Transfer learning for covariate-adaptive randomized trials with high-dimensional generalized linear models

Yujia Gu[1], Hanzhong Liu[1], Wei Ma[2]

[1]*Tsinghua University and* [2]*Renmin University of China*

Supplementary Material

The supplementary material is organized as follows. In Section S1, we construct consistent variance estimators for all four proposed estimators. In Section S2, we provide additional details on the generation of the semi-synthetic clinical trial data described in Section 7. Sections S3 and S4 present proofs of the main results and auxiliary lemmas.

S1 Consistent variance estimator

In this section, we construct consistent variance estimators for valid inference. Specifically, we consider the difference in the potential outcomes for any two treatments b and c. Let e_{bc} be an A-dimensional vector, such that all elements are zero, except for the bth element, which is 1, and the cth element, which is -1. Let $\tau_{bc} = E\{Y(b) - Y(c)\} = \tau_b - \tau_c,\ b, c \in \mathcal{A}$, be the real difference in potential outcomes. To simplify our notations, we use the

following expressions of $\hat{\tau}_{\text{glm}}$ and $\hat{\tau}_{\text{tl}}$, which is

$$\hat{\tau}^{\text{naive}}_{\dagger,a} = \sum_{k\in\mathcal{K}} p_{n[k]} \Bigg[\left\{ \bar{Y}_{[k]a} - \frac{1}{n_{[k]a}} \sum_{i\in[k]} \left(I(A_i = a) - \frac{n_{[k]a}}{n_{[k]}} \right) \hat{\Psi}^{\top}_{[k],\dagger}(X_i) \hat{\beta}^{\text{naive}}_{[k]}(a) \right\} - \left\{ \bar{Y}_{[k]0} - \frac{1}{n_{[k]0}} \sum_{i\in[k]} \left(I(A_i = 0) - \frac{n_{[k]0}}{n_{[k]}} \right) \hat{\Psi}^{\top}_{[k],\dagger}(X_i) \hat{\beta}^{\text{naive}}_{[k]}(0) \right\} \Bigg],$$

where $\hat{\beta}^{\text{naive}}_{[k]}(a) = \beta^{\text{naive}}_{[k]}(a) = e_a$. Let $\hat{\tau}^{*}_{\dagger,bc}$ be the estimator for τ_{bc} generated from the estimators above, for $\dagger = \text{glm}, \text{tl}, * = \text{naive}, \text{proj}$. Based on simple calculations, we have

$$\sqrt{n}(\sigma^{*}_{\text{glm,bc}})^{-1}(\hat{\tau}^{*}_{\dagger,bc} - \tau_{bc}) \xrightarrow{d} \text{N}(0,1), \quad \dagger = \text{glm}, \text{tl}, * = \text{naive}, \text{proj},$$

where

$$(\sigma^{\text{naive}}_{\text{glm,bc}})^2 = e_{bc}^{\top} V_{\text{glm}} e_{bc} = V_{\text{glm}}(b,b) - 2V_{\text{glm}}(b,c) + V_{\text{glm}}(c,c),$$

$$(\sigma^{\text{proj}}_{\text{glm,bc}})^2 = e_{bc}^{\top} V^{\text{proj}}_{\text{glm}} e_{bc} = V^{\text{proj}}_{\text{glm}}(b,b) - 2V^{\text{proj}}_{\text{glm}}(b,c) + V^{\text{proj}}_{\text{glm}}(c,c).$$

Here, for any matrix V, $V(i,j)$ denotes the element in the ith row and jth column. By using $r_i(a)$ as the transformed outcomes again and $\hat{r}_i(a)$ as an estimator of $r_i(a)$, $\hat{r}_i = \sum_{a\in\mathcal{A}_0} I(A_i = a)\hat{r}_i(a)$. Let $\bar{\hat{r}}_{[k]a} = 1/n_{[k]a} \sum_{i=1}^{n} I(A_i = a, B_i = k)\hat{r}_i$. For the estimators above, we use $\hat{r}^{*}_{i,\dagger} = Y_i - \hat{\Psi}_{[k],\dagger}(X_i)^{\top}\hat{\beta}^{*}_{[k]}(a)$, where $\dagger = \text{glm}, \text{tl}, * = \text{naive}, \text{proj}$, respectively, for $A_i = a$, $B_i = k$. We define

$$\hat{\sigma}^2_{r^{*}_{\dagger}(a)} = \sum_{k\in\mathcal{K}} \frac{p_{n[k]} n_{[k]}}{n_{[k]a}} \left\{ \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)(\hat{r}^{*}_{i,\dagger} - \bar{\hat{r}}^{*}_{[k]a,\dagger})^2 \right\},$$

where $\dagger = \text{glm}, \text{tl}, * = \text{naive}, \text{proj}$, and

$$\hat{\sigma}^2_{\text{bc},HY} = \sum_{k\in\mathcal{K}} p_{n[k]} \left\{ \left(\bar{Y}_{[k]b} - \sum_{k'\in\mathcal{K}} p_{n[k']} \bar{Y}_{[k']b} \right) - \left(\bar{Y}_{[k]c} - \sum_{k'\in\mathcal{K}} p_{n[k']} \bar{Y}_{[k']c} \right) \right\}^2.$$

Let

$$S_{\dagger,[k]\Psi\Psi} = \frac{1}{n_{[k]}} \sum_{i=1}^{n} I(B_i = k) \left\{ \hat{\Psi}_{[k],\dagger}(X_i) - \bar{\hat{\Psi}}'_{[k],\dagger}(X_i) \right\} \left\{ \hat{\Psi}_{[k],\dagger}(X_i) - \bar{\hat{\Psi}}'_{[k],\dagger}(X_i) \right\}^\top$$

denote the sample covariance matrix for $\Psi_{[k],\dagger}(\tilde{X})$. Also, let

$$S_{\dagger,[k]\Psi Y(a)} = \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k) \left\{ \hat{\Psi}_{[k],\dagger}(X_i) - \bar{\hat{\Psi}}'_{[k]a,\dagger}(X_i) \right\} (Y_i - \bar{Y}_{[k]a})$$

denote the sample stratum-specific covariance vector between $\Psi_{[k],\dagger}(X_i)$ and $Y_i(a)$. Let $(\hat{\sigma}^{\text{naive}}_{\dagger,\text{bc}})^2$ be the estimator of $\sigma^2_{\text{glm,bc}}$ and $(\hat{\sigma}^{\text{proj}}_{\dagger,\text{bc}})^2$ be the estimator of $(\sigma^{\text{proj}}_{\text{glm,bc}})^2$ for $\dagger = \text{glm}, \text{tl}$, respectively.

Theorem S1. *1. Under Assumptions 1–6, if* $r_{i,[k]\text{glm}}(a) \in \mathcal{R}_2$, $E\{r^2_{i,\text{glm}}(a)\} < \infty$ *for* $a \in \mathcal{A}_0$, *the variance estimator*

$$\begin{aligned}\hat{\sigma}^2_{\text{glm,bc}} &= \hat{\sigma}^2_{r^{\text{naive}}_{\text{glm}}(b)} + \hat{\sigma}^2_{r^{\text{naive}}_{\text{glm}}(c)} + \hat{\sigma}^2_{\text{bc},HY} \\ &- \sum_{k\in\mathcal{K}} p_{n[k]} \Big[\big\{ \beta^{\text{naive}}_{[k]}(b) - \beta^{\text{naive}}_{[k]}(c) \big\}^\top S_{\text{glm},[k]\Psi\Psi} \big\{ \beta^{\text{naive}}_{[k]}(b) - \beta^{\text{naive}}_{[k]}(c) \big\} \\ &\qquad - 2 \big\{ \beta^{\text{naive}}_{[k]}(b) - \beta^{\text{naive}}_{[k]}(c) \big\}^\top (S_{\text{glm},[k]\Psi Y(b)} - S_{\text{glm},[k]\Psi Y(c)}) \Big]\end{aligned}$$

is a consistent estimator for $\sigma^2_{\text{glm,bc}}$, *and*

$$(\hat{\sigma}^{\text{proj}}_{\text{glm,bc}})^2 = \hat{\sigma}^2_{r^{\text{proj}}_{\text{glm}}(b)} + \hat{\sigma}^2_{r^{\text{proj}}_{\text{glm}}(c)} + \hat{\sigma}^2_{\text{bc},HY}$$

$$+\sum_{k\in\mathcal{K}} p_{n[k]}\big\{\hat{\beta}^{\text{proj}}_{[k]}(b)-\hat{\beta}^{\text{proj}}_{[k]}(c)\big\}^\top S_{\text{glm},[k]\Psi\Psi}\big\{\hat{\beta}^{\text{proj}}_{[k]}(b)-\hat{\beta}^{\text{proj}}_{[k]}(c)\big\}$$

is a consistent estimator for $(\sigma^{\text{proj}}_{\text{glm,bc}})^2$.

2. *Under conditions in Theorem 2, the variance estimator*

$$\begin{aligned}\hat{\sigma}^2_{\text{tl,bc}} &= \hat{\sigma}^2_{r^{\text{naive}}_{\text{tl}}(b)}+\hat{\sigma}^2_{r^{\text{naive}}_{\text{tl}}(c)}+\hat{\sigma}^2_{\text{bc},HY}\\ &\quad-\sum_{k\in\mathcal{K}} p_{n[k]}\Big[\big\{\beta^{\text{naive}}_{[k]}(b)-\beta^{\text{naive}}_{[k]}(c)\big\}^\top S_{\text{tl},[k]\Psi\Psi}\big\{\beta^{\text{naive}}_{[k]}(b)-\beta^{\text{naive}}_{[k]}(c)\big\}\\ &\qquad\qquad-2\big\{\beta^{\text{naive}}_{[k]}(b)-\beta^{\text{naive}}_{[k]}(c)\big\}^\top(S_{\text{tl},[k]\Psi Y(b)}-S_{\text{tl},[k]\Psi Y(c)})\Big]\end{aligned}$$

is a consistent estimator for $\sigma^2_{\text{glm,bc}}$, *and*

$$\begin{aligned}(\hat{\sigma}^{\text{proj}}_{\text{tl,bc}})^2 &= \hat{\sigma}^2_{r^{\text{proj}}_{\text{tl}}(b)}+\hat{\sigma}^2_{r^{\text{proj}}_{\text{tl}}(c)}+\hat{\sigma}^2_{\text{bc},HY}\\ &\quad+\sum_{k\in\mathcal{K}} p_{n[k]}\big\{\hat{\beta}^{\text{proj}}_{[k]}(b)-\hat{\beta}^{\text{proj}}_{[k]}(c)\big\}^\top S_{\text{tl},[k]\Psi\Psi}\big\{\hat{\beta}^{\text{proj}}_{[k]}(b)-\hat{\beta}^{\text{proj}}_{[k]}(c)\big\}\end{aligned}$$

is a consistent estimator for $(\sigma^{\text{proj}}_{\text{glm,bc}})^2$.

S2 Details of semi-synthetic clinical trial data

In this section, we provide more details in generating the semi-synthetic trial data in Section 7. Recall that we analyze a semi-synthetic clinical trial dataset constructed based on a trial using nefazodone and the Cognitive Behavioral-Analysis System of Psychotherapy (CBASP) (Keller et al., 2000). The aim of the trial is to compare the performance of nefazodone,

CBASP, and their combination in chronic depression. We use the CBASP group as the treatment group, indexed by $A_i = 1$, and the combination group as the control group, indexed by $A_i = 0$. The binary outcome of interest is FinalHAMD, the final score on the 24-item Hamilton Rating Scale for Depression. The continuous score is dichotomized at a threshold of 18 to obtain a binary outcome. The treatment effect is evaluated by the risk difference. The semi-synthetic data for the binary outcome are generated by fitting a logistic regression model with nonparametric spline terms, using the function `bigssg` from the R package `bigsplines`. We stratify the data by GENDER and select eight additional covariates as baseline variables for the spline terms, including AGE, HAMD17, HAMD24, HAMD_CONGID, HAMA_SOMATI, Mstatus2, NDE, and TreatPD. We define the target population as individuals with HAMD17 < 20 and the source population as those with HAMD17 ≥ 20. In each iteration, we generate semi-synthetic data by sampling 200 units from the target population and 1200 units from the source population. Additional covariates included in the estimators are constructed using linear, quadratic, cubic, and interaction terms for continuous covariates; linear and interaction terms for binary covariates; and interactions between continuous and binary covariates, resulting in a total of $p = 135$ covariates. We perform stratified block randomization in each

dataset, with GENDER as the stratification variable and a 1:1 allocation ratio between treatment and control groups.

S3 Proof of main theorems

S3.1 Proof of Proposition 1

Recall that $\hat{\tau}_{\text{glm,a}}$ is

$$\hat{\tau}_{\text{glm,a}} = \sum_{k=1}^{K} p_{n[k]} \Bigg\{ \Big[\bar{Y}_{[k]a} - \{ \bar{\psi}'_{[k]a}(X_i^\top \hat{\beta}_{[k]\text{glm}}(a)) - \bar{\psi}'_{[k]}(X_i^\top \hat{\beta}_{[k]\text{glm}}(a)) \} \Big] - \Big[\bar{Y}_{[k]0} - \{ \bar{\psi}'_{[k]0}(X_i^\top \hat{\beta}_{[k]\text{glm}}(0)) - \bar{\psi}'_{[k]}(X_i^\top \hat{\beta}_{[k]\text{glm}}(0)) \} \Big] \Bigg\}$$

for $a \in \mathcal{A}$. Then we have

$$\begin{aligned}
\hat{\tau}_{\text{glm}} = \Big(& \sum_{k \in \mathcal{K}} p_{n[k]} \Big[\Big\{ \bar{Y}_{[k]a} - (\bar{\psi}'_{[k]a}(X_i^\top \hat{\beta}_{[k]\text{glm}}(a)) - \bar{\psi}'_{[k]}(X_i^\top \hat{\beta}_{[k]\text{glm}}(a))) \Big\} \\
& - \Big\{ \bar{Y}_{[k]0} - (\bar{\psi}'_{[k]0}(X_i^\top \hat{\beta}_{[k]\text{glm}}(0)) - \bar{\psi}'_{[k]}(X_i^\top \hat{\beta}_{[k]\text{glm}}(0))) \Big\} \Big] : a \in \mathcal{A} \Big) \\
= \Big(& \sum_{k \in \mathcal{K}} p_{n[k]} \big[\{ \bar{Y}_{[k]a} - (\bar{\psi}'_{[k]a}(X_i^\top \beta_{[k]\text{glm}}(a)) - E(\psi'(X_i^\top \beta_{[k]\text{glm}}(a)) | B_i = k)) \} \\
& - \{ \bar{Y}_{[k]0} - (\bar{\psi}'_{[k]0}(X_i^\top \hat{\beta}_{[k]\text{glm}}(0)) - E(\psi'(X_i^\top \beta_{[k]\text{glm}}(0)) | B_i = k)) \} \big] : a \in \mathcal{A} \Big) \qquad \text{(S3.1)} \\
+ \Big(& \sum_{k \in \mathcal{K}} p_{n[k]} \{ (\bar{\psi}'_{[k]\text{glm}}(X_i^\top \beta_{[k]\text{glm}}(a)) - \bar{\psi}'_{[k]\text{glm}}(X_i^\top \beta_{[k]\text{glm}}(0))) \\
& - (E(\psi'(X_i^\top \beta_{[k]\text{glm}}(a)) | B_i = k) - E(\psi'(X_i^\top \beta_{[k]\text{glm}}(0)) | B_i = k)) \} : a \in \mathcal{A} \Big) \qquad \text{(S3.2)}
\end{aligned}$$

$$
\begin{aligned}
&- \Big(\sum_{k\in\mathcal{K}} p_{n[k]} \big[\{\bar{\psi}'_{[k]a}(X_i^\top \hat{\beta}_{[k]\mathrm{glm}}(a)) - \bar{\psi}'_{[k]}(X_i^\top \hat{\beta}_{[k]\mathrm{glm}}(a))\} \\
&\quad - \{\bar{\psi}'_{[k]a}(X_i^\top \beta_{[k]\mathrm{glm}}(a)) - \bar{\psi}'_{[k]}(X_i^\top \beta_{[k]\mathrm{glm}}(a))\} \big] \\
&\quad - \big[\{\bar{\psi}'_{[k]0}(X_i^\top \hat{\beta}_{[k]\mathrm{glm}}(0)) - \bar{\psi}'_{[k]}(X_i^\top \hat{\beta}_{[k]\mathrm{glm}}(0))\} \\
&\quad - \{\bar{\psi}'_{[k]0}(X_i^\top \beta_{[k]\mathrm{glm}}(0)) - \bar{\psi}'_{[k]}(X_i^\top \beta_{[k]\mathrm{glm}}(0))\} \big] : a \in \mathcal{A} \Big). \qquad \text{(S3.3)}
\end{aligned}
$$

We first analyze the (S3.3). By Lemmas S1 and S2, we have

$$
\|\hat{\beta}_{[k]\mathrm{glm}}(a) - \beta_{[k]\mathrm{glm}}(a)\|_1 = O_p\left(s_{[k]} \sqrt{\frac{M_{n[k]} \log p}{n}} \right)
$$

and

$$
\|\bar{X}_{[k]a} - \bar{X}_{[k]}\|_\infty = O_p\left(\sqrt{\frac{\log p}{n}} \right),
$$

for any $a \in \mathcal{A}_0$ and $k \in \mathcal{K}$. By Taylor expansion, for some $t_i \in (0,1)$, we have

$$
\begin{aligned}
&\psi'(X_i^\top \hat{\beta}_{[k]\mathrm{glm}}(a)) - \psi'(X_i^\top \beta_{[k]\mathrm{glm}}(a)) \\
&= \psi''(X_i^\top \beta_{[k]\mathrm{glm}}(a) + t_i X_i^\top \{\hat{\beta}_{[k]\mathrm{glm}}(a) - \beta_{[k]\mathrm{glm}}(a)\}) X_i^\top (\hat{\beta}_{[k]\mathrm{glm}}(a) - \beta_{[k]\mathrm{glm}}(a)).
\end{aligned}
$$

Denote

$$
v_i = X_i^\top \beta_{[k]\mathrm{glm}}(a) + t_i X_i^\top \{\hat{\beta}_{[k]\mathrm{glm}}(a) - \beta_{[k]\mathrm{glm}}(a)\},
$$

then we have

$$
\begin{aligned}
&\Big| \sqrt{n} \big[\{\bar{\psi}'_{[k]a}(X_i^\top \hat{\beta}_{[k]\mathrm{glm}}(a)) - \bar{\psi}'_{[k]}(X_i^\top \hat{\beta}_{[k]\mathrm{glm}}(a))\} - \{\bar{\psi}'_{[k]a}(X_i^\top \beta_{[k]\mathrm{glm}}(a)) - \bar{\psi}'_{[k]}(X_i^\top \beta_{[k]\mathrm{glm}}(a))\} \big] \Big| \\
&= \sqrt{n} \Big| \frac{1}{n_{[k]a}} \sum_{i\in(a,k)} \psi''(v_i) X_i^\top \{\hat{\beta}_{[k]\mathrm{glm}}(a) - \beta_{[k]\mathrm{glm}}(a)\} - \frac{1}{n_{[k]}} \sum_{i\in[k]} \psi''(v_i) X_i^\top \{\hat{\beta}_{[k]\mathrm{glm}}(a) - \beta_{[k]\mathrm{glm}}(a)\} \Big|
\end{aligned}
$$

$$\leq \sqrt{n}\Big|\frac{1}{n_{[k]a}}\sum_{i\in(a,k)}\psi''(X_i^\top\beta_{[k]\mathrm{glm}}(a))X_i^\top\{\hat{\beta}_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\}$$

$$-\frac{1}{n_{[k]}}\sum_{i\in[k]}\psi''(X_i^\top\beta_{[k]\mathrm{glm}}(a))X_i^\top\{\hat{\beta}_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\}\Big| \tag{S3.4}$$

$$+\sqrt{n}\Big|\frac{1}{n_{[k]a}}\sum_{i\in(a,k)}\{\psi''(v_i)-\psi''(X_i^\top\beta_{[k]\mathrm{glm}}(a))\}X_i^\top\{\hat{\beta}_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\}\Big| \tag{S3.5}$$

$$+\sqrt{n}\Big|\frac{1}{n_{[k]}}\sum_{i\in[k]}\{\psi''(v_i)-\psi''(X_i^\top\beta_{[k]\mathrm{glm}}(a))\}X_i^\top\{\hat{\beta}_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\}\Big|. \tag{S3.6}$$

For (S3.4), by Assumptions 3 and 4, ψ'' and X_i are uniformly bounded, then we can retreat $\check{X}_i=\psi''(X_i^\top\beta_{[k]\mathrm{glm}}(a))X_i$ as a uniformly bounded random variable. Then by Lemmas S1 and S2, we have

$$\begin{aligned}
&\sqrt{n}\Big|\frac{1}{n_{[k]a}}\sum_{i\in(a,k)}\psi''(X_i^\top\beta_{[k]\mathrm{glm}}(a))X_i^\top\{\hat{\beta}_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\}\\
&\qquad-\frac{1}{n_{[k]}}\sum_{i\in[k]}\psi''(X_i^\top\beta_{[k]\mathrm{glm}}(a))X_i^\top\{\hat{\beta}_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\}\Big|\\
&=\sqrt{n}\Big|\frac{1}{n_{[k]a}}\sum_{i\in(a,k)}\check{X}_i^\top\{\hat{\beta}_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\}-\frac{1}{n_{[k]}}\sum_{i\in[k]}\check{X}_i^\top\{\hat{\beta}_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\}\Big|\\
&\leq\sqrt{n}\|\bar{\check{X}}_{[k]a}-\bar{\check{X}}_{[k]}\|_\infty\|\hat{\beta}_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\|_1\\
&=O_p\bigg(s_{[k]}\log p\sqrt{\frac{M_{n[k]}}{n}}\bigg)\\
&=o_p(1).
\end{aligned}$$

