UNIVERSITY OF CAMBRIDGE
Faculty of Economics

JANEWAY INSTITUTE CAMBRIDGE

CAMBRIDGE WORKING PAPERS IN ECONOMICS
JANEWAY INSTITUTE WORKING PAPERS

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

Liang Jian
Fudan University

Oliver B. Linton
University of Cambridge

Haihan Tang
Fudan University

Yichong Zhang
Singapore Management University

Abstract

We investigate how to improve efficiency using regression adjustments with covariates in covariate-adaptive randomizations (CARs) with imperfect subject compliance. Our regression-adjusted estimators, which are based on the doubly robust moment for local average treatment effects, are consistent and asymptotically normal even with heterogeneous probabilities of assignment and misspecified regression adjustments. We propose an optimal but potentially misspecified linear adjustment and its further improvement via a nonlinear adjustment, both of which lead to more efficient estimators than the one without adjustments. We also provide conditions for nonparametric and regularized adjustments to achieve the semiparametric efficiency bound under CARs.

Reference Details

2366	Cambridge Working Papers in Economics
2315	Janeway Institute Working Paper Series
Published	25 October 2023
Key Words	Randomized experiment, Covariate-adaptive randomization, High-dimensional data, Local average treatment effects, Regression adjustment
JEL Codes	C14, C21, I21
Websites	www.econ.cam.ac.uk/cwpe www.janeway.econ.cam.ac.uk/working-papers

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

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

October 25, 2023

Abstract

We investigate how to improve efficiency using regression adjustments with covariates in covariate-adaptive randomizations (CARs) with imperfect subject compliance. Our regression-adjusted estimators, which are based on the doubly robust moment for local average treatment effects, are consistent and asymptotically normal even with heterogeneous probabilities of assignment and misspecified regression adjustments. We propose an optimal but potentially misspecified linear adjustment and its further improvement via a nonlinear adjustment, both of which lead to more efficient estimators than the one without adjustments. We also provide conditions for nonparametric and regularized adjustments to achieve the semiparametric efficiency bound under CARs.

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

JEL codes: C14, C21, I21

*We thank the editor and three anonymous referees for extensive comments which have led to many improvements. We are also grateful to Federico Bugni and participants at NCSU CIMA seminar for their comments. Yichong Zhang acknowledges the financial support from Singapore Ministry of Education Tier 2 grant under grant MOE2018-T2-2-169, the NSFC under the grant No. 72133002, and a Lee Kong Chian fellowship. We thank Mingxin Zhang for excellent research assistance. The `R` package "drcarlate" to implement all the adjustment methods discussed in the paper can be downloaded from https://cran.r-project.org/web/packages/drcarlate. 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 become increasingly popular in economic research. One commonly used randomization method employed by economists to ensure 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 pretreatment variables. However, subject compliance with the random assignment is usually imperfect. We survey all publications using randomized experiments in eight leading economics journals from 2015 to 2022 and identify ten papers that used CARs with imperfect compliance.[1]

When subjects do not comply with the assignment in CARs, researchers usually estimate the local average treatment effects (LATEs) for the compliers using the two-stage least squares (TSLS) method with treatment assignment as an instrumental variable and covariates and strata fixed effects as exogenous controls. Actually, all ten papers mentioned above estimate the LATE in this way. We simply denote this estimator as TSLS. Recently, Ansel, Hong, and Li (2018) proposed an S estimator (denoted as S) which aggregates IV estimators for strata. Bugni and Gao (2023) proposed a fully saturated estimator with strata dummies, which we call the unadjusted estimator (NA) as it does not use covariates. The standard theory for the consistency of TSLS requires both correct specification of the conditional mean model and homogeneous treatment effect. In contrast, both S and NA estimators are consistent under CARs without requiring correct specifications, homogeneous treatment effect, or identical treatment assignment probability across strata. Ansel et al. (2018) further show the S estimator is the most efficient among all the estimators discussed in their paper (Proposition 7).

Nevertheless, the existing literature lacks a systematic study or comparison of various LATE estimators under CARs. TSLS and S estimators assume different linear conditional mean models, which can be viewed as different types of linear regression adjustments.[2] Then, under what conditions the TSLS estimator, like the S estimator, is consistent

[1] See Section 2.3 for more details.

[2] We use the plural of the word "adjustment" for an estimator because each estimator consists of working models for both the first and second stages.

even when the regression adjustments are misspecified? How is the efficiency comparison among TSLS, S, and NA estimators when all of them are consistent? Is the S estimator the most efficient among all linearly adjusted LATE estimators? Can other potentially misspecified *nonlinear* regression adjustments lead to more efficient LATE estimators? Last, what is the semiparametric efficiency bound (SEB) for LATE estimation under (CARs) and how can we achieve it?

In this paper, we provide answers to all these questions. Specifically, we follow the framework that was recently established by Bugni, Canay, and Shaikh (2018) to study causal inference under CARs, which allows for heterogeneous assignment probabilities and treatment effects. We first show that (1) TSLS with both strata dummies and covariates as exogenous controls is inconsistent if *both* the assignment probabilities *and* treatment effects are heterogeneous across strata; (2) even when TSLS is consistent (especially when the treatment assignment probabilities are homogeneous), its usual heteroskedasticity robust standard error is conservative due to the cross-sectional dependence introduced by CARs;[3] (3) the correct asymptotic variance of the TSLS estimator may be greater than that of the NA estimator, which defeats the purpose of using covariates in the regression.

We then propose a general regression-adjusted estimator using the doubly robust moment for LATE with a consistent estimator of the assignment probability and potentially misspecified regression adjustments based on covariates. The doubly robust moment for LATE has been derived by Frölich (2007) and used for estimating LATE by Słoczyński, Uysal, and Wooldridge (2022) and Heiler (2022). But we are the first to apply it under CARs and investigate the potential efficiency improvements when the regression adjustments are misspecified. We show that our inference method (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 SEB when the adjustments are correctly specified. The SEB for LATE under CARs is also new to the literature and complements those bounds derived by Frölich (2007) and Armstrong (2022).[4]

[3]This point is consistent with the result in Ansel et al. (2018) for their estimator $\hat{\beta}_2$. However, $\hat{\beta}_2$ is computed by TSLS with only strata dummies under the assumption of homogeneous assignment probabilities, but no covariates as exogenous control variables.

[4]Frölich (2007) derived the SEB for LATE assuming i.i.d. data. However, CARs can introduce cross-sectional dependence, and thus, violate the independence assumption. Armstrong (2022) derived the

Finally, we compare the efficiency of our LATE estimators with three specific parametric forms of regression adjustments: (1) the optimal linear adjustments (denoted as L), which yield the most efficient estimator among all linearly adjusted estimators, (2) the nonlinear logistic adjustments (denoted as NL), and (3) a combination of linear and nonlinear adjustments (denoted as F) which is more efficient than both linear and nonlinear adjustments and new to the literature. We also extend Ansel et al. (2018) by showing that their S estimator is asymptotically equivalent to our estimator L, and is thus optimal among the linearly adjusted estimators but less efficient than estimator F. We further give conditions under which estimators with nonparametric (denoted as NP) and regularized (denoted as R) regression adjustments achieve the SEB. Figure 1 in Section 2.5 visualizes the partial order of efficiency of these estimators.

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 the average treatment effect (ATE) or quantile treatment effect (QTE) under CARs without considering covariates. Bugni et al. (2018); Bugni, Canay, and Shaikh (2019); Bloniarz, Liu, Zhang, Sekhon, and Yu (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 schemes. 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) devised optimal randomization designs that may produce an ATE estimator with the lowest variance. Bugni and Gao (2023) further examined the optimal design with imperfect compliance. All the above works, except Bugni and Gao (2023), assumed perfect compliance, while we contribute to the literature by studying the LATE estimators in the context of CARs and regression adjustments, which allows imperfect compliance. Ren and Liu (2021) studied the

SEB for average treatment effect under CARs but without covariates. The SEB for LATE under CARs but without covariates is a byproduct of our result by letting our covariates be an empty set.

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 *superpopulation* asymptotics. Finally, our paper also connects to a vast literature on 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.

Acronyms. In this paper, we refer to the optimally linearly adjusted, nonlinearly (logistic) adjusted, and nonparametrically adjusted estimators, introduced in Sections 5.1.1, 5.1.2, and S.C of the Online Supplement, as L, NL, and NP, respectively. We also use NA and S to denote the fully saturated and S estimators proposed by Bugni and Gao (2023) and Ansel et al. (2018), respectively. F denotes the estimator with adjustments that improve upon both optimal linear and nonlinear adjustments (Section 5.1.3), while R denotes the estimator with regularized adjustments (Section 5.2). We will provide more details about these estimators below.

2 Setting and Empirical Practice

2.1 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 $Y_i(1)$ and $Y_i(0)$ are the potential treated and untreated outcomes for the individual i, respectively, and D_i is a binary random variable indicating whether the individual i received 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 outcome 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 $Y_i = Y_i(D_i(1))A_i + Y_i(D_i(0))(1-A_i)$. Individual i belongs to stratum S_i and possesses covariate vector X_i, where X_i does not include the constant term. The support of the vectors $\{X_i\}_{i=1}^n$ is denoted $\text{Supp}(X)$, while the support of $\{S_i\}_{i=1}^n$ is $\mathcal{S}$, which is a finite set. Without loss of generality, suppose that $\mathcal{S} = \{1, \ldots, S\}$ for some integer $S > 0$.

A researcher can observe the data $\{Y_i, D_i, A_i, S_i, X_i\}_{i=1}^n$. 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 with respect to* 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,* $c < \min_{s\in\mathcal{S}} \pi(s) \le \max_{s\in\mathcal{S}} \pi(s) < 1 - c$ *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) $\max_{a=0,1,s\in\mathcal{S}} \mathbb{E}(|Y_i(a)|^q | S_i = s) \le C < \infty$ *for some* $q \ge 4$.

Several remarks are in order. First, Assumption 1(i) allows for the treatment assignment $A^{(n)}$, and thus, the observed outcome $\{Y_i\}_{i\in[n]}$ to be cross-sectionally dependent, which is usually the case for CARs. Second, Assumption 1(ii) implies that the treatment assignment $A^{(n)}$ are generated only based on strata indicators. Third, Assumption 1(iii) imposes that the strata sizes are roughly balanced. Fourth, 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).[5] Note that we only

[5]For completeness, we briefly repeat their descriptions in Appendix S.A.

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). Fifth, Assumption 1(v) implies there are no defiers. Last, Assumption 1(vi) is a standard moment condition.

Throughout the paper, we are interested in estimating the *local average treatment effect* (LATE), which is denoted by 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).

2.2 Examples of Economics Datasets

To motivate our work, we give three examples of prominent economic datasets that use CARs and have imperfect compliance.

Example 1. *Atkin, Khandelwal, and Osman (2017) conducted a randomized experiment with a CAR design to identify the impact of exporting on firm performance.*[6] *They had two samples of firms. In sample 1, they randomized firms into treatment or control with a target probability of 0.5 in each of the strata named: Goublan, Tups and Duble. In sample 2, they randomly select firms for the treatment group with a target probability of 0.25 in stratum Duble. They then combined the two samples together, which makes the probabilities of assignment into treatment ($\pi(s)$) in their joint sample heterogeneous across strata. Firms with assignment into treatment were offered an initial opportunity to sell to high-income markets, but only 62.16% of them managed to secure large and lasting orders.*

Example 2. *Dupas, Karlan, Robinson, and Ubfal (2018) studied how rural households benefit from free bank accounts.*[7] *They randomly assigned 2,160 households to treatment or control groups within each of the 41 strata. The targeted assignment probability for treatment for each stratum is 0.5. Households with assignment into treatment received*

[6] The dataset can be found at https://doi.org/10.7910/DVN/QOGMVI.

[7] The dataset is available at https://www.openicpsr.org/openicpsr/project/116346/version/V1/view.

vouchers to open accounts, but only 41.87% of them did so and deposited money within 2 years.

Example 3. *Jha and Shayo (2019) examined how financial market participation affects political views and voting behavior.*[8] *They used CAR to randomly assign 1345 participants to treatment or control groups within each stratum, with a target probability for treatment of 0.75. Participants with assignment into treatment were offered to trade assets, but only 81.08% of them made a trade.*

2.3 Survey of Empirical Practice

[**Insert Table 1 here**.]

We survey the common practice for analyzing experiments in the empirical economics literature. Our survey is limited to articles that contain the term "experiment" in their title or abstract and are published between January 2015 and December 2022 in eight journals: the *American Economic Journal: Applied Economics* (AEJ: Applied), *American Economic Journal: Economic Policy* (AEJ: Policy), *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics* (QJE), *Review of Economics and Statistics* (ReStat), and *Review of Economic Studies*. We then manually select the articles that use CARs and report imperfect compliance. Table 1 tabulates the articles found in our survey. It shows that all the papers in our sample use TSLS with covariates and strata fixed effects to estimate the LATE. This finding motivates us to study the statistical properties of this commonly used TSLS estimator in Section 2.4 before proposing our new estimator.

2.4 TSLS with Covariates and Strata Fixed Effects

Our survey shows that empirical researchers using CARs usually estimate LATE via TSLS regressions with strata dummies and covariates. The first and second stages of the

[8] The dataset can be found at https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA16385.

TSLS regression can be formed as

$$D_i \sim \gamma A_i + \sum_{s\in\mathcal{S}} a_s 1\{S_i = s\} + X_i^\top \theta, \quad Y_i \sim \tau D_i + \sum_{s\in\mathcal{S}} \alpha_s 1\{S_i = s\} + X_i^\top \delta, \tag{2.1}$$

where $\{a_s\}_{s\in\mathcal{S}}$ and $\{\alpha_s\}_{s\in\mathcal{S}}$ are the strata fixed effects.

Denote the TSLS estimator of τ by $\hat{\tau}_{TSLS}$. To study the asymptotic properties of $\hat{\tau}_{TSLS}$, we follow Bugni et al. (2018) and Ansel et al. (2018) and make the following additional assumption on the treatment assignment mechanism.

Assumption 2. *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* $0 \leq \gamma(s) \leq \pi(s)(1-\pi(s))$.

Three remarks are in order. First, Assumption 2 is used to analyze the TSLS estimator only and is not needed for all the analyses in later sections in the paper. Second, it implies Assumption 1(iv). Third, we have $\gamma(s) = \pi(s)(1-\pi(s))$ for SRS and $\gamma(s) < \pi(s)(1-\pi(s))$ for the other three randomization designs mentioned after Assumption 1. Specifically, for BCD and SBR, we have $\gamma(s) = 0$, which means the assignment rules achieve the strong balance.

Following empirical researchers, we also consider the usual IV heteroskedasticity-robust standard error estimator for TSLS estimator $\hat{\tau}_{TSLS}$, which is denoted as $\hat{\sigma}_{TSLS,naive}/\sqrt{n}$.[9] We compare $\hat{\tau}_{TSLS}$ with Bugni and Gao's (2023) fully saturated estimator (denoted as $\hat{\tau}_{NA}$) for τ under CAR, which does not use any covariates X_i. The asymptotic variance of $\hat{\tau}_{NA}$ is then denoted as σ^2_{NA}, which is given in Bugni and Gao (2023). In Section 3, we further show that $\hat{\tau}_{NA}$ is a special case of our general estimator whose asymptotic variance is derived in the proof of Theorem 3.1.

[9]The detailed definition of $\hat{\sigma}_{TSLS,naive}$ can be found in the proof of Theorem 2.1.

Theorem 2.1. *Suppose Assumption 1 holds. Then, we have*

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

If $\pi(s)$ or $\frac{\mathbb{E}(Y_i(D_i(1))|S_i=s)-\mathbb{E}(Y_i(D_i(0))|S_i=s)}{\mathbb{E}(D_i(1)|S_i=s)-\mathbb{E}(D_i(0)|S_i=s)}$ is the same across $s \in \mathcal{S}$, then $\hat{\tau}_{TSLS} \xrightarrow{p} \tau$. If $\pi(s) = \pi$ for all $s \in \mathcal{S}$ and Assumptions 1 and 2 hold, then

$$\sqrt{n}(\hat{\tau}_{TSLS} - \tau) \rightsquigarrow \mathcal{N}(0, \sigma^2_{TSLS}) \quad \text{and} \quad \hat{\sigma}^2_{TSLS,naive} \xrightarrow{p} \sigma^2_{TSLS,naive},$$

where the definitions of σ^2_{TSLS} and $\sigma^2_{TSLS,naive}$ can be found in the proof, $\sigma^2_{TSLS} \leq \sigma^2_{TSLS,naive}$, and the inequality is strict if $\gamma(s) < \pi(1-\pi)$. Last, it is possible to have $\sigma^2_{TSLS} > \sigma^2_{NA}$.

Theorem 2.1 highlights one advantage and three limitations of the commonly used TSLS estimator under CARs. The advantage is that the TSLS estimator can consistently estimate the LATE under certain conditions without assuming the linear regression in (2.1) is correctly specified. Hence, the reason for incorporating covariates in the regression is to improve estimation efficiency. The first limitation is that the TSLS estimator is inconsistent when both the treatment effect and the probabilities of treatment assignment vary across strata. To ensure its consistency, economists should thus keep the target assignment probability ($\pi(s)$) equal across all strata in the experimental design stage, which may not be the case in practice (see, for example, the first dataset in Section 2.2). The second limitation is that the heteroskedasticity-robust standard error reported by standard software such as STATA is conservative and inconsistent unless $\gamma(s) = \pi(1-\pi)$. However, this condition is violated when treatment is not assigned independently, such as BCD and SBR, which are widely used in RCTs. The third limitation is that the asymptotic variance σ^2_{TSLS} may not be smaller than that of the unadjusted estimator, which goes against the purpose of using covariates in the regression.

In this paper, we develop estimators that have the same advantage as TSLS but avoid all these limitations. Specifically, our proposed LATE estimators are (1) consistent even under misspecification of regression models, (2) consistent even when the probabilities of treatment assignment are heterogeneous across strata, and (3) guaranteed to be weakly more efficient than the unadjusted estimator. We also provide consistent estimators of the asymptotic variances for our LATE estimators.

