# Semi-supervised Regression Analysis with Model Misspecification and High-dimensional Data

*Ye Tian*[a], *Peng Wu*[b] *and Zhiqiang Tan*[a*]

[a]*Rutgers University,*

[b]*Beijing Technology and Business University*

## Supplementary Material

This Supplementary Material consists of Sections S1–S10. Section S1 provides a heuristic discussion on conditions for the proposed estimator to be $\sqrt{N}$-consistent and asymptotic normal. Section S2 provides a comparison of our paper with several related papers with regression of $Y$ on high-dimensional $\boldsymbol{Z} = \boldsymbol{X}$. Section S3 introduces regularity conditions for the analysis of asymptotic properties of the proposed estimators. Section S4 contains technical tools used in proofs of lemmas in Section S5. Section S5 presents lemmas used in proofs of Proposition 2 and Theorem 1. Section S6 gives the technical proofs of Proposition 1 and Proposition 2. Section S7 provides the proof of Theorem 1. Section S8 includes contents of the extension to stratified sampling settings, including the proof of Proposition 3 and the variance comparison. Sections S9 and S10 present details of the numerical implementation

∗Correspondence to: ztan@stat.rutgers.edu.

and application, respectively.

## S1 Discussion on asymptotic normality of the proposed estimator

First, we explain why, in case (a) of $\boldsymbol{Z} = 1$, $\sqrt{N}$-consistency and asymptotic normality can be achieved using a linear OR working model, provided that either the OR model or the PS model is correctly specified. In the general case of $\boldsymbol{Z} \in \mathbb{R}^m$, under certain regularity conditions, we have the following expansion

$$\begin{aligned}
&\tau(\boldsymbol{O}, \hat{\alpha}, \hat{\beta}, \hat{\gamma}) - \tau(\boldsymbol{O}, \bar{\alpha}, \beta^*, \bar{\gamma}) \\
&= \frac{\partial \tau(\boldsymbol{O}, \alpha, \beta, \gamma)}{\partial \alpha}\bigg|^{\mathrm{T}}_{(\alpha=\bar{\alpha}, \beta=\beta^*, \gamma=\bar{\gamma})} (\hat{\alpha} - \bar{\alpha}) + \frac{\partial \tau(\boldsymbol{O}, \alpha, \beta, \gamma)}{\partial \gamma}\bigg|^{\mathrm{T}}_{(\alpha=\bar{\alpha}, \beta=\beta^*, \gamma=\bar{\gamma})} (\hat{\gamma} - \bar{\gamma}) + o_p(\frac{1}{\sqrt{N}}).
\end{aligned}$$

To obtain $\sqrt{N}$-consistency, we need the population estimating equation in $(\alpha, \gamma)$:

$$\begin{cases}
\mathbb{E}\left[\left\{\dfrac{\partial \tau(\boldsymbol{O}, \alpha, \beta, \gamma)}{\partial \gamma}\right\}^{\mathrm{T}}\Big|_{(\alpha=\bar{\alpha}, \beta=\beta^*, \gamma=\bar{\gamma})}\right] = -\mathbb{E}\left[R\dfrac{1-\pi(X;\bar{\gamma})}{\pi(X;\bar{\gamma})}\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\boldsymbol{Z}\bigotimes\boldsymbol{F}\right] = 0, \\
\mathbb{E}\left[\boldsymbol{Z}\left\{\dfrac{\partial \tau(\boldsymbol{O}, \alpha, \beta, \gamma)}{\partial \alpha}\right\}^{\mathrm{T}}\Big|_{(\alpha=\bar{\alpha}, \beta=\beta^*, \gamma=\bar{\gamma})}\right] = \mathbb{E}\left[\left\{1 - \dfrac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\right\}\psi'(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\boldsymbol{Z}\bigotimes\boldsymbol{G}\right] = 0.
\end{cases} \tag{S1.1}$$

similarly as discussed in Tan (2020a), Section 3.2. From the system of equations (S1.1), there are a total of $m \times (p+1) + m \times (q+1)$ equations, but $(p+1) + (q+1)$

unknown parameters, i.e., $(\alpha, \gamma)$. Therefore, $\alpha$ and $\gamma$ cannot be identified from the equations without additional conditions.

When $\boldsymbol{Z} = 1$ (i.e., case (a)) and the OR model is linear (i.e., $\psi(t) = t$ is the identity function), the system of equations (S1.1) becomes

$$\begin{cases} \mathbb{E}\left\{ R\dfrac{1-\pi(\boldsymbol{X};\bar{\gamma})}{\pi(\boldsymbol{X};\bar{\gamma})}(Y - \bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\boldsymbol{F} \right\} = 0, \\ \mathbb{E}\left\{ \Big(1 - \dfrac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\Big)\boldsymbol{G} \right\} = 0. \end{cases}$$

In this case, with $\boldsymbol{G} = \boldsymbol{F}$ as in Tan (2020a), the above system is just-identified, with the same number of equations and number of parameters, and naturally serves as population estimating equations for $\bar{\alpha}$ and $\bar{\gamma}$. This is the core idea of Tan (2020a), explaining why a $\sqrt{N}$-consistent and asymptotically normal estimator can be achieved in the case where $\boldsymbol{Z} = 1$, $\boldsymbol{G} = \boldsymbol{F}$, and a linear OR model is used, provided that either the OR model or the PS model is correctly specified.

Second, we explain why, for a general choice of $\boldsymbol{Z}$, the correct specification of the propensity score model is assumed to obtain a $\sqrt{N}$-consistent and asymptotically normal estimator.

From the above discussion of (S1.1), there are two key challenges for a general choice of $\boldsymbol{Z}$:

- When $\boldsymbol{Z}$ is multi-dimensional, the number of equations and the number of parameters in (S1.1) do not match.

- When the OR model is nonlinear or $\boldsymbol{F} \neq \boldsymbol{G}$, the two equations in (S1.1) are coupled in terms of dependency on $\bar{\alpha}$ and $\bar{\gamma}$, so that a direct approach of defining $\bar{\alpha}$ and $\bar{\gamma}$ from the two equations would not work.

In our paper, we address the first challenge by carefully specifying the form of $\boldsymbol{G}$ as follows

$$\boldsymbol{G} = [\boldsymbol{F}^{\mathrm{T}}, \{\boldsymbol{Z} \otimes \boldsymbol{F}\}^{\mathrm{T}}]^{\mathrm{T}}.$$

Then we circumvent the second challenge by assuming that the PS model is correctly specified. Alternative strategies can be considered by extending the iterative approach of Ghosh and Tan (2022). Investigating these topics would be interesting for future work.

Finally, it is worth pointing out that correct specification of the PS model is automatically satisfied in the classic SSL and stratified sampling setup. In such cases, the PS model is constant and known, as noted in Remark 1.

## S2 Comparison with additional related papers

We provide a comparison of our paper with several related papers with regression of $Y$ on high-dimensional $\boldsymbol{Z} = \boldsymbol{X}$ as mentioned in Section 2.1.

**First**, we compare our paper with Chakrabortty et al. (2019). Chakrabortty et al. (2019) consider estimating a high-dimensional parameter $\theta^*$, which is defined

as the minimizer of

$$\mathbb{L}(\theta) := \mathbb{E}[L(Y, \boldsymbol{X}, \theta)],$$

where $L(Y, \boldsymbol{X}, \theta)$ is a convex loss function. Under the setting of *semi-supervised learning with covariate shift* (i.e., the conditional distributions of $Y$ given $\boldsymbol{X}$ in the labeled and unlabeled datasets are assumed to be the same, whereas the marginal distributions of $\boldsymbol{X}$ are different), Chakrabortty et al. (2019) identify $\mathbb{L}(\theta)$ using the doubly robust form (their equation (2.1)):

$$\mathbb{L}(\theta) = \mathbb{E}[\phi(\boldsymbol{X}, \theta)] + \mathbb{E}\left[\frac{T}{\pi(\boldsymbol{X})}\{L(Y, \boldsymbol{X}, \theta) - \phi(\boldsymbol{X}, \theta)\}\right], \tag{S2.2}$$

where $T \in \{0, 1\}$ is the indicator of $Y$ being observed, $\phi(\boldsymbol{X}, \theta) = \mathbb{E}[L(Y, \boldsymbol{X}, \theta)|\boldsymbol{X}]$ and $\pi(\boldsymbol{X}) = \mathbb{P}(T = 1|\boldsymbol{X})$. Instead of dealing with the general loss function (S2.2), they focus on a specific yet common case where $L(Y, \boldsymbol{X}, \theta)$ satisfies (their equation (2.3)):

$$\nabla L(Y, \boldsymbol{X}, \theta) \equiv \frac{\partial}{\partial \theta} L(Y, \boldsymbol{X}, \theta) = \boldsymbol{h(X)}\{Y - g(\boldsymbol{X}, \theta)\} \tag{S2.3}$$

for some functions $h(\boldsymbol{X})$ with the same dimension as $\theta$. In this case, $\frac{\partial}{\partial \theta}\phi(\boldsymbol{X}, \theta) = \boldsymbol{h(X)}\{m(\boldsymbol{X}) - g(\boldsymbol{X}, \theta)\}$, where $m(\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X})$. Note that (S2.3) can be motivated from a conditional mean model $m(\boldsymbol{X}) = g(\boldsymbol{X}, \theta^*)$, but such a model is allowed to be misspecified. The authors propose a debiased and doubly robust (DDR) estimator $\hat{\theta}^{\text{DDR}}$ for $\theta^*$, as a minimizer of the following objective function plus a LASSO

penalty on $\theta$ (after ignoring cross-fitting):

$$\hat{\mathbb{L}}(\theta) = \tilde{\mathbb{E}}\left[\hat{\phi}(\boldsymbol{X},\theta) + \frac{T}{\hat{\pi}(\boldsymbol{X})}\{L(Y,\boldsymbol{X},\theta) - \hat{\phi}(\boldsymbol{X},\theta)\}\right],$$

where $\hat{\pi}(\boldsymbol{X})$ is an estimator of $\pi(\boldsymbol{X})$ through a PS model, and $\hat{\phi}(\boldsymbol{X},\theta)$ is an estimator of $\phi(\boldsymbol{X},\theta)$, satisfying $\frac{\partial}{\partial\theta}\hat{\phi}(\boldsymbol{X},\theta) = \boldsymbol{h}(\boldsymbol{X})\{\hat{m}(\boldsymbol{X}) - g(\boldsymbol{X},\theta)\}$, and $\hat{m}(\boldsymbol{X})$ is an estimator of $m(\boldsymbol{X})$ through an OR model. By simple calculation, the gradient of $\hat{\mathbb{L}}(\theta)$ is

$$\begin{aligned}\frac{\partial}{\partial\theta}\hat{\mathbb{L}}(\theta) &= \tilde{\mathbb{E}}\left[\boldsymbol{h}(\boldsymbol{X})\{\hat{m}(\boldsymbol{X}) - g(\boldsymbol{X},\theta)\} + \frac{T}{\hat{\pi}(\boldsymbol{X})}\boldsymbol{h}(\boldsymbol{X})\{Y - \hat{m}(\boldsymbol{X})\}\right] \\ &= \tilde{\mathbb{E}}\left[\frac{T}{\hat{\pi}(\boldsymbol{X})}\boldsymbol{h}(\boldsymbol{X})\{Y - g(\boldsymbol{X},\theta)\} + \left(1 - \frac{T}{\hat{\pi}(\boldsymbol{X})}\right)\boldsymbol{h}(\boldsymbol{X})\{\hat{m}(\boldsymbol{X}) - g(\boldsymbol{X},\theta)\}\right].\end{aligned}$$

This matches the AIPW estimating function (2.4) in our paper, by linking $\hat{\pi}(\boldsymbol{X})$, $\hat{m}(\boldsymbol{X})$, $g(\boldsymbol{X},\theta)$, and $\boldsymbol{h}(\boldsymbol{X})$ above with $\hat{\pi}(\boldsymbol{X})$, $\hat{m}(\boldsymbol{X})$, $\psi(\beta^{\mathrm{T}}\boldsymbol{Z})$, and $\boldsymbol{Z}$ in our paper. Hence, Chakrabortty et al. (2019) and our paper intersects in the idea of using AIPW estimating functions.

Our paper, however, differs from Chakrabortty et al. (2019) in two important technical aspects. (i) Chakrabortty et al. (2019) allow $\theta$ to be a high-dimensional parameter and study the $L_1$ and $L_2$ convergence rates for the DDR estimator $\hat{\theta}^{\mathrm{DDR}}$, which are slower than $N^{-1/2}$. (ii) Chakrabortty et al. (2019) then propose a desparsified DDR estimator $\hat{\theta}^{\text{D-DDR}}$, and prove that each *scalar* element of $\hat{\theta}^{\text{D-DDR}}$ is $\sqrt{N}$-consistent and asymptotically normal (called entry-wise asymptotic normality) if

both OR and PS models are correct such that $\hat{\pi}(\cdot)$ and $\hat{m}(\cdot)$ converge to the *true values*, $\pi(\cdot)$ and $m(\cdot)$, at fast enough rates (specifically, the product of the estimation errors is smaller than $N^{-1/2}$). In contrast, our setting corresponds to a regression model of $Y$ on a low-dimensional sub-vector $\boldsymbol{Z}$ of $\boldsymbol{X}$ with coefficient vector $\theta^*$, but our proposed estimator is shown to be $\sqrt{N}$-consistent and asymptotically normal for $\theta^*$ (not just entry-wise), even when $\hat{m}(\cdot)$ is inconsistent for the true value and $\hat{\pi}(\cdot)$ converges to the true value at a rate slower than $N^{-1/2}$ (but faster than $N^{-1/4}$), with misspecified OR model and correctly specified PS model. Hence the aforementioned product of estimation errors may be greater than $N^{-1/2}$. Note that $\hat{m}(\cdot)$ is consistent for the *target value*, different from the true value, and may also converge slower than $N^{-1/2}$ (but faster than $N^{-1/4}$), with misspecified OR model; see Section 3.2 of our paper. This is the main advantage of regularized calibrated estimation in our method. Therefore, our paper is not a special case of Chakrabortty et al. (2019).

**Second**, we compare our work with Deng et al. (2023), Zhang et al. (2023), and Chen and Zhang (2023). The settings of these works also differ from ours. We summarize the differences below, which demonstrate our unique contributions.

- Deng et al. (2023) consider estimating $\theta^*$ defined by

$$\theta^* = \text{argmin}_\theta \ \ \mathbb{E}(Y - \theta^{\mathrm{T}}\boldsymbol{X})^2,$$

where $\boldsymbol{X}$ is a high-dimensional covariate vector and $Y$ is the response. Under *the classical semi-supervised setup without covariate shift* (i.e., $(\boldsymbol{X}, Y)$ have the same joint distributions in the labeled and unlabeled datasets), the authors propose several Lasso/Dantzig selector-based estimators for $\theta$, denoted as $\hat{\theta}$, and then derive the convergence rates of $\hat{\theta}$ to the target value $\theta^*$ in terms of the $L_q$-norm, $||\hat{\theta} - \theta^*||_q$, for $q = 1, 2$. In particular, the semi-supervised Lasso estimator in Deng et al. (2023) (equation 3.6 in their paper), after ignoring the Lasso penalty on $\theta$, can be shown to be a solution to the following equation

$$\frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{m}(\boldsymbol{X}_i)]\boldsymbol{X}_i + \frac{1}{N}\sum_{i=1}^{N}[\hat{m}(\boldsymbol{X}_i) - \theta^{\mathrm{T}}\boldsymbol{X}_i]\boldsymbol{X}_i = 0,$$

where $\hat{m}(\boldsymbol{X})$ is an estimator of $m(\boldsymbol{X})$ as in the earlier discussion. This matches the AIPW estimating equation (5.1) in the classical semi-supervised setting in our paper, by linking $\hat{m}(\boldsymbol{X}_i)$ and $\theta^{\mathrm{T}}\boldsymbol{X}_i$ with $\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)$ and $\psi(\beta^{\mathrm{T}}\boldsymbol{Z}_i)$ in our paper.

There are several differences between Deng et al. (2023) and our paper. First, Deng et al. (2023) focus on the classical semi-supervised setup without covariate shift, whereas we focus on a setup that incorporates covariate shift. Second, Deng et al. (2023) consider a linear model $\theta^{*\mathrm{T}}\boldsymbol{X}$ directly with high-dimensional $\boldsymbol{X}$, whereas we consider a generalized linear model $\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})$ with possibly nonlinear $\psi(\cdot)$ and low-dimensional $\boldsymbol{Z}$ as a subvector of $\boldsymbol{X}$. Both these models are

used as approximations and allowed to be misspecified. Finally, Deng et al. (2023) concentrate on analyzing $||\hat{\theta} - \theta^*||_q$, which is slower than $N^{-1/2}$, without dealing with inference such as confidence intervals, whereas our theoretical analysis provides $\sqrt{N}$-consistency and asymptotic normality for the proposed estimator of $\beta^*$, thereby enabling valid confidence intervals under suitable sparsity conditions.

- Zhang et al. (2023) consider a high-dimensional linear regression model:

$$Y = \theta^{*\mathrm{T}} \boldsymbol{X} + \epsilon,$$

where $\theta^*$ is an unknown coefficient, $\mathbb{E}(\epsilon|\boldsymbol{X}) = 0$, and $\mathbb{E}(\epsilon^2|\boldsymbol{X}) = \sigma_\epsilon^2$. Under the setting of semi-supervised learning with covariate shift, their goal is to estimate the linear regression function, $m(x) = \theta^{*\mathrm{T}} x$, for a given query vector $x$. The authors propose a debiased-Lasso type estimator for $m(x)$, depending on an estimator $\hat{\pi}(\boldsymbol{X})$ with a correctly specified PS model, and show its $\sqrt{N}$-consistency and asymptotic normality under suitable sparsity conditions.

There are also notable differences between Zhang et al. (2023) and our paper. First, Zhang et al. (2023) consider a linear model $\theta^{*\mathrm{T}} \boldsymbol{X}$ directly with high-dimensional $\boldsymbol{X}$, whereas we consider a generalized linear model $\psi(\beta^{*\mathrm{T}} \boldsymbol{Z})$ with possibly nonlinear $\psi(\cdot)$ and low-dimensional $\boldsymbol{Z}$ as a subvector of $\boldsymbol{X}$. This is similar to the second difference between Deng et al. (2023) and our paper

discussed earlier, but the linear regression model in Zhang et al. (2023) is assumed to be correctly specified along with homoscedastic noises. Second, while the goal of Zhang et al. (2023) is inference about the linear regression function $m(x) = \theta^{*\mathrm{T}}x$ at a given point $x$, our goal is inference about the entire vector of regression coefficients $\beta^*$. In fact, because $\beta^*$ is low-dimensional in our theory, the asymptotic normality of our proposed estimator $\hat{\beta}$ directly gives asymptotic normality of the estimator $\hat{\beta}^{\mathrm{T}}z$ and hence enables valid confidence intervals for $\beta^{*\mathrm{T}}z$ and $\psi(\beta^{*\mathrm{T}}z)$ at a given point $z$.

- Chen and Zhang (2023) consider the following regression models

$$Y_i = \theta^{*\mathrm{T}}\boldsymbol{X}_i + \epsilon_i = \phi^* V_i + \alpha^{*\mathrm{T}}\boldsymbol{W}_i + \epsilon_i, \tag{S2.4}$$

$$V_i = \gamma^{*\mathrm{T}}\boldsymbol{W}_i + \delta_i, \tag{S2.5}$$

where $\boldsymbol{X}_i = (V_i, \boldsymbol{W}_i^{\mathrm{T}})^{\mathrm{T}}$, with $V_i$ being a particular covariate, and $\theta^* = (\phi^*, \alpha^{*\mathrm{T}})^{\mathrm{T}}$. Model (S2.4) is allowed to be misspecified, with $\theta^*$ defined as

$$\theta^* = \mathrm{argmin}_\theta \mathbb{E}(Y - \theta^{\mathrm{T}}\boldsymbol{X})^2.$$

Model (S2.5) is assumed to be correctly specified with $E(\delta_i|\boldsymbol{W}_i) = 0$ for all $i$, which is automatically satisfied in the Gaussian design (i.e., $\boldsymbol{X}_i = (V_i, \boldsymbol{W}_i^{\mathrm{T}})^{\mathrm{T}}$ is jointly Gaussian). Chen and Zhang (2023) consider semi-supervised learning with covariate shift satisfying $R_i \perp\!\!\!\perp V_i|\boldsymbol{W}_i$, where $R_i = 1$ or 0 for $i$ from the

labeled or unlabeled dataset. The authors propose various estimators for $\phi^*$ and show $\sqrt{N}$-consistency and asymptotic normality under suitable sparsity conditions. In contrast with Chen and Zhang (2023), our paper considers semi-supervised learning with covariate shift satisfying $R_i \perp\!\!\!\perp Y_i | \boldsymbol{X}_i$ (Assumption 1), and studies inference for a generalized linear model $\psi(\beta^{*\mathrm{T}} \boldsymbol{Z})$ for $Y$ with possibly nonlinear $\psi(\cdot)$ and low-dimensional $\boldsymbol{Z}$ as a subvector of $\boldsymbol{X}$. Therefore, Chen and Zhang (2023) and our paper deal with different semi-supervised settings and objectives of estimation.