Then we analyze (S3.5). By Assumption 4, ψ'' satisfies a Lipschitz condi-

tion, then we have

$$\begin{aligned}
&\sqrt{n}\Big|\frac{1}{n_{[k]a}}\sum_{i\in(a,k)}\{\psi''(v_i)-\psi''(X_i^\top\beta_{[k]\mathrm{glm}}(a))\}X_i^\top\{\hat\beta_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\}\Big| \\
&\quad\le \sqrt{n}\Big[\frac{1}{n_{[k]a}}\sum_{i\in(a,k)}\Big|\psi''(v_i)-\psi''(X_i^\top\beta_{[k]\mathrm{glm}}(a))\Big|\Big|X_i^\top\{\hat\beta_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\}\Big|\Big] \\
&\quad\le \sqrt{n}L\Big[\frac{1}{n_{[k]a}}\sum_{i\in(a,k)}|t_i|\Big|X_i^\top\{\hat\beta_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\}\Big|^2\Big] \\
&\quad\le \sqrt{n}L\{\hat\beta_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\}^\top\Big(\frac{1}{n_{[k]a}}\sum_{i\in(a,k)}X_iX_i^\top\Big)\{\hat\beta_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\}.
\end{aligned}$$

Using similar arguments in Lemma S4, we can show that the restricted largest eigenvalue of the sample covariance matrix can be upper bounded by a constant c_U, provided that $\lambda_{\max}(Q_{[k]XX})\le\kappa_u<+\infty$ in Assumption 5. Therefore, we have

$$\begin{aligned}
&\sqrt{n}\Big|\frac{1}{n_{[k]a}}\sum_{i\in(a,k)}\{\psi''(v_i)-\psi''(X_i^\top\beta_{[k]\mathrm{glm}}(a))\}X_i^\top\{\hat\beta_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\}\Big| \\
&\quad\le \sqrt{n}L\{\hat\beta_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\}^\top\Big(\frac{1}{n_{[k]a}}\sum_{i\in(a,k)}X_iX_i^\top\Big)\{\hat\beta_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\} \\
&\quad\le \sqrt{n}Lc_U\|\hat\beta_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\|_2^2 \\
&\quad= O_p\Big(s_{[k]}\frac{\log pM_{n[k]}}{\sqrt{n}}\Big) \\
&\quad= o_p(1).
\end{aligned}$$

We can apply the same procedure for (S3.6) to derive that

$$\sqrt{n}\Big|\frac{1}{n_{[k]}}\sum_{i\in[k]}\{\psi''(v_i)-\psi''(X_i^\top\beta_{[k]\mathrm{glm}}(a))\}X_i^\top\{\hat\beta_{[k]\mathrm{glm}}(a)-\beta_{[k]\mathrm{glm}}(a)\}\Big|=o_p(1).$$

Then we can obtain that

$$\Big|\sqrt{n}\big[\{\bar{\psi}'_{[k]a}(X_i^\top\hat{\beta}_{[k]\mathrm{glm}}(a))-\bar{\psi}'_{[k]}(X_i^\top\hat{\beta}_{[k]\mathrm{glm}}(a))\}$$
$$-\{\bar{\psi}'_{[k]a}(X_i^\top\beta_{[k]\mathrm{glm}}(a))-\bar{\psi}'_{[k]}(X_i^\top\beta_{[k]\mathrm{glm}}(a))\}\big]\Big|=o_p(1).$$

Then we deal with (S3.1) and (S3.2). Simple calculations show that

$$E(\psi'(X_i^\top\beta_{[k]\mathrm{glm}}(a))|B_i=k)=\frac{1}{n_{[k]}}\sum_{i=1}^{n}I(B_i=k)E(\psi'(X_i^\top\beta_{[k]\mathrm{glm}}(a))|B_i=k)$$
$$=\frac{1}{n_{[k]a'}}\sum_{i=1}^{n}I(A_i=a',B_i=k)E(\psi'(X_i^\top\beta_{[k]\mathrm{glm}}(a))|B_i=k).$$

Recall that

$$\tilde{r}_i(a)=r_i(a)-E\{r_i(a)|B_i=k\},$$

where $r_i(a)$ is a transformed outcome, which can be $Y_i(a)$, $\psi'(X_i^\top\beta_{[k]\mathrm{glm}}(a))$ or $r_{i,[k]\mathrm{glm}}(a)$. Define

$$\mathbb{L}^{(1)}_{n[k],r_{i,\mathrm{glm}}(a)}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}I(A_i=a,B_i=k)\tilde{r}_{i,[k]\mathrm{glm}}(a),$$
$$\mathbb{L}^{(1)}_{n[k],\psi(a)}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}I(B_i=k)\tilde{\psi}'(X_i^\top\beta_{[k]\mathrm{glm}}(a)),$$
$$\mathbb{L}^{(2)}_{n[k]}=\sqrt{n}(p_{[k]}-p_{n[k]}).$$

Some calculation for the first two parts of $\hat{\tau}_{\mathrm{glm}}$ leads to

$$\sqrt{n}(\hat{\tau}_{\mathrm{glm}}-\tau)=\Big\{\sum_{k\in\mathcal{K}}\Big(\frac{1}{\pi_{[k]a}}\mathbb{L}^{(1)}_{n[k],r_{i,\mathrm{glm}}(a)}-\frac{1}{\pi_{[k]0}}\mathbb{L}^{(1)}_{n[k],r_{i,\mathrm{glm}}(0)}\Big),a\in\mathcal{A}\Big\}$$
$$+\Big\{\sum_{k\in\mathcal{K}}\mathbb{L}^{(1)}_{n[k],\psi(a)}-\mathbb{L}^{(1)}_{n[k],\psi(0)},a\in\mathcal{A}\Big\}$$
$$+\Big\{\sum_{k\in\mathcal{K}}\mathbb{L}^{(2)}_{n[k]}\{m_{[k]}(a)-m_{[k]}(0)\},a\in\mathcal{A}\Big\}+o_p(1),$$

where $m_{[k]}(a) = E[Y_i(a) \mid B_i = k] - E[Y_i(a)]$. By Lemma S3 and some additional calculations, we have

$$V^{-1/2}\begin{pmatrix} \{\sum_{k\in\mathcal{K}}(\frac{1}{\pi_{[k]a}}\mathbb{L}^{(1)}_{n[k],r_{i,\mathrm{glm}}(a)} - \frac{1}{\pi_{[k]0}}\mathbb{L}^{(1)}_{n[k],r_{i,\mathrm{glm}}(0)}), a\in\mathcal{A}\} \\ \{\sum_{k\in\mathcal{K}}\mathbb{L}^{(1)}_{n[k],\psi(a)} - \mathbb{L}^{(1)}_{n[k],\psi(0)}, a\in\mathcal{A}\} \\ \{\sum_{k\in\mathcal{K}}\mathbb{L}^{(2)}_{n[k]}\{m_{[k]}(a) - m_{[k]}(0)\}, a\in\mathcal{A}\} \end{pmatrix} \xrightarrow{d} \mathrm{N}\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix} I_A & 0 & 0 \\ 0 & I_A & 0 \\ 0 & 0 & I_A \end{pmatrix}\right),$$

where

$$V = \begin{pmatrix} V_{\tilde{r}_{\mathrm{glm}}} & V_{\tilde{r}_{\mathrm{glm}}\tilde{\psi}_{a0}} & 0 \\ V^\top_{\tilde{r}_{\mathrm{glm}}\tilde{\psi}_{a0}} & V_{\tilde{\psi}_{a0}} & 0 \\ 0 & 0 & V_{H\tilde{Y}} \end{pmatrix},$$

and

$$\begin{aligned}
V_{H\tilde{Y}} &= \sum_{k\in\mathcal{K}} p_{[k]}\left(m_{[k]}(a) - m_{[k]}(0) : a\in\mathcal{A}\right)\times\left(m_{[k]}(a') - m_{[k]}(0) : a'\in\mathcal{A}\right)^\top, \\
V_{\tilde{r}_{\mathrm{glm}}} &= \sum_{k\in\mathcal{K}}\frac{p_{[k]}}{\pi_{[k]0}}\sigma^2_{[k]\tilde{r}_{\mathrm{glm}}(0)}1_A1_A^\top + \mathrm{diag}\left\{\sum_{k\in\mathcal{K}}\frac{p_{[k]}}{\pi_{[k]a}}\sigma^2_{[k]\tilde{r}_{\mathrm{glm}}(a)} : a\in\mathcal{A}\right\}, \\
V_{\tilde{\psi}_{a0}} &= \Big\{\sum_{k\in\mathcal{K}} p_{[k]}\mathrm{Cov}\{\psi'(X_i^\top\beta_{[k]\mathrm{glm}}(a)) - \psi'(X_i^\top\beta_{[k]\mathrm{glm}}(0)), \psi'(X_i^\top\beta_{[k]\mathrm{glm}}(a')) \\
&\qquad - \psi'(X_i^\top\beta_{[k]\mathrm{glm}}(0))\} : (a,a')\in\mathcal{A}\times\mathcal{A}\Big\}, \\
V_{\tilde{r}_{\mathrm{glm}}\tilde{\psi}_{a0}} &= \Big\{\sum_{k\in\mathcal{K}} p_{[k]}\mathrm{Cov}\big[\{Y_i(a) - Y_i(0)\} - \{\psi'(X_i^\top\beta_{[k]\mathrm{glm}}(a)) - \psi'(X_i^\top\beta_{[k]\mathrm{glm}}(0))\}, \\
&\qquad \psi'(X_i^\top\beta_{[k]\mathrm{glm}}(a') - \psi'(X_i^\top\beta_{[k]\mathrm{glm}}(0))\big] : (a,a')\in\mathcal{A}\times\mathcal{A}\Big\}.
\end{aligned}$$

Let $V_{\mathrm{glm}} = V_{H\tilde{Y}} + V_{\tilde{r}_{\mathrm{glm}}} + V_{\tilde{\psi}_{a0}} + V_{\tilde{r}_{\mathrm{glm}}\tilde{\psi}_{a0}} + V^\top_{\tilde{r}_{\mathrm{glm}}\tilde{\psi}_{a0}}$, we have

$$\sqrt{n}\; V_{\mathrm{glm}}^{-1/2}(\hat{\tau}_{\mathrm{glm}} - \tau) \xrightarrow{d} \mathrm{N}(0, I_A).$$

S3.2 Proof of Theorem 1

Transfer Step: Analog to $\hat{L}_{[k]a}(\beta_{[k]\text{glm}}(a))$ defined in the proof of Lemma S1 , we define

$$\hat{L}^{(s)}_{[k]a}(\beta_{[k]\text{glm}}(a)) = \frac{1}{n^{(s)}_{[k]a}} \sum_{i=1}^{n^{(s)}} I(A^{(s)}_i = a, B^{(s)}_i = k)\{\psi(X^{(s)\top}_i \beta) - Y^{(s)}_i X^{(s)\top}_i \beta\}.$$

By the definition of $\hat{\beta}^{(s)}_{[k]\text{glm}}(a)$, we have

$$\hat{L}^{(s)}_{[k]a}(\hat{\beta}^{(s)}_{[k]\text{glm}}(a)) + \lambda^{(s)}_{[k]a}\|\hat{\beta}^{(s)}_{[k]\text{glm}}(a)\|_1 \leq \hat{L}^{(s)}_{[k]a}(\beta^{(s)}_{[k]\text{glm}}(a)) + \lambda^{(s)}_{[k]a}\|\beta^{(s)}_{[k]\text{glm}}(a)\|_1.$$

Let $\Delta^{(s)} = \hat{\beta}^{(s)}_{[k]\text{glm}}(a) - \beta^{(s)}_{[k]\text{glm}}(a)$. Similar to the proof for Lemma S1, we define,

$$\delta\hat{L}^{(s)}_{[k]a}(\Delta) = \hat{L}^{(s)}_{[k]a}(\beta^{(s)}_{[k]\text{glm}}(a)+\Delta^{(s)}) - \hat{L}^{(s)}_{[k]a}(\beta^{(s)}_{[k]\text{glm}}(a)) - \{\nabla\hat{L}^{(s)}(\beta^{(s)}_{[k]\text{glm}}(a)\}^\top \Delta^{(s)},$$

where $\nabla\hat{L}^{(s)}(\beta^{(s)}_{[k]\text{glm}}(a))$ is the gradient of $\hat{L}^{(s)}_{[k]a}(\beta^{(s)}_{[k]\text{glm}}(a))$ with respect to $\beta^{(s)}_{[k]\text{glm}}(a)$. Then we can go through the proof until we have

$$\delta\hat{L}^{(s)}_{[k]a}(\Delta^{(s)}) \leq \frac{3}{2}\lambda^{(s)}_{[k]a}\|[\Delta^{(s)}]_{S_{[k]}}\|_1 + 2\lambda^{(s)}_{[k]a}\|[\beta^{(s)}_{[k]\text{glm}}(a)]_{S^C_{[k]}}\|_1 - \frac{1}{2}\lambda^{(s)}_{[k]a}\|[\Delta^{(s)}]_{S^C_{[k]}}\|_1.$$

Then we compare the first term and the second term on the RHS.

Case 1 :

$$\frac{3}{2}\lambda^{(s)}_{[k]a}\|[\Delta^{(s)}]_{S_{[k]}}\|_1 \geq 2\lambda^{(s)}_{[k]a}\|[\beta^{(s)}_{[k]\text{glm}}(a)]_{S^C_{[k]}}\|_1.$$

Then we have $\|[\Delta^{(s)}]_{S^C_{[k]}}\|_1 \leq 6\|[\Delta^{(s)}]_{S_{[k]}}\|_1$. By Lemma S4, we have

$$c_0\|\Delta^{(s)}\|_2^2 \leq 3\lambda^{(s)}_{[k]a}\|[\Delta^{(s)}]_{S_{[k]}}\|_1 \leq 3\sqrt{s_{[k]}}\lambda^{(s)}_{[k]a}\|[\Delta^{(s)}]_{S_{[k]}}\|_2,$$

which leads to

$$\|\Delta^{(s)}\|_2 \leq C_1\sqrt{s_{[k]}}\lambda^{(s)}_{[k]a}$$

and

$$\|\Delta^{(s)}\|_1 \leq 7\|[\Delta^{(s)}]_{S_{[k]}}\|_1 \leq 7\sqrt{s_{[k]}}\|[\Delta^{(s)}]_{S_{[k]}}\|_2 \leq C_2 s_{[k]}\lambda^{(s)}_{[k]a}.$$

Case 2 :

$$\frac{3}{2}\lambda^{(s)}_{[k]a}\|[\Delta^{(s)}]_{S_{[k]}}\|_1 < 2\lambda^{(s)}_{[k]a}\|[\beta^{(s)}_{[k]\mathrm{glm}}(a)]_{S^C_{[k]}}\|_1.$$

Then

$$\delta\hat{L}^{(s)}_{[k]a}(\Delta^{(s)}) \leq 4\lambda^{(s)}_{[k]a}\|[\beta^{(s)}_{[k]\mathrm{glm}}(a)]_{S^C_{[k]}}\|_1 - \frac{1}{2}\lambda^{(s)}_{[k]a}\|[\Delta^{(s)}]_{S^C_{[k]}}\|_1.$$

A direct bound for $\|\Delta^{(s)}\|_1$ is

$$\|\Delta^{(s)}\|_1 \leq \|[\Delta^{(s)}]_{S_{[k]}}\|_1 + \|[\Delta^{(s)}]_{S^C_{[k]}}\|_1 \leq 9\|[\beta^{(s)}_{[k]\mathrm{glm}}(a)]_{S^C_{[k]}}\|_1 \leq 9\|[\delta_{[k]}(a)]_{S^C_{[k]}}\|_1 \leq 9h.$$

The third inequality holds as

$$\|[\beta^{(s)}_{[k]\mathrm{glm}}(a)]_{S^C_{[k]}}\|_1 \leq \|[\beta_{[k]\mathrm{glm}}(a)]_{S^C_{[k]}}\|_1 + \|[\delta_{[k]}(a)]_{S^C_{[k]}}\|_1 = \|[\delta_{[k]}(a)]_{S^C_{[k]}}\|_1.$$

Here we can also drive a direct bound for $\|\Delta^{(s)}\|_2$

$$\|\Delta^{(s)}\|_2 \leq \|\Delta^{(s)}\|_1 \leq 9h.$$

Therefore, we have

$$\|\Delta^{(s)}\|_1 \leq C_2 s_{[k]}\lambda^{(s)}_{[k]a} + C_4 h.$$

and

$$\|\Delta^{(s)}\|_2 \leq C_1\sqrt{s_{[k]}}\lambda^{(s)}_{[k]a} + C_4 h.$$

Debiasing Step:

Recall that in the proof of Lemma S1, we define

$$\hat{L}_{[k]a}(\beta) = \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)\{\psi(X_i^\top \beta) - Y_i X_i^\top \beta\}.$$

By the definition of $\hat{\delta}_{[k]}(a)$ and $\hat{\beta}_{[k]\mathrm{tl}}(a) = \hat{\beta}^{(s)}_{[k]\mathrm{glm}}(a) + \hat{\delta}_{[k]a}$, we have

$$\hat{L}_{[k]a}(\hat{\beta}_{[k]\mathrm{tl}}(a)) + \lambda_{[k]a}\|\hat{\delta}_{[k]}(a)\|_1 \leq \hat{L}_{[k]a}(\beta_{[k]\mathrm{glm}}(a)) + \lambda_{[k]a}\|\beta_{[k]\mathrm{glm}}(a) - \hat{\beta}^{(s)}_{[k]\mathrm{glm}}(a)\|_1. \tag{S3.7}$$

Let $\Delta = \hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{glm}}(a)$ and

$$\delta\hat{L}_{[k]a}(\Delta) = \hat{L}_{[k]a}(\hat{\beta}_{[k]\mathrm{tl}}(a)) - \hat{L}_{[k]a}(\beta_{[k]\mathrm{glm}}(a)) - \nabla L(\beta_{[k]\mathrm{glm}}(a))^\top \Delta.$$

Also, let

$$E_2 = \left\{ \|\{\nabla \hat{L}(\beta_{[k]\mathrm{glm}}(a))\}\|_\infty \leq \frac{\lambda_{[k]a}}{2} \right\} \cap E_1^{(s)}.$$

where $E_1^{(s)}$ is analog to E_1 defined in the proof of Lemma S1, which is

$$E_1^{(s)} = \left\{ \|\nabla \hat{L}^{(s)}(\beta^{(s)}_{[k]\mathrm{glm}}(a))\|_\infty \leq \frac{1}{2}\lambda^{(s)}_{[k]a} \right\}.$$

By Lemma S5 we have

$$\Pr(E_2) \geq 1 - \frac{1}{M_{n[k]}} \to 1.$$

Under E_2, we have

$$\begin{aligned}
\delta\hat{L}_{[k]a}(\Delta) &\leq \lambda_{[k]a}(\|\beta_{[k]\mathrm{glm}}(a) - \hat{\beta}^{(s)}_{[k]\mathrm{glm}}(a)\|_1 - \|\hat{\delta}_{[k]}(a)\|_1) - \nabla L(\beta_{[k]\mathrm{glm}}(a))^\top \Delta \\
&\leq \lambda_{[k]a}(\|\beta_{[k]\mathrm{glm}}(a) - \hat{\beta}^{(s)}_{[k]\mathrm{glm}}(a)\|_1 - \|\hat{\delta}_{[k]}(a)\|_1) + \|\nabla L(\beta_{[k]\mathrm{glm}}(a))\|_\infty \|\Delta\|_1
\end{aligned}$$

$$\leq \lambda_{[k]a}(\|\beta_{[k]\text{glm}}(a) - \hat{\beta}^{(s)}_{[k]\text{glm}}(a)\|_1 - \|\hat{\delta}_{[k]}(a)\|_1) + \frac{\lambda_{[k]a}}{2}\|\Delta\|_1.$$

By the triangle inequality, we have

$$\|\Delta\|_1 \leq \|\beta_{[k]\text{glm}}(a) - \hat{\beta}^{(s)}_{[k]\text{glm}}(a)\|_1 + \|\hat{\delta}_{[k]}(a)\|_1,$$

$$\|\beta_{[k]\text{glm}}(a) - \hat{\beta}^{(s)}_{[k]\text{glm}}(a)\|_1 = \|\delta_{[k]}(a) + \Delta^{(s)}\|_1 \leq \|\delta_{[k]}(a)\|_1 + \|\Delta^{(s)}\|_1.$$

Then we have

$$\begin{aligned}\delta\hat{L}_{[k]a}(\Delta) &\leq \lambda_{[k]a}(\|\beta_{[k]\text{glm}}(a) - \hat{\beta}^{(s)}_{[k]\text{glm}}(a)\|_1 - \|\hat{\delta}_{[k]}(a)\|_1) + \frac{\lambda_{[k]a}}{2}\|\Delta\|_1 \\ &\leq \frac{3}{2}\lambda_{[k]a}\|\delta_{[k]}(a)\|_1 - \frac{1}{2}\lambda_{[k]a}\|\hat{\delta}_{[k]}(a)\|_1 + \frac{3}{2}\lambda_{[k]a}\|\Delta^{(s)}\|_1. \qquad \text{(S3.8)}\end{aligned}$$

By Taylor expansion and $\psi'' > 0$, $\delta\hat{L}_{[k]a}(\Delta) \geq 0$, we have

$$\|\hat{\delta}_{[k]}(a)\|_1 \leq 3\|\delta_{[k]}(a)\|_1 + 3\|\Delta^{(s)}\|_1 \leq 3h + 3\|\Delta^{(s)}\|_1.$$

