Inference on Finite Population Treatment Effects Under Limited Overlap*

Han Hong† Michael P. Leung‡ Jessie Li§

August 20, 2019

This paper studies inference on finite population average and local average treatment effects under limited overlap, meaning some strata have a small proportion of treated or untreated units. We model limited overlap in an asymptotic framework sending the propensity score to zero (or one) with the sample size. We derive the asymptotic distribution of analog estimators of the treatment effects under two common randomization schemes: conditionally independent and stratified block randomization. Under either scheme, the limit distribution is the same and conventional standard error formulas remain asymptotically valid, but the rate of convergence is slower the faster the propensity score degenerates. The practical import of these results is twofold. When overlap is limited, standard methods can perform poorly in smaller samples, as asymptotic approximations are inadequate due to the slower rate of convergence. However, in larger samples, standard methods can work quite well even when the propensity score is small.

JEL Codes: C14, C21, C26

Keywords: treatment effects, overlap, instrumental variables, stratified randomization, propensity score.

1 Introduction

A well-known concern with estimating treatment effects under unconfoundedness is limited overlap, the possibility that in some strata, the proportion of treated or untreated units is small. This is a common problem with observational data, for example

*First draft: March 2018. We acknowledge funding from the National Science Foundation (SES 1658950), SIEPR, and a Faculty Research Grant awarded by the Committee on Research from the University of California, Santa Cruz.

†Department of Economics, Stanford University.

‡Department of Economics, University of Southern California. E-mail: leungm@usc.edu

§Department of Economics, UC Santa Cruz. E-mail: jeqli@ucsc.edu.

when only a few states pass a law of interest. It can also be a concern in experimental settings where, for instance, treatment is expensive to procure or other institutional constraints exist, so that only a small proportion of units in certain subpopulations can feasibly be treated. Several papers discuss the problem of limited overlap informally (e.g. Dehejia and Wahba, 1999; Heckman et al., 1997), but to our knowledge, none propose a rigorous definition and study the asymptotic properties of conventional estimators under limited overlap. This paper seeks to fill this gap in the literature.

We derive our main results in the context of the finite population model of treatment effects, where the set of observed units constitutes the entire population and potential outcomes and covariates are nonrandom quantities. This model dates back to Neyman (1923) and is widely used in the causal inference literature (e.g. Li and Ding, 2017; Freedman, 2008; Hinkelmann and Kempthorne, 2008; Imbens and Rubin, 2015; Rosenbaum, 2002). It stands in contrast to the superpopulation model, where the set of observed units is a small, random subsample from a larger population, and potential outcomes and covariates are modeled as i.i.d. draws from a superpopulation distribution. The relevance of the finite population model in experimental and observational settings, is discussed in, for example, Abadie et al. (2014), Imbens and Rubin (2015), and Reichardt and Gollob (1999).

We focus on finite population analogs of the average treatment effect (ATE) and local average treatment effect (LATE), the latter originally defined in Imbens and Angrist (1994) under a superpopulation model. We study the ATE in the context of the conditionally independent (CI) model where treatment assignment is independent across strata. We study the LATE in the instrumental variables (IV) model where instead a binary instrument satisfies this distributional assumption, and variation in the instrument induces take up of a binary treatment. In the IV model, only the instrument is random, and take up decisions are fixed. Such a model is relevant, for example, in experimental settings with noncompliance.

We consider two randomization schemes for treatment assignment in the CI model and the instrument in the IV model: conditionally independent and stratified block randomization. By conditionally independent randomization we mean that within each stratum, units are assigned to treatment in an i.i.d. fashion. By stratified block randomization, we mean that in stratum x, exactly m_x out of n_x units are assigned to treatment. Most of the econometric literature appears to focus on conditionally independent randomization, while most of the statistics literature seems to consider

stratified block randomization but only for the case of a single stratum. Allowing for only one stratum is a serious practical limitation, since the main motivation for stratification is to ensure balance, which requires independent randomization across different strata. Hence, an important feature of our results is that they allow for multiple strata. A technical contribution of this paper is a new CLT under stratified block randomization that may be of independent interest.

Note that despite our use of the word "randomization" above, our assumptions in the CI model correspond to the usual unconfoundedness condition, since the probability that unit i selects treatment is not a function of her identity, and hence, her potential outcomes. The model is therefore relevant for observational data when the econometrician is solely interested in inference on the observed set of units. This applies to settings in which no obvious superpopulation exists, for example when the set of observed units is the fifty states ((Abadie et al., 2014)). That said, we also derive analogous results for the superpopulation model.

We now define limited overlap. Let $p_n(x)$ denote the propensity score, where x denotes a stratum and n denotes the sample size. In the context of the CI model, $p_n(x)$ is the proportion of units assigned to treatment, whereas in the IV model, it is the proportion of units assigned to a particular value of a binary instrument. We say there is limited overlap if there exists some stratum x such that $p_n(x)$ degenerates to zero or one as the sample size diverges. This formalizes the notion of the proportion of treated or untreated units being small in an asymptotic framework. It stands in contrast to the conventional assumption that the propensity score is bounded away from zero or one (e.g. Firpo, 2007; Hirano et al., 2003). We emphasize that this model of limited overlap is not intended to be a realistic description of how real-world decisions to select into or assign treatment evolve as the population size grows but rather to provide an asymptotic approximation to a finite-sample phenomenon. This is the same idea behind high-dimensional asymptotics in statistical learning theory and weak instruments asymptotics.

Our theoretical results are as follows. (1) Under limited overlap, we find that the "effective" sample size is smaller in that estimators converge at the much slower rate $(n \min_x p_n(x)(1 - p_n(x)))^{-1/2}$. If this quantity tends to zero, which is necessary for consistent estimation and nests the standard sufficient overlap case, then standard estimators of the ATE and LATE are asymptotically normal after proper scaling, and the asymptotic variance is the same under both randomization schemes. Even

in the case of sufficient overlap, the results for the finite population LATE model appear to be new. (2) Conventional variance estimators remain valid and are therefore robust to limited overlap. (3) Interestingly, while under sufficient overlap, the variance estimators are well-known to be conservative in the finite population setup, under limited overlap, they are in fact asymptotically exact because a problematic covariance term that cannot be consistently estimated vanishes in the limit. (4) The proof of asymptotic normality relies on a new CLT for stratified block randomization that may be of independent interest. (5) While the focus of the main text is the finite population case, we also derive analogous results for the superpopulation model (see section A.4 of the appendix). Mirroring standard results, we show that the variance estimator is always consistent, regardless of limited overlap.

In a simulation study, we find that limited overlap can lead to undercoverage in smaller sample sizes, but coverage reaches nominal levels when n is sufficiently large. This is the case even if the proportion treated is empirically quite small. In our simulation results, when $p_n(x) = n^{-1/2}$, we obtain close to the target level of coverage when n is above 1000, despite the fact that $p_n(x)$ is then less than 4 percent. Our simulation results also indicate that in small samples, coverage can be substantially more conservative under stratified block randomization. Intuitively this is because conventional variance estimators are derived under asymptotic independence, but block randomization induces negative correlation in treatment assignment across units in finite samples. This conservativeness actually has the interesting advantage of partially correcting for undercoverage that results from limited overlap.

Thus, the practical import of our results is twofold.

1. Standard methods can perform poorly in smaller samples under limited overlap due to the usual intuition that the effective sample size is smaller. We provide formal justification for this intuition, showing that the rate of convergence is slowed by limited overlap, and thus the normal approximation is inadequate in small samples.

2. In larger samples, however, standard methods can work quite well even when the propensity score is fairly close to zero or one, as shown by our simulation results. We recommend reporting conventional estimates and standard errors together with other commonly used corrections that remove strata with limited overlap but have the effect of changing the target estimand (reviewed below).

Related Literature. Rothe (2017) studies inference on the ATE in the presence of limited overlap. He shows that the asymptotic coverage error of the conventional confidence interval can be large when overlap is limited. We complement this result by quantifying the rate of convergence. Rothe proposes an inference procedure valid in finite samples that relies on the assumption of normally distributed potential outcomes. In contrast, we impose no distributional assumptions, treat potential outcomes as fixed, and consider asymptotic inference. Sasaki and Ura (2018) and more recently Ma and Wang (2018) study trimming for inverse probability weighting with "small denominators," a related model for formalizing small propensity scores. All of these papers focus on superpopulation models with full compliance, conditionally indepenent randomization, and a propensity score that does not vary with n. In particular for the latter paper, when the denominator is not too small ($\gamma_0 > 2$), the estimator attains a $\sqrt{n}$ rate of convergence, whereas in our case the convergence rate can be slower due to the vanishing propensity score.

Several papers propose corrections for limited overlap in the CI model by removing observations in strata for which overlap is too limited, which mimics empirical practice (e.g. Crump et al., 2009; Dehejia and Wahba, 1999; Heckman et al., 1997; Ho et al., 2007). A disadvantage of these results is that they necessarily change the target estimand to a conditional treatment effect. Our results are complementary to this literature, since they imply that the rate of convergence can be substantially improved by removing strata with limited overlap. Our simulation evidence indicates that these estimators are particularly useful when sample sizes are small, since conventional estimators perform very poorly due to the slow rate of convergence. However, if the target estimand is truly an unconditional treatment effect, and if the sample size is large, then our results suggest that conventional estimators can still perform well.

Our results for stratified block randomization are new, even in the CI model without limited overlap. Li and Ding (2017) derive central limit theorems (CLTs) relevant for the finite population ATE when there exists only a single stratum. We derive a new CLT for the multiple-stratum case by generalizing the CLT for finite population simple random samples in Appendix 4 of Lehmann and D'Abrera (2006). Chapter 9.6 of Imbens and Rubin (2015) discusses results for stratified block randomization, but these are limited to linear regression estimators of the ATE in the superpopulation model. Bugni et al. (2018) and Bugni et al. (2017) study inference for the superpopulation ATE in two-sample and linear regression t-tests under a variety of sampling

schemes. Ansel et al. (2018) extend their results to the LATE and also study the efficiency of linear regression estimators. We focus on conditionally independent and stratified block randomization and study finite population estimands, which requires a new large sample theory that does not rely on randomness of potential outcomes. Also note that none of the papers above allow for limited overlap.

Outline. The next section studies the ATE in the conditional independence model. Section 3 considers the LATE in the instrumental variables model. We conduct a simulation study in section 4, and section 5 concludes. All proofs are given in the appendix.

2 Average Treatment Effect

This section studies inference on the ATE in the standard finite population potential outcomes model under conditional independence (CI) or unconfoundedness.

2.1 Setup

There are n observed units, and each unit i is endowed with treatment assignment D_i, a random variable supported on $\{0,1\}$. Let $Y_i(d)$ denote the potential outcome of unit i under treatment assignment $d \in \{0,1\}$ and $W_i \in \mathbb{R}^k$ a vector of baseline covariates. The elements $\{(Y_i(1), Y_i(0), W_i)\}_{i=1}^n$ are constants, so the only source of randomness is treatment assignment. The econometrician observes $\{(Y_i, D_i, W_i)\}_{i=1}^n$, where

$$Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i).$$

Stratified block randomization is implemented in practice using a small set of strata. Following Bugni et al. (2018) and Bugni et al. (2017), we assume strata are obtained from baseline covariates according to some mapping $S : \mathbb{R}^k \to \mathbb{X}$, where $\mathbb{X}$ is a finite set. We let $X_i = S(W_i)$ denote the stratum of observation i. Assume there exists a "mass function" $f(x)$ satisfying

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i = x\} \to f(x)$$

as $n \to \infty$ for all $x \in \mathbb{X}$.

We consider randomization schemes for which treatment assignment is independent across strata and identically distributed within strata. The latter implies unconfoundedness, since the probability that unit i is assigned treatment is not a function of her identity, and hence, her nonrandom potential outcomes. We focus on the following two schemes.

(a) Conditionally independent randomization: within stratum x, each unit is assigned to treatment with probability p_x, independently across units.

(b) Stratified block randomization: within stratum x, exactly m_x out of the n_x units are assigned to treatment.

Define the propensity score as $p_n(x) = \mathbf{E}[D_1 \,|\, X_1 = x]$. Thus for conditionally independent randomization, $p_n(x) = p_x$, and for stratified block randomization, $p_n(x) = m_x/n_x$. We assume throughout that $\frac{m_x/n_x}{p_x} \to 1$, so that the propensity score is asymptotically equivalent under the two randomization schemes. Define the sample propensity score

$$\hat{p}_n(x) = \frac{\sum_{i=1}^n D_i \mathbf{1}\{X_i = x\}}{\sum_{i=1}^n \mathbf{1}\{X_i = x\}}.$$

We next allow for the possibility that $p_n(x) \to c \in \{0, 1\}$ as $n \to \infty$ for some or all values of $x \in \mathbb{X}$ in order to capture limited overlap.

Definition 1. Let $a_n = \min_{x \in \mathbb{X}} p_n(x)(1 - p_n(x))$. We say there is *limited overlap* if $a_n \to 0$.

