Supplementary Material for "Optimal Covariate Balancing Conditions in Propensity Score Estimation"

A Locally Semiparametric Efficient Estimator

For clarification, we reproduce the following definition of locally semiparametric efficient estimator given in Robins et al. (1994),

Definition A.1. Given a semiparametric model, say A, and an additional restriction R on the joint distribution of the data not imposed by the model, we say that an estimator $\widehat{\alpha}$ is locally semiparametric efficient in model A at R if $\widehat{\alpha}$ is a semiparametric estimator in model A whose asymptotic variance attains the semiparametric variance bound for model A when R is true.

In our setting, the semiparametric model A corresponds to the joint distribution of the observed data $(T_i, Y_i, \boldsymbol{X}_i)$ subject to the strong ignorability of the treatment assignment $\{Y_i(1), Y_i(0)\} \perp T_i \mid \boldsymbol{X}_i$; see Hahn (1998). The semiparametric variance bound for model A is V_{opt}. The restriction R is the intersection of R_1 and R_2 (denoted by $R_1 \cap R_2$), where R_1 is the model that satisfies the first condition in Theorem 3.1 (i.e., the propensity score is correctly specified) and R_2 is the model that satisfies the second condition in Theorem 3.1 (i..e, $K(\boldsymbol{X}_i) = \boldsymbol{\alpha}_1^\top \mathbf{M}_1 \boldsymbol{h}_1(\boldsymbol{X}_i)$ and $L(\boldsymbol{X}_i) = \boldsymbol{\alpha}_2^\top \mathbf{M}_2 \boldsymbol{h}_2(\boldsymbol{X}_i))$. In Corollary 3.2, we show that the asymptotic variance of our estimator of ATE $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is V_{opt} when $R_1 \cap R_2$ is true. From the above definition of locally semiparametric efficient estimator, we can claim that $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is locally semiparametric efficient at $R_1 \cap R_2$.

B Preliminaries

To simplify the notation, we use $\pi_i^* = \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)$ and $\pi_i^o = \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)$. For any vector $\boldsymbol{C} \in \mathbb{R}^K$, we denote $|\boldsymbol{C}| = (|C_1|, ..., |C_K|)^\top$ and write $\boldsymbol{C} \leq \boldsymbol{B}$ for $C_k \leq B_k$ for any $1 \leq k \leq K$.

Assumption B.1. (Regularity Conditions for CBPS in Section 2)

1. There exists a positive definite matrix $\mathbf{W}^*$ such that $\widehat{\mathbf{W}} \xrightarrow{p} \mathbf{W}^*$.
2. The minimizer $\boldsymbol{\beta}^o = \mathrm{argmin}_{\boldsymbol{\beta}} \mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}))^\top \mathbf{W}^* \mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}))$ is unique.
3. $\boldsymbol{\beta}^o \in \mathrm{int}(\Theta)$, where Θ is a compact set.

4. $\pi_{\boldsymbol{\beta}}(\boldsymbol{X})$ is continuous in $\boldsymbol{\beta}$.

5. There exists a constant $0 < c_0 < 1/2$ such that with probability tending to one, $c_0 \leq \pi_{\boldsymbol{\beta}}(\boldsymbol{X}) \leq 1 - c_0$, for any $\boldsymbol{\beta} \in \mathrm{int}(\Theta)$.

6. $\mathbb{E}|f_j(\boldsymbol{X})| < \infty$ for $1 \leq j \leq m$ and $\mathbb{E}|Y(1)|^2 < \infty$, $\mathbb{E}|Y(0)|^2 < \infty$.

7. $\mathbf{G}^* := \mathbb{E}(\partial \boldsymbol{g}(\boldsymbol{\beta}^o)/\partial \boldsymbol{\beta})$ exists and there is a q-dimensional function $C(\boldsymbol{X})$ and a small constant $r > 0$ such that $\sup_{\boldsymbol{\beta} \in \mathbb{B}_r(\boldsymbol{\beta}^o)} |\partial \pi_{\boldsymbol{\beta}}(\boldsymbol{X})/\partial \boldsymbol{\beta}| \leq C(\boldsymbol{X})$ and $\mathbb{E}(|f_j(\boldsymbol{X})|C(\boldsymbol{X})) < \infty$ for $1 \leq j \leq m$, where $\mathbb{B}_r(\boldsymbol{\beta}^o)$ is a ball in $\mathbb{R}^q$ with radius r and center $\boldsymbol{\beta}^o$. In addition, $\mathbb{E}(|Y|C(\boldsymbol{X})) < \infty$.

