2-2024

Improving estimation efficiency via regression-adjustment in covariate-adaptive randomizations with imperfect compliance

Liang JIANG
Fudan University

Oliver B. LINTON
University of Cambridge

Haihan TANG
Fudan University

Yichong ZHANG
Singapore Management University, yczhang@smu.edu.sg

Improving Estimation Efficiency via Regression-Adjustment in Covariate-Adaptive Randomizations with Imperfect Compliance*

Liang Jiang† Oliver B. Linton‡ Haihan Tang§ Yichong Zhang¶

September 7, 2022

Abstract

We study how to improve efficiency via regression adjustments with additional covariates under covariate-adaptive randomizations (CARs) when subject compliance is imperfect. We first establish the semiparametric efficiency bound for the local average treatment effect (LATE) under CARs. Second, we develop a general regression-adjusted LATE estimator which allows for parametric, nonparametric, and regularized adjustments. Even when the adjustments are misspecified, our proposed estimator is still consistent and asymptotically normal, and their inference method still achieves the exact asymptotic size under the null. When the adjustments are correctly specified, our estimator achieves the semiparametric efficiency bound. Third, we derive the optimal linear adjustment that leads to the smallest asymptotic variance among all linear adjustments. We then show the commonly used two stage least squares estimator is not optimal in the class of LATE estimators with linear adjustments while Ansel, Hong, and Li's (2018) estimator is. Fourth, we show how to construct a LATE estimator with nonlinear adjustments which is more efficient than those with the optimal linear adjustment. Fifth, we give conditions under which LATE estimators with nonparametric and regularized adjustments achieve the semiparametric efficiency bound. Last, simulation evidence and empirical application confirm efficiency gains achieved by regression adjustments relative to both the estimator without adjustment and the standard two-stage least squares estimator.

Keywords: Randomized experiment, Covariate-adaptive randomization, High-dimensional data, Local average treatment effects, Regression adjustment.

JEL codes: C14, C21, I21

*Yichong Zhang acknowledges the financial support from Singapore Ministry of Education Tier 2 grant under grant MOE2018-T2-2-169 and the Lee Kong Chian fellowship. Any and all errors are our own.

†Fudan University. E-mail address: jiangliang@fudan.edu.cn.

‡University of Cambridge. E-mail address: obl20@cam.ac.uk.

§Fudan University. E-mail address: hhtang@fudan.edu.cn.

¶Singapore Management University. E-mail address: yczhang@smu.edu.sg.

1 Introduction

Randomized experiments have seen increasing use in economic research. In existing experiments, one of the popular randomization methods applied by economists to achieve balance between treatment and control is covariate-adaptive randomization (CAR) (Bruhn and McKenzie, 2009), in which subjects are randomly assigned to treatment and control within strata formed by a few key pre-treatment variables. However, subject compliance with random assignment is usually imperfect. Recent studies in economics using CAR and with imperfect compliance include, for example, Hirshleifer, McKenzie, Almeida, and Ridao-Cano (2016), Blattman and Dercon (2018), Dupas, Karlan, Robinson, and Ubfal (2018), and Banerjee, Hanna, Kyle, Olken, and Sumarto (2018).

In CARs with imperfect compliance, the local average treatment effects (LATEs)–the average treatment effects among those who comply with the assignment, can be identified, as formulated in the seminal work by Imbens and Angrist (1994). To estimate the LATE, researchers usually run two-stage least squares (TSLS) with additional covariates as exogenous regressors and treatment assignment as instrumental variable, in which heteroscedasticity-robust standard errors is used for inference. Two issues arise with such practice. First, Ansel et al. (2018) and Bugni and Gao (2021) recently point out that, due to the cross-sectional dependence introduced by the CARs, the standard errors of various TSLS estimators are inconsistent. They then propose consistent alternatives. Second, Freedman (2008a,b) point out that simply including extra covariates as linear regressors may lead to degradation of the estimation precision. Ansel et al. (2018) address this issue by interacting additional covariates with stratum dummies and treatment assignment (the S estimator of Ansel et al. (2018) hereafter), which is guaranteed to be more efficient than various TSLS estimators discussed in their paper (their proposition 7).

However, is the S estimator proposed by Ansel et al. (2018) the most efficient among all linear adjusted LATE estimators? Can nonlinear, nonparametric, and regularized adjustments lead to further efficiency improvement? More profoundly, what is the semiparametric efficiency bound for the LATE estimation under CARs? What is a systematic but flexible method to use additional covariates that is guaranteed to improve the estimation precision and can potentially achieve the efficiency bound?

To answer these questions, after briefly analyzing existing methods, we first derive the semiparametric efficiency bound for the LATE under CARs with additional covariates. Such a bound is new to the literature because the treatment statuses generated by CARs are not independent. It complements the recent work by Armstrong (2022), which derives the semiparametric efficiency bound for ATE without additional covariates under CARs. We also observe that Ansel et al.'s (2018) S estimator does not achieve the efficiency bound in general.

Second, we propose a flexible way to make regression adjustments in the estimation of LATE. The new method can accommodate linear, nonlinear, nonparametric, and regularized adjustments. Our theoretical analysis follows a new asymptotic framework that was recently established by

Bugni, Canay, and Shaikh (2018) to study ATE estimators under CARs. This framework accounts for the cross-sectional dependence caused by the randomization. We thus develop an inference method that (1) achieves the exact asymptotic size under the null despite the cross-sectional dependence introduced by CARs, (2) is robust to adjustment misspecification, and (3) achieves the semiparametric efficiency bound when the adjustments are correctly specified.

Third, we further investigate efficiency gains brought by regression adjustments in parametric (both linear and nonlinear), nonparametric, and regularized forms. When adjustments are linear, we drive the most efficient estimator (denoted as LP) among all linearly adjusted LATE estimators, and in particular, show that it is weakly more efficient than the estimator without adjustment (denoted as NA, which coincides with Bugni and Gao's (2021) fully saturated estimator) and the commonly used TSLS estimator. We prove that the S estimator proposed by Ansel et al. (2018) is asymptotically equivalent to LP and thus linearly optimal. However, because the linear model is likely misspecified for the binary treatment status, it is expected that nonlinearly adjusted LATE estimator can be more efficient than the one proposed by Ansel et al. (2018). In fact, we further construct a new estimator (denoted as F) which combines the both linear and logistic adjustments (denoted as LG) and show it is weakly more efficient than both the optimally linearly-adjusted and the unadjusted estimators. The new estimator is also guaranteed to be weakly more efficient than the TSLS estimator. Last, we study the nonparametric (denoted as NP) and regularized adjustments (denoted as R) which are completely new to the literature and provide conditions under which the two methods achieve the semiparametric efficiency bound as if the adjustments are correctly specified. Figure 1 visualizes the partial order of efficiency for these estimators.

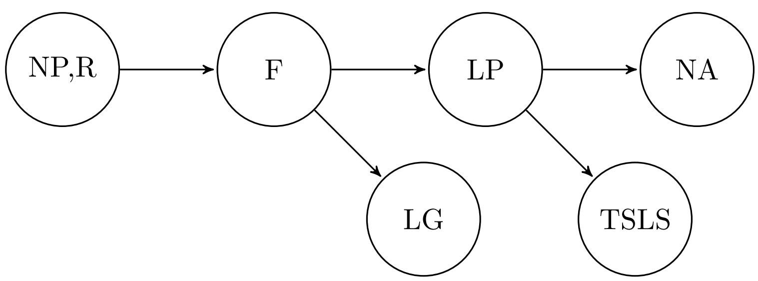

Figure 1: Efficiency of Various LATE Estimators (from the most efficient to the least)

The final contribution of the paper is to provide simulation evidence and empirical support for the efficiency gains achieved by our regression-adjusted LATE estimator. We compare it with both the one without any adjustment and the one obtained by the commonly used TSLS and confirm sizable efficiency gains that can be achieved by regression adjustments. In the empirical application, we revisit the experiment with a CAR design in Dupas et al. (2018). We find that by just using the same two covariates adopted in that paper, over nine outcome variables, the standard errors of our adjusted LATE estimators are on average around 7% lower than those without adjustments.

For some outcome variables, regression adjustments can reduce the standard errors by about 15%. Compared with the TSLS estimators, the standard errors of our estimators are generally smaller as well, although by a smaller margin.

Our paper is related to several lines of research. Hu and Hu (2012); Ma, Hu, and Zhang (2015); Ma, Qin, Li, and Hu (2020); Olivares (2021); Shao and Yu (2013); Zhang and Zheng (2020); Ye (2018); Ye and Shao (2020) studied inference of either ATEs or QTEs under CARs without considering additional covariates. Bugni et al. (2018); Bugni, Canay, and Shaikh (2019); Bloniarz, Liu, Zhang, Sekhon, and (2016); Fogarty (2018); Lin (2013); Lu (2016); Lei and Ding (2021); Li and Ding (2020); Liu, Tu, and Ma (2020); Liu and Yang (2020); Negi and Wooldridge (2020); Shao, Yu, and Zhong (2010); Ye, Yi, and Shao (2021); Zhao and Ding (2021) studied the estimation and inference of ATEs using a variety of regression methods under various randomization. Jiang, Phillips, Tao, and Zhang (2022) examined regression-adjusted estimation and inference of QTEs under CARs. Based on pilot experiments, Tabord-Meehan (2021) and Bai (2020) devise optimal randomization designs that may produce an ATE estimator with the lowest variance. All the works above assume perfect compliance while we contribute to the literature by studying the LATE estimators in the context of CARs and regression adjustment, which allows for imperfect compliance. Ren and Liu (2021) studied the regression-adjusted LATE estimator in completely randomized experiments for a *binary* outcome using the finite population asymptotics. We differ from their work by considering the regression-adjusted estimator in *covariate-adaptive* randomizations for a *general* outcome using the *superpoluation* asymptotics. We also derive the semiparametric efficiency bound of the LATE under CARs with additional covariates. Such a result is new to the literature of regression adjustments under various randomization schemes. Finally, our paper also connects to a vast literature on the estimation and inference in randomized experiments, including Hahn, Hirano, and Karlan (2011); Athey and Imbens (2017); Abadie, Chingos, and West (2018); Tabord-Meehan (2021); Bai, Shaikh, and Romano (2021); Bai (2020); Jiang, Liu, Phillips, and Zhang (2021) among many others.

The rest of the paper is organized as follows. Section 2 lays out our setup. Section 3 studies the existing methods to estimate LATE with additional covariates. Section 4 derives the semiparametric efficiency bound for the LATE under CARs. Section 5 introduces our flexible regression-adjusted LATE estimator $\hat{\tau}$. We establish the asymptotic properties of $\hat{\tau}$ in Section 5 under high-level conditions. We then examine efficiency of $\hat{\tau}$ in contexts of parametric, nonparametric, and regularized adjustments in Sections 6, 7 and 8, respectively. We conduct Monte Carlo simulations in Section 9 and an empirical application in Section 10. Section 11 concludes. Some implementation details for sieve and Lasso regressions and the proofs of the theoretical results are included in the Online Supplement.

2 Setup

Let Y_i denote the observed outcome of interest for individual i; write $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$, where $Y_i(1), Y_i(0)$ are the individual i's hypothetical treated and untreated outcomes, respectively, and D_i is a binary random variable indicating whether individual i received the treatment ($D_i = 1$) or not ($D_i = 0$) in the actual study. One could link D_i to the treatment assignment A_i in the following way: $D_i = D_i(1)A_i + D_i(0)(1 - A_i)$, where $D_i(a)$ is the individual i's treatment decision upon receiving treatment status $A_i = a$ for $a = 0, 1$; $D_i(a)$ is a binary random variable. Define $Y_i(D_i(a)) := Y_i(1)D_i(a) + Y_i(0)(1 - D_i(a))$, so we can write $Y_i = Y_i(D_i(1))A_i + Y_i(D_i(0))(1 - A_i)$.

Consider a CAR with n individuals; that is, a researcher can observe the data $\{Y_i, D_i, A_i, S_i, X_i\}_{i=1}^n$. The support of vectors $\{X_i\}_{i=1}^n$ is denoted $\text{Supp}(X)$. Define $[n] := \{1, 2, ...n\}$, $p(s) := \mathbb{P}(S_i = s)$, $n(s) := \sum_{i\in[n]} 1\{S_i = s\}$, $n_1(s) := \sum_{i\in[n]} A_i 1\{S_i = s\}$, $n_0(s) := n(s) - n_1(s)$, $S^{(n)} := (S_1, \ldots, S_n)$, $X^{(n)} := (X_1, \ldots, X_n)$, and $A^{(n)} := (A_1, \ldots, A_n)$. We make the following assumptions on the data generating process (DGP) and the treatment assignment rule.

Assumption 1. *(i)* $\{Y_i(1), Y_i(0), D_i(0), D_i(1), S_i, X_i\}_{i=1}^n$ *is i.i.d. over* i*. For each* i*, We allow* X_i *and* S_i *to be dependent.*

(ii) $\{Y_i(1), Y_i(0), D_i(0), D_i(1), X_i\}_{i=1}^n \perp\!\!\!\perp A^{(n)} | S^{(n)}$.

(iii) Suppose that $p(s)$ *is fixed w.r.t.* n *and positive for every* $s \in \mathcal{S}$.

(iv) Let $\pi(s)$ *denote the propensity score for stratum* s *(i.e., the targeted assignment probability for stratum* s*). Then,* $c < \min_{s\in\mathcal{S}} \pi(s) \leq \max_{s\in\mathcal{S}} \pi(s) < 1 - c$ *for some constant* $c \in (0, 0.5)$ *and* $\frac{B_n(s)}{n(s)} = o_p(1)$ *for* $s \in \mathcal{S}$*, where* $B_n(s) := \sum_{i=1}^n (A_i - \pi(s))1\{S_i = s\}$.

(v) Suppose $\mathbb{P}(D(1) = 0, D(0) = 1) = 0$.

(vi) $\max_{a=0,1,s\in\mathcal{S}} \mathbb{E}(|Y_i(a)|^q | S_i = s) \leq C < \infty$ *for* $q \geq 4$.

Several remarks are in order. First, Assumption 1(ii) implies that the treatment assignment $A^{(n)}$ are generated only based on strata indicators. Second, Assumption 1(iii) imposes that the sizes of strata are balanced. Third, Bugni et al. (2018) show that Assumption 1(iv) holds under several covariate-adaptive treatment assignment rules such as simple random sampling (SRS), biased-coin design (BCD), adaptive biased-coin design (WEI), and stratified block randomization (SBR). For completeness, we briefly repeat their descriptions below. Note that we only require $B_n(s)/n(s) = o_p(1)$, which is weaker than the assumption imposed by Bugni et al. (2018) but the same as that imposed by Bugni et al. (2019) and Zhang and Zheng (2020). Fourth, Assumption 1(v) implies there are no defiers. Last, Assumption 1(vi) is a standard moment condition.

Example 1 (SRS)**.** *Let* $\{A_i\}_{i=1}^n$ *be drawn independently across* i *and of* $\{S_i\}_{i=1}^n$ *as Bernoulli random variables with success rate* π*, i.e., for* $k = 1, \cdots, n$,

$$\mathbb{P}\left(A_k = 1 \middle| \{S_i\}_{i=1}^n, \{A_j\}_{j=1}^{k-1}\right) = \mathbb{P}(A_k = 1) = \pi.$$

Then, Assumption 1(iv) holds with $\pi(s) = \pi$ for all s.

Example 2 (WEI)**.** *This design was first proposed by Wei (1978). Let $n_{k-1}(S_k) = \sum_{i=1}^{k-1} 1\{S_i = S_k\}$, $B_{k-1}(S_k) = \sum_{i=1}^{k-1}\left(A_i - \frac{1}{2}\right) 1\{S_i = S_k\}$, and*

$$\mathbb{P}\left(A_k = 1 \middle| \{S_i\}_{i=1}^{k}, \{A_i\}_{i=1}^{k-1}\right) = f\left(\frac{2B_{k-1}(S_k)}{n_{k-1}(S_k)}\right),$$

where $f(\cdot) : [-1, 1] \mapsto [0, 1]$ is a pre-specified non-increasing function satisfying $f(-x) = 1 - f(x)$. Here, $\frac{B_0(S_1)}{n_0(S_1)}$ and $B_0(S_1)$ are understood to be zero. Then, Bugni et al. (2018) show that Assumption 1(iv) holds with $\pi(s) = \frac{1}{2}$ for all s. Recently, Hu (2016) generalized the adaptive biased-coin design to multiple treatment values and unequal target factions.

Example 3 (BCD)**.** *The treatment status is determined sequentially for $1 \leq k \leq n$ as*

$$\mathbb{P}\left(A_k = 1 | \{S_i\}_{i=1}^{k}, \{A_i\}_{i=1}^{k-1}\right) = \begin{cases} \frac{1}{2} & \text{if } B_{k-1}(S_k) = 0 \\ \lambda & \text{if } B_{k-1}(S_k) < 0 \\ 1 - \lambda & \text{if } B_{k-1}(S_k) > 0, \end{cases}$$

where $B_{k-1}(s)$ is defined as above and $\frac{1}{2} < \lambda \leq 1$. Then, Bugni et al. (2018) show that Assumption 1(iv) holds with $\pi(s) = \frac{1}{2}$ for all s.

Example 4 (SBR)**.** *For each stratum, $\lfloor \pi(s) n(s) \rfloor$ units are assigned to treatment and the rest are assigned to control. Then obviously Assumption 1(iv) holds as $n(s) \to \infty$.*

Throughout the paper, we are interested in estimating the *local average treatment effect* (LATE) which is denoted τ and defined as

$$\tau := \mathbb{E}\left[Y(1) - Y(0) | D(1) > D(0)\right];$$

that is, we are interested in the ATE for the compliers (Angrist and Imbens (1994)).

3 Existing Methods to Estimate LATE with Extra Covariates

3.1 Two Stage Least Squares

Empirical researchers using CARs usually estimate LATE via two stage least squares (TSLS) regressions with strata dummies and additional covariates (e.g. baseline variables) as exogenous regressors. Examples include Bruhn, Ibarra, and McKenzie (2014), McKenzie, Assaf, and Cusolito (2017), Banerjee et al. (2018), and Dupas et al. (2018).

The form of the TSLS regression commonly taken in empirical studies using CARs is

$$D_i = \gamma A_i + \alpha_s + X_i'\theta + e_i,$$
$$Y_i = \tau D_i + \alpha_s + X_i'\delta + \varepsilon_i$$

where $\{\alpha_s\}_{s\in\mathcal{S}}$ are the strata fixed effects, and $\{e_i\}_{i\in[n]}$ and $\{\varepsilon_i\}_{i\in[n]}$ are the error terms in the first and second stage regressions, respectively.

Denote the TSLS estimator as $\hat{\tau}_{tsls}$. Under Assumptions 1 and 1(iv') in the Online Supplement, we derive the probability limit of $\hat{\tau}_{tsls}$ in Theorem B.1 in Section B of the Online Supplement and show it equals τ when $\pi(s)$ is homogeneous across $s \in \mathcal{S}$. In this case, we further establish its asymptotic normality with an asymptotic variance σ^2_{tsls}.

3.2 The S Estimator in Ansel et al. (2018)

Ansel et al. (2018) propose a LATE estimator adjusted with extra covariates. It takes the form

$$\hat{\tau}_S := \frac{\sum_{s\in\mathcal{S}} \hat{p}(s)(\hat{\gamma}^Y_{1s} - \hat{\gamma}^Y_{0s} + (\hat{\nu}^Y_{1s} - \hat{\nu}^Y_{0s})^\top \bar{X}_s)}{\sum_{s\in\mathcal{S}} \hat{p}(s)(\hat{\gamma}^D_{1s} - \hat{\gamma}^D_{0s} + (\hat{\nu}^D_{1s} - \hat{\nu}^D_{0s})^\top \bar{X}_s)},$$

where $\hat{p}(s) := n(s)/n$, $\bar{X}_s := \frac{1}{n\hat{p}(s)} \sum_{i\in[n]} X_i 1\{S_i = s\}$, and $(\hat{\gamma}^Y_{as}, \hat{\gamma}^D_{as}, \hat{\nu}^Y_{as}, \hat{\nu}^D_{as})$ for $a = 0, 1$ are the estimated coefficients of four sets of cluster-specific regressions using only the s cluster:

$$(1 - A_i)Y_i = (1 - A_i)(\gamma^Y_{0s} + X_i^\top \nu^Y_{0s} + e^Y_{0i}), \quad A_iY_i = A_i(\gamma^Y_{1s} + X_i^\top \nu^Y_{1s} + e^Y_{1i})$$
$$(1 - A_i)D_i = (1 - A_i)(\gamma^D_{0s} + X_i^\top \nu^D_{0s} + e^D_{0i}), \quad A_iD_i = A_i(\gamma^D_{1s} + X_i^\top \nu^D_{1s} + e^D_{1i}).$$

Under Assumption 1 and Assumption 1(iv') in the Online Supplement, Ansel et al. (2018) show that $\hat{\tau}_S$ is the consistent estimator of τ and the most efficient among the estimators studied in their paper. In Section C in the Online Supplement, we further derive the explicit expression for $\hat{\tau}_S$'s asymptotic variance in our notation, which is denoted as σ^2_S.

Theorem 3.1. *Suppose that Assumption 1 and Assumption 1(iv') in the Online Supplement hold. Moreover, suppose that $\pi(s)$ is the same across $s \in \mathcal{S}$. Then $\hat{\tau}_S$ is more efficient than $\hat{\tau}_{tsls}$ in the sense that $\sigma^2_S \leq \sigma^2_{tsls}$.*

Both $\hat{\tau}_{tsls}$ and $\hat{\tau}_S$ use linear adjustments of X_i, but Theorem 3.1 states that $\hat{\tau}_S$ is more efficient than $\hat{\tau}_{tsls}$. In Theorem 6.3 below, we further show that $\hat{\tau}_S$ achieves the minimum asymptotic variance among the class of estimators with linear adjustments.

On the other hand, nonlinear adjustments may be more efficient than the optimal linear adjustment. In the next section, we establish the semiparametric efficiency bound for τ, which has not been previously calculated in the literature. Comparing the semiparametric efficiency bound and the expression of σ^2_S derived in Section C in the Online Supplement, we observe that the S estimator in general does not reach the efficiency bound. In Section 5, we propose a general regression-adjusted LATE estimator which allows for more general forms of adjustments, and thus can potentially achieve the semiparametric efficiency bound.

4 Semiparametric Efficiency Bound

Define $\mu^D(a,s,x) := \mathbb{E}\left[D(a)|S=s, X=x\right]$ and $\mu^Y(a,s,x) := \mathbb{E}\left[Y(D(a))|S=s, X=x\right]$ for $a = 0,1$ as the true specifications. In practice, the true specifications are unknown and empirical researchers employ working models $\overline{\mu}^D(a,s,x)$ and $\overline{\mu}^Y(a,s,x)$, which may differ from the true specifications. Let $\mathcal{D}_i := \{Y_i(1), Y_i(0), D_i(1), D_i(0), X_i\}$,

$$
\begin{aligned}
W_i &:= Y_i(D_i(1)), \quad Z_i := Y_i(D_i(0)), \\
\tilde{W}_i &:= W_i - \mathbb{E}[W_i|S_i], \quad \tilde{Z}_i := Z_i - \mathbb{E}[Z_i|S_i], \quad \tilde{X}_i := X_i - \mathbb{E}[X_i|S_i] \\
\tilde{D}_i(a) &:= D_i(a) - \mathbb{E}[D_i(a)|S_i], \quad a = 0,1.
\end{aligned}
$$

For $a = 0,1$, further let

$$
\tilde{\mu}^Y(a,S_i,X_i) := \overline{\mu}^Y(a,S_i,X_i) - \mathbb{E}\left[\overline{\mu}^Y(a,S_i,X_i)|S_i\right], \tag{4.1}
$$

$$
\tilde{\mu}^D(a,S_i,X_i) := \overline{\mu}^D(a,S_i,X_i) - \mathbb{E}\left[\overline{\mu}^D(a,S_i,X_i)|S_i\right],
$$

$$
\begin{aligned}
\Xi_1(\mathcal{D}_i,S_i) := & \left[\left(1-\frac{1}{\pi(S_i)}\right)\tilde{\mu}^Y(1,S_i,X_i) - \tilde{\mu}^Y(0,S_i,X_i) + \frac{\tilde{W}_i}{\pi(S_i)}\right] \\
& - \tau\left[\left(1-\frac{1}{\pi(S_i)}\right)\tilde{\mu}^D(1,S_i,X_i) - \tilde{\mu}^D(0,S_i,X_i) + \frac{\tilde{D}_i(1)}{\pi(S_i)}\right],
\end{aligned} \tag{4.2}
$$

$$
\begin{aligned}
\Xi_0(\mathcal{D}_i,S_i) := & \left[\left(\frac{1}{1-\pi(S_i)}-1\right)\tilde{\mu}^Y(0,S_i,X_i) + \tilde{\mu}^Y(1,S_i,X_i) - \frac{\tilde{Z}_i}{1-\pi(S_i)}\right] \\
& - \tau\left[\left(\frac{1}{1-\pi(S_i)}-1\right)\tilde{\mu}^D(0,S_i,X_i) + \tilde{\mu}^D(1,S_i,X_i) - \frac{\tilde{D}_i(0)}{1-\pi(S_i)}\right],
\end{aligned} \tag{4.3}
$$

$$
\Xi_2(S_i) := \left(\mathbb{E}[W_i - Z_i|S_i] - \mathbb{E}[W_i - Z_i]\right) - \tau\left(\mathbb{E}[D_i(1)-D_i(0)|S_i] - \mathbb{E}[D_i(1)-D_i(0)]\right). \tag{4.4}
$$

Theorem 4.1. *Suppose Assumptions 1 and E in the Online Supplement hold. For $a = 0,1$, define $\underline{\Xi}_1(\mathcal{D}_i,S_i)$, $\underline{\Xi}_0(\mathcal{D}_i,S_i)$ and $\underline{\Xi}_2(S_i)$ as $\Xi_1(\mathcal{D}_i,S_i)$, $\Xi_0(\mathcal{D}_i,S_i)$ and $\Xi_2(S_i)$ in (4.2)–(4.4), respectively, with the researcher-specified working model $\overline{\mu}^b(a,s,x)$ equal to the true specification $\mu^b(a,s,x)$ for all $(a,b,s,x) \in \{0,1\}\times\{D,Y\}\times\mathcal{SX}$, where $\mathcal{SX}$ is the joint support of (S,X). Then the semiparametric efficiency bound for τ is*

$$
\underline{\sigma}^2 := \frac{\underline{\sigma}_1^2 + \underline{\sigma}_0^2 + \underline{\sigma}_2^2}{\mathbb{P}(D(1) > D(0))^2},
$$

where

$$
\underline{\sigma}_1^2 := \mathbb{E}\left[\pi(S_i)\underline{\Xi}_1^2(\mathcal{D}_i,S_i)\right], \quad \underline{\sigma}_0^2 := \mathbb{E}\left[(1-\pi(S_i))\underline{\Xi}_0^2(\mathcal{D}_i,S_i)\right], \quad \underline{\sigma}_2^2 := \mathbb{E}\underline{\Xi}_2^2(S_i).
$$

Several remarks on in order. First, heuristically, Theorem 4.1 implies that the asymptotic variance of any root-n consistent and asymptotically normal semiparametric estimator of LATE is lower bounded by $\underline{\sigma}^2$. Second, the proof of Theorem 4.1 follows the argument in Armstrong (2022) which takes into consideration that $\{A_i\}_{i\in[n]}$ is cross-sectionally dependent. Third, the efficiency bound here is slightly different from what derived by Frölich (2007) under unconfoundedness for observational data because the additional covariates X_i only enter the conditional mean models (i.e., $\mu^b(a,s,x)$ for $a=0,1$, $b=\{D,Y\}$) but not the "propensity score" $\pi(\cdot)$. Fourth, Theorem 4.1 implies that various CARs (with or without achieving strong balance) lead to the same efficiency for LATE estimation. Such a result is consistent with what is discovered for ATE under general randomization schemes by Armstrong (2022). Fifth, Ansel et al.'s (2018) $\hat{\tau}_S$ with asymptotic variance defined in (C.1) in the Online Supplement may not achieve this efficiency bound because the linear adjustments may be misspecified. In Section 5, we propose a general regression-adjusted estimator which allows for nonlinear, nonparametric, and regularized adjustments and that can potentially achieves the efficiency bound.

5 The General Estimator and its Asymptotic Properties

In this section, we propose a general regression-adjusted LATE estimator for τ. Recall researchers specify working models $\overline{\mu}^D(a,s,x)$ and $\overline{\mu}^Y(a,s,x)$, which may differ from the true specifications. They then proceed to estimate the working models with estimators $\hat{\mu}^D(a,s,x)$ and $\hat{\mu}^Y(a,s,x)$. Again, as the working models are potentially misspecified, their estimators are potentially inconsistent to the true specifications.

In CAR, the propensity score is usually known or can be consistently estimated by $\hat{\pi}(s) := \frac{n_1(s)}{n(s)}$.[1] Then our proposed estimator of LATE is

$$\hat{\tau} := \left(\frac{1}{n}\sum_{i\in[n]}\Xi_{H,i}\right)^{-1}\left(\frac{1}{n}\sum_{i\in[n]}\Xi_{G,i}\right), \quad \text{where} \tag{5.1}$$

$$\Xi_{H,i} := \frac{A_i(D_i-\hat{\mu}^D(1,S_i,X_i))}{\hat{\pi}(S_i)} - \frac{(1-A_i)(D_i-\hat{\mu}^D(0,S_i,X_i))}{1-\hat{\pi}(S_i)} + \hat{\mu}^D(1,S_i,X_i) - \hat{\mu}^D(0,S_i,X_i), \tag{5.2}$$

$$\Xi_{G,i} := \frac{A_i(Y_i-\hat{\mu}^Y(1,S_i,X_i))}{\hat{\pi}(S_i)} - \frac{(1-A_i)(Y_i-\hat{\mu}^Y(0,S_i,X_i))}{1-\hat{\pi}(S_i)} + \hat{\mu}^Y(1,S_i,X_i) - \hat{\mu}^Y(0,S_i,X_i). \tag{5.3}$$

[1] This is because

$$\hat{\pi}(s)-\pi(s) = \frac{1}{n(s)}\sum_{i=1}^{n}(A_i-\pi(s))1\{S_i=s\} = \frac{B_n(s)}{n(s)} = o_p(1).$$

This estimator takes the form of doubly robust moments (see Robins, Rotnitzky, and Zhao (1994), Robins and Rotnitzky (1995), Scharfstein, Rotnitzky, and Robins (1999), Robins, Rotnitzky, and van der Laan (2000), Hirano and Imbens (2001), Frölich (2007), Wooldridge (2007), Rothe and Firpo (2019) etc; see Słoczyński and Wooldridge (2018) and Seaman and Vansteelandt (2018) for recent reviews). To the best of our knowledge, we are the first to apply doubly robust methods to study the LATE under CARs. Our analysis takes into account the cross-sectional dependence of the treatment statuses caused by the randomization and is therefore different from the double robustness literature that mostly focuses on the observational data with independent treatment statuses.

We make the following high-level assumptions on the regression adjustments.

Assumption 2. *(i) For $a = 0, 1$ and $s \in \mathcal{S}$, define*

$$\Delta^Y(a, s, X_i) := \hat{\mu}^Y(a, s, X_i) - \overline{\mu}^Y(a, s, X_i), \quad \Delta^D(a, s, X_i) := \hat{\mu}^D(a, s, X_i) - \overline{\mu}^D(a, s, X_i), \quad \text{and}$$
$$I_a(s) := \left\{i \in [n] : A_i = a, S_i = s\right\}.$$

Then, for $a = 0, 1$, $b = D, Y$, we have

$$\max_{s \in \mathcal{S}} \left| \frac{\sum_{i \in I_1(s)} \Delta^b(a, s, X_i)}{n_1(s)} - \frac{\sum_{i \in I_0(s)} \Delta^b(a, s, X_i)}{n_0(s)} \right| = o_p(n^{-1/2}).$$

(ii) For $a = 0, 1$ and $b = D, Y$, $\frac{1}{n}\sum_{i=1}^n \Delta^{b,2}(a, S_i, X_i) = o_p(1)$.

(iii) Suppose $\max_{a=0,1,s\in\mathcal{S}} \mathbb{E}(\overline{\mu}^{b,2}(a, S_i, X_i)|S_i = s) \leq C < \infty$ for $b = D, Y$ and some constant C.

Assumption 2 is mild. Consider a linear working model $\overline{\mu}^Y(a, s, X_i) = X_i^\top \beta_{a,s}$, where the coefficient $\beta_{a,s}$ may vary across treatment assignments and strata. Its estimator $\hat{\mu}^Y(a, s, X_i)$ can be written as $X_i^\top \hat{\beta}_{a,s}$ where $\hat{\beta}_{a,s}$ is an estimator of $\beta_{a,s}$. Then, Assumption 2(i) requires

$$\max_{s \in \mathcal{S}, a=0,1} \left| \left(\frac{1}{n_1(s)} \sum_{i \in I_1(s)} X_i - \frac{1}{n_0(s)} \sum_{i \in I_0(s)} X_i \right)^\top (\hat{\beta}_{a,s} - \beta_{a,s}) \right| = o_p(n^{-1/2}), \tag{5.4}$$

which holds whenever $\hat{\beta}_{a,s} \xrightarrow{p} \beta_{a,s}$. Similar remark applies to Assumption 2(ii).

Then Theorem 5.1 below shows that $\hat{\tau}$ is asymptotically normal with the asymptotic variance

$$\sigma^2 := \frac{\sigma_1^2 + \sigma_0^2 + \sigma_2^2}{\mathbb{P}(D(1) > D(0))^2}, \tag{5.5}$$

where

$$\sigma_1^2 := \mathbb{E}\left[\pi(S_i)\Xi_1^2(\mathcal{D}_i, S_i)\right], \quad \sigma_0^2 := \mathbb{E}\left[(1 - \pi(S_i))\Xi_0^2(\mathcal{D}_i, S_i)\right], \quad \sigma_2^2 := \mathbb{E}\left[\Xi_2^2(S_i)\right],$$

and $\Xi_1(\mathcal{D}_i, S_i)$, $\Xi_0(\mathcal{D}_i, S_i)$, and $\Xi_2(S_i)$ are defined in (4.2)–(4.4), respectively.

Next, we propose an estimator $\hat{\sigma}^2$ of σ^2. Recalling $\Xi_{H,i}$ defined in (5.2), we define $\hat{\sigma}^2$ as

$$\hat{\sigma}^2 = \frac{\frac{1}{n}\sum_{i=1}^n \left[A_i\hat{\Xi}_1^2(\mathcal{D}_i, S_i) + (1-A_i)\hat{\Xi}_0^2(\mathcal{D}_i, S_i) + \hat{\Xi}_2^2(S_i)\right]}{\left(\frac{1}{n}\sum_{i=1}^n \Xi_{H,i}\right)^2},$$

where

$$\begin{aligned}
\hat{\Xi}_1(\mathcal{D}_i, s) &:= \tilde{\Xi}_1(\mathcal{D}_i, s) - \frac{1}{n_1(s)}\sum_{i\in I_1(s)} \tilde{\Xi}_1(\mathcal{D}_i, s),\\
\hat{\Xi}_0(\mathcal{D}_i, s) &:= \tilde{\Xi}_0(\mathcal{D}_i, s) - \frac{1}{n_0(s)}\sum_{i\in I_0(s)} \tilde{\Xi}_0(\mathcal{D}_i, s),\\
\hat{\Xi}_2(s) &:= \left(\frac{1}{n_1(s)}\sum_{i\in I_1(s)}(Y_i - \hat{\tau}D_i)\right) - \left(\frac{1}{n_0(s)}\sum_{i\in I_0(s)}(Y_i - \hat{\tau}D_i)\right),\\
\tilde{\Xi}_1(\mathcal{D}_i, s) &:= \left[\left(1 - \frac{1}{\hat{\pi}(s)}\right)\hat{\mu}^Y(1, s, X_i) - \hat{\mu}^Y(0, s, X_i) + \frac{Y_i}{\hat{\pi}(s)}\right]\\
&\quad - \hat{\tau}\left[\left(1 - \frac{1}{\hat{\pi}(s)}\right)\hat{\mu}^D(1, s, X_i) - \hat{\mu}^D(0, S_i, X_i) + \frac{D_i}{\hat{\pi}(s)}\right], \quad \text{and}\\
\tilde{\Xi}_0(\mathcal{D}_i, s) &:= \left[\left(\frac{1}{1-\hat{\pi}(s)} - 1\right)\hat{\mu}^Y(0, s, X_i) + \hat{\mu}^Y(1, s, X_i) - \frac{Y_i}{1-\hat{\pi}(s)}\right]\\
&\quad - \hat{\tau}\left[\left(\frac{1}{1-\hat{\pi}(s)} - 1\right)\hat{\mu}^D(0, s, X_i) + \hat{\mu}^D(1, s, X_i) - \frac{D_i}{1-\hat{\pi}(s)}\right].
\end{aligned}$$

Theorem 5.1. *(i) Suppose Assumptions 1 and 2 hold, then*

$$\sqrt{n}(\hat{\tau} - \tau) \rightsquigarrow \mathcal{N}(0, \sigma^2) \quad \textit{and} \quad \hat{\sigma}^2 \xrightarrow{p} \sigma^2.$$

(ii) In addition, if the working models are correctly specified, i.e., $\overline{\mu}^b(a,s,x) = \mu^b(a,s,x)$ *for all* $(a,b,s,x) \in \{0,1\}\times\{D,Y\}\times\mathcal{SX}$*, where* $\mathcal{SX}$ *is the joint support of* (S,X)*, then the asymptotic variance* σ^2 *achieves the semiparametric efficiency bound.*

Theorem 5.1(i) establishes limit distribution of our adjusted LATE estimator and gives a consistent estimator of its asymptotic variance. Such a variance depends on the working model $\overline{\mu}^b(a,s,x)$ for $(a,b) \in \{0,1\}\times\{D,Y\}$. Theorem 5.1(ii) further shows our general regression-adjusted estimator achieves the semiparametric efficiency bound $\underline{\sigma}^2$ derived in Theorem 4.1 when the working model is correctly specified.