Therefore,

$$\begin{aligned}\|\Delta\|_1 &= \|\hat{\beta}^{(s)}_{[k]\text{glm}}(a) - \beta^{(s)}_{[k]\text{glm}}(a) + \hat{\delta}_{[k]}(a) - \delta_{[k]}(a)\|_1 \\ &\leq \|\Delta^{(s)}\|_1 + \|\hat{\delta}_{[k]}(a)\|_1 + \|\delta_{[k]}(a)\|_1 \\ &\leq 4\|\Delta^{(s)}\|_1 + 4h \\ &\leq C_5 s_{[k]}\sqrt{\frac{M^{(s)}_{n[k]}\log p}{n + n^{(s)}}} + C_6 h.\end{aligned}$$

By Assumption 5 and Proposition 2 in Negahban et al. (2012) and using the same coupling argument in the proof of Lemma S4, with probability

$1 - c_1 \exp(-c_2 n)$, there exist positive constants κ_1 and κ_2 such that

$$\delta \hat{L}_{[k]a}(\Delta) \geq \kappa_1 \|\Delta\|_2^2 - \kappa_1 \kappa_2 \sqrt{\frac{\log p}{n_{[k]a}}} \|\Delta\|_1 \|\Delta\|_2 \geq \frac{\kappa_1}{2} \|\Delta\|_2^2 - \frac{\kappa_1 \kappa_2^2}{2} \frac{\log p}{n_{[k]a}} \|\Delta\|_1^2.$$

The last inequality is derived by basic inequality

$$\kappa_2 \sqrt{\frac{\log p}{n_{[k]a}}} \|\Delta\|_1 \|\Delta\|_2 \leq \frac{\|\Delta\|_2^2}{2} + \frac{\kappa_2^2}{2} \frac{\log p}{n_{[k]a}} \|\Delta\|_1^2.$$

Then in (S3.8), we have

$$\kappa_1 \|\Delta\|_2^2 - \kappa_1 \kappa_2^2 \frac{\log p}{n_{[k]a}} \|\Delta\|_1^2 \leq \frac{3}{2} \lambda_{[k]a} \|\delta_{[k]}(a)\|_1 - \frac{1}{2} \lambda_{[k]a} \|\hat{\delta}_{[k]}(a)\|_1 + \frac{3}{2} \lambda_{[k]a} \|\Delta^{(s)}\|_1.$$

Because $\lambda_{[k]a}$ has the order of $O\{(M_{n[k]} \log p/n)^{1/2}\}$ and $\|\Delta\|_1$ has the same order with $\|\Delta^{(s)}\|_1$, we have

$$\begin{aligned} \kappa_1 \|\Delta\|_2^2 &\leq \kappa_1 \kappa_2^2 \frac{\log p}{n_{[k]a}} \|\Delta\|_1^2 + \frac{3}{2} \lambda_{[k]a} \|\delta_{[k]}(a)\|_1 + \frac{3}{2} \lambda_{[k]a} \|\Delta^{(s)}\|_1 \\ &= O_p\left(h \sqrt{\frac{M_{n[k]} \log p}{n}} + \frac{s_{[k]} \log p \sqrt{M_{n[k]} M_{n[k]}^{(s)}}}{\sqrt{n n^{(s)}}} \right). \end{aligned} \tag{S3.9}$$

S3.3 Proof of Theorem 2

Like the Proof of Proposition 1, we can decompose $\hat{\tau}_{\text{tl}}$ as we did for $\hat{\tau}_{\text{glm}}$. We only need to show that

$$\begin{aligned} \sqrt{n} \big[\{ \bar{\psi}'_{[k]a}(X_i^\top \hat{\beta}_{[k]\text{tl}}(a)) - \bar{\psi}'_{[k]}(X_i^\top \hat{\beta}_{[k]\text{tl}}(a)) \} \\ - \{ \bar{\psi}'_{[k]a}(X_i^\top \beta_{[k]\text{glm}}(a)) - \bar{\psi}'_{[k]}(X_i^\top \beta_{[k]\text{glm}}(a)) \} \big] = o_p(1) \end{aligned}$$

for each $a \in \mathcal{A}_0$ and $k \in \mathcal{K}$ to finish the proof. Denote

$$q_i = X_i^\top \beta_{[k]\mathrm{glm}}(a) + t_i X_i^\top \{\hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{glm}}(a)\},$$

then similar to the proof of Proposition 1, we have

$$\begin{aligned}
&\Big|\sqrt{n}\big[\{\bar{\psi}'_{[k]a}(X_i^\top \hat{\beta}_{[k]\mathrm{tl}}(a)) - \bar{\psi}'_{[k]}(X_i^\top \hat{\beta}_{[k]\mathrm{tl}}(a))\} - \{\bar{\psi}'_{[k]a}(X_i^\top \beta_{[k]\mathrm{glm}}(a)) - \bar{\psi}'_{[k]}(X_i^\top \beta_{[k]\mathrm{glm}}(a))\}\big]\Big| \\
&= \sqrt{n}\Big|\frac{1}{n_{[k]a}} \sum_{i\in(a,k)} \psi''(q_i) X_i^\top \{\hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{glm}}(a)\} - \frac{1}{n_{[k]}} \sum_{i\in[k]} \psi''(q_i) X_i^\top \{\hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{glm}}(a)\}\Big| \\
&\le \sqrt{n}\Big|\frac{1}{n_{[k]a}} \sum_{i\in(a,k)} \psi''(X_i^\top \beta_{[k]\mathrm{glm}}(a)) X_i^\top \{\hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{glm}}(a)\} \\
&\qquad - \frac{1}{n_{[k]}} \sum_{i\in[k]} \psi''(X_i^\top \beta_{[k]\mathrm{glm}}(a)) X_i^\top \{\hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{glm}}(a)\}\Big| \qquad \text{(S3.10)} \\
&\quad + \sqrt{n}\Big|\frac{1}{n_{[k]a}} \sum_{i\in(a,k)} \{\psi''(q_i) - \psi''(X_i^\top \beta_{[k]\mathrm{glm}}(a))\} X_i^\top \{\hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{glm}}(a)\}\Big| \qquad \text{(S3.11)} \\
&\quad + \sqrt{n}\Big|\frac{1}{n_{[k]}} \sum_{i\in[k]} \{\psi''(q_i) - \psi''(X_i^\top \beta_{[k]\mathrm{glm}}(a))\} X_i^\top \{\hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{glm}}(a)\}\Big|. \qquad \text{(S3.12)}
\end{aligned}$$

Formula (S3.10) is similar with (S3.4), which has the order of $O_p\{h(\log p)^{1/2} + (s_{[k]} \log p)(M^{(s)}_{n[k]}/n^{(s)})^{1/2}\}$. Then we analyze (S3.11). By Assumption 4, ψ'' satisfies a Lipschitz condition, then we have

$$\begin{aligned}
&\sqrt{n}\Big|\frac{1}{n_{[k]a}} \sum_{i\in(a,k)} \{\psi''(q_i) - \psi''(X_i^\top \beta_{[k]\mathrm{glm}}(a))\} X_i^\top \{\hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{glm}}(a)\}\Big| \\
&\qquad \le \sqrt{n}\Big[\frac{1}{n_{[k]a}} \sum_{i\in(a,k)} \Big|\psi''(q_i) - \psi''(X_i^\top \beta_{[k]\mathrm{glm}}(a))\Big| \Big|X_i^\top \{\hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{glm}}(a)\}\Big|\Big]
\end{aligned}$$

$$
\begin{aligned}
&\leq \sqrt{n}L\Big[\frac{1}{n_{[k]a}}\sum_{i\in(a,k)}|t_i|\Big|X_i^\top\{\hat{\beta}_{[k]\mathrm{tl}}(a)-\beta_{[k]\mathrm{glm}}(a)\}\Big|^2\Big]\\
&\leq \sqrt{n}L\{\hat{\beta}_{[k]\mathrm{tl}}(a)-\beta_{[k]\mathrm{glm}}(a)\}^\top\Big(\frac{1}{n_{[k]a}}\sum_{i\in(a,k)}X_iX_i^\top\Big)\{\hat{\beta}_{[k]\mathrm{tl}}(a)-\beta_{[k]\mathrm{glm}}(a)\}\\
&\leq \sqrt{n}Lc_U\|\hat{\beta}_{[k]\mathrm{tl}}(a)-\beta_{[k]\mathrm{glm}}(a)\|_2^2\\
&= O_p\left(s_{[k]}\log p\sqrt{\frac{M_{n[k]}M_{n[k]}^{(s)}}{n^{(s)}}}+h\sqrt{M_{n[k]}\log p}\right).
\end{aligned}
$$

Here the second-to-last equation is derived from equation (S3.9) in the proof of Theorem 1. Then, as $n^{(s)} > n$ and $M_{n[k]}^{(s)}$ has the same diverging function form as $M_{n[k]}$, e.g., $M_{n[k]}^{(s)} = \log\log n^{(s)}$ and $M_{n[k]} = \log\log n$, we have $M_{n[k]}^{(s)} > M_{n[k]}$. Therefore, by Assumption 9, we have

$$
s_{[k]}\log p\sqrt{\frac{M_{n[k]}M_{n[k]}^{(s)}}{n^{(s)}}} \leq s_{[k]}\log p\frac{M_{n[k]}^{(s)}}{\sqrt{n^{(s)}}} = o(1).
$$

Combining the condition $h\sqrt{M_{n[k]}\log p} = o(1)$, we have

$$
\sqrt{n}\Big|\frac{1}{n_{[k]a}}\sum_{i\in(a,k)}\{\psi''(q_i)-\psi''(X_i^\top\beta_{[k]\mathrm{glm}}(a))\}X_i^\top\{\hat{\beta}_{[k]\mathrm{tl}}(a)-\beta_{[k]\mathrm{glm}}(a)\}\Big| = o_P(1).
$$

Then we can apply the same procedure to (S3.12) to have

$$
\begin{aligned}
&\sqrt{n}\Big|\frac{1}{n_{[k]}}\sum_{i\in[k]}\{\psi''(q_i)-\psi''(X_i^\top\beta_{[k]\mathrm{glm}}(a))\}X_i^\top\{\hat{\beta}_{[k]\mathrm{tl}}(a)-\beta_{[k]\mathrm{glm}}(a)\}\Big|\\
&= o_P(1).
\end{aligned}
$$

Then we have

$$
\sqrt{n}\big[\{\bar{\psi}'_{[k]a}(X_i^\top\hat{\beta}_{[k]\mathrm{tl}}(a))-\bar{\psi}'_{[k]}(X_i^\top\hat{\beta}_{[k]\mathrm{tl}}(a))\}
$$

$$- \{\bar{\psi}'_{[k]a}(X_i^\top \beta_{[k]\mathrm{glm}}(a)) - \bar{\psi}'_{[k]}(X_i^\top \beta_{[k]\mathrm{glm}}(a))\}\big] = o_P(1)$$

for each $a \in \mathcal{A}_0$ and $k \in \mathcal{K}$. Therefore, we can follow the proof of Proposition 1 and obtain

$$\sqrt{n}\, V_{\mathrm{glm}}^{-1/2}(\hat{\tau}_{\mathrm{tl}} - \tau) \xrightarrow{d} \mathrm{N}(0, I_A).$$

S3.4 Proof of Theorem 3

We only prove that $\sqrt{n}(\hat{\tau}_{\mathrm{glm}}^{\mathrm{proj}} - \tau) \xrightarrow{d} \mathrm{N}(0, V_{\mathrm{glm}}^{\mathrm{proj}})$. The proof of the asymptotic normality for $\hat{\tau}_{\mathrm{tl}}^{\mathrm{proj}}$ is the same as we can use Theorem 2 instead of Proposition 1 for $\hat{\beta}_{[k]\mathrm{tl}}(a)$. We first write $\hat{\tau}_{\mathrm{glm}}^{\mathrm{proj}}$ as

$$\hat{\tau}_{\mathrm{glm,a}}^{\mathrm{proj}} = \sum_{k\in\mathcal{K}} p_{n[k]} \Bigg[\Big\{ \bar{Y}_{[k]a} - \{\bar{\hat{\Psi}}_{[k]a}(X_i) - \bar{\hat{\Psi}}_{[k]}(X_i)\}^\top \hat{\beta}_{[k]}^{\mathrm{proj}}(a) \Big\} - \Big\{ \bar{Y}_{[k]0} - \{\bar{\hat{\Psi}}_{[k]0}(X_i) - \bar{\hat{\Psi}}_{[k]}(X_i)\}^\top \hat{\beta}_{[k]}^{\mathrm{proj}}(0) \Big\} \Bigg].$$

Then we have

$$\begin{aligned}
\hat{\tau}_{\mathrm{glm}}^{\mathrm{proj}} &= \Big(\sum_{k\in\mathcal{K}} p_{n[k]} \Big[\Big\{ \bar{Y}_{[k]a} - \{\bar{\hat{\Psi}}_{[k]a}(X_i) - \bar{\hat{\Psi}}_{[k]}(X_i)\}^\top \hat{\beta}_{[k]}^{\mathrm{proj}}(a) \Big\} \\
&\quad - \Big\{ \bar{Y}_{[k]0} - \{\bar{\hat{\Psi}}_{[k]0}(X_i) - \bar{\hat{\Psi}}_{[k]}(X_i)\}^\top \hat{\beta}_{[k]}^{\mathrm{proj}}(0) \Big\} \Big] : a \in \mathcal{A} \Big) \\
&= \Big(\sum_{k\in\mathcal{K}} p_{n[k]} \big[\{\bar{Y}_{[k]a} - [\{\bar{\Psi}_{[k]a}(X_i) - E\{\Psi_{[k]}(X_i)|B_i = k\}]^\top \beta_{[k]}^{\mathrm{proj}}(a)\} \\
&\quad - \{\bar{Y}_{[k]0} - [\{\bar{\Psi}_{[k]0}(X_i) - E\{\Psi_{[k]}(X_i)|B_i = k\}]^\top \beta_{[k]}^{\mathrm{proj}}(0)\} \big] : a \in \mathcal{A} \Big) \\
&\quad + \Big(\sum_{k\in\mathcal{K}} p_{n[k]} \big[\bar{\Psi}_{[k]}(X_i) - E\{\Psi_{[k]}(X_i)|B_i = k\} \big]^\top \{\beta_{[k]}^{\mathrm{proj}}(a) - \beta_{[k]}^{\mathrm{proj}}(0)\} : a \in \mathcal{A} \Big)
\end{aligned}$$

$$- \Big(\sum_{k \in \mathcal{K}} p_{n[k]} \big\{ [\{\bar{\hat{\Psi}}_{[k]a}(X_i) - \bar{\hat{\Psi}}_{[k]}(X_i)\}^\top \hat{\beta}^{\text{proj}}_{[k]}(a) - \{\bar{\Psi}_{[k]a}(X_i) - \bar{\Psi}_{[k]}(X_i)\}^\top \beta^{\text{proj}}_{[k]}(a)]$$
$$- [\{\bar{\hat{\Psi}}_{[k]0}(X_i) - \bar{\hat{\Psi}}_{[k]}(X_i)\}^\top \hat{\beta}^{\text{proj}}_{[k]}(0) - \{\bar{\Psi}_{[k]0}(X_i) - \bar{\Psi}_{[k]}(X_i)\}^\top \beta^{\text{proj}}_{[k]}(0)] \big\} : a \in \mathcal{A} \Big).$$

We first consider the last part, which we aim to show that it is $o_p(1/\sqrt{n})$.

Let $\delta^{\psi}(X_i, a) = \psi'\{X_i^\top \hat{\beta}_{[k]\text{glm}}(a)\} - \psi'\{X_i^\top \beta_{[k]\text{glm}}(a)\}$ and

$$\delta^{\Psi}(X_i) = (\delta^{\Psi}(X_i, 0), \delta^{\Psi}(X_i, 1), \ldots, \delta^{\Psi}(X_i, A))^\top.$$

For any treatment $a \in \mathcal{A}$, the last part can be written as

$$\text{LP} = \Big[\{\bar{\delta}^{\Psi}_{[k]a}(X_i) - \bar{\delta}^{\Psi}_{[k]}(X_i)\}^\top \{\hat{\beta}^{\text{proj}}_{[k]}(a) - \beta^{\text{proj}}_{[k]}(a)\}$$
$$+ \{\bar{\Psi}_{[k]a}(X_i) - \bar{\Psi}_{[k]}(X_i)\}^\top \{\hat{\beta}^{\text{proj}}_{[k]}(a) - \beta^{\text{proj}}_{[k]}(a)\} + \{\bar{\delta}^{\Psi}_{[k]a}(X_i) - \bar{\delta}^{\Psi}_{[k]}(X_i)\}^\top \beta^{\text{proj}}_{[k]}(a) \Big]. \tag{S3.13}$$

By Proposition 1 and Lemma S6, we have already obtained that

$$\hat{\beta}^{\text{proj}}_{[k]}(a) \xrightarrow{P} \beta^{\text{proj}}_{[k]}(a),$$
$$\sqrt{n}\{\bar{\delta}^{\Psi}_{[k]a}(X_i) - \bar{\delta}^{\Psi}_{[k]}(X_i)\} = o_p(1),$$

for each $k \in \mathcal{K}$ and $a \in \mathcal{A}_0$. Therefore, we have

$$\sqrt{n}\Big[\{\bar{\delta}^{\Psi}_{[k]a}(X_i) - \bar{\delta}^{\Psi}_{[k]}(X_i)\}^\top \{\hat{\beta}^{\text{proj}}_{[k]}(a) - \beta^{\text{proj}}_{[k]}(a)\} \Big] = o_p(1),$$
$$\sqrt{n}\Big[\{\bar{\delta}^{\Psi}_{[k]a}(X_i) - \bar{\delta}^{\Psi}_{[k]}(X_i)\}^\top \beta^{\text{proj}}_{[k]}(a) \Big] = o_p(1).$$

Because treatments are finite, we can deal with each term element-wise to obtain the first and third parts of (S3.13) are $o_p(1/\sqrt{n})$. For the second

term in (S3.13), we take $\Psi\{X_i\}I(B_i = k)$ as the transformed outcome. By Lemma B.1 in Bugni (2019), we have

$$\sqrt{n}\Big\{\bar{\Psi}_{[k]a}(X_i) - \bar{\Psi}_{[k]}(X_i)\Big\} = O_p(1), \text{ for } a \in \mathcal{A}_0$$

and hence

$$\sqrt{n}\{\bar{\Psi}_{[k]a}(X_i) - \bar{\Psi}_{[k]}(X_i)\}^\top\{\hat{\beta}^{\text{proj}}_{[k]}(a) - \beta^{\text{proj}}_{[k]}(a)\} = o_p(1).$$

Then we deal with the first two parts in $\hat{\tau}^{\text{proj}}_{\text{glm}}$. Using similar arguments in the proof of Proposition 1, we define

$$\begin{aligned}
\mathbb{L}^{(1)}_{n[k],r^{\text{proj}}_{i,\text{glm}}(a)} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I(A_i = a, B_i = k)\tilde{r}^{\text{proj}}_{i,\text{glm}}(a), \\
\mathbb{L}^{(1)}_{n[k],\Psi(a)} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I(B_i = k)\tilde{\Psi}^\top_{[k]}(X_i)\beta^{\text{proj}}_{[k]}(a), \\
\mathbb{L}^{(2)}_{n[k]} &= \sqrt{n}(p_{[k]} - p_{n[k]}).
\end{aligned}$$

Some calculation for the first two parts of $\hat{\tau}^{\text{proj}}_{\text{glm}}$ leads to

$$\begin{aligned}
\sqrt{n}(\hat{\tau}^{\text{proj}}_{\text{glm}} - \tau) =& \Big\{\sum_{k\in\mathcal{K}}\Big(\frac{1}{\pi_{[k]a}}\mathbb{L}^{(1)}_{n[k],r^{\text{proj}}_{i,\text{glm}}(a)} - \frac{1}{\pi_{[k]0}}\mathbb{L}^{(1)}_{n[k],r^{\text{proj}}_{i,\text{glm}}(0)}\Big), a \in \mathcal{A}\Big\} \\
&+ \Big\{\sum_{k\in\mathcal{K}}\mathbb{L}^{(1)}_{n[k],\Psi(a)} - \mathbb{L}^{(1)}_{n[k],\Psi(0)}, a \in \mathcal{A}\Big\} \\
&+ \Big\{\sum_{k\in\mathcal{K}}\mathbb{L}^{(2)}_{n[k]}\{m_{[k]}(a) - m_{[k]}(0)\}, a \in \mathcal{A}\Big\} + o_p(1),
\end{aligned}$$

where $m_{[k]}(a) = E[Y_i(a) \mid B_i = k] - E[Y_i(a)]$. By Lemma S7 and some additional calculations, we have

$$V_{\text{proj}}^{-1/2}\begin{pmatrix}\{\sum_{k\in\mathcal{K}}(\frac{1}{\pi_{[k]a}}\mathbb{L}^{(1)}_{n[k],r^{\text{proj}}_{i,\text{glm}}(a)}-\frac{1}{\pi_{[k]0}}\mathbb{L}^{(1)}_{n[k],r^{\text{proj}}_{i,\text{glm}}(0)}),a\in\mathcal{A}\}\\ \{\sum_{k\in\mathcal{K}}\mathbb{L}^{(1)}_{n[k],\Psi(a)}-\mathbb{L}^{(1)}_{n[k],\Psi(0)},a\in\mathcal{A}\}\\ \{\sum_{k\in\mathcal{K}}\mathbb{L}^{(2)}_{n[k]}\{m_{[k]}(a)-m_{[k]}(0)\},a\in\mathcal{A}\}\end{pmatrix}\xrightarrow{d}\text{N}\left(\begin{pmatrix}0\\0\\0\end{pmatrix},\begin{pmatrix}I_A&0&0\\0&I_A&0\\0&0&I_A\end{pmatrix}\right),$$

where

$$V_{\text{proj}}=\begin{pmatrix}V_{\tilde{r}_{\text{proj}}}&0&0\\0&V_{\psi_{\text{proj}}}&0\\0&0&V_{H\tilde{Y}}\end{pmatrix},$$

and

$$V_{H\tilde{Y}}=\sum_{k\in\mathcal{K}}p_{[k]}\left(m_{[k]}(a)-m_{[k]}(0):a\in\mathcal{A}\right)\times\left(m_{[k]}(a')-m_{[k]}(0):a'\in\mathcal{A}\right)^{\top},$$

$$V_{\tilde{r}_{\text{proj}}}=\sum_{k\in\mathcal{K}}\frac{p_{[k]}}{\pi_{[k]0}}\sigma^2_{[k]\tilde{r}_{\text{proj}}(0)}1_A1_A^{\top}+\text{diag}\left\{\sum_{k\in\mathcal{K}}\frac{p_{[k]}}{\pi_{[k]a}}\sigma^2_{[k]\tilde{r}_{\text{proj}}(a)}:a\in\mathcal{A}\right\},$$

$$V_{\psi_{\text{proj}}}=\Big\{\sum_{k\in\mathcal{K}}p_{[k]}\{\beta^{\text{proj}}_{[k]}(a)-\beta^{\text{proj}}_{[k]}(0)\}^{\top}\Sigma_{[k]\Psi\Psi}\{\beta^{\text{proj}}_{[k]}(a')-\beta^{\text{proj}}_{[k]}(0)\}:(a,a')\in\mathcal{A}\times\mathcal{A}\Big\}.$$

Let $V^{\text{proj}}_{\text{glm}}=V_{r_{\text{proj}}}+V_{\psi_{\text{proj}}}+V_{HY}$. Then we have

$$\sqrt{n}\ (V^{\text{proj}}_{\text{glm}})^{-1/2}(\hat{\tau}^{\text{proj}}_{\text{glm}}-\tau)\xrightarrow{d}\text{N}(0,I_A).$$

S3.5 Proof of Corollary 1

The proof is similar to that of Corollary 1 in Gu et al. (2023). We can replace X_i with $\Psi(X_i)$ and $\beta(a)$ with $\beta^{\text{proj}}_{[k]}(a)$ in each stratum k, then we can go through the proof and reach our conclusion. Therefore, we omit the proof.