8. $\mathbf{G}^{*\top}\mathbf{W}^*\mathbf{G}^*$ and $\mathbb{E}(\boldsymbol{g}_{\boldsymbol{\beta}^o}(T_i, \boldsymbol{X}_i)\boldsymbol{g}_{\boldsymbol{\beta}^o}(T_i, \boldsymbol{X}_i)^\top)$ are nonsingular.

9. In the locally misspecified model (2.1), assume $|u(\boldsymbol{X}; \boldsymbol{\beta}^*)| \leq C$ almost surely for some constant $C > 0$.

Lemma B.1 (Lemma 2.4 in Newey and McFadden (1994))**.** Assume that the data Z_i are i.i.d., Θ is compact, $a(Z, \theta)$ is continuous for $\theta \in \Theta$, and there is $D(Z)$ with $|a(Z, \theta)| \leq D(Z)$ for all $\theta \in \Theta$ and $\mathbb{E}(D(Z)) < \infty$, then $\mathbb{E}(a(Z, \theta))$ is continuous and $\sup_{\theta \in \Theta} |n^{-1} \sum_{i=1}^n a(Z_i, \theta) - \mathbb{E}(a(Z, \theta))| \xrightarrow{p} 0$.

Lemma B.2. Under Assumption B.1 (or Assumptions 3.1), we have $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^o$.

Proof of Lemma B.2. The proof of $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^o$ follows from Theorem 2.6 in Newey and McFadden (1994). Note that their conditions (i)–(iii) follow directly from Assumption 3.1 (1)–(4). We only need to verify their condition (iv), i.e., $\mathbb{E}(\sup_{\boldsymbol{\beta} \in \Theta} |g_{\boldsymbol{\beta} j}(T_i, \boldsymbol{X}_i)|) < \infty$ where

$$g_{\boldsymbol{\beta} j}(T_i, \boldsymbol{X}_i) = \Big(\frac{T_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - \frac{1 - T_i}{1 - \pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}\Big) f_j(\boldsymbol{X}_i),$$

By Assumption B.1 (5), we have $|g_{\boldsymbol{\beta} j}(T_i, \boldsymbol{X}_i)| \leq 2|f_j(\boldsymbol{X}_i)|/c_0$ and thus $\mathbb{E}(\sup_{\boldsymbol{\beta} \in \Theta} |g_{\boldsymbol{\beta} j}(T_i, \boldsymbol{X}_i)|) < \infty$ by Assumption B.1 (6). In addition, for the proof of Theorem 3.1, we similarly verify the following conditions to prove this lemma for the oCBPS estimator, i.e., $\mathbb{E}(\sup_{\boldsymbol{\beta} \in \Theta} |g_{1\boldsymbol{\beta} j}(T_i, X_i)|) < \infty$ and $\mathbb{E}(\sup_{\boldsymbol{\beta} \in \Theta} |g_{2\boldsymbol{\beta} j}(T_i, \boldsymbol{X}_i)|) < \infty$, where

$$g_{1\boldsymbol{\beta} j}(T_i, \boldsymbol{X}_i) = \Big(\frac{T_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - \frac{1 - T_i}{1 - \pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}\Big) h_{1j}(\boldsymbol{X}_i), \quad \text{and} \quad g_{2\boldsymbol{\beta} j}(T_i, \boldsymbol{X}_i) = \Big(\frac{T_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - 1\Big) h_{2j}(\boldsymbol{X}_i).$$

We have $|g_{1\boldsymbol{\beta} j}(T_i, \boldsymbol{X}_i)| \leq 2|h_{1j}(\boldsymbol{X}_i)|/c_0$ and thus $\mathbb{E}(\sup_{\boldsymbol{\beta} \in \Theta} |g_{1\boldsymbol{\beta} j}(T_i, \boldsymbol{X}_i)|) < \infty$. Similarly, we can prove $\mathbb{E}(\sup_{\boldsymbol{\beta} \in \Theta} |g_{2\boldsymbol{\beta} j}(T_i, \boldsymbol{X}_i)|) < \infty$. This completes the proof. □