Second, when there are no adjustments so that $\overline{\mu}^Y(\cdot)$ and $\overline{\mu}^D(\cdot)$ are just zero, we have

$$\sigma^2 = \frac{\sum_{s\in S}\frac{p(s)}{\pi(s)}Var(W - \tau D(1)|S=s) + \sum_{s\in S}\frac{p(s)}{1-\pi(s)}Var(Z - \tau D(0)|S=s) + \sigma_2^2}{\mathbb{P}(D(1) > D(0))^2}.$$

In this case, our estimator coincides with Bugni and Gao's (2021) fully saturated estimator. Indeed, we can verify, by some tedious calculation, that σ^2 defined above is the same as the asymptotic variance of the fully saturated estimator derived by Bugni and Gao (2021).[2] Then Bugni and Gao (2021) have shown that our estimator without adjustments is weakly more efficient than the strata fixed effects and two-sample IV estimators. In the next section, we show that, with adjustments, we can further improve the efficiency, even when the working models are potentially misspecified.

6 Parametric Adjustments

In this section, we consider estimating $\overline{\mu}^b(a,s,x)$ for $a = 0,1$, $s \in \mathcal{S}$, and $b = D, Y$ via parametric regressions. Note that we do not require $\overline{\mu}^b(a,s,x)$ to be correctly specified. Suppose

$$\overline{\mu}^Y(a,S_i,X_i) = \sum_{s\in\mathcal{S}} 1\{S_i = s\}\Lambda^Y_{a,s}(X_i,\theta_{a,s}) \quad \text{and} \quad \overline{\mu}^D(a,S_i,X_i) = \sum_{s\in\mathcal{S}} 1\{S_i = s\}\Lambda^D_{a,s}(X_i,\beta_{a,s}), \tag{6.1}$$

where $\Lambda^b_{a,s}(\cdot)$ for $(a,b,s) \in \{0,1\} \times \{D,Y\} \times \mathcal{S}$ is a known function of X_i up to some finite-dimensional parameter (i.e., $\theta_{a,s}$ and $\beta_{a,s}$). The researchers have the freedom to choose the functional forms of $\Lambda^b_{a,s}(\cdot)$, the parameter values of $(\theta_{a,s},\beta_{a,s})$, and the ways they are estimated. In fact, as the parametric models are potentially misspecified, different estimation methods of the same model can lead to distinctive pseudo true values. We will discuss several detailed examples in Sections 6.1, 6.2, and 6.3 below. Here, we first focus on the general setup.

Define the estimators of $(\theta_{a,s},\beta_{a,s})$ as $(\hat{\theta}_{a,s},\hat{\beta}_{a,s})$, and hence the corresponding feasible parametric regression adjustments as

$$\hat{\mu}^Y(a,s,X_i) = \Lambda^Y_{a,s}(X_i,\hat{\theta}_{a,s}) \quad \text{and} \quad \hat{\mu}^D(a,s,X_i) = \Lambda^D_{a,s}(X_i,\hat{\beta}_{a,s}). \tag{6.2}$$

Assumption 3. *(i) Suppose* $\max_{a=0,1,s\in\mathcal{S}} ||\hat{\theta}_{a,s}-\theta_{a,s}||_2 \xrightarrow{p} 0$ *and* $\max_{a=0,1,s\in\mathcal{S}} ||\hat{\beta}_{a,s}-\beta_{a,s}||_2 \xrightarrow{p} 0$.

(ii) There exist a positive random variable L_i and a positive constant $C > 0$ such that for all $a = 0,1$ and $s \in \mathcal{S}$,

$$\left\|\frac{\partial \Lambda^Y_{a,s}(X_i,\theta_{a,s})}{\partial \theta_{a,s}}\right\|_2 \le L_i, \quad ||\Lambda^Y_{a,s}(X_i,\theta_{a,s})||_2 \le L_i$$
$$\left\|\frac{\partial \Lambda^D_{a,s}(X_i,\beta_{a,s})}{\partial \beta_{a,s}}\right\|_2 \le L_i, \quad ||\Lambda^D_{a,s}(X_i,\beta_{a,s})||_2 \le L_i,$$

almost surely and $\mathbb{E}(L_i^q|S_i = s) \le C$ for some $q > 2$.

[2]Derivation is available upon request.

Assumption 3(i) means that our estimators are consistent. Assumption 3(ii) means that the parametric models are smooth in their parameters, which is true for many widely used regression models such as linear, logit, and probit regressions. This restriction can be further relaxed to allow for non-smoothness under less intuitive entropy conditions.

Theorem 6.1. *Suppose Assumption 3 hold. Then $\overline{\mu}^b(a,s,X_i)$ and $\hat{\mu}^b(a,s,X_i)$ defined in (6.1) and (6.2), respectively, satisfy Assumption 2.*

Theorem 6.1 generalizes the intuition in (5.4) to general parametric models. It means that Assumption 2 holds for parametric models as long as the parameters are consistently estimated.

6.1 Optimal Linear Adjustments

Suppose, for $a = 0,1$ and $s \in \mathcal{S}$, $\overline{\mu}^Y(a,s,X) = \Psi_{i,s}^\top t_{a,s}$ and $\overline{\mu}^D(a,s,X) = \Psi_{i,s}^\top b_{a,s}$, where $\Psi_{i,s} = \Psi_s(X_i)$ is a function and the functional form can vary across $s \in \mathcal{S}$. The restriction that the function $\Psi_s(\cdot)$ does not depend on $a = 0,1$ is innocuous as if it does, we can stack them up and denote $\Psi_{i,s} = (\Psi_{1,s}^\top(X_i), \Psi_{0,s}^\top(X_i))^\top$. Similarly, it is also innocuous to impose that the function $\Psi_s(\cdot)$ is the same for modeling $\overline{\mu}^Y(a,s,X)$ and $\overline{\mu}^D(a,s,X)$.

The asymptotic variance of the adjusted LATE estimator $\hat{\tau}$ is denoted as σ^2, which depends on $(\overline{\mu}^Y(a,s,X), \overline{\mu}^D(a,s,X))$, and thus, depends on $(t_{a,s}, b_{a,s})$. The following theorem characterizes the optimal linear coefficients that minimize the asymptotic variance of $\hat{\tau}$ over all possible $(t_{a,s}, b_{a,s})$. Let

$$\Theta^* := \begin{pmatrix} (\theta^*_{a,s}, \beta^*_{a,s})_{a=0,1,s\in\mathcal{S}} : \\ (\theta^*_{a,s}, \beta^*_{a,s})_{a=0,1,s\in\mathcal{S}} \in \arg\min_{(t_{a,s},b_{a,s})_{a=0,1,s\in\mathcal{S}}} \sigma^2((t_{a,s}, b_{a,s})_{a=0,1,s\in\mathcal{S}}). \end{pmatrix}$$

Assumption 4. *Suppose $\mathbb{E}(||\Psi_{i,s}||_2^q | S_i = s) \leq C < \infty$ for constants C and $q > 2$. Denote $\tilde{\Psi}_{i,s} := \Psi_{i,s} - \mathbb{E}(\Psi_{i,s}|S_i = s)$ for $s \in \mathcal{S}$. Then there exist constants $0 < c < C < \infty$ such that*

$$c < \lambda_{\min}(\mathbb{E}(\tilde{\Psi}_{i,s}\tilde{\Psi}_{i,s}^\top)) \leq \lambda_{\max}(\mathbb{E}(\tilde{\Psi}_{i,s}\tilde{\Psi}_{i,s}^\top)) \leq C,$$

where for a generic symmetric matrix A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of A, respectively.

Assumption 4 requires the regressor $\Psi_{i,s}$ does not contain the constant term. In fact, (5.2) and (5.3) imply that our estimator is numerically invariant to stratum-specific location shift because by definition,

$$\sum_{i=1}^n \left(\frac{A_i}{\hat{\pi}(S_i)} - 1\right) 1\{S_i = s\} = 0 \quad \text{and} \quad \sum_{i=1}^n \left(\frac{1-A_i}{1-\hat{\pi}(S_i)} - 1\right) 1\{S_i = s\} = 0.$$

Theorem 6.2. *Suppose Assumptions 1 and 4 hold. Then, we have*

$$\Theta^* = \left(\begin{array}{c} (\theta^*_{a,s}, \beta^*_{a,s})_{a=0,1,s\in\mathcal{S}}: \\ \sqrt{\frac{1-\pi(s)}{\pi(s)}}(\theta^*_{1,s} - \tau\beta^*_{1,s}) + \sqrt{\frac{\pi(s)}{1-\pi(s)}}(\theta^*_{0,s} - \tau\beta^*_{0,s}) \\ = \sqrt{\frac{1-\pi(s)}{\pi(s)}}(\theta^{LP}_{1,s} - \tau\beta^{LP}_{1,s}) + \sqrt{\frac{\pi(s)}{1-\pi(s)}}(\theta^{LP}_{0,s} - \tau\beta^{LP}_{0,s}). \end{array} \right),$$

where

$$\begin{aligned} \theta^{LP}_{a,s} &= [\mathbb{E}(\tilde{\Psi}_{i,s}\tilde{\Psi}^\top_{i,s}|S_i = s)]^{-1}[\mathbb{E}(\tilde{\Psi}_{i,s}Y_i(D_i(a))|S_i = s)] \\ \beta^{LP}_{a,s} &= [\mathbb{E}(\tilde{\Psi}_{i,s}\tilde{\Psi}^\top_{i,s}|S_i = s)]^{-1}[\mathbb{E}(\tilde{\Psi}_{i,s}D_i(a)|S_i = s)]. \end{aligned} \tag{6.3}$$

The optimality result in Theorem 6.2 rely on two key restrictions: (1) the regressor $\Psi_{i,s}$ is the same for treated and control units and (2) both the adjustments $\overline{\mu}^Y(a,s,X)$ and $\overline{\mu}^D(a,s,X)$ are linear. The first restriction is innocuous as we can stack up regressors for treated and control units as previously mentioned. The second restriction means that it is possible to have nonlinear adjustments that are more efficient. We will come back to this point in Sections 6.2, 6.3, and 7.

In view of Theorem 6.1, the optimal linear coefficients are not unique. In order to achieve the optimality, we only need to consistently estimate one point in Θ^*. For the rest of the section, we choose $(\theta^{LP}_{a,s}, \beta^{LP}_{a,s})$ with the corresponding optimal linear adjustments (i.e., linear probability model)

$$\overline{\mu}^Y(a,s,X_i) = \Psi^\top_{i,s}\theta^{LP}_{a,s} \quad \text{and} \quad \overline{\mu}^D(a,s,X_i) = \Psi^\top_{i,s}\beta^{LP}_{a,s}. \tag{6.4}$$

We estimate $(\theta^{LP}_{a,s}, \beta^{LP}_{a,s})$ by $(\hat{\theta}^{LP}_{a,s}, \hat{\beta}^{LP}_{a,s})$, where

$$\begin{aligned} \dot{\Psi}_{i,a,s} &:= \Psi_{i,s} - \frac{1}{n_a(s)} \sum_{i\in I_a(s)} \Psi_{i,s} \\ \hat{\theta}^{LP}_{a,s} &:= \left(\frac{1}{n_a(s)} \sum_{i\in I_a(s)} \dot{\Psi}_{i,a,s}\dot{\Psi}^\top_{i,a,s}\right)^{-1} \left(\frac{1}{n_a(s)} \sum_{i\in I_a(s)} \dot{\Psi}_{i,a,s}Y_i\right) \\ \hat{\beta}^{LP}_{a,s} &:= \left(\frac{1}{n_a(s)} \sum_{i\in I_a(s)} \dot{\Psi}_{i,a,s}\dot{\Psi}^\top_{i,a,s}\right)^{-1} \left(\frac{1}{n_a(s)} \sum_{i\in I_a(s)} \dot{\Psi}_{i,a,s}D_i\right). \end{aligned} \tag{6.5}$$

Then, the feasible linear adjustments can be defined as

$$\hat{\mu}^Y(a,s,X_i) = \Psi^\top_{i,s}\hat{\theta}^{LP}_{a,s} \quad \text{and} \quad \hat{\mu}^D(a,s,X_i) = \Psi^\top_{i,s}\hat{\beta}^{LP}_{a,s}. \tag{6.6}$$

Suppose $\mathcal{S} = \{1, \cdots, S\}$ for some integer $S > 0$. It is clear that $\hat{\theta}^{LP}_{a,s}$ and $\hat{\beta}^{LP}_{a,s}$ are the estimated slopes of the following four cluster-specific regressions using only the s cluster:

$$(1-A_i)Y_i = (1-A_i)(\gamma^Y_{0,s} + \Psi_{i,s}^\top \theta_{0,s} + e^Y_{0,i}), \quad A_iY_i = A_i(\gamma^Y_{1,s} + \Psi_{i,s}^\top \theta_{1,s} + e^Y_{1,i}) \tag{6.7}$$
$$(1-A_i)D_i = (1-A_i)(\gamma^D_{0,s} + \Psi_{i,s}^\top \beta_{0,s} + e^D_{0,i}), \quad A_iD_i = A_i(\gamma^D_{1,s} + \Psi_{i,s}^\top \beta_{1,s} + e^D_{1,i}).$$

Theorem 6.3. *Suppose Assumptions 1 and 4 holds. Then,*

$$\{\overline{\mu}^b(a,s,X_i)\}_{b=D,Y,a=0,1,s\in\mathcal{S}} \quad \text{and} \quad \{\hat{\mu}^b(a,s,X_i)\}_{b=D,Y,a=0,1,s\in\mathcal{S}}$$

defined in (6.4) and (6.6), respectively, satisfy Assumption 2. Denote the adjusted LATE estimator with adjustment $\{\overline{\mu}^b(a,s,X_i)\}_{b=D,Y,a=0,1,s\in\mathcal{S}}$ defined in (6.6) as $\hat{\tau}^{LP}$. Then, all the results in Theorem 5.1(i) hold for $\hat{\tau}^{LP}$. In addition, $\hat{\tau}^{LP}$ is the most efficient among all linearly adjusted LATE estimators, and in particular, weakly more efficient than the LATE estimator with no adjustments.

The asymptotic variance of the LATE estimator with the optimal linear adjustments ($\hat{\tau}^{LP}$) takes the form of (5.5) with $\{\overline{\mu}^b(a,s,X_i)\}_{b=D,Y,a=0,1,s\in\mathcal{S}}$ in (4.2)–(4.4) defined in (6.4). When $\Psi_{i,s} = X_i$, such an asymptotic variance is the same as that for Ansel et al.'s (2018) $\hat{\tau}_S$ defined in Section C in the Online Supplement. This fact implies $\hat{\tau}_S$ is the most efficient LATE estimator adjusted by linear functions of X_i. Consequently, both $\hat{\tau}_S$ and $\hat{\tau}^{LP}$ are weakly more efficient than $\hat{\tau}_{tsls}$.

6.2 Linear and Logistic Regressions

It is also possible to consider a linear model for $\overline{\mu}^Y(a,s,X_i)$ and a logistic model for $\overline{\mu}^D(a,s,X_i)$, i.e.,

$$\overline{\mu}^Y(a,s,X_i) = \mathring{\Psi}_{i,s}^\top t_{a,s} \quad \text{and} \quad \overline{\mu}^D(a,s,X_i) = \lambda(\mathring{\Psi}_{i,s}^\top b_{a,s}),$$

where $\mathring{\Psi}_{i,s} = (1, \Psi_{i,s})$, $\Psi_{i,s} = \Psi_s(X_i)$ and $\lambda(u) = \exp(u)/(1+\exp(u))$ is the logistic CDF. As the model for $\overline{\mu}^D(a,s,X_i)$ is non-linear, the optimality result established in the previous section does not apply. We can consider fitting the linear and logistic models by OLS and MLE, respectively, and call this method the OLS-MLE adjustment. Specifically, define

$$\hat{\mu}^Y(a,s,X_i) = \mathring{\Psi}_{i,s}^\top \hat{\theta}^{OLS}_{a,s} \quad \text{and} \quad \hat{\mu}^D(a,s,X_i) = \lambda(\mathring{\Psi}_{i,s}^\top \hat{\beta}^{MLE}_{a,s}), \tag{6.8}$$

where

$$\hat{\theta}^{OLS}_{a,s} = \left(\frac{1}{n_a(s)} \sum_{i\in I_a(s)} \mathring{\Psi}_{i,s}\mathring{\Psi}_{i,s}^\top\right)^{-1} \left(\frac{1}{n_a(s)} \sum_{i\in I_a(s)} \mathring{\Psi}_{i,s} Y_i\right) \quad \text{and}$$
$$\hat{\beta}^{MLE}_{a,s} = \arg\max_b \frac{1}{n_a(s)} \sum_{i\in I_a(s)} \left[D_i \log(\lambda(\mathring{\Psi}_{i,s}^\top b)) + (1-D_i)\log(1-\lambda(\mathring{\Psi}_{i,s}^\top b))\right]. \tag{6.9}$$

It is clear that $\hat{\theta}^{OLS}_{a,s}$ and $\hat{\beta}^{MLE}_{a,s}$ are the OLS and MLE estimates of the following four cluster specific (logistic) regressions using only the s cluster:

$$(1-A_i)Y_i = (1-A_i)(\mathring{\Psi}_{i,s}^\top \theta_{0,s} + e_{0,i}^Y), \quad A_iY_i = A_i(\mathring{\Psi}_{i,s}^\top \theta_{1,s} + e_{1,i}^Y)$$
$$(1-A_i)D_i = (1-A_i)1\{\mathring{\Psi}_{i,s}^\top \beta_{0,s} \geq e_{0,i}^D\}, \quad A_iD_i = A_i 1\{\mathring{\Psi}_{i,s}^\top \beta_{1,s} \geq e_{1,i}^D\},$$

where $e_{0,i}^D$ and $e_{1,i}^D$ are logistic errors.[3]

In the OLS-MLE adjustment, we do allow the regressor $\mathring{\Psi}_{i,s}$ to contain the constant term. Suppose $\hat{\theta}_{a,s}^{OLS} = (\hat{h}_{a,s}^{OLS}, \underline{\hat{\theta}}_{a,s}^{OLS,\top})^\top$, where $\hat{h}_{a,s}^{OLS}$ are the coefficients of the constant terms in $\mathring{\Psi}_{i,s}$. Then, because our adjusted LATE estimator is invariant to the stratum-specific location shift of the adjustment term, using $\hat{\mu}^Y(a,s,X_i) = \mathring{\Psi}_{i,s}^\top \hat{\theta}_{a,s}^{OLS} = \hat{h}_{a,s}^{OLS} + \Psi_{i,s}^\top \underline{\hat{\theta}}_{a,s}^{OLS}$ and $\hat{\mu}^Y(a,s,X_i) = \Psi_{i,s}^\top \underline{\hat{\theta}}_{a,s}^{OLS}$ produce the exact same LATE estimator. In addition, we can obtain $\underline{\hat{\theta}}_{a,s}^{OLS}$ from the estimated slopes of the following cluster specific regressions using only the s cluster

$$(1-A_i)Y_i = (1-A_i)(\gamma_{0,s}^Y + \Psi_{i,s}^\top \theta_{0,s} + e_{0,i}^Y), \qquad A_iY_i = A_i(\gamma_{1,s}^Y + \Psi_{i,s}^\top \theta_{1,s} + e_{1,i}^Y)$$

which is exactly the same as (6.7). This implies $\underline{\hat{\theta}}_{a,s}^{OLS} = \hat{\theta}_{a,s}^{LP}$. On the other hand, because the logistic regression is nonlinear, the non-intercept part of $\hat{\beta}_{a,s}^{MLE}$ does not equal $\hat{\beta}_{a,s}^{LP}$.

The limits of $\hat{\theta}_{a,s}^{OLS}$ and $\hat{\beta}_{a,s}^{MLE}$ are defined as

$$\theta_{a,s}^{OLS} = \left(\mathbb{E}(\mathring{\Psi}_{i,s}\mathring{\Psi}_{i,s}^\top | S_i = s)\right)^{-1} \left(\mathbb{E}(\mathring{\Psi}_{i,s} Y_i(D_i(a)) | S_i = s)\right) \quad \text{and}$$
$$\beta_{a,s}^{MLE} = \arg\max_b \mathbb{E}\left(\left[D_i(a)\log(\lambda(\mathring{\Psi}_{i,s}^\top b)) + (1-D_i(a))\log(1-\lambda(\mathring{\Psi}_{i,s}^\top b))\right] | S_i = s\right),$$

which imply that

$$\overline{\mu}^Y(a,s,X_i) = \mathring{\Psi}_{i,s}^\top \theta_{a,s}^{OLS} \quad \text{and} \quad \overline{\mu}^D(a,s,X_i) = \lambda(\mathring{\Psi}_{i,s}^\top \beta_{a,s}^{MLE}). \tag{6.10}$$

Assumption 5. *(i) For $a = 0,1$ and $s \in \mathcal{S}$, suppose $\mathbb{E}(\mathring{\Psi}_{i,s}\mathring{\Psi}_{i,s}^\top | S_i = s)$ is invertible and*

$$\mathbb{E}\left(\left[D_i(a)\log(\lambda(\mathring{\Psi}_{i,s}^\top b)) + (1-D_i(a))\log(1-\lambda(\mathring{\Psi}_{i,s}^\top b))\right] | S_i = s\right)$$

has $\beta_{a,s}^{MLE}$ as its unique maximizer.

(ii) There exists a constant $C < \infty$ such that $\max_{a=0,1,s\in\mathcal{S}} \mathbb{E}||\mathring{\Psi}_{i,s}||_2^q \leq C < \infty$ for some $q > 2$.

Theorem 6.4. *Suppose Assumptions 1 and 5 hold. Then,*

$$\{\overline{\mu}^b(a,s,X_i)\}_{b=D,Y,a=0,1,s\in\mathcal{S}} \quad \textit{and} \quad \{\hat{\mu}^b(a,s,X_i)\}_{b=D,Y,a=0,1,s\in\mathcal{S}}$$

defined in (6.10) and (6.8), respectively, satisfy Assumption 2. Denote the adjusted LATE estimator with adjustment $\{\hat{\mu}^b(a,s,X_i)\}_{b=D,Y,a=0,1,s\in\mathcal{S}}$ defined in (6.8) as $\hat{\tau}^{LG}$. Then, all the results in

[3]Note we do not need the regression models to be correctly specified.

Theorem 5.1(i) hold for $\hat{\tau}^{LG}$.

Several remarks are in order. First, the OLS-MLE adjustment is not optimal in the sense that it does not necessarily minimize the asymptotic variance of the corresponding LATE estimator. Second, the OLS-MLE adjustment is not necessarily less efficient than the optimal linear adjustment studied in Section 6.1 as $\mu^D(a, s, X_i)$ could be nonlinear. In fact, as Theorem 5.1 shows, if the adjustments are correctly specified, then the adjusted LATE estimator can achieve the semiparametric efficiency bound. Compared with the linear probability model considered in Section 6.1, the logistic model is expected to be less misspecified, especially when the regressor Ψ_i contains technical terms of X_i such as interactions and quadratic terms. Third, we will further justify the above intuition in Section 7 below, in which we let $\Psi_{i,s}$ be the sieve basis functions with an increasing dimension and show that the OLS-MLE method can consistently estimate the correct specification under some regularity conditions. Fourth, one theoretical shortcoming of the OLS-MLE adjustment is that, unlike the optimal linear adjustment, it is not guaranteed to be more efficient than no adjustment. We address this issue in Section 6.3 below.

6.3 Further Efficiency Improvement

Let $\theta_{a,s}^{OLS} = (h_{a,s}^{OLS}, \underline{\theta}_{a,s}^{OLS})$, where $h_{a,s}^{OLS}$ is the intercept. If $\beta_{a,s}^{MLE}$ were known, the OLS-MLE adjustment can be viewed as a linear adjustment. Specifically, denote

$$\Phi_{i,s} := (\Psi_{i,s}^\top, \lambda(\mathring{\Psi}_{i,s}^\top \beta_{1,s}^{MLE}), \lambda(\mathring{\Psi}_{i,s}^\top \beta_{0,s}^{MLE}))^\top \tag{6.11}$$

$$t_{a,s}^{LG} := a \begin{pmatrix} \underline{\theta}_{1,s}^{OLS} \\ 0 \\ 0 \end{pmatrix} + (1-a) \begin{pmatrix} \underline{\theta}_{0,s}^{OLS} \\ 0 \\ 0 \end{pmatrix}, \quad b_{a,s}^{LG} := a \begin{pmatrix} 0_{d_\Psi} \\ 1 \\ 0 \end{pmatrix} + (1-a) \begin{pmatrix} 0_{d_\Psi} \\ 0 \\ 1 \end{pmatrix},$$

where d_Ψ is the dimension of $\Psi_{i,s}$. Then, the OLS-MLE adjustment can be written as

$$\overline{\mu}^Y(a, s, X_i) = \Phi_{i,s}^\top t_{a,s}^{LG} + h_{a,s}^{OLS} \quad \text{and} \quad \overline{\mu}^D(a, s, X_i) = \Phi_{i,s}^\top b_{a,s}^{LG}.$$

Because our estimator is invariant to stratum-level location shift of adjustments, the OLS-MLE adjustments and the linear adjustments

$$\overline{\mu}^Y(a, s, X_i) = \Phi_{i,s}^\top t_{a,s}^{LG} \quad \text{and} \quad \overline{\mu}^D(a, s, X_i) = \Phi_{i,s}^\top b_{a,s}^{LG}$$

produce the same estimator. Similarly, we can replicate no adjustments and the optimal linear adjustments with $\Phi_{i,s}$ defined in (6.11) as regressors by letting

$$\overline{\mu}^Y(a, s, X_i) = \Phi_{i,s}^\top t_{a,s} \quad \text{and} \quad \overline{\mu}^D(a, s, X_i) = \Phi_{i,s}^\top b_{a,s}$$

with $(t_{a,s}, b_{a,s}) = 0$ and $(t_{a,s}, b_{a,s}) = (t_{a,s}^{LP}, b_{a,s}^{LP})$, respectively, where

$$t_{a,s}^{LP} := a \begin{pmatrix} \theta_{1,s}^{LP} \\ 0 \\ 0 \end{pmatrix} + (1-a) \begin{pmatrix} \theta_{0,s}^{LP} \\ 0 \\ 0 \end{pmatrix}, \quad b_{a,s}^{LP} := a \begin{pmatrix} \beta_{1,s}^{LP} \\ 0 \\ 0 \end{pmatrix} + (1-a) \begin{pmatrix} \beta_{0,s}^{LP} \\ 0 \\ 0 \end{pmatrix}.$$

Based on Theorem 6.2, we can further improve all three types of adjustments by setting the linear coefficients of $\Phi_{i,s}$ as

$$\theta_{a,s}^{F} := \left(\mathbb{E}[\tilde{\Phi}_{i,s}\tilde{\Phi}_{i,s}^{\top}|S_i = s]\right)^{-1} \left([\mathbb{E}\tilde{\Phi}_{i,s}Y_i(D_i(a))|S_i = s]\right),$$
$$\beta_{a,s}^{F} := \left(\mathbb{E}[\tilde{\Phi}_{i,s}\tilde{\Phi}_{i,s}^{\top}|S_i = s]\right)^{-1} \left([\mathbb{E}\tilde{\Phi}_{i,s}D_i(a)|S_i = s]\right),$$

where $\tilde{\Phi}_{i,s} = \Phi_{i,s} - \mathbb{E}(\Phi_{i,s}|S_i = s)$. The final linear adjustments with $\theta_{a,s}^{F}$ and $\beta_{a,s}^{F}$ are

$$\overline{\mu}^{Y}(a, s, X_i) = \Phi_{i,s}^{\top}\theta_{a,s}^{F} \quad \text{and} \quad \overline{\mu}^{D}(a, s, X_i) = \Phi_{i,s}^{\top}\beta_{a,s}^{F}. \tag{6.12}$$

Because $\beta_{a,s}^{MLE}$ is unknown, we can replace it by its estimate proposed in Section 6.2, i.e., define

$$\hat{\Phi}_{i,s} := (\Psi_{i,s}, \lambda(\mathring{\Psi}_{i,s}^{\top}\hat{\beta}_{1,s}^{MLE}), \lambda(\mathring{\Psi}_{i,s}^{\top}\hat{\beta}_{0,s}^{MLE}))^{\top} \quad \text{and} \quad \breve{\Phi}_{i,a,s} := \hat{\Phi}_{i,s} - \frac{1}{n_a(s)}\sum_{i \in I_a(s)} \hat{\Phi}_{i,s}.$$

Then, we define the estimators of $\theta_{a,s}^{F}$ and $\beta_{a,s}^{F}$ as

$$\hat{\theta}_{a,s}^{F} := \left(\frac{1}{n_a(s)}\sum_{i \in I_a(s)} \breve{\Phi}_{i,a,s}\breve{\Phi}_{i,a,s}^{\top}\right)^{-1} \left(\frac{1}{n_a(s)}\sum_{i \in I_a(s)} \breve{\Phi}_{i,a,s}Y_i\right),$$
$$\hat{\beta}_{a,s}^{F} := \left(\frac{1}{n_a(s)}\sum_{i \in I_a(s)} \breve{\Phi}_{i,a,s}\breve{\Phi}_{i,a,s}^{\top}\right)^{-1} \left(\frac{1}{n_a(s)}\sum_{i \in I_a(s)} \breve{\Phi}_{i,a,s}D_i\right). \tag{6.13}$$

The corresponding feasible adjustments are

$$\hat{\mu}^{Y}(a, s, X_i) = \hat{\Phi}_{i,s}^{\top}\hat{\theta}_{a,s}^{F} \quad \text{and} \quad \hat{\mu}^{D}(a, s, X_i) = \hat{\Phi}_{i,s}^{\top}\hat{\beta}_{a,s}^{F}. \tag{6.14}$$

Assumption 6. *Suppose Assumption 4 holds for $\Phi_{i,s}$ defined in (6.11).*

Theorem 6.5. *Suppose Assumptions 1, 5, and 6 hold. Then,*

$$\{\overline{\mu}^{b}(a, s, X_i)\}_{b=D,Y,a=0,1,s\in\mathcal{S}} \quad \textit{and} \quad \{\hat{\mu}^{b}(a, s, X_i)\}_{b=D,Y,a=0,1,s\in\mathcal{S}}$$

defined in (6.12) and (6.14), respectively, satisfy Assumption 2. Denote the LATE estimator with regression adjustments $\{\hat{\mu}^{b}(a, s, X_i)\}_{b=D,Y,a=0,1,s\in\mathcal{S}}$ defined in (6.14) as $\hat{\tau}^{F}$. Then, all the results in Theorem 5.1(i) holds for $\hat{\tau}^{F}$. In addition, $\hat{\tau}^{F}$ is weakly more efficient than $\hat{\tau}^{LP}$, $\hat{\tau}^{LG}$, and the LATE estimator with no adjustments.

Theorem 6.5 shows that by refitting OLS-MLE adjustment in a linear regression with optimal linear coefficients, we can further improve the efficiency of the adjusted LATE estimator. As a byproduct, $\hat{\tau}^F$ is guaranteed to be weakly more efficient than the LATE estimator without any adjustments.

7 Nonparametric Adjustments

In this section, we consider the nonparametric regression as the adjustments for our LATE estimator. Specifically, we use linear and logistic sieve regressions to estimate the true specifications $\mu^Y(a,s,X_i)$ and $\mu^D(a,s,X_i)$, respectively. For implementation, the nonparametric adjustment is exactly the same as OLS-MLE adjustment studied in Section 6.2. Theoretically, we will let the regressors $\mathring{\Psi}_{i,s}$ in (6.8) be sieve basis functions whose dimensions will diverge to infinity as sample size increases. For notation simplicity, we suppress the subscript s and denote the sieve regressors as $\mathring{\Psi}_{i,n} \in \Re^{h_n}$, where the dimension h_n can diverge with the sample size. The corresponding feasible regression adjustments are

$$\hat{\mu}^Y(a,s,X_i) = \mathring{\Psi}_{i,n}^\top \hat{\theta}_{a,s}^{NP} \quad \text{and} \quad \hat{\mu}^D(a,s,X_i) = \lambda(\mathring{\Psi}_{i,n}^\top \hat{\beta}_{a,s}^{NP}), \tag{7.1}$$

where

$$\hat{\theta}_{a,s}^{NP} = \left(\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \mathring{\Psi}_{i,n} \mathring{\Psi}_{i,n}^\top\right)^{-1} \left(\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \mathring{\Psi}_{i,n} Y_i\right) \quad \text{and}$$
$$\hat{\beta}_{a,s}^{NP} = \arg\max_b \frac{1}{n_a(s)} \sum_{i \in I_a(s)} \left[D_i \log(\lambda(\mathring{\Psi}_{i,n}^\top b)) + (1-D_i)\log(1-\lambda(\mathring{\Psi}_{i,n}^\top b))\right].$$

We finally denote the corresponding adjusted LATE estimator as $\hat{\tau}^{NP}$.

Assumption 7. *(i) There exist constants $0 < c < C < \infty$ such that with probability approaching one,*

$$c \le \lambda_{\min}\left(\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \mathring{\Psi}_{i,n} \mathring{\Psi}_{i,n}^\top\right) \le \lambda_{\max}\left(\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \mathring{\Psi}_{i,n} \mathring{\Psi}_{i,n}^\top\right) \le C$$

and

$$c \le \lambda_{\min}\left(\mathbb{E}[\mathring{\Psi}_{i,n} \mathring{\Psi}_{i,n}^\top | S_i = s]\right) \le \lambda_{\max}\left(\mathbb{E}[\mathring{\Psi}_{i,n} \mathring{\Psi}_{i,n}^\top | S_i = s]\right) \le C.$$

(ii) For $a = 0, 1$, there exist $h_n \times 1$ vectors $\theta_{a,s}^{NP}$ and $\beta_{a,s}^{NP}$ such that for

$$R^Y(a,s,x) := \mathbb{E}\left[Y_i(D_i(a)) | S_i = s, X_i = x\right] - \mathring{\Psi}_{i,n}^\top \theta_{a,s}^{NP} \quad \text{and}$$
$$R^D(a,s,x) := \mathbb{P}\left(D_i(a) = 1 | S_i = s, X_i = x\right) - \lambda(\mathring{\Psi}_{i,n}^\top \beta_{a,s}^{NP}),$$

we have $\sup_{a=0,1,b\in\{D,Y\},s\in\mathcal{S},x\in Supp(X)} |R^b(a,s,x)| = o_p(1)$,

$$\sup_{a=0,1,b\in\{D,Y\},s\in\mathcal{S},x\in Supp(X)} \frac{1}{n_a(s)} \sum_{i\in I_a(s)} \left(R^b(a,s,X_i)\right)^2 = O_p\left(\frac{h_n \log n}{n}\right),$$

and

$$\sup_{a=0,1,b\in\{D,Y\},s\in\mathcal{S}} \mathbb{E}\left[\left(R^b(a,s,X_i)\right)^2 |S_i = s\right] = O\left(\frac{h_n \log n}{n}\right).$$

(iii) For $a = 0,1$, *there exists a constant* $c \in (0, 0.5)$ *such that*

$$\begin{aligned} c \le & \inf_{a=0,1,s\in\mathcal{S},x\in Supp(X)} \mathbb{P}\left(D_i(a) = 1|S_i = s, X_i = x\right) \\ \le & \sup_{a=0,1,s\in\mathcal{S},x\in Supp(X)} \mathbb{P}\left(D_i(a) = 1|S_i = s, X_i = x\right) \le 1 - c. \end{aligned}$$

(iv) Suppose $\mathbb{E}[\mathring{\Psi}^2_{i,n,k}|S_i = s] \le C < \infty$ *for some constant* $C > 0$, *where* $\mathring{\Psi}_{i,n,k}$ *denotes the kth coordinate of* $\mathring{\Psi}_{i,n}$. $\max_{i\in[n]} ||\mathring{\Psi}_{i,n}||_2 \le \zeta(h_n)$ *a.s., where* $\zeta(\cdot)$ *is a deterministic increasing function satisfying* $\zeta^2(h_n)h_n \log n = o(n)$. *Also* $h_n^2 \log^2 n = o(n)$.

Assumption 7 is standard for linear and logistic sieve regressions. We refer to Hirano, Imbens, and Ridder (2003) and Chen (2007) for more discussions. The quantity $\zeta(h_n)$ in Assumption 7(iv) depends on the choice of basis functions. For example, $\zeta(h_n) = O(h_n^{1/2})$ for splines and $\zeta(h_n) = O(h_n)$ for power series.

Theorem 7.1. *Suppose Assumptions 1 and 7 hold. Then* $\{\hat{\mu}^b(a,s,X_i)\}_{b=D,Y,a=0,1,s\in\mathcal{S}}$ *defined in* (7.1) *with* $\overline{\mu}^b(a,s,X) = \mu^b(a,s,X)$ *satisfy Assumption 2. Then, all the results in Theorem 5.1(i) hold for* $\hat{\tau}^{NP}$. *In addition,* $\hat{\tau}^{NP}$ *achieves the minimum asymptotic variance characterized in Theorem 5.1(ii).*

The OLS-MLE and nonparametric adjustments are numerically identical if the same set of regressors are used. Theorem 7.1 then shows that the OLS-MLE adjustment with technical regressors performs well because it can closely approximate the correct specification. Under the asymptotic framework that the dimension of the regressors diverges to infinity and the approximation error converges to zero, the OLS-MLE adjustment can be viewed as the nonparametric adjustment, which achieves the minimum asymptotic variance of the adjusted LATE estimator.