## S3 Regularity assumptions

In this section, we introduce some regularity conditions for the analysis of asymptotic properties of the proposed estimators.

The following Assumption is taken from Tan (2020a), which are plausible as discussed there.

**Assumption 1** (Regularity conditions for $\hat{\gamma}$)**.** Suppose that the following conditions are satisfied:

(i) $\max_{j=0,\ldots,p} |f_j(\boldsymbol{X})| \leq C_0$ a.s. for a constant $C_0 \geq 1$;

(ii) $\bar{\gamma}^{\mathrm{T}} \boldsymbol{F} \geq B_0$ a.s. for a constant $B_0 \in \mathbb{R}$, that is, $\pi(\boldsymbol{X}; \bar{\gamma})$ is bounded from below by $\{1 + \exp(-B_0)\}^{-1}$;

(iii) the compatibility condition holds for $\boldsymbol{\Sigma}_{\bar{\gamma}}$ with the subset $S_{\bar{\gamma}} = \{0\} \cup \{j :$

$\bar{\gamma}_j \neq 0, j = 1, \ldots, p\}$ and some constants $\nu_0 > 0$ and $\xi_0 > 1$, where $\boldsymbol{\Sigma}_{\bar{\gamma}} = \mathbb{E}\{Rw(\boldsymbol{X}; \bar{\gamma})\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}\}$ is the Hessian of $\mathbb{E}\{\ell_{\mathrm{CAL}}(\gamma)\}$ at $\gamma = \bar{\gamma}$;

(iv) $|S_{\bar{\gamma}}|\lambda_0 \leq \zeta_0$ for a sufficiently small constant $\zeta_0 > 0$, depending only on $(A_0, C_0, \xi_0, \nu_0)$, where $\lambda_0 = c_\gamma \sqrt{\ln\{(1+p)/\epsilon\}/N}$, $c_\gamma$ is a constant only depending on $(B_0, C_0)$ and $A_0 > (\xi_0 + 1)/(\xi_0 - 1)$ is a constant.

For studying the properties of $\hat{\alpha}$, we make the following Assumption 2.

**Assumption 2** (Regularity conditions for $\hat{\alpha}$)**.** Let $\psi_1(u)$ denote the derivative of $\psi(u)$. Suppose that the following conditions are satisfied:

(i) $C_1 \leq \psi_1(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) \leq C_2$ a.s. for positive constants $(C_1, C_2)$;

(ii) $\psi_1(u) \leq \psi_1(u') \exp(C_3|u - u'|)$ for any $(u, u')$ and certain constant $C_3 \geq 0$;

(iii) $\max_{j=0,\ldots,q} |g_j(\boldsymbol{X})| \leq C_4$ a.s. for a constant $C_4 \geq 1$;

(iv) $Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})$ is uniformly sub-Gaussian given $\boldsymbol{X}$: $D_0^2 \mathbb{E}(\exp[\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2/D_0^2] - 1|\boldsymbol{X}) \leq D_1^2$ for some positive constants $(D_0, D_1)$;

(v) the compatibility condition holds for $\boldsymbol{\Sigma}_{\bar{\alpha}}$ with the subset $S_{\bar{\alpha}} = \{0\} \cup \{j : \bar{\alpha}_j \neq 0, j = 0, \ldots, q\}$ and some constant $\nu_1 > 0$ and $\xi_1 > 1$, where $\boldsymbol{\Sigma}_{\bar{\alpha}} = \mathbb{E}\left[Rw(\boldsymbol{X}; \bar{\gamma})\psi_1(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\boldsymbol{G}\boldsymbol{G}^{\mathrm{T}}\right]$;

(vi) $(1 + \xi_1)^2 \nu_1^{-2} |S_{\bar{\alpha}}| \lambda_1 \leq \zeta_1$ for a sufficiently small constant $\zeta_1 > 0$, where $\lambda_1 = \max[\lambda_0, c_\alpha \sqrt{\ln\{(1+q)/\epsilon\}/N}]$, $c_\alpha$ is a constant depending on $(B_0, C_2, C_4, D_0, D_1)$;

$c_\gamma$ is defined in Section S6.1.
$c_\alpha$ is defined in Section S6.2.

(vii) let $A_1 > (\xi_1 + 1)/(\xi_1 - 1)$ be a constant. There exist $0 \leq \eta_2, \eta_3 < 1$, such that $\tilde{c}_\alpha |S_{\bar{\alpha}}| \lambda_1 \leq \eta_2$ and $\tilde{c}_\gamma |S_{\bar{\gamma}}| \lambda_0 \leq \eta_3$, where $\tilde{c}_\alpha$ and $\tilde{c}_\gamma$ are both constants.

Assumptions 2(i)–(ii) are mild conditions on the smoothness of the inverse link function $\psi$. Commonly used functions like the identity and logit functions satisfy these requirements. Assumptions 2(iii)–(vii) are similar to those used in related analysis by Tan (2020a). A subtle difference is that the compatibility condition in Tan (2020a) [Assumption 2(ii)] is assumed for $\boldsymbol{\Sigma}_{\bar{\gamma}}$ with $\boldsymbol{F} = \boldsymbol{G}$, whereas our compatibility condition is assumed for $\boldsymbol{\Sigma}_{\bar{\alpha}}$. In our setting, $\boldsymbol{G}$ has a higher dimension than $\boldsymbol{F}$ (except in the case where $\boldsymbol{Z} = 1$).

**Assumption 3.** Suppose that the following conditions are satisfied:

(i) $\boldsymbol{Z} = \{Z_{0,}, \ldots, Z_{m-1}\} \in \mathbb{R}^m$, and $\max_{j=0,\ldots,m-1} |Z_j| \leq C_5$ almost surely for a constant $C_5 \geq 1$, where $m$ is fixed as $N$ increases;

(ii) $\beta \in \Theta_\beta \subset \mathbb{R}^m$, and for $\forall \epsilon > 0$, $\inf_{\beta \in \Theta_\beta : \|\beta - \beta^*\|_1 \geq \epsilon} \mathbb{E}\{\|\tau(\boldsymbol{O}, \bar{\alpha}, \beta, \bar{\gamma})\|_1\} > 0$;

(iii) $\mathbb{E}\{\sup_{\beta \in \Theta_\beta} \|\tau(\boldsymbol{O}, \bar{\alpha}, \beta, \bar{\gamma})\|_1\} < \infty$;

(iv) $\psi_1(\beta^{*\mathrm{T}} \boldsymbol{Z}) \leq C_6$ a.e., for some constant $C_6 > 0$;

(v) $|S_{\bar{\gamma}}| \sqrt{\ln(1+p)\ln(1+q)} = o(N^{1/2})$ and $|S_{\bar{\alpha}}| \ln(1+q) = o(N^{1/2})$.

Assumptions 3(ii)–(iii) are standard conditions in the asymptotic theory of estimating equations. Assumption 3(iv) is a mild condition on the smoothness of the

$\tilde{c}_\alpha$ and $\tilde{c}_\gamma$ depend on $(A_0, A_1, C_0, C_3, C_4, M_0, \nu_1, \xi_1, \zeta_0, \zeta_1)$ and $(A_0, A_1, C_0, C_1, C_3, C_4, D_0, D_1, M_0, \xi_1, \zeta_0)$, respectively. They are defined in Section S6.2.

inverse link function at the value of $\beta^{*\mathrm{T}}\boldsymbol{Z}$. Assumption 3(v) is comparable to the sparsity requirement in Tan (2020a).

## S4 Technical tools

We state the following concentration inequalities, to facilitate proofs of lemmas in Section S5, which can be obtained from Bühlmann and Van De Geer (2011), Lemmas 14.11, 14.16 & 14.9.

**Lemma 1.** Let $(Y_1, \ldots, Y_N)$ be independent variables such that $\mathbb{E}(Y_i) = 0$ for $i = 1, \ldots, n$ and $\max_{i=1,\ldots,n} |Y_i| \le c_0$ for some constant $c_0$. Then for any $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} Y_i\right| > t\right) \le 2\exp\left(-\frac{nt^2}{2c_0^2}\right).$$

**Lemma 2.** Let $(Y_1, \ldots, Y_N)$ be independent variables such that $\mathbb{E}(Y_i) = 0$ for $i = 1, \ldots, n$ and $(Y_1, \ldots, Y_N)$ are uniformly sub-gaussian: $\max_{i=1,\ldots,n} c_1^2 \mathbb{E}\{\exp(Y_i^2/c_1^2) - 1\} \le c_2^2$. Then for any $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} Y_i\right| > t\right) \le 2\exp\left\{-\frac{nt^2}{8(c_1^2 + c_2^2)}\right\}.$$

**Lemma 3.** Let $(Y_1, \ldots, Y_N)$ be independent variables such that $\mathbb{E}(Y_i) = 0$ for $i = 1, \ldots, n$ and,

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}(|Y_i|^k) \le \frac{k!}{2} c_3^{k-2} c_4^2, \quad k = 2, 3, \ldots,$$

for some constants $(c_3, c_4)$. Then for any $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} Y_i\right| > c_3 t + c_4\sqrt{2t}\right) \leq 2\exp\left(-nt\right).$$

**Lemma 4.** Suppose that $|X| \leq c_5$, $c_5$ is some constant, and $Y$ is sub-gaussian: $c_1^2\mathbb{E}\{\exp\left(X^2/c_1^2\right) - 1\} \leq c_2^2$ for some constants $(c_1, c_2)$. Then $Z = XY^2$ satisfies

$$\mathbb{E}\{|Z - \mathbb{E}(Z)|^k\} \leq \frac{k!}{2} c_6^{k-2} c_7^2, \quad k = 2, 3, \ldots,$$

for $c_6 = 2c_5c_1^2$ and $c_7 = 2c_5c_1c_2$.

**Lemma 5.** Suppose that $Y$ is sub-gaussian: $c_1^2\mathbb{E}\{\exp\left(Y^2/c_1^2\right) - 1\} \leq c_2^2$ for some constants $(c_1, c_2)$. Then

$$\mathbb{E}(|Y|^k) \leq \Gamma(\frac{k}{2} + 1)(c_1^2 + c_2^2)c_1^{k-2}, \quad k = 2, 3 \ldots.$$

# S5 Technical lemmas

## S5.1 Lemmas for the parameter in the PS model

The following Lemmas 6–8 will be used in proofs of Proposition 1 and Theorem 1. Lemma 9 would be used in proofs of lemmas in Section S5.2 and Theorem 1.

**Lemma 6.** Under Assumptions 1(i) and 1(ii), the following statements hold:

(i) Denoted by $\Omega_{00}$ the event that

$$\sup_{j=0,\ldots,p} \left| \tilde{\mathbb{E}} \left[ \{ -Rw(\boldsymbol{X}; \bar{\gamma}) + (1-R) \} f_j(\boldsymbol{X}) \right] \right| \leq \lambda_0.$$

If $\lambda_0 \geq \sqrt{2}\{1+\exp(-B_0)\}C_0\sqrt{\ln\{(1+p)/\epsilon\}/N}$, then $\mathbb{P}(\Omega_{00}) \geq 1-2\epsilon$.

(ii) Denote by $\Omega_{01}$ the event that

$$\sup_{j,k=0,\ldots,p} \left| (\tilde{\boldsymbol{\Sigma}}_\gamma)_{jk} - (\boldsymbol{\Sigma}_\gamma)_{jk} \right| \leq \lambda_0, \tag{S5.6}$$

where $\tilde{\boldsymbol{\Sigma}}_\gamma$ is the empirical version of $\boldsymbol{\Sigma}_\gamma$. If $\lambda_0 \geq 4\exp(-B_0)C_0^2\sqrt{\ln\{(1+p)/\epsilon\}/N}$, then $\mathbb{P}(\Omega_{01}) \geq 1-2\epsilon^2$.

The result of Lemma 6(ii) is taken from Tan (2020b), Lemma 1(ii), the result of Lemma 6(i) can be shown similarly using Lemma 1 in Section S4 and the union bound.

Let $\boldsymbol{\Sigma}_{m^1} = \mathbb{E}\{Rw(\boldsymbol{X}; \bar{\gamma}) | Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) | \boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}\}$, and $\tilde{\boldsymbol{\Sigma}}_{m^1}$ be the sample version of $\boldsymbol{\Sigma}_{m^1}$.

**Lemma 7.** Let $\Omega_{02}$ denote the event that

$$\sup_{j,k=0,\ldots,p} |(\boldsymbol{\Sigma}_{m^1})_{jk} - (\tilde{\boldsymbol{\Sigma}}_{m^1})_{jk}| \leq \sqrt{D_0^2 + D_1^2}\lambda_0. \tag{S5.7}$$

Under Assumptions 1(i), 1(ii) and 2(iv), if

$$\lambda_0 \geq 4\exp(-B_0)C_0^2\sqrt{\frac{\ln\left\{\frac{(1+p)}{\epsilon}\right\}}{N}},$$

then $\mathbb{P}(\Omega_{02}) \geq 1 - 2\epsilon^2$.

**Proof.** Since $|Rw(\boldsymbol{X};\bar{\gamma})f_j(\boldsymbol{X})f_k(\boldsymbol{X})| \leq \exp(-B_0)C_0^2$ for $j,k = 0,\ldots,p$ by Assumptions 1(i), and 1(ii), and $|Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})|$ is uniformly sub-gaussian by Assumption 2(iv), $Rw(\boldsymbol{X};\bar{\gamma})|Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})|f_j(\boldsymbol{X})f_k(\boldsymbol{X})$ is uniformly sub-gaussian. By Lemma 2, we have

$$\mathbb{P}\left\{|(\boldsymbol{\Sigma}_{m^1})_{jk} - (\tilde{\boldsymbol{\Sigma}}_{m^1})_{jk}| > t\right\} \leq \frac{2\epsilon^2}{(1+p)^2},$$

for $j,k = 0,\ldots,p$, where $t = \exp(-B_0)C_0^2\sqrt{8(D_0^2+D_1^2)}\sqrt{\ln\{(1+p)^2/\epsilon^2\}/N}$. By union bounds, (S5.7) holds. □

Let $\boldsymbol{\Sigma}_{m^2} = \mathbb{E}[Rw(\boldsymbol{X};\bar{\gamma})\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}]$, the sample version of $\boldsymbol{\Sigma}_{m^2}$, $\tilde{\boldsymbol{\Sigma}}_{m^2} = \tilde{\mathbb{E}}[Rw(\boldsymbol{X};\bar{\gamma})\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}]$.

**Lemma 8.** Denote by $\Omega_{03}$ the event that

$$\sup_{j,k=0,\ldots,p}|(\boldsymbol{\Sigma}_{m^2})_{jk} - (\tilde{\boldsymbol{\Sigma}}_{m^2})_{jk}| \leq (D_0^2 + D_0D_1)\lambda_0, \tag{S5.8}$$

Under Assumptions 1(i), 1(ii) and 2(iv), if

$$(D_0^2 + D_0 D_1)\lambda_0 \geq 4\exp(-B_0)C_0^2[D_0^2 \ln\{(1+p)/\epsilon\}/N + D_0 D_1\sqrt{\ln\{(1+p)/\epsilon\}/N}],$$

then, $\mathbb{P}(\Omega_{03}) \geq 1 - 2\epsilon^2$. Furthermore, if we assume that $\ln\{(1+p)/\epsilon\}/N < 1$, the above condition reduces to $\lambda_0 \geq 4\exp(-B_0)C_0^2\sqrt{\ln\{(1+p)/\epsilon\}/N}$.

**Proof.** For $j, k = 0, \ldots, p$, the variable $Rw(\boldsymbol{X}; \bar{\gamma})\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2 f_j(\boldsymbol{X}) f_k(\boldsymbol{X})$ is the product of $Rw(\boldsymbol{X}; \bar{\gamma}) f_j(\boldsymbol{X}) f_k(\boldsymbol{X})$ and $\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2$, where $|Rw(\boldsymbol{X}; \bar{\gamma}) f_j(\boldsymbol{X}) f_k(\boldsymbol{X})| \leq \exp(-B_0)C_0^2$ by Assumptions 1(i) and 1(ii); and $\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}$ is sub-gaussian by Assumption 2(iv). By Lemmas 3 and 4 in Section S4, we have

$$\mathbb{P}\left\{|(\tilde{\boldsymbol{\Sigma}}_{m^2})_{j,k} - \boldsymbol{\Sigma}_{m^2})_{j,k}| > 2e^{(-B_0)}C_0^2 D_0^2 t + 2e^{(-B_0)}C_0^2 D_0 D_1\sqrt{2t}\right\} \leq 2\frac{\epsilon^2}{(1+p)^2},$$

for $j, k = 0, \ldots, p$, where $t = \ln\{(1+p)^2/\epsilon^2\}/N$. The result then follows from the union bound. □

**Lemma 9.** Under the conditions of Proposition 1, in the event $\Omega_{00} \cap \Omega_{01}$, we have

$$\tilde{\mathbb{E}}\left[Rw(\boldsymbol{X}; \bar{\gamma})\{(\hat{\gamma} - \bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}^2\right] \leq \exp(\eta_{01})M_0|S_{\bar{\gamma}}|\lambda_0^2, \tag{S5.9}$$

where $\eta_{01} = (A_0 - 1)^{-1}M_0 C_0 \zeta_0$.

**Proof.** By Lemma 6 and Proposition 1, in the event $\Omega_{00} \cap \Omega_{01}$, (4.1) holds, we

obtain

$$\|\hat{\gamma} - \bar{\gamma}\|_1 \leq (A_0 - 1)^{-1} M_0 |S_{\bar{\gamma}}| \lambda_0 \leq (A_0 - 1)^{-1} M_0 \zeta_0, \tag{S5.10}$$

the second inequality holds due to Assumption 1(iv). By the definition of $D^{\dagger}_{\mathrm{CAL}}(\cdot, \cdot)$, we obtain

$$\begin{aligned} D^{\dagger}_{\mathrm{CAL}}(\hat{\gamma}^{\mathrm{T}} \boldsymbol{F}, \bar{\gamma}^{\mathrm{T}} \boldsymbol{F}) =& - \tilde{\mathbb{E}}[R\{\exp(-\hat{\gamma}^{\mathrm{T}} \boldsymbol{F}) - \exp(-\bar{\gamma}^{\mathrm{T}} \boldsymbol{F})\}\{(\hat{\gamma} - \bar{\gamma})^{\mathrm{T}} \boldsymbol{F}\}] \\ =& \tilde{\mathbb{E}}(Rw(\boldsymbol{X}; \bar{\gamma}) \exp[-u\{(\hat{\gamma} - \bar{\gamma})^{\mathrm{T}} \boldsymbol{F}\}]\{(\hat{\gamma} - \bar{\gamma})^{\mathrm{T}} \boldsymbol{F}\}^2) \\ \geq & \exp(-C_0 \|\hat{\gamma} - \bar{\gamma}\|_1) \tilde{\mathbb{E}}[Rw(\boldsymbol{X}; \bar{\gamma})\{(\hat{\gamma} - \bar{\gamma})^{\mathrm{T}} \boldsymbol{F}\}^2] \\ \geq & \exp(-\eta_{01}) \tilde{\mathbb{E}}[Rw(\boldsymbol{X}; \bar{\gamma})\{(\hat{\gamma} - \bar{\gamma})^{\mathrm{T}} \boldsymbol{F}\}^2]. \end{aligned}$$

The second equality holds by the mean value theorem and $u$ is some scalar in $(0, 1)$. Combining the inequality with (4.1), we obtain

$$\tilde{\mathbb{E}}\left[Rw(\boldsymbol{X}; \bar{\gamma})\{(\hat{\gamma} - \bar{\gamma})^{\mathrm{T}} \boldsymbol{F}\}^2\right] \leq \exp(\eta_{01}) D^{\dagger}_{\mathrm{CAL}}(\hat{\gamma}^{\mathrm{T}} \boldsymbol{F}, \bar{\gamma}^{\mathrm{T}} \boldsymbol{F}) \leq \exp(\eta_{01}) M_0 |S_{\bar{\gamma}}| \lambda_0^2.$$

□

## S5.2 Lemmas for the parameter in the OR model

The following Lemmas 10–17 will be used in the proofs of Proposition 2 and Theorem 1.