The finite population average treatment effect is given by

$$\tau_n = \frac{1}{n} \sum_{i=1}^n \big(Y_i(1) - Y_i(0)\big).$$

To motivate our estimator, we note that τ_n can alternatively be represented as

$$\tau_n = \sum_{x \in \mathbb{X}} \hat{f}(x) \tau_n(x),$$

where $\tau_n(x)$ is the finite population "conditional" ATE

$$\tau_n(x) = \mu_n(1,x) - \mu_n(0,x) = \frac{\sum_{i=1}^n Y_i(1)\mathbf{1}\{X_i = x\}}{\sum_{i=1}^n \mathbf{1}\{X_i = x\}} - \frac{\sum_{i=1}^n Y_i(0)\mathbf{1}\{X_i = x\}}{\sum_{i=1}^n \mathbf{1}\{X_i = x\}}.$$

Our estimator for the ATE is the sample analog

$$\hat{\tau}_n = \frac{1}{n}\sum_{i=1}^n \hat{\tau}_n(X_i) = \sum_{x \in \mathbb{X}} \hat{f}(x)\hat{\tau}_n(x),$$

where

$$\hat{\tau}_n(x) = \hat{\mu}_n(1,x) - \hat{\mu}_n(0,x) = \frac{\sum_{i=1}^n Y_i D_i \mathbf{1}\{X_i = x\}}{\sum_{i=1}^n D_i \mathbf{1}\{X_i = x\}} - \frac{\sum_{i=1}^n Y_i(1-D_i)\mathbf{1}\{X_i = x\}}{\sum_{i=1}^n (1-D_i)\mathbf{1}\{X_i = x\}}.$$

Note that because the set of strata $\mathbb{X}$ is finite, $\hat{\tau}_n$ is equivalent to the inverse propensity score weighted and doubly robust estimators (Robins et al., 1994); see Proposition 6 in the appendix.

Remark 1. This paper considers the case of binary-valued treatment effects. In the case of multi-valued treatment effects, we would redefine $a_n = \min_{x \in \mathbb{X}} \Pi_{q=1}^Q p_n(x,q)$ where $p_n(x,q) = \mathbf{E}[D_i(q) \,|\, X_1 = x]$ with $D_i(q) = 1$ if treatment is level $q \in \{0, 1, ..., Q\}$.[1] Bugni et al. (forthcoming) and Cattaneo (2010) study efficient inference under full overlap. When Q is finite, it should be straightforward to extend our univariate CLT results in the next subsection to corresponding multivariate CLTs for $\sqrt{na_n}\Sigma_n^{-1/2}(\hat{\tau}_\mathbf{n} - \tau_\mathbf{n})$, where $\tau_\mathbf{n} = (\tau_n(1), ..., \tau_n(Q))$ for $\tau_n(q) = n^{-1}\sum_{i=1}^n (Y_i(q) - Y_i(0))$, and $\hat{\tau}_\mathbf{n} = (\hat{\tau}_n(1), ..., \hat{\tau}_n(Q))$ for $\hat{\tau}_n(q) = n^{-1}\sum_{i=1}^n \hat{\tau}_n(X_i, q)$ with $\hat{\tau}_n(x,q) = \frac{\sum_{i=1}^n Y_i D_i(q)\mathbf{1}\{X_i = x\}}{\sum_{i=1}^n D_i(q)\mathbf{1}\{X_i = x\}} - \frac{\sum_{i=1}^n Y_i D_i(0)\mathbf{1}\{X_i = x\}}{\sum_{i=1}^n D_i(0)\mathbf{1}\{X_i = x\}}$. The case $Q \to \infty$ is an interesting direction for future research.

2.2 Asymptotic Theory

We next derive the asymptotic distribution of $\hat{\tau}_n$ and an estimator for its asymptotic variance. In the standard case of sufficient overlap, $\hat{\tau}_n$ is well-known to be $\sqrt{n}$-consistent. However, under limited overlap, we find that the rate of convergence will be slower and depend on a_n. Our CLT below shows that the correct rate is $\sqrt{na_n}$, which indicates that limited overlap can substantially reduce the effective sample size.

[1] We thank a referee for this comment.

Let

$$\sigma_n^2 = \text{Var}\left(\sqrt{na_n}\frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathbb{X}} D_i v_i(x)\right) = \text{Var}\left(\sqrt{na_n}\frac{1}{n}\sum_{i=1}^{n} D_i \tilde{v}_i(X_i)\right),$$

where

$$v_i(x) = \tilde{v}_i(x)\,\mathbf{1}\{X_i = x\}, \quad \tilde{v}_i(x) = \left(\frac{Y_i(1)-\mu_n(1,x)}{p_n(x)} + \frac{Y_i(0)-\mu_n(0,x)}{1-p_n(x)}\right).$$

This is the variance of an asymptotically linear representation of $\hat{\tau}_n$ and therefore will be equal to the asymptotic variance.

Theorem 1. *Suppose treatment is assigned according to conditionally independent or stratified block randomization. Assume $na_n \to \infty$,*

$$\limsup_{n\to\infty} \max_{d\in\{0,1\}} \frac{1}{n}\sum_{i=1}^{n} |Y_i(d)|^{2+\varepsilon} < \infty \quad \textit{for some } \varepsilon > 0, \tag{1}$$

and

$$\liminf_{n\to\infty} \sigma_n^2 > 0. \tag{2}$$

Then $\sqrt{na_n}(\hat{\tau}_n - \tau_n)/\sigma_n \xrightarrow{d} \mathcal{N}(0,1)$.

Remark 2. We can explicitly characterize σ_n^2 as follows. Lemma A.1 in the appendix shows that $\hat{\tau}_n$ can be linearized as

$$\hat{\tau}_n - \tau_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathbb{X}} D_i v_i(x) + o_p\left((na_n)^{-1/2}\right) = \frac{1}{n}\sum_{i=1}^{n} D_i\tilde{v}_i(X_i) + o_p\left((na_n)^{-1/2}\right),$$

The variance of the first term on the right-hand side gives us σ_n^2. By Lemma A.2 in the appendix, under either randomization scheme, this variance equals

$$\sigma_n^2 \equiv a_n\left[\frac{1}{n}\sum_{i=1}^{n} p_n(X_i)^{-1}\tilde{Y}_i(1,X_i)^2 + \frac{1}{n}\sum_{i=1}^{n}(1-p_n(X_i))^{-1}\tilde{Y}_i(0,X_i)^2 - \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{Y}_i(1,X_i)-\tilde{Y}_i(0,X_i)\right)^2\right], \tag{3}$$

where $\tilde{Y}_i(d,x) = Y_i(d) - \mu_n(d,x)$. This coincides with the usual expression for the

asymptotic variance, except from the presence of a_n to account for limited overlap (for the case of block randomization, see e.g. Imbens and Rubin, 2015, p. 202). Block randomization induces negative correlation between D_i and D_j, since i being treated reduces the chance that j is treated. However, in large samples, this reduction in j's treatment probability is negligible, hence the equivalence between variances.

Remark 3. Equation (1) is a standard moment condition, while (2) requires a non-degenerate variance. To assess the reasonableness of the latter condition, first consider (3) in the standard case where $p_n(x) \to \rho(x) \in (0,1)$ for all $x \in \mathbb{X}$. Then $a_n \to \alpha \in (0,1)$, which corresponds to the standard case without limited overlap and a $\sqrt{n}$ rate of convergence. Thus, (2) reduces to the usual nondegeneracy condition.

If instead $p_n(x) \to \rho(x)' \in \{0,1\}$ for some x, then $a_n \to 0$, and there must exist some $x' \in \mathbb{X}$ and $d' \in \{0,1\}$ for which $a_n/(p_n(x)^{d'}(1-p_n(x))^{1-d'}) \to 1$. For this stratum and treatment, overlap is the most limited. For all other strata and treatments, the terms $a_n/p_n(x)$, $a_n/(1-p_n(x))$, and a_n vanish in (3), intuitively because they converge to their limits faster than the strata with the least overlap. Then it is straightforward to see that (2) holds under the primitive condition

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \tilde{Y}_i(d',x')^2 \mathbf{1}\{X_i = x'\} > 0,$$

which requires nondegeneracy of the conditional second moments of potential outcomes.

Remark 4. The requirement $na_n \to \infty$ states that overlap cannot be too limited. Otherwise, there is insufficient information in the sample, and $\hat{\tau}_n$ is not even consistent. Consistency is also required in Sasaki and Ura (2018), who use a different notion of small propensity scores. When this condition fails, the limit distribution can be shown to be non-normal. One can potentially test this condition as follows.[2] Given a consistent variance estimator $\hat{\sigma}_n^2$ (discussed below) our theorem establishes normality of $T_n \equiv \sqrt{na_n}(\hat{\tau}_n - \tau_n)/\hat{\sigma}_n$ under $na_n \to \infty$. When this condition fails, the distribution can be shown to be non-normal. This suggests implementing any conventional test for normality (e.g. a KS test) by bootstrapping the distribution of T_n.

[2]We thank the editor for this suggestion.

Remark 5. The result for stratified block randomization relies on an argument generalizing the proof of Theorem 6, Appendix 4 of Lehmann and D'Abrera (2006), allowing for independent randomization across strata. This appears to be new, even in the case without limited overlap. See Lemma A.4 in the appendix.

In order to construct confidence intervals using Theorem 1, we need estimates of σ_n^2 and a_n. We can consistently estimate the latter using $\hat{a}_n = \min_i \hat{p}_n(X_i)(1-\hat{p}_n(X_i))$, since $\hat{p}_n(\cdot)$ is uniformly consistent over $\mathbb{X}$, as shown in the proof of Theorem 1. We consider the following estimator for the variance:

$$\hat{\sigma}_n^2 = \frac{\hat{a}_n}{n} \sum_{i=1}^{n} \left[\frac{D_i \left(Y_i - \hat{\mu}_n\left(1, X_i\right)\right)^2}{\hat{p}_n\left(X_i\right)^2} + \frac{\left(1 - D_i\right) \left(Y_i - \hat{\mu}_n\left(0, X_i\right)\right)^2}{\left(1 - \hat{p}_n\left(X_i\right)\right)^2} \right].$$

This is a standard expression, except for the presence of $\hat{a}_n$. To be more specific, note that under the superpopulation model $\hat{\sigma}_n^2$ is consistent for the asymptotic variance of $\hat{\tau}$ minus

$$\tau_n^x \equiv \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}[Y_i(1) - Y_i(0) \,|\, X_i] \tag{4}$$

(see section A.4 of the appendix). This is well-known in the case of sufficient overlap (e.g. Imbens and Wooldridge, 2009). However, in the finite population model, it is also well-known that there is an additional term in the asymptotic variance

$$\beta_n = a_n \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{Y}_i(1, X_i) - \tilde{Y}_i(0, X_i)\right)^2$$

that cannot be estimated due to its dependence on $\tilde{Y}_i(1, X_i)\tilde{Y}_i(0, X_i)$, which is unobserved. Since β_n is nonnegative, if it does not vanish in the limit, we can expect that $\hat{\sigma}_n^2$ is generally conservative in the finite population model, as shown formally in the next proposition.

Proposition 1. *Suppose treatment is assigned according to conditionally independent or stratified block randomization. Assume* $na_n \to \infty$, (2) *holds, and*

$$\limsup_{n\to\infty} \max_{d\in\{0,1\}} \frac{1}{n} \sum_{i=1}^{n} Y_i(d)^4 < \infty. \tag{5}$$

Then $(\hat{\sigma}_n^2 - \beta_n)/\sigma_n^2 \xrightarrow{p} 1$.

Remark 6. Using Proposition 1, the following is a (potentially conservative) $100 * (1-\alpha)\%$ confidence interval for τ_n:

$$\hat{\tau}_n \pm z_{1-\alpha/2} \frac{\hat{\sigma}_n}{\sqrt{n\hat{a}_n}}, \tag{6}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. Note that this coincides with the conventional confidence interval for (4), since $\hat{a}_n$ cancels out of $\hat{\sigma}_n$, resulting in the usual variance estimator. Thus, our results show that conventional i.i.d. standard errors are valid under limited overlap and stratified block sampling.

Remark 7. Proposition 1 implies that $\hat{\sigma}_n^2$ is asymptotically exact when there is limited overlap, since $\beta_n \to 0$ when $a_n \to 0$. The only general condition in the existing literature under which the estimator is exact is the strong requirement of homogeneity in potential outcomes ($Y_i(d) = Y_j(d)$ for all i, j such that $X_i = X_j$).

Since $\mathbb{X}$ is finite, it can be shown that $\hat{\tau}_n$ is numerically identical to the coefficient on D_i in a regression of Y_i on an intercept, D_i, the dummies $V_{ij} \equiv \mathbf{1}\{X_i = j\}$ for $j = 2, \ldots, J$, and all interaction terms between D_i and the centered dummies $V_j - \hat{f}(j)$. For example, in `Stata`,

$$\text{reg } Y_i \; D_i \; V_{i2} \; \ldots \; V_{iJ} \; D_i(V_{i2} - \hat{f}(2)) \; \ldots D(V_{iJ} - \hat{f}(J)). \tag{7}$$

Furthermore, the robust (Eicker-White) standard errors coincide with the standard errors in (6).

Proposition 2. $\hat{\sigma}_n^2/(\hat{a}_n n)$ *is numerically identical to the robust variance for the coefficient of* D_i *in regression* (7).

Hence, by Remark 7, in the finite-population model, a confidence interval for τ_n computed using robust standard errors is conservative under sufficient overlap but asymptotically exact under limited overlap. In the superpopulation model, if the target parameter is $\tau_n^x \equiv$(4), the confidence interval is known to be exact, and if the target parameter is $\mathbf{E}[Y_i(1) - Y_i(0)] \neq \tau_n^x$, it is known to undercover ((Imbens and

Wooldridge, 2009; Ansel et al., 2018)).