8 Regularized Adjustments

In this section, we consider the case that the regressor $\mathring{\Psi}_{i,n} \in \Re^{p_n}$ in which we allow p_n to be much higher than n. In this case, we can no longer use the OLS-MLE (nonparametric) adjustment method. Instead, we need to regularize the least squares and logistic regressions. Specifically, let

$$\hat{\mu}^Y(a,s,X_i) = \mathring{\Psi}_{i,n}^\top \hat{\theta}^R_{a,s} \quad \text{and} \quad \hat{\mu}^D(a,s,X_i) = \lambda(\mathring{\Psi}_{i,n}^\top \hat{\beta}^R_{a,s}), \tag{8.1}$$

and the corresponding adjusted LATE estimator is denoted as $\hat{\tau}^R$, where

$$\hat{\theta}^R_{a,s} = \arg\min_t \frac{-1}{n_a(s)} \sum_{i \in I_a(s)} \left(Y_i - \mathring{\Psi}_{i,n}^\top t\right)^2 + \frac{\varrho_{n,a}(s)}{n_a(s)} ||\hat{\Omega}^Y t||_1,$$

$$\hat{\beta}^R_{a,s} = \arg\min_b \frac{-1}{n_a(s)} \sum_{i \in I_a(s)} \left[D_i \log\left(\lambda(\mathring{\Psi}_{i,n}^\top b)\right) + (1-D_i)\log(1-\lambda\left(\mathring{\Psi}_{i,n}^\top b\right))\right] + \frac{\varrho_{n,a}(s)}{n_a(s)} ||\hat{\Omega}^D b||_1,$$

$\{\varrho_{n,a}(s)\}_{a=0,1,s\in\mathcal{S}}$ are tuning parameters, and $\hat{\Omega}^b = \text{diag}(\hat{\omega}_1^b, \cdots, \hat{\omega}_{p_n}^b)$ is a diagonal matrix of data-dependent penalty loadings for $b = D, Y$. We provide more detail about $\hat{\Omega}^b$ in Section A.

We maintain the following assumptions for Lasso and logistic Lasso regressions.

Assumption 8. *(i) For $a = 0, 1$. Suppose*

$$\mathbb{E}\left[Y_i(D_i(a))|X_i, S_i = s\right] = \mathring{\Psi}_{i,n}^\top \theta^R_{a,s} + R^Y(a,s,X_i) \quad \textit{and}$$
$$\mathbb{P}(D_i(a) = 1|X_i, S_i = s) = \lambda(\mathring{\Psi}_{i,n}^\top \beta^R_{a,s}) + R^D(a,s,X_i)$$

such that $\max_{a=0,1,s\in\mathcal{S}} \max(||\theta^R_{a,s}||_0, ||\beta^R_{a,s}||_0) \leq h_n$, where $||a||_0$ denotes the number of nonzero components in a.

(ii) Suppose $\sup_{i\in[n]} ||\mathring{\Psi}_{i,n}||_\infty \leq \zeta_n$ a.s. and $\sup_{h\in[p_n]} \mathbb{E}\left[|\mathring{\Psi}^q_{i,n,h}||S_i = s\right] < \infty$ for $q > 2$.

(iii) Suppose

$$\max_{a=0,1,b=D,Y,s\in\mathcal{S}} \frac{1}{n_a(s)} \sum_{i\in I_a(s)} (R^b(a,s,X_i))^2 = O_p(h_n \log p_n/n),$$

$$\max_{a=0,1,b=D,Y,s\in\mathcal{S}} \mathbb{E}\left[(R^b(a,s,X_i))^2|S_i = s\right] = O(h_n \log p_n/n),$$

and

$$\sup_{a=0,1,b=D,Y,s\in\mathcal{S},x\in\mathcal{X}} |R^b(a,s,X)| = O(\sqrt{\zeta_n^2 h_n^2 \log p_n/n}).$$

(iv) Suppose $\frac{\log(p_n)\zeta_n^2 h_n^2}{n} \to 0$ and $\frac{\log^2(p_n)\log^2(n)h_n^2}{n} \to 0$.

(v) There exists a constant $c \in (0, 0.5)$ such that

$$c \leq \inf_{a=0,1,s\in\mathcal{S},x\in Supp(X)} \mathbb{P}(D_i(a) = 1|S_i = s, X_i = x)$$
$$\leq \sup_{a=0,1,s\in\mathcal{S},x\in Supp(X)} \mathbb{P}(D_i(a) = 1|S_i = s, X_i = x) \leq 1 - c.$$

(vi) Let ℓ_n be a sequence that diverges to infinity. Then there exist two constants κ_1 and κ_2 such

that with probability approaching one,

$$0 < \kappa_1 \leq \inf_{a=0,1,s\in\mathcal{S},||v||_0\leq h_n\ell_n} \frac{v^\top \left(\frac{1}{n_a(s)}\sum_{i\in I_a(s)} \mathring{\Psi}_{i,n}\mathring{\Psi}_{i,n}^\top\right) v}{||v||_2^2}$$
$$\leq \sup_{a=0,1,s\in\mathcal{S},||v||_0\leq h_n\ell_n} \frac{v^\top \left(\frac{1}{n_a(s)}\sum_{i\in I_a(s)} \mathring{\Psi}_{i,n}\mathring{\Psi}_{i,n}^\top\right) v}{||v||_2^2} \leq \kappa_2 < \infty,$$

and

$$0 < \kappa_1 \leq \inf_{a=0,1,s\in\mathcal{S},||v||_0\leq h_n\ell_n} \frac{v^\top \mathbb{E}\left[\mathring{\Psi}_{i,n}\mathring{\Psi}_{i,n}^\top | S_i = s\right] v}{||v||_2^2}$$
$$\leq \sup_{a=0,1,s\in\mathcal{S},||v||_0\leq h_n\ell_n} \frac{v^\top \mathbb{E}\left[\mathring{\Psi}_{i,n}\mathring{\Psi}_{i,n}^\top | S_i = s\right] v}{||v||_2^2} \leq \kappa_2 < \infty.$$

(vii) For $a = 0, 1$, let $\varrho_{n,a}(s) = c\sqrt{n_a(s)}F_N^{-1}\left(1 - 0.1/\left[\log(n_a(s))4p_n\right]\right)$ where $F_N(\cdot)$ is the standard normal CDF and $c > 0$ is a constant.

Assumption 8 is standard in the literature and we refer interested readers to Belloni, Chernozhukov, Fernández-V (2017) for more discussion.

Theorem 8.1. *Suppose Assumptions 1 and 8 hold. Then $\{\hat{\mu}^b(a, s, X_i)\}_{b=D,Y,a=0,1,s\in\mathcal{S}}$ defined in (8.1) and $\overline{\mu}^b(a, s, X) = \mu^b(a, s, X)$ satisfy Assumption 2. Then all the results in Theorem 5.1(i) hold for $\hat{\tau}^R$. In addition, $\hat{\tau}^R$ achieves the minimum asymptotic variance characterized in Theorem 5.1(ii).*

Based on the approximate sparsity, Lasso can consistently estimate the correct specification, which then implies that the adjusted LATE estimator achieves the minimum variance because of Theorem 5.1(ii).

9 Simulations

9.1 Data Generating Processes

Three data generating processes (DGPs) are used to assess the finite sample performance of the estimation and inference methods introduced in the paper. Suppose that

$$Y_i(1) = a_1 + \alpha(X_i, Z_i) + \varepsilon_{1,i}$$
$$Y_i(0) = a_0 + \alpha(X_i, Z_i) + \varepsilon_{2,i}$$
$$D_i(0) = 1\{b_0 + \gamma(X_i, Z_i) > c_0\varepsilon_{3,i}\}$$
$$D_i(1) = \begin{cases} 1\{b_1 + \gamma(X_i, Z_i) > c_1\varepsilon_{4,i}\} & \text{if } D_i(0) = 0 \\ 1 & \text{otherwise} \end{cases}$$

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

where $\{X_i, Z_i\}_{i\in[n]}, \alpha(\cdot,\cdot)$, $\{a_i, b_i, c_i\}_{i=0,1}$ and $\{\varepsilon_{j,i}\}_{j\in[4],i\in[n]}$ are separately specified as follows.

(i) Let Z_i be i.i.d. according to standardized Beta(2, 2), $S_i = \sum_{j=1}^{4} 1\{Z_i \leq g_j\}$, and $(g_1, g_2, g_3, g_4) = (-0.25\sqrt{20}, 0, 0.25\sqrt{20}, 0.5\sqrt{20})$. $X_i := (X_{1,i}, X_{2,i})^\top$, where $X_{1,i}$ follows a uniform distribution on $[-2, 2]$, $X_{2,i} := Z_i + N(0, 1)$, and $X_{1,i}$ and $X_{2,i}$ are independent. Further define

$$\alpha(X_i, Z_i) = 0.7X_{1,i}^2 + X_{2,i} + 4Z_i$$
$$\gamma(X_i, Z_i) = 0.5X_{1,i}^2 - 0.5X_{2,i}^2 - 0.5Z_i^2,$$

$a_1 = 2, a_0 = 1, b_1 = 1.3, b_0 = -1, c_1 = c_0 = 3$, and $(\varepsilon_{1,i}, \varepsilon_{2,i}, \varepsilon_{3,i}, \varepsilon_{4,i})^\top \overset{i.i.d}{\sim} N(0, \Sigma)$, where

$$\Sigma = \begin{pmatrix} 1 & 0.5 & 0.5^2 & 0.5^3 \\ 0.5 & 1 & 0.5 & 0.5^2 \\ 0.5^2 & 0.5 & 1 & 0.5 \\ 0.5^3 & 0.5^2 & 0.5 & 1 \end{pmatrix}$$

(ii) Let Z be i.i.d. according to uniform$[-2, 2]$, $S_i = \sum_{j=1}^{4} 1\{Z_i \leq g_j\}$, and $(g_1, g_2, g_3, g_4) = (-1, 0, 1, 2)$. Let $X_i := (X_{1,i}, X_{2,i})^\top$, where $X_{1,i}$ follows a uniform distribution on $[-2, 2]$, $X_{2,i}$ follows a standard normal distribution, and $X_{1,i}$ and $X_{2,i}$ are independent. Further define

$$\alpha(X_i, Z_i) = -0.8X_{1,i} \cdot X_{2,i} + Z_i^2 + Z_i \cdot X_{1,i}$$
$$\gamma(X_i, Z_i) = 0.5X_{1,i}^2 - 0.5X_{2,i}^2 - 0.5Z_i^2,$$

$a_1 = 2, a_0 = 1, b_1 = 1, b_0 = -1, c_1 = c_0 = 3$, and $(\varepsilon_{1,i}, \varepsilon_{2,i}, \varepsilon_{3,i}, \varepsilon_{4,i})^\top$ are defined in DGP(i).

(iii) Let Z be i.i.d. according to standardized Beta(2, 2), $S_i = \sum_{j=1}^{4} 1\{Z_i \leq g_j\}$, and $(g_1, g_2, g_3, g_4) = (-0.25\sqrt{20}, 0, 0.25\sqrt{20}, 0.5\sqrt{20})$. Let $X_i := (X_{1,i}, \cdots, X_{20,i})^\top$, where $X_i \overset{i.i.d}{\sim} N(0_{20\times 1}, \Omega)$ where Ω is the Toeplitz matrix

$$\Omega = \begin{pmatrix} 1 & 0.5 & 0.5^2 & \cdots & 0.5^{19} \\ 0.5 & 1 & 0.5 & \cdots & 0.5^{18} \\ 0.5^2 & 0.5 & 1 & \cdots & 0.5^{17} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.5^{19} & 0.5^{18} & 0.5^{17} & \cdots & 1 \end{pmatrix}$$

Further define $\alpha(X_i, Z_i) = \sum_{k=1}^{20} X_{k,i}\beta_k + Z_i$, $\gamma(X_i, Z_i) = \sum_{k=1}^{20} X_{k,i}^\top \gamma_k - Z_i$, with $\beta_k = \sqrt{6}/k^2$ and $\gamma_k = -2/k^2$. Moreover, $a_1 = 2, a_0 = 1, b_1 = 2, b_0 = -1, c_1 = c_0 = \sqrt{7}$, and

$(\varepsilon_{1,i}, \varepsilon_{2,i}, \varepsilon_{3,i}, \varepsilon_{4,i})^\top$ are defined in DGP(i).

For each data generating process, we consider the following four randomization schemes as in Zhang and Zheng (2020) with $\pi(s) = 0.5$ for $s \in \mathcal{S}$:

(i) SRS: Treatment assignment is generated as in Example 1.

(ii) WEI: Treatment assignment is generated as in Example 2 with $f(x) = (1 - x)/2$.

(iii) BCD: Treatment assignment is generated as in Example 3 with $\lambda = 0.75$.

(iv) SBR: Treatment assignment is generated as in Example 4.

We compute the true LATE effect τ_0 using Monte Carlo simulations, with sample size being 10000 and the number of Monte Carlo simulations being 1000. We test the true hypothesis

$$H_0 : \tau = \tau_0$$

by the test described in Theorem 5.1 in order to gauge the size of the test. The power is investigated by the hypothesis

$$H_0 : \tau = \tau_0 + 1.$$

All the tests are carried out at 5% level of significance.

9.2 Estimators for Comparison

For DGPs(i)-(ii), we consider the following estimators.

(i) NA: the estimator with no adjustments, i.e., setting $\bar{\mu}^b(a, s, x) = \hat{\mu}^b(a, s, x) = 0$ for $b = D, Y$, $a = 0, 1$, all s and all x.

(ii) TSLS: the two-stage least squares (TSLS) estimator of τ, i.e., $\hat{\tau}_{tsls}$ defined in Section 3. We use the IV heteroskedasticity-robust standard error for inference.

(iii) LP: the optimal linear estimator with $\Psi_{i,s} = X_i$ and the pseudo true values being estimated by $\hat{\theta}^{LP}_{a,s}$ and $\hat{\beta}^{LP}_{a,s}$ defined in (6.5). This is asymptotically equivalent to $\hat{\tau}_S$ proposed by Ansel et al. (2018).

(iv) LG: the OLS-MLE estimator with $\Psi_{i,s} = X_i$, and the pseudo true values being estimated by $\hat{\theta}^{OLS}_{a,s}$ and $\hat{\beta}^{MLE}_{a,s}$ defined in (6.9).

(v) F: the further efficiency improving estimator with $\Psi_{i,s} = X_i$, and the pseudo true values being estimated by $\hat{\theta}^{F}_{a,s}$ and $\hat{\beta}^{F}_{a,s}$ defined in (6.13).

(vi) NP: the nonparametric estimator outlined in Section 7. The sieve regressors are

$$\mathring{\Psi}_{i,n} = \Big(1, X_{1,i}, X_{2,i}, X_{1,i}^2, X_{2,i}^2, X_{1,i}1\{X_{1,i} > t_1\}, X_{2,i}1\{X_{2,i} > t_2\}, X_{1,i}X_{2,i},$$
$$X_{1,i}1\{X_{1,i} > t_1\}X_{2,i}1\{X_{2,i} > t_2\}\Big)^\top$$

where t_1 and t_2 are the sample medians of $\{X_{1,i}\}_{i\in[n]}$ and $\{X_{2,i}\}_{i\in[n]}$, respectively. The adjustments are computed as in (7.1). The implementation details are given in Section A.

(vii) R: a regularized estimator. The nonparametric estimator outlined in Section 7 (NP) might not have a good size when the sample size is small, so we propose to use Lasso to select the sieve regressors. The sieve regressors are

$$\mathring{\Psi}_{i,n} = \Big(1, X_{1,i}, X_{2,i}, X_{1,i}^2, X_{2,i}^2, X_{1,i}1\{X_{1,i} > t_1\}, X_{2,i}1\{X_{2,i} > t_2\}, X_{1,i}X_{2,i},$$
$$X_{1,i}1\{X_{1,i} > t_1\}X_{2,i}1\{X_{2,i} > t_2\}, X_{1,i}^2 1\{X_{1,i} > t_1\}, X_{2,i}^2 1\{X_{2,i} > t_2\}\Big)^\top$$

where t_1 and t_2 are the sample medians of $\{X_{1,i}\}_{i\in[n]}$ and $\{X_{2,i}\}_{i\in[n]}$, respectively. The adjustments are computed as in (8.1). The implementation details are given in Section A.

For GDP(iii), we consider the estimator with no adjustments (NA), and the lasso estimators $\hat{\theta}_{a,s}^R$ and $\hat{\beta}_{a,s}^R$ defined in (8.1) with $\Psi_{i,n} = X_i$. The implementation details are given in Section A.

9.3 Simulation Results

Table 1 presents the empirical sizes and powers of the true null $H_0 : \tau = \tau_0$ and false null $H_0 : \tau = \tau_0 + 1$, respectively, under DGPs (i)-(iii). Note that none of the working models is correctly specified. Consider DGP (i). When $N = 200$, the NA estimator is slightly under-sized. The NP estimator is over-sized because the number of sieve regressors is relatively large compared to the sample size, while the R estimator has the correct size thanks to the Lasso selection of the sieve regressors. All other estimators have sizes close to the nominal level 5%. This confirms that our estimation and inference procedures are robust to misspecification.

In terms of power, the NA estimator has the lowest power, corroborating the belief that one should carry out the regression adjustment whenever covariates correlate with the potential outcomes. Powers of the TSLS, LP, LG, F, NP and R estimators are much higher. In particular, power of the LP estimator is slightly higher than that of the LG estimator even though a logistic model is less misspecified. This shows some robustness of the LP estimator. Power of the F estimator is higher than those of the NA, TSLS, LP and LG estimators, which is consistent with our theory that the F estimator is weakly more efficient than those estimators. The NP and R estimators enjoy the highest powers as a nonparametric model could approximate the true specification very well. NP has more size distortion than R when sample size is 200. When the sample size is increased to

Table 1: NA, TSLS, LP, LG, F, NP, R stand for the no-adjustment, TSLS, optimal linear, OLS-MLE, further efficiency improving, nonparametric and regularized estimators, respectively.

	$N = 200$				$N = 400$			
Methods	SRS	WEI	BCD	SBR	SRS	WEI	BCD	SBR
Panel A: DGP(i)								
A.1: Size								
NA	0.033	0.031	0.031	0.034	0.041	0.043	0.042	0.039
TSLS	0.035	0.034	0.032	0.038	0.041	0.040	0.044	0.042
LP	0.042	0.041	0.041	0.045	0.045	0.044	0.047	0.047
LG	0.043	0.042	0.040	0.045	0.045	0.045	0.047	0.047
F	0.054	0.052	0.049	0.053	0.053	0.048	0.052	0.050
NP	0.106	0.094	0.092	0.090	0.064	0.062	0.067	0.062
R	0.056	0.051	0.051	0.054	0.052	0.049	0.050	0.047
A.2: Power								
NA	0.164	0.169	0.170	0.170	0.290	0.289	0.291	0.294
TSLS	0.246	0.254	0.260	0.255	0.436	0.433	0.443	0.436
LP	0.264	0.264	0.273	0.268	0.443	0.440	0.447	0.444
LG	0.257	0.257	0.267	0.261	0.436	0.435	0.443	0.439
F	0.285	0.292	0.296	0.293	0.462	0.454	0.466	0.463
NP	0.300	0.285	0.291	0.282	0.513	0.506	0.510	0.509
R	0.311	0.314	0.320	0.318	0.519	0.512	0.509	0.508
Panel B: DGP(ii)								
B.1: Size								
NA	0.030	0.031	0.029	0.030	0.041	0.043	0.043	0.041
TSLS	0.033	0.033	0.031	0.033	0.042	0.044	0.045	0.040
LP	0.044	0.040	0.044	0.038	0.044	0.047	0.046	0.046
LG	0.043	0.039	0.042	0.037	0.044	0.047	0.046	0.046
F	0.052	0.047	0.048	0.043	0.048	0.049	0.051	0.049
NP	0.102	0.087	0.087	0.081	0.062	0.063	0.065	0.062
R	0.059	0.053	0.050	0.051	0.048	0.051	0.050	0.048
B.2: Power								
NA	0.211	0.208	0.208	0.206	0.339	0.351	0.351	0.344
TSLS	0.214	0.212	0.211	0.210	0.344	0.352	0.354	0.346
LP	0.342	0.331	0.342	0.340	0.517	0.527	0.524	0.516
LG	0.333	0.324	0.336	0.333	0.514	0.524	0.523	0.514
F	0.375	0.374	0.379	0.376	0.563	0.568	0.566	0.561
NP	0.387	0.382	0.386	0.388	0.654	0.664	0.654	0.656
R	0.427	0.426	0.432	0.430	0.652	0.661	0.655	0.646
Panel C: DGP(iii)								
C.1: Size								
NA	0.048	0.043	0.046	0.048	0.040	0.047	0.045	0.047
R	0.049	0.044	0.044	0.041	0.047	0.050	0.049	0.051
C.2: Power								
NA	0.172	0.170	0.170	0.177	0.224	0.237	0.235	0.239
R	0.426	0.436	0.445	0.439	0.739	0.745	0.748	0.746

400, virtually all the sizes and powers of the estimators improve, and all the observations continue to hold.

Most observations uncovered in DGP (i) carry forward to DGP (ii). One new pattern is that powers of the LP, LG, F, NP and R estimators are much higher than those of the NA and TSLS estimators. This is probably because the true specifications for $Y_i(a)$ become more nonlinear.

We now consider DGP (iii). In this setting, only the NA and R estimators are feasible. When $N = 200$, both estimators have the correct sizes but the R estimator has considerably higher power. When $N = 400$, the sizes of these two estimators remain relatively unchanged, while their powers improve with a diverging gap.

9.4 Practical Recommendation

If researchers are restricted to parametric adjustments, we suggest using estimation method F which is guaranteed to be weakly more efficient than simple TSLS, LP, and LG. We can include linear, quadratic and interaction terms of the original covariates as regressors $\Psi_{i,s}$. If the dimension of covariates is high relative to the sample size and/or both covariates and their technical terms are used in the adjustments, we suggest using estimation method NP or R.

10 Empirical Application

Banking the unbanked is considered to be the first step toward broader financial inclusion – the focus of the World Bank's Universal Financial Access 2020 initiative.[4] In a field experiment with a CAR design, Dupas et al. (2018) examined the impact of expanding access to basic saving accounts for rural households living in three countries: Uganda, Malawi, and Chile. In particular, apart from the intent-to-treat effects for the whole sample, they also studied the local average treatment effects for the households who actively used the accounts. This section presents an application of our regression adjusted estimators to the same dataset to examine the LATEs of opening bank accounts on savings – a central outcome of interest in their study.

We focus on the experiment conducted in Uganda. The sample consists of 2,160 households who were randomized with a CAR design. Specifically, within each of 41 strata formed by gender, occupation, and bank branch, half of households were randomly allocated to the treatment group, the other half to the control one. Households in the treatment group were then offered a voucher to open bank accounts with no financial costs. However, not every treated household ever opened and used the saving accounts for deposit. In fact, among the treated households, only 41.87% of them opened the accounts and made at least one deposit within 2 years. Subject compliance is therefore imperfect in this experiment.

The randomization design apparently satisfies statements (i), (ii), and (iii) in Assumption 1.

[4]https://www.worldbank.org/en/topic/financialinclusion/brief/achieving-universal-financial-access-by-2020

The target fraction of treated households is 1/2. Because $\max_{s \in \mathcal{S}} |\frac{B_n(s)}{n(s)}| \approx 0.056$, it is plausible to claim that Assumption 1(iv) is also satisfied. Since households in the control group need to pay for the fees of opening accounts while the treated ones bear no financial costs, no-defiers statement in Assumption 1(v) holds plausibly in this case.

One of the key analyses in Dupas et al. (2018) is to estimate the treatment effects on savings for active users – households who actually opened the accounts and made at least one deposit within 2 years. We follow their footprints to estimate the same LATEs at savings balance.[5] Specifically, for each item in the savings balance, we estimate the LATEs on savings for active users by the methods "NA", "TSLS", "LP", "LG", "F", and "NP". To maintain comparability, for each outcome variable, we keep X_i similar to those used in Dupas et al. (2018) for all the adjusted estimators.[6]

Table 2 presents the LATE estimates and their standard errors (in parentheses) estimated by these methods. These results lead to four observations. First, consistent with the theoretical and simulation results, the standard errors for the LATE estimates with regression adjustments are lower than those without adjustments. This observation holds for all the outcome variables and all the regression adjustment methods. Over the nine outcome variables, the standard errors estimated by regression adjustments are on average around 7% lower than those without adjustment. In particular, when the outcome variable is total informal savings, the standard errors obtained via the further improvement adjustment – "F" method is about 14.9% lower than those without adjustment. This means that regression adjustments, just with the same two covariates used in Dupas et al. (2018), can achieve sizable efficiency gains in estimating the LATEs.

Second, the standard errors for the regression-adjusted LATE estimates are mostly lower than those obtained by the usual TSLS procedure. Especially, when the outcome variable is savings in friends/family, the standard error estimated by the optimal linear adjustment – "LP" method is around 6.9% lower than that obtained by TSLS. This means that, compared with our regression-adjusted methods, TSLS is less efficient to estimate the LATEs under CAR.

Third, the standard errors for the LATE estimates with regression adjustments are similar in terms of magnitude. This implies that all the regression adjustments achieve similar efficiency gain in this case.

Finally, as in Dupas et al. (2018), for the households who actively use bank accounts, we find that reducing the cost of opening a bank account can significantly increase their savings in formal

[5] Savings balance includes savings in formal financial intuitions, mobile money, cash at home or in secret place, savings in ROSCA/VSLA, savings with friends/family, other cash savings, total formal savings, total informal savings, and total savings (See Dupas et al. (2018) for details). Our analysis uses these variables obtained from the first follow-up survey.

[6] The description of these estimators is similar to that in Section 9. Except for savings in formal financial institutions, mobile money, and total formal savings, X_i includes baseline value for outcome of interest and dummy for missing observations. For savings in formal financial institutions, mobile money, and total formal savings, since their baseline values are all zero, we set X_i as baseline value of total savings and dummy for missing observations. For the nonparametric adjustment "NP", we choose the bases $\{1, X_i, (X_{i1} - qx_{1,0.3})\, 1\{X_{i1} > qx_{1,0.3}\}, (X_{i1} - qx_{1,0.5})\, 1\{X_{i1} > qx_{1,0.5}\}\}$ where X_{i1} denotes the continuous X variable. $qx_{1,0.3}$ and $qx_{1,0.5}$ are 30%th and 50%th quantiles of $\{X_{i1}\}_{i \in [n]}$.

institutions. We also observe the evidence of crowd-out – mainly moving cash from saving at home to saving in bank.

Table 2: Impacts on Saving Stocks in 2010 US Dollars

Y	n	NA	TSLS	LP	LG	F	NP
Formal fin. inst.	1968	20.558 (3.067)	20.824 (3.042)	22.192 (3.028)	22.070 (3.018)	23.033 (3.057)	22.367 (3.024)
Mobile	1972	-0.208 (0.223)	-0.197 (0.223)	-0.351 (0.217)	-0.349 (0.216)	-0.363 (0.215)	-0.374 (0.215)
Total formal	1966	20.399 (3.089)	20.678 (3.064)	21.894 (3.053)	21.758 (3.041)	22.737 (3.084)	22.228 (3.070)
Cash at home	1971	-10.826 (5.003)	-8.785 (4.559)	-9.285 (4.506)	-9.153 (4.443)	-8.978 (4.482)	-8.135 (4.390)
ROSCA/ VSLA	1975	-1.933 (1.971)	-2.575 (1.866)	-1.714 (1.784)	-1.769 (1.841)	-1.609 (1.823)	-1.769 (1.841)
Friends/ family	1974	-3.621 (2.040)	-3.472 (1.997)	-2.436 (1.860)	-2.539 (1.946)	-2.525 (1.873)	-2.540 (1.946)
Other cash	1965	0.027 (0.046)	0.040 (0.045)	0.033 (0.044)	0.033 (0.044)	0.033 (0.044)	0.033 (0.044)
Total informal	1960	-17.643 (6.200)	-15.714 (5.454)	-16.894 (5.329)	-16.805 (5.298)	-16.177 (5.276)	-17.973 (5.352)
Total savings	1952	2.787 (7.290)	5.355 (6.502)	6.061 (6.340)	6.023 (6.301)	6.820 (6.292)	5.639 (6.283)

Notes: The table reports the LATE estimates of opening bank accounts on saving stocks. NA, TSLS, LP, LG, F, NP, stand for the no-adjustment, TSLS, optimal linear, OLS-MLE, further efficiency improving, nonparametric estimators, respectively. n is the number of households. Standard errors are in parentheses.

11 Conclusion

In this paper, we address the problem of estimation and inference of local average treatment effects under covariate-adaptive randomizations using regression adjustments. We first derive the semi-parametric efficiency bound for the LATE under CARs. We then propose a regression-adjusted

LATE estimator under CARs. We derive its limit theory and show that, even under the potential misspecification of adjustments, our estimator maintains its consistency and its inference method still achieves an asymptotic size equal to the nominal level under the null. When the adjustment is correctly specified, our LATE estimator achieves the semiparametric efficiency bound. We also examine the efficiency gains brought by regression adjustments in parametric (both linear and nonlinear), nonparametric, and regularized forms. When the adjustment is parametrically misspecified, we construct a new estimator by combining the linear and nonlinear adjustments. This new estimator is shown to be weakly more efficient than all the parametrically adjusted estimators, including the one without any adjustment. Simulations and empirical application confirm efficiency gains that materialize from regression adjustments relative to both the estimator without adjustment and the widely used two-stage least squares estimator.

A Implementation Details for Sieve and Lasso Regressions

Sieve regressions. We provide more details on the sieve basis. Recall $\mathring{\Psi}_{i,n} \equiv (b_{1,n}(x), \cdots, b_{h_n,n}(x))^\top$, where $\{b_{h,n}(\cdot)\}_{h\in[h_n]}$ are h_n basis functions of a linear sieve space, denoted as $\mathcal{B}$. Given that all the elements of vector X are continuously distributed, the sieve space $\mathcal{B}$ can be constructed as follows.

1. For each element $X^{(l)}$ of X, $l = 1, \cdots, d_x$, where d_x denotes the dimension of vector X, let $\mathcal{B}_l$ be the univariate sieve space of dimension J_n. One example of $\mathcal{B}_l$ is the linear span of the J_n dimensional polynomials given by

$$\mathcal{B}_l = \left\{ \sum_{k=0}^{J_n} \alpha_k x^k, x \in \mathrm{Supp}(X^{(l)}), \alpha_k \in \mathbb{R} \right\};$$

   Another example is the linear span of r-order splines with J_n nodes given by

$$\mathcal{B}_l = \left\{ \sum_{k=0}^{r-1} \alpha_k x^k + \sum_{j=1}^{J_n} b_j [\max(x - t_j, 0)]^{r-1}, x \in \mathrm{Supp}(X^{(l)}), \alpha_k, b_j \in \mathbb{R} \right\},$$

   where the grid $-\infty = t_0 \leq t_1 \leq \cdots \leq t_{J_n} \leq t_{J_n+1} = \infty$ partitions $\mathrm{Supp}(X^{(l)})$ into $J_n + 1$ subsets $I_j = [t_j, t_{j+1}) \cap \mathrm{Supp}(X^{(l)})$, $j = 1, \cdots, J_n - 1$, $I_0 = (t_0, t_1) \cap \mathrm{Supp}(X^{(l)})$, and $I_{J_n} = (t_{J_n}, t_{J_n+1}) \cap \mathrm{Supp}(X^{(l)})$.

2. Let $\mathcal{B}$ be the tensor product of $\{\mathcal{B}_l\}_{l=1}^{d_x}$, which is defined as a linear space spanned by the functions $\prod_{l=1}^{d_x} g_l$, where $g_l \in \mathcal{B}_l$. The dimension of $\mathcal{B}$ is then $K \equiv d_x J_n$ if $\mathcal{B}_l$ is spanned by J_n dimensional polynomials.

We refer interested readers to Hirano et al. (2003) and Chen (2007) for more details about the implementation of sieve estimation. Given the sieve basis, we can compute the $\{\hat{\mu}^b(a, s, X_i)\}_{a=0,1,b=D,Y,s\in\mathcal{S}}$ following (7.1).

Lasso regressions. We follow the estimation procedure and the choice of tuning parameter proposed by Belloni et al. (2017). We provide details below for completeness. Recall $\varrho_{n,a}(s) = c\sqrt{n_a(s)}F_N^{-1}(1-1/(p_n\log(n_a(s))))$. We set $c = 1.1$ following Belloni et al. (2017). We then implement the following algorithm to estimate $\hat{\theta}^R_{a,s}$ and $\hat{\beta}^R_{a,s}$:

(i) Let $\hat{\sigma}_h^{Y,(0)} = \frac{1}{n_a(s)}\sum_{i\in I_a(s)}(Y_i - \bar{Y}_{a,s})^2\mathring{\Psi}^2_{i,n,h}$ and $\hat{\sigma}_h^{D,(0)} = \frac{1}{n_a(s)}\sum_{i\in I_a(s)}(D_i - \bar{D}_{a,s})^2\mathring{\Psi}^2_{i,n,h}$ for $h \in [p_n]$, where $\bar{Y}_{a,s} = \frac{1}{n_a(s)}\sum_{i\in I_a(s)} Y_i$ and $\bar{D}_{a,s} = \frac{1}{n_a(s)}\sum_{i\in I_a(s)} D_i$. Estimate

$$\hat{\theta}^{R,0}_{a,s} = \arg\min_t \frac{-1}{n_a(s)}\sum_{i\in I_a(s)}\left(Y_i - \mathring{\Psi}^\top_{i,n}t\right)^2 + \frac{\varrho_{n,a}(s)}{n_a(s)}\sum_{h\in[p_n]}\hat{\sigma}_h^{Y,(0)}|t_h|,$$

$$\hat{\beta}^{R,0}_{a,s} = \arg\min_b \frac{-1}{n_a(s)}\sum_{i\in I_a(s)}\left[D_i\log(\lambda(\mathring{\Psi}^\top_{i,n}b)) + (1-D_i)\log(1-\lambda(\mathring{\Psi}^\top_{i,n}b))\right] + \frac{\varrho_{n,a}(s)}{n_a(s)}\sum_{h\in[p_n]}\hat{\sigma}_h^{D,(0)}|b_h|.$$

(ii) For $k = 1,\cdots,K$, obtain $\hat{\sigma}_h^{Y,(k)} = \sqrt{\frac{1}{n}\sum_{i\in[n]}(\mathring{\Psi}_{i,n,h}\hat{\varepsilon}_i^{Y,(k)})^2}$, where $\hat{\varepsilon}_i^{Y,(k)} = Y_i - \mathring{\Psi}^\top_{i,n}\hat{\theta}^{R,k-1}_{a,s}$ and $\hat{\sigma}_h^{D,(k)} = \sqrt{\frac{1}{n}\sum_{i\in[n]}(\mathring{\Psi}_{i,n,h}\hat{\varepsilon}_i^{D,(k)})^2}$, where $\hat{\varepsilon}_i^{D,(k)} = D_i - \lambda(\mathring{\Psi}^\top_{i,n}\hat{\beta}^{R,k-1}_{a,s})$. Estimate

$$\hat{\theta}^{R,k}_{a,s} = \arg\min_t \frac{-1}{n_a(s)}\sum_{i\in I_a(s)}\left(Y_i - \mathring{\Psi}^\top_{i,n}t\right)^2 + \frac{\varrho_{n,a}(s)}{n_a(s)}\sum_{h\in[p_n]}\hat{\sigma}_h^{Y,(k-1)}|t_h|,$$

$$\hat{\beta}^{R,k}_{a,s} = \arg\min_b \frac{-1}{n_a(s)}\sum_{i\in I_a(s)}\left[D_i\log(\lambda(\mathring{\Psi}^\top_{i,n}b)) + (1-D_i)\log(1-\lambda(\mathring{\Psi}^\top_{i,n}b))\right] + \frac{\varrho_{n,a}(s)}{n_a(s)}\sum_{h\in[p_n]}\hat{\sigma}_h^{D,(k-1)}|b_h|.$$

(iii) Let $\hat{\theta}^R_{a,s} = \hat{\theta}^{R,K}_{a,s}$ and $\hat{\beta}^R_{a,s} = \hat{\beta}^{R,K}_{a,s}$.

B Asymptotic Properties of Two Stage Least Squares Estimators

As Bugni et al. (2018) and Ansel et al. (2018), we need the next assumption on the treatment assignment mechanism.

Assumption 1(*iv′*). *Suppose* $\pi(s) \in (0,1)$ *and*

$$\left\{\left\{\frac{B_n(s)}{\sqrt{n}}\right\}_{s\in\mathcal{S}}\middle|\{S_i\}_{i\in[n]}\right\} \rightsquigarrow \mathcal{N}(0,\Sigma_B),$$

where $B_n(s) = \sum_{i=1}^n(A_i - \pi(s))1\{S_i = s\}$, $\Sigma_B = diag(p(s)\gamma(s) : s \in \mathcal{S})$, *and* $0 \le \gamma(s) \le \pi(s)(1-\pi(s))$.

Under Assumptions 1 and 1(iv'), and when $\pi(s)$ is homogeneous across $s \in \mathcal{S}$, we can state the following result.