S3.6 Proof of Theorem S1

Here we only prove the consistency of $\hat{\sigma}^2_{\text{glm,bc}}$ and $(\hat{\sigma}^{\text{proj}}_{\text{glm,bc}})^2$. As for the transfer learning estimator for variance $\hat{\sigma}^2_{\text{tl}}$ and its corresponding projection variance estimator $(\hat{\sigma}^{\text{proj}}_{\text{tl}})^2$, we can replace $\hat{\beta}_{[k]\text{glm}}(a)$ with $\hat{\beta}_{[k]\text{tl}}(a)$ and complete the proof using the same strategy below. We first show that $(\hat{\sigma}^{\text{proj}}_{\text{glm,bc}})^2$ is consistent for $(\sigma^{\text{proj}}_{\text{glm,bc}})^2$. First, by Lemma S6, we have

$$S_{\text{glm},[k]\Psi\Psi} \xrightarrow{P} \Sigma_{[k]\Psi\Psi},\ \hat{\beta}^{\text{proj}}_{[k]\text{glm}}(a) \xrightarrow{P} \beta^{\text{proj}}_{[k]\text{glm}}(a).$$

Also, we have $p_{n[k]} \xrightarrow{P} p_{[k]},\ n_{[k]a}/n_{[k]} \xrightarrow{P} \pi_{[k]a}$. Therefore, we only need to show that

$$\hat{\sigma}^2_{r^{\text{proj}}_{\text{glm}}(b)} \xrightarrow{P} \sigma^2_{r^{\text{proj}}_{\text{glm}}(b)},$$
$$\hat{\sigma}^2_{\text{bc},HY} \xrightarrow{P} \sigma^2_{\text{bc},HY}.$$

For the second result, the proof is the same as Lemma 3 in Gu et al. (2023), so we omit the proof. Then we focus on $\hat{\sigma}^2_{r^{\text{proj}}_{\text{glm}}(b)}$. It is equivalent to showing that

$$\frac{1}{n_{[k]a}}\sum_{i=1}^{n} I(A_i = a, B_i = k)(\hat{r}^{\text{proj}}_{i,\text{glm}} - \bar{\hat{r}}^{\text{proj}}_{[k]a,\text{glm}})^2 \xrightarrow{P} \text{var}\left\{r^{\text{proj}}_{i,\text{glm}}(a)|B_i = k\right\}$$

where $\hat{r}^{\text{proj}}_{i,\text{glm}} = Y_i - \hat{\Psi}^{\top}_{[k]}(X_i)\hat{\beta}^{\text{proj}}_{[k]}(a)$. Recall that $r_i(a) = Y_i - \Psi^{\top}_{[k]}(X_i)\beta^{\text{proj}}_{[k]}(a)$. Also, we define

$$\Psi^c_{[k]}(X_i) = \Psi_{[k]}(X_i) - \bar{\Psi}_{[k]a}(X_i)$$

as the centered vector and $\hat{\Psi}^c_{[k]}(X_i)$ as its estimator. Then we have

$$\begin{aligned}
&\frac{1}{n_{[k]a}}\sum_{i=1}^{n} I(A_i=a,B_i=k)(\hat{r}^{\text{proj}}_{i,\text{glm}}-\bar{\hat{r}}^{\text{proj}}_{[k]a,\text{glm}})^2\\
&=\frac{1}{n_{[k]a}}\sum_{i=1}^{n} I(A_i=a,B_i=k)(r^{\text{proj}}_{i,\text{glm}}-\bar{r}^{\text{proj}}_{[k]a,\text{glm}})^2\\
&\quad-\frac{2}{n_{[k]a}}\sum_{i=1}^{n} I(A_i=a,B_i=k)(r^{\text{proj}}_{i,\text{glm}}-\bar{r}^{\text{proj}}_{[k]a,\text{glm}})\left[\left\{\hat{\Psi}^{c\top}_{[k]}(X_i)\hat{\beta}^{\text{proj}}_{[k]}(a)\right\}-\left\{\Psi^{c\top}_{[k]}(X_i)\beta^{\text{proj}}_{[k]}(a)\right\}\right]\\
&\quad+\frac{1}{n_{[k]a}}\sum_{i=1}^{n} I(A_i=a,B_i=k)\left[\left\{\hat{\Psi}^{c\top}_{[k]}(X_i)\hat{\beta}^{\text{proj}}_{[k]}(a)\right\}-\left\{\Psi^{c\top}_{[k]}(X_i)\beta^{\text{proj}}_{[k]}(a)\right\}\right]^2.
\end{aligned}$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\frac{1}{n_{[k]a}}\sum_{i\in(a,k)}(r^{\text{proj}}_{i,\text{glm}}-\bar{r}^{\text{proj}}_{[k]a,\text{glm}})\left[\left\{\hat{\Psi}^{c\top}_{[k]}(X_i)\hat{\beta}^{\text{proj}}_{[k]}(a)\right\}-\left\{\Psi^{c\top}_{[k]}(X_i)\beta^{\text{proj}}_{[k]}(a)\right\}\right]\le\\
&\sqrt{\frac{1}{n_{[k]a}}\sum_{i\in(a,k)}(r^{\text{proj}}_{i,\text{glm}}-\bar{r}^{\text{proj}}_{[k]a,\text{glm}})^2}\sqrt{\frac{1}{n_{[k]a}}\sum_{i\in(a,k)}\left[\left\{\hat{\Psi}^{c\top}_{[k]}(X_i)\hat{\beta}^{\text{proj}}_{[k]}(a)\right\}-\left\{\Psi^{c\top}_{[k]}(X_i)\beta^{\text{proj}}_{[k]}(a)\right\}\right]^2},
\end{aligned}$$

where $i\in(a,k)$ indicates unit i in stratum k and treatment a. As

$$\sqrt{\frac{1}{n_{[k]a}}\sum_{i\in(a,k)}(r^{\text{proj}}_{i,\text{glm}}-\bar{r}^{\text{proj}}_{[k]a,\text{glm}})^2}=O_p(1),$$

we only need to show that

$$\frac{1}{n_{[k]a}}\sum_{i=1}^{n} I(A_i=a,B_i=k)\left[\left\{\hat{\Psi}^{c\top}_{[k]}(X_i)\hat{\beta}^{\text{proj}}_{[k]}(a)\right\}-\left\{\Psi^{c\top}_{[k]}(X_i)\beta^{\text{proj}}_{[k]}(a)\right\}\right]^2=o_p(1),$$

so that the last two terms vanish under our conditions. By Lemma S6, we have shown that

$$\frac{1}{n_{[k]a}}\sum_{i=1}^{n} I(B_i=k,A_i=a)\|\hat{\Psi}_{[k]}(X_i)-\Psi_{[k]}(X_i)\|_2^2=o_p(1),\qquad\text{(S3.14)}$$

and

$$\hat{\beta}^{\text{proj}}_{[k]}(a) \xrightarrow{P} \beta^{\text{proj}}_{[k]}(a). \tag{S3.15}$$

Then we have

$$\begin{aligned}
&\frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k) \left[\left\{ \hat{\Psi}^{c\top}_{[k]}(X_i) \hat{\beta}^{\text{proj}}_{[k]}(a) \right\} - \left\{ \Psi^{c\top}_{[k]}(X_i) \beta^{\text{proj}}_{[k]}(a) \right\} \right]^2 \\
&\quad = \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k) \Big[\left\{ \hat{\Psi}^{c}_{[k]}(X_i) - \Psi^{c}_{[k]}(X_i) \right\}^{\top} \{ \hat{\beta}^{\text{proj}}_{[k]}(a) - \beta^{\text{proj}}_{[k]}(a) \} \\
&\qquad + \Psi^{c\top}_{[k]}(X_i) \left\{ \hat{\beta}^{\text{proj}}_{[k]}(a) - \beta^{\text{proj}}_{[k]}(a) \right\} + \left\{ \hat{\Psi}^{c}_{[k]}(X_i) - \Psi^{c}_{[k]}(X_i) \right\}^{\top} \beta^{\text{proj}}_{[k]}(a) \Big]^2.
\end{aligned}$$

By Cauchy-Schwarz inequality, we only need to deal with three squared terms, and they are $o_p(1)$ due to (S3.14) and (S3.15). Therefore, we have

$$\begin{aligned}
&\frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k) (\hat{r}^{\text{proj}}_{i,\text{glm}} - \bar{\hat{r}}^{\text{proj}}_{[k]a,\text{glm}})^2 \\
&\qquad = \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k) (r^{\text{proj}}_{i,\text{glm}} - \bar{r}^{\text{proj}}_{[k]a,\text{glm}})^2 + o_p(1).
\end{aligned}$$

For the consistency of $\hat{\sigma}^2_{\text{glm,bc}}$, we can replace both $\hat{\beta}^{\text{proj}}_{[k]}(a)$ and $\beta^{\text{proj}}_{[k]}(a)$ with e_a, and then follow the procedure above to reach the conclusion.

S4 Lemmas with proof

Lemma S1. *Under Assumptions 1–6, we have*

$$\|\hat{\beta}_{[k]\text{glm}}(a) - \beta_{[k]\text{glm}}(a)\|_1 = O_p\left(s_{[k]} \sqrt{\frac{M_{n[k]} \log p}{n}} \right).$$

Proof. We first define

$$\hat{L}_{[k]a}(\beta) = \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)\{\psi(X_i^\top \beta) - Y_i X_i^\top \beta\}.$$

By the definition of $\hat{\beta}_{[k]\mathrm{glm}}(a)$, we have

$$\hat{L}_{[k]a}(\hat{\beta}_{[k]\mathrm{glm}}(a)) + \lambda_{[k]a}\|\hat{\beta}_{[k]\mathrm{glm}}(a)\|_1 \leq \hat{L}_{[k]a}(\beta_{[k]\mathrm{glm}}(a)) + \lambda_{[k]a}\|\beta_{[k]\mathrm{glm}}(a)\|_1. \tag{S4.16}$$

Let

$$\delta\hat{L}_{[k]a}(\Delta) = \hat{L}_{[k]a}(\beta_{[k]\mathrm{glm}}(a) + \Delta) - \hat{L}_{[k]a}(\beta_{[k]\mathrm{glm}}(a)) - \{\triangledown\hat{L}(\beta_{[k]\mathrm{glm}}(a))\}^\top \Delta,$$

where $\triangledown\hat{L}(\beta_{[k]\mathrm{glm}}(a))$ is the gradient of $\hat{L}_{[k]a}(\beta_{[k]\mathrm{glm}}(a))$ with respect to $\beta_{[k]\mathrm{glm}}(a)$ and $\Delta = \hat{\beta}_{[k]\mathrm{glm}}(a) - \beta_{[k]\mathrm{glm}}(a)$. Then we have

$$\delta\hat{L}_{[k]a}(\Delta) \leq \lambda_{[k]a}\|\beta_{[k]\mathrm{glm}}(a)\|_1 - \lambda_{[k]a}\|\hat{\beta}_{[k]\mathrm{glm}}(a)\|_1 - \{\triangledown\hat{L}(\beta_{[k]\mathrm{glm}}(a))\}^\top \Delta.$$

Let

$$E_1 = \Big\{\|\triangledown\hat{L}(\beta_{[k]\mathrm{glm}}(a))\|_\infty \leq \frac{1}{2}\lambda_{[k]a}\Big\}.$$

By Lemma S5, $\Pr(E_1) \to 1$. Meanwhile, as $S_{[k]}$ is the union of the support of $\beta_{[k]\mathrm{glm}}(a)$ for $a \in \mathcal{A}_0$, we have

$$\|[\beta_{[k]\mathrm{glm}}(a)]_{S_{[k]}^C}\|_1 = 0.$$

Therefore, we split the ℓ_1 norm into the parts in $S_{[k]}$ and $S_{[k]}^C$, and by the

triangle inequality:

$$\|[[\beta_{[k]\text{glm}}(a)]_{S_{[k]}}\|_1 - \|[\hat{\beta}_{[k]\text{glm}}(a)]_{S_{[k]}}\|_1 \leq \|[\Delta]_{S_{[k]}}\|_1,$$

$$\|[\hat{\beta}_{[k]\text{glm}}(a)]_{S^C_{[k]}}\|_1 \geq \|[\Delta]_{S^C_{[k]}}\|_1 - \|[\beta_{[k]\text{glm}}(a)]_{S^C_{[k]}}\|_1.$$

Then under E_1, we can write (S4.16) as

$$\begin{aligned}
\delta\hat{L}_{[k]a}(\Delta) &\leq \lambda_{[k]a}\|\beta_{[k]\text{glm}}(a)\|_1 - \lambda_{[k]a}\|\hat{\beta}_{[k]\text{glm}}(a)\|_1 - \{\nabla\hat{L}(\beta_{[k]\text{glm}}(a))\}^\top\Delta \\
&\leq \lambda_{[k]a}\|\beta_{[k]\text{glm}}(a)\|_1 - \lambda_{[k]a}\|\hat{\beta}_{[k]\text{glm}}(a)\|_1 + |\{\nabla\hat{L}(\beta_{[k]\text{glm}}(a))\}^\top\Delta| \\
&\leq \lambda_{[k]a}(\|[\beta_{[k]\text{glm}}(a)]_{S_{[k]}}\|_1 - \|[\hat{\beta}_{[k]\text{glm}}(a)]_{S_{[k]}}\|_1) \\
&\quad + \lambda_{[k]a}(\|[\beta_{[k]\text{glm}}(a)]_{S^C_{[k]}}\|_1 - \|[\hat{\beta}_{[k]\text{glm}}(a)]_{S^C_{[k]}}\|_1) + \frac{1}{2}\lambda_{[k]a}\|\Delta\|_1 \\
&\leq \frac{3}{2}\lambda_{[k]a}\|[\Delta]_{S_{[k]}}\|_1 - \frac{1}{2}\lambda_{[k]a}\|[\Delta]_{S^C_{[k]}}\|_1.
\end{aligned}$$

By Taylor expansion, for some $t \in (0,1)$, by Assumptions 4 and 5, we have

$$\begin{aligned}
\delta\hat{L}_{[k]a}(\Delta) &= \hat{L}_{[k]a}(\beta_{[k]\text{glm}}(a)+\Delta) - \hat{L}_{[k]a}(\beta_{[k]\text{glm}}(a)) - \{\nabla\hat{L}(\beta_{[k]\text{glm}}(a))\}^\top\Delta \\
&= \Delta^\top\{\nabla^2\hat{L}(\beta_{[k]\text{glm}}(a)+t\Delta)\}\Delta \\
&= \Delta^\top\left(\frac{1}{n_{[k]a}}\sum_{i=1}^{n} I(A_i=a, B_i=k)X_iX_i^\top\psi''(X_i^\top(\beta_{[k]\text{glm}}(a)+t\Delta))\right)\Delta \\
&\geq 0,
\end{aligned}$$

where $\nabla^2\hat{L}(\beta_{[k]\text{glm}}(a))$ is the Hessian matrix of $\hat{L}_{[k]a}(\beta_{[k]\text{glm}}(a))$ with respect to $\beta_{[k]\text{glm}}(a)$. Therefore, we have $\frac{3}{2}\lambda_{[k]a}\|[\Delta]_{S_{[k]}}\|_1 - \frac{1}{2}\lambda_{[k]a}\|[\Delta]_{S^C_{[k]}}\|_1 \geq 0$, which leads to $\|[\Delta]_{S^C_{[k]}}\|_1 \leq 3\|[\Delta]_{S_{[k]}}\|_1$. Let

$$E_3 = \{\Delta \in V : \delta\hat{L}_{[k]a}(\Delta) \geq \kappa\|\Delta\|_2^2\},$$

where $V = \{u \in \mathbb{R}^p : \ \|u_{S_{[k]}^C}\|_1 \leq 3\|u_{S_{[k]}}\|_1\}$. By Lemma S4,

$$\Pr(E_3) \to 1.$$

Under E_3, we have

$$\kappa\|\Delta\|_2^2 \leq \frac{3}{2}\lambda_{[k]a}\|\Delta_{S_{[k]}}\|_1 \leq \frac{3}{2}\sqrt{s_{[k]}}\lambda_{[k]a}\|\Delta_{S_{[k]}}\|_2 \leq \frac{3}{2}\sqrt{s_{[k]}}\lambda_{[k]a}\|\Delta\|_2.$$

That is

$$\|\Delta\|_2 \leq \frac{3}{2\kappa}\sqrt{s_{[k]}}\lambda_{[k]a}.$$

Then

$$\|\Delta\|_1 = \|\Delta_{S_{[k]}}\|_1 + \|\Delta_{S_{[k]}^C}\|_1 \leq 4\|\Delta_{S_{[k]}}\|_1 \leq 4\sqrt{s_{[k]}}\|\Delta\|_2 \leq \frac{6}{\kappa}s_{[k]}\lambda_{[k]a}.$$

Therefore, as $\lambda_{[k]a} = O_p(\sqrt{M_{n[k]}\log p/n})$, we have

$$\|\hat{\beta}_{[k]\mathrm{glm}}(a) - \beta_{[k]\mathrm{glm}}(a)\|_1 = O_p\left(s_{[k]}\sqrt{\frac{M_{n[k]}\log p}{n}}\right).$$

□

Lemma S2. *Under Assumptions 1–3, we have*

$$\|\bar{X}_{[k]a} - \bar{X}_{[k]}\|_\infty = O_p\left\{\left(\frac{\log p}{n}\right)^{1/2}\right\},$$

for all $a \in \mathcal{A}_0,\ k \in \mathcal{K}$.