Remark 8. (Cattaneo et al., 2018) consider the case of "many" covariates in the superpopulation model and show that confidence intervals using the Eicker-White standard errors will undercover when the number of covariates grows with the sample size. The authors also examine confidence intervals using the HCk class of standard errors and show that HC3 will provide conservatively valid inference. In principle, our assumption that $\mathbb{X}$ is finite rules out the possibility of "many" covariates. However, if the cardinality of $\mathbb{X}$ is large in practice, it may be necessary to rework our asymptotics allowing the number of covariates to grow with the sample size. We leave this topic for further research.

Remark 9. The bootstrap is generally known to be valid for asymptotic inference under the superpopulation model for the parameter $\mathbf{E}[Y_i(1)-Y_i(0)]$. Thus, bootstrap confidence intervals are more conservative for (4) relative to (6). We therefore expect that the bootstrap is conservative for τ_n, although a formal result is beyond the scope of this paper. We also conjecture that the bootstrap is asymptotically exact under limited overlap when $a_n \to 0$, in light of the previous remark.

Remark 10. In completely randomized experiments, $p_n(x)$ does not vary across x. In this case, taking a simple difference of the two subsample means, which coincides with a linear regression Y_i on D_i, is also consistent for both the finite- and superpopulation ATEs. However, this estimator less efficient than regression (7). This is because the regression estimator is equivalent to $\hat{\tau}_n$, which in turn is equivalent to an inverse propensity score weighting estimator with an estimated propensity score (see section A.3 of the appendix). The latter estimator is known to reach the semiparametric efficiency bound (Hirano et al., 2003).

Relatedly, if $p_n(x)$ is known, the inverse propensity score weighting estimator using the known propensity score is consistent. However, it is more efficient to estimate it using regression (7) because the regression estimator is equivalent to $\hat{\tau}_n$.

3 Local Average Treatment Effect

In this section, we extend the results of the previous section to a finite population analog of the Imbens and Angrist (1994) instrumental variables (IV) model, where we consider inference on the LATE instead of the ATE.

3.1 Setup

As in the previous section, we treat $\{(Y_i(1), Y_i(0), W_i)\}_{i=1}^n$ as constants. Each unit i is now also endowed with an instrument Z_i, a random variable supported on $\{0, 1\}$. Let $D_i(z) \in \{0, 1\}$ represent unit i's take up choice when the instrument Z_i equals z. The elements $\{(D_i(1), D_i(0))\}_{i=1}^n$ are constants, so the only random element in the model is the instrument. The econometrician observes $\{(Y_i, D_i, Z_i, W_i)\}_{i=1}^n$, where

$$D_i = D_i(1)Z_i + D_i(0)(1 - Z_i).$$

As in section 2, we let $X_i = S(W_i)$ be the stratum of observation i, and we continue to assume that the set of strata $\mathbb{X}$ is finite. We again consider conditionally independent and stratified block randomization and allow for limited overlap, all defined analogously to the previous section, with D_i replaced by Z_i in these definitions.

We define the finite population LATE as

$$\lambda_n^* = \frac{\sum_{i=1}^n \big(Y_i(1) - Y_i(0)\big)\mathbf{1}\{D_i(1) > D_i(0)\}}{\sum_{i=1}^n \mathbf{1}\{D_i(1) > D_i(0)\}},$$

in complete analogy with the definition of the superpopulation estimand in Imbens and Angrist (1994). We next provide conditions under which λ_n^* is identified. Let

$$Y_i^*(z) = Y_i(1)D_i(z) + Y_i(0)(1 - D_i(z)),$$

so that $Y_i = Y_i^*(1)Z_i + Y_i^*(0)(1 - Z_i)$. For $z \in \{0, 1\}$ and $x \in \mathbb{X}$ let

$$\mu_n^*(z, x) = \frac{\sum_{i=1}^n Y_i^*(z)\mathbf{1}\{X_i = x\}}{\sum_{i=1}^n \mathbf{1}\{X_i = x\}}, \quad \gamma_n(z, x) = \frac{\sum_{i=1}^n D_i(z)\mathbf{1}\{X_i = x\}}{\sum_{i=1}^n \mathbf{1}\{X_i = x\}}.$$

Also let $\mu_n^*(z) = \sum_{x\in\mathbb{X}} \hat{f}(x)\mu_n^*(z,x)$ and $\gamma_n(z) = \sum_{x\in\mathbb{X}} \hat{f}(x)\gamma_n(z,x)$. Then define

$$\lambda_n = \frac{\mu_n^*(1) - \mu_n^*(0)}{\gamma_n(1) - \gamma_n(0)}.$$

This parameter can be estimated by its sample analog, defined below. As the next proposition shows, it is also asymptotically equivalent to the target parameter λ_n^* under standard identification conditions.

Proposition 3. *Suppose* (1) *holds, as well as the following conditions.*

(a) (Monotonicity) $\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{D_i(0) > D_i(1)\} \to 0$.

(b) (Compliers) $\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{D_i(1) > D_i(0)\} > 0$.

Then $|\lambda_n^* - \lambda_n| \to 0$.

The proof of this proposition is similar to the proof of Theorem 1 in Frölich (2007) and is therefore omitted. Assumption (b) requires the existence of compliers, so that λ_n^* is asymptotically well-defined. Assumption (a) rules out defiers in the limit. Observe that if there are absolutely no defiers ($\sum_{i=1}^n \mathbf{1}\{D_i(0) > D_i(1)\} = 0$) and some complier exists ($\sum_{i=1}^n \mathbf{1}\{D_i(1) > D_i(0)\} > 0$), then $\lambda_n^* = \lambda_n$.

We next define an estimator for λ_n. Let

$$\hat{\mu}_n^*(z,x) = \frac{\sum_{i=1}^n Y_i Z_i^z (1-Z_i)^{1-z}\mathbf{1}\{X_i = x\}}{\sum_{i=1}^n Z_i^z (1-Z_i)^{1-z}\mathbf{1}\{X_i = x\}},$$
$$\hat{\gamma}_n(z,x) = \frac{\sum_{i=1}^n D_i Z_i^z (1-Z_i)^{1-z}\mathbf{1}\{X_i = x\}}{\sum_{i=1}^n Z_i^z (1-Z_i)^{1-z}\mathbf{1}\{X_i = x\}}.$$

Also let $\hat{\mu}_n^*(z) = \sum_{x\in\mathbb{X}} \hat{f}(x)\hat{\mu}_n^*(z,x)$ and $\hat{\gamma}_n(z) = \sum_{x\in\mathbb{X}} \hat{f}(x)\hat{\gamma}_n(z,x)$. Our estimator for the LATE is

$$\hat{\lambda}_n = \frac{\hat{\mu}_n^*(1) - \hat{\mu}_n^*(0)}{\hat{\gamma}_n(1) - \hat{\gamma}_n(0)}.$$

3.2 Asymptotic Theory

Note that the IV model completely nests the CI model, the latter of which is obtained when $D_i(1) = 1 - D_i(0) = 1$ for all units i. We will obtain results analogous to those in section 2.

Define $p_n^*(x)$ as the proportion of treated units in stratum x (the analog of the propensity score for the instrument). Let $\Delta_n = n^{-1}\sum_{i=1}^n (D_i(1) - D_i(0))$ and

$$\tilde{v}_{\mu,i}(x) = \left(\frac{Y_i^*(1) - \mu_n^*(1,x)}{p_n^*(x)} + \frac{Y_i^*(0) - \mu_n^*(0,x)}{1 - p_n^*(x)}\right),$$
$$\tilde{v}_{\gamma,i}(x) = \left(\frac{D_i(1) - \gamma_n(1,x)}{p_n^*(x)} + \frac{D_i(0) - \gamma_n(0,x)}{1 - p_n^*(x)}\right).$$

The asymptotic variance will be given by

$$\sigma_{\lambda,n}^2 = \mathrm{Var}\left(\sqrt{na_n}\frac{1}{n\Delta_n}\sum_{i=1}^n Z_i\big(\tilde{v}_{\mu,i}(X_i) - \lambda_n^*\tilde{v}_{\gamma,i}(X_i)\big)\right).$$

Finally, define the sample analog of $p_n^*(x)$,

$$\hat{p}_n^*(x) = \frac{\sum_{i=1}^n Z_i \mathbf{1}\{X_i = x\}}{\sum_{i=1}^n \mathbf{1}\{X_i = x\}}.$$

Theorem 2. *Suppose the instrument is generated according to conditionally independent or stratified block randomization. Assume $na_n \to \infty$, assumptions (a) and (b) of Proposition 3 and* (1) *hold, and*

$$\liminf_{n\to\infty} \sigma_{\lambda,n}^2 > 0. \tag{8}$$

Then $\sqrt{na_n}(\hat{\lambda}_n - \lambda_n^*)/\sigma_{\lambda,n} \xrightarrow{d} \mathcal{N}(0,1)$.

Remark 11. Assumption (8) requires both an asymptotically nondegenerate variance, as in (2), as well as the usual rank condition that the instrument nontrivially affects the take up choice. The rank condition is equivalent to the compliers assumption in Proposition 3.

Lemma A.6 in the appendix derives an explicit form for $\sigma_{\lambda,n}^2$ in (23). Its sample

analog can be rewritten as

$$(\hat{\gamma}(1)-\hat{\gamma}(0))^{-2}\frac{\hat{a}_n}{n}\sum_{i=1}^{n}\left[\frac{Z_i\left(Y_i-\hat{\mu}_n^*(1,X_i)-\hat{\lambda}_n(D_i-\hat{\gamma}_n(1,X_i))\right)^2}{\hat{p}_n^*(X_i)^2}\right.$$
$$\left.+\frac{(1-Z_i)\left(Y_i-\hat{\mu}_n^*(0,X_i)-\hat{\lambda}_n(D_i-\hat{\gamma}_n(0,X_i))\right)^2}{(1-\hat{p}_n^*(X_i))^2}\right].$$

As in the CI model, under the superpopulation model, this is the sample analog of the asymptotic variance of $\hat{\lambda}_n$ minus

$$\lambda_n^x \equiv \frac{\sum_{i=1}^n \mathbf{E}[(Y_i(1)-Y_i(0))\mathbf{1}\{D_i(1)>D_i(0)\}\,|\,X_i]}{\sum_{i=1}^n \mathbf{P}(D_i(1)>D_i(0)\,|\,X_i)} \tag{9}$$

and is therefore consistent in that setting (see section A.4 of the appendix). On the other hand, if the target parameter is $\lambda^* = \mathbf{E}[Y_i(1)-Y_i(0)\,|\,D_i(1)>D_i(0)]$, then $\hat{\sigma}^2_{\lambda,n}$ is known to be an underestimate of the variance of $\hat{\lambda}_n-\lambda^*$ ((Ansel et al., 2018)).

In the finite population model, the estimator can be conservative because it does not account for the term

$$\beta_{\lambda,n} = a_n\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{Y}_i^*(1,X_i)-\tilde{Y}_i^*(0,X_i)-\lambda_n^*\left(\tilde{D}_i(1,X_i)-\tilde{D}_i(0,X_i)\right)\right)^2$$

in the asymptotic variance, which cannot be consistently estimated.

Proposition 4. *Suppose the instrument is generated according to conditionally independent or stratified block randomization. Assume $na_n \to \infty$ and assumptions (a) and (b) of Proposition 3 and (5) and (8) hold. Then $(\hat{\sigma}^2_{\lambda,n}-\beta_{\lambda,n})/\sigma^2_{\lambda,n} \xrightarrow{p} 1$.*

Remark 12. Similar to Proposition 1, we see that $\hat{\sigma}^2_{\lambda,n}$ can be conservative, but it is consistent for $\sigma^2_{\lambda,n}$ when potential outcomes and take up decisions are homogeneous or overlap is limited.

It can be shown that $\hat{\lambda}_n$ is numerically identical to the coefficient on D_i in an instrumental variable regression of Y_i on an intercept, D_i, the dummies $(V_{ij}, j = 2,\ldots,J)$ defined for (7), and the interaction terms $D_i(V_{ij}-\hat{f}(j))$, where D_i is instru-

mented by Z_i. The following proposition shows that the associated robust standard errors coincide with those constructed using $\hat{\sigma}^2_{\lambda,n}$. This is analogous to Proposition 2 for the CI model.

Proposition 5. *$\hat{\sigma}^2_{\lambda,n}/(\hat{a}_n n)$ is numerically identical to the robust variance for the coefficient of D_i in the aformentioned instrumental variable regression.*

In light of Remark 12, this implies that the `Stata` robust standard errors are asymptotically conservative under sufficient overlap and exact under limited overlap. Also, this regression is efficient for reasons analogous to those discussed in Remark 10 for the CI model.

4 Simulation Study

We study the finite-sample coverage of the confidence interval suggested by results in section 2. Suppose X_i is binary, and let $p_n(1) = 0.5$ and

$$p_n(0) = \min\{n^{-\delta}, 0.5\}.$$

We will display results for several values of $\delta \geqslant 0$ under both randomization schemes. In the case of stratified block randomization, we will define $m_x = \lceil p_n(x) n_x \rceil$.