2.5 Efficiency Comparison of LATE Estimators: Preview

Before delving into the details of our proposed estimators, we provide a preview of the efficiency comparison among various LATE estimators mentioned in the paper. Figure 1 illustrates their relationship, with the most efficient on the right and the least efficient on the left. A dashed circle around an estimator indicates that this estimator is not always consistent. The least efficient estimator in Figure 1 is the strata fixed effects IV (SFE IV) estimator, as proposed by Bugni and Gao (2023). Bugni and Gao (2023) showed that SFE IV is consistent only when the probability of assignment into treatment is the same across strata. Even when SFE IV is consistent, they showed that it is no more efficient than NA. There are no arrows between NA and TSLS because TSLS can be less efficient than NA even when it is consistent. Ansel et al.'s (2018) S estimator is asymptotically equivalent to our estimator L (i.e., the optimal linear adjustments). Since both NA and TSLS (whenever it is consistent) have linear adjustments (NA has linear adjustments with zero coefficients), they are less efficient than S or L. There is no clear winner between NL and L because even the optimal linear adjustments can be misspecified and thus potentially less efficient than some nonlinear adjustments. Theoretically, the logistic regression adjustments can be even less efficient than NA depending on how severe the misspecification is. However, the F estimator is guaranteed to be more efficient than both L and NL by construction. Last, as NP and R achieve the SEB, they are more efficient than F. Notice that all the comparisons, except for those with the TSLS or SFE IV, are made under the same set of assumptions (Assumptions 1 and 3 later). As for those with TSLS or SFE IV, the comparisons are made whenever TSLS or SFE IV is consistent.

[**Insert Figure 1 here.**]

3 The General Estimator and its Asymptotic Properties

In this section, we propose a general regression-adjusted LATE estimator for τ. 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, these 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. We then proceed to estimate the working models with estimators $\hat{\mu}^D(a,s,x)$ and $\hat{\mu}^Y(a,s,x)$. As the working models are potentially misspecified, their estimators are potentially inconsistent for the true specifications.

To further differentiate $\mu^b(\cdot)$, $\overline{\mu}^b(\cdot)$, and $\hat{\mu}^b(\cdot)$ for $b \in \{D,Y\}$, we consider an example that $\mu^D(a,s,x)$ follows a probit model, i.e., $\mu^D(a,s,x) = F_N(\tilde{\alpha}_{a,s} + x^\top \tilde{\beta}_{a,s})$, where $F_N(\cdot)$ is the standard normal CDF, and $\tilde{\alpha}_{a,s}$ and $\tilde{\beta}_{a,s}$ are the regression coefficients which are allowed to depend on assignment a and stratum s. However, the researcher does not know the correct specification and instead uses a logit model $\overline{\mu}^D(a,s,x) = \lambda(\alpha_{a,s} + x^\top \beta_{a,s})$ as the working model, where $\lambda(\cdot)$ is the logistic CDF. Then $(\alpha_{a,s}, \beta_{a,s})$ are the pseudo true values that depend on how they are estimated and can be defined as the probability limits of the chosen estimator $(\hat{\alpha}_{a,s}, \hat{\beta}_{a,s})$. For instance, we can estimate the regression coefficients in the logistic model via logistic quasi MLE or nonlinear least squares. As the logistic model is misspecified, the two estimation methods lead to two different pseudo true values. Suppose we estimate $(\alpha_{a,s}, \beta_{a,s})$ by quasi MLE and denote their estimators as $(\hat{\alpha}_{a,s}, \hat{\beta}_{a,s})$. The estimator of the working model is then $\hat{\mu}^D(a,s,x) = \lambda(\hat{\alpha}_{a,s} + x^\top \hat{\beta}_{a,s})$.

In CAR, the targeted assignment probability for stratum s, $\pi(s)$, is usually known or can be consistently estimated by $\hat{\pi}(s) := \frac{n_1(s)}{n(s)}$. Then our proposed estimator of LATE based on the doubly robust moments[10] is

$$\hat{\tau} := \left(\frac{1}{n}\sum_{i\in[n]} \Xi_{H,i}\right)^{-1} \left(\frac{1}{n}\sum_{i\in[n]} \Xi_{G,i}\right), \quad \text{where} \tag{3.1}$$

[10] For reference 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.

$$\Xi_{H,i} := \frac{A_i(D_i - \hat{\mu}^D(1,S_i,X_i))}{\hat{\pi}(S_i)} - \frac{(1-A_i)(D_i - \hat{\mu}^D(0,S_i,X_i))}{1-\hat{\pi}(S_i)} + \hat{\mu}^D(1,S_i,X_i) - \hat{\mu}^D(0,S_i,X_i), \tag{3.2}$$

$$\Xi_{G,i} := \frac{A_i(Y_i - \hat{\mu}^Y(1,S_i,X_i))}{\hat{\pi}(S_i)} - \frac{(1-A_i)(Y_i - \hat{\mu}^Y(0,S_i,X_i))}{1-\hat{\pi}(S_i)} + \hat{\mu}^Y(1,S_i,X_i) - \hat{\mu}^Y(0,S_i,X_i). \tag{3.3}$$

Given the double robustness and the consistency of $\hat{\pi}(s)$, our estimator $\hat{\tau}$ is consistent even when the working models $(\hat{\mu}^D(\cdot), \hat{\mu}^Y(\cdot))$ are misspecified. Our analysis also 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. Furthermore, our general adjusted estimator is numerically invariant to the stratum-specific location shift because

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

Therefore, using adjustments $\hat{\mu}^b(a,S_i,X_i)$ and $\hat{\mu}^b(a,S_i,X_i) - \mathbb{E}(\mu^b(a,S_i,X_i)|S_i)$ for $b \in \{D,Y\}$ are numerically equivalent.

Assumption 3. *(i) For $a = 0,1$ and $s \in \mathcal{S}$, define $I_a(s) := \big\{i \in [n] : A_i = a, S_i = s\big\}$,*

$$\Delta^Y(a,s,X_i) := \hat{\mu}^Y(a,s,X_i) - \overline{\mu}^Y(a,s,X_i), \quad \textit{and}$$
$$\Delta^D(a,s,X_i) := \hat{\mu}^D(a,s,X_i) - \overline{\mu}^D(a,s,X_i).$$

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

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

Assumption 3 requires $\hat{\mu}^b(\cdot)$ to be a consistent estimator of $\overline{\mu}^b(\cdot)$ for $b = D, Y$. For instance, we can consider a linear working model $\overline{\mu}^Y(a, s, X_i) = X_i^\top \beta_{a,s}$, where the pseudo true value $\beta_{a,s}$ is defined as the probability limit of the OLS estimator $\hat{\beta}_{a,s}$ from regressing Y_i on X_i using observations in $I_a(s)$. Then, the estimator $\hat{\mu}^Y(a, s, X_i)$ can be written as $X_i^\top \hat{\beta}_{a,s}$, and Assumption 3(i) reduces to

$$\max_{s\in\mathcal{S},a=0,1}\left|\left(\frac{1}{n_1(s)}\sum_{i\in I_1(s)} X_i - \frac{1}{n_0(s)}\sum_{i\in I_0(s)} X_i\right)^\top (\hat{\beta}_{a,s} - \beta_{a,s})\right| = o_p(n^{-1/2}), \tag{3.4}$$

which holds automatically because by definition, $\hat{\beta}_{a,s} \xrightarrow{p} \beta_{a,s}$, and we will assume $\mathbb{E}X_i^2 < \infty$. This example shows that we do not need to assume the working model $\overline{\mu}^Y(a, s, X_i) = X_i^\top \beta_{a,s}$ is correctly specified. A similar remark applies to Assumption 3(ii) and nonlinear working models such as the logistic regression mentioned earlier. We verify Assumption 3 for general parametric adjustments in Section 5.1 below.

To state our first main result below, we need to introduce extra notation. Let $\mathcal{D}_i := \{Y_i(1), Y_i(0), D_i(1), D_i(0), X_i\}$, $W_i := Y_i(D_i(1))$, $Z_i := Y_i(D_i(0))$, $\tilde{W}_i := W_i - \mathbb{E}[W_i|S_i]$, $\tilde{Z}_i := Z_i - \mathbb{E}[Z_i|S_i]$, $\tilde{X}_i := X_i - \mathbb{E}[X_i|S_i]$, $\tilde{D}_i(a) := D_i(a) - \mathbb{E}[D_i(a)|S_i]$ for $a = 0, 1$, and

$$\tilde{\mu}^b(a, S_i, X_i) := \overline{\mu}^b(a, S_i, X_i) - \mathbb{E}\left[\overline{\mu}^Y(a, S_i, X_i)|S_i\right], \quad b \in \{D, Y\}. \tag{3.5}$$

Theorem 3.1. *(i) Suppose Assumptions 1 and 3 hold, then*

$$\sqrt{n}(\hat{\tau} - \tau) \rightsquigarrow \mathcal{N}(0, \sigma^2), \quad \textit{where} \quad \sigma^2 := \frac{\sigma_1^2 + \sigma_0^2 + \sigma_2^2}{\mathbb{P}(D(1) > D(0))^2}, \tag{3.6}$$

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

and $\Xi_1(\mathcal{D}_i, S_i)$, $\Xi_0(\mathcal{D}_i, S_i)$, and $\Xi_2(S_i)$ are defined as

$$\begin{aligned}\Xi_1(\mathcal{D}_i, S_i) := &\left[\left(1 - \frac{1}{\pi(S_i)}\right)\tilde{\mu}^Y(1, S_i, X_i) - \tilde{\mu}^Y(0, S_i, X_i) + \frac{\tilde{W}_i}{\pi(S_i)}\right] \\ &- \tau\left[\left(1 - \frac{1}{\pi(S_i)}\right)\tilde{\mu}^D(1, S_i, X_i) - \tilde{\mu}^D(0, S_i, X_i) + \frac{\tilde{D}_i(1)}{\pi(S_i)}\right],\end{aligned} \tag{3.7}$$

$$\Xi_0(\mathcal{D}_i, S_i) := \left[\left(\frac{1}{1-\pi(S_i)} - 1\right)\tilde{\mu}^Y(0, S_i, X_i) + \tilde{\mu}^Y(1, S_i, X_i) - \frac{\tilde{Z}_i}{1-\pi(S_i)}\right]$$
$$- \tau\left[\left(\frac{1}{1-\pi(S_i)} - 1\right)\tilde{\mu}^D(0, S_i, X_i) + \tilde{\mu}^D(1, S_i, X_i) - \frac{\tilde{D}_i(0)}{1-\pi(S_i)}\right], \tag{3.8}$$
$$\Xi_2(S_i) := \left(\mathbb{E}[W_i - Z_i|S_i] - \mathbb{E}[W_i - Z_i]\right) - \tau\left(\mathbb{E}[D_i(1) - D_i(0)|S_i] - \mathbb{E}[D_i(1) - D_i(0)]\right). \tag{3.9}$$

(ii) Next, we define $\hat{\sigma}^2$ *as*

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

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

Then, we have $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$.

(iii) 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 SEB.*

Several remarks are in order. First, Theorem 3.1(i) establishes the limiting distribution of our adjusted LATE estimator, which also implies its consistency. Our estimator inherits the advantage of the TSLS estimator because it remains consistent even when the adjustment $\overline{\mu}^b(\cdot)$ is misspecified, but avoids its limitation because our estimator remains consistent when $\pi(s)$ varies across strata. Additionally, the terms σ_0^2, σ_1^2, and σ_2^2 in the asymptotic variance of our regression-adjusted LATE estimator represent the sampling variations from the control units within each stratum, the treatment units within each stratum, and the strata itself, respectively.

Second, Theorem 3.1(ii) gives a consistent estimator of this asymptotic variance, which depends on the working model $\overline{\mu}^b(a, s, x)$ for $(a, b) \in \{0, 1\} \times \{D, Y\}$. Different working models lead to different estimation efficiencies.

Third, Theorem 3.1(iii) further shows that our general regression-adjusted estimator achieves the semiparametric efficiency bound $\underline{\sigma}^2$ derived in Theorem 4.1 below when the working models are correctly specified.

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

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

In this case, our estimator coincides numerically with Bugni and Gao's (2023) fully saturated estimator (i.e., NA). Indeed, we can verify that σ^2 defined above is the same as the asymptotic variance of the fully saturated estimator derived by Bugni and Gao (2023).

4 Semiparametric Efficiency Bound

Theorem 4.1. *Suppose that Assumption 1 and the regularity conditions in Assumption S.H.1 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 (3.7)–(3.9), 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 SEB 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 $\underline{\sigma}_1^2 := \mathbb{E}\left[\pi(S_i)\underline{\Xi}_1^2(\mathcal{D}_i, S_i)\right]$, $\underline{\sigma}_0^2 := \mathbb{E}\left[(1 - \pi(S_i))\underline{\Xi}_0^2(\mathcal{D}_i, S_i)\right]$, and $\underline{\sigma}_2^2 := \mathbb{E}\underline{\Xi}_2^2(S_i)$.*

Several remarks are in order. First, Theorem 4.1 suggests that the asymptotic variance of any regular root-n consistent and asymptotically normal semiparametric estimator of

LATE is bounded from below by $\underline{\sigma}^2$. Second, the proof of Theorem 4.1 follows the arguments of Armstrong (2022), which accounts for the cross-sectional dependence of $\{A_i\}_{i\in[n]}$. Third, the efficiency bound here matches the one derived by Frölich (2007) under uncounfoundedness for observational data if their propensity score depends only on S_i. Fourth, Theorem 4.1 implies that various CARs (with or without achieving strong balance) lead to the same SEB for LATE estimation. Such a result is consistent with what Armstrong (2022) found for ATE under general randomization schemes.

5 Specific Adjustment Frameworks

5.1 Parametric Working Model

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 that

$$\overline{\mu}^Y(a,S_i,X_i) = \sum_{s\in\mathcal{S}} 1\{S_i = s\}\Lambda^Y_{a,s}(X_i,\theta_{a,s}) \quad \text{and} \quad \overline{\mu}^D(a,S_i,X_i) = \sum_{s\in\mathcal{S}} 1\{S_i = s\}\Lambda^D_{a,s}(X_i,\beta_{a,s}), \tag{5.1}$$

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 methods of estimation. As mentioned above, because 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 5.1.1, 5.1.2, and 5.1.3 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

$$\hat{\mu}^Y(a,s,X_i) = \Lambda^Y_{a,s}(X_i,\hat{\theta}_{a,s}) \quad \text{and} \quad \hat{\mu}^D(a,s,X_i) = \Lambda^D_{a,s}(X_i,\hat{\beta}_{a,s}). \tag{5.2}$$

Assumption 4. *(i) Suppose that* $\max_{a=0,1,s\in\mathcal{S}} ||\hat{\theta}_{a,s}-\theta_{a,s}||_2 \xrightarrow{p} 0$ *and* $\max_{a=0,1,s\in\mathcal{S}} ||\hat{\beta}_{a,s}-\beta_{a,s}||_2 \xrightarrow{p} 0$*, where* $\|\cdot\|_2$ *is the Euclidean norm.*

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

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

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

Assumption 4(i) means that $(\hat{\theta}_{a,s}, \hat{\beta}_{a,s})$ are consistent estimators for $(\theta_{a,s}, \beta_{a,s})$. Assumption 4(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 5.1. *Suppose Assumption 4 hold. Then* $\overline{\mu}^b(a, s, X_i)$ *and* $\hat{\mu}^b(a, s, X_i)$ *defined in* (5.1) *and* (5.2)*, respectively, satisfy Assumption 3.*

Theorem 5.1 generalizes the intuition in (3.4) and shows that Assumption 3 holds for general parametric models as long as the parameters are consistently estimated.

5.1.1 Optimal Linear Adjustments

In this section, we consider working models that are linear in $\Psi_{i,s}$ where $\Psi_{i,s} = \Psi_s(X_i)$ is a function of X_i and its functional form can vary across $s \in \mathcal{S}$. Specifically, suppose, for $a = 0, 1$ and $s \in \mathcal{S}$, that $\overline{\mu}^Y(a, s, X) = \Psi_{i,s}^\top t_{a,s}$ and $\overline{\mu}^D(a, s, X) = \Psi_{i,s}^\top b_{a,s}$, where $t_{a,s}$ and $b_{a,s}$ are the regression coefficients whose values are freely chosen by the researchers. 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 $\overline{\mu}^D(a, s, X)$.

Given that all values of $t_{a,s}$ and $b_{a,s}$ lead to consistent estimators of LATE, a natural question to ask is what values give the most precise estimator. Let the asymptotic variance of the adjusted LATE estimator $\hat{\tau}$ be as σ^2, which depends on $(\overline{\mu}^Y(a,s,X), \overline{\mu}^D(a,s,X))$, and thus, $(t_{a,s}, b_{a,s})$. Let Θ^* be the collection of optimal linear coefficients that minimize the asymptotic variance of $\hat{\tau}$ over all possible $(t_{a,s}, b_{a,s})$, i.e.,

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

Assumption 5. *Suppose that $\mathbb{E}(||\Psi_{i,s}||_2^q|S_i = s) \leq C < \infty$ 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 5 requires that the regressor $\Psi_{i,s}$ does not contain a constant term. In fact, (3.2) and (3.3) imply that our estimator is numerically invariant to a stratum-specific location shift. The following theorem characterizes the set of optimal linear coefficients.

Theorem 5.2. *Suppose that Assumptions 1 and 5 hold. Then, we have*

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

$$\begin{aligned} \theta^L_{a,s} &= [\mathbb{E}(\tilde{\Psi}_{i,s}\tilde{\Psi}_{i,s}^\top|S_i = s)]^{-1}[\mathbb{E}(\tilde{\Psi}_{i,s}Y_i(D_i(a))|S_i = s)] \\ \beta^L_{a,s} &= [\mathbb{E}(\tilde{\Psi}_{i,s}\tilde{\Psi}_{i,s}^\top|S_i = s)]^{-1}[\mathbb{E}(\tilde{\Psi}_{i,s}D_i(a)|S_i = s)]. \end{aligned} \tag{5.3}$$

The optimality result in Theorem 5.2 relies 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. It is possible to have nonlinear adjustments that are more efficient. We will come back to this point in Sections 5.1.2, 5.1.3, and S.C.