Theorem B.1. *Suppose Assumption 1 holds. Then, we have*

$$\hat{\tau}_{tsls} \xrightarrow{p} \frac{\mathbb{E}\pi(S_i)(1-\pi(S_i))\left[\mathbb{E}(Y_i(D_i(1))|S_i) - \mathbb{E}(Y_i(D_i(0))|S_i)\right]}{\mathbb{E}\pi(S_i)(1-\pi(S_i))\left[\mathbb{E}(D_i(1)|S_i) - \mathbb{E}(D_i(0)|S_i)\right]},$$

In addition, if $\pi(s)$ is the same across $s \in \mathcal{S}$ and Assumption 1(iv') holds, then

$$\sqrt{n}(\hat{\tau}_{tsls} - \tau) \rightsquigarrow \mathcal{N}(0, \sigma^2_{tsls}),$$

where $\mathbb{X}_i = (X_i^\top, \{1\{S_i = s\}\}_{s\in\mathcal{S}})^\top$,

$$\begin{aligned}
\sigma^2_{tsls} &= \frac{\sigma^2_{tsls,0} + \sigma^2_{tsls,1} + \sigma^2_{tsls,2} + \sigma^2_{tsls,3}}{(\mathbb{E}(D_i(1) - D_i(0)))^2},\\
\sigma^2_{tsls,1} &= \frac{\mathbb{E}\left[Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^* - \mathbb{E}[Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*|S_i]\right]^2}{\pi}\\
\sigma^2_{tsls,0} &= \frac{\mathbb{E}\left[Y_i(D_i(0)) - D_i(0)\tau - \mathbb{X}_i^\top\lambda^* - \mathbb{E}[Y_i(D_i(0)) - D_i(0)\tau - \mathbb{X}_i^\top\lambda^*|S_i]\right]^2}{1-\pi},\\
\sigma^2_{tsls,2} &= \mathbb{E}\left[\mathbb{E}\left[Y(D(1)) - Y(D(0)) - (D(1) - D(0))\tau|S_i\right]\right]^2,\\
\sigma^2_{tsls,3} &= \mathbb{E}\left\{\gamma(S_i)\left(\mathbb{E}\left[\frac{Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*}{\pi} + \frac{Y_i(D_i(0)) - D_i(0)\tau - \mathbb{X}_i^\top\lambda^*}{1-\pi}\middle|S_i\right]\right)^2\right\},\\
\lambda^* &= \left(\mathbb{E}\mathbb{X}_i\mathbb{X}_i^\top\right)^{-1}\mathbb{E}\mathbb{X}_i\left[\pi(Y_i(D_i(1)) - D_i(1)\tau) + (1-\pi)(Y_i(D_i(0)) - D_i(0)\tau)\right],
\end{aligned}$$

and λ^*_x is the first d_x elements of λ^*.

The proof of Theorem B.1 is shown in Section N in the Supplement. Several remarks are in order. First, for all the results in the rest of the paper, we do not assume Assumption 1(iv'). Second, when $\pi(s)$ is heterogeneous across $s \in \mathcal{S}$, $\hat{\tau}_{tsls}$ may not be consistent. Third, Ansel et al. (2018) studied the asymptotic properties of three TSLS estimators under CARs as well. Specifically, their TSLS estimators are

1. $\hat{\tau}_1$: the coefficient of D from the TSLS of Y on D and 1, with D instrumented by A.

2. $\hat{\tau}_2$: the coefficient of D from the TSLS of Y on D and $\{1\{S = s\}\}_{s\in\mathcal{S}}$, with D instrumented by A.

3. $\hat{\tau}_3$: the coefficient of D from the TSLS of Y on D, $\{1\{S = s\}\}_{s\in\mathcal{S}}$ and

$$\left\{A\cdot\left(1\{S = s\} - n^{-1}\sum_{i\in[n]}1\{S_i = s\}\right)\right\}_{s\in\mathcal{S}},$$

with D instrumented by A.

Under Assumptions 1 and 1(iv'), Ansel et al. (2018) show that $\hat{\tau}_1$ and $\hat{\tau}_2$ are consistent when $\pi(s) = \pi$ for $s \in \mathcal{S}$, while $\hat{\tau}_3$ is consistent even with heterogeneous $\pi(s)$ across strata.

C The Explicit Expression for σ_S^2

To define the explicit expression for σ_S^2, we need to introduce addition notation. For $s \in \mathcal{S}$, let $\tilde{X}_{is} = X_i - \mathbb{E}(X_i|S_i = s)$,

$$\begin{aligned}
\sigma_{S1}^2 &:= \mathbb{E}\Bigg[\pi(S_i)\Bigg\{\frac{Y_i(D_i(1)) - D_i(1)\tau - X^\top \nu_{1s}^{YD} - \mathbb{E}[Y_i(D_i(1)) - D_i(1)\tau - X_i^\top \nu_{1s}^{YD}|S]}{\pi(S)} \\
&\quad + X_i^\top(\nu_{1s}^{YD} - \nu_{0s}^{YD}) - \mathbb{E}\left[X_i^\top(\nu_{1s}^{YD} - \nu_{0s}^{YD})|S_i\right]\Bigg\}^2\Bigg] \\
\sigma_{S0}^2 &:= \mathbb{E}\Bigg[(1-\pi(S_i))\Bigg\{\frac{Y_i(D_i(0)) - D_i(0)\tau - X_i^\top \nu_{0s}^{YD} - \mathbb{E}[Y_i(D_i(0)) - D_i(0)\tau - X_i^\top \nu_{0s}^{YD}|S_i]}{1-\pi(S_i)} \\
&\quad - \left(X^\top(\nu_{1s}^{YD} - \nu_{0s}^{YD}) - \mathbb{E}\left[X_i^\top(\nu_{1s}^{YD} - \nu_{0s}^{YD})|S\right]\right)\Bigg\}^2\Bigg], \\
\sigma_{S2}^2 &:= \mathbb{E}\left[\left(\mathbb{E}\left[Y_i(D_i(1)) - Y_i(D_i(0)) - \tau(D_i(1) - D_i(0))|S_i\right]\right)^2\right], \\
\nu_{1s}^{YD} &:= \left[\mathbb{E}(\tilde{X}_{is}\tilde{X}_{is}^\top|S_i = s)\right]^{-1}\mathbb{E}\left(\tilde{X}_{is}\left[Y_i\left(D_i(1)\right) - D_i(1)\tau\right]|S_i = s\right), \\
\nu_{0s}^{YD} &:= \left[\mathbb{E}(\tilde{X}_{is}\tilde{X}_{is}^\top|S_i = s)\right]^{-1}\mathbb{E}\left(\tilde{X}_{is}\left[Y\left(D_i(0)\right) - D_i(0)\tau\right]|S_i = s\right).
\end{aligned}$$

Theorem C.1. *Suppose Assumptions 1 and 1(iv') hold. Then,*

$$\sigma_S^2 = \frac{\sigma_{S1}^2 + \sigma_{S0}^2 + \sigma_{S2}^2}{(\mathbb{E}(D(1) - D(0)))^2}, \tag{C.1}$$

It can be shown that $\sigma_{Sa}^2 \geq \underline{\sigma}_{Sa}^2$ for $a = 0, 1$ and $\sigma_{S2}^2 = \underline{\sigma}_{S2}^2$, where the inequalities are strict except special cases such as $\mathbb{E}(Y_i(D_i(a)) - D_i(a)\tau|X_i, S_i = s)$ is linear in X_i. This implies in general, the S estimator is not semiparametrically most efficient.

D Proof of Theorem 3.1

Observe that $\sigma_{S1}^2 + \sigma_{S0}^2$ and $\sigma_{tsls,1}^2 + \sigma_{tsls,0}^2$ can be written as

$$\mathbb{E}\left[\pi(S_i)\Xi_1(\mathcal{D}_i, S_i)^2 + (1-\pi(S_i))\Xi_0(\mathcal{D}_i, S_i)^2\right], \tag{D.1}$$

where

$$\Xi_1(\mathcal{D}_i, S_i) := \left[\left(1 - \frac{1}{\pi(S_i)}\right)\tilde{X}_i^\top \theta_{1s} - \tilde{X}_i^\top \theta_{0s} + \frac{\tilde{W}_i}{\pi(S_i)}\right]$$
$$- \tau \left[\left(1 - \frac{1}{\pi(S_i)}\right)\tilde{X}_i^\top \beta_{1s} - \tilde{X}_i^\top \beta_{0s} + \frac{\tilde{D}_i(1)}{\pi(S_i)}\right],$$
$$\Xi_0(\mathcal{D}_i, S_i) := \left[\left(\frac{1}{1-\pi(S_i)} - 1\right)\tilde{X}_i^\top \theta_{0s} + \tilde{X}_i^\top \theta_{1s} - \frac{\tilde{Z}_i}{1-\pi(S_i)}\right]$$
$$- \tau \left[\left(\frac{1}{1-\pi(S_i)} - 1\right)\tilde{X}_i^\top \beta_{0s} + \tilde{X}_i^\top \beta_{1s} - \frac{\tilde{D}_i(0)}{1-\pi(S_i)}\right],$$

with

$$\theta_{as} = \left[\mathbb{E}(\tilde{X}_{is}\tilde{X}_{is}^\top | S_i = s)\right]^{-1}\left[\mathbb{E}(\tilde{X}_{is} Y_i(D_i(a)) | S_i = s)\right] \quad \text{and}$$
$$\beta_{as} = \left[\mathbb{E}(\tilde{X}_{is}\tilde{X}_{is}^\top | S_i = s)\right]^{-1}\left[\mathbb{E}(\tilde{X}_{is} D_i(a) | S_i = s)\right], \quad a = 0, 1$$

for $\hat{\tau}_S$, while $\theta_{1s} = \theta_{0s}$, $\beta_{1s} = \beta_{0s}$ and $\theta_{1s} - \tau\beta_{1s} = \lambda_x^*$ for $\hat{\tau}_{tsls}$. It is obvious that

$$\sigma_{S1}^2 + \sigma_{S0}^2 \leq \sigma_{tsls,1}^2 + \sigma_{tsls,0}^2, \quad \sigma_{S2}^2 = \sigma_{tsls,2}^2, \quad \text{and} \quad 0 \leq \sigma_{tsls,3}^2,$$

which implies $\hat{\tau}_S$ is more efficient than $\hat{\tau}_{tsls}$.

E Proof of Theorem 4.1

Without loss of generality, we assume $A_i = \phi_i(\{S_i\}_{i\in[n]}, U)$, where $\phi_i(\cdot)$ is a deterministic function and U is a random variable (vector) with density $P_U(\cdot)$ and is independent of everything else in the data. Further denote $\mathcal{Y}_i(a) = \{Y_i(D_i(a)), D_i(a), X_i\}$. We consider parametric submodels indexed by a generic parameter θ. The likelihoods of S_i evaluated at s and $\mathcal{Y}_i(a)$ given $S_i = s$ evaluated at $\overline{y}$ are written as $f_S(s;\theta)$ and $f_{\mathcal{Y}(a)|S}(\overline{y}|s;\theta)$ for $a = 0,1$, respectively. The density of U does not depend on θ. Let $\theta_n = \theta^* + h/\sqrt{n}$, where θ^* indexes the true underlying DGP.

By Assumption 1, the joint likelihood of $\{Y_i, X_i, S_i, A_i\}_{i\in[n]}$ under θ can be written as

$$P_U(u)\Pi_{i\in[n]}\left[f_S(s_i;\theta)\Pi_{a=0,1} f_{\mathcal{Y}(a)|S}(\tilde{y}_i(a)|s_i;\theta)^{1\{\phi_i(s_1,\cdots,s_n,u)=a\}}\right]$$

where $(x_i, y_i(d_i(a)), d_i(a), u, s_i)$ are the realizations $(X_i, Y_i(D_i(a)), D_i(a), U, S_i)$ for $i \in [n]$ and $\tilde{y}_i(a) = \{y_i(d_i(a)), d_i(a), x_i\}$. We make the following regularity assumptions with respect to the submodel.

Assumption *E.* *(i) Suppose $f_S(s;\theta)$ and $f_{\mathcal{Y}(a)|S}(\overline{y}|s;\theta)$ for $a = 0,1$ are differentiable in quadratic* *mean at θ^* with score functions $g_s(S_i)$ and $g_a(\mathcal{Y}_i(a)|S_i)$ for $a = 0, 1$, respectively, such that*

$$\dot{f}_{\mathcal{Y}(a)|S}(\overline{y}|s;\theta) = \frac{\partial \log(f_{\mathcal{Y}(a)|S}(\overline{y}|s;\theta))}{\partial\theta}, \quad \dot{f}_S(s;\theta) = \frac{\partial \log(f_S(s;\theta))}{\partial\theta},$$
$$\dot{f}_{\mathcal{Y}(a)|S}(\mathcal{Y}_i(a)|S_i;\theta^*) = g_a(\mathcal{Y}_i(a)|S_i), \quad \text{and} \quad \dot{f}_S(S_i;\theta^*) = g_s(S_i).$$

(ii) *Suppose $\dot{f}_{\mathcal{Y}(a)|S}(\overline{y}|s;\theta)$ and $\dot{f}_S(s;\theta)$ are continuous at θ^* so that there exist a sequence $t_n = o(1)$ and a function $L_a(\mathcal{Y}_i(a), S_i)$ such that*

$$|\dot{f}_{\mathcal{Y}(a)|S}(\mathcal{Y}_i(a)|S_i;\theta^* + h/\sqrt{n}) - g_a(\mathcal{Y}_i(a)|S_i)| + |\dot{f}_S(S_i;\theta^* + h/\sqrt{n}) - g_s(S_i)| \leq t_n L_a(\mathcal{Y}_i(a), S_i)$$

and $\mathbb{E}|Y_i(D_i(a))L_a(\mathcal{Y}_i(a), S_i)| < \infty$ for $a = 0, 1$.

(iii) *Suppose there exists a constant $C > 0$ such that*

$$\max_{s\in\mathcal{S}} \mathbb{E}\left[\left|\Xi_1(\mathcal{D}_i, S_i)g_1(\mathcal{Y}_i(1)|S_i)\right| + \left|\Xi_0(\mathcal{D}_i, S_i)g_0(\mathcal{Y}_i(0)|S_i)\right| \Big| S_i = s\right] \leq C$$
$$\max_{s\in\mathcal{S}} \mathbb{E}\left[\Xi_1(\mathcal{D}_i, S_i)^2 + \Xi_0(\mathcal{D}_i, S_i)^2 \Big| S_i = s\right] \leq C.$$

We denote $\tau(\theta) = \mathbb{E}_\theta(Y_i(1) - Y_i(0)|D_i(1) > D_i(0))$, where $\mathbb{E}_\theta(\cdot)$ means the expectation is taken with the parametric submodel indexed by θ. We further denote $\mathbb{E}(\cdot) = \mathbb{E}_{\theta^*}(\cdot)$, which is the expectation with respect to the true DGP.

Proof of Theorem 4.1. Following the same argument in Armstrong (2022), in order to show the semiparametric efficiency bound, we only need to show (1) local asymptotic normality of the log likelihood ratio for the parametric submodel with tangent set of the form

$$\mathbb{T} = \left(\begin{array}{c} \Psi(\mathcal{D}_i, S_i, A_i) = g_s(S_i) + A_i g_1(\mathcal{Y}_i(1)|S_i) + (1 - A_i)g_0(\mathcal{Y}_i(0)|S_i) : \\ \mathbb{E}\left[g_s^2(S_i) + \sum_{a=0,1} g_a^2(\mathcal{Y}_i(a)|S_i)\right] < \infty, \mathbb{E}g_s(S_i) = 0, \mathbb{E}(g_a(\mathcal{Y}_i(a)|S_i)|S_i) = 0, \\ \mathbb{E}(g_1(\mathcal{Y}_i(1)|S_i)|X_i, S_i) = \mathbb{E}(g_0(\mathcal{Y}_i(0)|S_i)|X_i, S_i) \end{array}\right). \tag{E.1}$$

and (2) $\sqrt{n}(\tau(\theta^* + h/\sqrt{n}) - \tau(\theta^*)) = \langle\tilde{\Psi}, \Psi\rangle_{\bar{\mathbb{P}}} h + o(1)$, where $\tilde{\Psi}(\mathcal{D}_i, S_i, A_i)$ is the efficient score defined as

$$\tilde{\Psi}(\mathcal{D}_i, S_i, A_i) = \left[\Xi_2(S_i) + A_i\Xi_1(\mathcal{D}_i, S_i) + (1 - A_i)\Xi_0(\mathcal{D}_i, S_i)\right] / \mathbb{E}[D_i(1) - D_i(0)] \tag{E.2}$$

and $\langle\tilde{\Psi}, \Psi\rangle_{\bar{\mathbb{P}}} = \frac{1}{n}\sum_{i\in[n]} \mathbb{E}\tilde{\Psi}(\mathcal{D}_i, S_i, A_i)\Psi(\mathcal{D}_i, S_i, A_i)$ is the inner product w.r.t. measure $\bar{\mathbb{P}} := \frac{1}{n}\sum_{i\in[n]} \mathbb{P}_i$. We establish these two results in two steps.

Step 1. Denote $\theta_n = \theta^* + h/\sqrt{n}$ where θ^* is fixed and $\mathbb{P}_{n,h}$ as the joint distribution of $\{Y_i, X_i, S_i, A_i\}_{i\in[n]}$ under θ_n. The log likelihood ratio for θ_n against θ^* is given by

$$\ell_{n,h} = \sum_{i\in[n]} \tilde{\ell}_s(S_i;\theta_n) + \sum_{a=0,1}\sum_{i\in[n]} 1\{A_i = a\}\tilde{\ell}_{\mathcal{Y}(a)|S}(\mathcal{Y}_i|S_i;\theta_n),$$

where $\mathcal{Y}_i = (Y_i, D_i, X_i)$, $\tilde{\ell}_s(S_i;\theta_n) = \log\left(\frac{f_S(S_i;\theta_n)}{f_S(S_i;\theta^*)}\right)$, and $\tilde{\ell}_{\mathcal{Y}(a)|S}(\mathcal{Y}_i|S_i;\theta_n) = \log\left(\frac{f_{\mathcal{Y}(a)|S}(\mathcal{Y}_i|S_i;\theta_n)}{f_{\mathcal{Y}(a)|S}(\mathcal{Y}_i|S_i;\theta^*)}\right)$ for $a = 0, 1$. Then, Armstrong (2022, Corollary 3.1) shows $\ell_{n,h}$ converges in distribution to a $\mathcal{N}(-h'\tilde{I}^*h/2, h'\tilde{I}^*h)$ law under θ^* where $\tilde{I}^*$ is the limit of

$$\mathbb{E}_{\theta^*} g_s^2(S_i) + \frac{1}{n}\sum_{i\in[n]}\sum_{a=0,1} 1\{A_i = a\}\mathbb{E}_{\theta^*}\left[g_a^2(\mathcal{Y}_i(a)|S_i)|S_i\right].$$

and the score for this parametric submodel can be written as

$$\Psi(\mathcal{D}_i, S_i, A_i) = g_s(S_i) + A_i g_1(\mathcal{Y}_i(1)|S_i) + (1 - A_i)g_0(\mathcal{Y}_i(0)|S_i). \tag{E.3}$$

We note that by definition, we have

$$\mathbb{E}g_s(S_i) = 0 \quad \text{and} \quad \mathbb{E}(g_a(\mathcal{Y}_i(a)|S_i)|S_i) = 0.$$

In addition, we have the equality that, for an arbitrary function $h(\cdot)$ of X such that $\mathbb{E}h^2(X) < \infty$,

$$\begin{aligned}
&\mathbb{E}_\theta(h(X)|S) = \int_x h(x) f_{X|S}(x|S;\theta)\mathrm{d}x \\
&= \int_x h(x)\left[\int_{y(d(a)),d(a)} f_{Y(D(a)),D(a)|X,S}(y(d(a)), d(a)|x, S;\theta)\mathrm{d}y(d(a))\mathrm{d}d(a)\right] f_{X|S}(x|S;\theta)\mathrm{d}x \\
&= \int_{y(d(a)),d(a),x} h(x) f_{\mathcal{Y}(a)|S}(y(d(a)), d(a), x|S;\theta)\mathrm{d}y(d(a))\mathrm{d}d(a)\mathrm{d}x
\end{aligned} \tag{E.4}$$

for $a = 0, 1$, where $f_{\mathcal{Y}(a)|S}(y(d(a)), d(a), x|s;\theta)$ is the joint likelihood of $(Y(D(a)), D(a), X)$ given S for $a = 0, 1$. We note that, for $a = 0, 1$,

$$\frac{\partial f_{\mathcal{Y}(a)|S}(y(d(a)), d(a), x|S;\theta^*)}{\partial\theta} = f_{\mathcal{Y}(a)|S}(y(d(a)), d(a), x|S;\theta^*)g_a(\mathcal{Y}(a)|S).$$

Therefore, taking derivatives of θ in (E.4) and evaluating the derivatives at θ^*, we have

$$\mathbb{E}\left[h(X)g_1(\mathcal{Y}(1)|S)|S\right] = \mathbb{E}\left[h(X)g_0(\mathcal{Y}(0)|S)|S\right],$$

which implies $\mathbb{E}\left[g_1(\mathcal{Y}(1)|S) - g_0(\mathcal{Y}(0)|S)|X, S\right] = 0$. Therefore, the tangent set can be written in (E.1).

Step 2. We have

$$\tau(\theta) = \frac{\mathbb{E}_\theta(Y_i(D_i(1)) - Y_i(D_i(0)))}{\mathbb{E}_\theta(D_i(1) - D_i(0))}.$$

By the mean-value theorem, we have

$$\tau(\theta^* + h/\sqrt{n}) - \tau(\theta^*) = \frac{\partial \tau(\theta)}{\partial \theta}\bigg|_{\theta=\tilde{\theta}} \frac{h}{\sqrt{n}}$$
$$= \frac{\partial \tau(\theta)}{\partial \theta}\bigg|_{\theta=\theta^*} \frac{h}{\sqrt{n}} + \left[\frac{\partial \tau(\theta)}{\partial \theta}\bigg|_{\theta=\tilde{\theta}} - \frac{\partial \tau(\theta)}{\partial \theta}\bigg|_{\theta=\theta^*} \right] \frac{h}{\sqrt{n}}.$$

Let $G(\theta) = \mathbb{E}_\theta\left[Y(D(1)) - Y(D(0))\right]$, $H(\theta) = \mathbb{E}_\theta\left[D(1) - D(0)\right]$, $G = G(\theta^*)$, and $H = H(\theta^*)$. Note that $\tau(\theta) = G(\theta)/H(\theta)$ and $\tau = G/H$. Then, we have

$$\frac{\partial G(\theta)}{\partial \theta} = \mathbb{E}_\theta\left[Y(D(1))(\dot{f}_{\mathcal{Y}(1)|S}(\mathcal{Y}(1)|S;\theta) + \dot{f}_S(S;\theta)\right] - \mathbb{E}_\theta\left[Y(D(0))(\dot{f}_{\mathcal{Y}(0)|S}(\mathcal{Y}(0)|S;\theta) + \dot{f}_S(S;\theta))\right]$$
$$\frac{\partial H(\theta)}{\partial \theta} = \mathbb{E}_\theta\left[D(1)(\dot{f}_{\mathcal{Y}(1)|S}(\mathcal{Y}(1)|S;\theta) + \dot{f}_S(S;\theta))\right] - \mathbb{E}_\theta\left[D(0)(\dot{f}_{\mathcal{Y}(0)|S}(\mathcal{Y}(0)|S;\theta) + \dot{f}_S(S;\theta))\right].$$

Therefore by Assumption E we can find a constant L such that

$$\left| \frac{\partial \tau(\theta)}{\partial \theta}\bigg|_{\theta=\tilde{\theta}} - \frac{\partial \tau(\theta)}{\partial \theta}\bigg|_{\theta=\theta^*} \right| = \left| \frac{H(\tilde{\theta})\frac{\partial G(\tilde{\theta})}{\partial \theta} - G(\tilde{\theta})\frac{\partial H(\tilde{\theta})}{\partial \theta}}{H^2(\tilde{\theta})} - \frac{H(\theta^*)\frac{\partial G(\theta^*)}{\partial \theta} - G(\theta^*)\frac{\partial H(\theta^*)}{\partial \theta}}{H^2(\theta^*)} \right|$$
$$\leq t_n L.$$

This implies

$$\sqrt{n}(\tau(\theta^* + h/\sqrt{n}) - \tau(\theta^*)) = \frac{\partial \tau(\theta)}{\partial \theta}\bigg|_{\theta=\theta^*} h + o(1). \tag{E.5}$$

In addition, following the calculation by Frölich (2007), we have

$$\frac{\partial \tau(\theta)}{\partial \theta}\bigg|_{\theta=\theta^*} = \frac{\left[\frac{\partial G(\theta)}{\partial \theta} - \tau \frac{\partial H(\theta)}{\partial \theta}\right]\Big|_{\theta=\theta^*}}{H}$$
$$= \frac{\mathbb{E}\left[(Y(D(1)) - \tau D(1))(g_1(\mathcal{Y}(1)|S) + g_s(S))\right]}{H} - \frac{\mathbb{E}\left[(Y(D(0)) - \tau D(0))(g_0(\mathcal{Y}(0)|S) + g_s(S))\right]}{H},$$

where for notation simplicity, we write $\mathbb{E}_{\theta^*}$ as $\mathbb{E}$. Let

$$\Gamma(X,S) = \Big[\pi(S)(\mathbb{E}(Z|X,S) - \mathbb{E}(Z|S)) + \big(1 - \pi(S)\big)\left(\mathbb{E}(W|X,S) - \mathbb{E}(W|S)\right)$$
$$- \tau\Big(\pi(S)(\mathbb{E}(D(0)|X,S) - \mathbb{E}(D(0)|S)) + \big(1 - \pi(S)\big)\left(\mathbb{E}(D(1)|X,S) - \mathbb{E}(D(1)|S)\right)\Big)\Big].$$

Then, we have

$$Y(D(1)) - \tau D(1) = \pi(S)\Xi_1(\mathcal{D},S) + \mathbb{E}(W - \tau D(1)|S) + \Gamma(X,S),$$
$$Y(D(0)) - \tau D(0) = -(1 - \pi(S))\Xi_0(\mathcal{D},S) + \mathbb{E}(Z - \tau D(0)|S) + \Gamma(X,S).$$

This implies

$$\begin{aligned}
&\mathbb{E}(Y(D(1)) - \tau D(1))(g_1(\mathcal{Y}(1)|S) + g_s(S)) \\
&= \mathbb{E}\pi(S)\Xi_1(\mathcal{D}, S)g_1(\mathcal{Y}(1)|S) + \mathbb{E}\Gamma(X, S)g_1(\mathcal{Y}(1)|S) + \mathbb{E}(\mathbb{E}(W - \tau D(1)|S)g_s(S)),
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}(Y(D(0)) - \tau D(0))(g_0(\mathcal{Y}(0)|S) + g_s(S)) \\
&= -\mathbb{E}(1 - \pi(S))\Xi_0(\mathcal{D}, S)g_0(\mathcal{Y}(0)|S) + \mathbb{E}\Gamma(X, S)g_0(\mathcal{Y}(0)|S) + \mathbb{E}(\mathbb{E}(Z - \tau D(0)|S)g_s(S)),
\end{aligned}$$

where we have used $\mathbb{E}[\Xi_a(\mathcal{D}, S)|S] = 0$, $\mathbb{E}[\Gamma(X, S)|S] = 0$ and $\mathbb{E}[g_a(\mathcal{Y}(a)|S)|S] = 0$ for $a = 0, 1$. Then

$$\begin{aligned}
\frac{\partial \tau(\theta)}{\partial \theta}\bigg|_{\theta=\theta^*} &= \frac{\mathbb{E}\pi(S)\Xi_1(\mathcal{D}, S)g_1(\mathcal{Y}(1)|S)}{H} + \frac{\mathbb{E}(1 - \pi(S))\Xi_0(\mathcal{D}, S)g_0(\mathcal{Y}(0)|S)}{H} + \frac{\mathbb{E}g_s(S)\Xi_2(S)}{H} \\
&\quad + \frac{\mathbb{E}\Gamma(X, S)(g_1(\mathcal{Y}(1)|S) - g_0(\mathcal{Y}(0)|S))}{H} \\
&= \frac{\mathbb{E}\pi(S)\Xi_1(\mathcal{D}, S)g_1(\mathcal{Y}(1)|S)}{H} + \frac{\mathbb{E}(1 - \pi(S))\Xi_0(\mathcal{D}, S)g_0(\mathcal{Y}(0)|S)}{H} + \frac{\mathbb{E}g_s(S)\Xi_2(S)}{H}.
\end{aligned} \tag{E.6}$$

where the last equality is due to (E.1).

On the other hand, we note that

$$\begin{aligned}
\langle \tilde{\Psi}, \Psi \rangle_{\bar{\mathbb{P}}} &= \frac{1}{n}\sum_{i\in[n]}\left[\frac{\mathbb{E}\left[g_s(S_i)\Xi_2(S_i)\right]}{H} + \frac{\mathbb{E}\left[A_i\Xi_1(\mathcal{D}_i, S_i)g_1(\mathcal{Y}_i(1)|S_i)\right]}{H} + \frac{\mathbb{E}\left[(1 - A_i)\Xi_0(\mathcal{D}_i, S_i)g_0(\mathcal{Y}_i(0)|S_i)\right]}{H}\right] \\
&= \frac{\partial \tau(\theta)}{\partial \theta}\bigg|_{\theta=\theta^*} + \frac{1}{n}\sum_{i\in[n]}\left[\frac{\mathbb{E}\left[(A_i - \pi(S_i))\Xi_1(\mathcal{D}_i, S_i)g_1(\mathcal{Y}_i(1)|S_i)\right]}{H} - \frac{\mathbb{E}\left[(A_i - \pi(S_i))\Xi_0(\mathcal{D}_i, S_i)g_0(\mathcal{Y}_i(0)|S_i)\right]}{H}\right].
\end{aligned}$$

In addition, by Assumption E, we have, for some constant $C > 0$, that

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i\in[n]}\left[\frac{\mathbb{E}(A_i - \pi(S_i))\Xi_1(\mathcal{D}_i, S_i)g_1(\mathcal{Y}_i(1)|S_i)}{H} - \frac{\mathbb{E}(A_i - \pi(S_i))\Xi_0(\mathcal{D}_i, S_i)g_0(\mathcal{Y}_i(0)|S_i)}{H}\right]\right| \\
&\leq \frac{C}{n}\sum_{s\in\mathcal{S}}\mathbb{E}|B_n(s)| = o(1),
\end{aligned}$$

where the inequality is by law of iterated expectation and Assumption E(iii) and the last equality is due to $\mathbb{E}|B_n(s)|/n = o(1)$.[7] This implies

$$\langle \tilde{\Psi}, \Psi \rangle_{\bar{\mathbb{P}}} = \frac{\partial \tau(\theta)}{\partial \theta}\bigg|_{\theta=\theta^*} + o(1). \tag{E.7}$$

[7]Since $|B_n(s)/n| \leq 1$, $\{B_n(s)/n\}$ is uniformly integrable. Then from $B_n(s)/n = o_p(1)$, we have $\mathbb{E}|B_n(s)|/n = o(1)$.

Combining (E.5), (E.6) and (E.7), we obtained the desired result for Step 2. Last, it is obvious from the previous calculation that

$$\langle \tilde{\Psi}, \tilde{\Psi} \rangle_{\bar{\mathbb{P}}} \to \underline{\sigma}^2.$$

F Proof of Theorem 5.1

Let

$$G := \mathbb{E}\left[\left(Y(1) - Y(0)\right)\left(D(1) - D(0)\right)\right],$$
$$H := \mathbb{E}\left[D(1) - D(0)\right],$$
$$\hat{G} := \frac{1}{n}\sum_{i\in[n]}\left[\frac{A_i(Y_i - \hat{\mu}^Y(1, S_i, X_i))}{\hat{\pi}(S_i)} - \frac{(1 - A_i)(Y_i - \hat{\mu}^Y(0, S_i, X_i))}{1 - \hat{\pi}(S_i)} + \hat{\mu}^Y(1, S_i, X_i) - \hat{\mu}^Y(0, S_i, X_i)\right],$$
$$\hat{H} := \frac{1}{n}\sum_{i\in[n]}\left[\frac{A_i(D_i - \hat{\mu}^D(1, S_i, X_i))}{\hat{\pi}(S_i)} - \frac{(1 - A_i)(D_i - \hat{\mu}^D(0, S_i, X_i))}{1 - \hat{\pi}(S_i)} + \hat{\mu}^D(1, S_i, X_i) - \hat{\mu}^D(0, S_i, X_i)\right].$$

Then, we have

$$\begin{aligned}\sqrt{n}(\hat{\tau} - \tau) &= \sqrt{n}\left(\frac{\hat{G}}{\hat{H}} - \frac{G}{H}\right)\\ &= \frac{1}{\hat{H}}\sqrt{n}(\hat{G} - G) - \frac{G}{\hat{H}H}\sqrt{n}(\hat{H} - H)\\ &= \frac{1}{\hat{H}}\left[\sqrt{n}(\hat{G} - G) - \tau\sqrt{n}(\hat{H} - H)\right]. \qquad \text{(F.1)}\end{aligned}$$

Next, we divide the proof into three steps. In the first step, we obtain the linear expansion of $\sqrt{n}(\hat{G} - G)$. Based on the same argument, we can obtain the linear expansion of $\sqrt{n}(\hat{H} - H)$. In the second step, we obtain the linear expansion of $\sqrt{n}(\hat{\tau} - \tau)$ and then prove the asymptotic normality. In the third step, we show the consistency of $\hat{\sigma}$. The second result in the Theorem is obvious given the semiparametric efficiency bound derived in Theorem 4.1.

Step 1. We have

$$\begin{aligned}\sqrt{n}(\hat{G} - G) = &\frac{1}{\sqrt{n}}\sum_{i\in[n]}\Bigg[\frac{A_i(Y_i - \hat{\mu}^Y(1, S_i, X_i))}{\hat{\pi}(S_i)} - \frac{(1 - A_i)(Y_i - \hat{\mu}^Y(0, S_i, X_i))}{1 - \hat{\pi}(S_i)}\\ &+ \hat{\mu}^Y(1, S_i, X_i) - \hat{\mu}^Y(0, S_i, X_i)\Bigg] - \sqrt{n}G\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\hat{\mu}^Y(1, S_i, X_i) - \frac{A_i\hat{\mu}^Y(1, S_i, X_i)}{\hat{\pi}(S_i)}\right]\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{(1-A_i)\hat{\mu}^Y(0,S_i,X_i)}{1-\hat{\pi}(S_i)}-\hat{\mu}^Y(0,S_i,X_i)\right] \\
&+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{A_iY_i}{\hat{\pi}(S_i)}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{(1-A_i)Y_i}{1-\hat{\pi}(S_i)}-\sqrt{n}G \\
&=: R_{n,1}+R_{n,2}+R_{n,3},
\end{aligned}
$$

where

$$
\begin{aligned}
R_{n,1} &:= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\hat{\mu}^Y(1,S_i,X_i)-\frac{A_i\hat{\mu}^Y(1,S_i,X_i)}{\hat{\pi}(S_i)}\right], \\
R_{n,2} &:= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{(1-A_i)\hat{\mu}^Y(0,S_i,X_i)}{1-\hat{\pi}(S_i)}-\hat{\mu}^Y(0,S_i,X_i)\right], \\
R_{n,3} &:= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{A_iY_i}{\hat{\pi}(S_i)}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{(1-A_i)Y_i}{1-\hat{\pi}(S_i)}-\sqrt{n}G.
\end{aligned}
$$

Lemma P.1 shows that

$$
\begin{aligned}
R_{n,1} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1-\frac{1}{\pi(S_i)}\right)A_i\tilde{\mu}^Y(1,S_i,X_i)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-A_i)\tilde{\mu}^Y(1,S_i,X_i)+o_p(1), \\
R_{n,2} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{1}{1-\pi(S_i)}-1\right)(1-A_i)\tilde{\mu}^Y(0,S_i,X_i)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i\tilde{\mu}^Y(0,S_i,X_i)+o_p(1), \\
R_{n,3} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{\pi(S_i)}\tilde{W}_iA_i-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1-A_i}{1-\pi(S_i)}\tilde{Z}_i+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\mathbb{E}[W_i-Z_i|S_i]-\mathbb{E}[W_i-Z_i]\right).
\end{aligned}
$$

This implies

$$
\begin{aligned}
\sqrt{n}(\hat{G}-G) = &\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\left(1-\frac{1}{\pi(S_i)}\right)\tilde{\mu}^Y(1,S_i,X_i)-\tilde{\mu}^Y(0,S_i,X_i)+\frac{\tilde{W}_i}{\pi(S_i)}\right]A_i\right. \\
&\left.+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\left(\frac{1}{1-\pi(S_i)}-1\right)\tilde{\mu}^Y(0,S_i,X_i)+\tilde{\mu}^Y(1,S_i,X_i)-\frac{\tilde{Z}_i}{1-\pi(S_i)}\right](1-A_i)\right\} \\
&+\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\mathbb{E}[W_i-Z_i|S_i]-\mathbb{E}[W_i-Z_i]\right)\right\}+o_p(1). \qquad \text{(F.2)}
\end{aligned}
$$

Similarly, we can show that

$$
\begin{aligned}
\sqrt{n}(\hat{H}-H) = &\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\left(1-\frac{1}{\pi(S_i)}\right)\tilde{\mu}^D(1,S_i,X_i)-\tilde{\mu}^D(0,S_i,X_i)+\frac{\tilde{D}_i(1)}{\pi(S_i)}\right]A_i\right. \\
&\left.+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\left(\frac{1}{1-\pi(S_i)}-1\right)\tilde{\mu}^D(0,S_i,X_i)+\tilde{\mu}^D(1,S_i,X_i)-\frac{\tilde{D}_i(0)}{1-\pi(S_i)}\right](1-A_i)\right\}
\end{aligned}
$$

$$+\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(\mathbb{E}[D_i(1)-D_i(0)|S_i]-\mathbb{E}[D_i(1)-D_i(0)]\big)\right\}+o_p(1), \tag{F.3}$$

where $\tilde{D}_i(a)=D_i(a)-\mathbb{E}(D_i(a)|S_i)$ for $a=0,1$ and $\tilde{\mu}^D(0,s,X_i)=\overline{\mu}^D(0,s,X_i)-\mathbb{E}(\overline{\mu}^D(0,S_i,X_i)|S_i=s)$.