**Lemma 10.** Let $\Omega_{10}$ denote the event that

$$\sup_{j=0,\ldots,q} \left| \tilde{\mathbb{E}} \left[ Rw(\boldsymbol{X}; \bar{\gamma}) \{ Y - \psi(\bar{\alpha}^{\mathrm{T}} \boldsymbol{G}) \} g_j(\boldsymbol{X}) \right] \right| \leq \lambda_1. \tag{S5.11}$$

Under Assumptions 1, 2(iii) and 2(iv), if

$$\lambda_1 \geq \exp(-B_0) C_4 \sqrt{8(D_0^2 + D_1^2)} \sqrt{\ln\{(1+q)/\epsilon\}/N},$$

then $\mathbb{P}(\Omega_{10}) \geq 1 - 2\epsilon$.

**Proof.** Let $S_j = Rw(\boldsymbol{X}; \bar{\gamma}) \{ Y - \psi(\bar{\alpha}^{\mathrm{T}} \boldsymbol{G}) \} g_j(\boldsymbol{X})$ for $j = 0, \ldots, q$. Then, $\mathbb{E}(S_j) = 0$ by the definition of $\bar{\alpha}$. Under Assumptions 1 and 2(iii), $|S_j| \leq \exp(-B_0) C_4 |R\{Y - \psi(\bar{\alpha}^{\mathrm{T}} \boldsymbol{G})\}|$. By Assumption 2(iv), the variables $(S_0, \ldots, S_q)$ are uniformly sub-gaussian: $\max_{j=0,\ldots,q} D_2^2 \mathbb{E}\{\exp(S_j^2/D_2^2) - 1\} \leq D_3^2$, with $D_2 = \exp(-B_0) C_4 D_0$, and $D_3 = e^{(-B_0)} C_4 D_1$. Therefore, by Lemma 2 in Section S4 and the union bound, $\mathbb{P}(\Omega_{10}) \geq 1 - 2\epsilon$, if $\lambda_1 \geq \exp(-B_0) C_4 \sqrt{8(D_0^2 + D_1^2)} \sqrt{\ln\{(1+q)/\epsilon\}/N}$. □

**Lemma 11.** Denote by $\Omega_{11}$ the event that

$$\sup_{j,k=0,\ldots,q} |(\boldsymbol{\Sigma}_{\bar{\alpha}})_{jk} - (\tilde{\boldsymbol{\Sigma}}_{\bar{\alpha}})_{jk}| \leq \lambda_1, \tag{S5.12}$$

where $\tilde{\boldsymbol{\Sigma}}_{\bar{\alpha}}$ is the empirical version of $\boldsymbol{\Sigma}_{\bar{\alpha}}$. Under Assumptions 1(ii), 2(i) and 2(iii), if $\lambda_1 \geq 4 \exp(-B_0) C_2 C_4^2 \sqrt{\ln\{(1+q)/\epsilon\}/N}$, then $\mathbb{P}(\Omega_{11}) \geq 1 - 2\epsilon^2$.

**Proof.** Notice that $|Rw(\boldsymbol{X}; \bar{\gamma}) \psi_1(\bar{\alpha}^{\mathrm{T}} \boldsymbol{G}) g_j(\boldsymbol{X}) g_k(\boldsymbol{X})| \leq e^{(-B_0)} C_2 C_4^2$ for $j, k = 0, 1, \ldots, q$,

by Assumptions 1(ii), 2(i), and 2(iii). Thus,

$$|Rw(\boldsymbol{X};\bar{\gamma})\psi_1(\bar{\alpha}^{\mathrm{T}}G)g_j(\boldsymbol{X})g_k(\boldsymbol{X}) - \mathbb{E}\left\{Rw(\boldsymbol{X};\bar{\gamma})\psi_1(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})g_j(\boldsymbol{X})g_k(\boldsymbol{X})\right\}| \leq 2\exp(-B_0)C_2C_4^2.$$

By union bounds and applying Lemma 1 yields (S5.12). □

**Lemma 12.** For any $\alpha \in \mathbb{R}^{q+1}$, we have

$$D^{\dagger}_{\mathrm{WL}}(\hat{\alpha}^T\boldsymbol{G}, \alpha^{\mathrm{T}}\boldsymbol{G};\hat{\gamma}) + \lambda_\alpha\|\hat{\alpha}_{1:q}\|_1 \leq (\hat{\alpha}-\alpha)^{\mathrm{T}}\tilde{\mathbb{E}}[Rw(\boldsymbol{X};\hat{\gamma})\{Y-\psi(\alpha^{\mathrm{T}}\boldsymbol{G})\}\boldsymbol{G}] + \lambda_\alpha\|\alpha_{1:q}\|_1. \tag{S5.13}$$

**Proof.** For any $u \in (0,1]$, by definition of $\hat{\alpha}$, we have

$$\ell_{\mathrm{WL}}(\hat{\alpha},\hat{\gamma}) + \lambda_\alpha\|\hat{\alpha}_{1:q}\|_1 \leq \ell_{\mathrm{WL}}\{(1-u)\hat{\alpha}+u\alpha;\hat{\gamma}\} + \lambda_\alpha\|(1-u)\hat{\alpha}_{1:q}+u\alpha_{1:q}\|_1,$$

which implies

$$\ell_{\mathrm{WL}}(\hat{\alpha},\hat{\gamma}) - \ell_{\mathrm{WL}}\{(1-u)\hat{\alpha}+u\alpha;\hat{\gamma}\} + \lambda_\alpha u\|\hat{\alpha}_{1:q}\|_1 \leq \lambda_\alpha u\|\alpha_{1:q}\|_1,$$

by the convexity of $L_1$-norm. Dividing both sides of the preceding inequality by $u$ and letting $u \to 0_+$ leads to

$$-\tilde{\mathbb{E}}\left[Rw(\boldsymbol{X};\hat{\gamma})\left\{Y-\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})\right\}\left\{(\hat{\alpha}-\alpha)^{\mathrm{T}}\boldsymbol{G}\right\}\right] + \lambda_\alpha\|\hat{\alpha}_{1:q}\|_1 \leq \lambda_\alpha\|\alpha_{1:q}\|_1,$$

which yields (S5.13) after rearranging using (4.3). □

**Lemma 13.** For any function $h(\boldsymbol{X})$, under the conditions of Proposition 1, in the event $\Omega_{00} \cap \Omega_{01}$,

$$D^{\dagger}_{\mathrm{WL}}\{\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}, h(\boldsymbol{X}); \hat{\gamma}\} \geq \exp(-\eta_{01}) D^{\dagger}_{\mathrm{WL}}\{\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}, h(\boldsymbol{X}); \bar{\gamma}\}. \tag{S5.14}$$

**Proof.** By the definition of $D^{\dagger}_{\mathrm{WL}}(\cdot, \cdot; \cdot)$,

$$\begin{aligned}
& D^{\dagger}_{\mathrm{WL}}\{\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}, h(\boldsymbol{X}); \hat{\gamma}\} \\
=& \tilde{\mathbb{E}}\left(Rw(\boldsymbol{X}; \hat{\gamma})[\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi\{h(\boldsymbol{X})\}]\{\hat{\alpha}^{\mathrm{T}}\boldsymbol{G} - h(\boldsymbol{X})\}\right) \\
=& \tilde{\mathbb{E}}\left(Rw(\boldsymbol{X}; \bar{\gamma}) \exp\{-(\hat{\gamma} - \bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}[\psi(\hat{\alpha}^{T}\boldsymbol{G}) - \psi\{h(\boldsymbol{X})\}]\{\hat{\alpha}^{\mathrm{T}}\boldsymbol{G} - h(\boldsymbol{X})\}\right) \\
\geq& \tilde{\mathbb{E}}\left(Rw(\boldsymbol{X}; \bar{\gamma}) \exp(-\eta_{01})[\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi\{h(\boldsymbol{X})\}]\{\hat{\alpha}^{\mathrm{T}}\boldsymbol{G} - h(\boldsymbol{X})\}\right) \\
=& \exp(-\eta_{01}) \tilde{\mathbb{E}}(Rw(\boldsymbol{X}; \bar{\gamma})[\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi\{h(\boldsymbol{X})\}]\{\hat{\alpha}^{\mathrm{T}}\boldsymbol{G} - h(\boldsymbol{X})\}) \\
=& \exp(-\eta_{01}) D^{\dagger}_{\mathrm{WL}}\{\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}, h(\boldsymbol{X}); \bar{\gamma}\}.
\end{aligned}$$

The inequality holds since in the event $\Omega_{00} \cap \Omega_{01}$, (4.1) holds. □

For functions $h(\boldsymbol{X})$ and $h'(\boldsymbol{X})$, let $Q_{\mathrm{WL}}\{h(\boldsymbol{X}), h'(\boldsymbol{X}); \cdot\} = \tilde{\mathbb{E}}[Rw(\boldsymbol{X}; \cdot)\{h(\boldsymbol{X}) - h'(\boldsymbol{X})\}^2]$.

**Lemma 14.** Suppose Assumption 2(iv) holds, in the event $\Omega_{00} \cap \Omega_{01} \cap \Omega_{03}$, (S5.13) implies

$$
\begin{aligned}
&\exp(-\eta_{01})D^{\dagger}_{\mathrm{WL}}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G},\bar{\alpha}^{\mathrm{T}}\boldsymbol{G};\bar{\gamma})+\lambda_{\alpha}\|\hat{\alpha}_{1:q}\|_1\\
\leq&(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\tilde{\mathbb{E}}\left[Rw(\boldsymbol{X};\bar{\gamma})\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\boldsymbol{G}\right]\\
&+\lambda_{\alpha}\|\bar{\alpha}_{1:q}\|_1+\exp(\eta_{01})\sqrt{M_1|S_{\bar{\gamma}}|\lambda_0^2}\{Q_{\mathrm{WL}}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G},\bar{\alpha}^{\mathrm{T}}\boldsymbol{G};\bar{\gamma})\}^{1/2},
\end{aligned}
$$

where $M_1=\left[(2D_0^2+D_1^2+D_0D_1)\left\{\frac{M_0^2\zeta_0}{(A_0-1)^2}\right\}+(D_0^2+D_1^2)\exp(\eta_{01})M_0\right]$.

**Proof.** Consider the following decomposition,

$$
\begin{aligned}
&(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\tilde{\mathbb{E}}\left[Rw(\boldsymbol{X};\hat{\gamma})\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\boldsymbol{G}\right]\\
=&(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\tilde{\mathbb{E}}\left[Rw(\boldsymbol{X};\bar{\gamma})\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\boldsymbol{G}\right]\\
&+\tilde{\mathbb{E}}\left[R\left\{w(\boldsymbol{X};\hat{\gamma})-w(\boldsymbol{X};\bar{\gamma})\right\}\left\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\right\}\{(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\boldsymbol{G}\}\right],
\end{aligned}
\tag{S5.15}
$$

denoted as $\Delta_0^0+\Delta_1^0$. By the mean value theorem and Cauchy-Schwartz inequality,

$$
\begin{aligned}
\Delta_1^0\leq&\exp(C_0\|\hat{\gamma}-\bar{\gamma}\|_1)\tilde{\mathbb{E}}^{1/2}[Rw(\boldsymbol{X};\bar{\gamma})\{(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\boldsymbol{G}\}^2]\\
&\times\tilde{\mathbb{E}}^{1/2}[Rw(\boldsymbol{X};\bar{\gamma})\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2\{(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}^2]\\
\leq&\exp(\eta_{01})\{Q_{\mathrm{WL}}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G},\bar{\alpha}^{\mathrm{T}}\boldsymbol{G};\bar{\gamma})\}^{1/2}\\
&\times\tilde{\mathbb{E}}^{1/2}[Rw(\boldsymbol{X};\bar{\gamma})\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2\{(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}^2].
\end{aligned}
\tag{S5.16}
$$

We bound the third term in (S5.16). By Assumption 2(iv) and Lemma 5, we have

$$
\mathbb{E}[\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2|\boldsymbol{X}]\leq D_0^2+D_1^2.
$$

Therefore,

$$\mathbb{E}[Rw(\boldsymbol{X};\bar{\gamma})\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2\{(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}^2] \le (D_0^2+D_1^2)\mathbb{E}[Rw(\boldsymbol{X};\bar{\gamma})\{(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}^2].$$

Let $(\tilde{\mathbb{E}}-\mathbb{E})(\boldsymbol{U})$ denote $\tilde{\mathbb{E}}\{\boldsymbol{U}-\mathbb{E}(\boldsymbol{U})\}$ for $\boldsymbol{U}$, which is a function of $(\boldsymbol{X}, R, Y)$. Then in the event $\Omega_{01}$, by (S5.6), we have

$$(\mathbb{E}-\tilde{\mathbb{E}})[Rw(\boldsymbol{X};\bar{\gamma})\{(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}^2] \le \lambda_0\|\hat{\gamma}-\bar{\gamma}\|_1^2.$$

In the event $\Omega_{03}$, by (S5.12), we have

$$(\tilde{\mathbb{E}}-\mathbb{E})[Rw(\boldsymbol{X};\bar{\gamma})\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2\{(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}^2] \le (D_0^2+D_0D_1)\lambda_0\|\hat{\gamma}-\bar{\gamma}\|_1^2.$$

Combining preceding inequalities, we obtain in $\Omega_{00}\cap\Omega_{01}\cap\Omega_{03}$,

$$\begin{aligned}
&\tilde{\mathbb{E}}[Rw(\boldsymbol{X};\bar{\gamma})\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2\{(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}^2]\\
\le&(D_0^2+D_0D_1)\lambda_0\|\hat{\gamma}-\bar{\gamma}\|_1^2+(D_0^2+D_1^2)\left\{\lambda_0\|\hat{\gamma}-\bar{\gamma}\|_1^2+\tilde{\mathbb{E}}[Rw(\boldsymbol{X};\bar{\gamma})\{(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}^2]\right\}\\
\le&(2D_0^2+D_1^2+D_0D_1)\left\{\frac{M_0^2\zeta_0}{(A_0-1)^2}\right\}|S_{\bar{\gamma}}|\lambda_0^2+(D_0^2+D_1^2)\exp(\eta_{01})M_0|S_{\bar{\gamma}}|\lambda_0^2\\
=&M_1|S_{\bar{\gamma}}|\lambda_0^2,
\end{aligned} \tag{S5.17}$$

where the last inequality holds due to (4.1) and (S5.9). Combining (S5.15)–(S5.17), we obtain

$$
\begin{aligned}
&(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\tilde{\mathbb{E}}\left[Rw(\boldsymbol{X};\hat{\gamma})\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\boldsymbol{G}\right] \\
\leq&(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\tilde{\mathbb{E}}\left[Rw(\boldsymbol{X};\bar{\gamma})\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\boldsymbol{G}\right] \\
&+\exp(\eta_{01})(M_1|S_{\bar{\gamma}}|\lambda_0^2)^{1/2}\{Q_{\mathrm{WL}}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G},\bar{\alpha}^{\mathrm{T}}\boldsymbol{G};\bar{\gamma})\}^{1/2}.
\end{aligned}
\tag{S5.18}
$$

The desired result follows by combining (S5.13), (S5.14) and (S5.18) in the event $\Omega_{00}\cap\Omega_{01}\cap\Omega_{03}$. □

**Lemma 15.** Denote $b=\hat{\alpha}-\bar{\alpha}$. Suppose Assumption 2(iv) holds. In the event, $\Omega_{00}\cap\Omega_{01}\cap\Omega_{03}\cap\Omega_{10}$, we have

$$
\begin{aligned}
&\exp(-\eta_{01})D^{\dagger}_{\mathrm{WL}}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G},\bar{\alpha}^{\mathrm{T}}\boldsymbol{G};\bar{\gamma})+(A_1-1)\lambda_1\|b\|_1 \\
\leq&2A_1\lambda_1\sum_{j\in S_{\bar{\alpha}}}|b_j|+\exp(\eta_{01})\sqrt{M_1|S_{\bar{\gamma}}|\lambda_0^2}\{Q_{\mathrm{WL}}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G},\bar{\alpha}^{\mathrm{T}}\boldsymbol{G};\bar{\gamma})\}^{1/2}.
\end{aligned}
\tag{S5.19}
$$

**Proof.** In the event $\Omega_{10}$ , we have

$$
b^{\mathrm{T}}\tilde{\mathbb{E}}[Rw(\boldsymbol{X};\bar{\gamma})\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\boldsymbol{G}]\leq\lambda_1\|b\|_1,
$$

by which and Lemma 14, we have in the event $\Omega_{00}\cap\Omega_{01}\cap\Omega_{03}\cap\Omega_{10}$,

$$
\begin{aligned}
&\exp(-\eta_{01})D^{\dagger}_{\mathrm{WL}}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G},\bar{\alpha}^{\mathrm{T}}\boldsymbol{G};\bar{\gamma})+A_1\lambda_1\|\hat{\alpha}_{1:q}\|_1 \\
\leq&\lambda_1\|b\|_1+A_1\lambda_1\|\bar{\alpha}_{1:q}\|_1+\exp(\eta_{01})\sqrt{M_1|S_{\bar{\gamma}}|\lambda_0^2}\{Q_{\mathrm{WL}}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G},\bar{\alpha}^{\mathrm{T}}\boldsymbol{G};\bar{\gamma})\}^{1/2}.
\end{aligned}
$$

Applying to the preceding inequality the identity $\hat{\alpha}_j = |\hat{\alpha}_j - \bar{\alpha}_j|$ for $j \notin S_{\bar{\alpha}}$ and the triangle inequality

$$|\hat{\alpha}_j| \geq |\bar{\alpha}_j| - |\hat{\alpha}_j - \bar{\alpha}_j|, \quad j \in S_{\bar{\alpha}} \setminus \{0\},$$

and rearranging the result gives

$$\begin{aligned}
&\exp(-\eta_{01}) D^{\dagger}_{\mathrm{WL}}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}, \bar{\alpha}^{\mathrm{T}}\boldsymbol{G}; \bar{\gamma}) + (A_1 - 1)\lambda_1 \|b_{1:q}\|_1 \\
\leq &\lambda_1 |b_0| + 2A_1\lambda_1 \sum_{j \in S_{\bar{\alpha}} \setminus \{0\}} |b_j| + \exp(\eta_{01}) \sqrt{M_1 |S_{\bar{\gamma}}| \lambda_0^2} \{Q_{\mathrm{WL}}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}, \bar{\alpha}^{\mathrm{T}}\boldsymbol{G}; \bar{\gamma})\}^{1/2} \\
\leq &(A_1 + 1)\lambda_1 |b_0| + 2A_1\lambda_1 \sum_{j \in S_{\bar{\alpha}} \setminus \{0\}} |b_j| + \exp(\eta_{01}) \sqrt{M_1 |S_{\bar{\gamma}}| \lambda_0^2} \{Q_{\mathrm{WL}}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}, \bar{\alpha}^{\mathrm{T}}\boldsymbol{G}; \bar{\gamma})\}^{1/2}.
\end{aligned}$$

By adding $(A_1 - 1)\lambda_1 |b_0|$ on both sides of the previous inequality, the conclusion follows. □

**Lemma 16.** Suppose Assumptions 2(ii) and 2(iii) hold. Then, for any $\alpha, \tilde{\alpha} \in \mathbb{R}^{q+1}$,

$$D^{\dagger}_{\mathrm{WL}}(\alpha^{\mathrm{T}}\boldsymbol{G}, \tilde{\alpha}^{\mathrm{T}}\boldsymbol{G}; \bar{\gamma}) \geq \{b^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}_{\alpha}(\tilde{\alpha}) b\} \frac{1 - \exp(-C_{40}\|b\|_1)}{C_{40}\|b\|_1},$$

where $b = \alpha - \tilde{\alpha}$, $C_{40} = C_3 C_4$ and $\tilde{\boldsymbol{\Sigma}}_{\alpha}(\cdot) = \tilde{\mathbb{E}}\left[Rw(\boldsymbol{X}; \bar{\gamma}) \psi_1\{(\cdot)^{\mathrm{T}}\boldsymbol{G}\} \boldsymbol{G}\boldsymbol{G}^{\mathrm{T}}\right]$. Throughout, set $\{1 - \exp(-c)\}/c = 1$, for $c = 0$.