In view of Theorem 5.2, 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 $(\theta^L_{a,s}, \beta^L_{a,s})$ with the corresponding optimal linear adjustments

$$\overline{\mu}^Y(a,s,X_i) = \Psi_{i,s}^\top \theta^L_{a,s} \quad \text{and} \quad \overline{\mu}^D(a,s,X_i) = \Psi_{i,s}^\top \beta^L_{a,s}. \tag{5.4}$$

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

$$\begin{aligned}
\dot{\Psi}_{i,a,s} &:= \Psi_{i,s} - \frac{1}{n_a(s)} \sum_{j \in I_a(s)} \Psi_{j,s} \\
\hat{\theta}^L_{a,s} &:= \left(\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \dot{\Psi}_{i,a,s} \dot{\Psi}_{i,a,s}^\top \right)^{-1} \left(\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \dot{\Psi}_{i,a,s} Y_i \right) \\
\hat{\beta}^L_{a,s} &:= \left(\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \dot{\Psi}_{i,a,s} \dot{\Psi}_{i,a,s}^\top \right)^{-1} \left(\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \dot{\Psi}_{i,a,s} D_i \right).
\end{aligned} \tag{5.5}$$

Then, the feasible linear adjustments can be defined as

$$\hat{\mu}^Y(a,s,X_i) = \Psi_{i,s}^\top \hat{\theta}^L_{a,s} \quad \text{and} \quad \hat{\mu}^D(a,s,X_i) = \Psi_{i,s}^\top \hat{\beta}^L_{a,s}. \tag{5.6}$$

It is clear that $\hat{\theta}^L_{a,s}$ and $\hat{\beta}^L_{a,s}$ are the OLS-estimated slopes of the following two linear regressions using observations in $I_a(s)$:

$$Y_i \sim \gamma^Y_{a,s} + \Psi_{i,s}^\top \theta_{a,s} \quad \text{and} \quad D_i \sim \gamma^D_{a,s} + \Psi_{i,s}^\top \beta_{a,s}. \tag{5.7}$$

Theorem 5.3. *Suppose that 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 (5.4) *and* (5.6)*, respectively, satisfy Assumption 3. 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* (5.6) *as* $\hat{\tau}_L$*. Then, all the results in Theorem 3.1(i)-(ii) hold for* $\hat{\tau}_L$*. In addition,* $\hat{\tau}_L$ *is the most efficient among* *all linearly adjusted LATE estimators, and in particular, weakly more efficient than the LATE estimator with no adjustments. In the special case that $\pi(s)$ is homogeneous across strata and $\Psi_{i,s} = X_i$ so that the TSLS estimator $\hat{\tau}_{TSLS}$ is consistent, $\hat{\tau}_L$ is also weakly more efficient than $\hat{\tau}_{TSLS}$.*

The asymptotic variance of the LATE estimator with the optimal linear adjustments ($\hat{\tau}_L$) takes the form of (3.6) with $\{\overline{\mu}^b(a,s,X_i)\}_{b=D,Y,a=0,1,s\in\mathcal{S}}$ in (3.7)–(3.9) defined in (5.4). It is also guaranteed to be weakly smaller than that of both $\hat{\tau}_{NA}$ and $\hat{\tau}_{TSLS}$, which addresses the Freedman's critique (Freedman, 2008a, 2008b). When $\Psi_{i,s} = X_i$, this asymptotic variance is the same as that of Ansel et al. (2018)'s S estimator, as shown in Section S.B of the Online Supplement. This implies the S estimator is the most efficient LATE estimator adjusted by linear functions of X_i, and thus, more efficient than $\hat{\tau}_{TSLS}$ and $\hat{\tau}_{NA}$.

5.1.2 Linear and Logistic Regressions

It is also common 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 $\mathring{\Psi}_{i,s} = (1, \Psi_{i,s}^\top)^\top$, $\Psi_{i,s} = \Psi_s(X_i)$ 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 (quasi) MLE, respectively, and call this method the nonlinear (logistic) adjustment. Specifically, define

$$\hat{\mu}^Y(a,s,X_i) = \mathring{\Psi}_{i,s}^\top \hat{\theta}_{a,s}^{OLS} \quad \text{and} \quad \hat{\mu}^D(a,s,X_i) = \lambda(\mathring{\Psi}_{i,s}^\top \hat{\beta}_{a,s}^{MLE}), \tag{5.8}$$

where

$$\hat{\theta}_{a,s}^{OLS} = \left(\frac{1}{n_a(s)} \sum_{i\in I_a(s)} \mathring{\Psi}_{i,s}\mathring{\Psi}_{i,s}^\top\right)^{-1} \left(\frac{1}{n_a(s)} \sum_{i\in I_a(s)} \mathring{\Psi}_{i,s} Y_i\right) \quad \text{and}$$

$$\hat{\beta}_{a,s}^{MLE} = \arg\max_{b} \frac{1}{n_a(s)} \sum_{i \in I_a(s)} \left[D_i \log(\lambda(\mathring{\Psi}_{i,s}^\top b)) + (1 - D_i) \log(1 - \lambda(\mathring{\Psi}_{i,s}^\top b)) \right]. \tag{5.9}$$

It is clear that $\hat{\theta}_{a,s}^{OLS}$ and $\hat{\beta}_{a,s}^{MLE}$ are the OLS and ML estimates of the following two stratum-specific (logistic) regressions using observations in $I_a(s)$:

$$Y_i \sim \mathring{\Psi}_{i,s}^\top \theta_{a,s} \quad \text{and} \quad D_i \sim \lambda^{-1}(\mathring{\Psi}_{i,s}^\top \beta_{a,s}). \tag{5.10}$$

In the logistic regression, 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 $\hat{h}_{a,s}^{OLS}$ is the intercept. 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 have $\underline{\hat{\theta}}_{a,s}^{OLS} = \hat{\theta}_{a,s}^{L}$ by construction. This means $\hat{\mu}^Y(a,s,X_i)$ used here is the same as that for the optimal linear adjustment. In contrast, because the logistic regression is nonlinear, the non-intercept part of $\hat{\beta}_{a,s}^{MLE}$ does not equal $\hat{\beta}_{a,s}^{L}$. 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 the working models are

$$\overline{\mu}^Y(a,s,X_i) = \mathring{\Psi}_{i,s}^\top \theta_{a,s}^{OLS} \quad \text{and} \quad \overline{\mu}^D(a,s,X_i) = \lambda(\mathring{\Psi}_{i,s}^\top \beta_{a,s}^{MLE}). \tag{5.11}$$

Assumption 6. *(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 5.4. *Suppose Assumptions 1 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 (5.11) and (5.8), respectively, satisfy Assumption 3. 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 (5.8) as $\hat{\tau}_{NL}$. Then, all the results in Theorem 3.1(i)-(ii) hold for $\hat{\tau}_{NL}$.

Several remarks are in order. First, the nonlinear (logistic) adjustment is not optimal in the sense that it does not necessarily minimize the asymptotic variance of the corresponding LATE estimator over the class of linear/logistic adjustments. Second, the nonlinear (logistic) adjustment is not necessarily less efficient than the optimal linear adjustment studied in Section 5.1.1 as the true specification $\mu^D(a,s,X_i)$ could be nonlinear. In fact, as Theorem 3.1 shows, if the adjustments are correctly specified, then $\hat{\tau}_{NL}$ can achieve the semiparametric efficiency bound. Compared with the linear probability model considered in Section 5.1.1, the logistic model is expected to be less misspecified, especially when the regressor $\Psi_{i,s}$ contains nonlinear transformations of X_i such as interactions and quadratic terms. Third, we will further justify the intuition above in Section S.C, in which we let $\Psi_{i,s}$ be the sieve basis functions with an increasing dimension and show that the nonlinear (logistic) method can consistently estimate the correct specification under some regularity conditions. Fourth, one theoretical shortcoming of the nonlinear (logistic) 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 5.1.3 below.

5.1.3 Further Efficiency Improvement

Following the lead of Cohen and Fogarty (2023), we can treat the nonlinear (logistic) adjustments as regressors and obtain the optimal linear coefficients as proposed in Section 5.1.1.[11] Let $\theta_{a,s}^{OLS} = (h_{a,s}^{OLS}, \underline{\theta}_{a,s}^{OLS})$ be the probability limit of $\hat{\theta}_{a,s}^{OLS}$ defined in (5.9). If $\beta_{a,s}^{MLE}$ were known, the nonlinear (logistic) adjustment can be viewed as a linear adjustment.

[11] Cohen and Fogarty (2023)'s setting is different from ours as they consider neither CARs nor LATE.

Specifically, denote

$$\Phi_{i,s} := (\Psi_{i,s}^\top, \lambda(\mathring{\Psi}_{i,s}^\top \beta_{1,s}^{MLE}), \lambda(\mathring{\Psi}_{i,s}^\top \beta_{0,s}^{MLE}))^\top \tag{5.12}$$

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

where d_Ψ is the dimension of $\Psi_{i,s}$. Then, the nonlinear (logistic) adjustment can be written as

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

Similarly, we can replicate no adjustments and the optimal linear adjustments with $\Phi_{i,s}$ defined in (5.12) 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 $(t_{a,s}, b_{a,s}) = (t_{a,s}^L, b_{a,s}^L)$, respectively, where

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

Based on Theorem 5.2, we can further improve all three types of adjustments by setting the linear coefficients of $\Phi_{i,s}$ as

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

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

$$\overline{\mu}^Y(a,s,X_i) = \Phi_{i,s}^\top \theta_{a,s}^F \quad \text{and} \quad \overline{\mu}^D(a,s,X_i) = \Phi_{i,s}^\top \beta_{a,s}^F. \tag{5.13}$$

Because $\beta_{a,s}^{MLE}$ is unknown, we can replace it by its estimate proposed in Section 5.1.2, i.e., define

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

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

$$\begin{aligned}
\hat{\theta}_{a,s}^F &:= \left(\frac{1}{n_a(s)}\sum_{i\in I_a(s)}\breve{\Phi}_{i,a,s}\breve{\Phi}_{i,a,s}^{\top}\right)^{-1}\left(\frac{1}{n_a(s)}\sum_{i\in I_a(s)}\breve{\Phi}_{i,a,s}Y_i\right), \\
\hat{\beta}_{a,s}^F &:= \left(\frac{1}{n_a(s)}\sum_{i\in I_a(s)}\breve{\Phi}_{i,a,s}\breve{\Phi}_{i,a,s}^{\top}\right)^{-1}\left(\frac{1}{n_a(s)}\sum_{i\in I_a(s)}\breve{\Phi}_{i,a,s}D_i\right).
\end{aligned} \tag{5.14}$$

The corresponding feasible adjustments are

$$\hat{\mu}^Y(a,s,X_i) = \hat{\Phi}_{i,s}^{\top}\hat{\theta}_{a,s}^F \quad \text{and} \quad \hat{\mu}^D(a,s,X_i) = \hat{\Phi}_{i,s}^{\top}\hat{\beta}_{a,s}^F. \tag{5.15}$$

Assumption 7. *Suppose Assumption 5 holds for $\Phi_{i,s}$ defined in (5.12).*

Theorem 5.5. *Suppose that Assumptions 1, 6, and 7 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 (5.13) and (5.15), respectively, satisfy Assumption 3. 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 (5.15) as $\hat{\tau}_F$. Then, all the results in Theorem 3.1(i)-(ii) hold for $\hat{\tau}_F$. In addition, $\hat{\tau}_F$ is weakly more efficient than $\hat{\tau}_L$, $\hat{\tau}_{NL}$ and $\hat{\tau}_{NA}$.

Theorem 5.5 shows that by refitting nonlinear (logistic) adjustment in a linear regression with optimal linear coefficients, we can further improve the efficiency of the adjusted LATE estimator. Moreover, $\hat{\tau}_F$ is guaranteed to be weakly more efficient than $\hat{\tau}_L$, $\hat{\tau}_{NL}$ and $\hat{\tau}_{NA}$.

5.2 Regularized Large Dimensional Regression

We consider the nonparametric regression as the adjustments for our LATE estimator in Section S.C of the Online Supplement, while in this section, we consider the case where the regressor $\mathring{\Psi}_{i,n} \in \Re^{p_n}$ has dimension p_n that can be much higher than n. In this case, we can no longer use the nonlinear (logistic) (nonparametric) adjustment method. Instead, we need to regularize the least squares and logistic regressions. Specifically, let

$$\hat{\mu}^Y(a,s,X_i) = \mathring{\Psi}_{i,n}^\top \hat{\theta}_{a,s}^R \quad \text{and} \quad \hat{\mu}^D(a,s,X_i) = \lambda(\mathring{\Psi}_{i,n}^\top \hat{\beta}_{a,s}^R), \tag{5.16}$$

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

$$\hat{\theta}_{a,s}^R = \arg\min_t \frac{-1}{n_a(s)} \sum_{i \in I_a(s)} \left(Y_i - \mathring{\Psi}_{i,n}^\top t\right)^2 + \frac{\varrho_{n,a}(s)}{n_a(s)} ||\hat{\Omega}^Y t||_1,$$
$$\hat{\beta}_{a,s}^R = \arg\min_b \frac{-1}{n_a(s)} \sum_{i \in I_a(s)} \left[D_i \log\left(\lambda(\mathring{\Psi}_{i,n}^\top b)\right) + (1-D_i)\log(1-\lambda\left(\mathring{\Psi}_{i,n}^\top b)\right)\right] + \frac{\varrho_{n,a}(s)}{n_a(s)} ||\hat{\Omega}^D b||_1,$$

where $\{\varrho_{n,a}(s)\}_{a=0,1,s\in\mathcal{S}}$ are tuning parameters, $\hat{\Omega}^b = \text{diag}(\hat{\omega}_1^b, \cdots, \hat{\omega}_{p_n}^b)$ is a diagonal matrix of data-dependent penalty loadings for $b = D, Y$, and $\|\cdot\|_1$ is the ℓ_1 norm.[12]

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

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

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

such that $\max_{a=0,1,s\in\mathcal{S}} \max(||\theta_{a,s}^R||_0, ||\beta_{a,s}^R||_0) \le h_n$, where $||a||_0$ denotes the number of nonzero components in a.

(ii) Suppose that for $q > 2$,

$$\sup_{i\in[n]} ||\mathring{\Psi}_{i,n}||_\infty \le \zeta_n \ a.s. \quad \textit{and} \quad \sup_{h\in[p_n]} \mathbb{E}\left[|\mathring{\Psi}_{i,n,h}^q||S_i = s\right] < \infty,$$

[12] We provide more details about $\hat{\Omega}^b$ in Section S.D of the Online Supplement.

where $\|\cdot\|_\infty$ *is the* ℓ_∞ *norm.*

(iii) Suppose that

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

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

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 that $\frac{\log(p_n)\zeta_n^2 h_n^2}{n} \to 0$ *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*

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

(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,*

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

and

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

(vii) For $a=0,1$, *let* $\varrho_{n,a}(s) = c\sqrt{n_a(s)}F_N^{-1}\left(1-1/\left[p_n\log(n_a(s))\right]\right)$ *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-Val, and Hansen (2017) for more discussion.

Theorem 5.6. *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* (5.16) *and* $\overline{\mu}^b(a,s,X) = \mu^b(a,s,X)$ *satisfy Assumption 3. All the results in Theorem 3.1(i)-(ii) hold for* $\hat{\tau}_R$*. In addition,* $\hat{\tau}_R$ *achieves the SEB.*

Due to the approximate sparsity, the Lasso method consistently estimates the correct specification, which explains why the corresponding estimator can achieve the SEB.

6 Simulations

6.1 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

$$
\begin{aligned}
&Y_i(d) = a_d + \alpha(X_i,Z_i) + \varepsilon_{d+1,i}, \ d=0,1, \quad D_i(0) = 1\{b_0 + \gamma(X_i,Z_i) > c_0\varepsilon_{3,i}\},\\
&D_i(1) = \begin{cases} 1\{b_1 + \gamma(X_i,Z_i) > c_1\varepsilon_{4,i}\} & \text{if } D_i(0)=0,\\ 1 & \text{otherwise},\end{cases}\\
&D_i = D_i(1)A_i + D_i(0)(1-A_i), \quad \text{and} \quad Y_i = Y_i(1)D_i + Y_i(0)(1-D_i),
\end{aligned}
$$

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 specified as follows.

(i) Let Z_i be i.i.d. according to standardized Beta(2, 2), $S_i = \sum_{j=1}^4 1\{Z_i \le g_j\}$, 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

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

$a_1 = 2, a_0 = 1, b_1 = 1.3, b_0 = -1, c_1 = c_0 = 3$, 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

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

$a_1 = 2, a_0 = 1, b_1 = 1, b_0 = -1, c_1 = c_0 = 3$, 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 four randomization schemes (SRS, WEI, BCD, SBR) defined as in Examples 1–4 in Appendix S.A, respectively. Specifically, for WEI and BCD, we set $f(x) = (1-x)/2$ and $\lambda = 0.75$, respectively.

We compute the true LATE effect τ_0 using Monte Carlo simulations, with sample size being 10,000 and the number of Monte Carlo simulations being 1,000. We gauge the size and power of various tests by testing the hypotheses $H_0 : \tau = \tau_0$ and $H_0 : \tau = \tau_0 + 1$, respectively. All the tests are carried out at 5% level of significance, and with the number of Monte Carlo simulations being 10,000.

6.2 Estimators for Comparison

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

(i) NA: the fully saturated estimator by Bugni and Gao (2023), which is equivalent to 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: $\hat{\tau}_{TSLS}$ defined in Section 2.4. We use the usual IV heteroskedasticity-robust standard error (i.e., $\hat{\sigma}_{TSLS,naive}/\sqrt{n}$) for inference.

(iii) L: the optimal linear estimator with $\Psi_{i,s} = X_i$ and the pseudo true values being estimated by $\hat{\theta}^L_{a,s}$ and $\hat{\beta}^L_{a,s}$ defined in (5.5).

(iv) S: Ansel et al.'s (2018) S estimator with X_i as regressor. We use the standard error of the S estimator (i.e., $\hat{\sigma}_S/\sqrt{n}$; see Section S.B of the Online Supplement for details) for inference.

(v) NL: the nonlinear (logistic) 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 (5.9).

(vi) 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 (5.14).

(vii) NP: the nonparametric estimator outlined in Section S.C of the Online Supplement. The following 9 bases of a spline of order 3 are chosen as the sieve regressors:

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

$$X_{1,i}1\{X_{1,i} > t_1\}X_{2,i}1\{X_{2,i} > t_2\}\Big)^\top, \tag{6.1}$$

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.[13] The adjustments are computed as in (S.C.1) of the Online Supplement.