Proof. The proof is similar to that of Lemma S4 in Liu et al. (2023). We can replace $A_i = 1$ with $A_i = a$ for all $a \in \mathcal{A}_0$ to show our results. Therefore, we omit the proof. □

Lemma S3. *Let* $\mathbb{L}_n^{(2)} = \{\mathbb{L}_{n[k]}^{(2)}, k \in \mathcal{K}\}$, *and*

$$\mathbb{L}_n = \left(\left\{ \sum_{k\in\mathcal{K}} \frac{1}{\pi_{[k]a}} \mathbb{L}^{(1)}_{n[k],r_{i,\mathrm{glm}}(a)}, a \in \mathcal{A}_0 \right\}, \left\{ \sum_{k\in\mathcal{K}} \mathbb{L}^{(1)\top}_{n[k],\psi(a)}, a \in \mathcal{A}_0 \right\}, \mathbb{L}_n^{(2)} \right).$$

Under Assumptions 1–4, we have

$$\mathbb{L}_n \xrightarrow{d} \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & \Sigma_{12} & 0 \\ \Sigma_{12}^\top & \Sigma_2 & 0 \\ 0 & 0 & \Sigma_3 \end{pmatrix} \right),$$

where

$$\Sigma_1 = \mathrm{diag}\left\{ \sum_{k\in\mathcal{K}} \frac{p_{[k]}}{\pi_{[k]a}} \sigma^2_{[k]\tilde{r}_{\mathrm{glm}}(a)}, a \in \mathcal{A}_0 \right\},$$

$$\Sigma_{12} = \left\{ \sum_{k\in\mathcal{K}} p_{[k]} \mathrm{Cov}\big\{ Y_i(a) - \psi'(X_i^\top \beta_{[k]\mathrm{glm}}(a)), \psi'(X_i^\top \beta_{[k]\mathrm{glm}}(a')) \big\}, (a,a') \in \mathcal{A}_0 \times \mathcal{A}_0 \right\},$$

$$\Sigma_2 = \left\{ \sum_{k\in\mathcal{K}} p_{[k]} \mathrm{Cov}\big\{ \psi'(X_i^\top \beta_{[k]\mathrm{glm}}(a)), \psi'(X_i^\top \beta_{[k]\mathrm{glm}}(a')) \big\}, (a,a') \in \mathcal{A}_0 \times \mathcal{A}_0 \right\},$$

$$\Sigma_3 = \mathrm{diag}\{p_{[k]} : k \in \mathcal{K}\} - \{p_{[k]} p_{[k']}, (k,k') \in \mathcal{K} \times \mathcal{K}\}.$$

Proof. Recall that

$$\Psi_{[k]}(X_i) = \left(\psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(0)\}, \psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(1)\}, \ldots, \psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(A)\} \right)^\top.$$

Without loss of generality, we assume that $\Sigma_{[k]\Psi\Psi} = \mathrm{var}\{\Psi_{[k]}(X_i) | B_i = k\}$ is positive definite. Also, we denote $e_a \in \mathbb{R}^{A+1}$ is a vector whose ath element is 1 and others are 0. Then we have

$$\Psi_{[k]}^\top(X_i) e_a = \psi'(X_i^\top \beta_{[k]\mathrm{glm}}(a)), \quad i \text{ in treatment group } a.$$

Therefore we can write the transformed outcome $r_{\text{glm,a}}$ as

$$r_{\text{glm,a}} = Y_i(a) - \Psi_{[k]}^{\top}(X_i)e_a.$$

First, by CLT, we have

$$\mathbb{L}_n^{(2)} \xrightarrow{d} \mathrm{N}(0, \Sigma_3). \tag{S4.17}$$

Let $Z_i(a) = (Y_i(a), X_i)^{\top}$. The proof can follow that of Lemma 5 in Gu et al. (2023) until we obtain

$$\mathbb{L}_{n\tilde{Z}}^{(1)} \xrightarrow{d} \mathrm{N}(0, \Sigma_Z),$$

where $\mathbb{L}_{n\tilde{Z}}^{(1)} = \{(1/\sqrt{n})\sum_{i=1}^n I(A_i = a, B_i = k)\tilde{Z}_i(a), (a,k) \in \mathcal{A}_0 \times \mathcal{K}\}$ is a $(p+1) \times \{K(A+1)\}$ vector, as we vectorize a $(p+1) \times \{K(A+1)\}$ matrix column-wise and stack its columns first by treatment $a \in \mathcal{A}_0$ and then by stratum $k \in \mathcal{K}$. Here, $\Sigma_Z = \mathrm{diag}\{\pi_{[k]a}p_{[k]}\Sigma_{[k]Z(a)Z(a)},\ (a,k) \in \mathcal{A}_0 \times \mathcal{K}\}$ is a covariance matrix where

$$\Sigma_{[k]Z(a)Z(a)} = \begin{pmatrix} \sigma^2_{[k]\tilde{Y}(a)} & \Sigma^{\top}_{[k]XY(a)} \\ \Sigma_{[k]XY(a)} & \Sigma_{[k]XX} \end{pmatrix}.$$

Then, we can replace X_i by $\Psi_{[k]}(X_i)$ in $Z_i(a)$, which we denote as $Z_i^{\Psi}(a) = (Y_i(a), \Psi_{[k]}(X_i))^{\top}$. By the delta method, we can obtain

$$\mathbb{L}_{n\tilde{Z}^{\Psi}}^{(1)} \xrightarrow{d} \mathrm{N}(0, \Sigma_{Z^{\Psi}}),$$

where $\mathbb{L}_{n\tilde{Z}^{\Psi}}^{(1)} = \{(1/\sqrt{n})\sum_{i=1}^n I(A_i = a, B_i = k)\tilde{Z}_i^{\Psi}(a), (a,k) \in \mathcal{A}_0 \times \mathcal{K}\}$ is a $(A+2) \times \{K(A+1)\}$ vector and $\Sigma_{Z^{\Psi}} = \mathrm{diag}\{\pi_{[k]a}p_{[k]}\Sigma_{[k]Z^{\Psi}(a)Z^{\Psi}(a)},\ (a,k) \in$

$\mathcal{A}_0 \times \mathcal{K}\}$ is a covariance matrix with

$$\Sigma_{[k]Z^{\Psi}(a)Z^{\Psi}(a)} = \begin{pmatrix} \sigma^2_{[k]\tilde{Y}(a)} & \Sigma^{\top}_{[k]\Psi Y(a)} \\ \Sigma_{[k]\Psi Y(a)} & \Sigma_{[k]\Psi\Psi} \end{pmatrix}.$$

As defined in the proof of Proposition 1,

$$\begin{aligned}
\sum_{k\in\mathcal{K}} \frac{1}{\pi_{[k](a)}} \mathbb{L}^{(1)}_{n[k],r_{i,\mathrm{glm}}(a)} &= \sum_{k\in\mathcal{K}} \frac{1}{\pi_{[k](a)}} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(A_i = a, B_i = k) \tilde{r}_{i,[k]\mathrm{glm}}(a) \right], \\
\sum_{k\in\mathcal{K}} \mathbb{L}^{(1)}_{n[k],\Psi(a)} &= \sum_{k\in\mathcal{K}} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(B_i = k) \tilde{\psi}'(X_i^{\top} \beta_{[k]\mathrm{glm}}(a)) \right] \\
&= \sum_{k\in\mathcal{K}} \left[\sum_{a'\in\mathcal{A}_0} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(A_i = a', B_i = k) \tilde{\psi}'(X_i^{\top} \beta_{[k]\mathrm{glm}}(a)) \right\} \right] \\
&= \sum_{k\in\mathcal{K}} \left[\sum_{a'\in\mathcal{A}_0} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(A_i = a', B_i = k) \tilde{\Psi}^{\top}_{[k]}(X_i) e_a \right\} \right]
\end{aligned}$$

are linear combinations of $\mathbb{L}^{(1)}_{n\tilde{Z}^{\psi}}$. Therefore,

$$\left(\left\{ \sum_{k\in\mathcal{K}} \frac{1}{\pi_{[k]a}} \mathbb{L}^{(1)}_{n[k],r_{i,\mathrm{glm}}(a)}, a \in \mathcal{A}_0 \right\}, \left\{ \sum_{k\in\mathcal{K}} \mathbb{L}^{(1)}_{n[k],\Psi(a)}, a \in \mathcal{A}_0 \right\} \right)$$

is jointly asymptotic normal and the asymptotic covariance matrix is

$$\begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12}^{\top} & \Sigma_2 \end{pmatrix}, \tag{S4.18}$$

where

$$\begin{aligned}
\Sigma_1 &= \mathrm{diag}\left\{ \sum_{k\in\mathcal{K}} \frac{p_{[k]}}{\pi_{[k]a}} \sigma^2_{[k]\tilde{r}_{\mathrm{glm}}(a)}, a \in \mathcal{A}_0 \right\}, \\
\Sigma_{12} &= \left\{ \sum_{k\in\mathcal{K}} p_{[k]} e_{a'}^{\top} \Sigma^{\top}_{[k]\Psi r_{\mathrm{glm}}(a)}, (a, a') \in \mathcal{A}_0 \times \mathcal{A}_0 \right\},
\end{aligned}$$

$$\Sigma_2 = \left\{ \sum_{k\in\mathcal{K}} p_{[k]} e_a^\top \Sigma_{[k]\Psi\Psi} e_{a'}, (a,a') \in \mathcal{A}_0 \times \mathcal{A}_0 \right\}.$$

Then we reach our conclusion by combining (S4.17) and (S4.18).

□

Lemma S4. *Under Assumptions 1–8, for any* $u \in V = \{u \in \mathbb{R}^p : \|u_{S_{[k]}^C}\|_1 \leq 3\|u_{S_{[k]}}\|_1\}$*, we have*

$$\delta \hat{L}_{[k]a}(u) \geq \kappa \|u\|_2^2$$

with probability $1 - c_1 \exp(-c_2 n)$*, for some positive constants* c_1, c_2, κ *independent of* n*.*

Proof. The proof is similar to that of Lemma 4 in Gu et al. (2024). Here, we need to deal with $\delta \hat{L}_{[k]a}(u)$ with Proposition 2 in Negahban et al. (2012). As $\{Y_i(0), \ldots, Y_i(A), B_i, X_i\}$ are i.i.d. data and $A^{(n)}$ is independent with $\{Y_i(0), \ldots, Y_i(A), X_i\}$ conditional on $B^{(n)}$. Therefore, given the $A^{(n)}$ and $B^{(n)}$, the distribution of $\delta \hat{L}_{[k]a}(u)$ is the same if we order the units first by strata and then by treatments within each strata. Then, independently with $k = 1, 2, \ldots, K$ and $(A^{(n)}, B^{(n)})$, let $\{Y_i^k(0), \ldots, Y_i^k(A), X_i^k\}$ has the same distribution with $\{Y_i(0), \ldots, Y_i(A), X_i\}$ given $B_i = k$. Conditional on $\{A^{(n)}, B^{(n)}\}$, $\delta \hat{L}_{[k]a}(u)$ has same marginal distribution with

$$\delta L_n^k(u) = L_n^k(\beta_{[k]\mathrm{glm}}(a) + u) - L_n^k(\beta_{[k]\mathrm{glm}}(a)) - \{\nabla L_n^k(\beta_{[k]\mathrm{glm}}(a))\}^\top u,$$

where

$$L_n^k(\beta) = \frac{1}{n_{[k]a}} \sum_{i=1}^{n_{[k]a}} \left[\psi(\{X_i^k\}^\top \beta) - Y_i^k(a) \{X_i^k\}^\top \beta \right]$$

and $\{X_i^k, Y_i^k(0), \ldots, Y_i^k(A)\}$ has the same marginal distribution with $\{X_i, Y_i(0), \ldots, Y_i(A)\}|B_i = k$. As X_i is uniformly bounded by M and X_i^k has the same distribution with X_i conditional on $B_i = k$, X_i^k is a sub-Gaussian vector. Therefore, by Proposition 2 in Negahban et al. (2012), for any $N \geq 1$, we have

$$\delta L_N^k(u) \geq \kappa_1 \|u\|_2 \left(\|u\|_2 - \kappa_2 \sqrt{\frac{\log p}{N}} \|u\|_1 \right), \ \|u\|_2 \leq 1 \tag{S4.19}$$

with probability $1 - c_1 \exp(-c_2 N)$. By Assumption 3 and 4, using Taylor expansion on $\beta_{[k]\text{glm}}(a)$, we have

$$\delta L_N^k(u) = u^\top \left(\frac{1}{N} \sum_{i=1}^{N} X_i^k \psi''\big(\{X_i^k\}^\top \{\beta_{[k]\text{glm}}(a) + tu\}\big) \{X_i^k\}^\top \right) u,$$

for some $t \in (0, 1)$. Then, we can divide $\|u\|_2$ on both sides of (S4.19) to ignore the condition $\|u\|_2 \leq 1$ to have

$$\delta L_N^k(u) \geq \kappa_1 \|u\|_2 \left(\|u\|_2 - \kappa_2 \sqrt{\frac{\log p}{N}} \|u\|_1 \right).$$

Then, we can follow the steps in the proof of Lemma 4 in Gu et al. (2024) such that

$$\kappa_1 \|u\|_2 \left(\|u\|_2 - \kappa_2 \sqrt{\frac{\log p}{N}} \|u\|_1 \right) \geq \left(1 - 4\kappa_2 \sqrt{\frac{s_{[k]} \log p}{N}} \right) \kappa_1 \|u\|_2^2.$$

Under the assumption $M_{n[k]}^2 s_{[k]}^2 \log^2 p/n \to 0$, we have

$$1 - 4\kappa_2 \sqrt{\frac{s_{[k]} \log p}{n_{[k]}}} > \frac{1}{2}$$

when $n \to \infty$. Therefore,

$$\Pr\left(\delta L_N^k(u) \geq \frac{1}{2}\kappa_1 \|u\|_2^2\right) \to 1, \tag{S4.20}$$

when $N \to \infty$. Using the almost sure representation theorem, we can construct $\tilde{n}_{[k]a}$ (independent of $\{Y_i^k(0), \ldots, Y_i^k(A), X_i^k\}$) such that $\tilde{n}_{[k]a}/n$ has the same distribution with $n_{[k]a}/n$ and $\tilde{n}_{[k]a}/n \to \pi_{[k]a} p_{[k]}$ almost surely. Therefore, for any $u \in V$,

$$\begin{aligned}
&\Pr\left(u^\top \left\{\frac{1}{n_{[k]a}} \sum_{i=1}^{n_{[k]a}} X_i^k \psi''\big(\{X_i^k\}^\top \{\beta_{[k]\mathrm{glm}}(a) + tu\}\big)\{X_i^k\}^\top\right\} u \geq \frac{1}{2}\kappa_1 \|u\|_2^2 \;\Big|\; A^{(n)}, B^{(n)}\right) \\
&= \Pr\left(u^\top \left\{\frac{1}{n\frac{\tilde{n}_{[k]a}}{n}} \sum_{i=1}^{n\frac{\tilde{n}_{[k]a}}{n}} X_i^k \psi''\big(\{X_i^k\}^\top \{\beta_{[k]\mathrm{glm}}(a) + tu\}\big)\{X_i^k\}^\top\right\} u \geq \frac{1}{2}\kappa_1 \|u\|_2^2 \;\Big|\; A^{(n)}, B^{(n)}\right) \\
&= E\Bigg[\Pr\left(u^\top \left\{\frac{1}{n\frac{\tilde{n}_{[k]a}}{n}} \sum_{i=1}^{n\frac{\tilde{n}_{[k]a}}{n}} X_i^k \psi''\big(\{X_i^k\}^\top \{\beta_{[k]\mathrm{glm}}(a) + tu\}\big)\{X_i^k\}^\top\right\} u \geq \frac{1}{2}\kappa_1 \|u\|_2^2 \right. \\
&\left. \;\Big|\; A^{(n)}, B^{(n)}, \frac{\tilde{n}_{[k]a}}{n}\right)\Bigg] \to 1,
\end{aligned}$$

where the convergence follows from the dominant convergence theorem, $n(\tilde{n}_{[k]a}/n) \to \infty$ a.s., independence of $\tilde{n}_{[k]a}/n$ and $\{Y_i^k(0), \ldots, Y_i^k(A), X_i^k\}$ and (S4.19). Therefore,

$$\Pr\left(\delta \hat{L}_{[k]a}(u) \geq \frac{1}{2}\kappa_1 \|u\|_2^2\right)$$

$$= \Pr\left(u^\top \left(\frac{1}{n_{[k]a}} \sum_{i=1}^{n_{[k]a}} X_i^k \psi''\big(\{X_i^k\}^\top \{\beta_{[k]\mathrm{glm}}(a) + tu\}\big)\{X_i^k\}^\top \right) u \geq \frac{1}{2}\kappa_1 \|u\|_2^2 \right)$$

$\to 1.$

□

Lemma S5. *Under Assumptions 1–6, we have*

$$\|\{\triangledown \hat{L}(\beta_{[k]\mathrm{glm}}(a))\}\|_\infty \leq \frac{\lambda_{[k]a}}{2}$$

with probability tending to one.

Proof. Let

$$\epsilon_{i,\mathrm{glm}}(a) = Y_i(a) - \psi'(X_i^\top \beta_{[k]\mathrm{glm}}(a)).$$

Recall that

$$\begin{aligned}
\triangledown \hat{L}(\beta_{[k]\mathrm{glm}}(a)) &= \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k) X_i (Y_i - \psi'(X_i^\top \beta_{[k]\mathrm{glm}}(a))) \\
&= \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k) X_i (Y_i(a) - \psi'(X_i^\top \beta_{[k]\mathrm{glm}}(a))) \\
&= \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k) X_i \epsilon_{i,\mathrm{glm}}(a).
\end{aligned}$$

Similar to the arguments in the proof of Lemma S4, conditional on $\{A^{(n)}, B^{(n)}\}$, $\triangledown \hat{L}(\beta_{[k]\mathrm{glm}}(a))$ has same distribution with

$$\frac{1}{n_{[k]a}} \sum_{i=1}^{n_{[k]a}} X_i^k \epsilon_{i,\mathrm{glm}}^k(a),$$

where $\{X_i^k, \epsilon_{i,\text{glm}}^k(a)\}$ are i.i.d. with marginal distribution being equal to the distribution of $\{X_i, \epsilon_{i,\text{glm}}(a)\}|B_i = k$, independent of $\{A^{(n)}, B^{(n)}\}$. Then by Nemirovski's inequality, for any integer $m \geq 1$, there exists a constant c independent of n, such that

$$\begin{aligned}
&E\Big[\|\frac{1}{m}\sum_{i=1}^{m} X_i^k \epsilon_{i,\text{glm}}^k(a) - E\{X_i^k \epsilon_{i,\text{glm}}^k(a)\}\|_\infty^2\Big] \\
&\leq \frac{c\log p}{m^2}\sum_{i=1}^{m} E\big[\|X_i^k \epsilon_{i,\text{glm}}^k(a) - E\{X_i^k \epsilon_{i,\text{glm}}^k(a)\|_\infty^2\big] \\
&\leq \frac{4c\log p}{m^2}\sum_{i=1}^{m} E\big[\|X_i^k \epsilon_{i,\text{glm}}^k(a)\|_\infty^2\big] \\
&\leq \frac{4M^2 c\log p}{m^2}\sum_{i=1}^{m} E\big\{\epsilon_{i,\text{glm}}^k(a)\big\}^2 \\
&\leq c_\lambda \frac{\log p}{m},
\end{aligned}$$

where $c_\lambda = 4M^2 c \max_{a\in\mathcal{A}_0} E\big\{\epsilon_{i,\text{glm}}^k(a)\big\}^2$ is independent of n. The third inequality is because $||X_i||_\infty \leq M$ and $E\{\epsilon_{i,\text{glm}}^2(a) < \infty\}$. Therefore, by Markov inequality, for any sequence $M_{n[k]}$ tending to infinity, we have

$$\begin{aligned}
&\Pr\left(\|\frac{1}{m}\sum_{i=1}^{m} X_i^k \epsilon_{i,\text{glm}}^k(a) - E\{X_i^k \epsilon_{i,\text{glm}}^k(a)\}\|_\infty \geq \sqrt{c_\lambda M_{n[k]}\frac{\log p}{m}}\right) \\
&\leq \frac{m}{c_\lambda M_{n[k]}\log p} E\Big[\|\frac{1}{m}\sum_{i=1}^{m} X_i^k \epsilon_{i,\text{glm}}^k(a) - E\{X_i^k \epsilon_{i,\text{glm}}^k(a)\}\|_\infty^2\Big] \\
&\leq \frac{1}{M_{n[k]}}.
\end{aligned}$$

Then following similar argument in the proof of Lemma S3, we have

$$\Pr\left(\|\nabla\hat{L}(\beta_{[k]\text{glm}}(a)) - E\{\nabla\hat{L}(\beta_{[k]\text{glm}}(a))\}\|_\infty \geq \sqrt{c_\lambda M_{n[k]}\frac{\log p}{n_{[k]a}}}\right) \leq \frac{1}{M_{n[k]}}.$$

As $\beta_{[k]\text{glm}}(a)$ is derived by minimizing the population-level loss function, we have

$$E\{\nabla \hat{L}(\beta_{[k]\text{glm}}(a))\} = 0.$$

Therefore, we have

$$\Pr\left(\|\{\nabla \hat{L}(\beta_{[k]\text{glm}}(a))\}\|_\infty \leq \frac{\lambda_{[k]a}}{2}\right) \geq 1 - \frac{1}{M_{n[k]}} \to 1.$$

□

Lemma S6. *Under conditions for $\hat{\beta}_{[k]\text{glm}}(a)$ in Proposition 1 or conditions for $\hat{\beta}_{[k]\text{tl}}(a)$ in Theorem 1, for their corresponding $\hat{\beta}^{\text{proj}}_{[k]}(a)$, we have*

$$\hat{\beta}^{\text{proj}}_{[k]}(a) \xrightarrow{P} \beta^{\text{proj}}_{[k]}(a).$$

Proof. We only show that this lemma is true for $\hat{\beta}^{\text{proj}}_{[k]}(a)$ derived by GLM projection coefficient $\hat{\beta}_{[k]\text{glm}}(a)$. The results for $\hat{\beta}_{[k]\text{tl}}(a)$ can be derived similarly by replacing the result in Proposition 1 with the one in Theorem 1. As $\hat{\beta}^{\text{proj}}_{[k]}(a) = S^{-1}_{[k]\Psi\Psi} S_{[k]\Psi Y(a)}$, we only need to show that

$$S_{[k]\Psi\Psi} \xrightarrow{P} \Sigma_{[k]\Psi\Psi},$$

$$S_{[k]\Psi Y(a)} \xrightarrow{P} \Sigma_{[k]\Psi Y(a)}.$$

We first show that $S_{[k]\Psi Y(a)} \xrightarrow{P} \Sigma_{[k]\Psi Y(a)}$. As

$S_{[k]\Psi Y(a)}$

$$
\begin{aligned}
&= \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(B_i = k, A_i = a) \big[\hat{\Psi}_{[k]}(X_i) - \bar{\hat{\Psi}}_{[k]a}(X_i)\big](Y_i - \bar{Y}_{[k]a}) \\
&= \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(B_i = k, A_i = a) \big[\Psi_{[k]}(X_i) - \bar{\Psi}_{[k]a}(X_i)\big](Y_i - \bar{Y}_{[k]a}) \\
&\quad + \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(B_i = k, A_i = a) \left(\big[\hat{\Psi}_{[k]}(X_i) - \bar{\hat{\Psi}}_{[k]a}(X_i)\big] - \big[\Psi_{[k]}(X_i) - \bar{\Psi}_{[k]a}(X_i)\big] \right) (Y_i - \bar{Y}_{[k]a}).
\end{aligned}
$$

For the first term, By Lemma 1 in Gu et al. (2023), for $f_i = f(Y_i(0), \ldots, Y_i(A), B_i, X_i)$,

we have

$$
\frac{1}{n} \sum_{i=1}^{n} f_i I(A_i = a, B_i = k) \xrightarrow{P} p_{[k]} \pi_{[k]a} E(f | B_i = k). \tag{S4.21}
$$

Therefore, we have

$$
\begin{aligned}
&\frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(B_i = k, A_i = a) \big[\Psi_{[k]}(X_i) - \bar{\Psi}_{[k]a}(X_i)\big](Y_i - \bar{Y}_{[k]a}) \\
&\qquad = \frac{n}{n_{[k]a}} \frac{1}{n} \sum_{i=1}^{n} I(B_i = k, A_i = a) \big[\Psi_{[k]}(X_i) - \bar{\Psi}_{[k]a}(X_i)\big](Y_i - \bar{Y}_{[k]a}) \\
&\qquad \xrightarrow{P} \Sigma_{[k]\Psi Y(a)}.
\end{aligned}
$$

For the second term, let $\delta^{\psi}_{[k]}(X_i, a) = \psi'\{X_i^\top \hat{\beta}_{[k]\mathrm{glm}}(a)\} - \psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(a)\}$.