Lemma B.3. Under Assumption B.1 (or Assumptions 3.1 and 3.2), we have

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o) = -(\boldsymbol{H}_{\mathbf{f}}^{*\top} \mathbf{W}^* \boldsymbol{H}_{\mathbf{f}}^*)^{-1} n^{1/2} \boldsymbol{H}_{\mathbf{f}}^{*\top} \mathbf{W}^* \bar{\boldsymbol{g}}_{\boldsymbol{\beta}^o}(\boldsymbol{T}, \boldsymbol{X}) + o_p(1), \tag{B.1}$$

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o) \xrightarrow{d} N(0, (\boldsymbol{H}_{\mathbf{f}}^{*\top} \mathbf{W}^* \boldsymbol{H}_{\mathbf{f}}^*)^{-1} \boldsymbol{H}_{\mathbf{f}}^{*\top} \mathbf{W}^* \boldsymbol{\Omega} \mathbf{W}^* \boldsymbol{H}_{\mathbf{f}}^* (\boldsymbol{H}_{\mathbf{f}}^{*\top} \mathbf{W}^* \boldsymbol{H}_{\mathbf{f}}^*)^{-1}), \tag{B.2}$$

where $\Omega = \mathrm{Var}(\boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i, \boldsymbol{X}_i))$. If the propensity score model is correctly specified with $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)$ and $\mathbf{W}^* = \boldsymbol{\Omega}^{-1}$ holds, then $n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o) \xrightarrow{d} N(0, (\boldsymbol{H}_{\mathbf{f}}^{*\top} \boldsymbol{\Omega}^{-1} \boldsymbol{H}_{\mathbf{f}}^*)^{-1})$.

Proof. The proof of (B.1) and (B.2) follows from Theorem 3.4 in Newey and McFadden (1994). Note that their conditions (i), (ii), (iii) and (v) are directly implied by our Assumption B.1 (3), (4), (2) and Assumption B.1 (1), respectively. In addition, their condition (iv), that is, $\mathbb{E}(\sup_{\boldsymbol{\beta} \in \mathcal{N}} |\partial \boldsymbol{g}_{\boldsymbol{\beta}^o}(T_i, \boldsymbol{X}_i)/\partial \beta_j|) < \infty$ for some small neighborhood $\mathcal{N}$ around $\boldsymbol{\beta}^o$, is also implied by our Assumption B.1. To see this, by Assumption B.1 some simple calculations show that

$$\sup_{\boldsymbol{\beta} \in \mathcal{N}} \Big| \frac{\partial \boldsymbol{g}_{\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)}{\partial \beta_j} \Big| \le \Big(\frac{T_i |\mathbf{f}(\boldsymbol{X}_i)|}{c_0^2} + \frac{(1 - T_i)|\mathbf{f}(\boldsymbol{X}_i)|}{c_0^2} \Big) \sup_{\boldsymbol{\beta} \in \mathcal{N}} \Big| \frac{\partial \pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}{\partial \beta_j} \Big| \le C_j(\boldsymbol{X}) |\mathbf{f}(\boldsymbol{X}_i)| / c_0^2,$$

for $\mathcal{N} \in \mathbb{B}_r(\boldsymbol{\beta}^o)$. Hence, $\mathbb{E}(\sup_{\boldsymbol{\beta} \in \mathcal{N}} |\partial \boldsymbol{g}_{\boldsymbol{\beta}^o}(T_i, \boldsymbol{X}_i)/\partial \beta_j|) < \infty$, by Assumption B.1 (7). Thus, condition (iv) in Theorem 3.4 in Newey and McFadden (1994) holds. In order to apply this lemma to the proofs in Section 3, we need to further verify this condition for $\boldsymbol{g}_{\boldsymbol{\beta}}(\cdot) = (\boldsymbol{g}_{1\boldsymbol{\beta}}^\top(\cdot), \boldsymbol{g}_{2\boldsymbol{\beta}}^\top(\cdot))^\top$, where

$$\boldsymbol{g}_{1\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i) = \Big(\frac{T_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - \frac{1 - T_i}{1 - \pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} \Big) \boldsymbol{h}_1(\boldsymbol{X}_i), \quad \text{and} \quad \boldsymbol{g}_{2\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i) = \Big(\frac{T_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - 1 \Big) \boldsymbol{h}_2(\boldsymbol{X}_i).$$