(viii) SNP: Ansel et al.'s (2018) S estimator with $\mathring{\Psi}_{i,n}$ defined in (6.1) as regressor. We use the standard error of the S estimator (i.e., $\hat{\sigma}_S/\sqrt{n}$; see Section S.B of the Online Supplement for details) for inference.

(ix) R: a regularized estimator. The nonparametric estimator outlined in Section S.C of the Online Supplement might not have a good size when the sample size is small, so we propose to use Lasso to select the sieve regressors. The sieves regressors $\mathring{\Psi}_{i,n}$ are the same as in (6.1). The adjustments are computed as in (5.16). The tuning parameter is chosen as: $\varrho_{n,a}(s) = 1.1\sqrt{n_a(s)}F_N^{-1}(1-1/(p_n\log(n_a(s))))$. We compute the data-driven penalty loading matrices $\hat{\Omega}^Y$ and $\hat{\Omega}^D$ following the iterative procedure proposed by Belloni et al. (2017).[14]

For DGP(iii), we consider the estimator with no adjustments (NA), and the lasso estimators $\hat{\theta}^R_{a,s}$ and $\hat{\beta}^R_{a,s}$ defined in (5.16) with $\mathring{\Psi}_{i,n} = (1, \Psi_{i,n}^\top)^\top = (1, X_i^\top)^\top$. The tuning parameters are choosing as: $\varrho_{n,a}(s) = 1.1\sqrt{n_a(s)}F_N^{-1}(1-1/(p_n\log(n_a(s))))$.

6.3 Simulation Results

Tables 2-4 present the empirical sizes and powers of the true null $H_0: \tau = \tau_0$ and false null $H_0: \tau = \tau_0 + 1$ under DGPs (i)-(iii), respectively. We also report the ratio of the median length of the confidence intervals of a particular estimator to that of the NA estimator is in the corresponding bracket. Note that none of the working models in DGPs (i)-(iii) is correctly specified. Consider DGP (i). When $n = 200$, both the NA and TSLS estimators are slightly under-sized. Both the NP and SNP estimators are oversized because the numbers of sieve regressors are relatively large compared to the

[13] The formal definition of spline is given in Section S.D of the Online Supplement.
[14] `Matlab` code provided by Belloni et al. (2017) and the `R` package "hdm" provide a built-in option for this iterative procedure.

sample size, while the R estimator has the correct size thanks to the Lasso selection of the sieve regressors. The L estimator performs the same as the S estimator. All other estimators have sizes close to the nominal level of 5%. This confirms that our estimation and inference procedures are robust to misspecification.

[**Insert Table 2 here**.]

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. The powers of the other estimators are much higher. In particular, the power of the F estimator is higher than those of the NA, TSLS, L, and NL estimators, which is consistent with our theory that the F estimator is weakly more efficient than those estimators. The NP, SNP, and R estimators enjoy the highest powers as a nonparametric model could approximate the true specification very well. The NP and SNP estimators have more size distortions than the R estimator when the sample size is 200. When the sample size is increased to 400, virtually all the sizes and powers of the estimators improve, and all the observations continue to hold.

[**Insert Table 3 here**.]

We also report the ratio of the median length of the confidence intervals of a particular estimator to that of the NA estimator in the corresponding parentheses. Generally speaking, the confidence intervals of the TSLS and adjusted estimators (L, NL, F, NP, and R) are 20%-30% shorter, in terms of the median, than that of the NA estimator.

Most observations uncovered in DGP (i) carry forward to DGP (ii). Two new patterns emerge. First, the powers of the L, S, NL, F, NP, SNP, and R estimators are much higher than those of the NA and TSLS estimators. Second, the ratio of the median length of the confidence intervals of the TSLS estimator is as wide as that of the NA estimator, whereas the confidence intervals of the adjusted estimators (L, NL, F, NP, and R) become 25%-40% shorter, in terms of the median, than that of the NA estimator. 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. The confidence intervals of the R estimator are 60%-65% shorter, in terms of the median, than that of the NA estimator.

[**Insert Table 4 here**.]

In Section S.R of the Online Supplement, we simulate data with heterogeneous $\{\pi(s)\}$. We find that all estimators except TSLS have their empirical rejection rates close to the nominal size of 5% under the null. TSLS, on the other hand, has around 15% rejection rate when $n = 1200$. This indicates the TSLS estimator can be inconsistent when $\{\pi(s)\}$ are heterogeneous, in line with Theorem 2.1.

6.4 Practical Recommendation

If researchers want to use parametric adjustments without tuning parameters, we recommend the F estimator, which is guaranteed to be weakly more efficient than TSLS, L, and NL estimators. Regressors $\Psi_{i,s}$ can include linear, quadratic and interaction terms of the original covariates. If researchers want to achieve the SEB by using sieve bases and/or the dimension of covariates is high relative to the sample size, we recommend the R estimator. In general, the F and R estimators tend to have similar size, but the R estimator tends to have better power.

7 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.[15] 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,

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

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 balance– 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 those households with treatment assignment, 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) of Assumption 1. The target fraction of treatment assignment 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.[16] To maintain comparability, for each outcome variable, we also keep X_i similar to those used in Dupas et al. (2018) for our adjusted estimators.[17] Due to the low

[16] 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). We use data from the first follow-up survey and exclude other cash savings because only 2% of the households in the sample reported having it.

[17] The description of these estimators is similar to that in Section 6. Except for savings in formal financial institutions, mobile money, and total formal savings, X_i includes baseline value for the outcome of interest, baseline value of total income, and a dummy for missing observations. For savings in formal financial institutions, mobile money, and total formal savings, since their baseline values are all zero, we

dimension of covariates used in the regression adjustments, we focus on the performance of the methods "NA", "TSLS", "L", "NL", and "F".

Table 5 presents the LATE estimates and their standard errors (in parentheses) estimated by these methods.[18] 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 eight outcome variables, the standard errors estimated by regression adjustments are on average around 8% 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 18% lower than those without adjustment. This means that regression adjustments, with the similar covariates used in Dupas et al. (2018), can achieve sizable efficiency gains in estimating the LATEs.

[**Insert Table 5 here.**]

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 variables are mobile money and total informal savings, the standard errors obtained via "F" method are about 7.1% and 5%, respectively, lower than those by TSLS. When the outcome variable is savings in friends/family, the standard error estimated by the optimal linear adjustment – "L" method is around 6.7% lower than that obtained by TSLS. This means that, compared with our regression-adjusted methods, TSLS is generally 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,

set X_i as the baseline value of total savings, baseline value of total income, and a dummy for missing observations.

[18] For each outcome variable, we filter out the observations with missing values of outcome variables or the strata with less than 10 observations. The total trimmed observations are less than 10% of the whole sample in Dupas et al. (2018)).

we find that reducing the cost of opening a bank account can significantly increase their savings in formal institutions. We also observe the evidence of crowd-out – mainly moving cash from saving at home to saving in bank.

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.

Atkin, D., A. K. Khandelwal, and A. Osman (2017). Exporting and firm performance: Evidence from a randomized experiment. *The quarterly journal of economics 132*(2), 551–615.

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.*

Beam, E. A. and S. Quimbo (2021). The impact of short-term employment for low-income youth: Experimental evidence from the philippines. *The Review of Economics and Statistics*, 1–45.

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

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.

Bolhaar, J., N. Ketel, and B. van Der Klaauw (2019). Job search periods for welfare applicants: Evidence from a randomized experiment. *American Economic Journal: Applied Economics 11*(1), 92–125.

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 (2023). Inference under covariate-adaptive randomization with imperfect compliance. *Journal of Econometrics forthcoming*.

Cohen, P. L. and C. B. Fogarty (2023). No-harm calibration for generalized oaxaca-blinder estimators. *Biometrika forthcoming*.

Davis, J. M. and S. B. Heller (2020). Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs. *Review of Economics and Statistics 102*(4), 664–677.

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.

Heiler, P. (2022). Efficient covariate balancing for the local average treatment effect. *Journal of Business & Economic Statistics 40*(4), 1569–1582.

Himmler, O., R. Jäckle, and P. Weinschenk (2019). Soft commitments, reminders, and academic performance. *American Economic Journal: Applied Economics 11*(2), 114–142.

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.

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

Jha, S. and M. Shayo (2019). Valuing peace: the effects of financial market exposure on votes and political attitudes. *Econometrica 87*(5), 1561–1588.

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.

Marx, B. M. and L. J. Turner (2019). Student loan nudges: Experimental evidence on borrowing and educational attainment. *American Economic Journal: Economic Policy 11*(2), 108–141.

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

Okunogbe, O. and V. Pouliquen (2022). Technology, taxation, and corruption: evidence from the introduction of electronic tax filing. *American Economic Journal: Economic Policy 14*(1), 341–72.

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.

Royer, H., M. Stehr, and J. Sydnor (2015). Incentives, commitments, and habit formation in exercise: evidence from a field experiment with workers at a fortune-500 company. *American Economic Journal: Applied Economics 7*(3), 51–84.

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., S. D. Uysal, and J. M. Wooldridge (2022). Doubly robust estimation of local average treatment effects using inverse probability weighted regression adjustment. *arXiv preprint arXiv:2208.01300*.

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*.

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.

Table 1: Empirical Papers Using CARs with Imperfect Compliance

	Journal	Method	Covariates	Strata fixed effects
Royer et al. (2015)	AEJ: Applied	TSLS	Yes	Yes
Atkin et al. (2017)	QJE	TSLS	Yes	Yes
Dupas et al. (2018)	AEJ: Applied	TSLS	Yes	Yes
Marx and Turner (2019)	AEJ: Policy	TSLS	Yes	Yes
Jha and Shayo (2019)	Econometrica	TSLS	Yes	Yes
Himmler et al. (2019)	AEJ: Applied	TSLS	Yes	Yes
Bolhaar et al. (2019)	AEJ: Applied	TSLS	Yes	Yes
Davis and Heller (2020)	ReStat	TSLS	Yes	Yes
Beam and Quimbo (2021)	ReStat	TSLS	Yes	Yes
Okunogbe and Pouliquen (2022)	AEJ: Policy	TSLS	Yes	Yes

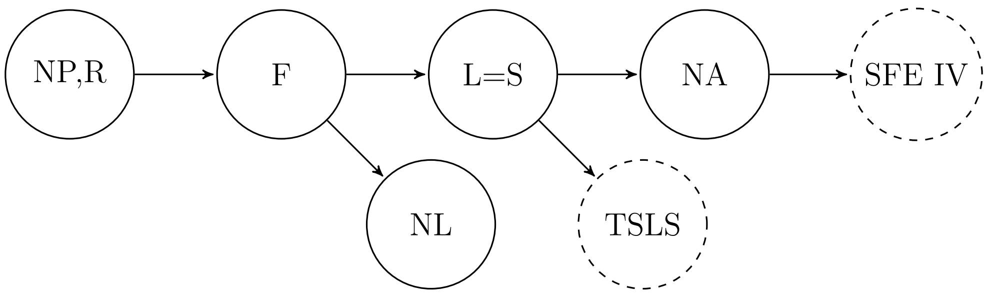

Figure 1: Efficiency of Various LATE Estimators (from the most efficient to the least).

Table 2: Size and Power for DGP(i)

	$n = 200$				$n = 400$			
Methods	SRS	WEI	BCD	SBR	SRS	WEI	BCD	SBR
Size								
NA	0.035	0.031	0.031	0.034	0.046	0.043	0.042	0.039
TSLS	0.036	0.034	0.032	0.038	0.045	0.040	0.044	0.042
	[77.8%]	[78.0%]	[77.6%]	[77.8%]	[78.0%]	[77.9%]	[77.8%]	[78.0%]
L	0.044	0.041	0.041	0.045	0.048	0.044	0.047	0.047
	[76.6%]	[76.6%]	[76.5%]	[76.5%]	[77.3%]	[77.1%]	[77.2%]	[77.4%]
S	0.044	0.041	0.041	0.045	0.048	0.044	0.047	0.047
	[76.6%]	[76.6%]	[76.5%]	[76.5%]	[77.3%]	[77.1%]	[77.2%]	[77.4%]
NL	0.044	0.042	0.040	0.045	0.049	0.045	0.047	0.047
	[77.5%]	[77.3%]	[77.2%]	[77.2%]	[77.6%]	[77.5%]	[77.5%]	[77.6%]
F	0.054	0.052	0.049	0.053	0.054	0.048	0.052	0.050
	[74.7%]	[74.9%]	[74.6%]	[74.5%]	[75.4%]	[75.3%]	[75.3%]	[75.6%]
NP	0.109	0.094	0.091	0.090	0.073	0.062	0.067	0.062
	[81.6%]	[80.5%]	[79.5%]	[79.0%]	[69.3%]	[69.4%]	[69.5%]	[69.4%]
SNP	0.100	0.091	0.090	0.085	0.070	0.061	0.063	0.060
	[73.2%]	[72.3%]	[72.0%]	[71.8%]	[68.0%]	[67.9%]	[68.1%]	[68.0%]
R	0.053	0.050	0.049	0.055	0.057	0.049	0.051	0.047
	[70.6%]	[70.3%]	[70.1%]	[70.1%]	[69.5%]	[69.5%]	[69.5%]	[69.6%]
Power								
NA	0.170	0.169	0.170	0.170	0.293	0.289	0.291	0.294
TSLS	0.260	0.254	0.260	0.255	0.430	0.433	0.443	0.436
	[77.8%]	[78.0%]	[77.6%]	[77.8%]	[78.0%]	[77.9%]	[77.8%]	[78.0%]
L	0.274	0.264	0.273	0.268	0.439	0.440	0.447	0.444
	[76.6%]	[76.6%]	[76.5%]	[76.5%]	[77.3%]	[77.1%]	[77.2%]	[77.4%]
S	0.274	0.264	0.273	0.268	0.439	0.440	0.447	0.444
	[76.6%]	[76.6%]	[76.5%]	[76.5%]	[77.3%]	[77.1%]	[77.2%]	[77.4%]
NL	0.268	0.257	0.267	0.261	0.434	0.435	0.443	0.439
	[77.5%]	[77.3%]	[77.2%]	[77.2%]	[77.6%]	[77.5%]	[77.5%]	[77.6%]
F	0.299	0.292	0.296	0.293	0.460	0.454	0.466	0.463
	[74.7%]	[74.9%]	[74.6%]	[74.5%]	[75.4%]	[75.3%]	[75.3%]	[75.6%]
NP	0.299	0.284	0.289	0.280	0.509	0.506	0.510	0.509
	[81.6%]	[80.5%]	[79.5%]	[79.0%]	[69.3%]	[69.4%]	[69.5%]	[69.4%]
SNP	0.344	0.331	0.340	0.333	0.532	0.526	0.533	0.532
	[73.2%]	[72.3%]	[72.0%]	[71.8%]	[68.0%]	[67.9%]	[68.1%]	[68.0%]
R	0.325	0.315	0.325	0.321	0.516	0.517	0.514	0.516
	[70.6%]	[70.3%]	[70.1%]	[70.1%]	[69.5%]	[69.5%]	[69.5%]	[69.6%]

Table 3: Size and Power for DGP(ii)

	$n = 200$				$n = 400$			
Methods	SRS	WEI	BCD	SBR	SRS	WEI	BCD	SBR
Size								
NA	0.033	0.031	0.029	0.030	0.045	0.042	0.043	0.041
TSLS	0.035	0.033	0.031	0.033	0.045	0.044	0.045	0.040
	[99.5%]	[99.4%]	[99.6%]	[99.4%]	[99.8%]	[99.8%]	[99.8%]	[99.7%]
L	0.044	0.040	0.044	0.038	0.049	0.047	0.046	0.046
	[74.7%]	[74.5%]	[74.8%]	[74.6%]	[75.5%]	[75.6%]	[75.6%]	[75.5%]
S	0.044	0.040	0.044	0.038	0.049	0.047	0.046	0.046
	[74.7%]	[74.5%]	[74.8%]	[74.6%]	[75.5%]	[75.6%]	[75.6%]	[75.5%]
NL	0.043	0.039	0.042	0.037	0.049	0.047	0.046	0.046
	[75.3%]	[75.1%]	[75.5%]	[75.2%]	[75.7%]	[75.8%]	[75.8%]	[75.6%]
F	0.052	0.047	0.048	0.043	0.050	0.049	0.051	0.049
	[70.2%]	[69.7%]	[70.3%]	[70.4%]	[70.8%]	[70.8%]	[70.9%]	[70.7%]
NP	0.100	0.084	0.087	0.079	0.062	0.063	0.065	0.062
	[69.2%]	[67.7%]	[67.3%]	[67.7%]	[60.4%]	[60.3%]	[60.4%]	[60.5%]
SNP	0.098	0.084	0.085	0.079	0.061	0.064	0.064	0.063
	[63.7%]	[62.8%]	[63.0%]	[62.8%]	[59.7%]	[59.8%]	[59.7%]	[59.8%]
R	0.055	0.051	0.049	0.049	0.052	0.051	0.048	0.045
	[63.3%]	[62.8%]	[63.2%]	[63.0%]	[62.1%]	[62.1%]	[62.2%]	[62.0%]
Power								
NA	0.202	0.208	0.208	0.206	0.350	0.351	0.351	0.345
TSLS	0.204	0.212	0.211	0.210	0.353	0.352	0.354	0.346
	[99.5%]	[99.4%]	[99.6%]	[99.4%]	[99.8%]	[99.8%]	[99.8%]	[99.7%]
L	0.334	0.331	0.342	0.340	0.512	0.526	0.524	0.516
	[74.7%]	[74.5%]	[74.8%]	[74.6%]	[75.5%]	[75.6%]	[75.6%]	[75.5%]
S	0.334	0.331	0.342	0.340	0.512	0.526	0.524	0.516
	[74.7%]	[74.5%]	[74.8%]	[74.6%]	[75.5%]	[75.6%]	[75.6%]	[75.5%]
NL	0.327	0.324	0.335	0.333	0.510	0.523	0.523	0.515
	[75.3%]	[75.1%]	[75.5%]	[75.2%]	[75.7%]	[75.8%]	[75.8%]	[75.6%]
F	0.372	0.374	0.379	0.375	0.562	0.568	0.566	0.561
	[70.2%]	[69.7%]	[70.3%]	[70.4%]	[70.8%]	[70.8%]	[70.9%]	[70.7%]
NP	0.378	0.381	0.387	0.386	0.649	0.663	0.653	0.655
	[69.2%]	[67.7%]	[67.3%]	[67.7%]	[60.4%]	[60.3%]	[60.4%]	[60.5%]
SNP	0.431	0.443	0.442	0.440	0.663	0.676	0.668	0.668
	[63.7%]	[62.8%]	[63.0%]	[62.8%]	[59.7%]	[59.8%]	[59.7%]	[59.8%]
R	0.419	0.429	0.431	0.432	0.644	0.661	0.657	0.648
	[63.3%]	[62.8%]	[63.2%]	[63.0%]	[62.1%]	[62.1%]	[62.2%]	[62.0%]