As we have shown in the proof of Proposition 1, by Taylor expansion, we

have

$$
\begin{aligned}
&\frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(B_i = k, A_i = a) \{\delta^{\psi}_{[k]}(X_i, a)\}^2 \\
&\leq M_\psi^2 \Big[\frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(B_i = k, A_i = a) \Big\{ X_i^\top (\hat{\beta}_{[k]\mathrm{glm}}(a) - \beta_{[k]\mathrm{glm}}(a)) \Big\}^2 \Big] \\
&\leq M_\psi^2 \|X_i\|_\infty^2 \|\hat{\beta}_{[k]\mathrm{glm}}(a) - \beta_{[k]\mathrm{glm}}(a))\|_1^2
\end{aligned}
$$

$$\leq M_\psi^2 M^2 \|\hat{\beta}_{[k]\mathrm{glm}}(a) - \beta_{[k]\mathrm{glm}}(a))\|_1^2$$

$$= o_p(1).$$

The last line is derived from the results of Proposition 1. We can use the same technique to show that even if unit i does not belong to treatment a, the conclusion is true, which we can write as

$$\frac{1}{n_{[k]a'}} \sum_{i=1}^{n} I(B_i = k, A_i = a')(\delta^\psi(X_i, a))^2 = o_p(1), \text{ for } (a, a') \in \mathcal{A}_0 \times \mathcal{A}_0.$$

Let

$$\delta^\psi(X_i) = (\delta^\psi(X_i, 0), \delta^\psi(X_i, 1), \ldots, \delta^\psi(X_i, A))^\top.$$

As we have finite treatments, we can obtain that

$$\frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(B_i = k, A_i = a) \|\delta^\psi(X_i)\|_2^2$$

$$= \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(B_i = k, A_i = a) \sum_{a' \in \mathcal{A}_0} (\delta^\psi(X_i, a'))^2$$

$$= o_p(1).$$

Therefore, by the Cauchy-Schwarz inequality, we have

$$\frac{1}{n_{[k]a}} \sum_{i \in (a,k)} (Y_i - \bar{Y}_{[k]a})(\delta^\psi(X_i) - \bar{\delta}^\psi_{[k]a}(X_i))$$

$$\leq \sqrt{\frac{1}{n_{[k]a}} \sum_{i \in (a,k)} (Y_i - \bar{Y}_{[k]a})^2} \sqrt{\frac{1}{n_{[k]a}} \sum_{i \in (a,k)} \|\delta^\psi(X_i) - \bar{\delta}^\psi_{[k]a}(X_i)\|_2^2}$$

$$= \{\sigma_{[k]Y(a)} + o_p(1)\} \cdot o_p(1)$$

$\xrightarrow{P} 0.$

The first convergence is due to (S4.21).

Then we can deal with $S_{[k]\Psi\Psi}$. We have

$$S_{[k]\Psi\Psi} = \frac{1}{n_{[k]a}} \sum_{i=1}^{n} \big[\hat{\Psi}_{[k]}(X_i) - \bar{\hat{\Psi}}_{[k]a}(X_i)\big]\big[\hat{\Psi}_{[k]}(X_i) - \bar{\hat{\Psi}}_{[k]a}(X_i)\big]^\top.$$

Similarly, as we have finite treatments, we can only focus on each element in $S_{[k]\Psi\Psi}$. Let $S_{[k]\Psi(a,a')}$ denote the (a,a')th element in $S_{[k]\Psi\Psi}$, which is

$$\begin{aligned} S_{[k]\Psi(a,a')} = \frac{1}{n_{[k]a}} \sum_{i\in(a,k)} &\big[\psi'\{X_i^\top \hat{\beta}_{[k]\mathrm{glm}}(a)\}) - \bar{\psi}'_{[k]a}\{X_i^\top \hat{\beta}_{[k]\mathrm{glm}}(a)\}\big] \\ &\times \big[\psi'\{X_i^\top \hat{\beta}_{[k]\mathrm{glm}}(a')\}) - \bar{\psi}'_{[k]a}\{X_i^\top \hat{\beta}_{[k]\mathrm{glm}}(a')\}\big]^\top. \end{aligned}$$

Then we have

$$\begin{aligned} &S_{[k]\Psi(a,a')} \\ &= \frac{1}{n_{[k]a}} \sum_{i\in(a,k)} [\psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(a) - \bar{\psi}'_{[k]a}\{X_i^\top \beta_{[k]\mathrm{glm}}(a)][\psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(a') - \bar{\psi}'_{[k]a}\{X_i^\top \beta_{[k]\mathrm{glm}}(a')] \\ &+ \frac{1}{n_{[k]a}} \sum_{i\in(a,k)} [\delta^{\psi}(X_i,a) - \bar{\delta}^{\psi}_{[k]a}(X_i,a)][\psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(a') - \bar{\psi}'_{[k]a}\{X_i^\top \beta_{[k]\mathrm{glm}}(a')] \\ &+ \frac{1}{n_{[k]a}} \sum_{i\in(a,k)} [\psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(a) - \bar{\psi}'_{[k]a}\{X_i^\top \beta_{[k]\mathrm{glm}}(a)][\delta^{\psi}(X_i,a') - \bar{\delta}^{\psi}_{[k]a}(X_i,a')] \\ &+ \frac{1}{n_{[k]a}} \sum_{i\in(a,k)} [\delta^{\psi}(X_i,a) - \bar{\delta}^{\psi}_{[k]a}(X_i,a)][\delta^{\psi}(X_i,a') - \bar{\delta}^{\psi}_{[k]a}(X_i,a')]. \end{aligned}$$

By (S4.21), Cauchy-Schwarz inequality and similar argument for the second

part of $S_{[k]\Psi Y(a)}$, we have

$$S_{[k]\Psi(a,a')} \xrightarrow{P} \sigma_{\Psi\Psi(a,a')}.$$

Then we have

$$S_{[k]\Psi\Psi} \xrightarrow{P} \Sigma_{[k]\Psi\Psi}.$$

□

Lemma S7. *Let* $\mathbb{L}_n^{(2)} = \{\mathbb{L}_{n[k]}^{(2)}, k \in \mathcal{K}\}$, *and*

$$\mathbb{L}_n^{\text{proj}} = \left(\left\{ \sum_{k\in\mathcal{K}} \frac{1}{\pi_{[k]a}} \mathbb{L}^{(1)}_{n[k],r^{\text{proj}}_{i,\text{glm}}(a)}, a \in \mathcal{A}_0 \right\}, \left\{ \sum_{k\in\mathcal{K}} \mathbb{L}^{(1)\top}_{n[k],\Psi(a)}, a \in \mathcal{A}_0 \right\}, \mathbb{L}_n^{(2)} \right).$$

Under conditions for $\hat{\beta}_{[k]\text{glm}}(a)$ *in Proposition 1 or conditions for* $\hat{\tau}_{\text{tl}}$ *in Theorem 2, we have*

$$\mathbb{L}_n^{\text{proj}} \xrightarrow{d} \text{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_1^{\text{proj}} & 0 & 0 \\ 0 & \Sigma_2^{\text{proj}} & 0 \\ 0 & 0 & \Sigma_3 \end{pmatrix} \right),$$

where

$$\Sigma_1^{\text{proj}} = \text{diag}\left\{ \sum_{k\in\mathcal{K}} \frac{p_{[k]}}{\pi_{[k]a}} \sigma^2_{[k]\tilde{r}^{\text{proj}}_{\text{glm}}(a)}, a \in \mathcal{A}_0 \right\},$$

$$\Sigma_2 = \left\{ \sum_{k\in\mathcal{K}} p_{[k]} \{\beta^{\text{proj}}_{[k]}(a)\}^\top \Sigma_{[k]\Psi\Psi} \beta^{\text{proj}}_{[k]}(a'), (a,a') \in \mathcal{A}_0 \times \mathcal{A}_0 \right\},$$

$$\Sigma_3 = \text{diag}\{p_{[k]} : k \in \mathcal{K}\} - \{p_{[k]} p_{[k']}, (k,k') \in \mathcal{K} \times \mathcal{K}\}.$$

Proof. The proof is similar to that of Lemma S3. We can replace e_a in the

proof of Lemma S3 with $\beta_{[k]}^{\text{proj}}(a)$ and go through the proof until we obtain

$$\left(\left\{ \sum_{k\in\mathcal{K}} \frac{1}{\pi_{[k]a}} \mathbb{L}^{(1)}_{n[k],r^{\text{proj}}_{i,\text{glm}}(a)}, a \in \mathcal{A}_0 \right\}, \left\{ \sum_{k\in\mathcal{K}} \mathbb{L}^{(1)}_{n[k],\Psi(a)}, a \in \mathcal{A}_0 \right\} \right)$$

is jointly asymptotically normal, and the asymptotic covariance matrix is

$$\begin{pmatrix} \Sigma_1^{\text{proj}} & \Sigma_{12}^{\text{proj}} \\ \{\Sigma_{12}^{\text{proj}}\}^\top & \Sigma_2^{\text{proj}} \end{pmatrix}, \tag{S4.22}$$

where

$$\begin{aligned}
\Sigma_1^{\text{proj}} &= \text{diag}\left\{ \sum_{k\in\mathcal{K}} \frac{p_{[k]}}{\pi_{[k]a}} \sigma^2_{[k]\tilde{r}^{\text{proj}}_{\text{glm}}(a)}, a \in \mathcal{A}_0 \right\}, \\
\Sigma_{12}^{\text{proj}} &= \left\{ \sum_{k\in\mathcal{K}} p_{[k]} \{\beta_{[k]}^{\text{proj}}(a')\}^\top \Sigma^\top_{[k]\Psi r^{\text{proj}}_{\text{glm}}(a)}, (a,a') \in \mathcal{A}_0 \times \mathcal{A}_0 \right\}, \\
\Sigma_2^{\text{proj}} &= \left\{ \sum_{k\in\mathcal{K}} p_{[k]} \{\beta_{[k]}^{\text{proj}}(a)\}^\top \Sigma_{[k]\Psi\Psi} \beta_{[k]}^{\text{proj}}(a'), (a,a') \in \mathcal{A}_0 \times \mathcal{A}_0 \right\}.
\end{aligned}$$

We notice that $\beta_{[k]}^{\text{proj}}(a)$ is derived by

$$\beta_{[k]}^{\text{proj}}(a) = \Sigma^{-1}_{[k]\Psi\Psi} \Sigma_{[k]\Psi Y(a)},$$

so we have

$$\begin{aligned}
\Sigma_{[k]\Psi r^{\text{proj}}_{\text{glm}}(a)} &= \text{Cov}\left(\tilde{\Psi}_{[k]}(X_i), \tilde{Y}_i(a) - \tilde{\Psi}^\top_{[k]}(X_i) \beta_{[k]}^{\text{proj}}(a) | A_i = a, B_i = k \right) \\
&= \Sigma_{[k]\Psi Y(a)} - E\left\{ \tilde{\Psi}_{[k]}(X_i) \tilde{\Psi}^\top_{[k]}(X_i) | B_i = k, A_i = a \right\} \beta_{[k]}^{\text{proj}}(a) \\
&= \Sigma_{[k]\Psi Y(a)} - \Sigma_{[k]\Psi\Psi} \beta_{[k]}^{\text{proj}}(a) \\
&= 0.
\end{aligned}$$

The third equality is because X_i is independent with A_i conditional on B_i. Therefore, we have

$$\left(\left\{ \sum_{k\in\mathcal{K}} \frac{1}{\pi_{[k]a}} \mathbb{L}^{(1)}_{n[k],r^{\text{proj}}_{i,\text{glm}}(a)}, a \in \mathcal{A}_0 \right\}, \left\{ \sum_{k\in\mathcal{K}} \mathbb{L}^{(1)}_{n[k],\Psi(a)}, a \in \mathcal{A}_0 \right\} \right)$$

is jointly asymptotically normal and the asymptotic covariance matrix is

$$\begin{pmatrix} \Sigma_1^{\text{proj}} & 0 \\ 0 & \Sigma_2^{\text{proj}} \end{pmatrix}. \tag{S4.23}$$

Then we combine (S4.17) and (S4.23) to reach our conclusion. □

S5 Additional simulation results

S5.1 Additional illustration under weak similarity case

In this section, we evaluate the performance of our proposed estimators under a weak similarity regime, explicitly defined as the scenario where the ℓ_1 estimation error of the transfer learning coefficient $\hat{\beta}_{[k]\text{tl}}(a)$ exceeds that of the target-only GLM coefficient $\hat{\beta}_{[k]\text{glm}}(a)$.

We consider a stratified block randomization design with three treatment groups and two strata (with a sample size ratio of $4:6$), where units are assigned with an allocation ratio of $1:1:1$ and a block size of 6. The potential outcomes in both the target and source datasets are generated

from the following logistic model:

$$\Pr(Y_i(a) = 1|X_i) = \text{expit}\Big\{\sum_{j=1}^{s} X_{ij}\beta_j\Big\}, \quad a = 0, 1, 2.$$

Here, the covariate vector $(X_{i1}, \ldots, X_{is})^\top$ consists of independent entries drawn from a Uniform$[-5, 5]$ distribution, with the true signal dimension $s \in \{2, 4, 6\}$. For the high-dimensional working model, we append $p - s$ independent noise covariates generated from a normal distribution $\mathcal{N}(0, 2)$ to reach a total dimension of $p = 100$.

We set the sample sizes to $n = 300$ for the target data and $n^{(s)} = 2700$ for the source data. The target parameters are set to $\beta_j = 2/s$ for $j = 1, \ldots, s$, producing an ℓ_1-norm of $\|\beta\|_1 = 2$. For the source data, we set $\beta_j^{(s)} = (2 + h)/s$, which results in a parameter distance of $\|\beta - \beta^{(s)}\|_1 = h$. The parameter h is chosen from a sequence ranging from 0 to 2, which explicitly controls the discrepancy between the source and target populations.

To identify the weak similarity regime, we first evaluate the average ratio of the ℓ_1-norm errors over $R = 1000$ replicates, defined as:

$$\text{Ratio} = \frac{1}{6}\sum_{a=0}^{2}\sum_{k=1}^{2}\frac{\|\hat{\beta}_{[k],\text{tl}}(a) - \beta\|_1}{\|\hat{\beta}_{[k]\text{glm}}(a) - \beta\|_1},$$

The weak similarity case can be illustrated as the settings where Ratio $\geq$ 1. Subsequently, we evaluate the estimation performance of our proposed

transfer learning estimators in terms of relative bias, sampling standard deviation, and empirical coverage probability. The results are as follows.

Figure S1 plots the empirical ℓ_1-norm error ratio across varying levels of the discrepancy parameter h. In this figure, the solid red horizontal line at 1.0 represents the target-only baseline, while the green dashed line illustrates the empirical error ratio of the transfer learning coefficient. The blue dash-dotted line denotes the ideal baseline ratio if the source data were perfectly compatible ($h = 0$). As expected, the error ratio of the transfer learning estimator monotonically increases as h grows, reflecting the impact of incorporating highly dissimilar source data. Notably, the critical intersection point where the transfer learning error surpasses the target-only error (Ratio ≥ 1) increases with the true signal dimension s increases. Therefore, when h increases, the weak similarity will degrade the performance on ℓ_1 error.

Under the same settings, the inference performance of our proposed estimators are presented in Figures S2, S3 and S4. First, we observe the robustness of the base transfer learning estimator ($\hat{\tau}_{\text{tl}}$) under weak similarity. Even when the transfer learning coefficient estimation is less accurate than the target-only GLM coefficient, $\hat{\tau}_{\text{tl}}$ consistently outperforms the baseline estimator $\hat{\tau}_{\text{glm}}$. When $s = 2$, $\hat{\tau}_{\text{tl}}$ exhibits slight under-coverage but still im-

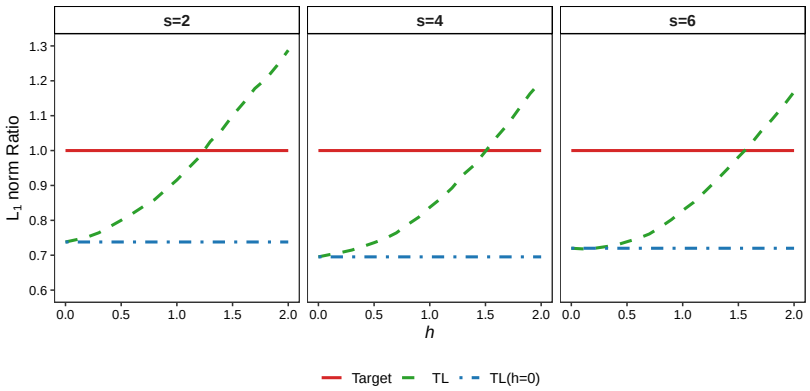


Figure S1: ℓ_1 norm ratio between each estimators.

proves upon $\hat{\tau}_{\text{glm}}$. As s increases, $\hat{\tau}_{\text{tl}}$ successfully reaches the nominal 95% coverage probability while maintaining superior efficiency compared to $\hat{\tau}_{\text{glm}}$.

Second, we evaluate the projection method under the extreme condition of a highly limited target sample size ($n = 300$). Theoretically, the projection step will guarantee efficiency improvements over the standard benchmark estimator $\hat{\tau}_{\text{ben}}$, which is also shown in Figures S2(B) and S3. However, because $\hat{\tau}_{\text{glm}}$ and $\hat{\tau}_{\text{tl}}$ already incorporate the generative working model, achieving further efficiency gains over these specific estimators via projection requires estimating the projection coefficients with sufficient precision on the target data. When the target sample size is exceptionally small, this estimation becomes noisy. Consequently, this finite-sample noise can occasionally render the projection estimator slightly less efficient than the unprojected estimator.

Finally, a noticeable fluctuation occurs in the empirical standard deviation for the target-only estimators ($\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{glm}}$, and $\hat{\tau}_{\text{glm}}^{\text{proj}}$) in the $s = 2$ scenario when $h > 1.7$. Because these benchmark estimators are theoretically independent of the source data and the parameter h, this fluctuation is a computational artifact rather than a theoretical breakdown. Specifically, the severely limited target units per treatment arm can lead to occasional complete separation in the logistic regression across the Monte Carlo replicates, causing extreme coefficient estimates that inflate the empirical variance. This artifact does not affect our overall conclusions; rather, it highlights that the proposed transfer learning estimators ($\hat{\tau}_{\text{tl}}$ and $\hat{\tau}_{\text{tl}}^{\text{proj}}$) effectively mitigate this computational instability through the regularizing effect of the source data, ensuring valid inference even at these boundary conditions.