Combining (F.1), (F.2), and (F.3), we obtain the linear expansion for $\hat{\tau}$ as

$$\begin{aligned}\sqrt{n}(\hat{\tau}-\tau)&=\frac{1}{\hat{H}}\left[\sqrt{n}(\hat{G}-G)-\tau\sqrt{n}(\hat{H}-H)\right]\\&=\frac{1}{\hat{H}}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Xi_1(\mathcal{D}_i,S_i)A_i+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Xi_0(\mathcal{D}_i,S_i)(1-A_i)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Xi_2(S_i)\right]+o_p(1),\end{aligned}$$

where $\mathcal{D}_i=\{Y_i(1),Y_i(0),D_i(1),D_i(0),X_i\}$,

$$\begin{aligned}\Xi_1(\mathcal{D}_i,S_i)&=\left[\left(1-\frac{1}{\pi(S_i)}\right)\tilde{\mu}^Y(1,S_i,X_i)-\tilde{\mu}^Y(0,S_i,X_i)+\frac{\tilde{W}_i}{\pi(S_i)}\right]\\&\quad-\tau\left[\left(1-\frac{1}{\pi(S_i)}\right)\tilde{\mu}^D(1,S_i,X_i)-\tilde{\mu}^D(0,S_i,X_i)+\frac{\tilde{D}_i(1)}{\pi(S_i)}\right],\\\Xi_0(\mathcal{D}_i,S_i)&=\left[\left(\frac{1}{1-\pi(S_i)}-1\right)\tilde{\mu}^Y(0,S_i,X_i)+\tilde{\mu}^Y(1,S_i,X_i)-\frac{\tilde{Z}_i}{1-\pi(S_i)}\right]\\&\quad-\tau\left[\left(\frac{1}{1-\pi(S_i)}-1\right)\tilde{\mu}^D(0,S_i,X_i)+\tilde{\mu}^D(1,S_i,X_i)-\frac{\tilde{D}_i(0)}{1-\pi(S_i)}\right],\\\Xi_2(S_i)&=\big(\mathbb{E}[W_i-Z_i|S_i]-\mathbb{E}[W_i-Z_i]\big)-\tau\left[\mathbb{E}[D_i(1)-D_i(0)|S_i]-\mathbb{E}[D_i(1)-D_i(0)]\right].\end{aligned}$$

Step 2. Lemma P.2 implies that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Xi_1(\mathcal{D}_i,S_i)A_i\rightsquigarrow\mathcal{N}(0,\sigma_1^2),\quad\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Xi_0(\mathcal{D}_i,S_i)(1-A_i)\rightsquigarrow\mathcal{N}(0,\sigma_0^2),\quad\text{and}$$
$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Xi_2(S_i)\rightsquigarrow\mathcal{N}(0,\sigma_2^2),$$

and the three terms are asymptotically independent, where

$$\sigma_1^2=\mathbb{E}\pi(S_i)\Xi_1^2(\mathcal{D}_i,S_i),\quad\sigma_0^2=\mathbb{E}(1-\pi(S_i))\Xi_0^2(\mathcal{D}_i,S_i),\quad\text{and}\quad\sigma_2^2=\mathbb{E}\Xi_2^2(S_i).$$

This further implies $\hat{H}\xrightarrow{p}H$ and

$$\sqrt{n}(\hat{\tau}-\tau)\rightsquigarrow\mathcal{N}\left(0,\frac{\sigma_1^2+\sigma_0^2+\sigma_2^2}{H^2}\right),$$

Step 3. We aim to show the consistency of $\hat{\sigma}^2$. First note that

$$\frac{1}{n}\sum_{i=1}^{n}\Xi_{H,i} = \hat{H} \xrightarrow{p} H = \mathbb{E}(D_i(1) - D_i(0)).$$

In addition, Lemma P.3 shows.

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\Xi}_1^2(\mathcal{D}_i, S_i) \xrightarrow{p} \sigma_1^2, \quad \frac{1}{n}\sum_{i=1}^{n}\hat{\Xi}_0^2(\mathcal{D}_i, S_i) \xrightarrow{p} \sigma_0^2, \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}\hat{\Xi}_2^2(\mathcal{D}_i, S_i) \xrightarrow{p} \sigma_2^2.$$

This implies $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$.

G Proof of Theorem 6.1

The proof is divided into two steps. In the first step, we show Assumption 2(i). In the second step, we establish Assumptions 2(ii) and 2(iii).

Step 1. Recall

$$\Delta^Y(a, s, X_i) = \hat{\mu}^Y(a, s, X_i) - \overline{\mu}^Y(a, s, X_i) = \Lambda_{a,s}^Y(X_i, \hat{\theta}_{a,s}) - \Lambda_{a,s}^Y(X_i, \theta_{a,s}),$$

and $\{X_i^s\}_{i\in[n]}$ is generated independently from the distribution of X_i given $S_i = s$, and so is independent of $\{A_i, S_i\}_{i\in[n]}$. Let $M_{a,s}(\theta_1, \theta_2) := \mathbb{E}[\Lambda_{a,s}^Y(X_i, \theta_1) - \Lambda_{a,s}^Y(X_i, \theta_2)|S_i = s] = \mathbb{E}[\Lambda_{a,s}^Y(X_i^s, \theta_1) - \Lambda_{a,s}^Y(X_i^s, \theta_2)]$. We have

$$\begin{aligned}
&\left|\frac{\sum_{i\in I_1(s)}\Delta^Y(a, s, X_i)}{n_1(s)} - \frac{\sum_{i\in I_0(s)}\Delta^Y(a, s, X_i)}{n_0(s)}\right| \\
&\leq \left|\frac{\sum_{i\in I_1(s)}[\Delta^Y(a, s, X_i) - M_{a,s}(\hat{\theta}_{a,s}, \theta_{a,s})]}{n_1(s)}\right| + \left|\frac{\sum_{i\in I_0(s)}[\Delta^Y(a, s, X_i) - M_{a,s}(\hat{\theta}_{a,s}, \theta_{a,s})]}{n_0(s)}\right| \\
&= o_p(n^{-1/2}). \qquad \text{(G.1)}
\end{aligned}$$

To see the last equality, we note that, for any $\varepsilon > 0$, with probability approaching one (w.p.a.1), we have

$$\max_{s\in\mathcal{S}} ||\hat{\theta}_{a,s} - \theta_{a,s}||_2 \leq \varepsilon.$$

Therefore, on the event $\mathcal{A}_n(\varepsilon) := \{\max_{s\in\mathcal{S}} ||\hat{\theta}_{a,s} - \theta_{a,s}||_2 \leq \varepsilon, \min_{s\in\mathcal{S}} n_1(s) \geq \varepsilon n\}$ we have

$$\begin{aligned}
&\left|\frac{\sum_{i\in I_1(s)}[\Delta^Y(a, s, X_i) - M_{a,s}(\hat{\theta}_{a,s}, \theta_{a,s})]}{n_1(s)}\right| \Bigg| \{A_i, S_i\}_{i\in[n]} \\
&\stackrel{d}{=} \left|\frac{\sum_{i=N(s)+1}^{N(s)+n_1(s)}[\Delta^Y(a, s, X_i^s) - M_{a,s}(\hat{\theta}_{a,s}, \theta_{a,s})]}{n_1(s)}\right| \Bigg| \{A_i, S_i\}_{i\in[n]} \leq ||\mathbb{P}_{n_1(s)} - \mathbb{P}||_{\mathcal{F}} \Bigg| \{A_i, S_i\}_{i\in[n]},
\end{aligned}$$

where

$$\mathcal{F} = \{\Lambda^Y_{a,s}(X_i^s,\theta_1) - \Lambda^Y_{a,s}(X_i^s,\theta_2) - M_{a,s}(\theta_1,\theta_2) : ||\theta_1-\theta_2||_2 \leq \varepsilon\}.$$

Therefore, for any $\delta > 0$ we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{\sum_{i\in I_1(s)}[\Delta^Y(a,s,X_i) - M_{a,s}(\hat\theta_{a,s},\theta_{a,s})]}{n_1(s)}\right| \geq \delta n^{-1/2}\right) \\
&\leq \mathbb{P}\left(\left|\frac{\sum_{i\in I_1(s)}[\Delta^Y(a,s,X_i) - M_{a,s}(\hat\theta_{a,s},\theta_{a,s})]}{n_1(s)}\right| \geq \delta n^{-1/2}, \mathcal{A}_n(\varepsilon)\right) + \mathbb{P}(\mathcal{A}_n^c(\varepsilon)) \\
&\leq \mathbb{E}\left[\mathbb{P}\left(\left|\frac{\sum_{i\in I_1(s)}[\Delta^Y(a,s,X_i) - M_{a,s}(\hat\theta_{a,s},\theta_{a,s})]}{n_1(s)}\right| \geq \delta n^{-1/2}, \mathcal{A}_n(\varepsilon)\middle|\{A_i,S_i\}_{i\in[n]}\right)\right] + \mathbb{P}(\mathcal{A}_n^c(\varepsilon)) \\
&\leq \sum_{s\in\mathcal{S}}\mathbb{E}\left[\mathbb{P}\left(||\mathbb{P}_{n_1(s)} - \mathbb{P}||_{\mathcal{F}} \geq \delta n^{-1/2}\middle|\{A_i,S_i\}_{i\in[n]}\right)1\{n_1(s)\geq n\varepsilon\}\right] + \mathbb{P}(\mathcal{A}_n^c(\varepsilon)) \\
&\leq \sum_{s\in\mathcal{S}}\mathbb{E}\left\{\frac{n^{1/2}\mathbb{E}\left[||\mathbb{P}_{n_1(s)} - \mathbb{P}||_{\mathcal{F}}|\{A_i,S_i\}_{i\in[n]}\right]1\{n_1(s)\geq n\varepsilon\}}{\delta}\right\} + \mathbb{P}(\mathcal{A}_n^c(\varepsilon)).
\end{aligned}$$

By Assumption 3, $\mathcal{F}$ is a VC-class with a fixed VC index and envelope L_i such that $\mathbb{E}(L_i^q|\{A_i,S_i\}_{i\in[n]}) \leq C < \infty$. This implies $\mathbb{E}\max_{i\in[n_1(s)]} L_i^2 \leq Cn_1^{2/q}(s)$. In addition,

$$\sup_{f\in\mathcal{F}}\mathbb{P}f^2 \leq \mathbb{E}L_i^2(\theta_1-\theta_2)^2 \leq C\varepsilon^2.$$

Invoke Chernozhukov, Chetverikov, and Kato (2014, Corollary 5.1) with A and ν being fixed constants, and σ^2, F, M being $C\varepsilon^2$, L, $\max_{1\leq i\leq n_1(s)} L_i$, respectively, in our setting. We have

$$\begin{aligned}
&n^{1/2}\mathbb{E}\left[||\mathbb{P}_{n_1(s)} - \mathbb{P}||_{\mathcal{F}}|\{A_i,S_i\}_{i\in[n]}\right]1\{n_1(s)\geq n\varepsilon\} \\
&\leq C\left(\sqrt{\frac{n}{n_1(s)}\varepsilon^2\log(1/\varepsilon)} + n^{1/2}n_1^{1/q-1}(s)\log(1/\varepsilon)\right)1\{n_1(s)\geq n\varepsilon\} \\
&\leq C(\varepsilon^{1/2}\log^{1/2}(1/\varepsilon) + n^{1/q-1/2}\varepsilon^{1/q-1}\log(1/\varepsilon)).
\end{aligned}$$

Therefore,

$$\mathbb{E}\left\{\frac{n^{1/2}\mathbb{E}\left[||\mathbb{P}_{n_1(s)} - \mathbb{P}||_{\mathcal{F}}|\{A_i,S_i\}_{i\in[n]}\right]1\{n_1(s)\geq n\varepsilon\}}{\delta}\right\} \leq C\mathbb{E}\left(\varepsilon^{1/2}\log^{1/2}(1/\varepsilon) + n^{1/q-1/2}\varepsilon^{1/q-1}\log(1/\varepsilon)\right)/\delta.$$

By letting $n\to\infty$ followed by $\varepsilon\to 0$, we have

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\frac{\sum_{i\in I_1(s)}[\Delta^Y(a,s,X_i) - M_{a,s}(\hat\theta_{a,s},\theta_{a,s})]}{n_1(s)}\right| \geq \delta n^{-1/2}\right) = 0,$$

Therefore,

$$\left|\frac{\sum_{i\in I_1(s)}[\Delta^Y(a,s,X_i)-M_{a,s}(\hat{\theta}_{a,s},\theta_{a,s})]}{n_1(s)}\right|=o_p(n^{-1/2}).$$

For the same reason, we have

$$\left|\frac{\sum_{i\in I_0(s)}[\Delta^Y(a,s,X_i)-M_{a,s}(\hat{\theta}_{a,s},\theta_{a,s})]}{n_0(s)}\right|=o_p(n^{-1/2}),$$

and (G.1) holds.

Step 2. We have

$$\frac{1}{n}\sum_{i=1}^n\Delta^{Y,2}(a,S_i,X_i)=\frac{1}{n}\sum_{i=1}^n\sum_{s\in\mathcal{S}}1\{S_i=s\}(\Lambda^Y_{a,s}(X_i,\hat{\theta}_{a,s})-\Lambda^Y_{a,s}(X_i,\theta_{a,s}))^2$$
$$\leq\left(\frac{1}{n}\sum_{i=1}^n L_i^2\right)C\max_{s\in\mathcal{S}}||\hat{\theta}_{a,s}-\theta_{a,s}||_2^2=o_p(1).$$

This verifies Assumption 2(ii). Assumption 2(iii) holds by Assumption 3(ii).

H Proof of Theorem 6.2

Let

$$\nu^Y(a,S_i,X_i)=\mathbb{E}(Y_i(D_i(a))|S_i,X_i)-\mathbb{E}(Y_i(D_i(a))|S_i)\quad\text{and}$$
$$\nu^D(a,S_i,X_i)=\mathbb{E}(D_i(a)|S_i,X_i)-\mathbb{E}(D_i(a)|S_i).\tag{H.1}$$

Also recall that $W_i=Y_i(D_i(1))$, $Z_i=Y_i(D_i(0))$, $\mu^Y(a,S_i,X_i)=\mathbb{E}(Y_i(D_i(a))|S_i,X_i)$. Then, we have

$$\mathbb{E}\pi(S_i)\Xi_1^2(\mathcal{D}_i,S_i)=\mathbb{E}\left\{\frac{\Big(W_i-\mu^Y(1,S_i,X_i)-\tau(D_i(1)-\mu^D(1,S_i,X_i))\Big)^2}{\pi(S_i)}\right\}$$
$$+\mathbb{E}\Bigg\{\pi(S_i)\Bigg[\frac{\nu^Y(1,S_i,X_i)-\tilde{\mu}^Y(1,S_i,X_i)-\tau(\nu^D(1,S_i,X_i)-\tilde{\mu}^D(1,S_i,X_i))}{\pi(S_i)}$$
$$+\tilde{\mu}^Y(1,S_i,X_i)-\tilde{\mu}^Y(0,S_i,X_i)-\tau(\tilde{\mu}^D(1,S_i,X_i)-\tilde{\mu}^D(0,S_i,X_i))\Bigg]^2\Bigg\}.$$

Similarly, we have

$$\mathbb{E}(1-\pi(S_i))\Xi_0^2(\mathcal{D}_i,S_i) = \mathbb{E}\left\{\frac{\Big(Z_i-\mu^Y(0,S_i,X_i)-\tau(D_i(0)-\mu^D(0,S_i,X_i))\Big)^2}{1-\pi(S_i)}\right\}$$
$$+\mathbb{E}\Big\{(1-\pi(S_i))\Big[\frac{\nu^Y(0,S_i,X_i)-\tilde{\mu}^Y(0,S_i,X_i)-\tau(\nu^D(0,S_i,X_i)-\tilde{\mu}^D(0,S_i,X_i))}{1-\pi(S_i)}$$
$$-\Big(\tilde{\mu}^Y(1,S_i,X_i)-\tilde{\mu}^Y(0,S_i,X_i)-\tau(\tilde{\mu}^D(1,S_i,X_i)-\tilde{\mu}^D(0,S_i,X_i))\Big)\Big]^2\Big\}.$$

Last, we have

$$\begin{aligned}\mathbb{E}\Xi_2^2(S_i) &= \mathbb{E}(\mu^Y(1,S_i,X_i)-\mu^Y(0,S_i,X_i)-\tau(\mu^D(1,S_i,X_i)-\mu^D(0,S_i,X_i)))^2\\ &\quad-\mathbb{E}(\nu^Y(1,S_i,X_i)-\nu^Y(0,S_i,X_i)-\tau(\nu^D(1,S_i,X_i)-\nu^D(0,S_i,X_i)))^2\end{aligned}$$

Let

$$\sigma_*^2 = (\mathbb{P}(D_i(1)>D_i(0)))^{-2}\Big\{\mathbb{E}\left[\frac{\big(W_i-\mu^Y(1,S_i,X_i)-\tau(D_i(1)-\mu^D(1,S_i,X_i))\big)^2}{\pi(S_i)}\right]$$
$$+\mathbb{E}\left[\frac{\big(Z_i-\mu^Y(0,S_i,X_i)-\tau[D_i(0)-\mu^D(0,S_i,X_i)]\big)^2}{1-\pi(S_i)}\right]$$
$$+\mathbb{E}\left(\mu^Y(1,S_i,X_i)-\mu^Y(0,S_i,X_i)-\tau[\mu^D(1,S_i,X_i)-\mu^D(0,S_i,X_i)]\right)^2\Big\},$$

which does not depend on the working models $\overline{\mu}^b(a,S_i,X_i)$ for $a=0,1$ and $b=D,Y$. Then, we have

$$\sigma^2((t_{a,s},b_{a,s})_{a=0,1,s\in\mathcal{S}}) = \frac{\sigma_*^2+V((t_{a,s},b_{a,s})_{a=0,1,s\in\mathcal{S}})}{\mathbb{P}(D_i(1)>D_i(0))^2},$$

where σ_*^2 does not depend on $(t_{a,s},b_{a,s})_{a=0,1,s\in\mathcal{S}}$ and

$$\begin{aligned}V((t_{a,s},b_{a,s})_{a=0,1,s\in\mathcal{S}}) &= \mathbb{E}\left(\sqrt{\frac{\pi(S_i)}{1-\pi(S_i)}}A_0(S_i,X_i)+\sqrt{\frac{1-\pi(S_i)}{\pi(S_i)}}A_1(S_i,X_i)\right)^2\\ &= \sum_{s\in\mathcal{S}}p(s)\mathbb{E}\left[\left(\sqrt{\frac{\pi(s)}{1-\pi(s)}}A_0(s,X_i)+\sqrt{\frac{1-\pi(s)}{\pi(s)}}A_1(s,X_i)\right)^2\Bigg|S_i=s\right]\end{aligned}$$

where for $a=0,1$,

$$\begin{aligned}A_a(s,x) &= \nu^Y(a,s,x)-\tilde{\mu}^Y(a,s,x)-\tau(\nu^D(a,s,x)-\tilde{\mu}^D(a,s,x))\\ &= (\nu^Y(a,s,x)-\tau\nu^D(a,s,x))-\tilde{\Psi}_{i,s}^\top(t_{a,s}-\tau b_{a,s}),\end{aligned}$$

and $(\tilde{\mu}^Y(a,s,x), \tilde{\mu}^D(a,s,x))$ and $(\nu^Y(a,s,x), \nu^D(a,s,x))$ are defined in (4.1) and (H.1), respectively. Specifically, we have

$$\tilde{\mu}^Y(a,s,x) = \tilde{\Psi}_{i,s}^\top t_{a,s}, \quad \tilde{\mu}^D(a,s,x) = \tilde{\Psi}_{i,s}^\top b_{a,s}, \quad \text{and} \quad \tilde{\Psi}_{i,s} = \Psi_{i,s} - \mathbb{E}(\Psi_{i,s}|S_i = s).$$

In order to minimize $V((t_{a,s}, b_{a,s})_{a=0,1,s\in\mathcal{S}})$, it suffices to minimize

$$\mathbb{E}\left[\left(\sqrt{\frac{\pi(s)}{1-\pi(s)}}A_0(s,X_i) + \sqrt{\frac{1-\pi(s)}{\pi(s)}}A_1(s,X_i)\right)^2 \middle| S_i = s\right]$$

for each $s \in \mathcal{S}$. In addition, we have

$$\mathbb{E}\left[\left(\sqrt{\frac{\pi(s)}{1-\pi(s)}}A_0(s,X_i) + \sqrt{\frac{1-\pi(s)}{\pi(s)}}A_1(s,X_i)\right)^2 \middle| S_i = s\right] = \mathbb{E}\left((\overline{y}_{i,s} - \tilde{\Psi}_{i,s}^\top \gamma_s)^2 \middle| S_i = s\right),$$

where

$$\overline{y}_{i,s} = \sqrt{\frac{1-\pi(s)}{\pi(s)}}(\nu^Y(1,s,X_i) - \tau\nu^D(1,s,X_i)) + \sqrt{\frac{\pi(s)}{1-\pi(s)}}(\nu^Y(0,s,X_i) - \tau\nu^D(0,s,X_i))$$

and

$$\gamma_s = \sqrt{\frac{1-\pi(s)}{\pi(s)}}(t_{1,s} - \tau b_{1,s}) + \sqrt{\frac{\pi(s)}{1-\pi(s)}}(t_{0,s} - \tau b_{0,s}).$$

By solving the first order condition, we find that

$$\begin{aligned}
\Theta^* &= \left(\begin{array}{c} (\theta^*_{a,s}, \beta^*_{a,s})_{a=0,1,s\in\mathcal{S}}: \\ \sqrt{\frac{1-\pi(s)}{\pi(s)}}(\theta^*_{1,s} - \tau\beta^*_{1,s}) + \sqrt{\frac{\pi(s)}{1-\pi(s)}}(\theta^*_{0,s} - \tau\beta^*_{0,s}) = \mathbb{E}(\tilde{\Psi}_{i,s}\tilde{\Psi}_{i,s}^\top|S_i = s)^{-1}\mathbb{E}(\tilde{\Psi}_{i,s}\overline{y}_{i,s}|S_i = s) \end{array}\right) \\
&= \left(\begin{array}{c} (\theta^*_{a,s}, \beta^*_{a,s})_{a=0,1,s\in\mathcal{S}}: \\ \sqrt{\frac{1-\pi(s)}{\pi(s)}}(\theta^*_{1,s} - \tau\beta^*_{1,s}) + \sqrt{\frac{\pi(s)}{1-\pi(s)}}(\theta^*_{0,s} - \tau\beta^*_{0,s}) \\ = \sqrt{\frac{1-\pi(s)}{\pi(s)}}(\theta^{LP}_{1,s} - \tau\beta^{LP}_{1,s}) + \sqrt{\frac{\pi(s)}{1-\pi(s)}}(\theta^{LP}_{0,s} - \tau\beta^{LP}_{0,s}). \end{array}\right),
\end{aligned}$$

where

$$\begin{aligned}
\theta^{LP}_{a,s} &= [\mathbb{E}(\tilde{\Psi}_{i,s}\tilde{\Psi}_{i,s}^\top|S_i = s)]^{-1}[\mathbb{E}(\tilde{\Psi}_{i,s}\nu^Y(a,s,X_i)|S_i = s)] \\
&= [\mathbb{E}(\tilde{\Psi}_{i,s}\tilde{\Psi}_{i,s}^\top|S_i = s)]^{-1}[\mathbb{E}(\tilde{\Psi}_{i,s}\mathbb{E}(Y_i(D_i(a))|S_i, X_i)|S_i = s)] \\
&= [\mathbb{E}(\tilde{\Psi}_{i,s}\tilde{\Psi}_{i,s}^\top|S_i = s)]^{-1}[\mathbb{E}(\tilde{\Psi}_{i,s}Y_i(D_i(a))|S_i = s)].
\end{aligned}$$

Similarly, we have

$$\beta^{LP}_{a,s} = [\mathbb{E}(\tilde{\Psi}_{i,s}\tilde{\Psi}_{i,s}^\top|S_i = s)]^{-1}[\mathbb{E}(\tilde{\Psi}_{i,s}D_i(a)|S_i = s)].$$

This concludes the proof.

I Proof of Theorem 6.3

In order to verify Assumption 2, by Theorem 6.1, it suffices to show that $\hat{\theta}^{LP}_{a,s} \xrightarrow{p} \theta^{LP}_{a,s}$ and $\hat{\beta}^{LP}_{a,s} \xrightarrow{p} \beta^{LP}_{a,s}$. We focus on the former with $a = 1$. Let $\{W_i^s, X_i^s\}_{i\in[n]}$ be generated independently from the joint distribution of $(Y_i(D_i(1)), X_i)$ given $S_i = s$ and denote $\Psi^s_{i,s} = \Psi_s(X_i^s)$, $\tilde{\Psi}^s_{i,s} = \Psi_s(X_i^s) - \mathbb{E}\Psi_s(X_i^s)$, $\dot{\Psi}^s_{i,1,s} = \Psi_s(X_i^s) - \frac{1}{n_1(s)}\sum_{i=N(s)+1}^{N(s)+n_1(s)} \Psi_s(X_i^s)$, and $\dot{\Psi}^s_{i,0,s} = \Psi_s(X_i^s) - \frac{1}{n_0(s)}\sum_{i=N(s)+n_1(s)+1}^{N(s)+n(s)} \Psi_s(X_i^s)$. Then, we have

$$\hat{\theta}^{LP}_{1,s} \overset{d}{=} \left(\frac{1}{n_1(s)}\sum_{i=N(s)+1}^{N(s)+n_1(s)} \dot{\Psi}^s_{i,1,s}\dot{\Psi}^{s,\top}_{i,1,s}\right)^{-1}\left(\frac{1}{n_1(s)}\sum_{i=N(s)+1}^{N(s)+n_1(s)} \dot{\Psi}^s_{i,1,s}W_i^s\right).$$

As $\frac{1}{n_1(s)}\sum_{i=N(s)+1}^{N(s)+n_1(s)} \Psi^s_{i,s} \xrightarrow{p} \mathbb{E}\Psi^s_{i,s} = \mathbb{E}(\Psi_s(X_i^s)) = \mathbb{E}(\Psi_s(X_i)|S_i = s)$ by the standard LLN, we have

$$\left(\frac{1}{n_1(s)}\sum_{i=N(s)+1}^{N(s)+n_1(s)} \dot{\Psi}^s_{i,1,s}\dot{\Psi}^{s,\top}_{i,1,s}\right) = \left(\frac{1}{n_1(s)}\sum_{i=N(s)+1}^{N(s)+n_1(s)} \tilde{\Psi}^s_{i,s}\tilde{\Psi}^{s,\top}_{i,s}\right) + o_p(1),$$
$$\left(\frac{1}{n_1(s)}\sum_{i=N(s)+1}^{N(s)+n_1(s)} \dot{\Psi}^s_{i,1,s}W_i^s\right) = \left(\frac{1}{n_1(s)}\sum_{i=N(s)+1}^{N(s)+n_1(s)} \tilde{\Psi}^s_{i,s}W_i^s\right) + o_p(1).$$

In addition, by the standard LLN,

$$\frac{1}{n_1(s)}\sum_{i=N(s)+1}^{N(s)+n_1(s)} \tilde{\Psi}^s_{i,s}\tilde{\Psi}^{s,\top}_{i,s} \xrightarrow{p} \mathbb{E}\tilde{\Psi}^s_{i,s}\tilde{\Psi}^{s,\top}_{i,s} = \mathbb{E}(\tilde{\Psi}_{i,s}\tilde{\Psi}^\top_{i,s}|S_i = s),$$
$$\frac{1}{n_1(s)}\sum_{i=N(s)+1}^{N(s)+n_1(s)} \tilde{\Psi}^s_{i,s}W_i^s \xrightarrow{p} \mathbb{E}\tilde{\Psi}^s_{i,s}W_i^s = \mathbb{E}(\tilde{\Psi}_{i,s}Y_i(D_i(1))|S_i = s).$$

Last, Assumption 4 implies $\mathbb{E}(\tilde{\Psi}_{i,s}\tilde{\Psi}^\top_{i,s}|S_i = s)$ is invertible, this means

$$\hat{\theta}^{LP}_{1,s} \xrightarrow{p} \left[\mathbb{E}(\tilde{\Psi}_{i,s}\tilde{\Psi}^\top_{i,s}|S_i = s)\right]^{-1}\mathbb{E}(\tilde{\Psi}_{i,s}Y_i(D_i(1))|S_i = s) = \theta^{LP}_{1,s}.$$

Similarly, we can show that $\hat{\theta}^{LP}_{0,s} \xrightarrow{p} \theta^{LP}_{0,s}$ and $\hat{\beta}^{LP}_{a,s} \xrightarrow{p} \beta^{LP}_{a,s}$ for $a = 0, 1$ and $s \in \mathcal{S}$. Therefore, Assumption 2 holds, and thus, all the results in Theorem 5.1 hold for $\hat{\tau}^{LP}$. Then, the optimality result in the second half of Theorem 6.3 is a direct consequence of Theorem 6.2.

J Proof of Theorem 6.4

Let $\{D_i^s(1), X_i^s\}_{i\in[n]}$ be generated independently from the joint distribution of $(D_i(1), X_i)$ given $S_i = s$, $\Psi_{i,s}^s = \Psi_s(X_i^s)$, and $\mathring{\Psi}_{i,s}^s = (1, \Psi_{i,s}^{s,\top})^\top$. Then, we have, pointwise in b,

$$
\begin{aligned}
&\frac{1}{n_1(s)} \sum_{i\in I_1(s)} \Big[D_i \log(\lambda(\mathring{\Psi}_{i,s}^\top b)) + (1-D_i)\log(1-\lambda(\mathring{\Psi}_{i,s}^\top b))\Big] \\
&\stackrel{d}{=} \frac{1}{n_1(s)} \sum_{i=N(s)+1}^{N(s)+n_1(s)} \Big[D_i^s(1) \log(\lambda(\mathring{\Psi}_{i,s}^{s,\top} b)) + (1-D_i^s(1))\log(1-\lambda(\mathring{\Psi}_{i,s}^{s,\top} b))\Big] \\
&\stackrel{p}{\longrightarrow} \mathbb{E}\Big[D_i^s(1) \log(\lambda(\mathring{\Psi}_{i,s}^{s,\top} b)) + (1-D_i^s(1))\log(1-\lambda(\mathring{\Psi}_{i,s}^{s,\top} b))\Big] \\
&= \mathbb{E}\Big[D_i(1) \log(\lambda(\mathring{\Psi}_{i,s}^\top b)) + (1-D_i(1))\log(1-\lambda(\mathring{\Psi}_{i,s}^\top b))|S_i = s\Big].
\end{aligned}
$$

As the logistic likelihood function is concave in b, the pointwise convergence in b implies uniform convergence, i.e.,

$$
\begin{aligned}
\sup_b \Bigg| &\frac{1}{n_1(s)} \sum_{i\in I_1(s)} \Big[D_i \log(\lambda(\mathring{\Psi}_{i,s}^\top b)) + (1-D_i)\log(1-\lambda(\mathring{\Psi}_{i,s}^\top b))\Big] \\
&- \mathbb{E}\Big[D_i(1) \log(\lambda(\mathring{\Psi}_{i,s}^\top b)) + (1-D_i(1))\log(1-\lambda(\mathring{\Psi}_{i,s}^\top b))|S_i = s\Big]\Bigg| \stackrel{p}{\longrightarrow} 0.
\end{aligned}
$$

Then, by the standard proof for the extremum estimation, we have $\hat{\beta}_{a,s}^{MLE} \stackrel{p}{\longrightarrow} \beta_{a,s}^{MLE}$. Similarly, we can show that $\hat{\theta}_{a,s}^{OLS} \stackrel{p}{\longrightarrow} \theta_{a,s}^{OLS}$. The verifies Assumption 3(i). Assumptions 3(ii) and 3(iii) follow from Assumption 5(ii). Then, the desired results hold due to Theorem 6.1.

K Proof of Theorem 6.5

We note that the adjustments proposed in Theorem 6.5 are still parametric. Specifically, we have

$$
\begin{aligned}
\overline{\mu}^Y(a,s,X_i) &= \Lambda_{a,s}^Y(X_i, \{\beta_{1,s}^{MLE}, \beta_{0,s}^{MLE}, \theta_{a,s}^F\}), \\
\overline{\mu}^D(a,s,X_i) &= \Lambda_{a,s}^D(X_i, \{\beta_{1,s}^{MLE}, \beta_{0,s}^{MLE}, \beta_{a,s}^F\}), \\
\hat{\mu}^Y(a,s,X_i) &= \Lambda_{a,s}^Y(X_i, \{\hat{\beta}_{1,s}^{MLE}, \hat{\beta}_{0,s}^{MLE}, \hat{\theta}_{a,s}^F\}), \quad \text{and} \\
\hat{\mu}^D(a,s,X_i) &= \Lambda_{a,s}^D(X_i, \{\hat{\beta}_{1,s}^{MLE}, \hat{\beta}_{0,s}^{MLE}, \hat{\beta}_{a,s}^F\}),
\end{aligned}
$$

where

$$
\Lambda_{a,s}^Y(X_i, \{b_1, b_0, t_a^*\}) = \begin{pmatrix} \Psi_{i,s}^\top \\ \lambda(\mathring{\Psi}_{i,s}^\top b_1) \\ \lambda(\mathring{\Psi}_{i,s}^\top b_0) \end{pmatrix}^\top t_a^* \quad \text{and} \quad \Lambda_{a,s}^D(X_i, \{b_1, b_0, b_a^*\}) = \begin{pmatrix} \Psi_{i,s}^\top \\ \lambda(\mathring{\Psi}_{i,s}^\top b_1) \\ \lambda(\mathring{\Psi}_{i,s}^\top b_0) \end{pmatrix}^\top b_a^*.
$$

Therefore, in view of Theorem 6.1, to verify Assumption 2, it suffices to show that $\hat{\theta}^F_{a,s} \xrightarrow{p} \theta^F_{a,s}$ and $\hat{\beta}^F_{a,s} \xrightarrow{p} \beta^F_{a,s}$, as we have already shown the consistency of $\hat{\beta}^{MLE}_{a,s}$ in the proof of Theorem 6.4. We focus on $\hat{\theta}^F_{a,s}$. Define $\dot{\Phi}_{i,a,s} := \Phi_{i,s} - \frac{1}{n_a(s)}\sum_{i\in I_a(s)}\Phi_{i,s}$, where

$$\Phi_{i,s} = \begin{pmatrix} \Psi_{i,s} \\ \lambda(\mathring{\Psi}^\top_{i,s}\beta^{MLE}_{1,s}) \\ \lambda(\mathring{\Psi}^\top_{i,s}\beta^{MLE}_{0,s}) \end{pmatrix}.$$

We first show that

$$\frac{1}{n_a(s)}\sum_{i\in I_a(s)} \breve{\Phi}_{i,a,s}\breve{\Phi}^\top_{i,a,s} = \frac{1}{n_a(s)}\sum_{i\in I_a(s)} \dot{\Phi}_{i,a,s}\dot{\Phi}^\top_{i,a,s} + o_p(1). \tag{K.1}$$

Let $v, u \in \Re^{d_\Psi+2}$ be two arbitrary vectors such that $||u||_2 = ||v||_2 = 1$. Then, we have

$$\begin{aligned}
&\left| v^\top \left[\frac{1}{n_a(s)}\sum_{i\in I_a(s)}\left(\breve{\Phi}_{i,a,s}\breve{\Phi}^\top_{i,a,s} - \dot{\Phi}_{i,a,s}\dot{\Phi}^\top_{i,a,s}\right)\right] u\right| \\
=&\left|\frac{1}{n_a(s)}\sum_{i\in I_a(s)}\left[(v^\top\breve{\Phi}_{i,a,s})(u^\top\breve{\Phi}_{i,a,s}) - (v^\top\dot{\Phi}_{i,a,s})(u^\top\dot{\Phi}_{i,a,s})\right]\right| \\
=&\left|\frac{1}{n_a(s)}\sum_{i\in I_a(s)}\left[v^\top(\breve{\Phi}_{i,a,s} - \dot{\Phi}_{i,a,s})(u^\top\breve{\Phi}_{i,a,s}) + (v^\top\dot{\Phi}_{i,a,s})u^\top(\breve{\Phi}_{i,a,s} - \dot{\Phi}_{i,a,s})\right]\right| \\
\leq&\frac{1}{n_a(s)}\sum_{i\in I_a(s)} ||\breve{\Phi}_{i,a,s} - \dot{\Phi}_{i,a,s}||_2(||\breve{\Phi}_{i,a,s}||_2 + ||\dot{\Phi}_{i,a,s}||_2)
\end{aligned} \tag{K.2}$$

where the first inequality is due to Cauchy-Schwarz inequality. We now show (K.2) is $o_p(1)$. First note that

$$||\breve{\Phi}_{i,a,s} - \dot{\Phi}_{i,a,s}||_2 \leq \sum_{a'=0,1}\|B_{a'}\|_2$$

where

$$B_{a'} := \lambda(\mathring{\Psi}^\top_{i,s}\hat{\beta}^{MLE}_{a',s}) - \lambda(\mathring{\Psi}^\top_{i,s}\beta^{MLE}_{a',s}) - \frac{1}{n_a(s)}\sum_{i\in I_a(s)}\left[\lambda(\mathring{\Psi}^\top_{i,s}\hat{\beta}^{MLE}_{a',s}) - \lambda(\mathring{\Psi}^\top_{i,s}\beta^{MLE}_{a',s})\right].$$

Note that

$$\lambda(\mathring{\Psi}^\top_{i,s}\hat{\beta}^{MLE}_{a',s}) - \lambda(\mathring{\Psi}^\top_{i,s}\beta^{MLE}_{a',s}) = \frac{\partial\lambda(\mathring{\Psi}^\top_{i,s}\tilde{\beta}^{MLE}_{a',s})}{\partial\beta_{a',s}}(\hat{\beta}^{MLE}_{a',s} - \beta^{MLE}_{a',s})$$

$$\frac{1}{n_a(s)}\sum_{i\in I_a(s)}\lambda(\mathring{\Psi}^\top_{i,s}\hat{\beta}^{MLE}_{a',s}) - \lambda(\mathring{\Psi}^\top_{i,s}\beta^{MLE}_{a',s}) = \left[\frac{1}{n_a(s)}\sum_{i\in I_a(s)}\frac{\partial\lambda(\mathring{\Psi}^\top_{i,s}\tilde{\beta}^{MLE}_{a',s})}{\partial\beta_{a',s}}\right](\hat{\beta}^{MLE}_{a',s} - \beta^{MLE}_{a',s})$$

where $\tilde{\beta}^{MLE}_{a',s}$ is a mid-point of $\hat{\beta}^{MLE}_{a',s}$ and $\beta^{MLE}_{a',s}$. Hence

$$\|B_{a'}\|_2 = \left\| \frac{\partial \lambda(\mathring{\Psi}_{i,s}^\top \tilde{\beta}^{MLE}_{a',s})}{\partial \beta_{a',s}} - \frac{1}{n_a(s)} \sum_{i \in I_a(s)} \frac{\partial \lambda(\mathring{\Psi}_{i,s}^\top \tilde{\beta}^{MLE}_{a',s})}{\partial \beta_{a',s}} \right\|_2 \|\hat{\beta}^{MLE}_{a',s} - \beta^{MLE}_{a',s}\|_2.$$

Since $\partial \lambda(u)/\partial u \leq 1$,

$$\left\| \frac{\partial \lambda(\mathring{\Psi}_{i,s}^\top \tilde{\beta}^{MLE}_{a',s})}{\partial \beta_{a',s}} \right\|_2 = \left\| \frac{\partial \lambda(u)}{\partial u} \Big|_{u = \mathring{\Psi}_{i,s}^\top \tilde{\beta}^{MLE}_{a',s}} \cdot \mathring{\Psi}_{i,s}^\top \right\|_2 \leq \|\mathring{\Psi}_{i,s}\|_2.$$

Thus,

$$\begin{aligned}
\|B_{a'}\|_2 &\leq \left(\|\mathring{\Psi}_{i,s}\|_2 + \frac{1}{n_a(s)} \sum_{i \in I_a(s)} \|\mathring{\Psi}_{i,s}\|_2 \right) \|\hat{\beta}^{MLE}_{a',s} - \beta^{MLE}_{a',s}\|_2 \\
&\leq \left(2 + \|\Psi_{i,s}\|_2 + \frac{1}{n_a(s)} \sum_{i \in I_a(s)} \|\Psi_{i,s}\|_2 \right) \|\hat{\beta}^{MLE}_{a',s} - \beta^{MLE}_{a',s}\|_2, \\
||\breve{\Phi}_{i,a,s} - \dot{\Phi}_{i,a,s}||_2 &\leq \left(2 + \|\Psi_{i,s}\|_2 + \frac{1}{n_a(s)} \cdot \sum_{i \in I_a(s)} \|\Psi_{i,s}\|_2 \right) \sum_{a'=0,1} \|\hat{\beta}^{MLE}_{a',s} - \beta^{MLE}_{a',s}\|_2. \qquad \text{(K.3)}
\end{aligned}$$

Moreover, we can show

$$||\breve{\Phi}_{i,a,s}||_2 + ||\dot{\Phi}_{i,a,s}||_2 \leq 2 \left(4 + ||\Psi_{i,s}||_2 + \frac{1}{n_a(s)} \sum_{i \in I_a(s)} \|\Psi_{i,s}\|_2 \right). \qquad \text{(K.4)}$$

Substituting (K.3), (K.4) and the fact that $||\hat{\beta}^{MLE}_{a,s} - \beta^{MLE}_{a,s}||_2 = o_p(1)$ into (K.2), we show that (K.2) is $o_p(1)$. As it holds for arbitrary u, v, it implies (K.1). Similarly, we can show that

$$\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \breve{\Phi}_{i,a,s} Y_i = \frac{1}{n_a(s)} \sum_{i \in I_a(s)} \dot{\Phi}_{i,a,s} Y_i + o_p(1). \qquad \text{(K.5)}$$

Following the same argument in the proof of Theorem 6.3, we can show that

$$\left[\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \dot{\Phi}_{i,a,s} \dot{\Phi}_{i,a,s}^\top \right]^{-1} \left[\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \dot{\Phi}_{i,a,s} Y_i \right] \xrightarrow{p} \theta^F_{a,s}.$$

In addition, by Assumption 6, with probability approaching one, there exists a constant $c > 0$ such that

$$\lambda_{\min} \left(\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \dot{\Phi}_{i,a,s} \dot{\Phi}_{i,a,s}^\top \right) \geq c. \qquad \text{(K.6)}$$

Combining (K.1), (K.5), and (K.6), we can show that

$$\hat{\theta}^F_{a,s} = \left[\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \breve{\Phi}_{i,a,s} \breve{\Phi}^\top_{i,a,s}\right]^{-1} \left[\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \breve{\Phi}_{i,a,s} Y_i\right]$$
$$= \left[\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \dot{\Phi}_{i,a,s} \dot{\Phi}^\top_{i,a,s}\right]^{-1} \left[\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \dot{\Phi}_{i,a,s} Y_i\right] + o_p(1) \xrightarrow{p} \theta^F_{a,s}.$$

Similarly, we have $\hat{\beta}^F_{a,s} \xrightarrow{p} \beta^F_{a,s}$, which implies all the results in Theorem 5.1 hold for $\hat{\tau}^F$. The optimality result in the second half of the theorem is a direct consequence of Theorem 6.2.