To this end, by Assumption 3.1 some simple calculations show that when

$$\sup_{\boldsymbol{\beta} \in \mathcal{N}} \Big| \frac{\partial \boldsymbol{g}_{1\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)}{\partial \beta_j} \Big| \le \Big(\frac{T_i |\boldsymbol{h}_1(\boldsymbol{X}_i)|}{c_0^2} + \frac{(1 - T_i)|\boldsymbol{h}_1(\boldsymbol{X}_i)|}{c_0^2} \Big) \sup_{\boldsymbol{\beta} \in \mathcal{N}} \Big| \frac{\partial \pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}{\partial \beta_j} \Big| \le C_j(\boldsymbol{X}) |\boldsymbol{h}_1(\boldsymbol{X}_i)| / c_0^2,$$

for $\mathcal{N} \in \mathbb{B}_r(\boldsymbol{\beta}^o)$. Hence, $\mathbb{E}(\sup_{\boldsymbol{\beta} \in \mathcal{N}} |\partial \boldsymbol{g}_{1\boldsymbol{\beta}^o}(T_i, \boldsymbol{X}_i)/\partial \beta_j|) < \infty$, by Assumption 3.1 (7). Following the similar arguments, we can prove that $\mathbb{E}(\sup_{\boldsymbol{\beta} \in \mathcal{N}} |\partial \boldsymbol{g}_{2\boldsymbol{\beta}^o}(T_i, \boldsymbol{X}_i)/\partial \beta_j|) < \infty$ holds. This completes the proof of (B.2). As shown in Lemma B.2, if $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)$ holds, the asymptotic normality of $n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)$ follows from (B.2). The proof is complete. □

C Proof of Results in Section 2

C.1 Proof of Theorem 2.1

Proof. First, we derive the bias of $\widehat{\boldsymbol{\beta}}$. By the arguments in the proof of Lemma B.3, we can show that $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^o + O_p(n^{-1/2})$, where $\boldsymbol{\beta}^o$ satisfies $\boldsymbol{\beta}^o = \mathrm{argmin}_{\boldsymbol{\beta}} \mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}))^\top \mathbf{W}^* \mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}))$. Let

$u_i^* = u(\boldsymbol{X}_i;\boldsymbol{\beta}^*)$. By the propensity score model and the fact that $|u(\boldsymbol{X}_i;\boldsymbol{\beta}^*)|$ is a bounded random variable and $\mathbb{E}|f_j(\boldsymbol{X}_i)| < \infty$, we can show that

$$\mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^o}) = \mathbb{E}\Big\{\frac{\pi_i^*(1+\xi u_i^*)\mathbf{f}(\boldsymbol{X}_i)}{\pi_i^o} - \frac{(1-\pi_i^*-\xi\pi_i^* u_i^*)\mathbf{f}(\boldsymbol{X}_i)}{1-\pi_i^o}\Big\} + O(\xi^2).$$

In addition, following the similar calculation, we have $\mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*}) = O(\xi)$. Therefore,

$$\lim_{n\to\infty} \mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*}(\boldsymbol{T},\boldsymbol{X}))^\top \mathbf{W}^* \mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*}(\boldsymbol{T},\boldsymbol{X})) = 0.$$

Clearly, this quadratic form $\mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X}))^\top \mathbf{W}^* \mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X}))$ must be nonnegative for any $\boldsymbol{\beta}$. By the uniqueness of $\boldsymbol{\beta}^o$, we have $\boldsymbol{\beta}^o - \boldsymbol{\beta}^* = o(1)$. Therefore, we can expand π_i^o around π_i^*, which yields

$$\mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^o}) = \mathbb{E}\Big\{\xi\Big(\frac{u_i^*}{1-\pi_i^*}\Big)\mathbf{f}(\boldsymbol{X}_i) + \boldsymbol{H}_{\mathbf{f}}^*(\boldsymbol{\beta}^o - \boldsymbol{\beta}^*)\Big\} + O(\xi^2 + \|\boldsymbol{\beta}^o - \boldsymbol{\beta}^*\|_2^2).$$