Table 4: Size and Power for DGP(iii)

	$n = 200$				$n = 400$			
Methods	SRS	WEI	BCD	SBR	SRS	WEI	BCD	SBR
Size								
NA	0.046	0.043	0.046	0.048	0.046	0.047	0.045	0.047
R	0.064	0.058	0.061	0.060	0.057	0.061	0.058	0.060
	[37.2%]	[36.9%]	[36.7%]	[36.8%]	[34.4%]	[34.4%]	[34.5%]	[34.5%]
Power								
NA	0.173	0.170	0.171	0.177	0.233	0.238	0.235	0.239
R	0.516	0.524	0.533	0.534	0.811	0.815	0.817	0.815
	[37.2%]	[36.9%]	[36.7%]	[36.8%]	[34.4%]	[34.4%]	[34.5%]	[34.5%]

Table 5: Impacts on Saving Stocks in 2010 US Dollars

Y	n	NA	TSLS	L	NL	F
Formal fin. inst.	1968	20.558	21.154	22.160	22.196	22.743
		(3.067)	(3.015)	(2.965)	(2.976)	(2.942)
		[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Mobile	1972	-0.208	-0.174	-0.291	-0.292	-0.302
		(0.223)	(0.224)	(0.212)	(0.213)	(0.208)
		[0.352]	[0.439]	[0.169]	[0.169]	[0.147]
Total formal	1966	20.399	21.097	21.924	21.986	22.335
		(3.089)	(3.034)	(2.979)	(2.994)	(2.956)
		[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Cash at home	1971	-10.826	-7.456	-9.004	-8.904	-8.373
		(5.003)	(4.404)	(4.401)	(4.355)	(4.354)
		[0.030]	[0.090]	[0.041]	[0.041]	[0.054]
ROSCA/ VSLA	1975	-1.933	-2.333	-1.242	-1.255	0.651
		(1.971)	(1.858)	(1.794)	(1.812)	(1.940)
		[0.327]	[0.209]	[0.489]	[0.488]	[0.737]
Friends/ family	1974	-3.621	-3.346	-1.428	-1.536	-2.067
		(2.040)	(1.999)	(1.866)	(2.015)	(2.042)
		[0.076]	[0.094]	[0.444]	[0.446]	[0.311]
Total informal	1960	-17.643	-14.317	-15.665	-15.693	-14.137
		(6.200)	(5.351)	(5.185)	(5.196)	(5.082)
		[0.004]	[0.007]	[0.003]	[0.003]	[0.005]
Total savings	1952	2.787	7.153	7.169	7.193	8.962
		(7.290)	(6.368)	(6.197)	(6.218)	(6.142)
		[0.702]	[0.261]	[0.247]	[0.247]	[0.145]

Notes: The table reports the LATE estimates of opening bank accounts on saving stocks. NA, TSLS, L, NL, and F stand for the no-adjustment, TSLS, optimal linear, nonlinear (logistic), and further efficiency improvement, respectively. n is the number of households. Standard errors are in parentheses. P-values are in square brackets.

Online Supplement for "Improving Estimation Efficiency via Regression-Adjustment in Covariate-Adaptive Randomizations with Imperfect Compliance"*

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

October 25, 2023

Abstract

Section S.A contains four commonly used Covariate-adaptive treatment assignment rules. Section S.B considers the S estimator proposed by Ansel, Hong, and Li (2018). We then examine the efficiency of $\hat{\tau}$ in the context of nonparametric adjustments in Section S.C. Section S.D provides the implementation details for sieve and Lasso regressions. Section S.E briefly discusses the regression adjustment under full compliance. Sections S.F–S.P prove Theorems 2.1–5.6, and S.B.1, respectively. Section S.Q collects technical lemmas used in the Proof of Theorem 3.1. Section S.R provides an additional simulation to demonstrate that when $\{\pi(s)\}$ are heterogeneous, the TSLS estimator could be inconsistent.

*We are grateful to Federico Bugni for his comments. Yichong Zhang acknowledges the financial support from Singapore Ministry of Education Tier 2 grant under grant MOE2018-T2-2-169, the NSFC under the grant No. 72133002, and a 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.

S.A Covariate-Adaptive Treatment Assignment Rules

Example 1 (SRS). *Let A_k be a Bernoulli random variable, independent of $\{S_i\}_{i=1,\neq k}^n$ and $\{A_i\}_{i=1}^{k-1}$, with success rate $\pi(s)$ when $S_k = s$ for $k = 1, \ldots, n$. That is,*

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

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*

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

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.

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 $\frac{1}{2} < \lambda \leq 1$.

Example 4 (SBR). *For each stratum, $\lfloor \pi(s) n(s) \rfloor$ units are assigned to treatment and the rest are assigned to control.*

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

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

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

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 $(\hat{\gamma}_{as}^Y, \hat{\gamma}_{as}^D, \hat{\nu}_{as}^Y, \hat{\nu}_{as}^D)$ for $a = 0, 1$ are the estimated coefficients of the four sets of stratum-specific regressions using only the s

stratum:

$$(1-A_i)Y_i = (1-A_i)(\gamma_{0s}^Y + X_i^\top \nu_{0s}^Y + e_{0i}^Y), \quad A_iY_i = A_i(\gamma_{1s}^Y + X_i^\top \nu_{1s}^Y + e_{1i}^Y)$$
$$(1-A_i)D_i = (1-A_i)(\gamma_{0s}^D + X_i^\top \nu_{0s}^D + e_{0i}^D), \quad A_iD_i = A_i(\gamma_{1s}^D + X_i^\top \nu_{1s}^D + e_{1i}^D).$$

Interpret $(\hat{\gamma}_{aS_i}^Y, \hat{\gamma}_{aS_i}^D, \hat{\nu}_{aS_i}^Y, \hat{\nu}_{aS_i}^D)$ for $a = 0, 1$ as the estimated coefficients of the four sets of stratum-specific regressions using only the S_i stratum.

Under Assumption 1 and Assumption 2 of our paper, Ansel et al. (2018) show that $\hat{\tau}_S$ is a consistent estimator of τ, asymptotically normal, and the most efficient among the estimators studied in their paper ($\pi(s)$ can be heterogenous across strata). To define the explicit expression for the asymptotic variance of $\hat{\tau}_S$, denoted as σ_S^2, we need to introduce addition notation. For $s \in \mathcal{S}$, let $\tilde{X}_{is} := X_i - \mathbb{E}(X_i|S_i = s)$,

$$\begin{aligned}
\rho_{iS_i}(1) &:= \frac{Y_i(D_i(1)) - D_i(1)\tau - X_i^\top \nu_{1S_i}^{YD}}{\pi(S_i)} + X_i^\top(\nu_{1S_i}^{YD} - \nu_{0S_i}^{YD}) \\
\rho_{iS_i}(0) &:= \frac{Y_i(D_i(0)) - D_i(0)\tau - X_i^\top \nu_{0S_i}^{YD}}{1-\pi(S_i)} - X_i^\top(\nu_{1S_i}^{YD} - \nu_{0S_i}^{YD}) \\
\nu_{1s}^{YD} &:= \left[\mathbb{E}(\tilde{X}_{is}\tilde{X}_{is}^\top|S_i = s)\right]^{-1} \mathbb{E}\left(\tilde{X}_{is}\left[Y_i\left(D_i(1)\right) - D_i(1)\tau\right]|S_i = s\right), \\
\nu_{0s}^{YD} &:= \left[\mathbb{E}(\tilde{X}_{is}\tilde{X}_{is}^\top|S_i = s)\right]^{-1} \mathbb{E}\left(\tilde{X}_{is}\left[Y_i\left(D_i(0)\right) - D_i(0)\tau\right]|S_i = s\right).
\end{aligned}$$

$$\begin{aligned}
\sigma_{S1}^2 &:= \mathbb{E}\left[\pi(S_i)\left\{\rho_{iS_i}(1) - \mathbb{E}[\rho_{iS_i}(1)|S_i]\right\}^2\right] \\
\sigma_{S0}^2 &:= \mathbb{E}\left[\left(1-\pi(S_i)\right)\left\{\rho_{iS_i}(0) - \mathbb{E}[\rho_{iS_i}(0)|S_i]\right\}^2\right] \\
\sigma_{S2}^2 &:= \mathbb{E}\left[\left(\mathbb{E}\left[Y_i(D_i(1)) - Y_i(D_i(0)) - \tau(D_i(1) - D_i(0))|S_i\right]\right)^2\right].
\end{aligned}$$

In addition, define

$$\begin{aligned}
\hat{\rho}_{iS_i}(1) &:= \frac{Y_i - D_i\hat{\tau}_S - X_i^\top \hat{\nu}_{1S_i}^{YD}}{\hat{\pi}(S_i)} + X_i^\top(\hat{\nu}_{1S_i}^{YD} - \hat{\nu}_{0S_i}^{YD}) \\
\hat{\rho}_{iS_i}(0) &:= \frac{Y_i - D_i\hat{\tau}_S - X_i^\top \hat{\nu}_{0S_i}^{YD}}{1-\hat{\pi}(S_i)} - X_i^\top(\hat{\nu}_{1S_i}^{YD} - \hat{\nu}_{0S_i}^{YD})
\end{aligned}$$

$$\hat{\sigma}^2_{S1} := \frac{1}{n}\sum_{i\in[n]} A_i \left[\hat{\rho}_{iS_i}(1) - \frac{1}{n_1(S_i)}\sum_{j\in I_1(S_i)} \hat{\rho}_{jS_j}(1)\right]^2$$

$$\hat{\sigma}^2_{S0} := \frac{1}{n}\sum_{i\in[n]} (1-A_i) \left[\hat{\rho}_{iS_i}(0) - \frac{1}{n_0(S_i)}\sum_{j\in I_0(S_i)} \hat{\rho}_{jS_j}(0)\right]^2$$

$$\hat{\sigma}^2_{S2} := \frac{1}{n}\sum_{i\in[n]} \left(\frac{1}{n_1(S_i)}\sum_{j\in I_1(S_i)} (Y_j - \hat{\tau}_S D_j) - \frac{1}{n_0(S_i)}\sum_{j\in I_0(S_i)} (Y_j - \hat{\tau}_S D_j)\right)^2$$

$$\hat{\sigma}^2_{S} := \frac{\hat{\sigma}^2_{S1} + \hat{\sigma}^2_{S0} + \hat{\sigma}^2_{S2}}{\left(\sum_{s\in\mathcal{S}} \hat{p}(s)(\hat{\gamma}^D_{1s} - \hat{\gamma}^D_{0s} + (\hat{\nu}^D_{1s} - \hat{\nu}^D_{0s})^\top \bar{X}_s)\right)^2}$$

where $\hat{\nu}^{YD}_{aS_i} := \hat{\nu}^{Y}_{aS_i} - \hat{\tau}_S \hat{\nu}^{D}_{aS_i}$ for $a = 0, 1$.

Theorem S.B.1. *Suppose Assumptions 1 and 2 hold. Then,*

(i)

$$\sigma^2_S = \frac{\sigma^2_{S1} + \sigma^2_{S0} + \sigma^2_{S2}}{(\mathbb{E}[D_i(1) - D_i(0)])^2}. \tag{S.B.1}$$

(ii)

$$\hat{\sigma}^2_S \xrightarrow{p} \sigma^2_S.$$

It can be shown that $\sigma^2_{Sa} \geq \underline{\sigma}^2_a$ for $a = 0, 1$ and $\sigma^2_{S2} = \underline{\sigma}^2_2$, 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, and $\underline{\sigma}^2_a$ for $a = 0, 1, 2$ are defined in Theorem 4.1. This implies in general, the S estimator is not semiparametrically most efficient.

Theorem S.B.2. *Suppose that Assumptions 1 and 2 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 $\sigma^2_S \leq \sigma^2_{TSLS}$.*

Theorem S.B.2 could be deduced from Theorem 5.3. Both $\hat{\tau}_{TSLS}$ and $\hat{\tau}_S$ use linear adjustments of X_i, but Theorem S.B.2 states that $\hat{\tau}_S$ is more efficient than $\hat{\tau}_{TSLS}$. In the discussion following Theorem 5.3, 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.

S.C 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 adjustments are exactly the same as nonlinear (logistic) adjustments studied in Section 5.1.2. Theoretically, we will let the regressors $\mathring{\Psi}_{i,s}$ in (5.8) be sieve basis functions whose dimensions will diverge to infinity as the sample size increases. For notational 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

$$\hat{\mu}^Y(a,s,X_i) = \mathring{\Psi}_{i,n}^\top \hat{\theta}_{a,s}^{NP} \quad \text{and} \quad \hat{\mu}^D(a,s,X_i) = \lambda(\mathring{\Psi}_{i,n}^\top \hat{\beta}_{a,s}^{NP}), \tag{S.C.1}$$

where

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

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

Assumption S.C.1. *(i) There exist constants $0 < c < C < \infty$ such that with probability approaching one,*

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

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

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

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

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

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

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

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

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

(iv) Suppose that $\mathbb{E}[\mathring{\Psi}^2_{i,n,k}|S_i = s] \leq C$ *for some constant* $C > 0$, *where* $\mathring{\Psi}_{i,n,k}$ *denotes the* k*th element of* $\mathring{\Psi}_{i,n}$. $\max_{i\in[n]} ||\mathring{\Psi}_{i,n}||_2 \leq \zeta(h_n)$ *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 S.C.1 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 S.C.1(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 S.C.1. *Suppose Assumptions 1 and S.C.1 hold. Then* $\{\hat{\mu}^b(a,s,X_i)\}_{b=D,Y,a=0,1,s\in\mathcal{S}}$ *defined in* (S.C.1) *with* $\overline{\mu}^b(a,s,X) = \mu^b(a,s,X)$ *satisfy Assumption 3. All the results in Theorem 3.1(i)-(ii) hold for* $\hat{\tau}_{NP}$. *In addition,* $\hat{\tau}_{NP}$ *achieves the SEB.*

The nonlinear (logistic) and nonparametric adjustments are numerically identical if the same set of regressors are used. Theorem S.C.1 then shows that the nonlinear (logistic) 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 nonlinear (logistic) adjustment can be viewed as the nonparametric adjustment, which achieves the SEB. In fact, if we estimate both $\mu^Y(a,s,X)$ and $\mu^D(a,s,X)$ by linear sieve regressions, under similar conditions to Assumption S.C.1, we can show that such an adjusted estimator also achieves the SEB. So does Ansel et al.'s (2018) S estimator when their X_i is replaced by sieve bases of X_i because it is asymptotically equivalent to our estimator L with optimal linear adjustment.

S.D 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 (S.C.1).

Lasso regressions. We follow the estimation procedure and the choice of tuning parameter proposed by Belloni, Chernozhukov, Fernández-Val, and Hansen (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 $\hat{\sigma}_h^{Y,(0)} = \frac{1}{n_a(s)} \sum_{i \in I_a(s)} (Y_i - \bar{Y}_{a,s})^2 \mathring{\Psi}_{i,n,h}^2$ and $\hat{\sigma}_h^{D,(0)} = \frac{1}{n_a(s)} \sum_{i \in I_a(s)} (D_i - \bar{D}_{a,s})^2 \mathring{\Psi}_{i,n,h}^2$ 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

$$\begin{aligned}
\hat{\theta}_{a,s}^{R,0} &= \arg\min_t \frac{-1}{n_a(s)} \sum_{i \in I_a(s)} \left(Y_i - \mathring{\Psi}_{i,n}^\top t \right)^2 + \frac{\varrho_{n,a}(s)}{n_a(s)} \sum_{h \in [p_n]} \hat{\sigma}_h^{Y,(0)} |t_h|, \\
\hat{\beta}_{a,s}^{R,0} &= \arg\min_b \frac{-1}{n_a(s)} \sum_{i \in I_a(s)} \left[D_i \log(\lambda(\mathring{\Psi}_{i,n}^\top b)) + (1 - D_i) \log(1 - \lambda(\mathring{\Psi}_{i,n}^\top b)) \right] \\
&\quad + \frac{\varrho_{n,a}(s)}{n_a(s)} \sum_{h \in [p_n]} \hat{\sigma}_h^{D,(0)} |b_h|.
\end{aligned}$$

(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 $\hat{\varepsilon}_i^{Y,(k)} = Y_i - \mathring{\Psi}_{i,n}^\top \hat{\theta}_{a,s}^{R,k-1}$ 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 $\hat{\varepsilon}_i^{D,(k)} = D_i - \lambda(\mathring{\Psi}_{i,n}^\top \hat{\beta}_{a,s}^{R,k-1})$. Estimate

$$\begin{aligned}
\hat{\theta}_{a,s}^{R,k} &= \arg\min_t \frac{-1}{n_a(s)} \sum_{i \in I_a(s)} \left(Y_i - \mathring{\Psi}_{i,n}^\top t \right)^2 + \frac{\varrho_{n,a}(s)}{n_a(s)} \sum_{h \in [p_n]} \hat{\sigma}_h^{Y,(k-1)} |t_h|, \\
\hat{\beta}_{a,s}^{R,k} &= \arg\min_b \frac{-1}{n_a(s)} \sum_{i \in I_a(s)} \left[D_i \log(\lambda(\mathring{\Psi}_{i,n}^\top b)) + (1 - D_i) \log(1 - \lambda(\mathring{\Psi}_{i,n}^\top b)) \right] \\
&\quad + \frac{\varrho_{n,a}(s)}{n_a(s)} \sum_{h \in [p_n]} \hat{\sigma}_h^{D,(k-1)} |b_h|.
\end{aligned}$$

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

S.E Regression Adjustment under Full Compliance

In this section, we briefly discuss the regression adjustment under full compliance. We aim to construct consistent and efficient estimators for the average treatment effect (ATE). Under full compliance, we have $D(a) = a$ for $a = 0, 1$ so that $D = A$. The estimator $\hat{\mu}^D(a, s, x) = a$ is correctly specified. Then, our proposed estimator of ATE is

$$\hat{\tau}_{ATE} := \frac{1}{n} \sum_{i \in [n]} \left[\frac{A_i (Y_i - \hat{\mu}^Y(1, S_i, X_i))}{\hat{\pi}(S_i)} - \frac{(1 - A_i)(Y_i - \hat{\mu}^Y(0, S_i, X_i))}{1 - \hat{\pi}(S_i)} + \hat{\mu}^Y(1, S_i, X_i) - \hat{\mu}^Y(0, S_i, X_i) \right], \tag{S.E.1}$$

where $\hat{\mu}^Y(a,s,x)$ is an estimator of the working model $\overline{\mu}^Y(a,s,x)$.