**Proof.** By the definition of $D^{\dagger}_{\mathrm{WL}}(\cdot, \cdot; \cdot)$,

$$D^{\dagger}_{\mathrm{WL}}(\alpha^{\mathrm{T}}\boldsymbol{G}, \tilde{\alpha}^{\mathrm{T}}\boldsymbol{G}; \bar{\gamma})$$

$$
\begin{aligned}
=&\tilde{\mathbb{E}}\left[Rw(\boldsymbol{X};\bar{\gamma})\{\psi(\alpha^{\mathrm{T}}\boldsymbol{G})-\psi(\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G})\}(\alpha^{\mathrm{T}}\boldsymbol{G}-\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G})\right]\\
=&\tilde{\mathbb{E}}\left(Rw(\boldsymbol{X};\bar{\gamma})\left[\int_0^1\psi_1\{\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G}+u(\alpha^{\mathrm{T}}\boldsymbol{G}-\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\mathrm{d}u\right]\{\alpha^{\mathrm{T}}\boldsymbol{G}-\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G}\}^2\right).
\end{aligned}
$$

By Assumptions 2(ii), 2(iii), and the fact that $|\alpha^{\mathrm{T}}\boldsymbol{G}-\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G}|\le\{\sup_{j=0,\ldots,q}|g_j(\boldsymbol{X})|\}\|\alpha-\tilde{\alpha}\|_1\le C_4\|\alpha-\tilde{\alpha}\|_1$, it follows that

$$
\begin{aligned}
&D^{\dagger}_{\mathrm{WL}}(\alpha^{\mathrm{T}}\boldsymbol{G},\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G};\bar{\gamma})\\
\ge&\tilde{\mathbb{E}}\left[Rw(\boldsymbol{X};\bar{\gamma})\left\{\int_0^1\psi_1(\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G})\exp(-uC_3|\alpha^{\mathrm{T}}\boldsymbol{G}-\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G}|)\mathrm{d}u\right\}(\alpha^{\mathrm{T}}\boldsymbol{G}-\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G})^2\right]\\
\ge&\tilde{\mathbb{E}}\{Rw(\boldsymbol{X};\bar{\gamma})\psi_1(\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G})(\alpha^{\mathrm{T}}\boldsymbol{G}-\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G})^2\}\left\{\int_0^1\exp(-uC_{40}\|\tilde{\alpha}-\alpha\|_1)\mathrm{d}u\right\},
\end{aligned}
$$

which gives the desired result since $\int_0^1\exp(-cu)\mathrm{d}u=\{1-\exp(-c)\}/c$ for $c\ge0$. □

**Lemma 17.** Suppose that Assumption 2(v) holds. In the event $\Omega_{11}$, Assumption 2(vi) implies a compatibility condition for $\tilde{\boldsymbol{\Sigma}}_{\bar{\alpha}}$: for any vector $b=(b_0,\ldots,b_q)^{\mathrm{T}}$ such that $\sum_{j\notin S_{\bar{\alpha}}}|b_j|\le\xi_1\sum_{j\in S_{\bar{\alpha}}}|b_j|$, we have

$$
(1-\zeta_1)\nu_1^2\left(\sum_{j\in S_{\bar{\alpha}}}|b_j|\right)^2\le|S_{\bar{\alpha}}|(b^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_{\bar{\alpha}}b). \tag{S5.20}
$$

**Proof.** In the event $\Omega_{11}$, we have $|b^{\mathrm{T}}(\tilde{\boldsymbol{\Sigma}}_{\bar{\alpha}}-\boldsymbol{\Sigma}_{\bar{\alpha}})b|\le\lambda_1\|b\|_1^2$. Then Assumption 2(v) implies that for any $b=(b_0,\ldots,b_q)^{\mathrm{T}}$ satisfying $\sum_{j\notin S_{\bar{\alpha}}}|b_j|\le\xi_1\sum_{j\in S_{\bar{\alpha}}}|b_j|$,

$$
\begin{aligned}
\nu_1^2\|b_{S_{\bar{\alpha}}}\|_1^2\le&|S_{\bar{\alpha}}|(b^{\mathrm{T}}\boldsymbol{\Sigma}_{\bar{\alpha}}b)\le|S_{\bar{\alpha}}|(b^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_{\bar{\alpha}}b+\lambda_1\|b\|_1^2)\\
\le&|S_{\bar{\alpha}}|(b^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_{\bar{\alpha}}b)+|S_{\bar{\alpha}}|\lambda_1(1+\xi_1)^2\|b_{S_{\bar{\alpha}}}\|_1^2
\end{aligned}
$$

$$\leq |S_{\bar{\alpha}}|(b^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_{\bar{\alpha}}b) + \zeta_1\nu_1^2\|b_{S_{\bar{\alpha}}}\|_1^2$$

where $\|b_{S_{\bar{\alpha}}}\|_1 = \sum_{j\in S_{\bar{\alpha}}}|b_j|$; and the last inequality holds due to Assumption 2(vi), $(1+\xi_1)^2\nu_1^{-2}|S_{\bar{\alpha}}|\lambda_1 \leq \zeta_1$. Thus (S5.20) follows by rearrangement. □

# S6 Proofs of Propositions 1 and 2

## S6.1 Proof of Proposition 1

**Proof.** Let $c_\gamma$ in Assupmtion 2(iv) $= \max[\sqrt{2}\{1+\exp(-B_0)\}C_0, 4\exp(-B_0)C_0^2]$. This can be shown similarly to Tan (2020a), Theorem 1 by Lemmas 6–8 and Lemmas similar to 10–17. The small difference in probability is due to extra constraints of $\lambda_0$ on $\Omega_{02}$ and $\Omega_{03}$ from the sequential estimate, which is also demonstrated in Tan (2020a), Theorem 5. □

## S6.2 Proof of Proposition 2

**Proof.** To facilitate the proof, we first define some constants. Let $\nu_2 = \nu_1(1-\zeta_1)^{1/2}$, $\xi_2 = 1-2A_1/\{(\xi_1+1)(A_1-1)\}$, $\xi_3 = (\xi_1+1)(A_1-1)$, $c_\alpha$ in Assumption 2(vi) equals to $\max\{C_4\sqrt{8(D_0^2+D_1^2)}, 4C_2C_4^2\}\exp(-B_0)$, $\tilde{c}_\alpha$ and $\tilde{c}_\gamma$ in Assumption 2(vii) equal to $C_3C_4\exp(\eta_{01})(A_1-1)^{-1}\xi_3^2\nu_2^{-2}$ and $C_1^{-1}C_3C_4(A_1-1)^{-1}\xi_2^{-2}\exp(3\eta_{01})M_1$, respectively.

Denote $b = \hat{\alpha}-\bar{\alpha}$, $D_{\mathrm{WL}}^{\dagger} = D_{\mathrm{WL}}^{\dagger}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}, \bar{\alpha}^{\mathrm{T}}\boldsymbol{G}; \bar{\gamma})$, $Q_{\mathrm{WL}} = Q_{\mathrm{WL}}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}, \bar{\alpha}^{\mathrm{T}}\boldsymbol{G}; \bar{\gamma})$ and $D_{\mathrm{WL}}^{\ddagger} = \exp(-\eta_{01})D_{\mathrm{WL}}^{\dagger} + (A_1-1)\lambda_1\|b\|_1$.

In the event $\Omega_{00}\cap\Omega_{01}\cap\Omega_{03}\cap\Omega_{10}\cap\Omega_{11}$, (S5.19) in Lemma 15 leads to two possible cases: either

$$\xi_2 D_{\mathrm{WL}}^{\ddagger} \leq \exp(\eta_{01})(M_1|S_{\bar{\gamma}}|\lambda_0^2 Q_{\mathrm{WL}})^{1/2}, \tag{S6.21}$$

or $(1-\xi_2)D_{\mathrm{WL}}^{\ddagger} \leq 2A_1\lambda_1\sum_{j\in S_{\bar{\alpha}}}|b_j|$, i.e.,

$$D_{\mathrm{WL}}^{\ddagger} \leq (\xi_1+1)(A_1-1)\lambda_1\sum_{j\in S_{\bar{\alpha}}}|b_j| = \xi_3\lambda_1\sum_{j\in S_{\bar{\alpha}}}|b_j|. \tag{S6.22}$$

By Lemma 16, we have

$$D_{\mathrm{WL}}^{\dagger} \geq \{b^{\mathrm{T}}\tilde{\mathbf{\Sigma}}_{\alpha}(\bar{\alpha})b\}\frac{1-\exp(-C_{40}\|b\|_1)}{C_{40}\|b\|_1} = (b^{\mathrm{T}}\tilde{\mathbf{\Sigma}}_{\bar{\alpha}}b)\frac{1-\exp(-C_{40}\|b\|_1)}{C_{40}\|b\|_1}. \tag{S6.23}$$

If (S6.21) holds, notice that $D_{\mathrm{WL}}^{\dagger} \leq \exp(\eta_{01})D_{\mathrm{WL}}^{\ddagger}$ and by Assumption 2(i), $Q_{\mathrm{WL}} < C_1^{-1}(b^{\mathrm{T}}\tilde{\mathbf{\Sigma}}_{\alpha}b)$, which together with (S6.23) yields

$$D_{\mathrm{WL}}^{\ddagger} \leq \exp(3\eta_{01})\xi_2^{-2}C_1^{-1}(M_1|S_{\bar{\gamma}}|\lambda_0^2)\frac{C_{40}\|b\|_1}{1-\exp(-C_{40}\|b\|_1)}. \tag{S6.24}$$

Since $(A_1-1)\lambda_1\|b\|_1 \leq D_{\mathrm{WL}}^{\ddagger}$ and Assumption 2(vii) holds, (S6.24) implies that

$$1-\exp(-C_{40}\|b\|_1) \leq (A_1-1)^{-1}\exp(3\eta_{01})\xi_2^{-2}C_1^{-1}(M_1|S_{\bar{\gamma}}|\lambda_0)C_{40} \leq \eta_3 < 1.$$

As a result, $C_{40}\|b\|_1 \leq -\ln(1-\eta_3)$, which leads to

$$\frac{1-\exp(-C_{40}\|b\|_1)}{C_{40}\|b\|_1} = \int_0^1 \exp(-C_{40}\|b\|_1 u)\mathrm{d}u \geq \exp(-C_{40}\|b\|_1) \geq 1-\eta_3. \tag{S6.25}$$

Combining the inequality with (S6.24), we obtain

$$D_{\mathrm{WL}}^{\ddagger} \leq \exp(3\eta_{01})\xi_2^{-2}\{C_1(1-\eta_3)\}^{-1}(M_1|S_{\bar{\gamma}}|\lambda_0^2)$$

.

If (S6.22) holds, then $\sum_{j\notin S_{\bar{\alpha}}}|b_j| \leq \xi_1 \sum_{j\in S_{\bar{\alpha}}}|b_j|$, which, together with Assumptions 2(v)–2(vi), implies (S5.20) in Lemma 17, that is,

$$\sum_{j\in S_{\bar{\alpha}}}|b_j| \leq (1-\zeta_1)^{-1/2}\nu_1^{-1}|S_{\bar{\alpha}}|^{1/2}\left(b^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_{\bar{\alpha}}b\right)^{1/2}. \tag{S6.26}$$

Since $D_{\mathrm{WL}}^{\dagger} \leq \exp(\eta_{01})D_{\mathrm{WL}}^{\ddagger}$, combining (S6.22), (S6.23) and (S6.26) yields

$$D_{\mathrm{WL}}^{\ddagger} \leq \exp(\eta_{01})\xi_3^2(1-\zeta_1)^{-1}\nu_1^{-2}|S_{\bar{\alpha}}|\lambda_1^2\frac{C_{40}\|b\|_1}{1-\exp(-C_{40}\|b\|_1)}. \tag{S6.27}$$

Since $(A_1-1)\lambda_1\|b\|_1 \leq D_{\mathrm{WL}}^{\ddagger}$ and Assumption 2(ii) holds, (S6.27) implies that

$$1-\exp(-C_{40}\|b\|_1) \leq (A_1-1)^{-1}\exp(\eta_{01})\xi_3^2(1-\zeta_1)^{-1}\nu_1^{-2}|S_{\bar{\alpha}}|\lambda_1 C_{40} \leq \eta_2 < 1.$$

As a result, $C_{40}\|b\|_1 \leq -\ln(1-\eta_2)$, which leads to

$$\frac{1-\exp(-C_{40}\|b\|_1)}{C_{40}\|b\|_1} = \int_0^1 \exp(-C_{40}\|b\|_1 u)\mathrm{d}u \geq \exp(-C_{40}\|b\|_1) \geq 1-\eta_2.$$

Combining the inequality with (S6.27), we obtain $D_{\mathrm{WL}}^{\ddagger} \leq \exp(\eta_{01})\xi_3^2\nu_2^{-2}(1-\eta_2)^{-2}|S_{\bar{\alpha}}|\lambda_1^2$.

Therefore, we obtain

$$\begin{aligned} & D^{\dagger}_{\mathrm{WL}}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}, \bar{\alpha}^{\mathrm{T}}\boldsymbol{G}, \bar{\gamma}) + \exp(\eta_{01})(A_1 - 1)\lambda_1\|\hat{\alpha} - \bar{\alpha}\|_1 \\ \leq\ & \exp(4\eta_{01})\xi_2^{-2}\{C_1(1-\eta_3)\}^{-1}(M_1|S_{\bar{\gamma}}|\lambda_0^2) + \exp(2\eta_{01})\xi_3^2\nu_2^{-2}(1-\eta_2)^{-2}|S_{\bar{\alpha}}|\lambda_1^2, \end{aligned}$$

Let $M_{11} = \exp(4\eta_{01})\xi_2^{-2}\{C_1(1-\eta_3)\}^{-1}M_1$ and $M_{12} = \exp(2\eta_{01})\xi_3^2\nu_2^{-2}(1-\eta_2)^{-2}$, then (4.2) holds.

□

# S7 Proof of Theorem 1

## S7.1 Lemmas for the proposed estimator

**Lemma 18.** Suppose that Assumptions 1(i), 2(i), 2(ii), and 2(vi) hold, if

$$\lambda_0 \geq \sqrt{2}\{1 + \exp(-B_0)\}C_0\sqrt{\frac{\ln(\frac{1+P}{\epsilon})}{N}},$$

then, $\mathbb{P}(\Omega_{20}) \geq 1 - 2\epsilon$ for any $r \geq 0$, where $\Omega_{20}$ denotes the event

$$\sup_{\|\alpha-\bar{\alpha}\|_1 \leq r; j=0,\ldots,d} \left| (\tilde{\mathbb{E}} - \mathbb{E}) \left[ \left\{ \frac{R}{\pi(\boldsymbol{X}; \bar{\gamma})} - 1 \right\} \{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\} Z_j \right] \right| \leq B_1\lambda_0 r,$$

where $B_1$ is a positive constant, depending on $(C_0, C_2, C_3, C_5)$.

**Proof.** This can be shown similarly to Lemma 13 in the Supplement of Tan (2020a).

□

**Lemma 19.** Suppose Assumptions 1, 2 and 3 hold, if a function $h(\cdot)$ on a set of samples satisfying $h[\{\boldsymbol{X}_i\}_{i=1}^N] \leq M\{|S_{\bar{\gamma}}|\lambda_0(\epsilon)+|S_{\bar{\alpha}}|\lambda_1(\epsilon)\}$ for some constant $M$ with probability $1-c\epsilon$ for some constant $c>0$ and any $\epsilon>0$, then, $h[\{\boldsymbol{X}_i\}_{i=1}^N] = o_p(1)$.

**Proof.** For $\forall \epsilon > 0$, let $\Omega_\epsilon$ be the event where $h[\{\boldsymbol{X}_i\}_{i=1}^N] \leq M\{|S_{\bar{\gamma}}|\lambda_0(\epsilon/c) + |S_{\bar{\alpha}}|\lambda_1(\epsilon/c)\}$ holds. Suppose $|S_{\bar{\gamma}}|\lambda_0(\epsilon/c) < |S_{\bar{\alpha}}|\lambda_1(\epsilon/c)$, then, on $\Omega_\epsilon$,

$$\frac{h[\{\boldsymbol{X}_i\}_{i=1}^N]\sqrt{N}}{|S_{\bar{\alpha}}|\sqrt{\ln\{e(q+1)\}}} \leq 2M\sqrt{\frac{\ln(\frac{q+1}{\epsilon/c})}{\ln\{e(q+1)\}}} \leq 2M\sqrt{\frac{\ln(q+1)}{\ln(q+1)}+\ln\left(\frac{c}{\epsilon}\right)} \leq 2M\sqrt{1+\ln\left(\frac{c}{\epsilon}\right)}, \tag{S7.28}$$

which implies that

$$\mathbb{P}\left[\frac{h[\{\boldsymbol{X}_i\}_{i=1}^N]\sqrt{N}}{|S_{\bar{\alpha}}|\sqrt{\ln\{e(q+1)\}}} \leq 2M\sqrt{1+\ln\left(\frac{c}{\epsilon}\right)}\right] \geq \mathbb{P}(\Omega_\epsilon) = 1-\epsilon.$$

Similarly, suppose $|S_{\bar{\alpha}}|\lambda_1 < |S_{\bar{\gamma}}|\lambda_0$, then, on $\Omega_\epsilon$,

$$\mathbb{P}\left[\frac{h[\{\boldsymbol{X}_i\}_{i=1}^N]\sqrt{N}}{|S_{\bar{\gamma}}|\sqrt{\ln\{e(p+1)\}}} \leq 2M\sqrt{1+\ln\left(\frac{c}{\epsilon}\right)}\right] \geq \mathbb{P}(\Omega_\epsilon) = 1-\epsilon.$$

For $\forall \epsilon > 0$, we have

$$\mathbb{P}\left[\frac{h[\{\boldsymbol{X}_i\}_{i=1}^N]\sqrt{N}}{|S_{\bar{\alpha}}|\sqrt{\ln\{e(q+1)\}}+|S_{\bar{\gamma}}|\sqrt{\ln\{e(p+1)\}}} \leq 2M\sqrt{1+\ln\left(\frac{c}{\epsilon}\right)}\right] \geq 1-\epsilon,$$

We write $\lambda_0$, $\lambda_1$ as $\lambda_0(\epsilon)$, $\lambda_1(\epsilon)$, since we treat them as functions of $\epsilon$.

thus, $\frac{h[\{\boldsymbol{X}_i\}_{i=1}^N]\sqrt{N}}{|S_{\bar{\alpha}}|\sqrt{\ln\{e(q+1)\}}+|S_{\bar{\gamma}}|\sqrt{\ln\{e(p+1)\}}} = O_p(1)$, and by Assumption 3(vi), it follows that

$$h[\{\boldsymbol{X}_i\}_{i=1}^N] = o_p\left[\frac{|S_{\bar{\alpha}}|\sqrt{\ln\{e(q+1)\}}+|S_{\bar{\gamma}}|\sqrt{\ln\{e(p+1)\}}}{\sqrt{N}}\right] = o_p(1).$$

□

**Lemma 20.** Suppose Assumptions 1, 2 and 3 hold, if a function $h(\cdot)$ on a set of samples satisfying $h[\{\boldsymbol{X}_i\}_{i=1}^N] \leq M(|S_{\bar{\gamma}}|+|S_{\bar{\alpha}}|)\lambda_0(\epsilon)\lambda_1(\epsilon)$ for some constant $M$ with probability $1-c\epsilon$ for some constant $c$ and any $\epsilon > 0$, then, $h[\{\boldsymbol{X}_i\}_{i=1}^N] = o_p(1/\sqrt{N})$.

**Proof.** By similar trick used in the proof of Lemma 19, it can be easily shown that $h[\{\boldsymbol{X}_i\}_{i=1}^N] = o_p\left[\frac{(|S_{\bar{\gamma}}|+|S_{\bar{\alpha}}|)\sqrt{\ln\{e(q+1)\}}\sqrt{\ln\{e(p+1)\}}}{N}\right] = o_p(1/\sqrt{N})$ by Assumption 3(vi). □

**Lemma 21.** Under Assumptions 1, 2 and 3, suppose either the PS model (3.1) is correct or the OR model (3.2) is correct, the AIPW estimator $\hat{\beta} \xrightarrow{\mathbb{P}} \beta^*$.

**Proof.** First, we notice that when $\pi(\boldsymbol{X};\gamma)$ is correct or $\phi(\boldsymbol{X};\alpha) = \psi(\alpha^{\mathrm{T}}\boldsymbol{G})$ is correct, $\beta^*$ is the unique solution to $\mathbb{E}\{\tau(\boldsymbol{O},\bar{\alpha},\beta,\bar{\gamma})\} = 0$. If $\pi(\boldsymbol{X};\gamma)$ is correct, we obtain,

$$\begin{aligned}
&\mathbb{E}\{\tau(\boldsymbol{O},\bar{\alpha},\beta^*,\bar{\gamma})\}\\
&= \mathbb{E}\left[\frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\{Y-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z} + \left\{1-\frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\right\}\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z}\right]\\
&= \mathbb{E}_{\boldsymbol{X}}\left(\mathbb{E}\left[\frac{R\{Y-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}}{\pi(\boldsymbol{X};\bar{\gamma})} + \left\{1-\frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\right\}\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\right]\boldsymbol{Z}\middle|\boldsymbol{X}\right)
\end{aligned}$$

$$= \mathbb{E}\left[\{Y - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z}\right]$$

$$= 0.$$

If $\psi(\alpha^{\mathrm{T}}\boldsymbol{G})$ is correct, we obtain,

$$\begin{aligned}
&\mathbb{E}\{\tau(\boldsymbol{O}, \bar{\alpha}, \beta^*, \bar{\gamma})\} \\
&= \mathbb{E}\left[\frac{R}{\pi(\boldsymbol{X}; \bar{\gamma})}\{Y - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z} + \left\{1 - \frac{R}{\pi(\boldsymbol{X}; \bar{\gamma})}\right\}\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\,\boldsymbol{Z}\right] \\
&= \mathbb{E}\left[\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z}\right] \\
&= \mathbb{E}\left[\{Y - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z}\right] \\
&= 0.
\end{aligned}$$

The uniqueness is determined by the uniqueness of $\beta^*$.