We simulate a random-coefficients model, where outcomes $\mathbf{Y} = (Y_1, \ldots, Y_n)$ satisfy

$$\mathbf{Y} = \mathbf{W}\theta,$$

$\mathbf{W} = (\mathbf{1}, \mathbf{X}, \mathbf{D}, \mathbf{D} * \mathbf{X})$, $\mathbf{1}$ is an n-dimensional vector of ones, $\mathbf{X} = (X_1, \ldots, X_n)$, $\mathbf{D} = (D_1, \ldots, D_n)$, and "$*$" denotes the elementwise product. We draw $\{\theta_i\}_{i=1}^n \sim \mathcal{N}(\mu, \Sigma)$ for $\mu = (0, 0.5, 2, 1)$ and Σ the identity matrix, and $\{X_i\}_{i=1}^n \overset{iid}{\sim} \text{Ber}(0.5)$.

Table 1 displays results for conditionally independent randomization and Table 2 for stratified block randomization. Row "Cover" denotes the fraction of simulations for which the nominal 95-percent confidence interval (6) covers τ_n. We use 5000 simulations, redrawing only D in each simulation, since this is a finite population model. Row "P-score" gives the average value of $\hat{p}_n(0)$ across the simulations.

From the simulation results, we find that when δ is small, so that there is sufficient overlap, the confidence intervals are conservative, which corroborates Proposition 1.

Table 1: Conditionally Independent Randomization

δ	0			0.4			0.5		
n	100	1000	5000	100	1000	5000	100	1000	5000
Cover	0.962	0.970	0.968	0.932	0.956	0.949	0.896	0.931	0.942
P-score	0.50	0.50	0.50	0.16	0.06	0.03	0.10	0.03	0.01

Table 2: Stratified Block Randomization

δ	0			0.4			0.5		
n	100	1000	5000	100	1000	5000	100	1000	5000
Cover	0.956	0.961	0.969	0.949	0.954	0.951	0.930	0.936	0.948
P-score	0.50	0.50	0.50	0.17	0.06	0.03	0.11	0.03	0.01

Also coverage is similar across randomization schemes for large samples, which is consistent with Remark 2. When δ is larger, there is undercoverage for smaller sample sizes due to the small number of treated observations in stratum $x = 0$. In unreported results, we find that for larger values of δ, the sample size needs to be quite substantial for adequate coverage. This is unsurprising given that limited overlap reduces the rate of convergence by Theorem 1.

It is noteworthy to compare Tables 1 and 2 for the case $\delta = 0.5$. We see that coverage is more conservative for stratified block randomization relative to conditionally independent randomization when the sample size is small. This is because in finite samples, block randomization induces negative correlation in treatment assignment across units. When δ is large and n small, there are few treated units, so this negative correlation remains substantial. Since the standard errors are derived under asymptotic independence, they become conservative for block randomization in finite samples, relative to independent randomization. This conservativeness actually appears to be advantageous given that the confidence intervals tend to undercover substantially under conditionally independent randomization and limited overlap. Of course, when the sample size is sufficiently large, coverage approaches nominal levels.

5 Conclusion

We propose an asymptotic definition of limited overlap, that the propensity score tends to zero or one with the sample size for some strata. This provides a rigorous notion of overlap being "limited," analogous to how high-dimensional asymptotics take the number of covariates large as the sample size diverges. We study the properties of standard estimators for the ATE and LATE in a finite population model under this notion of limited overlap. We find that the estimators are asymptotically normal under both conditionally independent and stratified block randomization. In the case of block randomization, our results allow for independent randomization across multiple strata, while existing results only allow for a single stratum. We also find that limited overlap slows the rate of convergence, yet standard variance estimators remain valid.

In a simulation study, we show that in finite samples, limited overlap can lead to undercoverage, which, in addition to our rate result, provides motivation for alternative estimators that eliminate observations in strata with limited overlap, (e.g Crump et al., 2009; Ho et al., 2007). However, these estimators come at the cost of estimating a conditional average treatment effect instead. Our results suggest that researchers interested in the unconditional ATE can still use conventional estimators when sample sizes are large, since in this case, the estimators can perform well even when the propensity score is empirically quite small.

A Appendix

A.1 Proofs: ATE

This section provides proofs of results in section 2. The next lemma states high-level conditions for asymptotic normality that we will later verify for the cases of conditionally independent and stratified block randomization.

Lemma A.1. *For $na_n \to \infty$, under the following conditions,*

$$\sqrt{na_n}(\hat{\tau}_n - \tau_n)/\sigma_n \xrightarrow{d} \mathcal{N}(0,1).$$

(a) (Nondegenerate variance) $\sigma_n^{-1} = O(1)$.

(b) (P-score estimator) $\hat{p}_n(x)/p_n(x) \xrightarrow{p} 1$ *for all* $x \in \mathbb{X}$.

(c) (Rate of convergence) For all $d \in \{0,1\}$ *and* $x \in \mathbb{X}$,

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i(d) - \mu_n(d,x))\left(\frac{D_i}{p_n(x)}\right)^d\left(\frac{1-D_i}{1-p_n(x)}\right)^{1-d}\mathbf{1}\{X_i = x\} = O_p\left((na_n)^{-1/2}\right). \tag{10}$$

(d) (Normality of linear representation)

$$\sqrt{na_n}\frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathbb{X}}(D_i - p_n(x))v_i(x)/\sigma_n \xrightarrow{d} \mathcal{N}(0,1).$$

PROOF. Fix $x \in \mathbb{X}$. Let $\hat{\kappa}_n(1,x) = (\frac{1}{n}\sum_{i=1}^{n} D_i\mathbf{1}\{X_i = x\}/(p_n(x)f(x)))^{-1}$. Then

$$\hat{f}(x)(\hat{\mu}_n(1,x) - \mu_n(1,x)) = \hat{\kappa}_n(1,x)\frac{\hat{f}(x)}{f(x)}\frac{1}{n}\sum_{i=1}^{n}\tilde{Y}_i(1,x)\frac{D_i}{p_n(x)}\mathbf{1}\{X_i = x\}, \tag{11}$$

where $\tilde{Y}_i(d,x) = Y_i(d) - \mu_n(d,x)$. By assumption (b), $\hat{\kappa}_n(1,x) \xrightarrow{p} 1$, and by assumption (c), the average on the right-hand side of (11) is $O_p((na_n)^{-1/2})$. This and a similar derivation for $\hat{f}(x)(\hat{\mu}_n(0,x) - \mu_n(0,x))$ yield the following asymptotically linear representation of the conditional ATE (scaled by $\hat{f}(x)$):

$$\begin{aligned}\hat{f}(x)\left(\hat{\tau}_n(x) - \tau_n(x)\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{Y}_i(1,x) - \tilde{Y}_i(0,x)\right)\mathbf{1}\{X_i = x\} \\ &+ \frac{1}{n}\sum_{i=1}^{n}(D_i - p_n(x))\left(\frac{\tilde{Y}_i(1,x)}{p_n(x)} + \frac{\tilde{Y}_i(0,x)}{1-p_n(x)}\right)\mathbf{1}\{X_i = x\} + o_p((na_n)^{-1/2}). \end{aligned} \tag{12}$$

The first term on the right-hand side is identically zero. The result then follows from assumptions (a) and (d). ∎

We next compute the variance of the influence function, which gives us the asymptotic variance by the previous lemma.

Lemma A.2. *Under conditionally independent or stratified block randomization,*

$$\begin{aligned} &Var\left(\sqrt{na_n}\frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathbb{X}}(D_i - p_n(x))v_i(x)\right) \\ &\quad= \sum_{x\in\mathbb{X}}\left[\frac{a_n}{p_n(x)}\frac{1}{n}\sum_{i=1}^{n}\tilde{Y}_i(1,x)^2\mathbf{1}\{X_i = x\} + \frac{a_n}{1-p_n(x)}\frac{1}{n}\sum_{i=1}^{n}\tilde{Y}_i(0,x)^2\mathbf{1}\{X_i = x\}\right. \\ &\qquad\qquad\left. -a_n\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{Y}_i(1,x) - \tilde{Y}_i(0,x)\right)^2\mathbf{1}\{X_i = x\}\right]. \quad (13) \end{aligned}$$

Note that this is equivalent to (3)*.*

PROOF. Under either randomization scheme,

$$\sigma_n^2 = \frac{a_n}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{x\in\mathbb{X}}\mathrm{Cov}(D_i v_i(x), D_j v_j(x)).$$

Under conditionally independent randomization, $\mathrm{Cov}(D_i v_i(x), D_j v_j(x)) = 0$ if $i \neq j$ and otherwise equals $\mathrm{Var}(D_i v_i(x))$, which is given by

$$\begin{aligned} &\left(\frac{1-p_n(x)}{p_n(x)}\tilde{Y}_i(1,x)^2 + 2\tilde{Y}_i(1,x)\tilde{Y}_i(0,x) + \frac{p_n(x)}{1-p_n(x)}\tilde{Y}_i(0,x)^2\right)\mathbf{1}\{X_i = x\} \\ &\quad= \left(\frac{1}{p_n(x)}\tilde{Y}_i(1,x)^2 + \frac{1}{1-p_n(x)}\tilde{Y}_i(0,x)^2 - \left(\tilde{Y}_i(1,x) - \tilde{Y}_i(0,x)\right)^2\right)\mathbf{1}\{X_i = x\}, \end{aligned}$$

where $\tilde{Y}_i(d,x) = Y_i(d) - \mu_n(d,x)$. This establishes (13).

Under stratified block randomization, we also have to show that the following is $o(1)$:

$$\frac{a_n}{n}\sum_{i\neq j}\sum_{x\in\mathbb{X}}\mathrm{Cov}(D_i v_i(x), D_j v_j(x)) = \frac{a_n}{n}\sum_{i\neq j}\sum_{x\in\mathbb{X}} w_{ij}(x)\frac{\frac{m_x(m_x-1)}{n_x(n_x-1)} - p_n(x)^2}{p_n(x)^2(1-p_n(x))^2}, \quad (14)$$

where $w_{ij}(x) = p_n(x)^2(1-p_n(x))^2 v_i(x)v_j(x)$. Let $r_n(x)^2 = m_x(m_x-1)/(n_x(n_x-1))$. Then

$$(14) \leqslant \left(\frac{1}{n^2}\sum_{i\neq j}\sum_{x\in\mathbb{X}} w_{ij}(x)\right)\max_{x\in\mathbb{X}}\frac{a_n}{p_n(x)(1-p_n(x))}\frac{n(r_n(x)^2 - p_n(x)^2)}{p_n(x)(1-p_n(x))}.$$

Note that $a_n/(p_n(x)(1-p_n(x))) \leqslant 1$. Also, by (1),

$$\frac{1}{n^2}\sum_{i\neq j}\sum_{x\in\mathbb{X}} w_{ij}(x) = \sum_{x\in\mathbb{X}} \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} p_n(x)(1-p_n(x))v_i(x)\right)^2}_{0} - \underbrace{\frac{1}{n^2}\sum_{i=1}^{n} p_n(x)^2(1-p_n(x))^2 v_i(x)^2}_{o(1)}.$$

Furthermore,

$$\frac{n(r_n(x)^2 - p_n(x)^2)}{p_n(x)(1-p_n(x))} = \frac{n}{n_x}\frac{\frac{m_x(m_x-1)}{n_x-1} - \frac{m_x^2}{n_x}}{p_n(x)(1-p_n(x))} = \frac{1}{\hat{f}(x)}\frac{\frac{m_x}{n_x-1}(\frac{m_x}{n_x}-1)}{p_n(x)(1-p_n(x))} \to -\frac{1}{f(x)}.$$

Thus, (14) $\to 0$, as desired. ∎

The next two lemmas verify the conditions of Lemma A.1 for our two randomization schemes.

Lemma A.3. *Suppose treatment is assigned according to conditionally independent randomization. Assume $na_n \to \infty$ and (1) and (2) hold. Then $\sqrt{na_n}(\hat{\tau}_n - \tau_n)/\sigma_n \xrightarrow{d} \mathcal{N}(0,1)$.*

Proof. It suffices to verify assumptions (b), (c), and (d) of Lemma A.1. To verify (b), note that $\hat{p}_n(x)/p_n(x)$ has mean one and variance $(1-p_n(x))^2/(na_n\hat{f}(x))$, which tends to zero since $na_n \to \infty$.

To verify assumption (c), consider the case $d = 1$, with case $d = 0$ being similar. Then

$$na_n \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{Y}_i(1,x)\frac{D_i}{p_n(x)}\mathbf{1}\{X_i = x\}\right) = \frac{1}{n}\sum_{i=1}^{n}\tilde{Y}_i(1,x)^2\mathbf{1}\{X_i = x\}\frac{a_n(1-p_n(x))}{p_n(x)},$$

which is $O(1)$ by (1). Hence, the variance of the left-hand side of (10) is $O((na_n)^{-1})$. Since the left-hand side is also mean zero, this verifies (c).