L Proof of Theorem 7.1

We focus on verifying Assumption 2 for $\hat{\mu}^D(a,s,X_i)$. The proof for $\hat{\mu}^Y(a,s,X_i)$ is similar and hence omitted. Following the proof of Theorem 6.4, we note that, for each $a = 0,1$ and $s \in \mathcal{S}$, the data in cell $I_a(s)$, denoted $\{D^s_i(a), X^s_i\}_{i \in [n]}$, can be viewed as i.i.d. following the joint distribution of $(D_i(a), X_i)$ given $S_i = s$ conditionally on $\{A_i, S_i\}_{i \in [n]}$. Then following the standard logistic sieve regression in Hirano et al. (2003), we have

$$\max_{a=0,1,s \in \mathcal{S}} ||\hat{\beta}^{NP}_{a,s} - \beta^{NP}_{a,s}||_2 = O_p\left(\sqrt{h_n/n_a(s)}\right).$$

Then we have

$$\begin{aligned}
&\left|\frac{\sum_{i \in I_1(s)} \Delta^D(a,s,X_i)}{n_1(s)} - \frac{\sum_{i \in I_0(s)} \Delta^D(a,s,X_i)}{n_0(s)}\right| \\
&\leq \left|\frac{\sum_{i \in I_1(s)} \left(\lambda(\mathring{\Psi}^\top_{i,n}\hat{\beta}^{NP}_{a,s}) - \lambda(\mathring{\Psi}^\top_{i,n}\beta^{NP}_{a,s})\right)}{n_1(s)} - \frac{\sum_{i \in I_0(s)} \left(\lambda(\mathring{\Psi}^\top_{i,n}\hat{\beta}^{NP}_{a,s}) - \lambda(\mathring{\Psi}^\top_{i,n}\beta^{NP}_{a,s})\right)}{n_0(s)}\right| \\
&\quad + \left|\frac{1}{n_1(s)} \sum_{i \in I_1(s)} \Big(R^D(a,s,X_i) - \mathbb{E}[R^D(a,s,X_i)|S_i = s]\Big)\right| \\
&\quad + \left|\frac{1}{n_0(s)} \sum_{i \in I_0(s)} \Big(R^D(a,s,X_i) - \mathbb{E}[R^D(a,s,X_i)|S_i = s]\Big)\right| =: I + II + III.
\end{aligned} \tag{L.1}$$

To bound term I in (L.1), we define $M_{a,s}(\beta_1,\beta_2) := \mathbb{E}\left[\lambda(\mathring{\Psi}^\top_{i,n}\beta_1) - \lambda(\mathring{\Psi}^\top_{i,n}\beta_2)|S_i = s\right] = \mathbb{E}[\lambda(\mathring{\Psi}^{s,\top}_{i,n}\beta_1) - \lambda(\mathring{\Psi}^{s,\top}_{i,n}\beta_2)]$, where $\mathring{\Psi}^s_{i,n} = \mathring{\Psi}(X^s_i)$. Then we have

$$\begin{aligned}
I \leq &\left|\frac{\sum_{i \in I_1(s)} \left[\lambda(\mathring{\Psi}^\top_{i,n}\hat{\beta}^{NP}_{a,s}) - \lambda(\mathring{\Psi}^\top_{i,n}\beta^{NP}_{a,s}) - M_{a,s}(\hat{\beta}^{NP}_{a,s}, \beta^{NP}_{a,s})\right]}{n_1(s)}\right| \\
&+ \left|\frac{\sum_{i \in I_0(s)} \left[\lambda(\mathring{\Psi}^\top_{i,n}\hat{\beta}^{NP}_{a,s}) - \lambda(\mathring{\Psi}^\top_{i,n}\beta^{NP}_{a,s}) - M_{a,s}(\hat{\beta}^{NP}_{a,s}, \beta^{NP}_{a,s})\right]}{n_0(s)}\right| =: I_1 + I_2.
\end{aligned}$$

Following the argument in the proof of Theorem 6.1, in order to show $I_1 = o_p(n^{-1/2})$, we only need to show

$$n^{1/2}\mathbb{E}\left[||\mathbb{P}_{n_1(s)} - \mathbb{P}||_{\mathcal{F}}|\{A_i, S_i\}_{i\in[n]}\right] 1\{n_1(s) \geq n\varepsilon, n_0(s) \geq n\varepsilon\} = o(1),$$

where ε is an arbitrary but fixed constant, and

$$\mathcal{F} := \left\{\lambda(\mathring{\Psi}_{i,n}^\top \beta_1) - \lambda(\mathring{\Psi}_{i,n}^\top \beta_{a,s}^{NP}) : \beta_1 \in \Re^{h_n}, ||\beta_1 - \beta_{a,s}^{NP}||_2 \leq C\sqrt{h_n/n_a(s)}\right\},$$

for some constant $C > 0$. Furthermore, we note that $\mathcal{F}$ has a bounded envelope, is of the VC-type with VC-index upper bounded by Ch_n,[8] and has

$$\sup_{f\in\mathcal{F}} \mathbb{E}\left[f^2|\{A_i, S_i\}_{i\in[n]}\right] \leq \frac{Ch_n}{n_a(s)}.$$

Invoking Chernozhukov et al. (2014, Corollary 5.1) with A being a constant, $\nu = Ch_n$, $\sigma^2 = Ch_n/n_a(s)$, and F and M being $2h_n$, we have

$$\begin{aligned}
&n^{1/2}\mathbb{E}\left[||\mathbb{P}_{n_1(s)} - \mathbb{P}||_{\mathcal{F}}|\{A_i, S_i\}_{i\in[n]}\right] 1\{n_1(s) \geq n\varepsilon, n_0(s) \geq n\varepsilon\} \\
&\leq C\sqrt{\frac{n}{n_1(s)}}\left(\sqrt{\frac{h_n^2 \log n}{n_a(s)}} + \frac{h_n \log n}{\sqrt{n_1(s)}}\right) 1\{n_1(s) \geq n\varepsilon, n_0(s) \geq n\varepsilon\} \\
&\leq C\sqrt{\frac{1}{\varepsilon}}\left(\sqrt{\frac{h_n^2 \log n}{n\varepsilon}} + \frac{h \log n}{\sqrt{n\varepsilon}}\right) \to 0,
\end{aligned}$$

as $n \to \infty$.

Similarly, we can show $I_2 = o_p(n^{-1/2})$. In addition, we note that

$$II \stackrel{d}{=} \left|\frac{1}{n_1(s)} \sum_{i=N(s)+1}^{N(s)+n_1(s)} \left(R^D(a, s, X_i^s) - \mathbb{E}[R^D(a, s, X_i^s)]\right)\right| = o_p(n^{-1/2})$$

by the Chebyshev's inequality as $\mathbb{E}R^{D,2}(a, s, X_i^s) = \mathbb{E}[R^{D,2}(a, s, X_i)|S_i = s] = o(1)$ by Assumption 7(ii). Similarly we have $III = o_p(n^{-1/2})$. Combining the bounds of I, II, III with (L.1), we have

$$\left|\frac{\sum_{i\in I_1(s)} \Delta^D(a, s, X_i)}{n_1(s)} - \frac{\sum_{i\in I_0(s)} \Delta^D(a, s, X_i)}{n_0(s)}\right| = o_p(n^{-1/2}),$$

which verifies Assumption 2(i).

To verify Assumption 2(ii), we note that

$$\frac{1}{n}\sum_{i=1}^n \Delta^{D,2}(a, s, X_i) \leq \frac{2}{n}\sum_{i=1}^n \left(\lambda(\mathring{\Psi}_{i,n}^\top \hat{\beta}_{a,s}^{NP}) - \lambda(\mathring{\Psi}_{i,n}^\top \beta_{a,s}^{NP})\right)^2 + \frac{2}{n}\sum_{i=1}^n R^{D,2}(a, S_i, X_i)$$

[8] See van der Vaart and Wellner (1996, Section 2.6.5) for the calculation of the VC index.

$$\le \frac{2}{n}\sum_{i=1}^{n} \|\mathring{\Psi}_{i,n}\|_2^2 \|\hat{\beta}_{a,s}^{NP} - \beta_{a,s}^{NP}\|_2^2 + \frac{2}{n}\sum_{i=1}^{n} R^{D,2}(a, S_i, X_i)$$
$$= \frac{2}{n}\sum_{i=1}^{n} \|\mathring{\Psi}_{i,n}\|_2^2 \|\hat{\beta}_{a,s}^{NP} - \beta_{a,s}^{NP}\|_2^2 + o_p(1) \le 2\max_i \|\mathring{\Psi}_{i,n}\|_2^2 \max_s \|\hat{\beta}_{a,s}^{NP} - \beta_{a,s}^{NP}\|_2^2 + o_p(1)$$
$$= O_p\left(\zeta^2(h_n)h_n/n_a(s)\right) + o_p(1) = o_p(1)$$

where the first equality is due to Assumption 7(ii), and the second equality is due to Assumption 7(iv).

Last, Assumption 2(iii) is implied by Assumption 1(vi) via Jensen's inequality.

M Proof of Theorem 8.1

We focus on verifying Assumption 2 for $\hat{\mu}^D(a, s, X_i)$. The proof for $\hat{\mu}^Y(a, s, X_i)$ is similar and hence omitted. Following the proof of Theorem 6.4, we note that, for each $a = 0, 1$ and $s \in \mathcal{S}$, the data in cell $I_a(s)$, denoted $\{D_i^s(a), X_i^s\}_{i\in[n]}$, can be viewed as i.i.d. following the joint distribution of $(D_i(a), X_i)$ given $S_i = s$ conditionally on $\{A_i, S_i\}_{i\in[n]}$. Then following the standard logistic Lasso regression in Belloni et al. (2017), we have

$$\max_{a=0,1,s\in\mathcal{S}} ||\hat{\beta}_{a,s}^R - \beta_{a,s}^R||_2 = O_p\left(\sqrt{h_n \log p_n / n_a(s)}\right) \quad \text{and} \quad \max_{a=0,1,s\in\mathcal{S}} ||\hat{\beta}_{a,s}^R||_0 = O_p(h_n).$$

Then, we have

$$\begin{aligned}
&\left|\frac{\sum_{i\in I_1(s)} \Delta^D(a, s, X_i)}{n_1(s)} - \frac{\sum_{i\in I_0(s)} \Delta^D(a, s, X_i)}{n_0(s)}\right| \\
&\le \left|\frac{\sum_{i\in I_1(s)} \left(\lambda(\mathring{\Psi}_{i,n}^\top \hat{\beta}_{a,s}^R) - \lambda(\mathring{\Psi}_{i,n}^\top \beta_{a,s}^R)\right)}{n_1(s)} - \frac{\sum_{i\in I_0(s)} \left(\lambda(\mathring{\Psi}_{i,n}^\top \hat{\beta}_{a,s}^R) - \lambda(\mathring{\Psi}_{i,n}^\top \beta_{a,s}^R)\right)}{n_0(s)}\right| \\
&\quad + \left|\frac{1}{n_1(s)} \sum_{i\in I_1(s)} \left(R^D(a, s, X_i) - \mathbb{E}[R^D(a, s, X_i)|S_i = s]\right)\right| \\
&\quad + \left|\frac{1}{n_0(s)} \sum_{i\in I_0(s)} \left(R^D(a, s, X_i) - \mathbb{E}[R^D(a, s, X_i)|S_i = s]\right)\right| := I + II + III. \qquad \text{(M.1)}
\end{aligned}$$

To bound term I in (L.1), we define $M_{a,s}(\beta_1, \beta_2) := \mathbb{E}\left[\lambda(\mathring{\Psi}_{i,n}^\top \beta_1) - \lambda(\mathring{\Psi}_{i,n}^\top \beta_2)|S_i = s\right] = \mathbb{E}[\lambda(\mathring{\Psi}_{i,n}^{s,\top} \beta_1) - \lambda(\mathring{\Psi}_{i,n}^{s,\top} \beta_2)]$, where $\mathring{\Psi}_{i,n}^s = \mathring{\Psi}(X_i^s)$. Then we have

$$\begin{aligned}
I \le &\left|\frac{\sum_{i\in I_1(s)} \left[\lambda(\mathring{\Psi}_{i,n}^\top \hat{\beta}_{a,s}^R) - \lambda(\mathring{\Psi}_{i,n}^\top \beta_{a,s}^R) - M_{a,s}(\hat{\beta}_{a,s}^R, \beta_{a,s}^R)\right]}{n_1(s)}\right| \\
&+ \left|\frac{\sum_{i\in I_0(s)} \left[\lambda(\mathring{\Psi}_{i,n}^\top \hat{\beta}_{a,s}^R) - \lambda(\mathring{\Psi}_{i,n}^\top \beta_{a,s}^R) - M_{a,s}(\hat{\beta}_{a,s}^R, \beta_{a,s}^R)\right]}{n_0(s)}\right| =: I_1 + I_2.
\end{aligned}$$

Following the argument in the proof of Theorems 6.1 and 7.1, in order to show $I_1 = o_p(n^{-1/2})$, we only need to show

$$n^{1/2}\mathbb{E}\left[||\mathbb{P}_{n_1(s)} - \mathbb{P}||_{\mathcal{F}}|\{A_i, S_i\}_{i\in[n]}\right] 1\{n_1(s) \geq n\varepsilon, n_0(s) \geq n\varepsilon\} = o(1),$$

where ε is an arbitrary but fixed constant, and

$$\mathcal{F} := \left\{\lambda(\mathring{\Psi}_{i,n}^\top \beta_1) - \lambda(\mathring{\Psi}_{i,n}^\top \beta_{a,s}^R) : \beta_1 \in \Re^{h_n}, ||\beta_1 - \beta_{a,s}^R||_2 \leq C\sqrt{h_n \log(p_n)/n_a(s)}, ||\beta_1||_0 \leq Ch_n\right\},$$

for some constant $C > 0$. Furthermore, we note that $\mathcal{F}$ has a bounded envelope and

$$\sup_Q N(\mathcal{F}, e_Q, \varepsilon||F||_{Q,2}) \leq \left(\frac{c_1 p_n}{\varepsilon}\right)^{c_2 h_n},$$

where c_1, c_2 are two fixed constants, $N(\cdot)$ is the covering number, $e_Q(f,g) = \sqrt{Q|f-g|^2}$, and the supremum is taken over all discrete probability measures Q. Last, we have

$$\sup_{f\in\mathcal{F}} \mathbb{E}\left[f^2|\{A_i, S_i\}_{i\in[n]}\right] \leq \frac{Ch_n \log p_n}{n_a(s)}.$$

Invoking Chernozhukov et al. (2014, Corollary 5.1) with $A = Cp_n$, $\nu = Ch_n$, $\sigma^2 = Ch_n \log(p_n)/n_a(s)$, and F and M being 2, we have

$$\begin{aligned}
&n^{1/2}\mathbb{E}\left[||\mathbb{P}_{n_1(s)} - \mathbb{P}||_{\mathcal{F}}|\{A_i, S_i\}_{i\in[n]}\right] 1\{n_1(s) \geq n\varepsilon, n_0(s) \geq n\varepsilon\} \\
&\leq C\sqrt{\frac{n}{n_1(s)}}\left(\sqrt{h_n \frac{h_n \log p_n}{n_a(s)} \log\left(\frac{p_n}{\sqrt{\frac{h_n \log p_n}{n_a(s)}}}\right)} + \frac{h_n}{\sqrt{n_1(s)}} \log\left(\frac{p_n}{\sqrt{\frac{h_n \log p_n}{n_a(s)}}}\right)\right) 1\{n_1(s) \geq n\varepsilon, n_0(s) \geq n\varepsilon\} \\
&\leq C\left(\sqrt{\frac{n}{n_1(s)}}\right)\left(\frac{h_n \log(p_n)}{\sqrt{n_1(s) \wedge n_0(s)}}\right) 1\{n_1(s) \geq n\varepsilon, n_0(s) \geq n\varepsilon\} \to 0.
\end{aligned}$$

The bounds for I_2, II and III can be established following the same argument as in the proof of Theorem 7.1. We omit the detail for brevity. This leads to Assumption 2(i).

To verify Assumption 2(ii), we note that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \Delta^{D,2}(a, s, X_i) &\leq \frac{2}{n}\sum_{i=1}^n \left(\lambda(\mathring{\Psi}_{i,n}^\top \hat{\beta}_{a,s}^R) - \lambda(\mathring{\Psi}_{i,n}^\top \beta_{a,s}^R)\right)^2 + \frac{2}{n}\sum_{i=1}^n R^{D,2}(a, S_i, X_i) \\
&= \frac{2}{n}\sum_{i=1}^n \left(\lambda(\mathring{\Psi}_{i,n}^\top \hat{\beta}_{a,s}^R) - \lambda(\mathring{\Psi}_{i,n}^\top \beta_{a,s}^R)\right)^2 + o_p(1) = o_p(1),
\end{aligned}$$

where the first equality is due to Assumption 8(iii) and the second equality is by Assumption 8(vi) and the fact that

$$\frac{2}{n}\sum_{i=1}^{n}\left(\lambda(\mathring{\Psi}_{i,n}^\top\hat{\beta}_{a,s}^R)-\lambda(\mathring{\Psi}_{i,n}^\top\beta_{a,s}^R)\right)^2 \lesssim \frac{(\hat{\beta}_{a,s}^R-\beta_{a,s}^R)^\top}{n}\sum_{i=1}^{n}\mathring{\Psi}_{i,n}\mathring{\Psi}_{i,n}^\top(\hat{\beta}_{a,s}^R-\beta_{a,s}^R)\lesssim ||\hat{\beta}_{a,s}^R-\beta_{a,s}^R||_2^2=o_p(1),$$

where the first probability inequality is due to the fact that $\lambda(\cdot)$ is Lipschitz continuous with Lipschitz constant 1. Last, Assumption 2(iii) is implied by Assumption 1(vi) via Jensen's inequality.

N Proof of Theorem B.1

Define $\tilde{A}_i$ as the residual from the regression of A_i on X_i and $\{1\{S_i=s\}\}_{s\in\mathcal{S}}$. Then, we have

$$\hat{\tau}_4=\frac{\sum_{i\in[n]}\tilde{A}_iY_i}{\sum_{i\in[n]}\tilde{A}_iD_i}=\frac{\sum_{i\in[n]}\left(A_i-\pi(S_i)\right)Y_i+\sum_{i\in[n]}R_iY_i}{\sum_{i\in[n]}\left(A_i-\pi(S_i)\right)D_i+\sum_{i\in[n]}R_iD_i},$$

where $R_i=\tilde{A}_i-(A_i-\pi(S_i))$. We first suppose that

$$\frac{1}{n}\sum_{i\in[n]}R_iY_i=o_p(1)\quad\text{and}\quad\frac{1}{n}\sum_{i\in[n]}R_iD_i=o_p(1). \tag{N.1}$$

In addition, we note that

$$\frac{1}{n}\sum_{i\in[n]}(A_i-\pi(S_i))Y_i=\frac{1}{n}\sum_{i\in[n]}A_i(1-\pi(S_i))Y_i(D_i(1))-\frac{1}{n}\sum_{i\in[n]}(1-A_i)\pi(S_i)Y_i(D_i(0)).$$

For the first term on the RHS of the above display, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in[n]}A_i(1-\pi(S_i))Y_i(D_i(1))\\
&=\frac{1}{n}\sum_{i\in[n]}A_i(1-\pi(S_i))(Y_i(D_i(1))-\mathbb{E}(Y_i(D_i(1))|S_i))+\frac{1}{n}\sum_{i\in[n]}A_i(1-\pi(S_i))\mathbb{E}(Y_i(D_i(1))|S_i)\\
&=o_p(1)+\frac{1}{n}\sum_{i\in[n]}\pi(S_i)(1-\pi(S_i))\mathbb{E}(Y_i(D_i(1))|S_i)+\frac{1}{n}\sum_{s\in\mathcal{S}}B_n(s)(1-\pi(s))\mathbb{E}(Y_i(D_i(1))|S_i=s)\\
&=\mathbb{E}\pi(S_i)(1-\pi(S_i))\mathbb{E}(Y_i(D_i(1))|S_i)+o_p(1),
\end{aligned} \tag{N.2}$$

where the second equality is by conditional Chebyshev's inequality using the facts that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i\in[n]}A_i(1-\pi(S_i))(Y_i(D_i(1))-\mathbb{E}(Y_i(D_i(1))|S_i))\Big|\{A_i,S_i\}_{i\in[n]}\right]=0\\
&\mathbb{E}\left[\left(\frac{1}{n}\sum_{i\in[n]}A_i(1-\pi(S_i))(Y_i(D_i(1))-\mathbb{E}(Y_i(D_i(1))|S_i))\right)^2\Big|\{A_i,S_i\}_{i\in[n]}\right]
\end{aligned}$$

$$\leq \sum_{s\in\mathcal{S}} \frac{n_1(s)(1-\pi(s))^2\mathbb{E}(Y^2(D(1))|S_i=s)}{n^2} = o_p(1),$$

and the third equality is by Assumption 1(iv) and the usual LLN. For the same reason, we have

$$\frac{1}{n}\sum_{i\in[n]}(1-A_i)\pi(S_i)Y_i(D_i(0)) \xrightarrow{p} \mathbb{E}\pi(S_i)(1-\pi(S_i))\mathbb{E}(Y_i(D_i(0))|S_i),$$
$$\frac{1}{n}\sum_{i\in[n]}A_i(1-\pi(S_i))D_i(1) \xrightarrow{p} \mathbb{E}\pi(S_i)(1-\pi(S_i))\mathbb{E}(D_i(1)|S_i),$$
$$\frac{1}{n}\sum_{i\in[n]}(1-A_i)\pi(S_i)D_i(0) \xrightarrow{p} \mathbb{E}\pi(S_i)(1-\pi(S_i))\mathbb{E}D_i(0)|S_i),$$

and

$$\hat{\tau}_4 \xrightarrow{p} \frac{\mathbb{E}\pi(S_i)(1-\pi(S_i))(\mathbb{E}(Y_i(D_i(1))|S_i)-\mathbb{E}(Y_i(D_i(0))|S_i))}{\mathbb{E}\pi(S_i)(1-\pi(S_i))(\mathbb{E}(D_i(1)|S_i)-\mathbb{E}(D_i(0)|S_i))}.$$

Therefore, it is only left to show (N.1). Let $\mathbb{X}_i = (X_i^\top, \{1\{S_i=s\}\}_{s\in\mathcal{S}})^\top$, $\hat{\theta}$ be the OLS coefficient of regressing A_i on $\mathbb{X}_i$, and $\theta = (0_{d_x}^\top, \{\pi(s)\}_{s\in\mathcal{S}})^\top$, where d_x is the dimension of X_i. Then, we have $R_i = -\mathbb{X}_i^\top(\hat{\theta}-\theta)$. In order to show (N.1), it suffices to show $\hat{\theta} \xrightarrow{p} \theta$, or equivalently, $\frac{1}{n}\sum_{i\in[n]}\mathbb{X}_i(A_i-\pi(S_i)) \xrightarrow{p} 0$. We note that

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in[n]}\mathbb{X}_i(A_i-\pi(S_i)) = \frac{1}{n}\sum_{i\in[n]}(\mathbb{X}_i-\mathbb{E}(\mathbb{X}_i|S_i))(A_i-\pi(S_i)) + \frac{1}{n}\sum_{i\in[n]}\mathbb{E}(\mathbb{X}_i|S_i)(A_i-\pi(S_i))\\
&= \frac{1}{n}\sum_{i\in[n]}(\mathbb{X}_i-\mathbb{E}(\mathbb{X}_i|S_i))A_i(1-\pi(S_i)) - \frac{1}{n}\sum_{i\in[n]}(\mathbb{X}_i-\mathbb{E}(\mathbb{X}_i|S_i))(1-A_i)\pi(S_i) + \frac{1}{n}\sum_{s\in\mathcal{S}}\mathbb{E}(\mathbb{X}_i|S_i=s)B_n(s)\\
&= o_p(1), \qquad \text{(N.3)}
\end{aligned}$$

where the last equality holds following the similar argument in (N.2). This concludes the proof of the first statement.

For the second statement, let $\mathbb{X}_i = (X_i^\top, \{1\{S_i=s\}\}_{s\in\mathcal{S}})^\top$,

$$\hat{\theta} = \left(\frac{1}{n}\sum_{i\in[n]}\mathbb{X}_i\mathbb{X}_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i\in[n]}\mathbb{X}_iA_i\right),$$

$\widetilde{A}_i = A_i - \mathbb{X}_i^\top\hat{\theta}$, and $\theta = (0_{d_x}^\top, \pi, \cdots, \pi)^\top$. Then, we have

$$\sqrt{n}(\hat{\tau}_4-\tau) = \frac{\frac{1}{\sqrt{n}}\sum_{i\in[n]}\widetilde{A}_i(Y_i-D_i\tau)}{\frac{1}{n}\sum_{i\in[n]}\widetilde{A}_iD_i}.$$

By the same argument in the proof of the first statement of Theorem B.1, we have

$$\frac{1}{n}\sum_{i\in[n]}\widetilde{A}_i D_i \xrightarrow{p} \pi(1-\pi)\mathbb{E}(D(1)-D(0)).$$

Next, we turn to the numerator. We have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i\in[n]}\widetilde{A}_i(Y_i-D_i\tau) = \frac{1}{\sqrt{n}}\sum_{i\in[n]}(A_i-\mathbb{X}_i^\top\theta-\mathbb{X}_i^\top(\hat{\theta}-\theta))(Y_i-D_i\tau)\\
&=\frac{1}{\sqrt{n}}\sum_{i\in[n]}(A_i-\pi)(Y_i-D_i\tau)-\frac{1}{n}\sum_{i\in[n]}\mathbb{X}_i^\top(Y_i-D_i\tau)\left(\frac{1}{n}\sum_{i\in[n]}\mathbb{X}_i\mathbb{X}_i^\top\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i\in[n]}\mathbb{X}_i(A_i-\pi)\right).
\end{aligned}$$

where the second equality uses the facts that $\mathbb{X}_i^\top\theta=\pi$ and

$$\hat{\theta}-\theta=\left(\frac{1}{n}\sum_{i\in[n]}\mathbb{X}_i\mathbb{X}_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i\in[n]}\mathbb{X}_i(A_i-\mathbb{X}_i^\top\theta)\right)=\left(\frac{1}{n}\sum_{i\in[n]}\mathbb{X}_i\mathbb{X}_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i\in[n]}\mathbb{X}_i(A_i-\pi)\right).$$

We first consider the joint convergence of $\frac{1}{\sqrt{n}}\sum_{i\in[n]}(A_i-\pi)(Y_i-D_i\tau)$ and $\frac{1}{\sqrt{n}}\sum_{i\in[n]}\mathbb{X}_i(A_i-\pi)$. Let λ_1 be a scalar and $\lambda_2\in\Re^{d_x}$. Then, it suffices to consider the weak convergence of $\frac{1}{\sqrt{n}}\sum_{i\in[n]}(A_i-\pi)(\lambda_1(Y_i-D_i\tau)+\lambda_2^\top\mathbb{X}_i)$. Let $\rho_i=\lambda_1(Y_i-D_i\tau)+\lambda_2^\top\mathbb{X}_i$ and $\rho_i(a)=\lambda_1(Y_i(D_i(a))-D_i(a)\tau)+\lambda_2^\top\mathbb{X}_i$. Note that $\rho_i=A_i\rho_i(1)+(1-A_i)\rho_i(0)$. We have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i\in[n]}(A_i-\pi)\rho_i=\frac{1}{\sqrt{n}}\sum_{i\in[n]}\left[A_i(1-\pi)\rho_i(1)-(1-A_i)\pi\rho_i(0)\right]\\
&=\frac{1}{\sqrt{n}}\sum_{i\in[n]}\left[A_i(1-\pi)(\rho_i(1)-\mathbb{E}(\rho_i(1)|S_i))-(1-A_i)\pi(\rho_i(0)-\mathbb{E}(\rho_i(0)|S_i))\right]\\
&\quad+\frac{1}{\sqrt{n}}\sum_{i\in[n]}\left[A_i(1-\pi)\mathbb{E}(\rho_i(1)|S_i)-(1-A_i)\pi\mathbb{E}(\rho_i(0)|S_i)\right]\\
&=\frac{1}{\sqrt{n}}\sum_{i\in[n]}\left[A_i(1-\pi)(\rho_i(1)-\mathbb{E}(\rho_i(1)|S_i))-(1-A_i)\pi(\rho_i(0)-\mathbb{E}(\rho_i(0)|S_i))\right]\\
&\quad+\frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}B_n(s)\left[(1-\pi)\mathbb{E}(\rho_i(1)|S_i=s)+\pi\mathbb{E}(\rho_i(0)|S_i=s)\right]+\frac{\pi(1-\pi)}{\sqrt{n}}\sum_{i\in[n]}\mathbb{E}(\rho_i(1)-\rho_i(0)|S_i)\\
&=\frac{1}{\sqrt{n}}\sum_{i\in[n]}\left[A_i(1-\pi)(\rho_i(1)-\mathbb{E}(\rho_i(1)|S_i))-(1-A_i)\pi(\rho_i(0)-\mathbb{E}(\rho_i(0)|S_i))\right]\\
&\quad+\frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}B_n(s)\left[(1-\pi)\mathbb{E}(\rho_i(1)|S_i=s)+\pi\mathbb{E}(\rho_i(0)|S_i=s)\right]\\
&\quad+\frac{\pi(1-\pi)}{\sqrt{n}}\sum_{i\in[n]}\left(\mathbb{E}(\rho_i(1)-\rho_i(0)|S_i)-\mathbb{E}(\rho_i(1)-\rho_i(0))\right)\\
&\rightsquigarrow\mathcal{N}(0,\Sigma^2),
\end{aligned}\tag{N.4}$$

where

$$\Sigma^2 = (1-\pi)\pi\left[(1-\pi)\mathbb{E}\left[\rho_i(1)-\mathbb{E}(\rho_i(1)|S_i)\right]^2 + \pi\mathbb{E}\left[\rho_i(0)-\mathbb{E}(\rho_i(0)|S_i)\right]^2\right]$$
$$+\,\mathbb{E}\left[\gamma(S_i)\left(\mathbb{E}\left[(1-\pi)\rho_i(1)+\pi\rho_i(0)|S_i\right]\right)^2\right] + \pi^2(1-\pi)^2\mathbb{E}\left(\mathbb{E}\left[\rho_i(1)-\rho_i(0)|S_i\right]\right)^2,$$

the last convergence in distribution is by a similar argument in the proof of Bugni et al. (2018, Lemma B.2) and the fact that

$$\mathbb{E}(\rho_i(1)-\rho_i(0)) = \lambda_1\mathbb{E}(Y_i(D_i(1)) - Y_i(D_i(1)) - (D_i(1)-D_i(0))\tau) = 0.$$

Thus (N.4) implies both $\frac{1}{\sqrt{n}}\sum_{i\in[n]}(A_i-\pi)(Y_i-D_i\tau)$ and $\frac{1}{\sqrt{n}}\sum_{i\in[n]}\mathbb{X}_i(A_i-\pi)$ are $O_p(1)$. In addition, let $\hat{\lambda} = \left(\frac{1}{n}\sum_{i\in[n]}\mathbb{X}_i\mathbb{X}_i^\top\right)^{-1}\frac{1}{n}\sum_{i\in[n]}\mathbb{X}_i(Y_i-D_i\tau)$. We can show

$$\hat{\lambda} \xrightarrow{p} \lambda^* := \left(\mathbb{E}\mathbb{X}_i\mathbb{X}_i^\top\right)^{-1}\mathbb{E}\mathbb{X}_i\left[\pi(Y_i(D_i(1)) - D_i(1)\tau) + (1-\pi)(Y_i(D_i(0)) - D_i(0)\tau)\right].$$

Therefore, by letting $\lambda_1 = 1$ and $\lambda_2 = \lambda^*$, we have

$$\sqrt{n}(\hat{\tau}_4 - \tau) \rightsquigarrow \mathcal{N}(0, \sigma_4^2),$$

where

$$\sigma_4^2 = \frac{\sigma_{41}^2 + \sigma_{42}^2 + \sigma_{43}^2}{(\mathbb{E}(D_i(1)-D_i(0)))^2},$$
$$\sigma_{41}^2 = \frac{\mathbb{E}\left[Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^* - \mathbb{E}[Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*|S_i]\right]^2}{\pi}$$
$$+\,\frac{\mathbb{E}\left[Y_i(D_i(0)) - D_i(0)\tau - \mathbb{X}_i^\top\lambda^* - \mathbb{E}[Y_i(D_i(0)) - D_i(0)\tau - \mathbb{X}_i^\top\lambda^*|S_i]\right]^2}{1-\pi},$$
$$\sigma_{43}^2 = \mathbb{E}\left\{\gamma(S_i)\left(\mathbb{E}\left[\frac{Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*}{\pi} + \frac{Y_i(D_i(0)) - D_i(0)\tau - \mathbb{X}_i^\top\lambda^*}{1-\pi}\middle|S_i\right]\right)^2\right\},$$
$$\sigma_{42}^2 = \mathbb{E}\left[\mathbb{E}\left[Y(D(1)) - Y(D(0)) - (D(1)-D(0))\tau|S_i\right]\right]^2.$$

O Proof of Theorem C.1

Some part of the proof is due to Ansel et al. (2018) while some part of the proof is original. We first state the expression:

We now prove this. Let $U_i := (1, X_i^\top)^\top$ and $\hat{\lambda}_{as} := (\hat{\gamma}_{as}^b, \hat{\nu}_{as}^{b,\top})^\top$ for $a = 0, 1$ and $b = Y, D$.