Since Assumptions 3(ii) and 3(iii) hold, by standard argument of consistency (e.g. Van der Vaart (2000)), it suffices to show that

$$\tilde{\mathbb{E}}\{\tau(\boldsymbol{O}, \bar{\alpha}, \hat{\beta}, \bar{\gamma})\} = o_p(1). \tag{S7.29}$$

We consider the following decomposition,

$$\begin{aligned}
&\tilde{\mathbb{E}}\{\tau(\boldsymbol{O}, \bar{\alpha}, \hat{\beta}, \bar{\gamma})\} = \tilde{\mathbb{E}}\{\tau(\boldsymbol{O}, \bar{\alpha}, \hat{\beta}, \bar{\gamma})\} - \tilde{\mathbb{E}}\{\tau(\boldsymbol{O}, \hat{\alpha}, \hat{\beta}, \hat{\gamma})\} \\
=&\tilde{\mathbb{E}}\{\tau(\boldsymbol{O}, \bar{\alpha}, \hat{\beta}, \hat{\gamma})\} - \tilde{\mathbb{E}}\{\tau(\boldsymbol{O}, \hat{\alpha}, \hat{\beta}, \hat{\gamma})\} + \tilde{\mathbb{E}}\{\tau(\boldsymbol{O}, \bar{\alpha}, \hat{\beta}, \bar{\gamma})\} - \tilde{\mathbb{E}}\{\tau(\boldsymbol{O}, \bar{\alpha}, \hat{\beta}, \hat{\gamma})\}
\end{aligned}$$

$$=\Delta_0^1+\Delta_1^1,$$

where

$$\Delta_0^1=\tilde{\mathbb{E}}\{\tau(\boldsymbol{O},\bar{\alpha},\hat{\beta},\hat{\gamma})\}-\tilde{\mathbb{E}}\{\tau(\boldsymbol{O},\hat{\alpha},\hat{\beta},\hat{\gamma})\}=\tilde{\mathbb{E}}\left[\left\{\frac{R}{\pi(\boldsymbol{X};\hat{\gamma})}-1\right\}\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\boldsymbol{Z}\right],$$
$$\Delta_1^1=\tilde{\mathbb{E}}\{\tau(\boldsymbol{O},\bar{\alpha},\hat{\beta},\bar{\gamma})\}-\tilde{\mathbb{E}}\{\tau(\boldsymbol{O},\bar{\alpha},\hat{\beta},\hat{\gamma})\}=\tilde{\mathbb{E}}\left[R\left\{\frac{1}{\pi(\boldsymbol{X};\bar{\gamma})}-\frac{1}{\pi(\boldsymbol{X};\hat{\gamma})}\right\}\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\boldsymbol{Z}\right].$$

By Assumptions 1(iv) and 2(vi), we know that $\exists\ \zeta_{10}<0$ a constant, such that $|S_{\bar{\gamma}}|\lambda_0+|S_{\bar{\alpha}}|\lambda_1\leq\zeta_{10}$. By (4.2) in Proposition 2, we know that $\exists\ M_0^0>0$ a constant, such that in the event $\Omega_{00}\cap\Omega_{01}\cap\Omega_{03}\cap\Omega_{10}\cap\Omega_{11}$, $\|\hat{\alpha}-\bar{\alpha}\|_1\leq M_0^0(|S_{\bar{\gamma}}|\lambda_0+|S_{\bar{\alpha}}|\lambda_1)$. Consider the $j$-th coordinate of $\Delta_0^1$,

$$\begin{aligned}
|\Delta_{0,j}^1|=&\left|\tilde{\mathbb{E}}\left[\left\{\frac{R}{\pi(\boldsymbol{X};\hat{\gamma})}-1\right\}\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}Z_j\right]\right|\\
\leq&\tilde{\mathbb{E}}^{1/2}\left[\left\{\frac{R}{\pi(\boldsymbol{X};\hat{\gamma})}-1\right\}^2Z_j^2\right]\times\tilde{\mathbb{E}}^{1/2}\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2\\
\leq&\tilde{\mathbb{E}}^{1/2}\left(\left[R-1+Rw(\boldsymbol{X};\bar{\gamma})\exp\{-(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}\right]^2Z_j^2\right)\\
&\times\sqrt{2}\tilde{\mathbb{E}}^{1/2}\left[\{\psi(\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\psi_1(\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G})(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\boldsymbol{G}\right]\\
\leq&\sqrt{2}C_5\tilde{\mathbb{E}}^{1/2}\left[(R-1)^2+Rw^2(\boldsymbol{X};\bar{\gamma})\exp\{-2(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}\right]\\
&\times\tilde{\mathbb{E}}^{1/2}\left[\int_0^1\psi_1\{\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}+u(\tilde{\alpha}-\bar{\alpha})^{\mathrm{T}}\boldsymbol{G}\}du\psi_1(\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G})\{(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\boldsymbol{G}\}^2\right]\\
\leq&\sqrt{2}C_5\{1+\exp(-2B_0+2C_0\|\hat{\gamma}-\bar{\gamma}\|_1)\}^{\frac{1}{2}}\\
&\times\tilde{\mathbb{E}}^{1/2}[\psi_1^2(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\exp(2C_3|\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G}-\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}|)\{(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\boldsymbol{G}\}^2]
\end{aligned}$$

$$
\begin{aligned}
&\leq 2C_5\{1+\exp(-2B_0+2\eta_{01})\}^{\frac{1}{2}} \times C_2C_4\exp(C_3C_4\|\hat{\alpha}-\bar{\alpha}\|_1)\|\hat{\alpha}-\bar{\alpha}\|_1\\
&\leq \sqrt{2}C_2C_4C_5\{1+\exp(-2B_0+2\eta_{01})\}^{\frac{1}{2}} \times \exp(C_3C_4M_0^0\zeta_{10})\|\hat{\alpha}-\bar{\alpha}\|_1\\
&= M_0^1(|S_{\bar{\gamma}}|\lambda_0+|S_{\bar{\alpha}}|\lambda_1)\\
&= o_p(1)
\end{aligned}
$$

where $M_0^1$ is a constant, $\tilde{\alpha}=u\hat{\alpha}+(1-u)\bar{\alpha}$ for some constant $u\in(0,1)$ and the last equality holds by Lemma 19.

We consider the $j$-th coordinate of $\Delta_1^1$. By Cauchy–Schwarz inequality, Lemma 8, and Lemma 5, since Assumption 2(iv) holds, we obtain

$$
\begin{aligned}
|\Delta_{1,j}^1| &= \left|\tilde{\mathbb{E}}\left[R\left\{\frac{1}{\pi(\boldsymbol{X};\bar{\gamma})}-\frac{1}{\pi(\boldsymbol{X};\hat{\gamma})}\right\}\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}Z_j\right]\right|\\
&= \left|\tilde{\mathbb{E}}\left[Rw(\boldsymbol{X};\bar{\gamma})\exp\{-u(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}(\bar{\gamma}-\hat{\gamma})^{\mathrm{T}}\boldsymbol{F}\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}Z_j\right]\right|\\
&\leq \tilde{\mathbb{E}}^{1/2}\left[Rw(\boldsymbol{X};\bar{\gamma})\exp\{-2u(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}\{(\bar{\gamma}-\hat{\gamma})^{\mathrm{T}}\boldsymbol{F}\}^2Z_j^2\right]\\
&\quad\times\tilde{\mathbb{E}}^{1/2}[Rw(\boldsymbol{X};\bar{\gamma})\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2]\\
&\leq \exp(-B_0/2+C_0\|\hat{\gamma}-\bar{\gamma}\|_1)C_0C_5\|\hat{\gamma}-\bar{\gamma}\|_1\\
&\quad\times\sqrt{\mathbb{E}[Rw(\boldsymbol{X};\bar{\gamma})\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2]+(D_0^2+D_0D_1)\lambda_0}\\
&\leq \exp(-B_0/2+\eta_{01})C_0C_5\sqrt{(D_0^2+D_1^2)\{\lambda_0+\exp(-B_0)\}}\|\hat{\gamma}-\bar{\gamma}\|_1\\
&\leq M_1^0(|S_{\bar{\gamma}}|\lambda_0+|S_{\bar{\alpha}}|\lambda_1)\\
&= o_p(1),
\end{aligned}
$$

where $u \in (0,1)$ and $M_1^0$ are both constants. Therefore, $\|\tilde{\mathbb{E}}\{\tau(\boldsymbol{O}, \bar{\alpha}, \hat{\beta}, \bar{\gamma})\}\|_\infty \leq \|\Delta_0^1\|_\infty + \|\Delta_1^1\|_\infty = o_p(1)$. Hence, $\hat{\beta} \xrightarrow{\mathbb{P}} \beta^*$. □

## S7.2 Proof of Theorem 1(i)

We show the asymptotic normality of $\sqrt{N}(\hat{\beta} - \beta^*)$. First, we consider the following decomposition,

$$
\begin{aligned}
&\tau(\boldsymbol{O}, \hat{\alpha}, \hat{\beta}, \hat{\gamma}) - \tau(\boldsymbol{O}, \bar{\alpha}, \beta^*, \bar{\gamma}) \\
=&\left[\frac{R}{\pi(\boldsymbol{X}; \hat{\gamma})}\{Y - \psi(\hat{\beta}^{\mathrm{T}}\boldsymbol{Z})\} - \left\{\frac{R}{\pi(\boldsymbol{X}; \hat{\gamma})} - 1\right\}\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\hat{\beta}^{\mathrm{T}}\boldsymbol{Z})\}\right]\boldsymbol{Z} \\
&- \left[\frac{R}{\pi(\boldsymbol{X}; \bar{\gamma})}\{Y - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\} - \left\{\frac{R}{\pi(\boldsymbol{X}; \bar{\gamma})} - 1\right\}\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\right]\boldsymbol{Z} \\
=&\{\psi(\beta^{*\mathrm{T}}\boldsymbol{Z}) - \psi(\hat{\beta}^{\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z} + R\left\{\frac{1}{\pi(\boldsymbol{X}; \hat{\gamma})} - \frac{1}{\pi(\boldsymbol{X}; \bar{\gamma})}\right\}\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\boldsymbol{Z} \\
&+ \left\{1 - \frac{R}{\pi(\boldsymbol{X}; \bar{\gamma})}\right\}\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\boldsymbol{Z} \\
&+ \left\{\frac{R}{\pi(\boldsymbol{X}; \bar{\gamma})} - \frac{R}{\pi(\boldsymbol{X}; \hat{\gamma})}\right\}\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\boldsymbol{Z},
\end{aligned}
$$

denoted as $\delta_0^0 + \delta_1^0 + \delta_2^0 + \delta_3^0$. Then, $-\tilde{\mathbb{E}}\tau(\boldsymbol{O}, \bar{\alpha}, \beta^*, \bar{\gamma}) = \Delta_0^2 + \Delta_1^2 + \Delta_2^2 + \Delta_3^2$, with $\Delta_i^2 = \tilde{\mathbb{E}}(\delta_i^0)$, $i = 0, 1, 2, 3$. First, we show that $\Delta_1^2 + \Delta_2^2 + \Delta_3^2 = o_p(1/\sqrt{N})$. To upper-bound $\Delta_1^2$, consider $\Delta_{1,j}^2$, the $j$-th coordinate of $\Delta_1^2$. By Taylor expansion in a neighborhood of $\bar{\gamma}$,

$$
\begin{aligned}
\Delta_{1,j}^2 =&\tilde{\mathbb{E}}\left[\left\{\frac{R}{\pi(\boldsymbol{X}; \hat{\gamma})} - \frac{R}{\pi(\boldsymbol{X}; \bar{\gamma})}\right\}\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}Z_j\right] \\
=&-(\hat{\gamma} - \bar{\gamma})^{\mathrm{T}}\tilde{\mathbb{E}}\left[\boldsymbol{F}Rw(\boldsymbol{X}; \bar{\gamma})\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}Z_j\right]
\end{aligned}
$$

$$+\frac{1}{2}(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\tilde{\mathbb{E}}\left[\boldsymbol{F}Rw(\boldsymbol{X},\tilde{\gamma}_j)\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}Z_j\boldsymbol{F}^{\mathrm{T}}\right](\hat{\gamma}-\bar{\gamma}),$$

denoted as $\Delta_{10,j}^2+\Delta_{11,j}^2$ where $\tilde{\gamma}_j=u_j\hat{\gamma}+(1-u_j)\bar{\gamma}$ for some $u_j\in(0,1)$.

In the event $\Omega_{00}\cap\Omega_{01}\cap\Omega_{10}$, by (4.1) and Lemma 10, we obtain

$$|\Delta_{10,j}^2|\le\|\hat{\gamma}-\bar{\gamma}\|_1\|\tilde{\mathbb{E}}\left[\boldsymbol{F}Rw(\boldsymbol{X};\bar{\gamma})\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}Z_j\right]\|_\infty\le M_{10,j}^1|S_{\bar{\gamma}}|\lambda_0\lambda_1,$$

for some constant $M_{10,j}^1>0$. In the event $\Omega_{00}\cap\Omega_{01}$, by (S5.9), we obtain

$$\begin{aligned}|\Delta_{11,j}^2|=&\frac{1}{2}|(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\tilde{\mathbb{E}}\left[\boldsymbol{F}Rw(\boldsymbol{X};\tilde{\gamma}_j)\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}Z_j\boldsymbol{F}^{\mathrm{T}}\right](\hat{\gamma}-\bar{\gamma})|\\ \le&\frac{1}{2}C_5\exp(\|\hat{\gamma}-\bar{\gamma}\|_1C_0)\tilde{\mathbb{E}}\left\{Rw(\boldsymbol{X};\bar{\gamma})|Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})||(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}|^2\right\}.\end{aligned}$$

We bound $\tilde{\mathbb{E}}\left\{Rw(\boldsymbol{X};\bar{\gamma})|Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})||(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}|^2\right\}$ by following steps. First, by Lemma 7, in the event $\Omega_{02}$,

$$(\tilde{\mathbb{E}}-\mathbb{E})\left[Rw(\boldsymbol{X};\bar{\gamma})|Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})|\{(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}^2\right]\le\sqrt{D_0^2+D_1^2}\|\hat{\gamma}-\bar{\gamma}\|_1^2\lambda_0.$$

Second, by Assumption 2(iv) and Lemma 5, we have

$$\mathbb{E}[\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2|\boldsymbol{X}]\le D_0^2+D_1^2.$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[Rw(\boldsymbol{X};\bar{\gamma})|Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})|\{(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}^2\right]\\
\leq&\mathbb{E}^{1/2}\left[Rw(\boldsymbol{X};\bar{\gamma})|Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})|^2\{(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}^2\right]\times\mathbb{E}^{1/2}\left[Rw(\boldsymbol{X};\bar{\gamma})\{(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}^2\right]\\
\leq&\sqrt{D_0^2+D_1^2}\mathbb{E}\left[Rw(\boldsymbol{X};\bar{\gamma})\{(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}^2\right].
\end{aligned}$$

Third, in the event $\Omega_{01}$, by (S5.6),

$$(\mathbb{E}-\tilde{\mathbb{E}})\left[Rw(\boldsymbol{X};\bar{\gamma})\{(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}^2\right]\leq\lambda_0\|\bar{\gamma}-\hat{\gamma}\|_1^2.$$

Combining preceding inequalities and (S5.9) in Lemma 9, in the event $\Omega_{00}\cap\Omega_{01}\cap\Omega_{03}\cap\Omega_{02}$,

$$\begin{aligned}
&\tilde{\mathbb{E}}\left\{Rw(\boldsymbol{X};\bar{\gamma})|Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})||(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}|^2\right\}\\
\leq&\sqrt{D_0^2+D_1^2}\|\hat{\gamma}-\bar{\gamma}\|_1^2\lambda_0+\sqrt{D_0^2+D_1^2}\{\lambda_0\|\bar{\gamma}-\hat{\gamma}\|_1^2+\exp(\eta_{01})M_0|S_{\bar{\gamma}}|\lambda_0^2\}\\
\leq&2\sqrt{D_0^2+D_1^2}\left(\frac{M_0}{A_0-1}\right)^2|S_{\bar{\gamma}}|^2\lambda_0^3+\sqrt{D_0^2+D_1^2}\exp(\eta_{01})M_0|S_{\bar{\gamma}}|\lambda_0^2\\
\leq&2\zeta_0\sqrt{D_0^2+D_1^2}\left(\frac{M_0}{A_0-1}\right)^2|S_{\bar{\gamma}}|\lambda_0^2+\sqrt{D_0^2+D_1^2}\exp(\eta_{01})M_0|S_{\bar{\gamma}}|\lambda_0^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
|\Delta_{11,j}^2|\leq&\frac{1}{2}C_5\exp(\|\hat{\gamma}-\bar{\gamma}\|_1C_0)\tilde{\mathbb{E}}\left\{Rw(\boldsymbol{X};\bar{\gamma})|Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})||(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}|^2\right\}\\
\leq&\frac{1}{2}C_5\exp(\eta_{01})\times\left\{2\zeta_0\sqrt{D_0^2+D_1^2}\left(\frac{M_0}{A_0-1}\right)^2|S_{\bar{\gamma}}|\lambda_0^2+\sqrt{D_0^2+D_1^2}\exp(\eta_{01})M_0|S_{\bar{\gamma}}|\lambda_0^2\right\}
\end{aligned}$$

$$=M_{11}^1|S_{\bar\gamma}|\lambda_0^2,$$

for some constant $M_{11}^1 > 0$. Hence, in the event $\Omega_{00} \cap \Omega_{01} \cap \Omega_{03} \cap \Omega_{02} \cap \Omega_{10}$,

$$\begin{aligned}
&\sup_{j=0,\ldots,m-1} |\Delta_{1,j}^2| \\
\leq &\sup_{j=0,\ldots,m-1} (|\Delta_{10,j}^2| + |\Delta_{11,j}^2|) \\
\leq &\sup_{j=0,\ldots,m-1} M_{10}^1|S_{\bar\gamma}|\lambda_0\lambda_1 + \sup_{j=0,\ldots,m-1} M_{11}^1|S_{\bar\gamma}|\lambda_0^2 \\
= &M_{10}^1|S_{\bar\gamma}|\lambda_0\lambda_1 + M_{11}^1|S_{\bar\gamma}|\lambda_0^2 \\
\leq &M_1^1(|S_{\bar\gamma}|\lambda_0\lambda_1 + |S_{\bar\gamma}|\lambda_0^2),
\end{aligned}$$

where $M_1^1 = \max(M_{10}^1, M_{11}^1)$.

To bound $\Delta_2^2$, consider $\Delta_{2,j}^2$, the $j$-th coordinate of $\Delta_2^2$, $\Delta_{2,j}^2$ can be decomposed as

$$\begin{aligned}
\Delta_{2,j}^2 = &(\tilde{\mathbb{E}} - \mathbb{E}) \left[ \left\{ 1 - \frac{R}{\pi(\boldsymbol{X}; \bar\gamma)} \right\} \{\psi(\hat\alpha^{\mathrm{T}}\boldsymbol{G}) - \psi(\bar\alpha^{\mathrm{T}}\boldsymbol{G})\} Z_j \right] \\
&+ \mathbb{E} \left[ \left\{ 1 - \frac{R}{\pi(\boldsymbol{X}; \bar\gamma)} \right\} \{\psi(\hat\alpha^{\mathrm{T}}\boldsymbol{G}) - \psi(\bar\alpha^{\mathrm{T}}\boldsymbol{G})\} Z_j \right],
\end{aligned}$$

denoted as $\Delta_{20,j}^2 + \Delta_{21,j}^2$. Since $\pi(\boldsymbol{X}; \bar\gamma)$ is correct, $\Delta_{21,j}^2 = 0$. By (4.2), $\exists M_\alpha > 0$, such that $\|\hat\alpha - \bar\alpha\|_1 \leq M_\alpha(|S_{\bar\gamma}|\lambda_0 + |S_{\bar\alpha}|\lambda_1)$. Take $r_\alpha = M_\alpha(|S_{\bar\gamma}|\lambda_0 + |S_{\bar\alpha}|\lambda_1)$ in Lemma 18, then in the event $\Omega_{00} \cap \Omega_{01} \cap \Omega_{03} \cap \Omega_{10} \cap \Omega_{11} \cap \Omega_{20}$, we have $\|\hat\alpha - \bar\alpha\|_1 \leq r_\alpha$ and

Note that $\lambda_0 \leq \lambda_1$.

hence

$$|\Delta_{20,j}^2| \leq B_1 M_\alpha (|S_{\bar{\gamma}}|\lambda_0 + |S_{\bar{\alpha}}|\lambda_1)\lambda_0. \tag{S7.30}$$

Thus, in the event $\Omega_{00} \cap \Omega_{01} \cap \Omega_{03} \cap \Omega_{10} \cap \Omega_{11} \cap \Omega_{20}$,

$$\sup_{j=0,\ldots,m-1} |\Delta_{2,j}^2| \leq B_1 M_\alpha (|S_{\bar{\gamma}}|\lambda_0 + |S_{\bar{\alpha}}|\lambda_1)\lambda_0 = M_2^1 (|S_{\bar{\gamma}}|\lambda_0 + |S_{\bar{\alpha}}|\lambda_1)\lambda_0, \tag{S7.31}$$

for some positive constant $M_2^1$.