Lastly, we verify assumption (d) of Lemma A.1. We check the Lindeberg condition:

for all $\delta > 0$,

$$\sum_{i=1}^{n} \mathbf{E}\left[W_i^2 \mathbf{1}\left\{|W_i| > \delta\right\}\right] \to 0,$$

where $W_i = (a_n/(n\sigma_n^2))^{1/2} \sum_{x\in\mathbb{X}} (D_i - p_n(x))v_i(x)$. Note that for any $\varepsilon > 0$,

$$\sum_{i=1}^{n} \mathbf{E}\left[W_i^2 \mathbf{1}\left\{|W_i| > \delta\right\}\right] \leqslant \sum_{i=1}^{n} \mathbf{E}\left[|W_i|^{2+\varepsilon} \frac{1}{W_i^{\varepsilon}} \mathbf{1}\left\{|W_i| > \delta\right\}\right] \leqslant \sum_{i=1}^{n} \mathbf{E}\left[|W_i|^{2+\varepsilon}\right] \delta^{-\varepsilon}. \tag{15}$$

Hence, it suffices to show that $\sum_{i=1}^{n} \mathbf{E}\left[|W_i|^{2+\varepsilon}\right]$ is $o(1)$ for some $\varepsilon > 0$. Now, this expectation equals

$$\begin{aligned}
&\left(\frac{a_n}{n\sigma_n^2}\right)^{\frac{2+\varepsilon}{2}} \sum_{i=1}^{n} \sum_{x\in\mathbb{X}} \left[p_n(x)(1-p_n(x))^{2+\varepsilon} + (1-p_n(x))p_n(x)^{2+\varepsilon}\right] |v_i(x)|^{2+\varepsilon} \\
&\qquad = (na_n)^{-\varepsilon/2} \sigma_n^{-1-\varepsilon/2} \sum_{x\in\mathbb{X}} \left[\left((1-p_n(x))\frac{a_n}{p_n(x)(1-p_n(x))}\right)^{1+\varepsilon}\right. \\
&\qquad\quad \left. + \left(p_n(x)\frac{a_n}{p_n(x)(1-p_n(x))}\right)^{1+\varepsilon}\right] \frac{1}{n} \sum_{i=1}^{n} |p_n(x)(1-p_n(x))v_i(x)|^{2+\varepsilon}.
\end{aligned} \tag{16}$$

Note that $na_n \to \infty$, $\sigma_n^{-1-\varepsilon/2} = O(1)$ by (2), and $\max_{x\in\mathbb{X}} a_n/(p_n(x)(1-p_n(x))) = 1$. Furthermore,

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} \sum_{x\in\mathbb{X}} |p_n(x)(1-p_n(x))v_i(x)|^{2+\varepsilon} \\
&\qquad = \frac{1}{n} \sum_{i=1}^{n} \sum_{x\in\mathbb{X}} 1\{X_i = x\} \left|\left((1-p_n(x))\tilde{Y}_i(1,x) + p_n(x)\tilde{Y}_i(0,x)\right)\right|^{2+\varepsilon},
\end{aligned}$$

which is $O(1)$ by assumption (1). Hence, (16) $\to 0$, as desired. ∎

Lemma A.4. *Suppose treatment is assigned according to stratified block randomization. Assume $na_n \to \infty$ and (1) and (2) hold. Then $\sqrt{na_n}(\hat{\tau}_n - \tau_n)/\sigma_n \xrightarrow{d} \mathcal{N}(0,1)$.*

PROOF. It suffices to verify assumptions (b), (c), and (d) of Lemma A.1. To verify

(b), note that $\hat{p}_n(x)/p_n(x)$ has mean one and variance

$$\left[\sum_{i=1}^n \mathbf{1}\{X_i = x\} p_n(x)(1-p_n(x)) + \sum_{i\neq j} \mathbf{1}\{X_i = X_j = x\}\left(\frac{m_x(m_x-1)}{n_x(n_x-1)} - p_n(x)^2\right)\right] \times p_n(x)^{-2}\left(\sum_{i=1}^n \mathbf{1}\{X_i = x\}\right)^{-2} = \frac{(1-p_n(x))^2}{np_n(x)(1-p_n(x))\hat{f}(x)} + p_n(x)^{-1}\left(\frac{m_x-1}{n_x-1} - 1\right),$$

which is $o(1)$, since $na_n \to \infty$.

Next, we verify assumption (c) of Lemma A.1. Consider case $d = 1$, with case $d = 0$ being similar. It suffices to show that the following is $O(1)$:

$$na_n \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n \tilde{Y}_i(1,x)\frac{D_i}{p_n(x)}\mathbf{1}\{X_i = x\}\right) = \frac{1}{n}\sum_{i=1}^n \tilde{Y}_i(1,x)^2 \mathbf{1}\{X_i = x\}\frac{a_n(1-p_n(x))}{p_n(x)} + \frac{a_n}{n}\sum_{i\neq j}\tilde{Y}_i(1,x)\tilde{Y}_j(1,x)\mathbf{1}\{X_i = X_j = x\}\frac{\frac{m_x(m_x-1)}{n_x(n_x-1)} - p_n(x)^2}{p_n(x)^2}.$$

The first term is $O(1)$ by (1). The second term equals

$$\frac{1}{n^2}\sum_{i\neq j}\tilde{Y}_i(1,x)\tilde{Y}_j(1,x)\mathbf{1}\{X_i = X_j = x\}\frac{a_n(1-p_n(x))}{p_n(x)}\frac{n}{n_x}\frac{\frac{m_x(m_x-1)}{n_x-1} - \frac{m_x^2}{n_x}}{p_n(x)(1-p_n(x))}. \tag{17}$$

To see that this is $o(1)$, first note that

$$\frac{1}{n^2}\sum_{i\neq j}\tilde{Y}_i(1,x)\tilde{Y}_j(1,x)\mathbf{1}\{X_i = X_j = x\} = \left(\frac{1}{n}\sum_{i=1}^n \tilde{Y}_i(1,x)\mathbf{1}\{X_i = x\}\right)^2 - \frac{1}{n^2}\sum_{i=1}^n \tilde{Y}_i(1,x)^2\mathbf{1}\{X_i = x\}.$$

The first term equals zero, while the second is $o(1)$ by (1). Second,

$$na_n(1-p_n(x))/(n_x p_n(x)) \leqslant \hat{f}(x)^{-1} \to f(x)^{-1}.$$

Third,

$$\frac{\frac{m_x(m_x-1)}{n_x-1} - \frac{m_x^2}{n_x}}{p_n(x)(1-p_n(x))} = \frac{\frac{m_x}{n_x-1}(p_n(x)-1)}{p_n(x)(1-p_n(x))} \to -1.$$

Thus, (17) $\to 0$, as desired.

Lastly, we verify assumption (d) of Lemma A.1. Let $\{D_i^*\}_{i=1}^n$ be i.i.d. conditionally independent Bernoulli random variables with success probability m_x/n_x. (Recall $p_n(x) = m_x/n_x$ under stratified block randomization.) Define $W_i^*(x) = v_i(x)D_i^*$.

To apply Corollary 2, Appendix 3 of Lehmann and D'Abrera (2006), we first verify (A.121) of the corollary and then apply the Lindeberg CLT to $\sum_{i=1}^n \sum_{x\in\mathbb{X}} W_i^*(x)$ (after centering and scaling). Equation (A.121) corresponds to

$$\frac{\mathbf{E}\left[\left(\sum_{i=1}^n \sum_{x\in\mathbb{X}}(W_i^*(x) - W_i(x))\right)^2\right]}{\mathrm{Var}\left(\sum_{i=1}^n \sum_{x\in\mathbb{X}} W_i^*(x)\right)} \to 0, \tag{18}$$

where $W_i(x) = v_i(x)D_i$. By the Cauchy-Schwarz inequality, the left-hand side is bounded above by

$$\sum_{x,y\in\mathbb{X}} \left(\frac{\mathbf{E}\left[\left(\sum_{i=1}^n (W_i^*(x) - W_i(x))\right)^2\right]}{\mathrm{Var}\left(\sum_{i=1}^n W_i^*(x)\right)} \frac{\mathbf{E}\left[\left(\sum_{i=1}^n (W_i^*(y) - W_i(y))\right)^2\right]}{\mathrm{Var}\left(\sum_{i=1}^n W_i^*(y)\right)} \right.$$
$$\left. \frac{\mathrm{Var}\left(\sum_{i=1}^n W_i^*(x)\right)}{\mathrm{Var}\left(\sum_{i=1}^n \sum_{x\in\mathbb{X}} W_i^*(x)\right)} \frac{\mathrm{Var}\left(\sum_{i=1}^n W_i^*(y)\right)}{\mathrm{Var}\left(\sum_{i=1}^n \sum_{x\in\mathbb{X}} W_i^*(x)\right)} \right)^{1/2}.$$

By Lemma 1, Appendix 4 of Lehmann and D'Abrera (2006),

$$\max_{x\in\mathbb{X}} \frac{\mathbf{E}\left[\left(\sum_{i=1}^n (W_i^*(x) - W_i(x))\right)^2\right]}{\mathrm{Var}\left(\sum_{i=1}^n W_i^*(x)\right)} \to 0.$$

Furthermore, since treatment assignment is independent across strata,

$$\frac{\mathrm{Var}\left(\sum_{i=1}^n W_i^*(x)\right)}{\mathrm{Var}\left(\sum_{i=1}^n \sum_{x\in\mathbb{X}} W_i^*(x)\right)} \leqslant 1$$

for all $x \in \mathbb{X}$. This establishes (18).

It remains to prove a CLT for $\sum_{i=1}^n \sum_{x\in\mathbb{X}} v_i(x)D_i^*$, i.e. to show that

$$\sqrt{na_n}\frac{1}{n}\sum_{i=1}^n\sum_{x\in\mathbb{X}}(D_i^* - p_n(x))v_i(x)/\sigma_n \xrightarrow{d} \mathcal{N}(0,1).$$

Since D_i^* is now conditionally i.i.d., it is enough to check the Lindeberg condition. We can apply the corresponding argument from the proof of Lemma A.3 verbatim. ■

Proof of Theorem 1. This follows from Lemmas A.3 and A.4. ■

Proof of Proposition 1. First observe that

$$\hat\sigma_n^2 = \sum_{x\in\mathbb{X}}\hat f(x)\left[\frac{\hat a_n}{\hat p_n(x)}\frac{\sum_{i=1}^n\big(Y_i-\hat\mu_n(1,x)\big)^2 D_i\mathbf{1}\{X_i=x\}}{\sum_{i=1}^n D_i\mathbf{1}\{X_i=x\}} + \frac{\hat a_n}{1-\hat p_n(x)}\frac{\sum_{i=1}^n\big(Y_i-\hat\mu_n(0,x)\big)^2(1-D_i)\mathbf{1}\{X_i=x\}}{\sum_{i=1}^n(1-D_i)\mathbf{1}\{X_i=x\}}\right]$$

(see the proof of Proposition 2). From the proof of Theorem A.3, $\hat p_n(x)/p_n(x) \xrightarrow{p} 1$ and $\hat a_n/a_n \xrightarrow{p} 1$. By (3), it suffices to show that for any $d\in\{0,1\}$,

$$\left|\hat f(x)\frac{\sum_{i=1}^n\big(Y_i-\hat\mu_n(d,x)\big)^2 D_i^d(1-D_i)^{1-d}\mathbf{1}\{X_i=x\}}{\sum_{i=1}^n D_i^d(1-D_i)^{1-d}\mathbf{1}\{X_i=x\}} - \frac{1}{n}\sum_{i=1}^n\tilde Y_i(d,x)^2\mathbf{1}\{X_i=x\}\right| \xrightarrow{p} 0, \tag{19}$$

where $\tilde Y_i(d,x) = Y_i(d) - \mu_n(d,x)$. We prove the case $d=1$, as $d=0$ is similar. Since $\hat f(x)/f(x)\to 1$, (19) holds if

$$\left|\frac{1}{n}\sum_{i=1}^n\tilde Y_i(d,x)^2\frac{D_i}{p_n(x)}\mathbf{1}\{X_i=x\} - \frac{1}{n}\sum_{i=1}^n\tilde Y_i(d,x)^2\mathbf{1}\{X_i=x\}\right| \xrightarrow{p} 0. \tag{20}$$

First consider conditionally independent randomization. Here,

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n\tilde Y_i(d,x)^2\frac{D_i}{p_n(x)}\mathbf{1}\{X_i=x\}\right) = \frac{1}{n^2}\sum_{i=1}^n\tilde Y_i(d,x)^4(1-p_n(x))\mathbf{1}\{X_i=x\},$$

which is $o(1)$ by (5). This establishes (19).