S5.2 Decomposition of estimation performance between transfer learning and projection

To more clearly illustrate the individual contributions of transfer learning and the projection method, we provide a separated pairwise comparison of the estimators evaluated in the simulation study in Section 6 of the main text. Specifically, Figures S5 and S6 present detailed pairwise comparisons

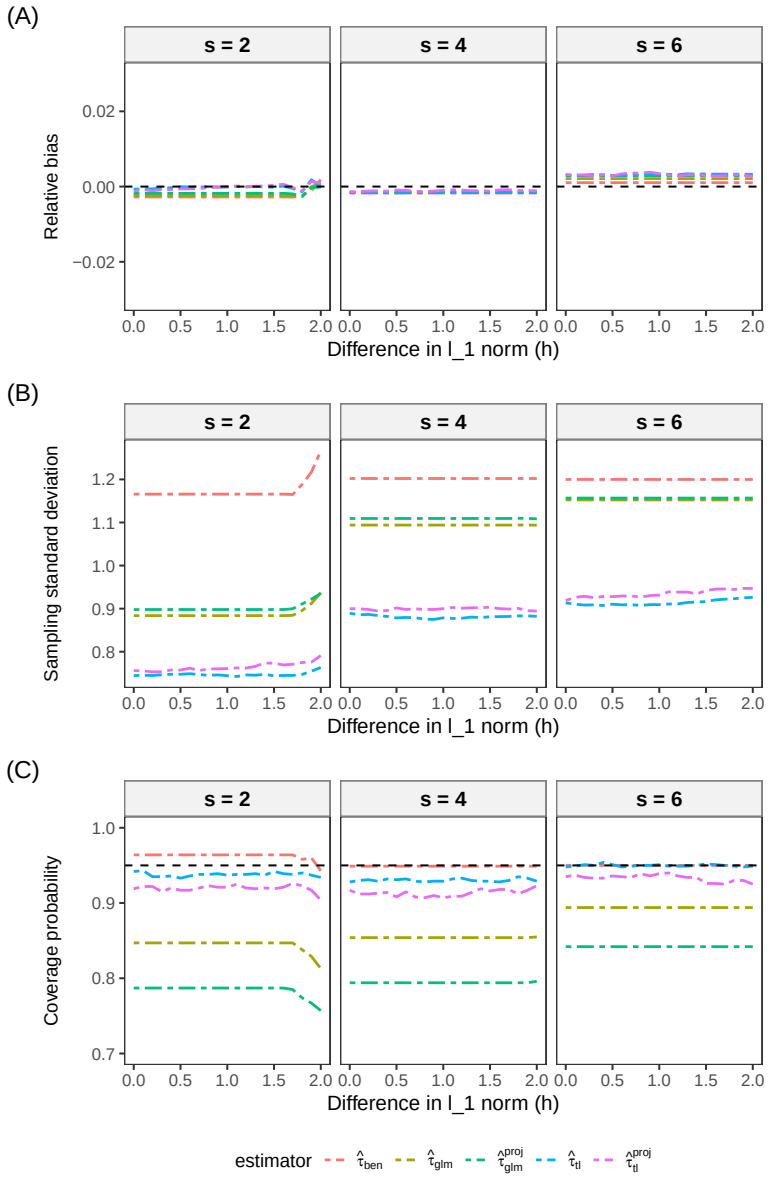


Figure S2: Simulation results for weak similarity case. Each column represents different s in the simulation settings. (A) Relative bias for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{glm}}$, $\hat{\tau}_{\text{glm}}^{\text{proj}}$, $\hat{\tau}_{\text{tl}}$ and $\hat{\tau}_{\text{tl}}^{\text{proj}}$. (B) Sampling standard deviation for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{glm}}$, $\hat{\tau}_{\text{glm}}^{\text{proj}}$, $\hat{\tau}_{\text{tl}}$ and $\hat{\tau}_{\text{tl}}^{\text{proj}}$. (C) Coverage probability for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{glm}}$, $\hat{\tau}_{\text{glm}}^{\text{proj}}$, $\hat{\tau}_{\text{tl}}$ and $\hat{\tau}_{\text{tl}}^{\text{proj}}$.

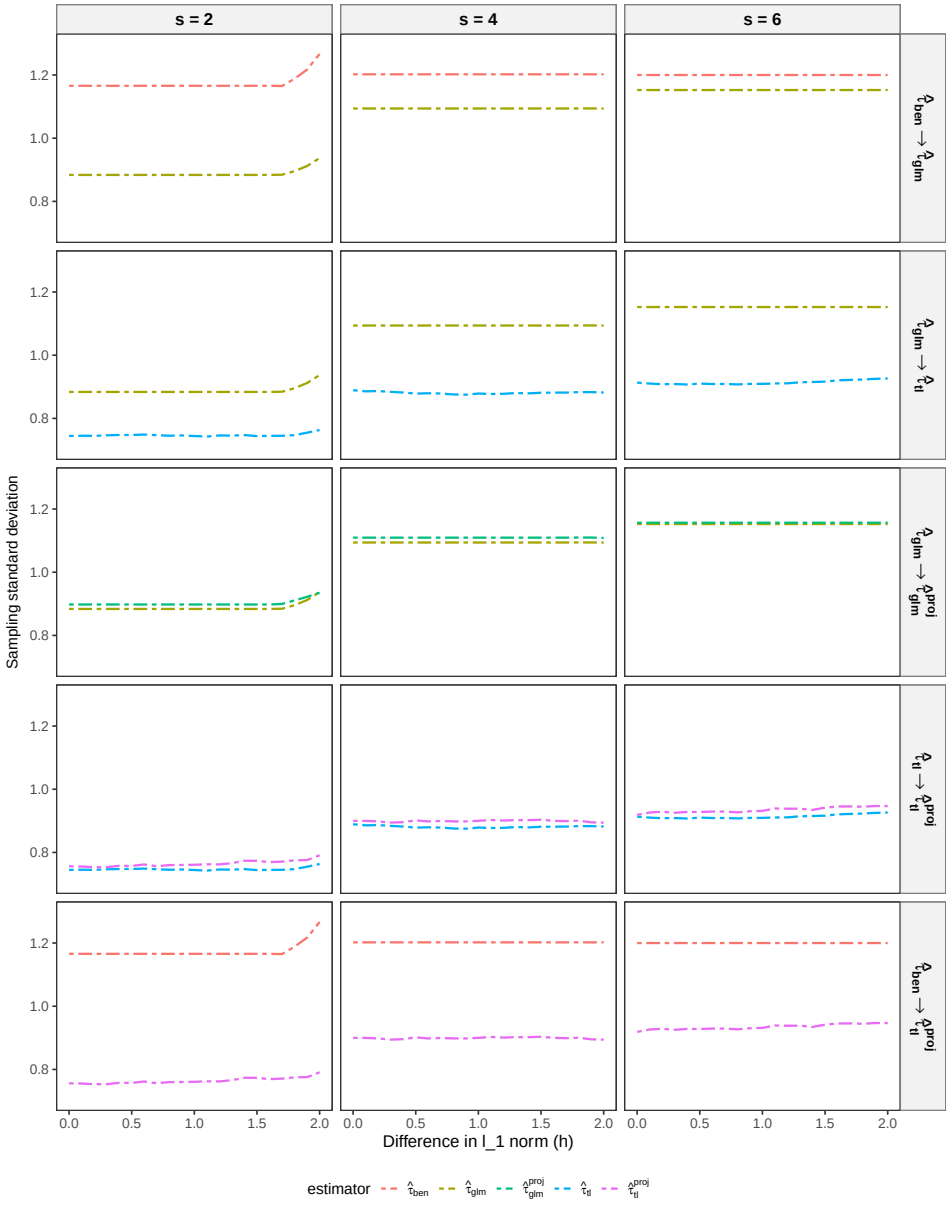


Figure S3: Detail comparison for the sampling standard deviation between each two estimators. Each column represents different s in the simulation settings.

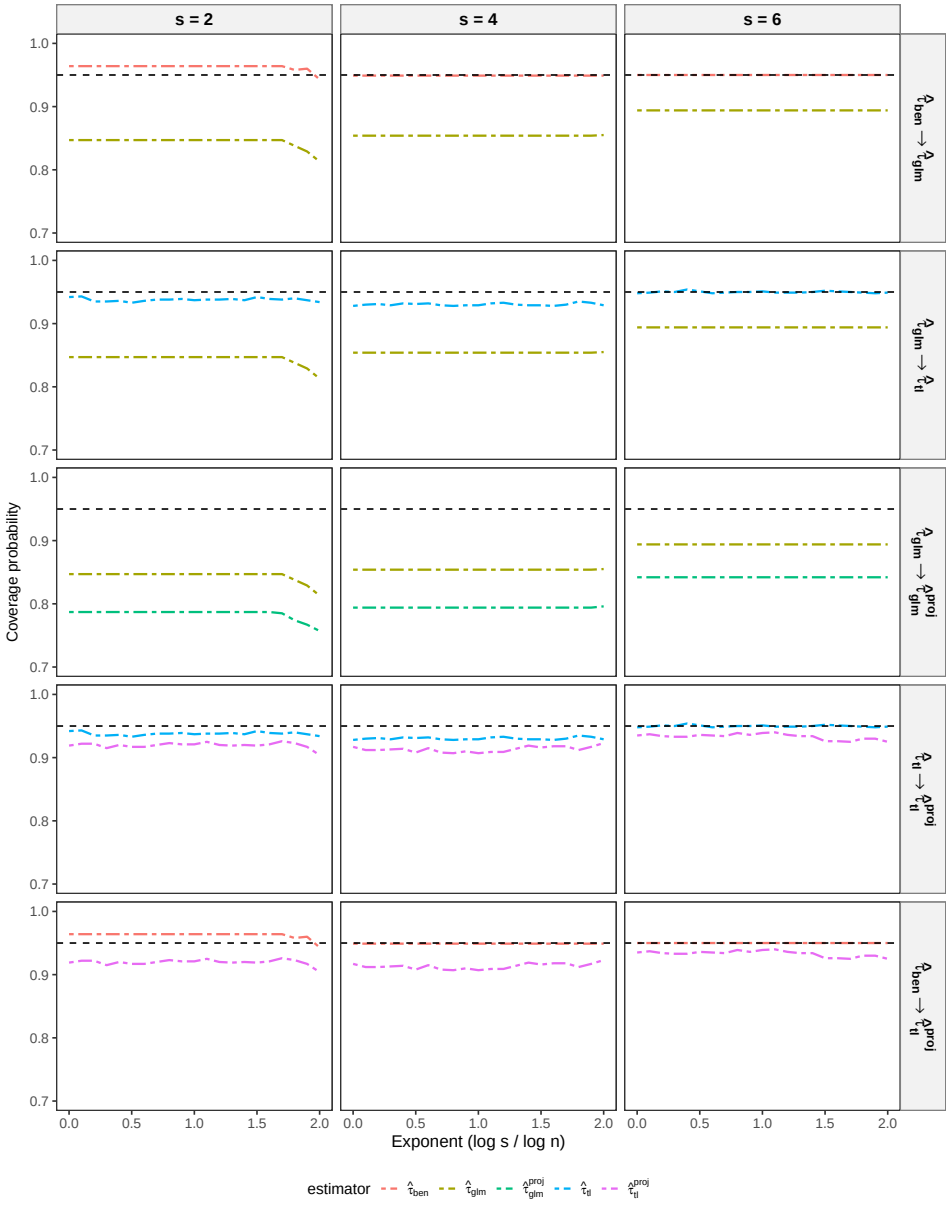


Figure S4: Detail comparison for the nominal 95% coverage probability between each two estimators. Each column represents different s in the simulation settings.

corresponding to Figure 1, while Figures S7 and S8 provide the detailed breakdowns for Figure 2. These figures isolate the incremental changes in sampling standard deviation and coverage probability, allowing the distinct effects of each methodological step to be evaluated directly. In Figures S5 and S7, the independent effect of the projection step is illustrated by comparing the GLM adjusted and the transfer learning estimators to their projected counterparts, namely $\hat{\tau}_{\text{glm}}$ versus $\hat{\tau}_{\text{glm}}^{\text{proj}}$ and $\hat{\tau}_{\text{tl}}$ versus $\hat{\tau}_{\text{tl}}^{\text{proj}}$. In both scenarios, the application of the projection method leads to a reduction in the sampling standard deviation, confirming the theoretical result in Corollary 1. Furthermore, the theoretical contribution of transfer learning is demonstrated in the separated coverage probability plots in Figures S6 and S8. While $\hat{\tau}_{\text{glm}}$ and $\hat{\tau}_{\text{glm}}^{\text{proj}}$ may suffer from severe under-coverage as the dimension of covariates s increases, the estimators incorporating transfer learning ($\hat{\tau}_{\text{tl}}$ and $\hat{\tau}_{\text{tl}}^{\text{proj}}$) successfully maintain coverage probabilities near the nominal 95% level across all evaluated settings. Ultimately, these separated comparisons clarify that the projection method serves as the primary driver for efficiency gains, whereas transfer learning plays the crucial role of guaranteeing valid inference and achieving nominal coverage in high-dimensional settings with small target sample sizes.

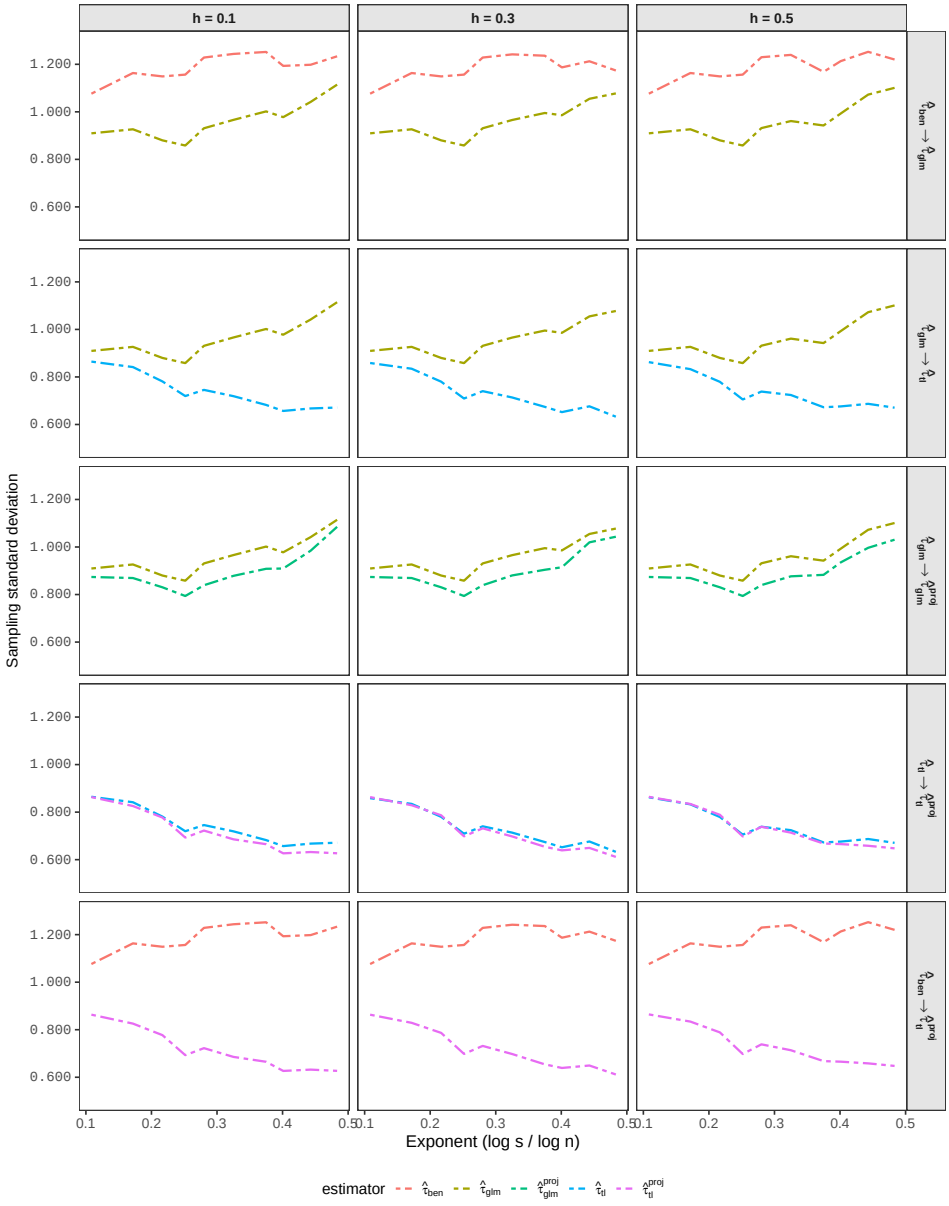


Figure S5: Detail comparison for sampling standard deviation between each two estimators under logistic regression model. Each column represents different h in the simulation settings.

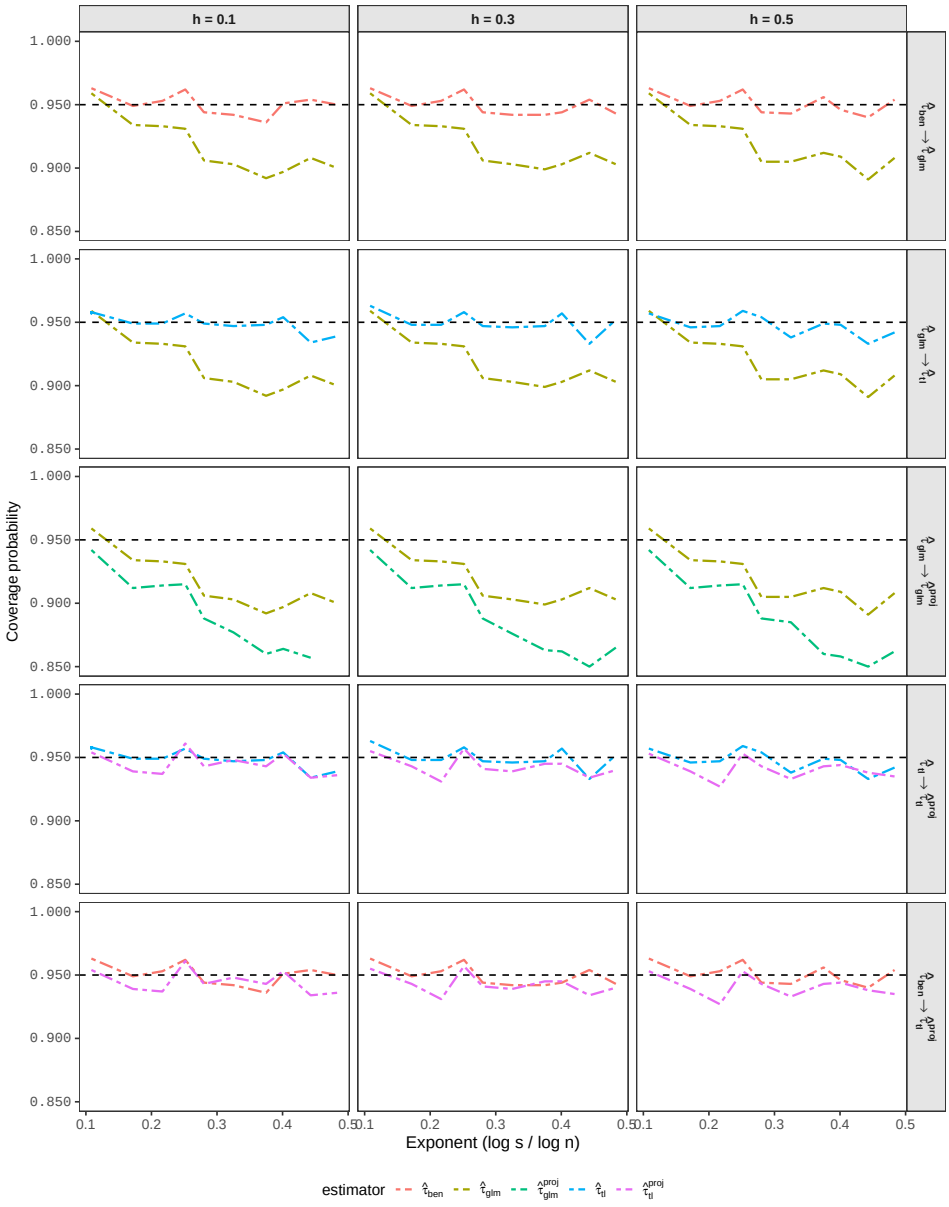


Figure S6: Detail comparison for nominal 95% coverage probability between each two estimators under logistic regression model. Each column represents different h in the simulation settings.

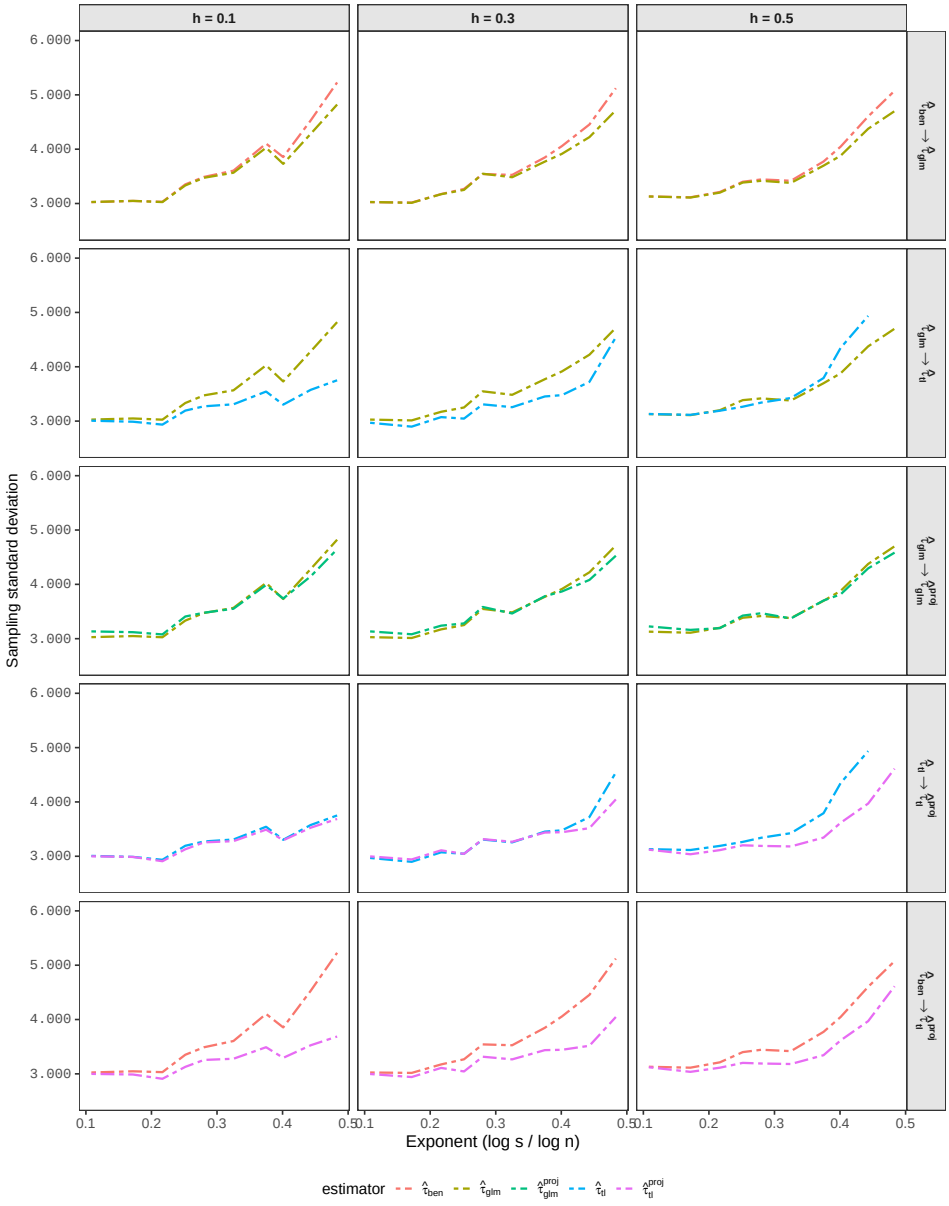


Figure S7: Detail comparison for sampling standard deviation between each two estimators under Poisson model. Each column represents different h in the simulation settings.