To deal with $\Delta_3^2$, first, by mean value theorem, we obtain for some $u \in (0,1)$,

$$\begin{aligned} \frac{1}{\pi(\boldsymbol{X};\hat{\gamma})} - \frac{1}{\pi(\boldsymbol{X};\bar{\gamma})} =& -\exp\{-u\hat{\gamma}^{\mathrm{T}}\boldsymbol{F} - (1-u)\bar{\gamma}^{\mathrm{T}}\boldsymbol{F}\}(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F} \\ =& -w(\boldsymbol{X};\bar{\gamma})\exp\{-u(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F} \end{aligned} \tag{S7.32}$$

and for some $\tilde{\alpha}$ lies between $\bar{\alpha}$ and $\hat{\alpha}$,

$$\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}) = -\psi_1(\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G})(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\boldsymbol{G}. \tag{S7.33}$$

Combining (S7.32) and (S7.33) and applying Cauchy-Schwartz inequality to $j$-th coordinate of $\Delta_3^2$ in the event $\Omega_{00} \cap \Omega_{01} \cap \Omega_{03} \cap \Omega_{10} \cap \Omega_{11} \cap \Omega_{20}$, we get

$$\begin{aligned} |\Delta_{3,j}^2| =& \left|\tilde{\mathbb{E}}\left[\left\{\frac{R}{\pi(\boldsymbol{X};\hat{\gamma})} - \frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\right\}\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})\}Z_j\right]\right| \\ =& |\tilde{\mathbb{E}}([Rw(\boldsymbol{X};\bar{\gamma})\exp\{-u_j(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}]\{\psi_1(\tilde{\alpha}_j^{\mathrm{T}}\boldsymbol{G})(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\boldsymbol{G}\})Z_j| \\ \leq& C_5\exp(\eta_{01})\tilde{\mathbb{E}}^{1/2}|Rw(\boldsymbol{X};\bar{\gamma})\{(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}^2| \times \tilde{\mathbb{E}}^{1/2}|Rw(\boldsymbol{X};\bar{\gamma})\psi_1^2(\tilde{\alpha}_j^{\mathrm{T}}\boldsymbol{G})\{(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\boldsymbol{G}\}^2| \end{aligned}$$

$$
\begin{aligned}
\leq & C_5 \exp(\eta_{01})\{\exp(\eta_{01}) M_0 |S_{\bar{\gamma}}| \lambda_0^2\}^{1/2} \\
& \times \tilde{\mathbb{E}}^{1/2} \left( Rw(\boldsymbol{X}; \bar{\gamma}) [\psi_1(\bar{\alpha}_j^{\mathrm{T}} \boldsymbol{G}) \exp\{C_3 |(\tilde{\alpha} - \bar{\alpha})^{\mathrm{T}} \boldsymbol{G}|\}]^2 \{(\hat{\alpha} - \bar{\alpha})^{\mathrm{T}} \boldsymbol{G}\}^2 \right) \\
\leq & C_5 \exp(\eta_{01})\{\exp(\eta_{01}) M_0 |S_{\bar{\gamma}}| \lambda_0^2\}^{1/2} C_2^{\frac{1}{2}} \exp(C_3 C_4 r_\alpha) \left\{ \frac{M^\dagger (|S_{\bar{\gamma}}| \lambda_0^2 + |S_{\bar{\alpha}}| \lambda_1^2)}{1 - \eta_3} \right\}^{\frac{1}{2}} \\
\leq & M_3^1 \{ |S_{\bar{\gamma}}| \lambda_0^2 + (|S_{\bar{\gamma}}| |S_{\bar{\alpha}}|)^{1/2} \lambda_0 \lambda_1 \}.
\end{aligned}
$$

The second inequality holds due to (S5.9) in Lemma 9. The third inequality holds by Assumption 2(ii), (S6.25), the facts that $\tilde{\mathbb{E}} \left[ Rw(\boldsymbol{X}; \bar{\gamma}) \psi_1(\bar{\alpha}_j^{\mathrm{T}} \boldsymbol{G}) \{(\hat{\alpha} - \bar{\alpha})^{\mathrm{T}} \boldsymbol{G}\}^2 \right] = b^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}_\alpha b$ in (S6.23) and that by (4.2) in Proposition 2, $\exists M^\dagger > 0$ a constant, such that $D_{\mathrm{WL}}^\dagger \leq M^\dagger (|S_{\bar{\gamma}}| \lambda_0^2 + |S_{\bar{\alpha}}| \lambda_1^2)$. Therefore,

$$
\sup_{j=0,\ldots,m-1} |\Delta_{3,j}^2| \leq M_3^1 \{ |S_{\bar{\gamma}}| \lambda_0^2 + (|S_{\bar{\gamma}}| |S_{\bar{\alpha}}|)^{1/2} \lambda_0 \lambda_1 \}, \tag{S7.34}
$$

for some constant $M_3^1$. Thus, on the event $\Omega_{00} \cap \Omega_{01} \cap \Omega_{03} \cap \Omega_{02} \cap \Omega_{10} \cap \Omega_{11} \cap \Omega_{20}$,

$$
\begin{aligned}
\|\Delta_1^2 + \Delta_2^2 + \Delta_3^2\|_\infty \leq & \sup_{j=0,\ldots,m-1} |\Delta_{1,j}^2| + \sup_{j=0,\ldots,m-1} |\Delta_{2,j}^2| + \sup_{j=0,\ldots,m-1} |\Delta_{3,j}^2| \\
\leq & M_1^1 (|S_{\bar{\gamma}}| \lambda_0 \lambda_1 + |S_{\bar{\gamma}}| \lambda_0^2) + M_2^1 (|S_{\bar{\gamma}}| \lambda_0 + |S_{\bar{\alpha}}| \lambda_1) \lambda_0 \\
& + M_3^1 \{ |S_{\bar{\gamma}}| \lambda_0^2 + (|S_{\bar{\gamma}}| |S_{\bar{\alpha}}|)^{1/2} \lambda_0 \lambda_1 \} \\
\leq & M^1 (|S_{\bar{\gamma}}| + |S_{\bar{\alpha}}|) \lambda_0 \lambda_1.
\end{aligned}
$$

By Lemma 20, $\|\Delta_1^2 + \Delta_2^2 + \Delta_3^2\|_\infty = o_p(1/\sqrt{N})$.

Then, we deal with $\Delta_0^2$. For the $j$-th coordinate of $\Delta_0^2$, $\Delta_{0,j}^2$, by mean value

theorem,

$$\Delta_{0,j}^2 = \tilde{\mathbb{E}}[\{\psi(\beta^{*\mathrm{T}}\boldsymbol{Z}) - \psi(\hat{\beta}^{\mathrm{T}}\boldsymbol{Z})\}Z_j] = -\tilde{\mathbb{E}}\{\psi_1(\tilde{\beta}_j^T\boldsymbol{Z})Z_j\boldsymbol{Z}^{\mathrm{T}}(\hat{\beta} - \beta^*)\},$$

where $\tilde{\beta}_j = (1-u_j)\beta^* + u_j\hat{\beta}$ for some $u_j \in (0,1)$. We first show that $\tilde{\mathbb{E}}\{\psi_1(\beta^{*\mathrm{T}}\boldsymbol{Z})Z_j\boldsymbol{Z}\} - \tilde{\mathbb{E}}\{\psi_1(\tilde{\beta}^T\boldsymbol{Z})Z_j\boldsymbol{Z}\} \xrightarrow{\mathbb{P}} 0$. By Assumption 2(ii), we know that for $\forall u, u'$, if $\psi_1(u) > \psi_1(u')$, since $\psi_1(u') \geq \psi_1(u)\exp(-C_3|u-u'|)$, then $|\psi_1(u) - \psi_1(u')| \leq \psi_1(u)\{1 - \exp(-C_3|u-u'|)\}$; if $\psi_1(u) < \psi_1(u')$, $\psi_1(u') \leq \psi_1(u)\exp(C_3|u-u'|)$, $|\psi_1(u) - \psi_1(u')| \leq \psi_1(u)\{\exp(C_3|u-u'|) - 1\}$; therefore, $|\psi_1(u) - \psi_1(u')| \leq \psi_1(u)\max\{1 - \exp(-C_3|u-u'|), \exp(C_3|u-u'|) - 1\}$. Consider the $i$-th element of the difference, if $C_3 = 0$, $\psi_1$ is a constant, then

$$|\tilde{\mathbb{E}}\{\psi_1(\beta^{*\mathrm{T}}X)Z_iZ_j\} - \tilde{\mathbb{E}}\{\psi_1(\tilde{\beta}_j^{\mathrm{T}}\boldsymbol{Z})Z_iZ_j\}| = 0.$$

Otherwise,

$$\begin{aligned}
&|\tilde{\mathbb{E}}\{\psi_1(\beta^{*\mathrm{T}}X)Z_iZ_j\} - \tilde{\mathbb{E}}\{\psi_1(\tilde{\beta}_j^{\mathrm{T}}\boldsymbol{Z})Z_iZ_j\}| \leq \tilde{\mathbb{E}}\{|\psi_1(\beta^{*\mathrm{T}}\boldsymbol{Z}) - \psi_1(\tilde{\beta}_j^{\mathrm{T}}\boldsymbol{Z})||Z_iZ_j|\} \\
\leq& C_5^2\tilde{\mathbb{E}}[\psi_1(\beta^{*\mathrm{T}}\boldsymbol{Z})\max\{1 - \exp(-C_3|(\hat{\beta} - \beta^*)^{\mathrm{T}}\boldsymbol{Z}|), \exp(C_3|(\hat{\beta} - \beta^*)^{\mathrm{T}}\boldsymbol{Z}|) - 1\}] \\
=& C_5^2C_6\tilde{\mathbb{E}}\{C_3|(\hat{\beta} - \beta^*)^{\mathrm{T}}\boldsymbol{Z}| + o_p(|(\hat{\beta} - \beta^*)^{\mathrm{T}}\boldsymbol{Z}|)\} \\
\leq& C_3C_5^3C_6\|\hat{\beta} - \beta^*\|_1 + o_p(\|\hat{\beta} - \beta^*\|_1) \\
=& o_p(1),
\end{aligned}$$

which leads to $-\tilde{\mathbb{E}}\{\psi_1(\tilde{\beta}_j^{\mathrm{T}}\boldsymbol{Z})Z_j\boldsymbol{Z}^{\mathrm{T}}\} + \tilde{\mathbb{E}}\{\psi_1(\beta^{*\mathrm{T}}\boldsymbol{Z})Z_j\boldsymbol{Z}^{\mathrm{T}}\} \xrightarrow{\mathbb{P}} 0$. Therefore, we consider the following decomposition,

$$
\begin{aligned}
&- \tilde{\mathbb{E}}\{\psi_1(\tilde{\beta}_j^{\mathrm{T}}\boldsymbol{Z})Z_j\boldsymbol{Z}\} + \mathbb{E}\{\psi_1(\beta^{*\mathrm{T}}\boldsymbol{Z})Z_j\boldsymbol{Z}\} \\
=&- \tilde{\mathbb{E}}\{\psi_1(\tilde{\beta}_j^{T}\boldsymbol{Z})Z_j\boldsymbol{Z}\} + \tilde{\mathbb{E}}\{\psi_1(\beta^{*\mathrm{T}}X)Z_j\boldsymbol{Z}\} - \tilde{\mathbb{E}}\{\psi_1(\beta^{*\mathrm{T}}\boldsymbol{Z})Z_j\boldsymbol{Z}\} + E\{\psi_1(\beta^{*\mathrm{T}}\boldsymbol{Z})Z_j\boldsymbol{Z}\} \\
\xrightarrow{\mathbb{P}}&0.
\end{aligned}
$$

Hence,

$$
\tilde{\mathbb{E}}\{\psi_1(\tilde{\beta}_j^{\mathrm{T}}\boldsymbol{Z})Z_j\boldsymbol{Z}\} \xrightarrow{\mathbb{P}} \mathbb{E}\{\psi_1(\beta^{*\mathrm{T}}\boldsymbol{Z})Z_j\boldsymbol{Z}\} = \boldsymbol{\Gamma}_j, \tag{S7.35}
$$

where $\boldsymbol{\Gamma}_j$ is the $j$-th row of $\boldsymbol{\Gamma}$, and (S7.35) holds for $j = 0, \ldots, m-1$. Hence, $\Delta_0^2 = \tilde{\mathbb{E}}(\tilde{\boldsymbol{Z}}\boldsymbol{Z}^{\mathrm{T}})(\hat{\beta} - \beta^*)$, where $\tilde{\boldsymbol{Z}}_j = -\psi_1(\tilde{\beta}_j^{\mathrm{T}}\boldsymbol{Z})Z_j$, and $\tilde{\mathbb{E}}(\tilde{\boldsymbol{Z}}\boldsymbol{Z}^{\mathrm{T}}) \xrightarrow{\mathbb{P}} -\boldsymbol{\Gamma}$. Suppose $\sqrt{N}(\hat{\beta} - \beta^*) \xrightarrow{d} G_2$, by continuous mapping theorem,

$$
\sqrt{N}\Delta_0^2 \xrightarrow{d} -\boldsymbol{\Gamma}G_2. \tag{S7.36}
$$

Besides, by central limit theorem,

$$
\sqrt{N}\Delta_0^2 = \sqrt{N}\tilde{\mathbb{E}}\left\{\tau(\boldsymbol{O}, \bar{\alpha}, \beta^*, \bar{\gamma})\right\} + o_p(1) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{\Lambda}), \tag{S7.37}
$$

where

$$
\boldsymbol{\Lambda} = \mathbb{E}\left\{\tau(\boldsymbol{O}, \bar{\alpha}, \beta^*, \bar{\gamma})\tau(\boldsymbol{O}, \bar{\alpha}, \beta^*, \bar{\gamma})^{\mathrm{T}}\right\}
$$

$$=\mathbb{E}\left(\left[\frac{1}{\pi(\boldsymbol{X};\bar{\gamma})}\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2+\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}^2\right]\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\right)$$
$$+2\mathbb{E}[\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}]. \tag{S7.38}$$

Therefore,

$$\sqrt{N}(\hat{\beta}-\beta^*)\xrightarrow{d}G_2\sim-\boldsymbol{\Gamma}^{-1}\mathrm{N}(0,\boldsymbol{\Lambda})\sim\mathrm{N}(0,\boldsymbol{\Sigma}),$$

where $\sim$ denotes "distributed as", i.e., for any two distributions $G_0$ and $G_1$, $G_0\sim G_1$ means the two distributions are the same.

### S7.3 Proof of Theorem 1(ii)

We show the consistency of $\hat{\boldsymbol{\Sigma}}$. First, if we let $\hat{\boldsymbol{\Gamma}}_j=\tilde{\mathbb{E}}\{\psi_1(\hat{\beta}^{\mathrm{T}}\boldsymbol{Z})Z_j\boldsymbol{Z}\}$, then

$$\hat{\boldsymbol{\Gamma}}_j\xrightarrow{\mathbb{P}}\mathbb{E}\{\psi_1(\beta^{*\mathrm{T}}\boldsymbol{Z})Z_j\boldsymbol{Z}\},$$

i.e., $\hat{\boldsymbol{\Gamma}}_j\xrightarrow{\mathbb{P}}\boldsymbol{\Gamma}_j$, which can be shown in the way similar to the proof of (S7.35). Then we get $\hat{\boldsymbol{\Gamma}}\xrightarrow{\mathbb{P}}\boldsymbol{\Gamma}$.

Next, we want to show that $\hat{\boldsymbol{\Lambda}}\xrightarrow{\mathbb{P}}\boldsymbol{\Lambda}$. Since $\boldsymbol{\Lambda}=\mathbb{E}\{\tau(\boldsymbol{O},\bar{\alpha},\beta^*,\bar{\gamma})\tau(\boldsymbol{O},\bar{\alpha},\beta^*,\bar{\gamma})^{\mathrm{T}}\}$ and $\hat{\boldsymbol{\Lambda}}=\tilde{\mathbb{E}}\{\tau(\boldsymbol{O},\hat{\alpha},\hat{\beta},\hat{\gamma})\tau(\boldsymbol{O},\hat{\alpha},\hat{\beta},\hat{\gamma})^{\mathrm{T}}\}$, it suffices to show that

$$\tilde{\mathbb{E}}\{\tau(\boldsymbol{O},\hat{\alpha},\hat{\beta},\hat{\gamma})\tau(\boldsymbol{O},\hat{\alpha},\hat{\beta},\hat{\gamma})^{\mathrm{T}}\}-\tilde{\mathbb{E}}\{\tau(\boldsymbol{O},\bar{\alpha},\beta^*,\bar{\gamma})\tau(\boldsymbol{O},\bar{\alpha},\beta^*,\bar{\gamma})^{\mathrm{T}}\}=o_p(1).$$

We consider the $i,j$-th element of the difference above:

$$
\begin{aligned}
&|[\tilde{\mathbb{E}}\{\tau(\boldsymbol{O},\hat{\alpha},\hat{\beta},\hat{\gamma})\tau(\boldsymbol{O},\hat{\alpha},\hat{\beta},\hat{\gamma})^{\mathrm{T}}\}-\tilde{\mathbb{E}}\{\tau(\boldsymbol{O},\bar{\alpha},\beta^*,\bar{\gamma})\tau(\boldsymbol{O},\bar{\alpha},\beta^*,\bar{\gamma})^{\mathrm{T}}\}]_{ij}| \\
=&\bigg|\tilde{\mathbb{E}}\bigg(\bigg[\frac{R}{\pi(\boldsymbol{X};\hat{\gamma})}\{Y-\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})\}+\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\hat{\beta}^{\mathrm{T}}\boldsymbol{Z})\}\bigg]^2 Z_iZ_j \\
&-\bigg[\frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}+\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\bigg]^2 Z_iZ_j\bigg)\bigg| \\
\le&\tilde{\mathbb{E}}\bigg(\bigg|\bigg[\frac{R}{\pi(\boldsymbol{X};\hat{\gamma})}\{Y-\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})\}+\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\hat{\beta}^{\mathrm{T}}\boldsymbol{Z})\}\bigg]^2 \\
&-\bigg[\frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}+\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\bigg]^2\bigg|\bigg|Z_iZ_j\bigg|\bigg) \\
\le&C_5^2\tilde{\mathbb{E}}\bigg(\bigg|\bigg[\frac{R}{\pi(\boldsymbol{X};\hat{\gamma})}\{Y-\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})\}+\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\hat{\beta}^{\mathrm{T}}\boldsymbol{Z})\}\bigg]^2 \\
&-\bigg[\frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}+\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\bigg]^2\bigg|\bigg) \\
\le&C_5^2\tilde{\mathbb{E}}\bigg(\bigg[\frac{R}{\pi(\boldsymbol{X};\hat{\gamma})}\{Y-\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})\}+\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\hat{\beta}^{\mathrm{T}}\boldsymbol{Z})\} \\
&-\frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}-\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\bigg]^2\bigg) \\
&+2C_5^2\tilde{\mathbb{E}}\bigg(\bigg[\frac{R}{\pi(\boldsymbol{X};\hat{\gamma})}\{Y-\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})\}+\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\hat{\beta}^{\mathrm{T}}\boldsymbol{Z})\} \\
&-\frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}-\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\bigg] \\
&\times\bigg[\frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}+\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\bigg]\bigg) \\
\le&C_5^2\tilde{\mathbb{E}}\bigg(\bigg[\frac{R}{\pi(\boldsymbol{X};\hat{\gamma})}\{Y-\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})\}+\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\hat{\beta}^{\mathrm{T}}\boldsymbol{Z})\} \\
&-\frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}-\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\bigg]^2\bigg) \\
&+2C_5^2\tilde{\mathbb{E}}^{\frac{1}{2}}\bigg(\bigg[\frac{R}{\pi(\boldsymbol{X};\hat{\gamma})}\{Y-\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})\}+\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\hat{\beta}^{\mathrm{T}}\boldsymbol{Z})\}
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}-\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\Big]^2\Big)\\
&\times \tilde{\mathbb{E}}^{\frac{1}{2}}\left(\left[\frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}+\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\right]^2\right),
\end{aligned}
$$

therefore, we only need to show that

$$
\begin{aligned}
&\tilde{\mathbb{E}}\Big(\Big[\frac{R}{\pi(\boldsymbol{X};\hat{\gamma})}\{Y-\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})\}+\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\hat{\beta}^{\mathrm{T}}\boldsymbol{Z})\}\\
&\quad-\frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}-\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\Big]^2\Big)\\
=&o_p(1).
\end{aligned}
$$

Consider the following decomposition:

$$
\begin{aligned}
&\frac{R}{\pi(\boldsymbol{X};\hat{\gamma})}\{Y-\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})\}+\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\hat{\beta}^{\mathrm{T}}\boldsymbol{Z})\}\\
&\quad-\frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}-\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\\
=&\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\left\{1-\frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\right\}\\
&\quad+R\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\left\{\frac{1}{\pi(\boldsymbol{X};\hat{\gamma})}-\frac{1}{\pi(\boldsymbol{X};\bar{\gamma})}\right\}\\
&\quad+R\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\left\{\frac{1}{\pi(\boldsymbol{X};\bar{\gamma})}-\frac{1}{\pi(\boldsymbol{X};\hat{\gamma})}\right\}\\
&\quad+\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})-\psi(\hat{\beta}^{\mathrm{T}}\boldsymbol{Z}),
\end{aligned}
$$

denoted as $\delta_0^1+\delta_1^1+\delta_2^1+\delta_3^1$. Let $\Delta_i^3=\tilde{\mathbb{E}}\{(\delta_i^1)^2\}$, $i=0,\ldots,3$, we only need to show that $\Delta_i^3=o_p(1)$, $i=0,\ldots,3$.