Next consider stratified block randomization. By the argument for conditionally

independent randomization, it suffices to show that the following covariance is $o(1)$:

$$\frac{1}{n^2}\sum_{i\neq j}\tilde{Y}_i(d,x)^2\tilde{Y}_j(d,x)^2\mathbf{1}\{X_i=X_j=x\}\frac{\frac{m_x(m_x-1)}{n_x(n_x-1)}-\frac{m_x^2}{n_x^2}}{p_n(x)^2}.$$

This tends to zero because $\frac{1}{n^2}\sum_{i\neq j}\tilde{Y}_i(d,x)^2\tilde{Y}_j(d,x)^2\mathbf{1}\{X_i=X_j=x\}=O(1)$ by (5), and

$$\frac{\frac{m_x(m_x-1)}{n_x(n_x-1)}-\frac{m_x^2}{n_x^2}}{p_n(x)^2}\to 0.$$

■

PROOF OF PROPOSITION 2. The normal equations of this regression take the form of $\frac{1}{n}\sum_{i=1}^n W_i\epsilon_i=0$, where the *instruments* W_i consist of 1, D_i, $V_{ij}-\bar{V}_j, j\geqslant 2$, $D_i\times\left(V_{ij}-\bar{V}_j\right), j\geqslant 2$, and where $\epsilon_i=y_i-\hat{\alpha}-D_i\hat{\tau}_n-\hat{\eta}'\left(V_i-\bar{V}\right)-\hat{\phi}'D_i\times\left(V_i-\bar{V}\right)$. By linearly transforming the normal equations, we can replace the second instrument D_i by

$$\sum_{j=1}^J 1\left(x_i=j\right)\frac{D_i-\hat{p}_n\left(j\right)}{\hat{p}_n\left(j\right)\left(1-\hat{p}_n\left(j\right)\right)}=\frac{D_i-\hat{p}_n\left(X_i\right)}{\hat{p}_n\left(X_i\right)\left(1-\hat{p}_n\left(X_i\right)\right)}$$

and denote the transformed instruments as $\tilde{W}_i$. Then the second row of the Jacobian matrix of the normal equations is

$$\frac{1}{n}\sum_{i=1}^n\tilde{W}_i\left[1\quad D_i\quad V_i-\bar{V}\quad D_i\times\left(V_i-\bar{V}\right)\right]=\left[0\quad 1\quad 0\quad 0\right]$$

By inspecting the other rows we find that the second row of the inverse of the Jacobian matrix is also $[0\quad 1\quad 0\quad 0]$. Hence the robust variance of $\hat{\tau}_n$ from Stata is given by

$$\text{se}=\sum_{i=1}^n\left(\frac{D_i-\hat{p}_n\left(X_i\right)}{\hat{p}_n\left(X_i\right)\left(1-\hat{p}_n\left(X_i\right)\right)}\right)^2\epsilon_i^2.$$

Next by reparameterization, we can rewrite, for $\zeta_{j1}=\hat{\mu}_n\left(1,j\right)$, $\zeta_{j0}=\hat{\mu}_n\left(0,j\right)$, and $V_{ij}=1\left(X_i=j\right)$,

$$\begin{aligned}\epsilon_i=&\quad y_i-\textstyle\sum_{j=1}^J\zeta_{j1}V_{ij}D_i-\sum_{j=1}^J\zeta_{j0}V_{ij}\left(1-D_i\right)\\=&\quad\textstyle\sum_{j=1}^J\left(y_i-\zeta_{j1}\right)V_{ij}D_i-\sum_{j=1}^J\left(y_i-\zeta_{j0}\right)V_{ij}\left(1-D_i\right)\end{aligned}$$

Then we can write

$$\begin{aligned}\text{se} = & \sum_{i=1}^{n}\left(\tfrac{D_i-\hat{p}_n(X_i)}{\hat{p}_n(X_i)(1-\hat{p}_n(X_i))}\right)^2\left[\sum_j V_{ij}D_i\left(y_i-\zeta_{j1}\right)^2+\sum_j V_{ij}\left(1-D_i\right)\left(y_i-\zeta_{j0}\right)^2\right]\\ = & \sum_{j=1}^{J}\sum_{i=1}^{n}\left(\tfrac{D_i-\hat{p}_n(j)}{\hat{p}_n(j)(1-\hat{p}_n(j))}\right)^2 V_{ij}D_i\left(y_i-\zeta_{j1}\right)^2\\ & +\sum_{j=1}^{J}\sum_{i=1}^{n}\left(\tfrac{D_i-\hat{p}_n(j)}{\hat{p}_n(j)(1-\hat{p}_n(X_i))}\right)^2 V_{ij}\left(1-D_i\right)\left(y_i-\zeta_{j0}\right)^2\end{aligned}$$

Next note that

$$\left(D_i-\hat{p}_n\left(j\right)\right)^2 D_i=D_i\left(1-\hat{p}_n\left(j\right)\right)^2,\quad \left(D_i-\hat{p}_n\left(j\right)\right)^2\left(1-D_i\right)=\left(1-D_i\right)\hat{p}_n\left(j\right)^2$$

We can further rewrite

$$\text{se}=\sum_{j=1}^{J}\frac{1}{\hat{p}_n\left(j\right)^2}\sum_{i=1}^{n}V_{ij}D_i\left(y_i-\zeta_{j1}\right)^2+\sum_{j=1}^{J}\frac{1}{\left(1-\hat{p}_n\left(j\right)\right)^2}\sum_{i=1}^{n}V_{ij}\left(1-D_i\right)\left(y_i-\zeta_{j0}\right)^2.$$

We can then manipulate $\hat{\sigma}_n^2/\left(n\hat{a}_n\right)$ to be exactly this same form, so that se $=$ $\hat{\sigma}_n^2/\left(n\hat{a}_n\right)$. ■

A.2 Proofs: LATE

The next four lemmas are respectively analogous to Lemmas A.1-A.4 in the main text.

Lemma A.5. *For $na_n\to\infty$, under the following conditions,*

$$\sqrt{na_n}(\hat{\lambda}_n-\lambda_n^*)/\sigma_{\lambda,n}\xrightarrow{d}\mathcal{N}(0,1).$$

(a) Assumptions (a) and (b) of Proposition 3 hold.

(b) (1) *and* (8) *hold.*

(c) $\hat{p}_n^*(x)/p_n^*(x)\xrightarrow{p}1$ for all $x\in\mathbb{X}$.

(d) For all $z \in \{0,1\}$ and $x \in \mathbb{X}$,

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i(z)-\mu_n^*(z,x))\left(\frac{Z_i}{p_n^*(x)}\right)^z\left(\frac{1-Z_i}{1-p_n^*(x)}\right)^{1-z}\mathbf{1}\{X_i=x\}=O_p\left((na_n)^{-1/2}\right),$$
$$\frac{1}{n}\sum_{i=1}^{n}(D_i(z)-\gamma_n(z,x))\left(\frac{Z_i}{p_n^*(x)}\right)^z\left(\frac{1-Z_i}{1-p_n^*(x)}\right)^{1-z}\mathbf{1}\{X_i=x\}=O_p\left((na_n)^{-1/2}\right).$$

(e) $\sqrt{na_n}\frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathbb{X}}\left((Z_i-p_n^*(x))(v_{\mu,i}(x)-\lambda_n^* v_{\gamma,i}(x))\right)/\sigma_{\lambda,n}\xrightarrow{d}\mathcal{N}(0,1)$, *where*

$$v_{\mu,i}(x)=\left(\frac{Y_i^*(1)-\mu_n^*(1,x)}{p_n^*(x)}+\frac{Y_i^*(0)-\mu_n^*(0,x)}{1-p_n^*(x)}\right)\mathbf{1}\{X_i=x\},$$
$$v_{\gamma,i}(x)=\left(\frac{D_i(1)-\gamma_n(1,x)}{p_n^*(x)}+\frac{D_i(0)-\gamma_n(0,x)}{1-p_n^*(x)}\right)\mathbf{1}\{X_i=x\}$$

PROOF. First note that

$$\hat{\lambda}_n-\lambda_n^*=\frac{\hat{\mu}_n^*(1)-\hat{\mu}_n^*(0)-\mu_n^*(1)+\mu_n^*(0)}{\hat{\gamma}_n(1)-\hat{\gamma}_n(0)}-\lambda_n^*\frac{\hat{\gamma}_n(1)-\hat{\gamma}_n(0)-\gamma_n(1)+\gamma_n(0)}{\hat{\gamma}_n(1)-\hat{\gamma}_n(0)}.$$

Second, by assumptions (b) and (c) and the arguments in the proof of Lemma A.1,

$$\hat{\mu}_n^*(z)-\mu_n^*(z)=\frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathbb{X}}(Z_i-p_n^*(x))v_{\mu,i}(x)+o_p\left((na_n)^{-1/2}\right),$$
$$\hat{\gamma}_n(z)-\gamma_n(z)=\frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathbb{X}}(Z_i-p_n^*(x))v_{\gamma,i}(x)+o_p\left((na_n)^{-1/2}\right).$$

Therefore,

$$\sqrt{na_n}\frac{\hat{\lambda}_n-\lambda_n^*}{\sigma_{\lambda,n}}=\frac{1}{\hat{\gamma}_n(1)-\hat{\gamma}_n(0)}\frac{1}{\sigma_{\lambda,n}}\sqrt{\frac{a_n}{n}}\left[\sum_{i=1}^{n}\sum_{x\in\mathbb{X}}(Z_i-p_n^*(x))(v_{\mu,i}(x)-\lambda_n^*v_{\gamma,i}(x))\right.$$
$$\left.+A_n-B_n\right]+o_p\left(\sigma_{\lambda,n}^{-1}(\hat{\gamma}_n(1)-\hat{\gamma}_n(0))^{-1}(1-\lambda_n^*)(na_n)^{-1/2}\right), \quad (21)$$

where for $\tilde{Y}_i^*(z,x) = Y_i^*(z) - \mu_n^*(z,x)$ and $\tilde{D}_i(z,x) = D_i(z) - \gamma_n(z,x)$,

$$A_n = \sum_{i=1}^n \sum_{x\in\mathbb{X}} \left(\tilde{Y}_i^*(1,x) - \tilde{Y}_i^*(0,x)\right) \mathbf{1}\{X_i = x\},$$
$$B_n = \lambda_n^* \sum_{i=1}^n \sum_{x\in\mathbb{X}} \left(\tilde{D}_i(1,x) - \tilde{D}_i(0,x)\right) \mathbf{1}\{X_i = x\}.$$

Note that $A_n = B_n = 0$.

We have $\lambda_n^* = O(1)$ by assumption (b) of Proposition 3 and (1). Furthermore,

$$\begin{aligned}
&\hat{\gamma}_n(1) - \hat{\gamma}_n(0) \\
&\quad = \sum_{x\in\mathbb{X}} \hat{f}(x) \left(\frac{\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i = x\} Z_i D_i(1)}{\hat{f}(x)\hat{p}_n(x)} - \frac{\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i = x\}(1-Z_i) D_i(0)}{\hat{f}(x)(1-\hat{p}_n(x))} \right) \\
&\qquad = \sum_{x\in\mathbb{X}} \hat{f}(x) \left(\frac{1}{n}\sum_{i=1}^n D_i(1) - \frac{1}{n}\sum_{i=1}^n D_i(0) \right) + o_p(1) = \Delta_n + o_p(1)
\end{aligned}$$

Δ_n's limit infimum is strictly positive by assumptions (a) and (b) of Proposition 3. Therefore, by (8),

$$\sigma_{\lambda,n}^{-1} (\hat{\gamma}_n(1) - \hat{\gamma}_n(0))^{-1} (1 - \lambda_n^*)(n a_n)^{-1/2} = O_p((n a_n)^{-1/2}).$$

The result then follows from assumption (d). ∎

Define $\tilde{Y}_i^*(z,x) = Y_i^*(z) - \mu_n^*(z,x)$, $\tilde{D}_i(z,x) = D_i(z) - \gamma_n(z,x)$, $\gamma_n(z) = \sum_{x\in\mathbb{X}} \hat{f}(x)\gamma_n(z,x)$, and

$$\Delta v_i(x) = v_{\mu,i}(x) - \lambda_n^* v_{\gamma,i}(x), \tag{22}$$

Lemma A.6. *Under conditionally independent or stratified block randomization,*

$$\begin{aligned}
\sigma_{\lambda,n}^2 = \Delta_n^{-2} \frac{1}{n} \sum_{i=1}^n \sum_{x\in\mathbb{X}} \Bigg[&\frac{a_n}{p_n^*(x)} \left(\tilde{Y}_i^*(1,x) - \lambda_n^* \tilde{D}_i(1,x)\right)^2 \\
&+ \frac{a_n}{1-p_n^*(x)} \left(\tilde{Y}_i^*(0,x) - \lambda_n^* \tilde{D}_i(0,x)\right)^2 \\
&- a_n \left(\tilde{Y}_i^*(1,x) - \tilde{Y}_i^*(0,x) - \lambda_n^* \left(\tilde{D}_i(1,x) - \tilde{D}_i(0,x)\right)\right)^2 \Bigg] \mathbf{1}\{X_i = x\}.
\end{aligned} \tag{23}$$

PROOF. Under either randomization scheme,

$$\sigma^2_{\lambda,n} = \Delta_n^{-2}\frac{a_n}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{x\in\mathbb{X}}\mathrm{Cov}(Z_i\Delta v_i(x), Z_j\Delta v_j(x)).$$

Under conditionally independent randomization, $\mathrm{Cov}(Z_i\Delta v_i(x), Z_j\Delta v_j(x)) = 0$, so some simple algebra establishes that

$$\Delta_n^{-2}\frac{a_n}{n}\sum_{i=1}^{n}\sum_{x\in\mathbb{X}}\mathrm{Var}(Z_i\Delta v_i(x)) = (23).$$

Under stratified block randomization, we also have to show that

$$\frac{a_n}{n}\sum_{i\neq j}\sum_{x\in\mathbb{X}}\mathrm{Cov}(Z_i\Delta v_i(x), Z_j\Delta v_j(x)) \to 0.$$

This can be proven using the argument in the proof of Lemma A.2 that establishes $(14) = o(1)$. ∎

Lemma A.7. *Suppose the instrument is generated according to conditionally independent randomization. Assume* $na_n \to \infty$ *and* (1) *and* (8) *hold. Then* $\sqrt{na_n}(\hat{\lambda}_n - \lambda_n^*)/\sigma_{\lambda,n} \xrightarrow{d} \mathcal{N}(0,1)$.

PROOF. It suffices to verify assumptions (b), (c), and (d) of Lemma A.5. Assumption (b) is shown in Lemma A.1. Verification of assumption (c) proceeds along the same lines as the argument in Lemma A.3 for verifying assumption (c) of Lemma A.1.