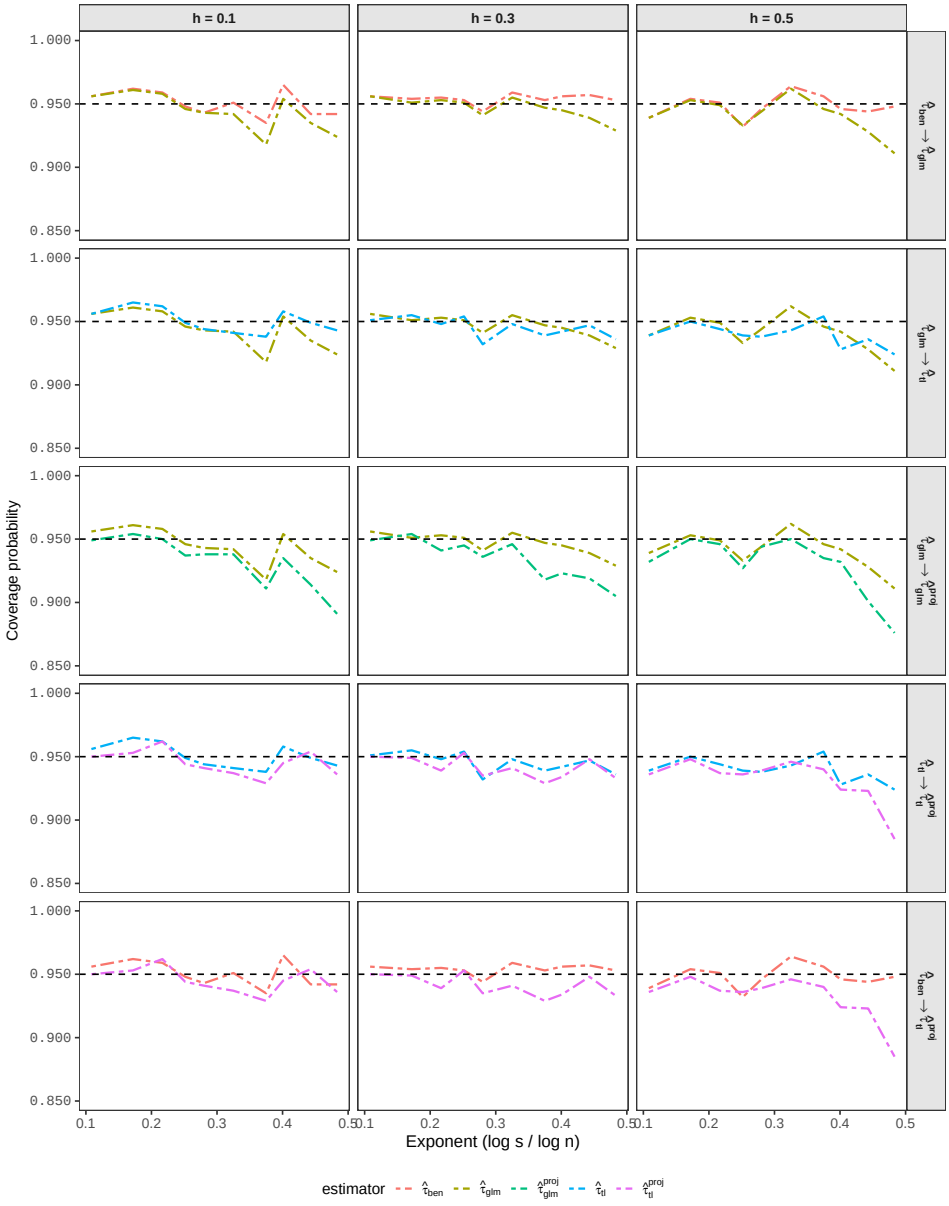


Figure S8: Detail comparison for nominal 95% coverage probability between each two estimators under Poisson model. Each column represents different h in the simulation settings.

S5.3 Additional illustration under misspecified data generation model

In this section, we evaluate the performance of our proposed estimators under a misspecified data generation model. We compare our methods with the benchmark estimator $\hat{\tau}_{\text{ben}}$ to demonstrate the robustness of the transfer learning and projection steps. The potential outcomes in both target and source data are generated via the following models:

$$\Pr(Y_i(a)=1|X_i)=\text{expit}\Big\{\mu_a+X_{i1}\beta_1(a)+\sum_{j=2}^{s}X_{ij}\beta_j(a)+\gamma_a\frac{1}{s-1}\sum_{j=2}^{s}X_{ij}^2\Big\},\quad a=0,1.$$
$$\Pr(Y_i(2)=1|X_i)=\text{expit}\Big\{\mu_2+X_{i1}\beta_1(2)+\sum_{j=2}^{s}X_{i1}X_{ij}\beta_j(a)\Big\}.$$

In both the target and source data, the common parameters are set as $\mu_0=0.7$, $\mu_1=\mu_2=0.2$, and $\gamma_0=-0.5$, $\gamma_1=0$. Let X_i be a s-dimensional covariate such that X_{i1} is a stratified random variable and $\{X_{ij}\}_{j=2}^{s}$ are regarded as additional baseline covariates. X_{i1} take values in $\{1,2\}$ with probabilities $\{0.4,0.6\}$ respectively. In addition, $(X_{i2},\ldots,X_{is})^\top$ is a covariate vector with independent entries, each drawn from a Uniform$[-5,5]$ distribution. We apply high-dimensional logistic regression as the working model. Specifically, in the working model, we use a covariate vector $X_i=(X_{i1},\ldots,X_{is},\ldots,X_{ip})^\top$ with a total dimension of $p=100$, where the

additional $p - s$ covariates follow a standard normal distribution independently. Due to the unmodeled quadratic term, the data generation model for $Y_i(0)$ is severely misspecified for both the target and source data, while the functional forms for $Y_i(1)$ and $Y_i(2)$ remain correctly specified.

We set $n = 600$ units in the target data and $n^{(s)} = 2400$ units in the source data. The units are assigned to each treatment group by stratified block randomization with ratio $1 : 1 : 1$. Let $s = \lceil n^r \rceil$, where $r \in \{0.16, 0.20, \ldots, 0.48, 0.52\}$. In the target data, we set $\beta_j(0) = 0.5$ and $\beta_j(1) = \beta_j(2) = 0.3$. For the source data, the parameters are generated as $\beta_j^{(s)}(a) = \beta_j(a) + hj/s$ for $j = 1, 2, \ldots, s$, and $a = 0, 1, 2$. The parameter $h \in \{0.1, 0.3, 0.5\}$ explicitly controls the source-target similarity, leading to a parameter distance of $\|\beta(a) - \beta^{(s)}(a)\|_1 = h(s+1)/2$. Larger values of h represent weaker similarity between the source and target populations. Our primary interest is to estimate the treatment effect between treatment groups 1 and 0. To this end, we evaluate the relative bias, sampling standard deviation, and nominal 95% coverage probabilities based on $R = 1000$ replicates for five estimators: the benchmark estimator $\hat{\tau}_{\text{ben}}$, the GLM-adjusted estimator $\hat{\tau}_{\text{glm}}$, the transfer learning estimator $\hat{\tau}_{\text{tl}}$, and their corresponding projection estimators $\hat{\tau}_{\text{glm}}^{\text{proj}}$ and $\hat{\tau}_{\text{tl}}^{\text{proj}}$. The overall results are presented in Figure S9, with separate pairwise comparisons for standard

deviation and coverage probability detailed in Figures S10 and S11.

The simulation results show behaviors consistent with the well-specified cases in Section 6 of the main text, while highlighting the robustness of our approach under model misspecification.

Regarding the relative bias, as shown in Figure S9(A), model misspecification introduces relative bias for the GLM-based estimators ($\hat{\tau}_{\text{glm}}$ and $\hat{\tau}_{\text{glm}}^{\text{proj}}$), which deviates further from zero as the dimension s increases. In contrast, the transfer learning estimators ($\hat{\tau}_{\text{tl}}$ and $\hat{\tau}_{\text{tl}}^{\text{proj}}$) remain unbiased, demonstrating that the transfer learning procedure effectively leverages source data even when the working model is misspecified.

The results in Figures S9(B) and S10 allow us to separate the effects of transfer learning and the projection step on efficiency. First, $\hat{\tau}_{\text{tl}}^{\text{proj}}$ consistently achieves the smallest sampling standard deviation across all settings, including less favorable scenarios with weaker similarity ($h = 0.5$). Second, the projection step provides consistent refinement regardless of the base estimator: $\hat{\tau}_{\text{glm}}^{\text{proj}}$ achieves efficiency gains over $\hat{\tau}_{\text{glm}}$ and $\hat{\tau}_{\text{ben}}$, and $\hat{\tau}_{\text{tl}}^{\text{proj}}$ further improves upon $\hat{\tau}_{\text{tl}}$. This indicates that the projection refinement (Corollary 1) is robust against model misspecification.

Finally, concerning the coverage probability, Figures S9(C) and S11 demonstrate that $\hat{\tau}_{\text{tl}}$ and $\hat{\tau}_{\text{tl}}^{\text{proj}}$ maintain coverage probabilities near the nom-

inal 95% level. Conversely, $\hat{\tau}_{\text{glm}}$ and $\hat{\tau}_{\text{glm}}^{\text{proj}}$ suffer from severe under-coverage as s grows. This confirms that the theoretical guarantees for valid inference hold under model misspecification.

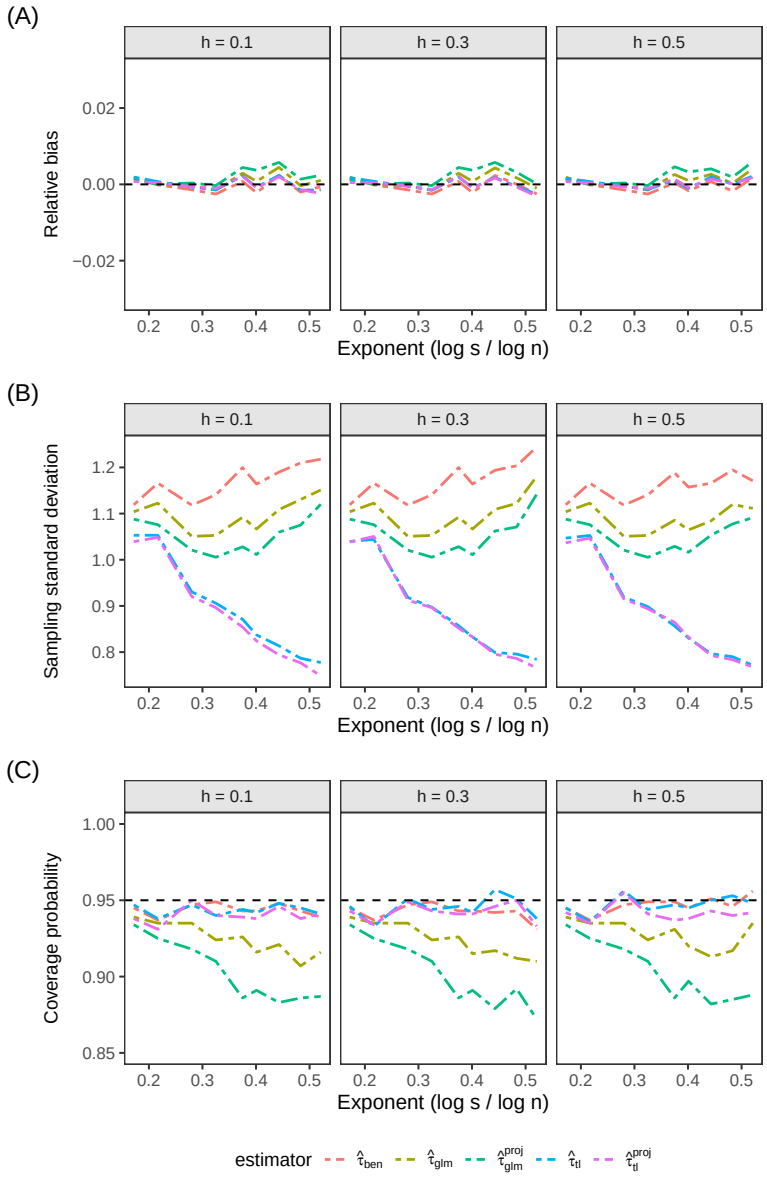


Figure S9: Simulation results for misspecified model. Each column represents different h in the simulation settings. (A) Relative bias for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{glm}}$, $\hat{\tau}^{\text{proj}}_{\text{glm}}$, $\hat{\tau}_{\text{tl}}$ and $\hat{\tau}^{\text{proj}}_{\text{tl}}$. (B) Sampling standard deviation for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{glm}}$, $\hat{\tau}^{\text{proj}}_{\text{glm}}$, $\hat{\tau}_{\text{tl}}$ and $\hat{\tau}^{\text{proj}}_{\text{tl}}$. (C) Coverage probability for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{glm}}$, $\hat{\tau}^{\text{proj}}_{\text{glm}}$, $\hat{\tau}_{\text{tl}}$ and $\hat{\tau}^{\text{proj}}_{\text{tl}}$.

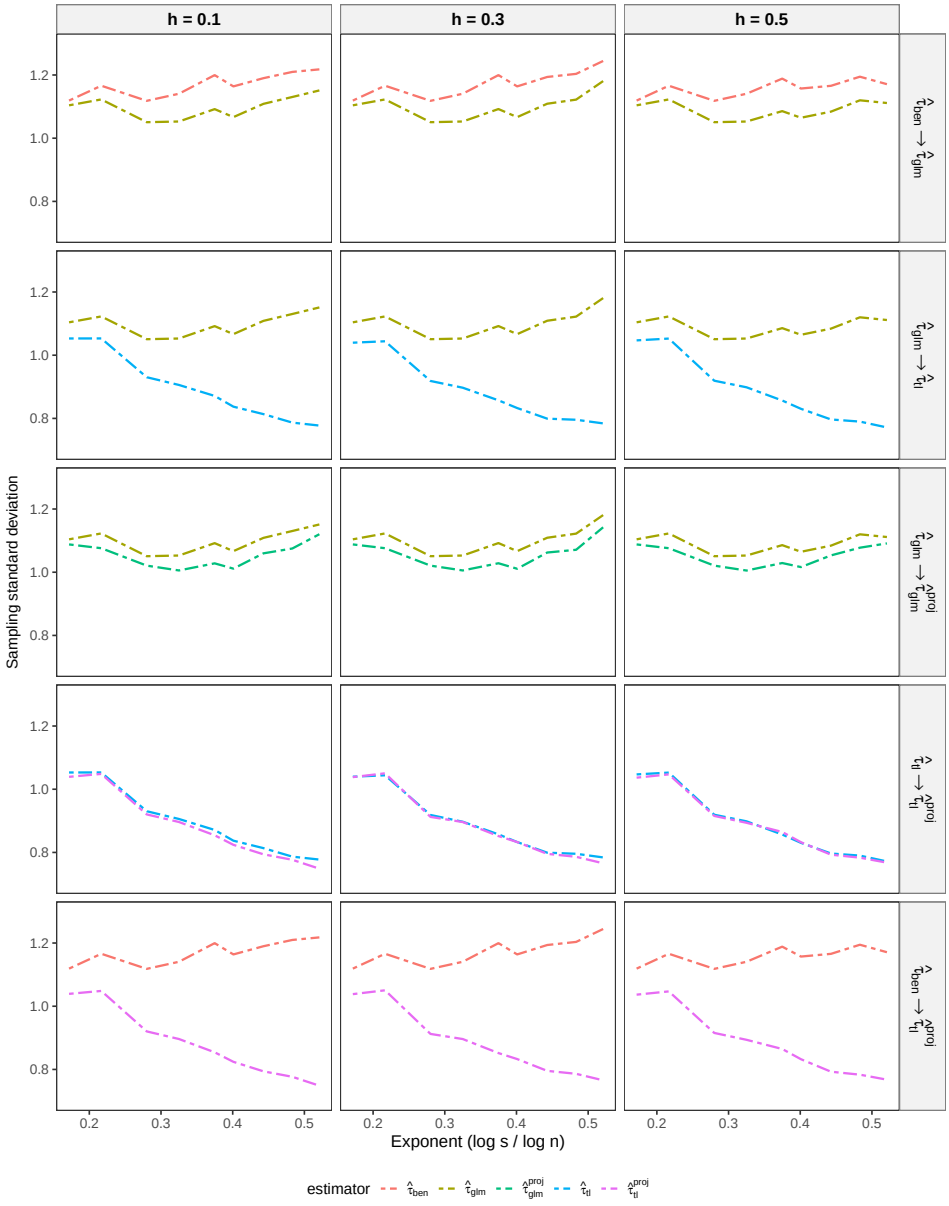


Figure S10: Detail comparison for sampling standard deviation between each two estimators under misspecified model. Each column represents different h in the simulation settings.

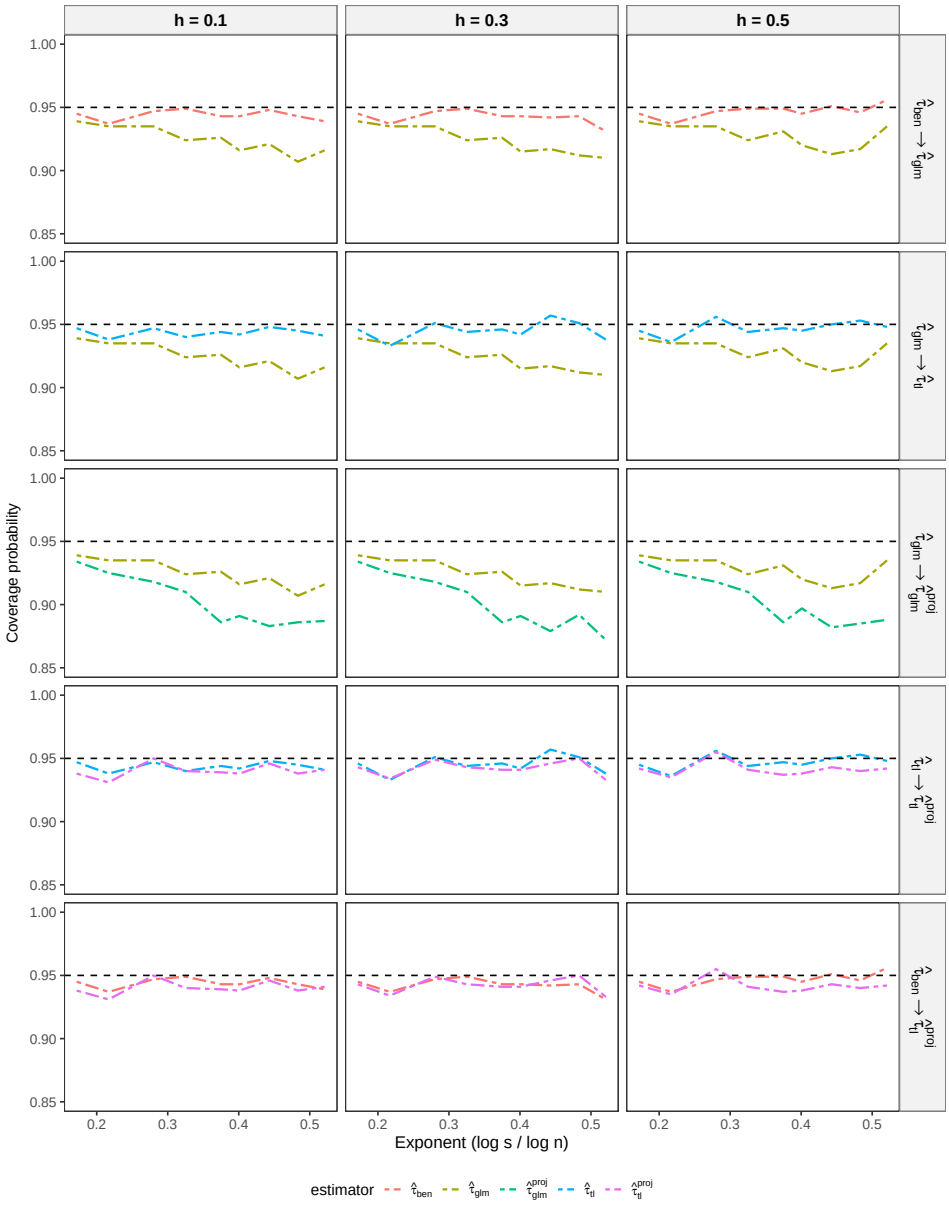


Figure S11: Detail comparison for the nominal 95% coverage probability between each two estimators under misspecified model. Each column represents different h in the simulation settings.