Consider $\hat{\lambda}_{0s}^D$ as an example; note that

$$\hat{\lambda}_{0s}^D = \left(\frac{1}{n}\sum_{i=1}^n (1-A_i)1\{S_i = s\}U_iU_i^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^n (1-A_i)1\{S_i = s\}U_iD_i.$$

Consider the denominator of $\hat{\lambda}_{0s}^D$:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (1-A_i)1\{S_i = s\}U_iU_i^\top &= \frac{1}{n}\sum_{i=1}^n (\pi(s)-A_i)1\{S_i = s\}U_iU_i^\top + \frac{1}{n}\sum_{i=1}^n (1-\pi(s))1\{S_i = s\}U_iU_i^\top \\
= \frac{1}{n}\sum_{i=1}^n (\pi(s)-A_i)1\{S_i = s\}\left(U_iU_i^\top - \mathbb{E}[U_iU_i^\top|S_i]\right) &+ \frac{1}{n}\sum_{i=1}^n (\pi(s)-A_i)1\{S_i = s\}\mathbb{E}[U_iU_i^\top|S_i] \\
+ \frac{1}{n}\sum_{i=1}^n (1-\pi(s))1\{S_i = s\}U_iU_i^\top. & \qquad \text{(O.1)}
\end{aligned}$$

Consider the first term of (O.1). Note that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\pi(s)-A_i)1\{S_i = s\}\left(U_iU_i^\top - \mathbb{E}[U_iU_i^\top|S_i]\right)|A^{(n)}, S^{(n)}\right] = 0.$$

Invoking the conditional Chebyshev's inequality, we have, for any $a > 0$, $1 \le k, \ell \le \dim(U_i)$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (\pi(s)-A_i)1\{S_i = s\}\left(U_{ik}U_{i\ell} - \mathbb{E}[U_{ik}U_{i\ell}|S_i]\right)\right| \ge a|A^{(n)}, S^{(n)}\right) \\
&\le \frac{1}{a^2}\operatorname{var}\left(\frac{1}{n}\sum_{i=1}^n (\pi(s)-A_i)1\{S_i = s\}\left(U_{ik}U_{i\ell} - \mathbb{E}[U_{ik}U_{i\ell}|S_i]\right)|A^{(n)}, S^{(n)}\right) \\
&= \frac{\sum_{i,j\in[n]}(\pi(s)-A_i)(\pi(s)-A_j)1\{S_i = s\}1\{S_j = s\}\mathbb{E}\left[\left(U_{ik}U_{i\ell} - \mathbb{E}[U_{ik}U_{i\ell}|S_i]\right)\left(U_{jk}U_{j\ell} - \mathbb{E}[U_{jk}U_{j\ell}|S_j]\right)|A^{(n)}, S^{(n)}\right]}{a^2n^2} \\
&= \frac{\sum_{i\in[n]}(\pi(s)-A_i)^2 1\{S_i = s\}\mathbb{E}\left[\left(U_{ik}U_{i\ell} - \mathbb{E}[U_{ik}U_{i\ell}|S_i]\right)^2|A^{(n)}, S^{(n)}\right]}{a^2n^2} \\
&\le \frac{\sum_{i\in[n]}(\pi(s)-A_i)^2 1\{S_i = s\}\mathbb{E}\left[U_{ik}^2U_{i\ell}^2|S_i\right]}{a^2n^2} \le \frac{\sum_{i\in[n]}\mathbb{E}\left[U_{ik}^2U_{i\ell}^2|S_i = s\right]}{a^2n^2} = o(1) \qquad \text{(O.2)}
\end{aligned}$$

where the second equality is due to

$$\begin{aligned}
&\mathbb{E}\left[\left(U_{ik}U_{i\ell} - \mathbb{E}[U_{ik}U_{i\ell}|S_i]\right)\left(U_{jk}U_{j\ell} - \mathbb{E}[U_{jk}U_{j\ell}|S_j]\right)|A^{(n)}, S^{(n)}\right] \\
&= \mathbb{E}\left[\left(U_{ik}U_{i\ell} - \mathbb{E}[U_{ik}U_{i\ell}|S_i]\right)\left(U_{jk}U_{j\ell} - \mathbb{E}[U_{jk}U_{j\ell}|S_j]\right)|S^{(n)}\right] \\
&= \mathbb{E}\left[U_{ik}U_{i\ell} - \mathbb{E}[U_{ik}U_{i\ell}|S_i]|S^{(n)}\right]\mathbb{E}\left[U_{jk}U_{j\ell} - \mathbb{E}[U_{jk}U_{j\ell}|S_j]|S^{(n)}\right] \\
&= \mathbb{E}\left[U_{ik}U_{i\ell} - \mathbb{E}[U_{ik}U_{i\ell}|S_i]|S_i\right]\mathbb{E}\left[U_{jk}U_{j\ell} - \mathbb{E}[U_{jk}U_{j\ell}|S_j]|S_j\right] = 0
\end{aligned}$$

for $i \ne j$. From (O.2), we deduce that the first term of (O.1) is $o_p(1)$. Consider the second term of

(O.1).

$$\frac{1}{n}\sum_{i=1}^{n}(\pi(s)-A_i)1\{S_i=s\}\mathbb{E}[U_iU_i^\top|S_i]=\mathbb{E}[UU^\top|S=s]\frac{1}{n}\sum_{i=1}^{n}(\pi(s)-A_i)1\{S_i=s\}$$
$$=\mathbb{E}[UU^\top|S=s]\frac{1}{n}B_n(s)=o_p(1).$$

Consider the third term of (O.1).

$$\frac{1}{n}\sum_{i=1}^{n}(1-\pi(s))1\{S_i=s\}U_iU_i^\top=(1-\pi(s))\frac{n(s)}{n}\frac{1}{n(s)}\sum_{i=1}^{n}1\{S_i=s\}U_iU_i^\top$$
$$\xrightarrow{p}(1-\pi(s))p(s)\mathbb{E}[UU^\top|S=s].$$

We hence have

$$\frac{1}{n}\sum_{i=1}^{n}(1-A_i)1\{S_i=s\}U_iU_i^\top\xrightarrow{p}(1-\pi(s))p(s)\mathbb{E}[UU^\top|S=s].$$

Similarly, we have

$$\frac{1}{n}\sum_{i=1}^{n}(1-A_i)1\{S_i=s\}U_iD_i\xrightarrow{p}(1-\pi(s))\hat{p}(s)\mathbb{E}[UD(0)|S=s]$$
$$\hat{\lambda}_{0s}^D\xrightarrow{p}\left(\mathbb{E}[UU^\top|S=s]\right)^{-1}\mathbb{E}[UD(0)|S=s]$$
$$\hat{\lambda}_{1s}^D\xrightarrow{p}\left(\mathbb{E}[UU^\top|S=s]\right)^{-1}\mathbb{E}[UD(1)|S=s].$$

Thus, we have

$$\sum_{s\in\mathcal{S}}\hat{p}(s)(\hat{\gamma}_{1s}^D-\hat{\gamma}_{0s}^D+(\hat{\nu}_{1s}^D-\hat{\nu}_{0s}^D)^\top\bar{X}_s)=\sum_{s\in\mathcal{S}}(\hat{\lambda}_{1s}^D-\hat{\lambda}_{0s}^D)^\top\begin{pmatrix}\frac{1}{n}\sum_{i\in[n]}1\{S_i=s\}\\ \frac{1}{n}\sum_{i\in[n]}X_i1\{S_i=s\}\end{pmatrix}$$
$$=\sum_{s\in\mathcal{S}}\frac{n(s)}{n}\frac{1}{n(s)}\sum_{i\in[n]}1\{S_i=s\}U_i^\top(\hat{\lambda}_{1s}^D-\hat{\lambda}_{0s}^D)$$
$$\xrightarrow{p}\sum_{s\in\mathcal{S}}p(s)\mathbb{E}[U^\top|S=s]\left(\mathbb{E}[UU^\top|S=s]\right)^{-1}\mathbb{E}\left[U\left(D(1)-D(0)\right)|S=s\right]$$
$$=\sum_{s\in\mathcal{S}}p(s)\mathbb{E}\left[D(1)-D(0)|S=s\right]=\mathbb{E}\left[D(1)-D(0)\right]$$

where the second last equality is due to $\mathbb{E}[U^\top|S=s]\left(\mathbb{E}[UU^\top|S=s]\right)^{-1}=(1,0,\ldots,0)$ (Ansel et al. (2018) p290). Thus, the denominator of $\sqrt{n}(\hat{\tau}_S-\tau)$ converges in probability to $\mathbb{E}[D(1)-D(0)]$.

We now consider the numerator of $\sqrt{n}(\hat{\tau}_S - \tau)$. Relying on a similar argument, we have

$$\hat{\lambda}^Y_{1s} = \left(\frac{1}{n}\sum_{i=1}^n A_i 1\{S_i = s\}U_i U_i^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^n A_i 1\{S_i = s\}U_i Y_i(D_i(1))$$
$$\xrightarrow{p} \left(\mathbb{E}[UU^\top|S=s]\right)^{-1}\mathbb{E}[UY(D(1))|S=s]$$
$$\hat{\lambda}^Y_{0s} = \left(\frac{1}{n}\sum_{i=1}^n (1-A_i) 1\{S_i = s\}U_i U_i^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^n (1-A_i) 1\{S_i = s\}U_i Y_i(D_i(0))$$
$$\xrightarrow{p} \left(\mathbb{E}[UU^\top|S=s]\right)^{-1}\mathbb{E}[UY(D(0))|S=s]$$

$$\hat{\eta}_{1s} := \hat{\lambda}^Y_{1s} - \tau\hat{\lambda}^D_{1s} \xrightarrow{p} \left(\mathbb{E}[UU^\top|S=s]\right)^{-1}\mathbb{E}\left[U\left[Y(D(1)) - \tau D(1)\right]|S=s\right] =: \eta_{1s}$$
$$\hat{\eta}_{0s} := \hat{\lambda}^Y_{0s} - \tau\hat{\lambda}^D_{0s} \xrightarrow{p} \left(\mathbb{E}[UU^\top|S=s]\right)^{-1}\mathbb{E}\left[U\left[Y(D(0)) - \tau D(0)\right]|S=s\right] =: \eta_{0s}.$$

The numerator of $\sqrt{n}(\hat{\tau}_S - \tau)$ could be written as

$$\sqrt{n}\sum_{s\in\mathcal{S}}\hat{p}(s)(\hat{\gamma}^Y_{1s} - \hat{\gamma}^Y_{0s} + (\hat{\nu}^Y_{1s} - \hat{\nu}^Y_{0s})^\top \bar{X}_s) - \sqrt{n}\sum_{s\in\mathcal{S}}\hat{p}(s)(\hat{\gamma}^D_{1s} - \hat{\gamma}^D_{0s} + (\hat{\nu}^D_{1s} - \hat{\nu}^D_{0s})^\top \bar{X}_s)\tau$$
$$= \sqrt{n}\sum_{s\in\mathcal{S}}\hat{p}(s)\frac{1}{n(s)}\sum_{i\in[n]} 1\{S_i = s\}U_i^\top\left[\hat{\lambda}^Y_{1s} - \tau\hat{\lambda}^D_{1s} - (\hat{\lambda}^Y_{0s} - \tau\hat{\lambda}^D_{0s})\right]$$
$$= \sqrt{n}\sum_{s\in\mathcal{S}}\hat{p}(s)\bar{U}_s^\top(\hat{\eta}_{1s} - \eta_{1s}) - \sqrt{n}\sum_{s\in\mathcal{S}}\hat{p}(s)\bar{U}_s^\top(\hat{\eta}_{0s} - \eta_{0s}) + \sqrt{n}\sum_{s\in\mathcal{S}}\hat{p}(s)\bar{U}_s^\top(\eta_{1s} - \eta_{0s}) \tag{O.3}$$

where $\bar{U}_s := \frac{1}{n(s)}\sum_{i\in[n]} 1\{S_i = s\}U_i \xrightarrow{p} \mathbb{E}[U|S=s]$. Consider the first term of (O.3).

$$\sqrt{n}\sum_{s\in\mathcal{S}}\hat{p}(s)\bar{U}_s^\top(\hat{\eta}_{1s} - \eta_{1s})$$
$$= \sum_{s\in\mathcal{S}}\hat{p}(s)\bar{U}_s^\top\left(\frac{1}{n}\sum_{i=1}^n A_i 1\{S_i = s\}U_i U_i^\top\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n A_i 1\{S_i = s\}U_i\left[Y_i(D_i(1)) - \tau D_i(1) - U_i^\top\eta_{1s}\right]$$
$$= \sum_{s\in\mathcal{S}}\hat{p}(s)\mathbb{E}[U^\top|S=s]\left(\pi(s)\hat{p}(s)\mathbb{E}[UU^\top|S=s]\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n A_i 1\{S_i = s\}U_i\left[Y_i(D_i(1)) - \tau D_i(1) - U_i^\top\eta_{1s}\right] + o_p(1)$$
$$= \sum_{s\in\mathcal{S}}\frac{1}{\pi(s)}\mathbb{E}[U^\top|S=s]\left(\mathbb{E}[UU^\top|S=s]\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n A_i 1\{S_i = s\}U_i\left[Y_i(D_i(1)) - \tau D_i(1) - U_i^\top\eta_{1s}\right] + o_p(1)$$
$$= \sum_{s\in\mathcal{S}}\frac{1}{\pi(s)}\frac{1}{\sqrt{n}}\sum_{i=1}^n A_i 1\{S_i = s\}\left[Y_i(D_i(1)) - \tau D_i(1) - U_i^\top\eta_{1s}\right] + o_p(1) \tag{O.4}$$

where the second equality is based on the conjecture that

$$n^{-1/2}\sum_{i=1}^n A_i 1\{S_i = s\}U_i\left[Y_i(D_i(1)) - \tau D_i(1) - U_i^\top\eta_{1s}\right] = O_p(1),$$

which we will prove shortly, and the last equality is due to $\mathbb{E}[U^\top|S=s]\left(\mathbb{E}[UU^\top|S=s]\right)^{-1} = (1,0,\ldots,0)$ (Ansel et al. (2018) p290). Likewise, the second term of (O.3)

$$\begin{aligned}&\sqrt{n}\sum_{s\in\mathcal{S}}\hat{p}(s)\bar{U}_s^\top(\hat{\eta}_{0s}-\eta_{0s})\\&=\sum_{s\in\mathcal{S}}\frac{1}{1-\pi(s)}\frac{1}{\sqrt{n}}\sum_{i=1}^n(1-A_i)1\{S_i=s\}\left[Y_i(D_i(0))-\tau D_i(0)-U_i^\top\eta_{0s}\right]+o_p(1).\end{aligned}\tag{O.5}$$

Note that

$$\eta_{as}=\begin{pmatrix}\mathbb{E}\left[Y(D(a))-\tau D(a)|S=s\right]-\mathbb{E}[X^\top\nu_{as}^{YD}|S=s]\\ \nu_{as}^{YD}\end{pmatrix}$$

for $a=0,1$ via the Frisch-Waugh Theorem. Hence

$$U_i^\top\eta_{as}=\mathbb{E}\left[Y(D(a))-\tau D(a)|S=s\right]+X_i^\top\nu_{as}^{YD}-\mathbb{E}[X^\top\nu_{as}^{YD}|S=s].\tag{O.6}$$

Substituting (O.4), (O.5) and (O.6) into (O.3), we could write the numerator of $\sqrt{n}(\hat{\tau}_S-\tau)$ as

$$\begin{aligned}&\sum_{s\in\mathcal{S}}\frac{1}{\pi(s)}\frac{1}{\sqrt{n}}\sum_{i=1}^n A_i1\{S_i=s\}\left[Y_i(D_i(1))-\tau D_i(1)-U_i^\top\eta_{1s}\right]\\&\quad-\sum_{s\in\mathcal{S}}\frac{1}{1-\pi(s)}\frac{1}{\sqrt{n}}\sum_{i=1}^n(1-A_i)1\{S_i=s\}\left[Y_i(D_i(0))-\tau D_i(0)-U_i^\top\eta_{0s}\right]\qquad\qquad(\text{O.7})\\&\quad+\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\sum_{i\in[n]}1\{S_i=s\}U_i^\top(\eta_{1s}-\eta_{0s})+o_p(1)\\&=\sum_{s\in\mathcal{S}}\frac{1}{\pi(s)}\frac{1}{\sqrt{n}}\sum_{i=1}^n A_i1\{S_i=s\}\Big[Y_i(D_i(1))-\tau D_i(1)-\mathbb{E}\left[Y(D(1))-\tau D(1)|S=s\right]-\left(X_i^\top\nu_{1s}^{YD}-\mathbb{E}[X^\top\nu_{1s}^{YD}|S=s]\right)\Big]\\&\quad-\sum_{s\in\mathcal{S}}\frac{1}{1-\pi(s)}\frac{1}{\sqrt{n}}\sum_{i=1}^n(1-A_i)1\{S_i=s\}\Big[Y_i(D_i(0))-\tau D_i(0)-\mathbb{E}\left[Y(D(0))-\tau D(0)|S=s\right]-\left(X_i^\top\nu_{0s}^{YD}-\mathbb{E}[X^\top\nu_{0s}^{YD}|S=s]\right)\Big]\\&\quad+\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\sum_{i\in[n]}1\{S_i=s\}\mathbb{E}\Big[Y(D(1))-Y(D(0))-\tau(D(1)-D(0))|S=s\Big]\\&\quad+\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\sum_{i\in[n]}1\{S_i=s\}\Big(X_i^\top(\nu_{1s}^{YD}-\nu_{0s}^{YD})-\mathbb{E}[X^\top(\nu_{1s}^{YD}-\nu_{0s}^{YD})|S=s]\Big)+o_p(1)\\&=\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^n A_i1\{S_i=s\}\left[\frac{Y_i(D_i(1))-\tau D_i(1)-X_i^\top\nu_{1s}^{YD}-\mathbb{E}\left[Y(D(1))-\tau D(1)-X^\top\nu_{1s}^{YD}|S=s\right]}{\pi(s)}\right]\\&\quad+\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\sum_{i\in[n]}A_i1\{S_i=s\}\Big(X_i^\top(\nu_{1s}^{YD}-\nu_{0s}^{YD})-\mathbb{E}[X^\top(\nu_{1s}^{YD}-\nu_{0s}^{YD})|S=s]\Big)\\&\quad-\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^n(1-A_i)1\{S_i=s\}\left[\frac{Y_i(D_i(0))-\tau D_i(0)-X_i^\top\nu_{0s}^{YD}-\mathbb{E}\left[Y(D(0))-\tau D(0)-X^\top\nu_{0s}^{YD}|S=s\right]}{1-\pi(s)}\right]\\&\quad+\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\sum_{i\in[n]}(1-A_i)1\{S_i=s\}\Big(X_i^\top(\nu_{1s}^{YD}-\nu_{0s}^{YD})-\mathbb{E}[X^\top(\nu_{1s}^{YD}-\nu_{0s}^{YD})|S=s]\Big)\\&\quad+\frac{1}{\sqrt{n}}\sum_{i\in[n]}\mathbb{E}\Big[Y(D(1))-Y(D(0))-\tau(D(1)-D(0))|S\Big]+o_p(1).\qquad\qquad(\text{O.8})\end{aligned}$$

Define

$$\rho_i(1) := \frac{Y_i(D_i(1)) - \tau D_i(1) - X_i^\top \nu_{1s}^{YD}}{\pi(s)} + X_i^\top(\nu_{1s}^{YD} - \nu_{0s}^{YD})$$
$$\rho_i(0) := \frac{Y_i(D_i(0)) - \tau D_i(0) - X_i^\top \nu_{0s}^{YD}}{1-\pi(s)} - X_i^\top(\nu_{1s}^{YD} - \nu_{0s}^{YD}).$$

Then the first four terms of (O.8) could be written compactly as

$$R_{n,1} := \sum_{s\in\mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} A_i 1\{S_i = s\} \left[\rho_i(1) - \mathbb{E}[\rho_i(1)|S_i = s]\right]$$
$$- \sum_{s\in\mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1-A_i) 1\{S_i = s\} \left[\rho_i(0) - \mathbb{E}[\rho_i(0)|S_i = s]\right].$$

Define $R_{n,2} := \frac{1}{\sqrt{n}} \sum_{i\in[n]} \mathbb{E}\left[Y(D(1)) - Y(D(0)) - \tau(D(1) - D(0))|S\right]$. To establish the asymptotic distribution of (O.8), we first argue that

$$(R_{n,1}, R_{n,2}) \stackrel{d}{=} (R_{n,1}^*, R_{n,2}) + o_p(1)$$

for a random variable $R_{n,1}^*$ that satisfies $R_{n,1}^* \perp\!\!\!\perp R_{n,2}$. Conditional on $\{S^{(n)}, A^{(n)}\}$, the distribution of $R_{n,1}$ is the same as the distribution of the same quantity where units are ordered by strata and then ordered by $A_i = 1$ first and $A_i = 0$ second within strata. To this end, define $N(s) := \sum_{i=1}^n 1\{S_i < s\}$ and $F(s) := \mathbb{P}(S_i < s)$. Furthermore, independently for each $s \in \mathcal{S}$ and independently of $\{S^{(n)}, A^{(n)}\}$, let $\left\{Y_i(1)^s, Y_i(0)^s, D_i(1)^s, D_i(0)^s, X_i^s : 1 \le i \le n\right\}$ be i.i.d. over i with distribution equal to that of $(Y(1), Y(0), D(1), D(0), X)|S = s$. Define

$$\tilde{\rho}_i(a) := \rho_i(a) - \mathbb{E}[\rho_i(a)|S_i = s], \qquad \tilde{\rho}_i^s(a) := \rho_i^s(a) - \mathbb{E}[\rho_i^s(a)|S_i = s],$$

where

$$\rho_i^s(1) := \frac{Y_i^s(D_i^s(1)) - \tau D_i^s(1) - X_i^{s,\top} \nu_{1s}^{YD}}{\pi(s)} + X_i^{s,\top}(\nu_{1s}^{YD} - \nu_{0s}^{YD})$$
$$\rho_i^s(0) := \frac{Y_i^s(D_i^s(0)) - \tau D_i^s(0) - X_i^{s,\top} \nu_{0s}^{YD}}{1-\pi(s)} - X_i^{s,\top}(\nu_{1s}^{YD} - \nu_{0s}^{YD}).$$

Then we have

$$R_{n,1} := \sum_{s\in\mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1\{S_i = s\} \left[A_i \tilde{\rho}_i(1) - (1-A_i)\tilde{\rho}_i(0)\right].$$

Define

$$\tilde{R}_{n,1} := \sum_{s\in\mathcal{S}} \left[\frac{1}{\sqrt{n}} \sum_{i=n\frac{N(s)}{n}+1}^{n\left(\frac{N(s)}{n}+\frac{n_1(s)}{n}\right)} \tilde{\rho}_i^s(1) - \frac{1}{\sqrt{n}} \sum_{n\left(\frac{N(s)}{n}+\frac{n_1(s)}{n}\right)+1}^{n\left(\frac{N(s)}{n}+\frac{n(s)}{n}\right)} \tilde{\rho}_i^s(0) \right]$$

$$R^*_{n,1} := \sum_{s\in\mathcal{S}} \left[\frac{1}{\sqrt{n}} \sum_{i=\lfloor nF(s)\rfloor+1}^{\lfloor n(F(s)+\pi(s)p(s))\rfloor} \tilde{\rho}_i^s(1) - \frac{1}{\sqrt{n}} \sum_{i=\lfloor n(F(s)+\pi(s)p(s))\rfloor+1}^{\lfloor n(F(s)+p(s))\rfloor} \tilde{\rho}_i^s(0) \right].$$

Thus $R_{n,1}|S^{(n)}, A^{(n)} \stackrel{d}{=} \tilde{R}_{n,1}|S^{(n)}, A^{(n)}$ (and as a by-product $R_{n,1} \stackrel{d}{=} \tilde{R}_{n,1}$). Since $R_{n,2}$ is a function of $\{S^{(n)}, A^{(n)}\}$, we have, arguing along the line of a joint distribution being the product of a conditional distribution and a marginal distribution, $(R_{n,1}, R_{n,2}) \stackrel{d}{=} (\tilde{R}_{n,1}, R_{n,2})$. Define the following partial sum process

$$g_n(u) := \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nu\rfloor} \tilde{\rho}_i^s(1).$$

Under our assumptions, $g_n(u)$ converges weakly to a suitably scaled Brownian motion. Next, by elementary properties of Brownian motion, we have that

$$g_n\left(F(s)+\pi(s)p(s)\right) - g_n\left(F(s)\right) = \frac{1}{\sqrt{n}} \sum_{i=\lfloor nF(s)\rfloor+1}^{\lfloor n(F(s)+\pi(s)p(s))\rfloor} \tilde{\rho}_i^s(1) \rightsquigarrow \mathcal{N}\left(0, \pi(s)p(s)\operatorname{var}\left(\rho(1)|S=s\right)\right). \tag{O.9}$$

Furthermore, since

$$\left(\frac{N(s)}{n}, \frac{n_1(s)}{n}\right) \xrightarrow{p} \left(F(s), \pi(s)p(s)\right),$$

it follows that

$$g_n\left(\frac{N(s)+n_1(s)}{n}\right) - g_n\left(\frac{N(s)}{n}\right) - \left[g_n\left(F(s)+\pi(s)p(s)\right) - g_n\left(F(s)\right)\right] \xrightarrow{p} 0 \tag{O.10}$$

where the convergence follows from the stochastic equicontinuity of the partial sum process. Using (O.9) and (O.10), we have:

$$(R_{n,1}, R_{n,2}) \stackrel{d}{=} (\tilde{R}_{n,1}, R_{n,2}) = (R^*_{n,1}, R_{n,2}) + o_p(1) \tag{O.11}$$

$$\begin{aligned} R^*_{n,1} &\rightsquigarrow \mathcal{N}\left(0, \sum_{s\in\mathcal{S}} \left[\pi(s)p(s)\operatorname{var}\left(\rho(1)|S=s\right) + \left[1-\pi(s)\right]p(s)\operatorname{var}\left(\rho(0)|S=s\right)\right]\right) \\ &= \mathcal{N}\left(0, \mathbb{E}\left[\pi(S)\left(\rho(1)-\mathbb{E}[\rho(1)|S=s]\right)^2 + (1-\pi(S))\left(\rho(0)-\mathbb{E}[\rho(0)|S=s]\right)^2\right]\right) \\ &=: \zeta_1 \end{aligned}$$

where the convergence in distribution is due to an analogous argument for $\tilde{\rho}_i^s(0)$ and the independence of $\{Y_i(1)^s, Y_i(0)^s, D_i(1)^s, D_i(0)^s, X_i^s : 1 \le i \le n, s \in \mathcal{S}\}$ across both i and s. Moreover, since $R_{n,1}^*$ is a function of $\{Y_i(1)^s, Y_i(0)^s, D_i(1)^s, D_i(0)^s, X_i^s : 1 \le i \le n, s \in \mathcal{S}\} \perp\!\!\!\perp S^{(n)}, A^{(n)}$, and $R_{n,2}$ is a function of $\{S^{(n)}, A^{(n)}\}$, we see that $R_{n,1}^* \perp\!\!\!\perp R_{n,2}$. Thus (O.11) implies

$$(R_{n,1}, R_{n,2}) \stackrel{d}{=} (R_{n,1}^*, R_{n,2}) + o_p(1) \rightsquigarrow (\zeta_1, \zeta_2)$$

where ζ_1 and ζ_2 are independent, with

$$\zeta_2 := \mathcal{N}\left(0, \mathbb{E}\left[\left(\mathbb{E}\left[Y(D(1)) - Y(D(0)) - \tau(D(1) - D(0))|S\right]\right)^2\right]\right).$$

We hence show that the asymptotic distribution of the numerator of $\sqrt{n}(\hat{\tau}_S - \tau)$ is $\zeta_1 + \zeta_2$. This completes the proof.

P Technical Lemmas Used in the Proof of Theorem 5.1

Lemma P.1. *Suppose assumptions in Theorem 5.1 hold. Then, we have*

$$R_{n,1} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(1 - \frac{1}{\pi(S_i)}\right) A_i \tilde{\mu}^Y(1, S_i, X_i) + \frac{1}{\sqrt{n}} \sum_{i=1}^n (1 - A_i) \tilde{\mu}^Y(1, S_i, X_i) + o_p(1),$$
$$R_{n,2} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\frac{1}{1 - \pi(S_i)} - 1\right) (1 - A_i) \tilde{\mu}^Y(0, S_i, X_i) - \frac{1}{\sqrt{n}} \sum_{i=1}^n A_i \tilde{\mu}^Y(0, S_i, X_i) + o_p(1),$$
$$R_{n,3} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{\pi(S_i)} \tilde{W}_i A_i - \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1 - A_i}{1 - \pi(S_i)} \tilde{Z}_i + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\mathbb{E}[W_i - Z_i|S_i] - \mathbb{E}[W_i - Z_i]\right) + o_p(1),$$

where for $a = 0, 1$,

$$\tilde{\mu}^Y(a, S_i, X_i) := \overline{\mu}^Y(a, S_i, X_i) - \overline{\mu}^Y(a, S_i), \quad \overline{\mu}^Y(a, S_i) := \mathbb{E}\left[\overline{\mu}^Y(a, S_i, X_i)|S_i\right],$$
$$W_i := Y_i(1)D_i(1) + Y_i(0)(1 - D_i(1)), \quad Z_i := Y_i(1)D_i(0) + Y_i(0)(1 - D_i(0)),$$
$$\tilde{W}_i := W_i - \mathbb{E}[W_i|S_i], \quad \textit{and} \quad \tilde{Z}_i := Z_i - \mathbb{E}[Z_i|S_i].$$

Proof. We have

$$\begin{aligned}
R_{n,1} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\hat{\mu}^Y(1, S_i, X_i) - \frac{A_i \hat{\mu}^Y(1, S_i, X_i)}{\hat{\pi}(S_i)}\right] \\
&= -\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{A_i - \hat{\pi}(S_i)}{\hat{\pi}(S_i)} \hat{\mu}^Y(1, S_i, X_i) \\
&= -\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{A_i - \hat{\pi}(S_i)}{\hat{\pi}(S_i)} \left[\hat{\mu}^Y(1, S_i, X_i) - \overline{\mu}^Y(1, S_i, X_i) + \overline{\mu}^Y(1, S_i, X_i)\right]
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{A_i-\hat{\pi}(S_i)}{\hat{\pi}(S_i)}\Delta^Y(1,S_i,X_i) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{A_i}{\hat{\pi}(S_i)}\overline{\mu}^Y(1,S_i,X_i) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\overline{\mu}^Y(1,S_i,X_i) \\
&= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{A_i-\hat{\pi}(S_i)}{\hat{\pi}(S_i)}\Delta^Y(1,S_i,X_i) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{A_i}{\hat{\pi}(S_i)}\tilde{\mu}^Y(1,S_i,X_i) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\mu}^Y(1,S_i,X_i),
\end{aligned} \tag{P.1}
$$

where the last equality is due to

$$
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{A_i}{\hat{\pi}(S_i)}\overline{\mu}^Y(1,S_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\overline{\mu}^Y(1,S_i).
$$

Consider the first term of (P.1).

$$
\begin{aligned}
&\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{A_i-\hat{\pi}(S_i)}{\hat{\pi}(S_i)}\Delta^Y(1,S_i,X_i)\right| = \left|\frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{i=1}^{n}\frac{A_i-\hat{\pi}(s)}{\hat{\pi}(s)}\Delta^Y(1,s,X_i)1\{S_i=s\}\right| \\
&= \frac{1}{\sqrt{n}}\left|\sum_{s\in\mathcal{S}}\frac{1}{\hat{\pi}(s)}\sum_{i=1}^{n}A_i\Delta^Y(1,s,X_i)1\{S_i=s\} - \sum_{s\in\mathcal{S}}\sum_{i=1}^{n}\Delta^Y(1,s,X_i)1\{S_i=s\}\right| \\
&= \frac{1}{\sqrt{n}}\left|\sum_{s\in\mathcal{S}}\sum_{i\in I_1(s)}\Delta^Y(1,s,X_i)\frac{n(s)}{n_1(s)} - \sum_{s\in\mathcal{S}}\sum_{i\in I_0(s)\cup I_1(s)}\Delta^Y(1,s,X_i)\right| \\
&= \frac{1}{\sqrt{n}}\left|\sum_{s\in\mathcal{S}}\sum_{i\in I_1(s)}\Delta^Y(1,s,X_i)\frac{n_0(s)}{n_1(s)} - \sum_{s\in\mathcal{S}}\sum_{i\in I_0(s)}\Delta^Y(1,s,X_i)\right| \\
&= \frac{1}{\sqrt{n}}\left|\sum_{s\in\mathcal{S}}n_0(s)\left[\frac{\sum_{i\in I_1(s)}\Delta^Y(1,s,X_i)}{n_1(s)} - \frac{\sum_{i\in I_0(s)}\Delta^Y(1,s,X_i)}{n_0(s)}\right]\right| \\
&\leq \frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}n_0(s)\left|\frac{\sum_{i\in I_1(s)}\Delta^Y(1,s,X_i)}{n_1(s)} - \frac{\sum_{i\in I_0(s)}\Delta^Y(1,s,X_i)}{n_0(s)}\right| = o_p(1)
\end{aligned}
$$

where the last equality is due to Assumption 2. Thus

$$
\begin{aligned}
R_{n,1} &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{A_i}{\hat{\pi}(S_i)}\tilde{\mu}^Y(1,S_i,X_i) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\mu}^Y(1,S_i,X_i) + o_p(1) \\
&= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{A_i}{\hat{\pi}(S_i)}\tilde{\mu}^Y(1,S_i,X_i) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i\tilde{\mu}^Y(1,S_i,X_i) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-A_i)\tilde{\mu}^Y(1,S_i,X_i) + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1-\frac{1}{\hat{\pi}(S_i)}\right)A_i\tilde{\mu}^Y(1,S_i,X_i) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-A_i)\tilde{\mu}^Y(1,S_i,X_i) + o_p(1).
\end{aligned} \tag{P.2}
$$

In addition, we note that

$$
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1-\frac{1}{\hat{\pi}(S_i)}\right)A_i\tilde{\mu}^Y(1,S_i,X_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1-\frac{1}{\pi(S_i)}\right)A_i\tilde{\mu}^Y(1,S_i,X_i)
$$

$$+ \sum_{s \in \mathcal{S}} \left(\frac{1}{\pi(s)} - \frac{1}{\hat{\pi}(s)} \right) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} A_i \tilde{\mu}^Y(1, s, X_i) 1\{S_i = s\}.$$

Note that under Assumption 1(i), conditional on $\{S^{(n)}, A^{(n)}\}$, the distribution of

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} A_i \tilde{\mu}^Y(1, s, X_i) 1\{S_i = s\}$$

is the same as the distribution of the same quantity where units are ordered by strata and then ordered by $A_i = 1$ first and $A_i = 0$ second within strata. To this end, define $N(s) := \sum_{i=1}^{n} 1\{S_i < s\}$ and $F(s) := \mathbb{P}(S_i < s)$. Furthermore, independently for each $s \in \mathcal{S}$ and independently of $\{S^{(n)}, A^{(n)}\}$, let $\left\{X_i^s : 1 \leq i \leq n\right\}$ be i.i.d with marginal distribution equal to the distribution of $X_i|S = s$. Define

$$\tilde{\mu}^b(a, s, X_i^s) := \overline{\mu}^b(a, s, X_i^s) - \mathbb{E}\left[\overline{\mu}^b(a, s, X_i^s)|S_i = s\right]$$

Then, we have, for $s \in \mathcal{S}$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} A_i \tilde{\mu}^Y(1, s, X_i) 1\{S_i = s\} \overset{d}{=} \frac{1}{\sqrt{n}} \sum_{i=N(s)+1}^{N(s)+n_1(s)} \tilde{\mu}^Y(1, s, X_i^s).$$

In addition, we have

$$\begin{aligned} \mathbb{E}\left[\left(\frac{1}{\sqrt{n}} \sum_{i=N(s)+1}^{N(s)+n_1(s)} \tilde{\mu}^Y(1, s, X_i^s)\right)^2 \middle| S^{(n)}, A^{(n)}\right] &= \frac{n_1(s)}{n} \mathbb{E}\left[\tilde{\mu}^{Y,2}(a, s, X_i^s)|S^{(n)}\right] \\ &\leq \frac{n_1(s)}{n} E\left[\overline{\mu}^{Y,2}(a, s, X_i)|S_i = s\right] = O_p(1), \end{aligned}$$

which implies

$$\max_{s \in \mathcal{S}} \left| \frac{1}{\sqrt{n}} \sum_{i=N(s)+1}^{N(s)+n_1(s)} \tilde{\mu}^Y(1, s, X_i^s) \right| = O_p(1).$$

Combining this with the facts that $\max_{s \in \mathcal{S}} |\hat{\pi}(s) - \pi(s)| = o_p(1)$ and $\min_{s \in \mathcal{S}} \pi(s) > c > 0$ for some constant c, we have

$$\sum_{s \in \mathcal{S}} \left(\frac{1}{\pi(s)} - \frac{1}{\hat{\pi}(s)} \right) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} A_i \tilde{\mu}^Y(1, s, X_i) 1\{S_i = s\} = o_p(1)$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(1 - \frac{1}{\hat{\pi}(S_i)}\right) A_i \tilde{\mu}^Y(1, S_i, X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(1 - \frac{1}{\pi(S_i)}\right) A_i \tilde{\mu}^Y(1, S_i, X_i) + o_p(1).$$

Therefore, we have

$$R_{n,1} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(1 - \frac{1}{\pi(S_i)}\right) A_i \tilde{\mu}^Y(1, S_i, X_i) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1 - A_i) \tilde{\mu}^Y(1, S_i, X_i) + o_p(1).$$

The linear expansion of $R_{n,2}$ can be established in the same manner.