By mean value theorem,

$$\begin{aligned}
\Delta_0^3 =&\tilde{\mathbb{E}}\left[\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2\left\{1-\frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\right\}^2\right]\\
=&2\tilde{\mathbb{E}}\left[\psi_1(\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G})\{\psi(\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\boldsymbol{G}\left\{1-\frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\right\}^2\right]\\
=&2\tilde{\mathbb{E}}\left(\psi_1(\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G})\left[\int_0^1\psi_1\{\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}+u(\tilde{\alpha}-\bar{\alpha})^{\mathrm{T}}\boldsymbol{G}\}\mathrm{d}u\right]\{(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\boldsymbol{G}\}^2\left\{1-\frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\right\}^2\right)\\
\le&2\tilde{\mathbb{E}}\left[\psi_1(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\exp(C_3C_4r_\alpha)\left\{\int_0^1\psi_1(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\exp(C_3C_4r_\alpha)\mathrm{d}u\right\}\{(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\boldsymbol{G}\}^2\left\{1-\frac{R}{\pi(\boldsymbol{X};\bar{\gamma})}\right\}^2\right]\\
\le&2C_2^2C_4^2\exp(2C_3C_4r_\alpha)\{1+\exp(-B_0)\}^2\|\hat{\alpha}-\bar{\alpha}\|_1^2\\
=&o_p(1).
\end{aligned}$$

$$\begin{aligned}
\Delta_1^3 =&\tilde{\mathbb{E}}\left[R\left\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\right\}^2\left\{\frac{1}{\pi(\boldsymbol{X};\hat{\gamma})}-\frac{1}{\pi(\boldsymbol{X};\bar{\gamma})}\right\}^2\right]\\
=&-2\tilde{\mathbb{E}}\Big(Rw(\boldsymbol{X};\bar{\gamma})^2[\exp\{-(\tilde{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}-1]\exp\{-(\tilde{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}\times\left\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\right\}^2(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{G}\Big)\\
\le&2\{1+\exp(\eta_{01})\}^2\exp(-B_0+\eta_{01})C_4\|\hat{\gamma}-\bar{\gamma}\|_1\tilde{\mathbb{E}}\left[Rw(\boldsymbol{X};\bar{\gamma})\left\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\right\}^2\right]\\
\le&2\{1+\exp(\eta_{01})\}^2\exp(-B_0+\eta_{01})C_4\|\hat{\gamma}-\bar{\gamma}\|_1\Big(\mathbb{E}\left[Rw(\boldsymbol{X};\bar{\gamma})\left\{Y-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\right\}^2\right]+(D_0^2+D_0D_1)\lambda_0\Big)\\
=&o_p(1).
\end{aligned}$$

By mean value theorem,

$$\begin{aligned}
\Delta_2^3 =&\tilde{\mathbb{E}}\left[R\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2\left\{\frac{1}{\pi(\boldsymbol{X};\bar{\gamma})}-\frac{1}{\pi(\boldsymbol{X};\hat{\gamma})}\right\}^2\right]\\
=&\tilde{\mathbb{E}}\left(R\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2 w(\boldsymbol{X};\bar{\gamma})^2[1-\exp\{-(\hat{\gamma}-\bar{\gamma})^{\mathrm{T}}\boldsymbol{F}\}]^2\right)\\
\leq&\{1+\exp(\eta_{01})\}^2\exp(-2B_0)\tilde{\mathbb{E}}\left[\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2\right]\\
=&2\{1+\exp(\eta_{01})\}^2\exp(-2B_0)\tilde{\mathbb{E}}\left[\{\psi(\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G})-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\psi_1(\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G})(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\boldsymbol{G}\right]\\
=&2\{1+\exp(\eta_{01})\}^2\exp(-2B_0)\times\tilde{\mathbb{E}}\left(\psi_1(\tilde{\alpha}^{\mathrm{T}}\boldsymbol{G})\left[\int_0^1\psi_1\{\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}+u(\tilde{\alpha}-\bar{\alpha})^{\mathrm{T}}\boldsymbol{G}\}\mathrm{d}u\right]\{(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\boldsymbol{G}\}^2\right)\\
\leq&2C_2^2C_4^2\{1+\exp(\eta_{01})\}^2\exp(2C_3C_4r_\alpha-2B_0)\|\hat{\alpha}-\bar{\alpha}\|_1^2\\
=&o_p(1).
\end{aligned}$$

$$\begin{aligned}
\Delta_3^3 =&\tilde{\mathbb{E}}[\{\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})-\psi(\hat{\beta}^{\mathrm{T}}\boldsymbol{Z})\}^2]\\
=&2\tilde{\mathbb{E}}\left[\{\psi(\tilde{\beta}^{\mathrm{T}}\boldsymbol{Z})-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\psi_1(\tilde{\beta}^{\mathrm{T}}\boldsymbol{Z})(\hat{\beta}-\beta^*)^{\mathrm{T}}\boldsymbol{Z}\right]\\
=&2\tilde{\mathbb{E}}\left(\psi_1(\tilde{\beta}^{\mathrm{T}}\boldsymbol{Z})\left[\int_0^1\psi_1\{\beta^{*\mathrm{T}}\boldsymbol{Z}+u(\tilde{\beta}-\beta^*)^{\mathrm{T}}\boldsymbol{Z}\}\mathrm{d}u\right]\{(\hat{\beta}-\beta^*)^{\mathrm{T}}\boldsymbol{Z}\}^2\right)\\
\leq&2C_5^2C_6^2\{1+\exp(\eta_{01})\}^2\exp(2C_3C_5\|\hat{\beta}-\beta^*\|_1)\|\hat{\beta}-\beta^*\|_1^2\\
=&o_p(1).
\end{aligned}$$

Therefore, $\hat{\boldsymbol{\Lambda}}\xrightarrow{\mathbb{P}}\boldsymbol{\Lambda}$. Then, by continuous mapping theorem, $\hat{\boldsymbol{\Gamma}}^{-1}\xrightarrow{\mathbb{P}}\boldsymbol{\Gamma}^{-1}$. Thus, by continuous mapping theorem again, $\hat{\boldsymbol{\Sigma}}\xrightarrow{\mathbb{P}}\boldsymbol{\Sigma}$.

# S8 Extension to the setting of stratified sampling

## S8.1 Proof of Proposition 3

Let

$$\tau^s(\boldsymbol{O},\alpha,\beta,\cdot)=\frac{1}{N}\sum_{i=1}^{n}\frac{1}{(\cdot)}\{Y_i-\psi(\alpha^{\mathrm{T}}\boldsymbol{G}_i)\}\boldsymbol{Z}_i+\frac{1}{N}\sum_{i=1}^{N}\{\psi(\alpha^{\mathrm{T}}\boldsymbol{G}_i)-\psi(\beta^{\mathrm{T}}\boldsymbol{Z}_i)\}\boldsymbol{Z}_i. \tag{S8.39}$$

We have

$$\begin{aligned}
&-\tau^s\{\boldsymbol{O},\bar{\alpha},\beta^*,\pi(\boldsymbol{X};\bar{\gamma})\}=-\tau^s\{\boldsymbol{O},\bar{\alpha},\beta^*,\pi^*(\boldsymbol{X})\}=\tau^s\{\boldsymbol{O},\hat{\alpha},\hat{\beta}^s,\hat{\pi}(\boldsymbol{X})\}-\tau^s\{\boldsymbol{O},\bar{\alpha},\beta^*,\pi^*(\boldsymbol{X})\}\\
=&\frac{1}{N}\sum_{i=1}^{N}[\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)-\psi(\hat{\beta}^{s\mathrm{T}}\boldsymbol{Z}_i)\}\boldsymbol{Z}_i]-\frac{1}{N}\sum_{i=1}^{N}[\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z}_i)\}\boldsymbol{Z}_i]\\
&+\frac{1}{N}\sum_{i=1}^{N}[\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z}_i)\}\boldsymbol{Z}_i]-\frac{1}{N}\sum_{i=1}^{N}[\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z}_i)\}\boldsymbol{Z}_i]\\
&+\frac{1}{n}\sum_{i=1}^{n}\{Y_i-\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)\}\boldsymbol{Z}_i-\frac{1}{n}\sum_{i=1}^{n}\{Y_i-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)\}\boldsymbol{Z}_i,
\end{aligned}$$

denoted as $\Delta_5^0+\Delta_5^1+\Delta_5^2$.

We first deal with $\Delta_5^0$. Let $Z_{ij}$ denote the $j$-th co-ordinate of the $i$-th sample $\boldsymbol{Z}_i$. Then, we consider the $j$-th coordinate of $\Delta_5^0$, $\Delta_{5,j}^0$

$$\begin{aligned}
\Delta_{5,j}^0=&\frac{1}{N}\sum_{i=1}^{N}[\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)-\psi(\hat{\beta}^{s\mathrm{T}}\boldsymbol{Z}_i)\}Z_{ij}]-\frac{1}{N}\sum_{i=1}^{N}[\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)-\psi(\beta^{*\mathrm{T}}\boldsymbol{Z}_i)\}Z_{ij}]\\
=&\frac{1}{N}\sum_{i=1}^{N}[\{\psi(\beta^{*\mathrm{T}}\boldsymbol{Z}_i)-\psi(\hat{\beta}^{s\mathrm{T}}\boldsymbol{Z}_i)\}Z_{ij}]
\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{N}\sum_{i=1}^{N}\{\psi_1(\tilde{\beta}_j^{\mathrm{T}}\boldsymbol{Z}_i)Z_{ij}(\beta^*-\hat{\beta}^s)^{\mathrm{T}}\boldsymbol{Z}_i\}\\
=&\frac{1}{N}\sum_{i=1}^{N}\{\psi_1(\beta^{*\mathrm{T}}\boldsymbol{Z}_i)Z_{ij}(\beta^*-\hat{\beta}^s)^{\mathrm{T}}\boldsymbol{Z}_i\}+O_p(\|\beta^*-\hat{\beta}^s\|_1^2),
\end{aligned}
$$

where $\tilde{\beta}_j = u_j\beta^* + (1-u_j)\hat{\beta}^s$ for some $u_j \in (0,1)$. The last equality holds since $\psi_1(\beta^{*\mathrm{T}}\boldsymbol{Z}_i)\exp(-C_3C_5\|\beta^*-\hat{\beta}^s\|_1) \le \psi_1(\tilde{\beta}_j^{\mathrm{T}}\boldsymbol{Z}_i) \le \psi_1(\beta^{*\mathrm{T}}\boldsymbol{Z}_i)\exp(C_3C_5\|\beta^*-\hat{\beta}^s\|_1)$, which leads to $\psi_1(\tilde{\beta}_j^{\mathrm{T}}\boldsymbol{Z}_i) = \psi_1(\beta^{*\mathrm{T}}\boldsymbol{Z}_i) + O_p(\|\beta^*-\hat{\beta}^s\|_1)$.

Then we deal with $\Delta_5^1$ and $\Delta_5^2$ together. Consider the $j$-th coordinate of $\Delta_5^1+\Delta_5^2$:

$$
\begin{aligned}
&|\Delta_{5,j}^1+\Delta_{5,j}^2|\\
=&\left|\frac{1}{n}\sum_{i=1}^{n}\left[\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)-\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)\}Z_{ij}\right]+\frac{1}{N}\sum_{i=1}^{N}[\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)-\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)\}Z_{ij}]\right|\\
=&\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\psi_1(\tilde{\alpha}_j^{\mathrm{T}}\boldsymbol{G}_i)(\bar{\alpha}-\hat{\alpha})^{\mathrm{T}}\boldsymbol{G}_iZ_{ij}\right\}+\frac{1}{N}\sum_{i=1}^{N}\{\psi_1(\tilde{\alpha}_j^{\mathrm{T}}\boldsymbol{G}_i)(\hat{\alpha}-\bar{\alpha})^{\mathrm{T}}\boldsymbol{G}_iZ_{ij}\}\right|\\
\le&\|\hat{\alpha}-\bar{\alpha}\|_1\left\|\frac{1}{n}\sum_{i=1}^{n}\left\{\psi_1(\tilde{\alpha}_j^{\mathrm{T}}\boldsymbol{G}_i)\boldsymbol{G}_iZ_{ij}\right\}-\frac{1}{N}\sum_{i=1}^{N}\{\psi_1(\tilde{\alpha}_j^{\mathrm{T}}\boldsymbol{G}_i)\boldsymbol{G}_iZ_{ij}\}\right\|_\infty,
\end{aligned}
$$

where $\tilde{\alpha}_j = u_j\hat{\alpha} + (1-u_j)\bar{\alpha}$, for some $0 < u_j < 1$, $j = 0,\ldots,m-1$. Consider the $k$-th coordinate of $\boldsymbol{G}_i$. For technical convenience, we assume that $n$ is divisible by $N$. Let

$$
V_{ijk} \triangleq \left\{\frac{n}{N}-1\right\}\psi_1(\tilde{\alpha}_j^{\mathrm{T}}\boldsymbol{G}_i)G_{ik}Z_{ij} + \frac{n}{N}\sum_{s=n+(N/n-1)(i+1)+1}^{n+i(N/n-1)}\psi_1(\tilde{\alpha}_j^{\mathrm{T}}\boldsymbol{G}_s)G_{sk}Z_{sj},
$$

then, $\exists$ constant $C_7 > 0$, such that $|V_{ijk}| \le C_7$. Moreover, we have $\mathbb{E}(V_{ijk}) = 0$;

therefore, $\mathbb{E}(V_{ijk}^2) = \mathrm{Var}(V_{ijk}) = \{(n-N)^2/N^2 + n^2/N^2(N/n-1)\}\mathrm{Var}\{\psi_1(\tilde{\alpha}_j^{\mathrm{T}}\boldsymbol{G})G_k Z_j\} \leq C_8$ for some constant $C_8 > 0$. Let $t = \sqrt{C_8 \ln\{(q+1)/\epsilon\}/n}$, since $\ln(q+1) = o(n)$, $\exists\, n$ large enough, such that $t^2/C_8 \leq 3t/C_7$, then by Bernstein's inequality, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} V_{ijk}\right| \geq t\right) = \mathbb{P}\left(\left|\sum_{i=1}^{n} V_{ijk}\right| \geq nt\right) \leq \exp\left(-\frac{nt^2}{C_8}\right) = \frac{\epsilon}{q+1}. \tag{S8.40}$$

Let $V_j$ denote $\sup_{k=0,\ldots,q} |1/n \sum_{i=1}^{n} V_{ijk}|$, and

$$V_j < t \Rightarrow \frac{V_j\sqrt{n}}{\sqrt{C_8 \ln\{e(q+1)\}}} \leq \sqrt{1 - \ln(\epsilon)}.$$

It follows that

$$\begin{aligned}
&\mathbb{P}\left[\frac{V_j\sqrt{n}}{\sqrt{C_8 \ln\{e(q+1)\}}} \leq \sqrt{1-\ln(\epsilon)}\right] \geq \mathbb{P}(V_j < t) \\
=&1 - \mathbb{P}(V_j \geq t) \geq 1 - \sum_{k=0}^{q} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} V_{ijk}\right| \geq t\right) \\
=&1 - \epsilon.
\end{aligned}$$

Therefore, $\frac{V_j\sqrt{n}}{\sqrt{C_8 \ln\{e(q+1)\}}} = O_p(1)$, for $j = 0, \ldots, m-1$, it follows that

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{\pi^s(\bar{\gamma})}\psi_1(\tilde{\alpha}_j^{\mathrm{T}}\boldsymbol{G}_i)\boldsymbol{G}_i Z_{ij}\right\} - \frac{1}{N}\sum_{i=1}^{N}\{\psi_1(\tilde{\alpha}_j^{\mathrm{T}}\boldsymbol{G}_i)\boldsymbol{G}_i Z_{ij}\}\right\|_\infty = V_j = O_p(\sqrt{\ln(q+1)/n}).$$

Therefore,

$$|\Delta_{5,j}^1 + \Delta_{5,j}^2| \leq O_p(\sqrt{ln(q+1)/n})O_p(|S_{\bar{\alpha}}|\sqrt{ln(q+1)/n}) = o_p(1/\sqrt{n}),$$

it follows that $\Delta_5^1 + \Delta_5^2 = o_p(1/\sqrt{n})$. Hence, by central limit theorem and continuous mapping theorem, we have

$$\sqrt{n}(\hat{\beta}^s - \beta^*) \xrightarrow{d} \sqrt{n}\Gamma^{-1}\tau^s\{\boldsymbol{O}, \bar{\alpha}, \beta^*, \pi^*(\boldsymbol{X})\} \sim \mathrm{N}(0, \boldsymbol{\Sigma}^s).$$

### S8.2 Variance comparison

Because the following relationship holds,

$$\begin{aligned}
\frac{1}{n}\boldsymbol{\Lambda}^s =& \frac{1}{n}\mathbb{E}\left[\left\{Y - \frac{N-n}{N}\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \frac{n}{N}\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\right\}^2 \boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\right] + \frac{(N-n)}{N^2}\mathbb{E}\left[\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}^2\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\right] \\
=& \frac{1}{n}\mathbb{E}\left[\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}} + \left(\frac{n}{N}\right)^2\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}^2\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\right] \\
&+ \frac{2}{N}\mathbb{E}\left[\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\right] + \frac{N-n}{N^2}\mathbb{E}\left[\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}^2\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\right] \\
=& \frac{1}{n}\mathbb{E}\left[\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\right] + \frac{2}{N}\mathbb{E}\left[\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\right] \\
&+ \frac{1}{N}\mathbb{E}\left[\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}^2\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\right] \\
=& \frac{1}{N}\boldsymbol{\Lambda},
\end{aligned}$$

we obtain

$$\boldsymbol{\Sigma}^s/n = \boldsymbol{\Gamma}^{-1}\left(\frac{1}{n}\boldsymbol{\Lambda}^s\right)\boldsymbol{\Gamma}^{-1} = \boldsymbol{\Gamma}^{-1}\left(\frac{1}{N}\boldsymbol{\Lambda}\right)\boldsymbol{\Gamma}^{-1} = \boldsymbol{\Sigma}/N.$$

## S9 Details of numerical implementation and simulation

In this section, we provide details of the numerical implementation and simulation. We consider the estimators of population mean for $\boldsymbol{Z} = 1$, regression coefficients in the mean model for $\boldsymbol{Z} = X_1$ and $\boldsymbol{Z} = \boldsymbol{X}$, respectively, and $\psi(\cdot)$ is assumed to be the identity function.

### S9.1 Data generating process

Throughout the simulation, we generate the covariates $\boldsymbol{X}$ as follows. We first generate a random vector from $\mathrm{N}(0, \boldsymbol{\Sigma})$, where the variance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times 3}$ has elements $\boldsymbol{\Sigma}_{i,j}$ defined as $2^{-|i-j|}$ for $i, j = 1, 2, 3$. Then we clamp each of its coordinates within $[-3, 3]$ to obtain $(X_1, X_2, X_3)$ and $\boldsymbol{X} = (1, X_1, X_2, X_3)$. In addition, the data source indicator $R$ follows a Bernoulli distribution with success probability $\pi(\boldsymbol{X})$, where $\pi(\boldsymbol{X}) = \{1 + \exp(-\gamma^{\mathrm{T}}\boldsymbol{F})\}^{-1}$, the parameter $\gamma = (-1.5, -0.8, -0.2, 0.3, 0, \ldots, 0)^{\mathrm{T}}$ and the basis functions $\boldsymbol{F}$ are described in Section 6.