Lastly, we verify assumption (d) of Lemma A.5. By (15), it suffices to show that

$$\sum_{i=1}^{n}\mathbf{E}\left[|W_i|^{2+\varepsilon}\right] \to 0. \tag{24}$$

Then, as in (16), the left-hand side of (24) equals

$$(na_n)^{-\varepsilon/2}\sigma_{\lambda,n}^{-1-\varepsilon/2}\sum_{x\in\mathbb{X}}\left[\left((1-p_n^*(x))\frac{a_n}{p_n^*(x)(1-p_n^*(x))}\right)^{1+\varepsilon}\right.$$
$$\left.+\left(p_n^*(x)\frac{a_n}{p_n^*(x)(1-p_n^*(x))}\right)^{1+\varepsilon}\right]\frac{1}{n}\sum_{i=1}^{n}|p_n^*(x)(1-p_n^*(x))\Delta v_i(x)|^{2+\varepsilon}, \quad (25)$$

where $\Delta v_i(x)$ is defined in (22). Note that $na_n \to \infty$, $\sigma_{\lambda,n}^{-1-\varepsilon/2} = O(1)$ by (8), and

$$\max_{x\in\mathbb{X}}\frac{a_n}{p_n^*(x)(1-p_n^*(x))} = 1.$$

Furthermore,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathbb{X}}|p_n^*(x)(1-p_n^*(x))\Delta v_i(x)|^{2+\varepsilon} \leqslant |2\lambda_n^*|^{2+\varepsilon}$$
$$+\frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathbb{X}}1\{X_i=x\}\left|\left(Y_i^*(1)-\mu_n^*(1,x)+Y_i^*(0)-\mu_n^*(0,x)\right)\right|^{2+\varepsilon},$$

which is $O(1)$ by assumption (1) and (2). Hence, (25) $\to 0$, as desired. ∎

Lemma A.8. *Suppose the instrument is generated according to stratified block randomization. Assume $na_n \to \infty$ and (1) and (8) hold. Then $\sqrt{na_n}(\hat{\lambda}_n - \lambda_n^*)/\sigma_{\lambda,n} \xrightarrow{d} \mathcal{N}(0,1)$.*

PROOF. Let $\{D_i^*\}_{i=1}^n$ be i.i.d. Bernoulli random variables with success probability m_x/n_x. (Recall $p_n^*(x) = m_x/n_x$ under stratified block randomization.) Define $W_i^*(x) = \Delta v_i(x)D_i^*$, where $\Delta v_i(x)$ is defined in (22).

As in the proof of Lemma A.4, we use Corollary 2, Appendix 3 of Lehmann and D'Abrera (2006) to reduce the problem to showing $\sum_{i=1}^n\sum_{x\in\mathbb{X}}W_i^*(x)$ is asymptotically normal after centering and scaling. To apply the corollary, we need to verify (18). The argument for this proceeds as in the proof of Lemma A.4.

It remains to prove a CLT for $\sum_{i=1}^n \sum_{x\in\mathbb{X}} \Delta v_i(x) D_i^*$, i.e. to show that

$$\sqrt{na_n}\frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathbb{X}}(D_i^* - p_n^*(x))\Delta v_i(x)/\sigma_n \xrightarrow{d} \mathcal{N}(0,1).$$

It suffices to verify assumptions (b), (c), and (d) of Lemma A.5. The argument for (d) is identical to the proof of Lemma A.7, while (b) is shown in the proof of Lemma A.4. Lastly, the argument for verifying assumption (c) is the same as the argument in Lemma A.4 for verifying assumption (c) of Lemma A.1. ∎

PROOF OF THEOREM 2. This follows from Lemmas A.7 and A.8. ∎

PROOF OF PROPOSITION 4. First observe that $\hat{\sigma}^2_{\lambda,n}$ equals

$$(\hat{\gamma}(1)-\hat{\gamma}(0))^{-2}\sum_{x\in\mathbb{X}}\hat{f}(x)\left[\frac{\hat{a}_n}{\hat{p}_n^*(x)}\frac{\sum_{i=1}^n\left(Y_i-\hat{\mu}_n^*(1,x)-\hat{\lambda}_n(D_i-\hat{\gamma}_n(1,x))\right)^2 Z_i\mathbf{1}\{X_i=x\}}{\sum_{i=1}^n Z_i\mathbf{1}\{X_i=x\}}\right.$$
$$\left.+\frac{\hat{a}_n}{1-\hat{p}_n^*(x)}\frac{\sum_{i=1}^n\left(Y_i-\hat{\mu}_n^*(0,x)-\hat{\lambda}_n(D_i-\hat{\gamma}_n(0,x))\right)^2(1-Z_i)\mathbf{1}\{X_i=x\}}{\sum_{i=1}^n(1-Z_i)\mathbf{1}\{X_i=x\}}\right]$$

(see the proof of Proposition 5). By Theorem 2, $|\hat{\lambda}_n - \lambda_n^*| \xrightarrow{p} 0$. Also, the proof of Lemma A.5 shows $(\hat{\gamma}_n(1)-\hat{\gamma}_n(0))/(\gamma_n(1)-\gamma_n(0)) \xrightarrow{p} 1$. The remainder of the proof is similar to the proof of Proposition 1. ∎

PROOF OF PROPOSITION 5. The estimating equations of this IV regression take the form of $\frac{1}{n}\sum_{i=1}^n W_i\epsilon_i = 0$, where the *instruments* W_i consists of 1, Z_i, $V_{ij}-\bar{V}_j, j\geqslant 2$, $Z_i\times\left(V_{ij}-\bar{V}_j\right), j\geqslant 2$, and where $\epsilon_i = y_i - \hat{\alpha} - D_i\hat{\lambda}_n - \hat{\eta}'\left(V_i-\bar{V}\right) - \hat{\phi}' Z_i\times\left(V_i-\bar{V}\right)$. By linearly transforming the normal equations, we can replace the second instrument Z_i by

$$\sum_{j=1}^{J} 1\left(x_i=j\right)\frac{Z_i-\hat{p}_n^*\left(j\right)}{\hat{p}_n^*\left(j\right)\left(1-\hat{p}_n^*\left(j\right)\right)} = \frac{Z_i-\hat{p}_n^*\left(X_i\right)}{\hat{p}_n^*\left(X_i\right)\left(1-\hat{p}_n^*\left(X_i\right)\right)}$$

and denote the transformed instruments as $\tilde{W}_i$. Then the second row of the Jacobian

matrix of the estimating equations is

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{W}_i\left[1 \;\; D_i \;\; V_i-\bar{V} \;\; Z_i\times\left(V_i-\bar{V}\right)\right]=\left[0 \;\; t \;\; 0 \;\; 0\right]$$

where $t=\hat{\gamma}_n(1)-\hat{\gamma}_n(0)$. By inspecting the other rows we find that the second row of the inverse of the Jacobian matrix is also $[0 \;\; t^{-1} \;\; 0 \;\; 0]$. To derive t, note that

$$\begin{aligned}
\hat{\gamma}_n(1)-\hat{\gamma}_n(0) &= \sum_j \hat{f}(j)\left[\frac{\sum_i D_i Z_i V_{ij}}{\sum_i Z_i V_{ij}}-\frac{\sum_i D_i(1-Z_i)V_{ij}}{\sum_i (1-Z_i)V_{ij}}\right] \\
&= \sum_j\left[\frac{\frac{1}{n}\sum_i D_i Z_i V_{ij}}{\hat{p}_n^*(j)}-\frac{\frac{1}{n}\sum_i D_i(1-Z_i)V_{ij}}{1-\hat{p}_n^*(j)}\right] \\
&= \frac{1}{n}\sum_i\left[D_i Z_i\frac{\sum_j V_{ij}}{\hat{p}_n^*(j)}-D_i(1-Z_i)\frac{\sum_j V_{ij}}{1-\hat{p}_n^*(j)}\right] \\
&= \frac{1}{n}\sum_i\left[\frac{D_i Z_i}{\hat{p}_n^*(X_i)}-\frac{D_i(1-Z_i)}{1-\hat{p}_n^*(X_i)}\right] \\
&= \frac{1}{n}\sum_i\frac{D_i\left(Z_i-\hat{p}_n^*(X_i)\right)}{\hat{p}_n^*(X_i)\left(1-\hat{p}_n^*(X_i)\right)}
\end{aligned}$$

Then we note that by equivalent linear transformation of the moment conditions, the instruments can also be transformed to

$$\bar{W}_i=\left(Z_i V_{ij}, j=1,\ldots,J,\;\; (1-Z_i)V_{ij}, j=1,\ldots,J\right)$$

Next we consider reparameterizing the residual term. Introduce $\phi_1\equiv 0$ and write

$$\begin{aligned}
\epsilon_i = &\;\; y_i-\hat{\alpha}-D_i\hat{\lambda}_n-\textstyle\sum_{j\geqslant 2}\hat{\eta}_j'\left(V_{ij}-\bar{V}_j\right)-\sum_{j\geqslant 2}\hat{\phi}_j' Z_i\times\left(V_{ij}-\bar{V}_j\right) \\
= &\;\; y_i-\hat{\alpha}-D_i\hat{\lambda}_n-\textstyle\sum_{j\geqslant 2}\hat{\eta}_j'\left(V_{ij}-\bar{V}_j\right)-\sum_{j\geqslant 1}\hat{\phi}_j' Z_i\times\left(V_{ij}-\bar{V}_j\right).
\end{aligned}$$

If we fix $\hat{\lambda}_n$ as an arbitrary number and take $\hat{\alpha},\eta_j,\phi_j, j\geqslant 1$ as parameters that are exactly identified by the moment equations, we can then reparameterize as

$$\epsilon_i=y_i-D_i\hat{\lambda}_n-\sum_j\zeta_{j1}V_{ij}Z_i-\sum_j\zeta_{j0}V_{ij}(1-Z_i).$$

where it can easily be shown that

$$\zeta_{j1} = \hat{\mu}_n^* (1, j) - \hat{\lambda}_n \hat{\gamma}_n (1, j), \quad \zeta_{j0} = \hat{\mu}_n^* (0, j) - \hat{\lambda}_n \hat{\gamma}_n (0, j).$$

Then the error term can be written as

$$\epsilon_i = \sum_j \left(y_i - \hat{\lambda}_n D_i - \zeta_{j1} \right) V_{ij} Z_i + \sum_j \left(y_i - \hat{\lambda}_n D_i - \zeta_{j0} \right) V_{ij} (1 - Z_i).$$

The nominal Stata robust variance for $\hat{\lambda}_n$ is then given by

$$\begin{aligned} &(\hat{\gamma}_n (1) - \hat{\gamma}_n (0))^{-2} \sum_i \left(\frac{Z_i - \hat{p}_n^* (X_i)}{\hat{p}_n^* (X_i) (1 - \hat{p}_n^* (X_i))} \right)^2 \epsilon_i^2 \\ &\qquad = \sum_{j=1}^J \frac{1}{\hat{p}_n^* (j)^2} \sum_{i=1}^n V_{ij} Z_i \left(y_i - \hat{\lambda}_n D_i - \zeta_{j1} \right)^2 \\ &\qquad\qquad + \sum_{j=1}^J \frac{1}{(1 - \hat{p}_n^* (j))^2} \sum_{i=1}^n V_{ij} (1 - Z_i) \left(y_i - \hat{\lambda}_n D_i - \zeta_{j0} \right)^2. \end{aligned}$$

where the last equality follows similarly to the end of the proof for Proposition 2. Finally, $\hat{\sigma}_{\lambda,n}^2$ can be rewritten to take exactly the same form. ■

A.3 Estimator Equivalence

Consider the IV model of section 3. Note that the CI model of section 2 is a special case, obtained when $D_i(1) = 1$ and $D_i(0) = 0$ for all i. We define three common estimators and show their computational equivalence. The projection estimator is given by

$$\hat{\lambda}_n = \frac{\frac{1}{n} \sum_{i=1}^n \left(\hat{\mu}_n^* (1, X_i) - \hat{\mu}_n^* (0, X_i) \right)}{\frac{1}{n} \sum_{i=1}^n \left(\hat{\gamma}_n (1, X_i) - \hat{\gamma}_n (0, X_i) \right)}.$$

Note that this is equivalent to the original definition of $\hat{\lambda}_n$ in section 3 by (26) below. The inverse probability weighting estimator is defined as

$$\hat{\lambda}_{IP} = \left(\frac{1}{n} \sum_{i=1}^n \left(\frac{D_i Z_i}{\hat{p}_n (X_i)} - \frac{D_i (1 - Z_i)}{1 - \hat{p}_n (X_i)} \right) \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \left(\frac{Y_i Z_i}{\hat{p}_n (X_i)} - \frac{Y_i (1 - Z_i)}{1 - \hat{p}_n (X_i)} \right) \right).$$

The doubly robust estimator is given by

$$\hat{\lambda}_{DR} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i(Y_i-\hat{\mu}_n^*(1,X_i))}{\hat{p}_n(X_i)} + \hat{\mu}_n^*(1,X_i) - \frac{(1-Z_i)(Y_i-\hat{\mu}_n^*(0,X_i))}{1-\hat{p}_n(X_i)} - \hat{\mu}_n^*(0,X_i)\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i(D_i-\hat{\gamma}_n(1,X_i))}{\hat{p}_n(X_i)} + \hat{\gamma}_n(1,X_i) - \frac{(1-Z_i)(D_i-\hat{\gamma}_n(0,X_i))}{1-\hat{p}_n(X_i)} - \hat{\gamma}_n(0,X_i)\right)}.$$

Equivalence of the three estimators is known in the ATE setting. The next proposition establishes the result in the more general LATE setting.