The optimal linear adjustment is $\hat{\mu}^Y(a,s,X_i) = \Psi_{i,s}^\top \hat{\theta}_{a,s}^L$, where

$$\dot{\Psi}_{i,a,s} := \Psi_{i,s} - \frac{1}{n_a(s)} \sum_{i \in I_a(s)} \Psi_{i,s}$$

$$\hat{\theta}_{a,s}^L := \left(\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \dot{\Psi}_{i,a,s} \dot{\Psi}_{i,a,s}^\top \right)^{-1} \left(\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \dot{\Psi}_{i,a,s} Y_i \right).$$

We can show that such an adjustment achieves the minimal variance of the ATE estimator that is adjusted by linear functions of $\Psi_{i,s}$.

Let $\mathring{\Psi}_{i,n}$ contains sieve bases of X_i. Then, the nonparametric adjustment can be written as $\hat{\mu}^Y(a,s,X_i) = \mathring{\Psi}_{i,n}^\top \hat{\theta}_{a,s}^{NP}$, where

$$\hat{\theta}_{a,s}^{NP} = \left(\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \mathring{\Psi}_{i,n} \mathring{\Psi}_{i,n}^\top \right)^{-1} \left(\frac{1}{n_a(s)} \sum_{i \in I_a(s)} \mathring{\Psi}_{i,n} Y_i \right).$$

Last, suppose $\mathring{\Psi}_{i,n}$ contains high-dimensional regressors of X_i. Then, the regularized adjustment can be written as $\hat{\mu}^Y(a,s,X_i) = \mathring{\Psi}_{i,n}^\top \hat{\theta}_{a,s}^R$, where

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

$\{\varrho_{n,a}(s)\}_{a=0,1,s\in\mathcal{S}}$ are tuning parameters, $\hat{\Omega}^Y = \text{diag}(\hat{\omega}_1^Y, \cdots, \hat{\omega}_{p_n}^Y)$ is a diagonal matrix of data-dependent penalty loadings as defined in Section S.D. Under similar conditions as in Assumptions S.C.1 and 8, we can show that the ATE estimator with both the nonparametric and regularized adjustments achieves the semiparametric efficiency bound.

S.F Proof of Theorem 2.1

We define $\hat{\sigma}^2_{TSLS,naive}$ as

$$\hat{\sigma}^2_{TSLS,naive} = e_1^\top [S_{\bar{Z},\bar{X}}^\top S_{\bar{Z},\bar{Z}}^{-1} S_{\bar{Z},\bar{X}}]^{-1} \left[S_{\bar{Z},\bar{X}}^\top S_{\bar{Z},\bar{Z}}^{-1} \left(\frac{1}{n} \sum_{i=1}^n (\bar{Z}_i \bar{Z}_i^\top \hat{\varepsilon}_i^2) \right) S_{\bar{Z},\bar{Z}}^{-1} S_{\bar{Z},\bar{X}} \right] [S_{\bar{Z},\bar{X}}^\top S_{\bar{Z},\bar{Z}}^{-1} S_{\bar{Z},\bar{X}}]^{-1} e_1,$$

where $\bar{X}_i = (D_i, \{1\{S_i = s\}\}_{s\in\mathcal{S}}, X_i^\top)^\top$ $\bar{Z}_i = (A_i, \{1\{S_i = s\}\}_{s\in\mathcal{S}}, X_i^\top)^\top$, $S_{\bar{Z},\bar{Z}} = \frac{1}{n}\sum_{i\in[n]} \bar{Z}_i\bar{Z}_i^\top$, $S_{\bar{Z},\bar{X}} = \frac{1}{n}\sum_{i\in[n]} \bar{Z}_i\bar{X}_i^\top$, e_1 is a vector with its first element being one and the rest being zero, $\hat{\varepsilon}_i = Y_i - \hat{\tau}_{TSLS}D_i - \sum_{s\in\mathcal{S}}\hat{\alpha}_{s,TSLS}1\{S_i = s\} - X_i^\top\hat{\delta}_{TSLS}$, and $(\hat{\tau}_{TSLS}, \hat{\alpha}_{s,TSLS}, \hat{\delta}_{TSLS})$ are the usual TSLS estimators.

Next, we define σ^2_{TSLS} and $\sigma^2_{TSLS,naive}$. Let $\mathbb{X}_i = (X_i^\top, \{1\{S_i = s\}\}_{s\in\mathcal{S}})^\top$,

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

Furthermore, define

$$\begin{aligned}
\sigma^2_{TSLS,naive} &= \frac{\sigma^2_{TSLS,0} + \sigma^2_{TSLS,1} + \sigma^2_{TSLS,2} + \tilde{\sigma}^2_{TSLS,3}}{(\mathbb{E}(D_i(1) - D_i(0)))^2}, \quad \text{where} \\
\tilde{\sigma}^2_{TSLS,3} &= \mathbb{E}\left\{\pi(1-\pi)\left(\mathbb{E}\left[\frac{Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*}{\pi} + \frac{Y_i(D_i(0)) - D_i(0)\tau - \mathbb{X}_i^\top\lambda^*}{1-\pi}\middle|S_i\right]\right)^2\right\}.
\end{aligned}$$

By definition, $\sigma^2_{TSLS} \le \sigma^2_{TSLS,naive}$. The inequality is strict if $\gamma(s) < \pi(1-\pi)$.

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

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

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

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

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

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

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

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i\in[n]}A_i(1-\pi(S_i))(Y_i(D_i(1))-\mathbb{E}(Y_i(D_i(1))|S_i))\middle|\{A_i,S_i\}_{i\in[n]}\right]=0\\
&\mathbb{E}\left[\left(\frac{1}{n}\sum_{i\in[n]}A_i(1-\pi(S_i))(Y_i(D_i(1))-\mathbb{E}(Y_i(D_i(1))|S_i))\right)^2\middle|\{A_i,S_i\}_{i\in[n]}\right]\\
&\leq\sum_{s\in\mathcal{S}}\frac{n_1(s)(1-\pi(s))^2\mathbb{E}(Y^2(D(1))|S_i=s)}{n^2}=o_p(1),
\end{aligned}$$

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

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

and

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

Therefore, it is only left to show (S.F.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 (S.F.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

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

where the last equality holds following the similar argument in (S.F.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

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

By the same argument in the proof of the first statement of Theorem 2.1, we have

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

Next, we turn to the numerator. We have

$$\frac{1}{\sqrt{n}}\sum_{i\in[n]}\widetilde{A}_i(Y_i - D_i\tau) = \frac{1}{\sqrt{n}}\sum_{i\in[n]}(A_i - \mathbb{X}_i^\top\theta - \mathbb{X}_i^\top(\hat{\theta} - \theta))(Y_i - D_i\tau)$$

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

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 $\varpi_i = \lambda_1(Y_i-D_i\tau)+\lambda_2^\top\mathbb{X}_i$ and $\varpi_i(a) = \lambda_1(Y_i(D_i(a))-D_i(a)\tau)+\lambda_2^\top\mathbb{X}_i$. Note that $\varpi_i = A_i\varpi_i(1)+(1-A_i)\varpi_i(0)$. We have

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

where

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

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

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

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

Thus (S.F.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

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

where

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

To see the second result, we note that $\bar{X}_i=(D_i,\mathbb{X}_i^\top)^\top$ and $\bar{Z}_i=(A_i,\mathbb{X}_i^\top)^\top$. Denote $\breve{Z}_i=(\tilde{A}_i,\mathbb{X}_i^\top)^\top$. Then, we have

$$\begin{aligned}
&e_1^\top[S_{\bar{X},\bar{Z}}S^{-1}_{\bar{Z},\bar{Z}}S_{\bar{Z},\bar{X}}]^{-1}\\
&=[S_{\bar{X},\breve{Z}}S^{-1}_{\breve{Z},\breve{Z}}S_{\breve{Z},\bar{X}}]^{-1}
\end{aligned}$$

$$= e_1^\top \left\{ \begin{pmatrix} \sum_{i\in[n]} D_i\tilde{A}_i/n & \sum_{i\in[n]} D_i\mathbb{X}_i^\top/n \\ 0 & \sum_{i\in[n]} \mathbb{X}_i\mathbb{X}_i^\top/n \end{pmatrix} \begin{pmatrix} \sum_{i\in[n]} \tilde{A}_i^2/n & 0 \\ 0 & \sum_{i\in[n]} \mathbb{X}_i\mathbb{X}_i^\top/n \end{pmatrix}^{-1} \right.$$

$$\left. \times \begin{pmatrix} \sum_{i\in[n]} D_i\tilde{A}_i/n & 0 \\ \sum_{i\in[n]} D_i\mathbb{X}_i/n & \sum_{i\in[n]} \mathbb{X}_i\mathbb{X}_i^\top/n \end{pmatrix} \right\}^{-1}$$

$$= e_1^\top \begin{pmatrix} \frac{(\sum_{i\in[n]} D_i\tilde{A}_i/n)^2}{(\sum_{i\in[n]} \tilde{A}_i^2/n)} + \sum_{i\in[n]} D_i\mathbb{X}_i^\top/n \left[\sum_{i\in[n]} \mathbb{X}_i\mathbb{X}_i^\top/n\right]^{-1} \sum_{i\in[n]} D_i\mathbb{X}_i/n & \sum_{i\in[n]} D_i\mathbb{X}_i^\top/n \\ \sum_{i\in[n]} D_i\mathbb{X}_i/n & \sum_{i\in[n]} \mathbb{X}_i\mathbb{X}_i^\top/n \end{pmatrix}^{-1}$$

$$\xrightarrow{p} [\pi(1-\pi)]^{-1}(\mathbb{E}(D(1)-D(0)))^{-2} \begin{pmatrix} 1 & -\gamma_D^\top \end{pmatrix}.$$

and

$$\begin{aligned} S_{\bar{X},\bar{Z}} S_{\bar{Z},\bar{Z}} \bar{Z}_i = S_{\bar{X},\breve{Z}} S_{\breve{Z},\breve{Z}} \breve{Z}_i \\ &= \begin{pmatrix} \sum_{i\in[n]} D_i\tilde{A}_i/n & \sum_{i\in[n]} D_i\mathbb{X}_i^\top/n \\ 0 & \sum_{i\in[n]} \mathbb{X}_i\mathbb{X}_i^\top/n \end{pmatrix} \begin{pmatrix} \sum_{i\in[n]} \tilde{A}_i^2/n & 0 \\ 0 & \sum_{i\in[n]} \mathbb{X}_i\mathbb{X}_i^\top/n \end{pmatrix}^{-1} \begin{pmatrix} \tilde{A}_i \\ \mathbb{X}_i \end{pmatrix} \\ &= \begin{pmatrix} (\sum_{i\in[n]} D_i\tilde{A}_i/n)(\sum_{i\in[n]} \tilde{A}_i^2/n)^{-1} & (\sum_{i\in[n]} D_i\mathbb{X}_i^\top/n)(\sum_{i\in[n]} \mathbb{X}_i\mathbb{X}_i^\top/n)^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} \tilde{A}_i \\ \mathbb{X}_i \end{pmatrix}, \end{aligned}$$

where $\gamma_D = (\mathbb{E}\mathbb{X}_i\mathbb{X}_i^\top)^{-1}\mathbb{E}(\mathbb{X}_i(\pi D_i(1) + (1-\pi)D_i(0)))$. Further note that

$$\begin{pmatrix} (\sum_{i\in[n]} D_i\tilde{A}_i/n)(\sum_{i\in[n]} \tilde{A}_i^2/n)^{-1} & (\sum_{i\in[n]} D_i\mathbb{X}_i^\top/n)(\sum_{i\in[n]} \mathbb{X}_i\mathbb{X}_i^\top/n)^{-1} \\ 0 & I \end{pmatrix} \xrightarrow{p} \begin{pmatrix} \mathbb{E}(D(1)-D(0)) & \gamma_D^\top \\ 0 & I \end{pmatrix}$$

and

$$\begin{aligned} \hat{\lambda}_{TSLS} \equiv \begin{pmatrix} \hat{\alpha}_{1,TSLS} \\ \vdots \\ \hat{\alpha}_{S,TSLS} \\ \hat{\theta}_{TSLS} \end{pmatrix} &= (\sum_{i\in[n]} \mathbb{X}_i\mathbb{X}_i^\top/n)^{-1}(\sum_{i\in[n]} \mathbb{X}_i(Y_i - D_i\hat{\tau}_{TSLS})/n) \\ &= \hat{\lambda} + (\sum_{i\in[n]} \mathbb{X}_i\mathbb{X}_i^\top/n)^{-1}(\sum_{i\in[n]} \mathbb{X}_i D_i/n)(\tau - \hat{\tau}_{TSLS}) \xrightarrow{p} \lambda^*. \end{aligned}$$

Then, we have

$$\hat{e}_i = e_i - D_i(\hat{\tau}_{TSLS} - \tau) - \mathbb{X}_i^\top(\hat{\lambda}_{TSLS} - \lambda^*),$$

where $e_i = Y_i - D_i\tau - \mathbb{X}_i^\top\lambda^*$. In addition, as shown above, we have $\tilde{A}_i = A_i - \pi - \mathbb{X}_i^\top(\hat{\theta} - \theta)$ and $\hat{\theta} \xrightarrow{p} \theta$. This implies,

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in[n]} \hat{e}_i^2 \begin{pmatrix} \tilde{A}_i^2 & \tilde{A}_i\mathbb{X}_i^\top \\ \tilde{A}_i\mathbb{X}_i & \mathbb{X}_i\mathbb{X}_i^\top \end{pmatrix} \\
&= \frac{1}{n}\sum_{i\in[n]} e_i^2 \begin{pmatrix} (A_i-\pi)^2 & (A_i-\pi)\mathbb{X}_i^\top \\ (A_i-\pi)\mathbb{X}_i & \mathbb{X}_i\mathbb{X}_i^\top \end{pmatrix} + o_P(1) \\
&= \frac{1}{n}\sum_{i\in[n]} \Big[A_i(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2 + (1-A_i)(Y_i(D_i(0)) - D_i(0)\tau - \mathbb{X}_i^\top\lambda^*)^2\Big] \\
&\times \begin{pmatrix} (A_i-\pi)^2 & (A_i-\pi)\mathbb{X}_i^\top \\ (A_i-\pi)\mathbb{X}_i & \mathbb{X}_i\mathbb{X}_i^\top \end{pmatrix} + o_P(1) \\
&= \frac{1}{n}\sum_{i\in[n]} A_i \begin{pmatrix} (1-\pi)^2(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2 & (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i^\top \\ (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i & (Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i\mathbb{X}_i^\top \end{pmatrix} \\
&+ \frac{1}{n}\sum_{i\in[n]} (1-A_i) \begin{pmatrix} \pi^2(Y_i(D_i(0)) - D_i(0)\tau - \mathbb{X}_i^\top\lambda^*)^2 & -\pi(Y_i(D_i(0)) - D_i(0)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i^\top \\ -\pi(Y_i(D_i(0)) - D_i(0)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i & (Y_i(D_i(0)) - D_i(0)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i\mathbb{X}_i^\top \end{pmatrix} + o_P(1).
\end{aligned}$$

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

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in[n]} A_i \begin{pmatrix} (1-\pi)^2(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2 & (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i^\top \\ (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i & (Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i\mathbb{X}_i^\top \end{pmatrix} \\
&= \frac{1}{n}\sum_{i\in[n]} A_i \Bigg\{ \begin{pmatrix} (1-\pi)^2(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2 & (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i^\top \\ (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i & (Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i\mathbb{X}_i^\top \end{pmatrix} \\
&- \mathbb{E}\left[\begin{pmatrix} (1-\pi)^2(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2 & (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i^\top \\ (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i & (Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i\mathbb{X}_i^\top \end{pmatrix} \Bigg| S_i\right] \Bigg\} \\
&+ \frac{1}{n}\sum_{i\in[n]} (A_i - \pi)\mathbb{E}\left[\begin{pmatrix} (1-\pi)^2(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2 & (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i^\top \\ (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i & (Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i\mathbb{X}_i^\top \end{pmatrix}\right. \\
&+ \frac{1}{n}\sum_{i\in[n]} \pi\mathbb{E}\left[\begin{pmatrix} (1-\pi)^2(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2 & (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i^\top \\ (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i & (Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i\mathbb{X}_i^\top \end{pmatrix} \Bigg| S_i\right]
\end{aligned}$$

$$
\begin{aligned}
&\xrightarrow{p} \pi\mathbb{E}\left[\begin{pmatrix} (1-\pi)^2(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2 & (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i^\top \\ (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i & (Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i\mathbb{X}_i^\top \end{pmatrix}\right] \\
&\equiv \Omega_1.
\end{aligned}
$$

To see the convergence in probability in the above display, we note that

$$
\begin{aligned}
&A_i\Bigg\{\begin{pmatrix} (1-\pi)^2(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2 & (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i^\top \\ (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i & (Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i\mathbb{X}_i^\top \end{pmatrix} \\
&- \mathbb{E}\left[\begin{pmatrix} (1-\pi)^2(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2 & (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i^\top \\ (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i & (Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i\mathbb{X}_i^\top \end{pmatrix}\middle| S_i\right]\Bigg\}
\end{aligned}
$$

is independent and conditionally mean zero given $(A^{(n)}, S^{(n)})$. Therefore, by the conditional Chebyshev's inequality, we have

$$
\begin{aligned}
&\frac{1}{n}\sum_{i\in[n]} A_i\Bigg\{\begin{pmatrix} (1-\pi)^2(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2 & (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i^\top \\ (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i & (Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i\mathbb{X}_i^\top \end{pmatrix} \\
&- \mathbb{E}\left[\begin{pmatrix} (1-\pi)^2(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2 & (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i^\top \\ (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i & (Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i\mathbb{X}_i^\top \end{pmatrix}\middle| S_i\right]\Bigg\} \\
&= o_P(1).
\end{aligned}
$$

