_j,p_{j,n},\zeta_n]\\
E[(\epsilon_i\epsilon_j)^2|\mathcal{F}_n] &= E[\sigma^2(X_i)|\psi_i,p_{i,n},\zeta_n]E[\sigma^2(X_j)|\psi_j,p_{j,n},\zeta_n]
\end{aligned}$$

Proof. We start with the first three inequalities. By applying Lemma 9.21 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 9.21 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 9.21 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

$$\begin{aligned} E[\epsilon_i^2|\mathcal{F}_n] &= E[\epsilon_i^2|\psi_i, p_{i,n}, \zeta_n] = E[E[\epsilon_i^2|X_i, \zeta_n]|\psi_i, p_{i,n}, \zeta_n] = E[E[\epsilon_i^2|X_i]|\psi_i, p_{i,n}, \zeta_n] \\ &= E[\sigma^2(X_i)|\psi_i, p_{i,n}, \zeta_n] \end{aligned}$$

The first line by applying Lemma 9.21, then tower law and independence. The other equalities in the display now again follow by applying the Lemma 9.21 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 $I = \{i, j\}$, $h_i(W_i, \zeta_n) = \epsilon_i^2$. □