Proposition 6. *If $\{x : f(x) > 0\}$ has finite cardinality, then $\hat{\lambda}_n = \hat{\lambda}_{IP} = \hat{\lambda}_{DR}$.*

Proof. First note that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_n^*(z,X_i) &= \sum_{x\in\mathbb{X}}\hat{f}(x)\frac{\sum_{i=1}^{n}Y_iZ_i^z(1-Z_i)^{1-z}\mathbf{1}\{X_i=x\}}{\sum_{i=1}^{n}Z_i^z(1-Z_i)^{1-z}\mathbf{1}\{X_i=x\}}, \qquad (26)\\
\frac{1}{n}\sum_{i=1}^{n}\hat{\gamma}_n(z,X_i) &= \frac{1}{n}\sum_{i=1}^{n}\frac{D_iZ_i^z(1-Z_i)^{1-z}}{\hat{p}_n^*(X_i)}.
\end{aligned}$$

Then $\hat{\lambda}_{IP} = \hat{\lambda}_n$ follows from the fact that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\frac{Y_iZ_i^z(1-Z_i)^{1-z}}{\hat{p}_n(X_i)} &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i=x\}\frac{Y_iZ_i^z(1-Z_i)^{1-z}}{\hat{p}_n^*(x)}\\
&= \sum_{x\in\mathbb{X}}\hat{f}(x)\frac{\sum_{i=1}^{n}Y_iZ_i^z(1-Z_i)^{1-z}\mathbf{1}\{X_i=x\}}{\sum_{i=1}^{n}Z_i\mathbf{1}\{X_i=x\}}.
\end{aligned}$$

To complete the proof, we show that $\hat{\lambda}_{DR} = \hat{\lambda}_n$. First observe that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i\left(Y_i - \hat{\mu}_n^*\left(1, X_i\right)\right)}{\hat{p}_n\left(X_i\right)} + \hat{\mu}_n^*\left(1, X_i\right)\right) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{\hat{p}_n\left(X_i\right)} + \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_n^*\left(1, X_i\right) - \frac{1}{n}\sum_{i=1}^{n} Z_i \frac{\hat{\mu}_n^*\left(1, X_i\right)}{\hat{p}_n\left(X_i\right)} \\
&\quad = 2\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_n^*\left(1, X_i\right) - \frac{1}{n}\sum_{i=1}^{n} Z_i \frac{\hat{\mu}_n^*\left(1, X_i\right)}{\hat{p}_n\left(X_i\right)} \\
&\quad = 2\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_n^*\left(1, X_i\right) - \frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathbb{X}} \mathbf{1}\left\{X_i = x\right\} Z_i \frac{\hat{\mu}_n^*\left(1, x\right)}{\hat{p}_n\left(x\right)} \\
&\quad = 2\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_n^*\left(1, X_i\right) - \sum_{x\in\mathbb{X}} \frac{\hat{\mu}_n^*\left(1, x\right)}{\hat{p}_n\left(x\right)} \frac{\frac{1}{n}\sum_{i=1}^{n} Z_i \mathbf{1}\left\{X_i = x\right\}}{\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{X_i = x\right\}} \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{X_i = x\right\} \\
&\quad = 2\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_n^*\left(1, X_i\right) - \sum_{x\in\mathbb{X}} \hat{\mu}_n^*\left(1, x\right)\hat{f}(x) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_n^*\left(1, X_i\right).
\end{aligned}$$

By a similar argument,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left(1 - Z_i\right)\left(Y_i - \hat{\mu}_n^*\left(0, X_i\right)\right)}{1 - \hat{p}_n\left(X_i\right)} - \hat{\mu}_n^*\left(0, X_i\right)\right) = \frac{1}{n}\sum_{i=1}^{n}\hat{\gamma}_n\left(0, X_i\right).$$

Hence, $\hat{\lambda}_{DR} = \hat{\lambda}_n$, as desired. ∎

A.4 Superpopulation Model

We show that $\hat{\sigma}^2_{\lambda,n}$ is consistent for the asymptotic variance of $\hat{\lambda}_n$ under the superpopulation IV model where $\{(Y_i(1), Y_i(0), D_i(1), D_i(0), X_i)\}_{i=1}^n$ is i.i.d., the probability mass function of X_1 is $f(x)$, and the following identification conditions hold:

- (Exclusion) $(Y_1(0), Y_1(1), D_1(z)) \perp\!\!\!\perp Z_1 \,|\, X_1$.
- (Monotonicity) $\mathbf{P}(D_1(0) > D_1(1)) = 0$.
- (Compliers) $\mathbf{P}(D_1(1) > D_1(0)) > 0$.

Note that this nests the CI model, which is obtained by setting $D_i(1) = 1 - D_i(0) = 1$ for all i. We will also need the following definitions:

- $\mu^*(z,x) = \mathbf{E}[Y_1^*(z) \,|\, X_1 = x]$, $\mu^*(z) = \sum_{x \in \mathbb{X}} f(x)\mu^*(z,x)$,
- $\gamma(z,x) = \mathbf{E}[D_1(z) \,|\, X_1 = x]$, $\gamma(z) = \sum_{x \in \mathbb{X}} f(x)\gamma(z,x)$,
- $\lambda_n^x = (9)$,
- $\tilde{Y}_i^*(z,x) = Y_i(z) - \mu^*(z,x)$, and $\tilde{D}_i(z,x) = D_i(z) - \gamma(z,x)$.
- $v_{\mu,i}(x) = \left(\frac{Y_i^*(1) - \mu^*(1,x)}{p_n^*(x)} + \frac{Y_i^*(0) - \mu^*(0,x)}{1 - p_n^*(x)} \right) \mathbf{1}\{X_i = x\}$
- $v_{\gamma,i}(x) = \left(\frac{D_i(1) - \gamma(1,x)}{p_n^*(x)} + \frac{D_i(0) - \gamma(0,x)}{1 - p_n^*(x)} \right) \mathbf{1}\{X_i = x\}$

By an argument similar to (21),

$$\sqrt{na_n}(\hat{\lambda}_n - \lambda_n^x) = \frac{1}{\hat{\gamma}(1) - \hat{\gamma}(0)} \sum_{x \in \mathbb{X}} \sqrt{\frac{a_n}{n}} \sum_{i=1}^n \kappa_i(x) + o_p(1),$$

where

$$\begin{aligned} \kappa_i(x) = &\left(\tilde{Y}_i^*(1,x) - \tilde{Y}_i^*(0,x) \right) \mathbf{1}\{X_i = x\} \\ &- \lambda_n^x \left(\tilde{D}_i(1,x) - \tilde{D}_i(0,x) \right) \mathbf{1}\{X_i = x\} \\ &+ (Z_i - p_n^*(x))(v_{\mu,i}(x) - \lambda_n^x v_{\gamma,i}(x)). \end{aligned}$$

By the law of large numbers, $\hat{\gamma}(1) - \hat{\gamma}(0) \xrightarrow{p} \gamma(1) - \gamma(0)$. Also,

$$\operatorname{Var}\left(\sum_{x \in \mathbb{X}} \sqrt{\frac{a_n}{n}} \sum_{i=1}^n \kappa_i(x) \right) = \frac{a_n}{n} \sum_{i=1}^n \sum_{j=1}^n \sum_{x \in \mathbb{X}} \operatorname{Cov}(\kappa_i(x), \kappa_j(x)).$$

In the case of conditionally independent randomization, the covariance is zero when $i \neq j$ and otherwise equals

$$f(x) \left[\frac{1}{p_n^*(x)} \operatorname{Var}(Y_1(1) \,|\, X_1 = x) + \frac{1}{1 - p_n^*(x)} \operatorname{Var}(Y_1(0) \,|\, X_1 = x) \right]. \tag{27}$$

Under stratified block randomization, one can show, using the argument for (14), that

$$\frac{a_n}{n}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{x\in\mathbb{X}}\mathrm{Cov}(\kappa_i(x),\kappa_j(x))=o(1).$$

Thus, the asymptotic variance is

$$\sigma^2_{\lambda,n}=\sum_{x\in\mathbb{X}}f(x)\left[\frac{a_n}{p^*_n(x)}\mathrm{Var}\left(Y_1^*(1)-\mu(1,x)-\lambda^x_n(D_1(1)-\gamma(1,x))\,|\,X_1=x\right)\right.$$
$$\left.+\frac{a_n}{1-p^*_n(x)}\mathrm{Var}\left(Y_1^*(0)-\mu(0,x)-\lambda^x_n(D_1(0)-\gamma(0,x))\,|\,X_1=x\right)\right]\big(\gamma(1)-\gamma(0)\big)^{-2},$$

which for the CI model reduces to

$$\sigma^2_n=\sum_{x\in\mathbb{X}}f(x)\left[\frac{a_n}{p^*_n(x)}\mathrm{Var}(Y_1(1)\,|\,X_1=x)+\frac{a_n}{1-p^*_n(x)}\mathrm{Var}(Y_1(0)\,|\,X_1=x)\right].$$

Since $\hat{f}(x)/f(x)\xrightarrow{p}1$ by the law of large numbers and $\hat{p}^*_n(x)/p^*_n(x)\xrightarrow{p}1$ by arguments in Lemmas A.3 and A.4, by simple mean-variance calculations, we obtain $\hat{\sigma}^2_{\lambda,n}/\sigma^2_{\lambda,n}\xrightarrow{p}1$ provided (8), the identification conditions, and the following analog of (5) holds:

$$\max_{d\in\{0,1\}}\mathbf{E}\left[Y_i(d)^4\right]<\infty.$$

References

Abadie, A., S. Athey, G. Imbens, and J. Wooldridge, "Finite Population Causal Standard Errors," *working paper*, 2014.

Ansel, J., H. Hong, and J. Li, "OLS and 2SLS in Randomized and Conditionally Randomized Experiments," *Jahrbücher für Nationalökonomie und Statistik (Journal of Economics and Statistics)*, 2018, *238* (3-4), 243–293.

Bugni, F., I. Canay, and A. Shaikh, "Inference Under Covariate-Adaptive Randomization with Multiple Treatments," *working paper*, 2017.

_ , _ , and _ , "Inference Under Covariate-Adaptive Randomization," *Journal of the American Statistical Association*, 2018, *113* (524), 1784–1796.

_ , _ , and _ , "Inference Under Covariate-Adaptive Randomization with Multiple Treatments," *Quantitative Economics*, forthcoming.

Cattaneo, M., "Efficient Semiparametric Estimation of Multi-Valued Treatment Effects Under Ignorability," *Journal of Econometrics*, 2010, *155* (2), 138–154.

_ , M. Jansson, and W. Newey, "Inference in Linear Regression Models with Many Covariates and Heteroscedasticity," *Journal of the American Statistical Association*, 2018, *113* (523), 1350–1361.

Crump, R., G. Hotz, G. Imbens, and O. Mitnik, "Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand," *Biometrika*, 2009, *96*, 187–199.

Dehejia, R. and S. Wahba, "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 1999, *94* (448), 1053–1062.

Firpo, S., "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 2007, *75* (1), 259–276.

Freedman, D., "On Regression Adjustments to Experimental Data," *Advances in Applied Mathematics*, 2008, *40* (2), 180–193.

Frölich, M., "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 2007, *139* (1), 35–75.

Heckman, J., H. Ichimura, and P. Todd, "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *The Review of Economic Studies*, 1997, *64* (4), 605–654.

Hinkelmann, K. and O. Kempthorne, *Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design, 2nd ed.*, John Wiley & Sons, 2008.

Hirano, K., G. Imbens, and G. Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 2003, *71* (4), 1161–1189.

Ho, D., K. Imai, G. King, and E. Stuart, "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," *Political Analysis*, 2007, *15* (3), 199–236.

Imbens, G and D. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press, 2015.

Imbens, G. and J. Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994, *62* (2), 467–475.

Imbens, Guido W and Jeffrey M Wooldridge, "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 2009, *47* (1), 5–86.

Lehmann, E. and H. D'Abrera, *Nonparametrics: Statistical Methods Based on Ranks*, Springer New York, 2006.

Li, X. and P. Ding, "General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference," *Journal of the American Statistical Association*, 2017, *112*, 1759–1769.

Ma, X. and J. Wang, "Robust Inference Using Inverse Probability Weighting," *working paper*, 2018.

Neyman, J., "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," *Annals of Agricultural Sciences*, 1923, *10*, 1–51.

Reichardt, C. and H. Gollob, "Justifying the Use and Increasing the Power of a *t*-Test for a Randomized Experiment with a Convenience Sample," *Psychological methods*, 1999, *4* (1), 117.

Robins, J., A. Rotnitzky, and L. Zhao, "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 1994, *89* (427), 846–866.

Rosenbaum, P., *Observational Studies, 2nd ed.*, Springer, 2002.

Rothe, C., "Robust Confidence Intervals for Average Treatment Effects Under Limited Overlap," *Econometrica*, 2017, *85* (2), 645–660.

Sasaki, Y. and T. Ura, "Estimation and Inference for Moments of Ratios with Robustness against Large Trimming Bias," *working paper*, 2018.