Also, by Assumption 2, we have

$$
\begin{aligned}
&\frac{1}{n}\sum_{i\in[n]}(A_i - \pi)\mathbb{E}\Bigg[\begin{pmatrix} (1-\pi)^2(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2 & (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i^\top \\ (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i & (Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i\mathbb{X}_i^\top \end{pmatrix}\Bigg| S_i \\
&= o_P(1).
\end{aligned}
$$

Last, by the usual Law of Large numbers for i.i.d. data, we have

$$
\begin{aligned}
&\frac{1}{n}\sum_{i\in[n]} \pi\mathbb{E}\left[\begin{pmatrix} (1-\pi)^2(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2 & (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i^\top \\ (1-\pi)(Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i & (Y_i(D_i(1)) - D_i(1)\tau - \mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i\mathbb{X}_i^\top \end{pmatrix}\middle| S_i\right] \\
&\xrightarrow{p} \Omega_1.
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
&\frac{1}{n}\sum_{i\in[n]}(1-A_i)\begin{pmatrix}\pi^2(Y_i(D_i(0))-D_i(0)\tau-\mathbb{X}_i^\top\lambda^*)^2 & -\pi(Y_i(D_i(0))-D_i(0)\tau-\mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i^\top \\ -\pi(Y_i(D_i(0))-D_i(0)\tau-\mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i & (Y_i(D_i(0))-D_i(0)\tau-\mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i\mathbb{X}_i^\top\end{pmatrix} \\
&\xrightarrow{p}(1-\pi)\mathbb{E}\left[\begin{pmatrix}\pi^2(Y_i(D_i(0))-D_i(0)\tau-\mathbb{X}_i^\top\lambda^*)^2 & -\pi(Y_i(D_i(0))-D_i(0)\tau-\mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i^\top \\ -\pi(Y_i(D_i(0))-D_i(0)\tau-\mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i & (Y_i(D_i(0))-D_i(0)\tau-\mathbb{X}_i^\top\lambda^*)^2\mathbb{X}_i\mathbb{X}_i^\top\end{pmatrix}\right] \\
&\equiv\Omega_0.
\end{aligned}
$$

Consequently, we have

$$
\begin{aligned}
\hat{\sigma}^2_{TSLS,naive}&\xrightarrow{p}\frac{\begin{pmatrix}1 & -\gamma_D^\top\end{pmatrix}\begin{pmatrix}\mathbb{E}(D(1)-D(0)) & \gamma_D^\top \\ 0 & I\end{pmatrix}(\Omega_1+\Omega_0)\begin{pmatrix}\mathbb{E}(D(1)-D(0)) & \gamma_D^\top \\ 0 & I\end{pmatrix}^\top\begin{pmatrix}1 \\ -\gamma_D\end{pmatrix}}{[\pi(1-\pi)]^2[\mathbb{E}(D(1)-D(0))]^4} \\
&=\frac{\pi^{-1}\mathbb{E}(Y_i(D_i(1))-D_i(1)\tau-\mathbb{X}_i^\top\lambda^*)^2+(1-\pi)^{-1}\mathbb{E}(Y_i(D_i(1))-D_i(1)\tau-\mathbb{X}_i^\top\lambda^*)^2}{[\mathbb{E}(D(1)-D(0))]^2} \\
&=\frac{\sigma^2_{TSLS,0}+\sigma^2_{TSLS,1}+\sigma^2_{TSLS,2}+\tilde{\sigma}^2_{TSLS,3}}{(\mathbb{E}(D_i(1)-D_i(0)))^2}.
\end{aligned}
$$

For the last result, by the proof Theorem 3.1 with $\overline{\mu}^b(a,s,x)=0$ for $a=0,1$ and $b=D,Y$ and $\pi(s)=\pi$, we have

$$
\begin{aligned}
\sigma^2_{NA}=&\frac{\sum_{s\in S}\frac{p(s)}{\pi}Var(Y(D(1))-\tau D(1)|S=s)+\sum_{s\in S}\frac{p(s)}{1-\pi}Var(Y(D(0))-\tau D(0)|S=s)}{\mathbb{P}(D(1)>D(0))^2} \\
&+\frac{Var(\mathbb{E}[W_i-Z_i|S_i]-\tau\left(\mathbb{E}[D_i(1)-D_i(0)|S_i]\right))}{\mathbb{P}(D(1)>D(0))^2} \\
=&\frac{\mathbb{E}\frac{1}{\pi}Var(Y(D(1))-\tau D(1)|S)+\frac{1}{1-\pi}Var(Y(D(0))-\tau D(0)|S)}{\mathbb{P}(D(1)>D(0))^2}+\frac{\sigma^2_{TSLS,2}}{\mathbb{P}(D(1)>D(0))^2}.
\end{aligned}
$$

Then, we have $\sigma^2_{NA}<\sigma^2_{TSLS}$ if and only if

$$
\mathbb{E}\left[\frac{1}{\pi}Var(Y(D(1))-\tau D(1)|S)+\frac{1}{1-\pi}Var(Y(D(0))-\tau D(0)|S)\right]<\sigma^2_{TSLS,0}+\sigma^2_{TSLS,1}+\sigma^2_{TSLS,3},
$$

which is equivalent to

$$
2\left[\frac{\mathbb{E}cov(Y_i(D_i(1))-D_i(1)\tau,\mathbb{X}_i^\top\lambda^*|S)}{\pi}+\frac{\mathbb{E}cov(Y_i(D_i(0))-D_i(0)\tau,\mathbb{X}_i^\top\lambda^*|S)}{1-\pi}\right]\leq\frac{\mathbb{E}Var(\mathbb{X}_i^\top\lambda^*|S)}{\pi(1-\pi)}+\sigma^2_{TSLS,3}.
$$

S.G Proof of Theorem 3.1

Let

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

Then, we have

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

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

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

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

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 S.Q.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

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

Similarly, we can show that

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

$$+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\left(\frac{1}{1-\pi(S_i)}-1\right)\tilde{\mu}^D(0,S_i,X_i)+\tilde{\mu}^D(1,S_i,X_i)-\frac{\tilde{D}_i(0)}{1-\pi(S_i)}\right](1-A_i)\Bigg\}$$
$$+\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(\mathbb{E}[D_i(1)-D_i(0)|S_i]-\mathbb{E}[D_i(1)-D_i(0)]\big)\right\}+o_p(1), \tag{S.G.3}$$

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 (S.G.1), (S.G.2), and (S.G.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 S.Q.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 S.Q.3 shows.

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

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

S.H 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 S.H.1. *(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

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

where $\Xi_1(\mathcal{D}_i,S_i)$, $\Xi_0(\mathcal{D}_i,S_i)$ and $\Xi_2(S_i)$ are defined as $\Xi_1(\mathcal{D}_i,S_i)$, $\Xi_0(\mathcal{D}_i,S_i)$ and $\Xi_2(S_i)$ in (3.7)–(3.9), 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 $(a,b,s,x) \in \{0,1\}\times\{D,Y\}\times\mathcal{SX}$.

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

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

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

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

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

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

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 $\mathbb{E}h^2(X) < \infty$,

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

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 (S.H.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 (S.H.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

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

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

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

Therefore by Assumption S.H.1 we can find a constant L such that

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

This implies

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

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

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

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

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

Then, we have

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

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

$$\left.\frac{\partial \tau(\theta)}{\partial \theta}\right|_{\theta=\theta^*} = \frac{\mathbb{E}\pi(S)\Xi_1(\mathcal{D},S)g_1(\mathcal{Y}(1)|S)}{H} + \frac{\mathbb{E}(1-\pi(S))\Xi_0(\mathcal{D},S)g_0(\mathcal{Y}(0)|S)}{H} + \frac{\mathbb{E}g_s(S)\Xi_2(S)}{H}$$

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

where the last equality is due to (S.H.1).

On the other hand, we note that

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

In addition, by Assumption S.H.1, 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 S.H.1(iii) and the last equality is due to $\mathbb{E}|B_n(s)|/n=o(1)$.[1] This implies

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

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

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

[1] 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 $\mathbb{E}|B_n(s)|/n=o(1)$.

S.I Proof of Theorem 5.1

The proof is divided into two steps. In the first step, we show Assumption 3(i). In the second step, we establish Assumptions 3(ii) and 3(iii).

Step 1. Recall

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

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 $M_{a,s}(\theta_1,\theta_2) := \mathbb{E}[\Lambda^Y_{a,s}(X_i,\theta_1) - \Lambda^Y_{a,s}(X_i,\theta_2)|S_i = s] = \mathbb{E}[\Lambda^Y_{a,s}(X_i^s,\theta_1) - \Lambda^Y_{a,s}(X_i^s,\theta_2)]$. We have

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

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

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

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 4, $\mathcal{F}$ 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 (S.I.1) holds.

Step 2. We have

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

This verifies Assumption 3(ii). Assumption 3(iii) holds by Assumption 4(ii).

S.J Proof of Theorem 5.2

Let

$$\begin{aligned}
&\nu^Y(a,S_i,X_i) = \mathbb{E}(Y_i(D_i(a))|S_i,X_i) - \mathbb{E}(Y_i(D_i(a))|S_i) \quad \text{and} \\
&\nu^D(a,S_i,X_i) = \mathbb{E}(D_i(a)|S_i,X_i) - \mathbb{E}(D_i(a)|S_i).
\end{aligned} \tag{S.J.1}$$

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

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

Similarly, we have

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

Last, we have

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

Let

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

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

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

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 (3.5) and (S.J.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

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

$$
= \begin{pmatrix} (\theta^*_{a,s}, \beta^*_{a,s})_{a=0,1,s\in\mathcal{S}}: \\ \sqrt{\frac{1-\pi(s)}{\pi(s)}}(\theta^*_{1,s} - \tau\beta^*_{1,s}) + \sqrt{\frac{\pi(s)}{1-\pi(s)}}(\theta^*_{0,s} - \tau\beta^*_{0,s}) \\ = \sqrt{\frac{1-\pi(s)}{\pi(s)}}(\theta^L_{1,s} - \tau\beta^L_{1,s}) + \sqrt{\frac{\pi(s)}{1-\pi(s)}}(\theta^L_{0,s} - \tau\beta^L_{0,s}). \end{pmatrix},
$$

where

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

Similarly, we have

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

This concludes the proof.

S.K Proof of Theorem 5.3

In order to verify Assumption 3, by Theorem 5.1, it suffices to show that $\hat{\theta}^L_{a,s} \xrightarrow{p} \theta^L_{a,s}$ and $\hat{\beta}^L_{a,s} \xrightarrow{p} \beta^L_{a,s}$. We focus on the former with $a = 1$. Let $\{W^s_i, X^s_i\}_{i\in[n]}$ be generated independently from the joint distribution of $(Y_i(D_i(1)), X_i)$ given $S_i = s$ and denote $\Psi^s_{i,s} = \Psi_s(X^s_i)$, $\tilde{\Psi}^s_{i,s} = \Psi_s(X^s_i) - \mathbb{E}\Psi_s(X^s_i)$, $\dot{\Psi}^s_{i,1,s} = \Psi_s(X^s_i) - \frac{1}{n_1(s)}\sum_{i=N(s)+1}^{N(s)+n_1(s)}\Psi_s(X^s_i)$, and $\dot{\Psi}^s_{i,0,s} = \Psi_s(X^s_i) - \frac{1}{n_0(s)}\sum_{i=N(s)+n_1(s)+1}^{N(s)+n(s)}\Psi_s(X^s_i)$. Then, we have

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

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

$$
\begin{aligned}
\left(\frac{1}{n_1(s)}\sum_{i=N(s)+1}^{N(s)+n_1(s)}\dot{\Psi}^s_{i,1,s}\dot{\Psi}^{s,\top}_{i,1,s}\right) &= \left(\frac{1}{n_1(s)}\sum_{i=N(s)+1}^{N(s)+n_1(s)}\tilde{\Psi}^s_{i,s}\tilde{\Psi}^{s,\top}_{i,s}\right) + o_p(1), \\
\left(\frac{1}{n_1(s)}\sum_{i=N(s)+1}^{N(s)+n_1(s)}\dot{\Psi}^s_{i,1,s}W^s_i\right) &= \left(\frac{1}{n_1(s)}\sum_{i=N(s)+1}^{N(s)+n_1(s)}\tilde{\Psi}^s_{i,s}W^s_i\right) + o_p(1).
\end{aligned}
$$

In addition, by the standard LLN,

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

Last, Assumption 5 implies $\mathbb{E}(\tilde{\Psi}_{i,s}\tilde{\Psi}^\top_{i,s}|S_i=s)$ is invertible, this means

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

Similarly, we can show that $\hat{\theta}^L_{0,s}\xrightarrow{p}\theta^L_{0,s}$ and $\hat{\beta}^L_{a,s}\xrightarrow{p}\beta^L_{a,s}$ for $a=0,1$ and $s\in\mathcal{S}$. Therefore, Assumption 3 holds, and thus, all the results in Theorem 3.1 hold for $\hat{\tau}_L$. Then, the optimality result in the second half of Theorem 5.3 is a direct consequence of Theorem 5.2.

Last, we compare the asymptotic variances of TSLS estimator and the estimator with the optimal linear adjustment with $\pi(s)=\pi$ for $s\in\mathcal{S}$ and $\Psi_{i,s}=X_i$. In this special case, we first note that the asymptotic variance of the estimator with the optimal linear adjustment is

$$\frac{\sigma_1^2+\sigma_0^2+\sigma_2^2}{[\mathbb{E}(D(1)-D(0))]^2},$$

where

$$\sigma_0^2=\mathbb{E}(1-\pi)\Xi_0^2(\mathcal{D}_i,S_i)$$
$$\Xi_0(\mathcal{D}_i,S_i):=\left[\left(\frac{1}{1-\pi}-1\right)\tilde{X}_i^\top\theta_{0s}+\tilde{X}_i^\top\theta_{1s}-\frac{\tilde{Z}_i}{1-\pi}\right]-\tau\left[\left(\frac{1}{1-\pi}-1\right)\tilde{X}_i^\top\beta_{0s}+\tilde{X}_i^\top\beta_{1s}-\frac{\tilde{D}_i(0)}{1-\pi}\right]$$
$$\sigma_1^2=\mathbb{E}\pi\Xi_1^2(\mathcal{D}_i,S_i),$$
$$\Xi_1(\mathcal{D}_i,S_i):=\left[\left(1-\frac{1}{\pi}\right)\tilde{X}_i^\top\theta_{1s}-\tilde{X}_i^\top\theta_{0s}+\frac{\tilde{W}_i}{\pi}\right]-\tau\left[\left(1-\frac{1}{\pi}\right)\tilde{X}_i^\top\beta_{1s}-\tilde{X}_i^\top\beta_{0s}+\frac{\tilde{D}_i(1)}{\pi}\right],$$
$$\sigma_2^2=\mathbb{E}\left[\mathbb{E}\left[Y(D(1))-Y(D(0))-(D(1)-D(0))\tau|S_i\right]\right]^2,$$

with

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

$$\beta_{as} = \left[\mathbb{E}(\tilde{X}_{is}\tilde{X}_{is}^\top | S_i = s)\right]^{-1} \left[\mathbb{E}(\tilde{X}_{is} D_i(a) | S_i = s)\right], \quad a = 0, 1. \tag{S.K.1}$$

Observe that $\sigma^2_{TSLS,1} + \sigma^2_{TSLS,0}$ can also be written as

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

where

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

with $\theta_{1s} = \theta_{0s}$, $\beta_{1s} = \beta_{0s}$ and $\theta_{1s} - \tau\beta_{1s} = \lambda_x^*$, where λ_x^* is the first d_x coefficients of λ^* defined in Theorem 2.1 where d_x is the dimension of X_i. By Theorem 5.2, we achieve the optimal linear adjustment when $\theta_{a,s}$ and $\beta_{a,s}$ satisfy (S.K.1), which implies

$$\sigma_1^2 + \sigma_0^2 \leq \sigma^2_{TSLS,1} + \sigma^2_{TSLS,0}.$$

In addition, we have $\sigma_2^2 = \sigma^2_{TSLS,2}$ and $0 \leq \sigma^2_{TSLS,3}$, which implies the desired result.

S.L Proof of Theorem 5.4

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,

$$\begin{aligned}
&\frac{1}{n_1(s)}\sum_{i\in I_1(s)}\left[D_i \log(\lambda(\mathring{\Psi}_{i,s}^\top b)) + (1-D_i)\log(1-\lambda(\mathring{\Psi}_{i,s}^\top b))\right] \\
&\stackrel{d}{=} \frac{1}{n_1(s)}\sum_{i=N(s)+1}^{N(s)+n_1(s)}\left[D_i^s(1)\log(\lambda(\mathring{\Psi}_{i,s}^{s,\top} b)) + (1-D_i^s(1))\log(1-\lambda(\mathring{\Psi}_{i,s}^{s,\top} b))\right]
\end{aligned}$$

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

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

$$\sup_b \left| \frac{1}{n_1(s)} \sum_{i \in I_1(s)} \left[D_i \log(\lambda(\mathring{\Psi}_{i,s}^{\top} b)) + (1-D_i)\log(1-\lambda(\mathring{\Psi}_{i,s}^{\top} b))\right] \right.$$
$$\left. - \mathbb{E}\left[D_i(1)\log(\lambda(\mathring{\Psi}_{i,s}^{\top} b)) + (1-D_i(1))\log(1-\lambda(\mathring{\Psi}_{i,s}^{\top} b))|S_i = s\right] \right| \xrightarrow{p} 0.$$

Then, by the standard proof for the extremum estimation, we have $\hat{\beta}_{a,s}^{MLE} \xrightarrow{p} \beta_{a,s}^{MLE}$. Similarly, we can show that $\hat{\theta}_{a,s}^{OLS} \xrightarrow{p} \theta_{a,s}^{OLS}$. The verifies Assumption 4(i). Assumptions 4(ii) and 4(iii) follow from Assumption 6(ii). Then, the desired results hold due to Theorem 5.1.