For $R_{n,3}$, note that

$$\begin{aligned} Y_i &= Y_i(1)\left[D_i(1)A_i + D_i(0)(1 - A_i)\right] + Y_i(0)\left[1 - D_i(1)A_i - D_i(0)(1 - A_i)\right] \\ &= \left[Y_i(1)D_i(1) - Y_i(0)D_i(1)\right] A_i + \left[Y_i(1)D_i(0) - Y_i(0)D_i(0)\right](1 - A_i) + Y_i(0). \end{aligned}$$

Then

$$\begin{aligned} A_iY_i &= \left[Y_i(1)D_i(1) + Y_i(0)(1 - D_i(1))\right] A_i, \\ (1 - A_i)Y_i &= \left[Y_i(1)D_i(0) + Y_i(0)(1 - D_i(0))\right](1 - A_i), \\ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{A_iY_i}{\hat{\pi}(S_i)} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{\hat{\pi}(S_i)} \left[Y_i(1)D_i(1) + Y_i(0)(1 - D_i(1))\right] A_i =: \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{\hat{\pi}(S_i)} W_iA_i, \\ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{(1 - A_i)Y_i}{1 - \hat{\pi}(S_i)} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\left[Y_i(1)D_i(0) + Y_i(0)(1 - D_i(0))\right](1 - A_i)}{1 - \hat{\pi}(S_i)} =: \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{Z_i(1 - A_i)}{1 - \hat{\pi}(S_i)}. \end{aligned}$$

Thus we have

$$\begin{aligned} R_{n,3} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{A_iY_i}{\hat{\pi}(S_i)} - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{(1 - A_i)Y_i}{1 - \hat{\pi}(S_i)} - \sqrt{n}G \\ &= \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{\hat{\pi}(S_i)} \tilde{W}_iA_i - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1 - A_i}{1 - \hat{\pi}(S_i)} \tilde{Z}_i \right\} \\ &\quad + \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{\hat{\pi}(S_i)} \mathbb{E}[W_i|S_i]A_i - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1 - A_i}{1 - \hat{\pi}(S_i)} \mathbb{E}[Z_i|S_i] - \sqrt{n}G \right\}. \end{aligned} \tag{P.3}$$

We now consider the second term on the RHS of (P.3). First note that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{\hat{\pi}(S_i)} \mathbb{E}[W_i|S_i]A_i = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{\pi(S_i)} \mathbb{E}[W_i|S_i]A_i - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\hat{\pi}(S_i) - \pi(S_i)}{\hat{\pi}(S_i)\pi(S_i)} \mathbb{E}[W_i|S_i]A_i,$$

$$\begin{aligned} &\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{\pi(S_i)} \mathbb{E}[W_i|S_i]A_i = \sum_{s \in \mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{\pi(s)} \mathbb{E}[W_i|S_i = s]A_i 1\{S_i = s\} \\ &= \sum_{s \in \mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\mathbb{E}[W_i|S_i = s]}{\pi(s)} (A_i - \pi(s)) 1\{S_i = s\} + \sum_{s \in \mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{\pi(s)} \mathbb{E}[W_i|S_i = s]\pi(s) 1\{S_i = s\} \end{aligned}$$

$$= \sum_{s\in\mathcal{S}} \frac{\mathbb{E}[W|S=s]}{\pi(s)\sqrt{n}} \sum_{i=1}^{n} (A_i - \pi(s)) 1\{S_i = s\} + \sum_{s\in\mathcal{S}} \frac{\mathbb{E}[W|S=s]}{\sqrt{n}} \sum_{i=1}^{n} 1\{S_i = s\}$$

$$= \sum_{s\in\mathcal{S}} \frac{\mathbb{E}[W|S=s]}{\pi(s)\sqrt{n}} B_n(s) + \sum_{s\in\mathcal{S}} \frac{\mathbb{E}[W|S=s]}{\sqrt{n}} n(s), \tag{P.4}$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\hat{\pi}(S_i) - \pi(S_i)}{\hat{\pi}(S_i)\pi(S_i)} \mathbb{E}[W_i|S_i] A_i = \sum_{s\in\mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\hat{\pi}(s) - \pi(s)}{\hat{\pi}(s)\pi(s)} \mathbb{E}[W_i|S_i = s] A_i 1\{S_i = s\}$$

$$= \sum_{s\in\mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{B_n(s)}{n(s)\hat{\pi}(s)\pi(s)} \mathbb{E}[W_i|S_i = s] A_i 1\{S_i = s\}$$

$$= \sum_{s\in\mathcal{S}} \frac{B_n(s)\mathbb{E}[W|S=s]}{\sqrt{n}n(s)\hat{\pi}(s)\pi(s)} \sum_{i=1}^{n} A_i 1\{S_i = s\} = \sum_{s\in\mathcal{S}} \frac{B_n(s)\mathbb{E}[W|S=s]}{\sqrt{n}n(s)\hat{\pi}(s)\pi(s)} n_1(s)$$

$$= \sum_{s\in\mathcal{S}} \frac{B_n(s)\mathbb{E}[W|S=s]}{\sqrt{n}\pi(s)}.$$

Therefore, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{\hat{\pi}(S_i)} \mathbb{E}[W_i|S_i] A_i = \sum_{s\in\mathcal{S}} \frac{\mathbb{E}[W|S=s]}{\sqrt{n}} n(s).$$

Similarly, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1 - A_i}{1 - \hat{\pi}(S_i)} \mathbb{E}[Z_i|S_i] = \sum_{s\in\mathcal{S}} \frac{\mathbb{E}[Z|S=s]}{\sqrt{n}} n(s)$$

Then, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{\hat{\pi}(S_i)} \mathbb{E}[W_i|S_i] A_i - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1 - A_i}{1 - \hat{\pi}(S_i)} \mathbb{E}[Z_i|S_i] - \sqrt{n}G$$

$$= \sum_{s\in\mathcal{S}} \frac{\mathbb{E}[W|S=s]}{\sqrt{n}} n(s) - \sum_{s\in\mathcal{S}} \frac{\mathbb{E}[Z|S=s]}{\sqrt{n}} n(s) - \sqrt{n}G$$

$$= \sum_{s\in\mathcal{S}} \sqrt{n} \left(\frac{n(s)}{n} - p(s) \right) \mathbb{E}[W - Z|S=s] + \sum_{s\in\mathcal{S}} \sqrt{n} p(s) \mathbb{E}[W - Z|S=s] - \sqrt{n}G$$

$$= \sum_{s\in\mathcal{S}} \sqrt{n} \left(\frac{n(s)}{n} - p(s) \right) \mathbb{E}[W - Z|S=s] + \sqrt{n}\mathbb{E}[W - Z] - \sqrt{n}G$$

$$= \sum_{s\in\mathcal{S}} \frac{n(s)}{\sqrt{n}} \mathbb{E}[W - Z|S=s] - \sqrt{n}\mathbb{E}[W - Z]$$

$$= \frac{1}{\sqrt{n}} \sum_{s\in\mathcal{S}} \sum_{i=1}^{n} \Big(1\{S_i = s\} \mathbb{E}[W_i - Z_i|S_i = s] \Big) - \sqrt{n}\mathbb{E}[W - Z]$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}[W_i - Z_i | S_i] - \sqrt{n}\mathbb{E}[W - Z]$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\mathbb{E}[W_i - Z_i | S_i] - \mathbb{E}[W_i - Z_i]\right). \tag{P.5}$$

Combining (P.3) and (P.5), we have

$$\begin{aligned} R_{n,3} &= \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{\hat{\pi}(S_i)} \tilde{W}_i A_i - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1 - A_i}{1 - \hat{\pi}(S_i)} \tilde{Z}_i \right\} + \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\mathbb{E}[W_i - Z_i | S_i] - \mathbb{E}[W_i - Z_i]\right) \right\} \\ &= \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{\pi(S_i)} \tilde{W}_i A_i - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1 - A_i}{1 - \pi(S_i)} \tilde{Z}_i \right\} \\ &+ \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\mathbb{E}[W_i - Z_i | S_i] - \mathbb{E}[W_i - Z_i]\right) \right\} + o_p(1), \end{aligned}$$

where the second equality holds because

$$\left(\frac{1}{\pi(s)} - \frac{1}{\hat{\pi}(s)}\right) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{W}_i A_i 1\{S_i = s\} = o_p(1) \quad \text{and}$$

$$\left(\frac{1}{\pi(s)} - \frac{1}{\hat{\pi}(s)}\right) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{Z}_i (1 - A_i) 1\{S_i = s\} = o_p(1)$$

due to the same argument used in the proofs of $R_{n,1}$. □

Lemma P.2. *Under the assumptions in Theorem 5.1, we have*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Xi_1(\mathcal{D}_i, S_i) A_i \rightsquigarrow \mathcal{N}\left(0, \mathbb{E}\pi(S_i)\Xi_1^2(\mathcal{D}_i, S_i)\right),$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Xi_0(\mathcal{D}_i, S_i)(1 - A_i) \rightsquigarrow \mathcal{N}\left(0, \mathbb{E}(1 - \pi(S_i))\Xi_0^2(\mathcal{D}_i, S_i)\right), \quad \textit{and}$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Xi_2(S_i) \rightsquigarrow \mathcal{N}(0, \mathbb{E}\Xi_2^2(S_i)),$$

and the three terms are asymptotically independent.

Proof. Note that under Assumption 1(i), conditional on $\{S^{(n)}, A^{(n)}\}$, the distribution of

$$\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Xi_1(\mathcal{D}_i, S_i) A_i, \; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Xi_0(\mathcal{D}_i, S_i)(1 - A_i) \right)$$

is the same as the distribution of the same quantity where units are ordered by strata and then ordered by $A_i = 1$ first and $A_i = 0$ second within strata. To this end, define $N(s) := \sum_{i=1}^n 1\{S_i < s\}$ and $F(s) := \mathbb{P}(S_i < s)$. Furthermore, independently for each $s \in \mathcal{S}$ and independently of $\{S^{(n)}, A^{(n)}\}$, let $\left\{\mathcal{D}_i^s : 1 \le i \le n\right\}$ be i.i.d with marginal distribution equal to the distribution of $\mathcal{D}|S = s$. Then, we have

$$\begin{aligned}&\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \Xi_1(\mathcal{D}_i, S_i)A_i,\ \frac{1}{\sqrt{n}}\sum_{i=1}^n \Xi_0(\mathcal{D}_i, S_i)(1-A_i)\right)\Bigg|S^{(n)}, A^{(n)}\\ &\stackrel{d}{=}\left(\frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{i=N(s)+1}^{N(s)+n_1(s)} \Xi_1(\mathcal{D}_i^s, s),\ \frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{N(s)+n_1(s)+1}^{N(s)+n(s)} \Xi_0(\mathcal{D}_i^s, s)\right)\Bigg|S^{(n)}, A^{(n)}.\end{aligned}$$

In addition, since $\Xi_2(S_i)$ is a function of $\{S^{(n)}, A^{(n)}\}$, we have, arguing along the line of a joint distribution being the product of a conditional distribution and a marginal distribution,

$$\begin{aligned}&\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \Xi_1(\mathcal{D}_i, S_i)A_i,\ \frac{1}{\sqrt{n}}\sum_{i=1}^n \Xi_0(\mathcal{D}_i, S_i)(1-A_i),\ \frac{1}{\sqrt{n}}\sum_{i=1}^n \Xi_2(S_i)\right)\\ &\stackrel{d}{=}\left(\frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{i=N(s)+1}^{N(s)+n_1(s)} \Xi_1(\mathcal{D}_i^s, s),\ \frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{N(s)+n_1(s)+1}^{N(s)+n(s)} \Xi_0(\mathcal{D}_i^s, s),\ \frac{1}{\sqrt{n}}\sum_{i=1}^n \Xi_2(S_i)\right).\end{aligned}$$

Define $\Gamma_{a,n}(u,s) = \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor un\rfloor} \Xi_a(\mathcal{D}_i^s, s)$ for $a = 0, 1, s \in \mathcal{S}$. We have

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{i=N(s)+1}^{N(s)+n_1(s)} \Xi_1(\mathcal{D}_i^s, s) &= \sum_{s\in\mathcal{S}}\left[\Gamma_{1,n}\left(\frac{N(s)+n_1(s)}{n}, s\right) - \Gamma_{1,n}\left(\frac{N(s)}{n}, s\right)\right],\\ \frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{N(s)+n_1(s)+1}^{N(s)+n(s)} \Xi_0(\mathcal{D}_i^s, s) &= \sum_{s\in\mathcal{S}}\left[\Gamma_{0,n}\left(\frac{N(s)+n(s)}{n}, s\right) - \Gamma_{0,n}\left(\frac{N(s)+n_1(s)}{n}, s\right)\right].\end{aligned}$$

In addition, the partial sum process (w.r.t. $u \in [0,1]$) is stochastic equicontinuous and

$$\left(\frac{N(s)}{n}, \frac{n_1(s)}{n}\right) \stackrel{p}{\to} \left(F(s), \pi(s)p(s)\right).$$

Therefore,

$$\begin{aligned}&\left(\frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{i=N(s)+1}^{N(s)+n_1(s)} \Xi_1(\mathcal{D}_i^s, s),\ \frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{N(s)+n_1(s)+1}^{N(s)+n(s)} \Xi_0(\mathcal{D}_i^s, s),\ \frac{1}{\sqrt{n}}\sum_{i=1}^n \Xi_2(S_i)\right)\\ &= \begin{pmatrix}\sum_{s\in\mathcal{S}}\left[\Gamma_{1,n}\left(F(s)+p(s)\pi(s), s\right) - \Gamma_{1,n}\left(F(s), s\right)\right],\\ \sum_{s\in\mathcal{S}}\left[\Gamma_{0,n}\left(F(s)+p(s), s\right) - \Gamma_{0,n}\left(F(s)+\pi(s)p(s), s\right)\right],\\ \frac{1}{\sqrt{n}}\sum_{i=1}^n \Xi_2(S_i)\end{pmatrix} + o_p(1)\end{aligned}$$

and by construction,

$$\sum_{s\in\mathcal{S}}\Big[\Gamma_{1,n}\left(F(s)+p(s)\pi(s),s\right)-\Gamma_{1,n}\left(F(s),s\right)\Big],$$
$$\sum_{s\in\mathcal{S}}\Big[\Gamma_{0,n}\left(F(s)+p(s),s\right)-\Gamma_{0,n}\left(F(s)+p(s)\pi(s),s\right)\Big],$$
$$\text{and}\quad \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Xi_2(S_i)$$

are independent. Last, we have

$$\sum_{s\in\mathcal{S}}\Big[\Gamma_{1,n}\left(F(s)+p(s)\pi(s),s\right)-\Gamma_{1,n}\left(F(s),s\right)\Big]\rightsquigarrow\mathcal{N}\Big(0,\mathbb{E}\pi(S_i)\Xi_1^2(\mathcal{D}_i,S_i)\Big)$$
$$\sum_{s\in\mathcal{S}}\Big[\Gamma_{0,n}\left(F(s)+p(s),s\right)-\Gamma_{0,n}\left(F(s)+p(s)\pi(s),s\right)\Big]\rightsquigarrow\mathcal{N}\Big(0,\mathbb{E}(1-\pi(S_i))\Xi_0^2(\mathcal{D}_i,S_i)\Big)$$
$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Xi_2(S_i)\rightsquigarrow\mathcal{N}\Big(0,\mathbb{E}\Xi_2^2(S_i)\Big).$$

This implies the desired result.

□

Lemma P.3. *Suppose assumptions in Theorem 5.1 hold. Then,*

$$\frac{1}{n}\sum_{i=1}^{n}A_i\hat{\Xi}_1^2(\mathcal{D}_i,S_i)\xrightarrow{p}\sigma_1^2,\quad \frac{1}{n}\sum_{i=1}^{n}(1-A_i)\hat{\Xi}_0^2(\mathcal{D}_i,S_i)\xrightarrow{p}\sigma_0^2,\quad \textit{and}\quad \frac{1}{n}\sum_{i=1}^{n}\hat{\Xi}_2^2(\mathcal{D}_i,S_i)\xrightarrow{p}\sigma_2^2.$$

Proof. To derive the limit of $\frac{1}{n}\sum_{i=1}^{n}A_i\hat{\Xi}_1^2(\mathcal{D}_i,S_i)$, we first define

$$\begin{aligned}\tilde{\Xi}_1^*(\mathcal{D}_i,s)=&\left[\left(1-\frac{1}{\pi(s)}\right)\overline{\mu}^Y(1,s,X_i)-\overline{\mu}^Y(0,s,X_i)+\frac{Y_i}{\pi(s)}\right]\\&-\tau\left[\left(1-\frac{1}{\pi(s)}\right)\overline{\mu}^D(1,s,X_i)-\overline{\mu}^D(0,s,X_i)+\frac{D_i}{\pi(s)}\right]\quad\text{and}\\ \breve{\Xi}_1(\mathcal{D}_i,s)=&\left[\left(1-\frac{1}{\hat{\pi}(s)}\right)\overline{\mu}^Y(1,s,X_i)-\overline{\mu}^Y(0,s,X_i)+\frac{Y_i}{\hat{\pi}(s)}\right]\\&-\hat{\tau}\left[\left(1-\frac{1}{\hat{\pi}(s)}\right)\overline{\mu}^D(1,s,X_i)-\overline{\mu}^D(0,s,X_i)+\frac{D_i}{\hat{\pi}(s)}\right]\end{aligned}$$

Then, we have

$$\left[\frac{1}{n_1(s)}\sum_{i\in I_1(s)}(\tilde{\Xi}_1^*(\mathcal{D}_i,s)-\tilde{\Xi}_1(\mathcal{D}_i,s))^2\right]^{1/2}$$

$$\leq \left[\frac{1}{n_1(s)}\sum_{i\in I_1(s)}(\tilde{\Xi}_1^*(\mathcal{D}_i,s)-\breve{\Xi}_1(\mathcal{D}_i,s))^2\right]^{1/2}+\left[\frac{1}{n_1(s)}\sum_{i\in I_1(s)}(\tilde{\Xi}_1(\mathcal{D}_i,s)-\breve{\Xi}_1(\mathcal{D}_i,s))^2\right]^{1/2}$$

$$\leq \frac{|\hat{\pi}(s)-\pi(s)|}{\hat{\pi}(s)\pi(s)}\left\{\left[\frac{1}{n_1(s)}\sum_{i\in I_1(s)}\overline{\mu}^{Y,2}(1,s,X_i)\right]^{1/2}+\left[\frac{1}{n_1(s)}\sum_{i\in I_1(s)}W_i^2\right]^{1/2}\right\}$$

$$+\left(|\hat{\tau}-\tau|+\frac{|\tau\hat{\pi}(s)-\hat{\tau}\pi(s)|}{\hat{\pi}(s)\pi(s)}\right)\left\{\left[\frac{1}{n_1(s)}\sum_{i\in I_1(s)}\overline{\mu}^{D,2}(1,s,X_i)\right]^{1/2}+\left[\frac{1}{n_1(s)}\sum_{i\in I_1(s)}D_i^2(1)\right]^{1/2}\right\}$$

$$+|\hat{\tau}-\tau|\left[\frac{1}{n_1(s)}\sum_{i\in I_1(s)}\overline{\mu}^{D,2}(0,s,X_i)\right]^{1/2}$$

$$+\left(\frac{1}{\hat{\pi}(s)}-1\right)\left\{\left[\frac{1}{n_1(s)}\sum_{i\in I_1(s)}\Delta^{Y,2}(1,s,X_i)\right]^{1/2}+|\hat{\tau}|\left[\frac{1}{n_1(s)}\sum_{i\in I_1(s)}\Delta^{D,2}(1,s,X_i)\right]^{1/2}\right\}$$

$$+\left[\frac{1}{n_1(s)}\sum_{i\in I_1(s)}\Delta^{Y,2}(0,s,X_i)\right]^{1/2}+|\hat{\tau}|\left[\frac{1}{n_1(s)}\sum_{i\in I_1(s)}\Delta^{D,2}(0,s,X_i)\right]^{1/2}=o_p(1),$$

where the second inequality holds by the triangle inequality and the fact that when $i\in I_1(s)$, $A_i=1$, $Y_i=W_i$, and $D_i=D_i(1)$, and the last equality is due to Assumption 2(ii) and the facts that $\hat{\pi}(s)\xrightarrow{p}\pi(s)$ and $\hat{\tau}\xrightarrow{p}\tau$. This further implies

$$\frac{1}{n_1(s)}\sum_{i\in I_1(s)}(\tilde{\Xi}_1^*(\mathcal{D}_i,s)-\tilde{\Xi}_1(\mathcal{D}_i,s))\xrightarrow{p}0,$$

by the Cauchy-Schwarz inequality and thus,

$$\left[\frac{1}{n_1(s)}\sum_{i\in I_1(s)}\hat{\Xi}_1^2(\mathcal{D}_i,s)\right]^{1/2}=\left[\frac{1}{n_1(s)}\sum_{i\in I_1(s)}\left(\tilde{\Xi}_1^*(\mathcal{D}_i,s)-\frac{1}{n_1}\sum_{i\in I_1(s)}\tilde{\Xi}_1^*(\mathcal{D}_i,s)\right)^2\right]^{1/2}+o_p(1).$$

Next, following the same argument in the proof of Lemma P.2, we have

$$\begin{aligned}\frac{1}{n_1(s)}\sum_{i\in I_1(s)}\tilde{\Xi}_1^*(\mathcal{D}_i,s)&\stackrel{d}{=}\frac{1}{n_1(s)}\sum_{i=N(s)+1}^{N(s)+n_1(s)}\left\{\left[\left(1-\frac{1}{\pi(s)}\right)\overline{\mu}^Y(1,s,X_i^s)-\overline{\mu}^Y(0,s,X_i^s)+\frac{W_i^s}{\pi(s)}\right]\right.\\&\left.\quad-\tau\left[\left(1-\frac{1}{\pi(s)}\right)\overline{\mu}^D(1,s,X_i^s)-\overline{\mu}^D(0,s,X_i^s)+\frac{D_i^s(1)}{\pi(s)}\right]\right\}\\&\xrightarrow{p}\mathbb{E}\left\{\left[\left(1-\frac{1}{\pi(S_i)}\right)\overline{\mu}^Y(1,S_i,X_i)-\overline{\mu}^Y(0,S_i,X_i)+\frac{W_i}{\pi(S_i)}\right]\right.\\&\left.\quad-\tau\left[\left(1-\frac{1}{\pi(S_i)}\right)\overline{\mu}^D(1,S_i,X_i)-\overline{\mu}^D(0,S_i,X_i)+\frac{D_i(1)}{\pi(S_i)}\right]|S_i=s\right\},\end{aligned}$$

This implies

$$
\begin{aligned}
&\left[\frac{1}{n_1(s)}\sum_{i\in I_1(s)}\left(\tilde{\Xi}_1^*(\mathcal{D}_i,s)-\frac{1}{n_1}\sum_{i\in I_1(s)}\tilde{\Xi}_1^*(\mathcal{D}_i,s)\right)^2\right]^{1/2}\\
=&\left[\frac{1}{n_1(s)}\sum_{i\in I_1(s)}\left(\tilde{\Xi}_1^*(\mathcal{D}_i,s)-\mathbb{E}\left\{\left[\left(1-\frac{1}{\pi(S_i)}\right)\overline{\mu}^Y(1,S_i,X_i)-\overline{\mu}^Y(0,S_i,X_i)+\frac{W_i}{\pi(S_i)}\right]\right.\right.\right.\\
&\left.\left.\left.-\tau\left[\left(1-\frac{1}{\pi(S_i)}\right)\overline{\mu}^D(1,S_i,X_i)-\overline{\mu}^D(0,S_i,X_i)+\frac{D_i(1)}{\pi(S_i)}\right]|S_i=s\right\}\right)^2\right]^{1/2}+o_p(1)\\
=&\left[\frac{1}{n_1(s)}\sum_{i\in I_1(s)}\left(\left[\left(1-\frac{1}{\pi(S_i)}\right)\tilde{\mu}^Y(1,S_i,X_i)-\tilde{\mu}^Y(0,S_i,X_i)+\frac{\tilde{W}_i}{\pi(S_i)}\right]\right.\right.\\
&\left.\left.-\tau\left[\left(1-\frac{1}{\pi(S_i)}\right)\tilde{\mu}^D(1,S_i,X_i)-\tilde{\mu}^D(0,S_i,X_i)+\frac{\tilde{D}_i(1)}{\pi(S_i)}\right]\right)^2\right]^{1/2}+o_p(1).
\end{aligned}
$$

Last, following the same argument in the proof of Lemma P.2, we have

$$
\begin{aligned}
&\left[\frac{1}{n_1(s)}\sum_{i\in I_1(s)}\left(\left[\left(1-\frac{1}{\pi(S_i)}\right)\tilde{\mu}^Y(1,S_i,X_i)-\tilde{\mu}^Y(0,S_i,X_i)+\frac{\tilde{W}_i}{\pi(S_i)}\right]\right.\right.\\
&\left.\left.-\tau\left[\left(1-\frac{1}{\pi(S_i)}\right)\tilde{\mu}^D(1,S_i,X_i)-\tilde{\mu}^D(0,S_i,X_i)+\frac{\tilde{D}_i(1)}{\pi(S_i)}\right]\right)^2\right]^{1/2}\\
\stackrel{d}{=}&\left[\frac{1}{n_1(s)}\sum_{i=N(s)+1}^{N(s)+n_1(s)}\left(\left[\left(1-\frac{1}{\pi(s)}\right)\tilde{\mu}^Y(1,s,X_i^s)-\tilde{\mu}^Y(0,s,X_i^s)+\frac{\tilde{W}_i^s}{\pi(s)}\right]\right.\right.\\
&\left.\left.-\tau\left[\left(1-\frac{1}{\pi(s)}\right)\tilde{\mu}^D(1,s,X_i^s)-\tilde{\mu}^D(0,s,X_i^s)+\frac{\tilde{D}_i^s(1)}{\pi(s)}\right]\right)^2\right]^{1/2}\\
\xrightarrow{p}&\left[\mathbb{E}(\Xi_1^2(\mathcal{D}_i,S_i)|S_i=s)\right]^{1/2},
\end{aligned}
$$

where $\tilde{W}_i^s=W_i^s-\mathbb{E}(W_i|S_i=s)$ and $\tilde{D}_i^s(1)=D_i^s(1)-\mathbb{E}(D_i(1)|S_i=s)$ and the last convergence is due to the fact that conditionally on $S^{(n)}$, $A^{(n)}$, $\{X_i^s,\tilde{W}_i^s,\tilde{D}_i^s(1)\}_{i\in I_1(s)}$ is a sequence of i.i.d. random variables so that the standard LLN is applicable. Combining all the results above, we have shown that

$$
\begin{aligned}
&\frac{1}{n_1(s)}\sum_{i\in I_1(s)}\hat{\Xi}_1^2(\mathcal{D}_i,S_i)\xrightarrow{p}\mathbb{E}(\Xi_1^2(\mathcal{D}_i,S_i)|S_i=s)\\
&\frac{1}{n}\sum_{i=1}^n A_i\hat{\Xi}_1^2(\mathcal{D}_i,S_i)=\sum_{s\in S}\frac{n_1(s)}{n}\left(\frac{1}{n_1(s)}\sum_{i\in I_1(s)}\hat{\Xi}_1^2(\mathcal{D}_i,S_i)\right)\\
&\qquad\xrightarrow{p}\sum_{s\in\mathcal{S}}p(s)\pi(s)\mathbb{E}(\Xi_1^2(\mathcal{D}_i,S_i)|S_i=s)=\mathbb{E}\pi(S_i)\mathbb{E}(\Xi_1^2(\mathcal{D}_i,S_i)|S_i)=\sigma_1^2.
\end{aligned}
$$

For the same reason, we can show that

$$\frac{1}{n}\sum_{i=1}^{n}(1-A_i)\hat{\Xi}_0^2(\mathcal{D}_i,S_i)\xrightarrow{p}\sigma_0^2.$$

Last, by the similar argument, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\hat{\Xi}_2^2(S_i) &= \sum_{s\in\mathcal{S}}\frac{n(s)}{n}\hat{\Xi}_2^2(s)\\ &= \sum_{s\in\mathcal{S}}\frac{n(s)}{n}(\mathbb{E}(W_i-\tau D_i(1)|S_i=s)-\mathbb{E}(Z_i-\tau D_i(0)|S_i=s))^2+o_p(1)\\ &= \sum_{s\in\mathcal{S}}\frac{n(s)}{n}\Xi_2^2(s)+o_p(1)\\ &\xrightarrow{p}\sum_{s\in\mathcal{S}}p(s)\Xi_2^2(s)=\mathbb{E}\Xi_2^2(S_i)=\sigma_2^2.\end{aligned}$$

□

References

Abadie, A., M. M. Chingos, and M. R. West (2018). Endogenous stratification in randomized experiments. *Review of Economics and Statistics 100*(4), 567–580.

Angrist, J. D. and G. Imbens (1994). Identification and estimation of local average treatment effects. *Econometrica 62*, 467–475.

Ansel, J., H. Hong, and J. Li (2018). Ols and 2sls in randomised and conditionally randomized experiments. *Journal of Economics and Statistics 238*, 243–293.

Armstrong, T. B. (2022). Asymptotic efficiency bounds for a class of experimental designs.

Athey, S. and G. W. Imbens (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments*, Volume 1, pp. 73–140. Elsevier.

Bai, Y. (2020). Optimality of matched-pair designs in randomized controlled trials. *Available at SSRN 3483834*.

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

Banerjee, A., R. Hanna, J. Kyle, B. A. Olken, and S. Sumarto (2018). Tangible information and citizen empowerment: Identification cards and food subsidy programs in indonesia. *Journal of Political Economy 126*(2), 451–491.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017). Program evaluation with high-dimensional data. *Econometrica 85*(1), 233–298.

Blattman, C. and S. Dercon (2018). The impacts of industrial and entrepreneurial work on income and health: Experimental evidence from ethiopia. *American Economic Journal: Applied Economics 10*, 1–38.

Bloniarz, A., H. Liu, C.-H. Zhang, J. S. Sekhon, and B. Yu (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences 113*(27), 7383–7390.

Bruhn, M., G. L. Ibarra, and D. McKenzie (2014). The minimal impact of a large-scale financial education program in mexico city. *Journal of Development Economics 108*, 184–189.

Bruhn, M. and D. McKenzie (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics 1*(4), 200–232.

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association 113*(524), 1741–1768.

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2019). Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics 10*(4), 1747–1785.

Bugni, F. A. and M. Gao (2021). Inference under covariate-adaptive randomization with imperfect compliance. *arXiv preprint arXiv: 2102.03937*.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In *Handbook of Econometrics*, Volume 6, pp. 5549–5632. Elsevier.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014). Gaussian approximation of suprema of empirical processes. *Annals of Statistics 42*(4), 1564–1597.

Dupas, P., D. Karlan, J. Robinson, and D. Ubfal (2018). Banking the unbanked? evidence from three countries. *American Economic Journal: Applied Economics 10*, 257–297.

Fogarty, C. B. (2018). Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika 105*(4), 994–1000.

Freedman, D. A. (2008a). On regression adjustments in experiments with several treatments. *Annals of Applied Statistics 2*(1), 176–196.

Freedman, D. A. (2008b). On regression adjustments to experimental data. *Advances in Applied Mathematics 40*(2), 180–193.

Frölich, M. (2007). Nonparametric iv estimation of local average treatment effects with covariates. *Journal of Econometrics 139*(1), 35–75.

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

Hirano, K. and G. Imbens (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services & Outcomes Research Methodology 2*, 259–278.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Hirshleifer, S., D. McKenzie, R. Almeida, and C. Ridao-Cano (2016). The impact of vocational training for the unemployed: experimental evidence from turkey. *The Economic Journal 126*(597), 2115–2146.

Hu, Y. (2016). Generalized Efron's biased coin design and its theoretical properties. *Journal of Applied Probability 53*(2), 327–340.

Hu, Y. and F. Hu (2012). Asymptotic properties of covariate-adaptive randomization. *Annals of Statistics 40*(3), 1794–1815.

Imbens, G. and J. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*, 467–475.

Jiang, L., X. Liu, P. C. B. Phillips, and Y. Zhang (2021). Bootstrap inference for quantile treatment effects in randomized experiments with matched pairs. *Review of Economics and Statistics, forthcoming.*

Jiang, L., P. Phillips, Y. Tao, and Y. Zhang (2022). Regression-adjusted estimation of quantile treatment effects under covariate-adaptive randomizations. *Journal of Econometrics, forthcoming.*

Lei, L. and P. Ding (2021). Regression adjustment in completely randomized experiments with a diverging number of covariates. *Biometrika, forthcoming.*

Li, X. and P. Ding (2020). Rerandomization and regression adjustment. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(1), 241–268.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. *Annals of Applied Statistics 7*(1), 295–318.

Liu, H., F. Tu, and W. Ma (2020). A general theory of regression adjustment for covariate-adaptive randomization: OLS, Lasso, and beyond. *arXiv preprint arXiv: 2011.09734*.

Liu, H. and Y. Yang (2020). Regression-adjusted average treatment effect estimates in stratified randomized experiments. *Biometrika 107*(4), 935–948.

Lu, J. (2016). Covariate adjustment in randomization-based causal inference for 2K factorial designs. *Statistics & Probability Letters 119*, 11–20.

Ma, W., F. Hu, and L. Zhang (2015). Testing hypotheses of covariate-adaptive randomized clinical trials. *Journal of the American Statistical Association 110*(510), 669–680.

Ma, W., Y. Qin, Y. Li, and F. Hu (2020). Statistical inference for covariate-adaptive randomization procedures. *Journal of the American Statistical Association 115*(531), 1488–1497.

McKenzie, D., N. Assaf, and A. P. Cusolito (2017). The additionality impact of a matching grant programme for small firms: experimental evidence from yemen. *Journal of Development Effectiveness 9*(1), 1–14.

Negi, A. and J. M. Wooldridge (2020). Revisiting regression adjustment in experiments with heterogeneous treatment effects. *Econometric Reviews 40*(5), 1–31.

Olivares, M. (2021). Robust permutation test for equality of distributions under covariate-adaptive randomization. *Working paper, University of Illinois at Urbana Champaign*.

Ren, J. and H. Liu (2021). Model-assisted complier average treatment effect estimates in randomized experiments with non-compliance and a binary outcome. *arXiv preprint arXiv:2111.10109*.

Robins, J., A. Rotnitzky, and M. van der Laan (2000). Comment. *Journal of the American Statistical Association 95*, 477–482.

Robins, J., A. Rotnitzky, and L. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association 89*, 846–866.

Robins, J. M. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association 90*(429), 122–129.

Rothe, C. and S. Firpo (2019). Properties of doubly robust estimators when nuisance functions are estimated nonparametrically. *Econometric Theory 35*, 1048–1087.

Scharfstein, D., A. Rotnitzky, and J. Robins (1999). Rejoinder. *Journal of the American Statistical Association 94*, 1135–1146.

Seaman, S. R. and S. Vansteelandt (2018). Introduction to double robust methods for incomplete data. *Statistical Science 33*(2), 184–197.

Shao, J. and X. Yu (2013). Validity of tests under covariate-adaptive biased coin randomization and generalized linear models. *Biometrics 69*(4), 960–969.

Shao, J., X. Yu, and B. Zhong (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika 97*(2), 347–360.

Słoczyński, T. and J. M. Wooldridge (2018). A general double robustness result for estimating average treatment effects. *Econometric Theory 34*, 112–133.

Tabord-Meehan, M. (2021). Stratification trees for adaptive randomization in randomized controlled trials. *arXiv preprint arXiv: 1806.05127*.

van der Vaart, A. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

Wei, L. (1978). An application of an urn model to the design of sequential controlled clinical trials. *Journal of the American Statistical Association 73*(363), 559–563.

Wooldridge, J. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics 141*, 1281–1301.

Ye, T. (2018). Testing hypotheses under covariate-adaptive randomisation and additive models. *Statistical Theory and Related Fields 2*(1), 96–101.

Ye, T. and J. Shao (2020). Robust tests for treatment effect in survival analysis under covariate-adaptive randomization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(5), 1301–1323.

Ye, T., Y. Yi, and J. Shao (2021). Inference on average treatment effect under minimization and other covariate-adaptive randomization methods. *Biometrika, forthcoming*.

Zhang, Y. and X. Zheng (2020). Quantile treatment effects and bootstrap inference under covariate-adaptive randomization. *Quantitative Economics 11*(3), 957–982.

Zhao, A. and P. Ding (2021). Covariate-adjusted fisher randomization tests for the average treatment effect. *Journal of Econometrics 225*, 278–294.