**Study I.** We first focus on the estimation of the population mean and consider two data-generating mechanisms:

- **Case 1.** The outcome $Y = -0.2 + 0.1\tilde{X}_1 + 0.4\tilde{X}_2 + 0.7\tilde{X}_3 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1)$

and $\tilde{X}_j = X_j \cdot |X|_j^{0.1} + X_j \cdot |X|_j^{0.3} + X_j \cdot |X|_j^{0.5}$, for $j = 1, 2, 3$. We set $\boldsymbol{Z} = 1$.

- **Case 2.** The outcome $Y = -0.2 + 0.1\tilde{X}_1 + 0.4\tilde{X}_2 + 0.7\tilde{X}_3 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1)$ and $\tilde{X}_j = |X_j| \exp(|X_j|^{0.1} + |X_j|^{0.3})$ for $j = 1, 2, 3$. We set $\boldsymbol{Z} = 1$.

In many scenarios, estimating the conditional mean given a subset of variables in $\boldsymbol{X}$ garners statistical interest. Accordingly, we design experiments in Study II to evaluate the performance of the proposed estimator in such setups.

**Study II.** We further consider three additional cases for estimating regression coefficients in the conditional mean outcome model.

- **Case 3.** The outcome $Y = 0.4\tilde{X}_1 + 0.2\tilde{X}_2 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1)$, $\tilde{X}_1 = X_1 + X_1^2$ and $\tilde{X}_2 = \cos(\pi/9 \cdot X_1 \cdot X_3)$. We set $\boldsymbol{Z} = X_1$.

- **Case 4.** The outcome $Y = 0.4\tilde{X}_1 + 0.2\tilde{X}_2 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1)$, $\tilde{X}_1 = X_1 \cdot \mathbb{I}\{X_1 > 0\}\sqrt{|X_1|}$ and $\tilde{X}_2 = X_1 \cdot X_2$. We set $\boldsymbol{Z} = X_1$.

- **Case 5.** The outcome $Y = -0.2 + 0.1\tilde{X}_1 + 0.4\tilde{X}_2 + 0.7\tilde{X}_3 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1)$, $\tilde{X}_1 = X_1 \cdot X_2$, $\tilde{X}_2 = X_2 \cdot X_3$ and $\tilde{X}_3 = X_1 \cdot X_3$. We set $\boldsymbol{Z} = \boldsymbol{X}$.

Cases 3 and 4 involve $\boldsymbol{Z}$ as a specific covariate $X_1$, while Case 5 involves $\boldsymbol{Z}$ as the full set of covariates $\boldsymbol{X}$. In addition, for all Cases 1–5, OR models are misspecified.

### S9.2 Implementation details

We compare the proposed method $\text{AIPW}_{\text{RCAL}}$ with IPW, $\text{AIPW}_{\text{RML}}$ and $\text{AIPW}_{\text{CF}}$ introduced in Section 6. Both the Lasso-regularized calibrated and maximum likeli-

hood estimators for the PS and OR models can be implemented using the R package `RCAL` (Tan and Sun, 2020). We employ 5-fold cross-fitting to select the optimal tuning parameters. In addition, by equation (2.2), $\beta^* = \mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}})^{-1}\mathbb{E}(Y\boldsymbol{Z})$. Thus the true value of $\beta^*$ is calculated as $\tilde{\mathbb{E}}(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}})^{-1}\tilde{\mathbb{E}}(Y\boldsymbol{Z})$ through a simulation with a sample size of 100,000. For $\boldsymbol{Z} = X_1$ and $\boldsymbol{Z} = \boldsymbol{X}$, we denote $\beta^* = \beta_1$ and $\beta^* = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}}$, respectively.

### S9.3 Summary of results

We present and analyze the simulation results for Study I and Study II. In the following tables, we compare various methods in terms of Bias, $\sqrt{\mathrm{Var}}$, $\sqrt{\mathrm{EVar}}$, CP90, and CP95. As discussed in the paragraph below Theorem 1, under high-dimensional settings, the IPW estimator is not $\sqrt{N}$-consistent, and its asymptotic normality is not well established. Therefore, we do not report its numerical results for $\sqrt{\mathrm{EVar}}$, CP90, and CP95.

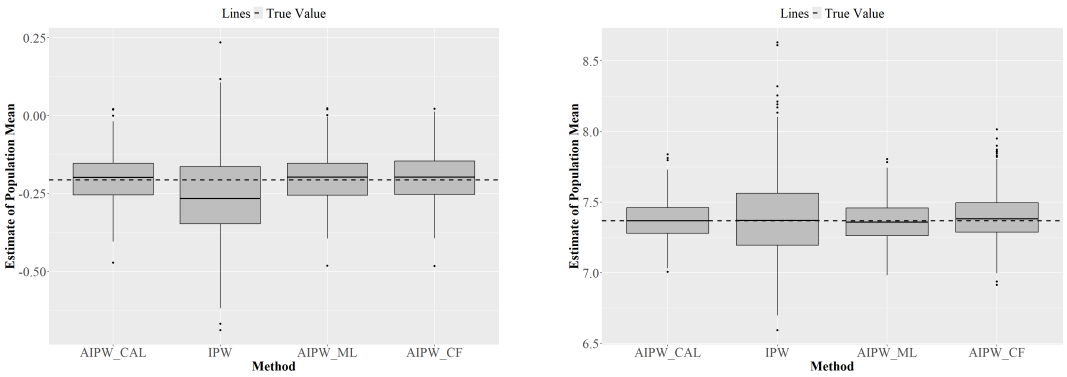
**Results for Study I.** Table 1 shows the numerical results for the estimation of population mean $\mathbb{E}(Y)$. From the table, the proposed method $\mathrm{AIPW}_{\mathrm{RCAL}}$ has the smallest $\sqrt{\mathrm{Var}}$ and $\sqrt{\mathrm{EVar}}$, and Bias. Moreover, CP90 and CP95 of the proposed method are more aligned with their nominal values of 0.90 and 0.95, respectively. This indicates the effectiveness of the proposed method in terms of estimating the population mean.

Figure 1 depicts the box plots of the estimates for Case 1 and Case 2, where the

Table 1: Summary of estimates of population mean for Study I.

| | **Case 1** | | | **Case 2** | | |
|---|---|---|---|---|---|---|
| | $\text{AIPW}_{\text{RCAL}}$ | $\text{AIPW}_{\text{RML}}$ | $\text{AIPW}_{\text{CF}}$ | $\text{AIPW}_{\text{RCAL}}$ | $\text{AIPW}_{\text{RML}}$ | $\text{AIPW}_{\text{CF}}$ |
| Bias | 0.004 | 0.004 | 0.006 | 0.002 | -0.006 | 0.031 |
| $\sqrt{\text{Var}}$ | 0.078 | 0.078 | 0.079 | 0.139 | 0.143 | 0.166 |
| $\sqrt{\text{EVar}}$ | 0.079 | 0.079 | 0.082 | 0.140 | 0.143 | 0.161 |
| CP90 | 0.904 | 0.908 | 0.918 | 0.898 | 0.886 | 0.896 |
| CP95 | 0.948 | 0.954 | 0.956 | 0.948 | 0.946 | 0.956 |

horizontal dash line indicates the true value. In both cases, our method $\text{AIPW}_{\text{RCAL}}$ exhibits the smallest biases, interquartile ranges, and whiskers, indicating the smallest variances compared to the other methods. In addition, $\text{AIPW}_{\text{CF}}$ shows more outliers than the other two methods, which is more apparent in the results of Study II, suggesting that cross-fitting may cause instability for the estimates.




(a) Case 1 (b) Case 2

Figure 1: Box plots for estimates of population mean.

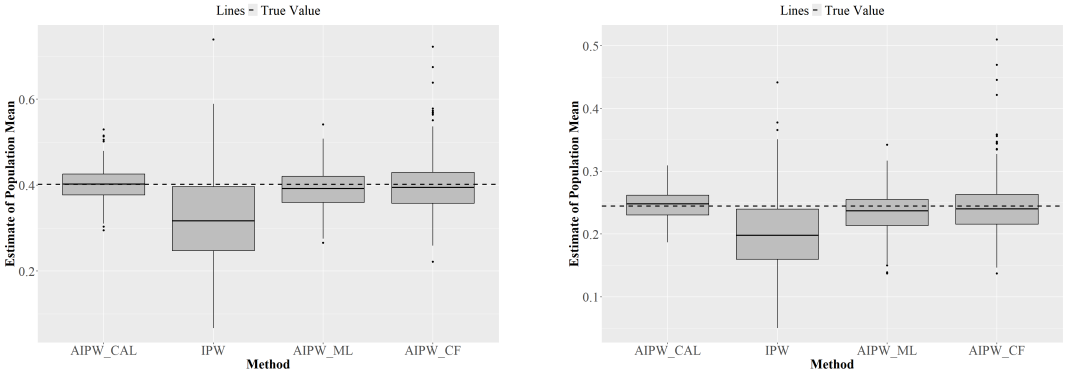
**Results for Study II.** The simulation results for Cases 3–5 are presented in Tables 2–3, and the corresponding box plots are displayed in Figures 2–3. We observe similar patterns as those in Cases 1 and 2: the proposed method performs well in terms of all metrics.

Table 2: Summary of estimates of $\beta_1$ in Cases 3 and 4.

| | **Case 3** | | | **Case 4** | | |
|---|---|---|---|---|---|---|
| | $\text{AIPW}_{\text{RCAL}}$ | $\text{AIPW}_{\text{RML}}$ | $\text{AIPW}_{\text{CF}}$ | $\text{AIPW}_{\text{RCAL}}$ | $\text{AIPW}_{\text{RML}}$ | $\text{AIPW}_{\text{CF}}$ |
| Bias | 0.000 | -0.011 | -0.004 | 0.002 | -0.010 | -0.002 |
| $\sqrt{\text{Var}}$ | 0.036 | 0.043 | 0.058 | 0.023 | 0.031 | 0.042 |
| $\sqrt{\text{EVar}}$ | 0.036 | 0.037 | 0.056 | 0.021 | 0.025 | 0.042 |
| CP90 | 0.884 | 0.824 | 0.840 | 0.866 | 0.774 | 0.838 |
| CP95 | 0.946 | 0.886 | 0.900 | 0.930 | 0.834 | 0.906 |

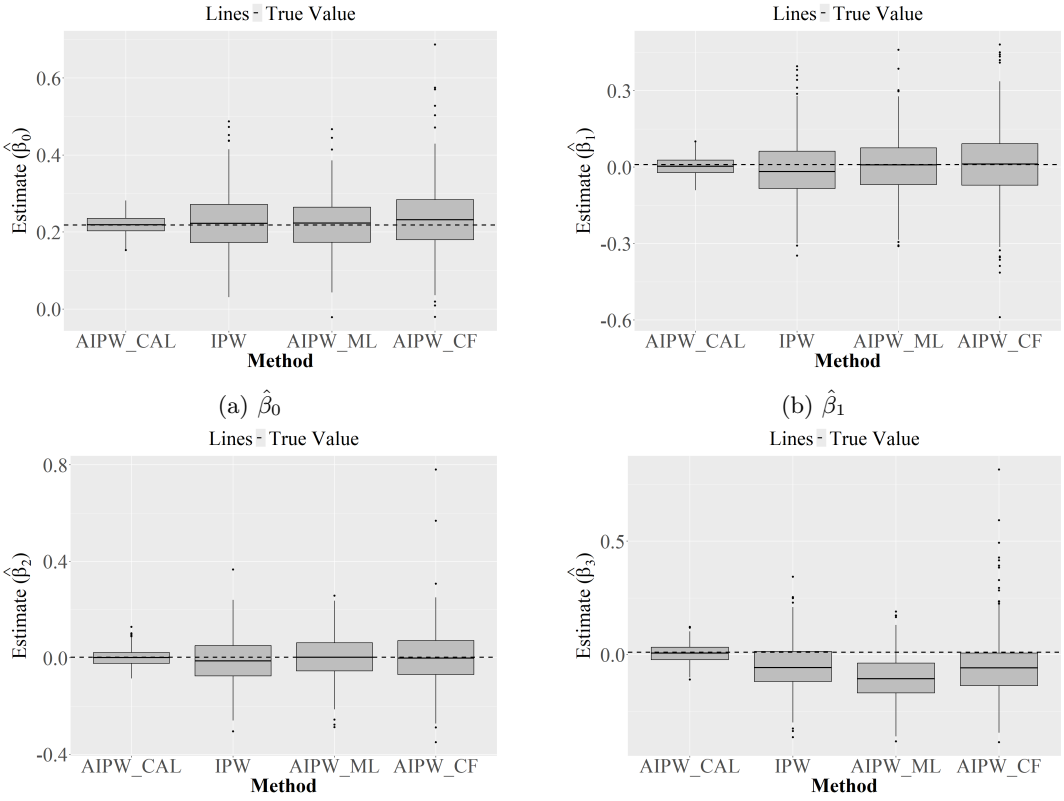
Table 3: Summary of estimates of $\beta_0, \beta_1, \beta_2, \beta_3$ in Case 5

| | $\beta_0$ | | | $\beta_1$ | | |
|---|---|---|---|---|---|---|
| | $\text{AIPW}_{\text{RCAL}}$ | $\text{AIPW}_{\text{RML}}$ | $\text{AIPW}_{\text{CF}}$ | $\text{AIPW}_{\text{RCAL}}$ | $\text{AIPW}_{\text{RML}}$ | $\text{AIPW}_{\text{CF}}$ |
| Bias | 0.001 | 0.001 | 0.013 | -0.007 | -0.004 | 0.000 |
| $\sqrt{\text{Var}}$ | 0.024 | 0.068 | 0.084 | 0.036 | 0.114 | 0.139 |
| $\sqrt{\text{EVar}}$ | 0.025 | 0.055 | 0.077 | 0.036 | 0.089 | 0.131 |
| CP90 | 0.910 | 0.800 | 0.846 | 0.886 | 0.774 | 0.830 |
| CP95 | 0.958 | 0.874 | 0.914 | 0.942 | 0.842 | 0.900 |
| | $\beta_2$ | | | $\beta_3$ | | |
| | $\text{AIPW}_{\text{RCAL}}$ | $\text{AIPW}_{\text{RML}}$ | $\text{AIPW}_{\text{CF}}$ | $\text{AIPW}_{\text{RCAL}}$ | $\text{AIPW}_{\text{RML}}$ | $\text{AIPW}_{\text{CF}}$ |
| Bias | -0.001 | 0.002 | 0.001 | -0.006 | 0.114 | -0.060 |
| $\sqrt{\text{Var}}$ | 0.033 | 0.086 | 0.111 | 0.040 | 0.095 | 0.132 |
| $\sqrt{\text{EVar}}$ | 0.034 | 0.075 | 0.108 | 0.041 | 0.095 | 0.133 |
| CP90 | 0.920 | 0.818 | 0.866 | 0.898 | 0.592 | 0.774 |
| CP95 | 0.958 | 0.902 | 0.938 | 0.940 | 0.682 | 0.842 |




(a) Case 3 (b) Case4

Figure 2: Box plots of estimates of $\beta_1$ in Cases 3 and 4.




(a) $\hat{\beta}_0$ (b) $\hat{\beta}_1$

(c) $\hat{\beta}_2$ (d) $\hat{\beta}_3$

Figure 3: Box plots of estimates of $\beta_0, \beta_1, \beta_2, \beta_3$ in Case 5

# S10 Details of the application

## S10.1 Pre-processing details of the community and crime dataset

We pre-process the data in following steps:

Step 1. remove 22 covariates missing 84% of data and 2 variables missing roughly 59% of data;

Step 2. remove covariates with weak linear relationships to the response `ViolentCrimesPerPop`

based on their correlation coefficients.

Step 3. remove covariates that exhibit multi-collinearity based on their values of variance inflation factors.

After the process, we obtain 1993 observations of 26 covariates.

## S10.2 Test results of the covariate shift

**Kernel two-sample test with maximum mean discrepancy**

- Kernel: $\exp(-\|\cdot\|_2^2)$
- MMD: 0.39227
- P-value: 0.001

**Bootstrap KS-tests for univariate covariates**

| Covariate | Bootstrap-KS P-value | KS-test Statistic | KS-test Approximate P-value |
|---|---|---|---|
| racePctHisp | 0.000 | 0.335 | 0.000 |
| pctWWage | 0.000 | 0.230 | 0.000 |
| pctWInvInc | 0.000 | 0.330 | 0.000 |
| blackPerCap | 0.000 | 0.421 | 0.000 |
| PctLess9thGrade | 0.010 | 0.119 | 0.010 |
| PctUnemployed | 0.000 | 0.231 | 0.000 |
| PctOccupManu | 0.000 | 0.205 | 0.000 |

| | | | |
|---|---|---|---|
| MalePctDivorce | 0.000 | 0.373 | 0.000 |
| MalePctNevMarr | 0.000 | 0.202 | 0.000 |
| PctTeen2Par | 0.000 | 0.289 | 0.000 |
| PctIlleg | 0.000 | 0.200 | 0.000 |
| NumImmig | 0.000 | 0.267 | 0.000 |
| PctImmigRec10 | 0.001 | 0.141 | 0.001 |
| PctHousLess3BR | 0.000 | 0.218 | 0.000 |
| MedNumBR | 0.000 | 0.138 | 0.000 |
| HousVacant | 0.000 | 0.184 | 0.000 |
| PctHousOccup | 0.000 | 0.195 | 0.000 |
| PctHousOwnOcc | 0.000 | 0.285 | 0.000 |
| PctVacantBoarded | 0.797 | 0.040 | 0.923 |
| PctHousNoPhone | 0.000 | 0.373 | 0.000 |
| PctWOFullPlumb | 0.000 | 0.146 | 0.000 |
| RentLowQ | 0.000 | 0.529 | 0.000 |
| MedRentPctHousInc | 0.062 | 0.089 | 0.099 |
| NumInShelters | 0.022 | 0.087 | 0.113 |
| NumStreet | 0.001 | 0.101 | 0.043 |
| PopDens | 0.000 | 0.266 | 0.000 |

Table 4: Bootstrap KS-tests for univariate covariates

### S10.3 Design of basis functions

In this application, we design basis functions in the following way: Given $\{X_{ij}\}_{j=1}^{N}$, i.e., N samples of the $i$-th coordinate of $\boldsymbol{X}$, let $\{\xi_{ij}\}_{j=1}^{n_k}$ be the $n_k$ points equally spaced within the $[-a_i, b_i]$, where $a_i = \min_{j=1,\ldots,N} X_{ij}$ and $b_i = \max_{j=1,\ldots,N} X_{ij}$. Let $f_{ij}(\boldsymbol{X})$ denote $(X_i - \xi_{ij})_+$, $i = 1, \ldots, d; j = 1, \ldots, n_k$. Let $\boldsymbol{F} = \{1, f_{11}(\boldsymbol{X}), \ldots, f_{1n_k}(\boldsymbol{X}), \ldots, f_{d1}(\boldsymbol{X}), \ldots, f_{dn_k}(\boldsymbol{X})\}^{\mathrm{T}}$ be the basis functions of the PS model, and let $\boldsymbol{G} = \{\boldsymbol{F}^{\mathrm{T}}, (\boldsymbol{Z} \bigotimes \boldsymbol{F})^{\mathrm{T}}\}^{\mathrm{T}}$. We choose $n_k = 4$ in this application.

## References

Bühlmann, P. and S. Van De Geer (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media.

Chakrabortty, A., J. Lu, T. T. Cai, and H. Li (2019). High dimensional m-estimation with missing outcomes: A semi-parametric framework. *arXiv preprint arXiv:1911.11345*.

Chen, K. and Y. Zhang (2023). Enhancing efficiency and robustness in high-dimensional linear regression with additional unlabeled data. *arXiv preprint arXiv:2311.17685*.

Deng, S., Y. Ning, J. Zhao, and H. Zhang (2023). Optimal and safe estimation for high-dimensional semi-supervised learning. *Journal of the American Statistical Association*, 1–12.

Ghosh, S. and Z. Tan (2022). Doubly robust semiparametric inference using regularized calibrated estimation with high-dimensional data. *Bernoulli 28*, 1675–1703.

Tan, Z. (2020a). Model-assisted inference for treatment effects using regularized calibrated estimation with

high-dimensional data. *The Annals of Statistics 48*, 811–837.

Tan, Z. (2020b). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika 107*, 137–158.

Tan, Z. and B. Sun (2020). *RCAL: Regularized calibrated estimation*. R package version 2.0.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Zhang, Y., A. Giessing, and Y.-C. Chen (2023). Efficient inference on high-dimensional linear models with missing outcomes. *arXiv preprint arXiv:2309.06429*.