S.M Proof of Theorem 5.5

We note that the adjustments proposed in Theorem 5.5 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 5.1, to verify Assumption 3, it suffices to show that $\hat{\theta}_{a,s}^F \xrightarrow{p} \theta_{a,s}^F$ and $\hat{\beta}_{a,s}^F \xrightarrow{p} \beta_{a,s}^F$, as we have already shown the consistency of $\hat{\beta}_{a,s}^{MLE}$ in the proof of Theorem

5.4. 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

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

We first show that

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

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

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

where the first inequality is due to Cauchy-Schwarz inequality. We now show (S.M.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

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

Note that

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

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

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

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

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

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

Thus,

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

Moreover, we can show

$$||\breve{\Phi}_{i,a,s}||_2+||\dot{\Phi}_{i,a,s}||_2\le 2\left(4+||\Psi_{i,s}||_2+\frac{1}{n_a(s)}\sum_{i\in I_a(s)}\|\Psi_{i,s}\|_2\right).\tag{S.M.4}$$

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

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

Following the same argument in the proof of Theorem 5.3, we can show that

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

In addition, by Assumption 7, with probability approaching one, there exists a constant $c>0$ such that

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

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

$$\begin{aligned}\hat{\theta}_{a,s}^F &= \left[\frac{1}{n_a(s)}\sum_{i\in I_a(s)}\breve{\Phi}_{i,a,s}\breve{\Phi}_{i,a,s}^\top\right]^{-1}\left[\frac{1}{n_a(s)}\sum_{i\in I_a(s)}\breve{\Phi}_{i,a,s}Y_i\right]\\ &= \left[\frac{1}{n_a(s)}\sum_{i\in I_a(s)}\dot{\Phi}_{i,a,s}\dot{\Phi}_{i,a,s}^\top\right]^{-1}\left[\frac{1}{n_a(s)}\sum_{i\in I_a(s)}\dot{\Phi}_{i,a,s}Y_i\right]+o_p(1)\xrightarrow{p}\theta_{a,s}^F.\end{aligned}$$

Similarly, we have $\hat{\beta}_{a,s}^F\xrightarrow{p}\beta_{a,s}^F$, which implies all the results in Theorem 3.1 hold for $\hat{\tau}_F$. The optimality result in the second half of the theorem is a direct consequence of Theorem 5.2.

S.N Proof of Theorem S.C.1

We focus on verifying Assumption 3 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 5.4, 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 sieve regression in Hirano et al. (2003), we have

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

Then we have

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

$$
\leq \left| \frac{\sum_{i \in I_1(s)} \left(\lambda(\mathring{\Psi}_{i,n}^\top \hat{\beta}_{a,s}^{NP}) - \lambda(\mathring{\Psi}_{i,n}^\top \beta_{a,s}^{NP})\right)}{n_1(s)} - \frac{\sum_{i \in I_0(s)} \left(\lambda(\mathring{\Psi}_{i,n}^\top \hat{\beta}_{a,s}^{NP}) - \lambda(\mathring{\Psi}_{i,n}^\top \beta_{a,s}^{NP})\right)}{n_0(s)} \right|
$$

$$
+ \left| \frac{1}{n_1(s)} \sum_{i \in I_1(s)} \left(R^D(a,s,X_i) - \mathbb{E}[R^D(a,s,X_i)|S_i = s] \right) \right|
$$

$$
+ \left| \frac{1}{n_0(s)} \sum_{i \in I_0(s)} \left(R^D(a,s,X_i) - \mathbb{E}[R^D(a,s,X_i)|S_i = s] \right) \right| =: I + II + III. \quad \text{(S.N.1)}
$$

To bound term I in (S.N.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

$$
I \leq \left| \frac{\sum_{i \in I_1(s)} \left[\lambda(\mathring{\Psi}_{i,n}^\top \hat{\beta}_{a,s}^{NP}) - \lambda(\mathring{\Psi}_{i,n}^\top \beta_{a,s}^{NP}) - M_{a,s}(\hat{\beta}_{a,s}^{NP}, \beta_{a,s}^{NP})\right]}{n_1(s)} \right|
$$

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

Following the argument in the proof of Theorem 5.1, 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,[2] 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

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

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

$$\leq C\sqrt{\frac{n}{n_1(s)}}\left(\sqrt{\frac{h_n^2\log n}{n_a(s)}}+\frac{h_n\log n}{\sqrt{n_1(s)}}\right)1\{n_1(s)\geq n\varepsilon, n_0(s)\geq n\varepsilon\}$$
$$\leq C\sqrt{\frac{1}{\varepsilon}}\left(\sqrt{\frac{h_n^2\log n}{n\varepsilon}}+\frac{h\log n}{\sqrt{n\varepsilon}}\right)\to 0,$$

as $n\to\infty$.

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

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

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 S.C.1(ii). Similarly we have $III=o_p(n^{-1/2})$. Combining the bounds of I, II, III with (S.N.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 3(i).

To verify Assumption 3(ii), we note that

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

where the first equality is due to Assumption S.C.1(ii), and the second equality is due to Assumption S.C.1(iv).

Last, Assumption 3(iii) is implied by Assumption 1(vi) via Jensen's inequality.

S.O Proof of Theorem 5.6

We focus on verifying Assumption 3 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 5.4, 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

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

Then, we have

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

To bound term I in (S.N.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 $\mathring{\Psi}^s_{i,n}=\mathring{\Psi}(X_i^s)$. Then we have

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

Following the argument in the proof of Theorems 5.1 and S.C.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

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

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

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

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

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

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

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

The bounds for I_2, II and III can be established following the same argument as in the proof of Theorem S.C.1. We omit the detail for brevity. This leads to Assumption 3(i).

To verify Assumption 3(ii), we note that

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

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 3(iii) is implied by Assumption 1(vi) via Jensen's inequality.

S.P Proof of Theorem S.B.1

Some part of the proof of part (i) is due to Ansel et al. (2018) while some part of the proof is original. 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 $\hat{\lambda}_{0s}^{D}$:

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

Consider the first term of (S.P.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 $a > 0$, $1 \leq k, \ell \leq \dim(U_i)$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(\pi(s)-A_i)1\{S_i = s\}\left(U_{ik}U_{i\ell}-\mathbb{E}[U_{ik}U_{i\ell}|S_i]\right)\right| \geq a|A^{(n)}, S^{(n)}\right)$$

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

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 $i \neq j$, where the second equality is due to that $U_{ik}U_{i\ell} - \mathbb{E}[U_{ik}U_{i\ell}|S_i]$ and $U_{jk}U_{j\ell} - \mathbb{E}[U_{jk}U_{j\ell}|S_j]$ are independent conditional on $S^{(n)}$. From (S.P.2), we deduce that the first term of (S.P.1) is $o_p(1)$. Consider the second term of (S.P.1).

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

Consider the third term of (S.P.1).

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

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

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

Thus, we have

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

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

$$\begin{aligned}
\hat{\lambda}_{1s}^{Y} &= \left(\frac{1}{n}\sum_{i=1}^{n}A_i1\{S_i=s\}U_iU_i^\top\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}A_i1\{S_i=s\}U_iY_i(D_i(1)) \\
&\xrightarrow{p} \left(\mathbb{E}[UU^\top|S=s]\right)^{-1}\mathbb{E}[UY(D(1))|S=s] \\
\hat{\lambda}_{0s}^{Y} &= \left(\frac{1}{n}\sum_{i=1}^{n}(1-A_i)1\{S_i=s\}U_iU_i^\top\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}(1-A_i)1\{S_i=s\}U_iY_i(D_i(0)) \\
&\xrightarrow{p} \left(\mathbb{E}[UU^\top|S=s]\right)^{-1}\mathbb{E}[UY(D(0))|S=s]
\end{aligned}$$

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

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

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

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

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 (S.P.3).

$$\begin{aligned}
&\sqrt{n}\sum_{s\in\mathcal{S}}\hat{p}(s)\bar{U}_s^{\top}(\hat{\eta}_{1s} - \eta_{1s}) \\
&= \sum_{s\in\mathcal{S}}\hat{p}(s)\bar{U}_s^{\top}\left(\frac{1}{n}\sum_{i=1}^{n}A_i 1\{S_i=s\}U_iU_i^{\top}\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i1\{S_i=s\}U_i\left[Y_i(D_i(1)) - \tau D_i(1) - U_i^{\top}\eta_{1s}\right] \\
&= \sum_{s\in\mathcal{S}}\hat{p}(s)\mathbb{E}[U^{\top}|S=s]\left(\pi(s)\hat{p}(s)\mathbb{E}[UU^{\top}|S=s]\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i1\{S_i=s\}U_i\left[Y_i(D_i(1)) - \tau D_i(1) - U_i^{\top}\eta_{1s}\right] + o_p(1) \\
&= \sum_{s\in\mathcal{S}}\frac{1}{\pi(s)}\mathbb{E}[U^{\top}|S=s]\left(\mathbb{E}[UU^{\top}|S=s]\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i1\{S_i=s\}U_i\left[Y_i(D_i(1)) - \tau D_i(1) - U_i^{\top}\eta_{1s}\right] + o_p(1) \\
&= \sum_{s\in\mathcal{S}}\frac{1}{\pi(s)}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i1\{S_i=s\}\left[Y_i(D_i(1)) - \tau D_i(1) - U_i^{\top}\eta_{1s}\right] + o_p(1)
\end{aligned} \tag{S.P.4}$$

where the second equality is based on that

$$n^{-1/2}\sum_{i=1}^{n}A_i1\{S_i=s\}U_i\left[Y_i(D_i(1)) - \tau D_i(1) - U_i^{\top}\eta_{1s}\right] = O_p(1),$$

which is implied by the asymptotic normality of (S.P.7), 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 (S.P.3)

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

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

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

Substituting (S.P.4), (S.P.5) and (S.P.6) into (S.P.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_i 1\{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 \text{(S.P.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_i 1\{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= \right. \\
&\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_i 1\{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_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 - \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 \text{(S.P.8)}
\end{aligned}$$

Define

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

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

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

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 (S.P.8), we first argue that

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

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

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

where

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

Then we have

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

Define

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

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

Thus $R_{n,1}|S^{(n)}, A^{(n)} \overset{d}{=} \tilde{R}_{n,1}|S^{(n)}, A^{(n)}$ (and as a by-product $R_{n,1} \overset{d}{=} \tilde{R}_{n,1}$). 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, $(R_{n,1}, R_{n,2}) \overset{d}{=} (\tilde{R}_{n,1}, R_{n,2})$. Define the following partial sum process

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

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

$$g_n\big(F(s)+\pi(s)p(s)\big) - g_n\big(F(s)\big) = \frac{1}{\sqrt{n}} \sum_{i=\lfloor nF(s)\rfloor+1}^{\lfloor n(F(s)+\pi(s)p(s))\rfloor} \tilde{\rho}^s_{is}(1) \rightsquigarrow \mathcal{N}\Big(0, \pi(s)p(s)\operatorname{var}\big(\rho_s(1)|S=s\big)\Big). \tag{S.P.9}$$

Furthermore, since

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

it follows that

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

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

$$
\begin{aligned}
(R_{n,1}, R_{n,2}) &\overset{d}{=} (\tilde{R}_{n,1}, R_{n,2}) = (R^*_{n,1}, R_{n,2}) + o_p(1) \qquad \text{(S.P.11)}\\
R^*_{n,1} &\rightsquigarrow \mathcal{N}\left(0, \sum_{s\in\mathcal{S}} \left[\pi(s)p(s)\operatorname{var}\left(\rho_s(1)|S=s\right) + \left[1-\pi(s)\right]p(s)\operatorname{var}\left(\rho_s(0)|S=s\right)\right]\right)\\
&= \mathcal{N}\left(0, \mathbb{E}\left[\pi(S)\left(\rho_S(1) - \mathbb{E}[\rho_S(1)|S]\right)^2 + (1-\pi(S))\left(\rho_S(0) - \mathbb{E}[\rho_S(0)|S]\right)^2\right]\right)\\
&=: \zeta_1
\end{aligned}
$$

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

$$
(R_{n,1}, R_{n,2}) \overset{d}{=} (R^*_{n,1}, R_{n,2}) + o_p(1) \rightsquigarrow \left(\zeta_1, \zeta_2\right)
$$

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 of part (i). The proof of part (ii), available upon request, is omitted in the interest of space as it is quite similar to that of part (ii) of Theorem 3.1.

S.Q Technical Lemmas Used in the Proof of Theorem 3.1

Lemma S.Q.1. *Suppose assumptions in Theorem 3.1 hold. Then, we have*

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

$$R_{n,2} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{1}{1-\pi(S_i)} - 1 \right) (1-A_i)\tilde{\mu}^Y(0, S_i, X_i) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} A_i \tilde{\mu}^Y(0, S_i, X_i) + o_p(1),$$

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

where for $a = 0, 1$,

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

Proof. We have

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

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 (S.Q.1).

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

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

where the last equality is due to Assumption 3. Thus

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

In addition, we note that

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

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 $\left\{X_i^s : 1 \le i \le n\right\}$ 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 $s \in \mathcal{S}$,

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

In addition, we have

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$
\begin{aligned}
&= \sum_{s \in \mathcal{S}} \frac{n(s)}{\sqrt{n}} \mathbb{E}[W - Z | S = s] - \sqrt{n}\mathbb{E}[W - Z] \\
&= \frac{1}{\sqrt{n}} \sum_{s \in \mathcal{S}} \sum_{i=1}^{n} \Big(1\{S_i = s\} \mathbb{E}[W_i - Z_i | S_i = s] \Big) - \sqrt{n}\mathbb{E}[W - Z] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}[W_i - Z_i | S_i] - \sqrt{n}\mathbb{E}[W - Z] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(\mathbb{E}[W_i - Z_i | S_i] - \mathbb{E}[W_i - Z_i] \big) . \qquad \text{(S.Q.5)}
\end{aligned}
$$

Combining (S.Q.3) and (S.Q.5), we have

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

where the second equality holds because

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

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

Lemma S.Q.2. *Under the assumptions in Theorem 3.1, we have*

$$
\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Xi_1(\mathcal{D}_i, S_i) A_i \rightsquigarrow \mathcal{N}\left(0, \mathbb{E}\pi(S_i)\Xi_1^2(\mathcal{D}_i, S_i)\right), \\
&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Xi_0(\mathcal{D}_i, S_i)(1 - A_i) \rightsquigarrow \mathcal{N}\left(0, \mathbb{E}(1 - \pi(S_i))\Xi_0^2(\mathcal{D}_i, S_i)\right), \quad \textit{and} \\
&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Xi_2(S_i) \rightsquigarrow \mathcal{N}(0, \mathbb{E}\Xi_2^2(S_i)),
\end{aligned}
$$

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

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

Therefore,

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

and by construction,

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

are independent. Last, we have

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

This implies the desired result. □

Lemma S.Q.3. *Suppose assumptions in Theorem 3.1 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

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

Then, we have

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

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 3(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,

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

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

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

This implies

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

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

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

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

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

For the same reason, we can show that

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

Last, by the similar argument, we have

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

$$
\begin{aligned}
&= \sum_{s \in \mathcal{S}} \frac{n(s)}{n} \Xi_2^2(s) + o_p(1) \\
&\xrightarrow{p} \sum_{s \in \mathcal{S}} p(s) \Xi_2^2(s) = \mathbb{E}\Xi_2^2(S_i) = \sigma_2^2.
\end{aligned}
$$

□

S.R An Additional Simulation

In this section, we use an additional simulation to demonstrate that when probabilities of treatment assignment $\{\pi(s)\}$ are heterogeneous across strata, the TSLS estimator could be inconsistent. The data generating process we consider here, denoted DGP(iv), is almost the same as DGP(i) in Section 6; the only difference is in $Y_i(a)$:

$$
\begin{aligned}
Y_i(1) =& 2 + S_i^2 + 0.7X_{1,i}^2 + X_{2,i} + 4Z_i + \varepsilon_{1,i} \\
Y_i(0) =& 1 + 0.7X_{1,i}^2 + X_{2,i} + 4Z_i + \varepsilon_{2,i}.
\end{aligned}
$$

The rationale for specifying this DGP is to allow a difference between the probabilistic limit of $\hat{\tau}_{TSLS}$ and τ. We consider randomization schemes SRS and SBR with $(\pi(1), \pi(2), \pi(3), \pi(4)) = (0.2, 0.2, 0.2, 0.5)$. We do not consider randomization scheme WEI or BCD because for these two, $\pi(s) = 0.5$ for all $s \in \mathcal{S}$. The rest of the simulation setting is the same as DGP(i) in Section 6. Table 1 presents the empirical sizes. We see that all estimators, except the TSLS estimator, have the empirical sizes converging to 0.05 as sample size increases. The size distortion of TSLS becomes larger as sample size increases due to its inconsistency in this setting.

References

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.

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

Table 1: NA, TSLS, L, S, NL, F, NP, SNP, R stand for the unadjusted estimator, TSLS estimator, optimally linearly adjusted estimator, Ansel et al. (2018)'s S estimator with X_i as regressor, nonlinearly (logistic) adjusted estiamtor, further efficiency improving estimator, nonparametrically adjusted estimator, Ansel et al. (2018)'s S estimator with $\mathring{\Psi}_{i,n}$ defined in (6.1) as regressor, and estimator with regularized adjustments, respectively.

	$n = 400$		$n = 800$		$n = 1200$	
Methods	SRS	SBR	SRS	SBR	SRS	SBR
			DGP(iv)			
Size						
NA	0.043	0.036	0.048	0.047	0.044	0.050
TSLS	0.069	0.072	0.114	0.113	0.149	0.142
L	0.064	0.060	0.056	0.055	0.053	0.054
S	0.064	0.060	0.056	0.055	0.053	0.054
NL	0.060	0.058	0.056	0.054	0.053	0.054
F	0.083	0.077	0.064	0.062	0.058	0.057
NP	0.201	0.174	0.107	0.096	0.080	0.078
SNP	0.201	0.190	0.102	0.099	0.076	0.077
R	0.078	0.079	0.063	0.061	0.060	0.061

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.

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.

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

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.

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.