This implies that the bias of $\boldsymbol{\beta}^o$ is

$$\boldsymbol{\beta}^o - \boldsymbol{\beta}^* = -\xi(\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\boldsymbol{H}_{\mathbf{f}}^*)^{-1}\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\mathbb{E}\Big\{\Big(\frac{u_i^*}{1-\pi_i^*}\Big)\mathbf{f}(\boldsymbol{X}_i)\Big\} + O(\xi^2). \tag{C.1}$$

Our next step is to derive the bias of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$. Similar to the proof of Theorem 3.2, we have

$$\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu = \frac{1}{n}\sum_{i=1}^n D_i + \mathbf{H}_y^{*\top}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o) + o_p(n^{-1/2}),$$

where

$$D_i = \frac{T_i Y_i(1)}{\pi_i^o} - \frac{(1-T_i)Y_i(0)}{1-\pi_i^o} - \mu,$$

and

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o) = -(\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\boldsymbol{H}_{\mathbf{f}}^*)^{-1} n^{1/2}\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^o}(\boldsymbol{T},\boldsymbol{X}) + o_p(1).$$

In addition, following the similar steps, we can show that $\mathbb{E}(D_i) = Bn^{-1/2} + o(n^{-1/2})$. Thus,

$$\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu = \frac{1}{n}\sum_{i=1}^n\{D_i - \mathbb{E}(D_i)\} + \mathbf{H}_y^{*\top}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o) + Bn^{-1/2} + o_p(n^{-1/2}).$$

Then the asymptotic normality of $\sqrt{n}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu)$ follows from the above asymptotic expansion and the central limit theorem. This completes the proof. □

C.2 Proof of Corollary 2.1

Proof. When $\boldsymbol{H}_{\mathbf{f}}^*$ is invertible, it is easy to show the bias term can be written as

$$B \;=\; \left[\mathbb{E}\left\{\frac{u(\boldsymbol{X}_i;\boldsymbol{\beta}^*)(K(\boldsymbol{X}_i)+(1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))L(\boldsymbol{X}_i))}{1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}\right\} + \boldsymbol{H}_y^*\boldsymbol{H}_{\mathbf{f}}^{*-1}\mathbb{E}\left(\frac{u(\boldsymbol{X}_i;\boldsymbol{\beta}^*)\mathbf{f}(\boldsymbol{X}_i)}{1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}\right)\right],$$

when the propensity score model is locally misspecified. If we choose the balancing function $\mathbf{f}(\boldsymbol{X})$ such that $\boldsymbol{\alpha}^\top\mathbf{f}(\boldsymbol{X}) = K(\boldsymbol{X}_i) + (1-\pi_i^*)L(\boldsymbol{X}_i)$ for some $\boldsymbol{\alpha}\in\mathbb{R}^q$, we have

$$\begin{aligned}\boldsymbol{H}_y^* &= -\mathbb{E}\left(\frac{K(\boldsymbol{X}_i)+(1-\pi_i^*)L(\boldsymbol{X}_i)}{\pi_i^*(1-\pi_i^*)}\cdot\frac{\partial\pi_i^*}{\partial\boldsymbol{\beta}}\right) = -\boldsymbol{\alpha}^\top\mathbb{E}\Big(\frac{\mathbf{f}(\boldsymbol{X}_i)}{\pi_i^*(1-\pi_i^*)}\Big(\frac{\partial\pi_i^*}{\partial\boldsymbol{\beta}}\Big)^\top\Big),\\ \boldsymbol{H}_{\mathbf{f}}^* &= -\mathbb{E}\left(\frac{\partial g_{\boldsymbol{\beta}^*}(T_i,\boldsymbol{X}_i)}{\partial\boldsymbol{\beta}}\right) = -\mathbb{E}\left(\frac{\mathbf{f}(\boldsymbol{X}_i)}{\pi_i^*(1-\pi_i^*)}\Big(\frac{\partial\pi_i^*}{\partial\boldsymbol{\beta}}\Big)^\top\right).\end{aligned}$$

So the bias becomes

$$B \;=\; \left[\boldsymbol{\alpha}^\top\mathbb{E}\left\{\frac{u(\boldsymbol{X}_i;\boldsymbol{\beta}^*)\mathbf{f}(\boldsymbol{X}_i)}{1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}\right\} + \boldsymbol{\alpha}^\top\boldsymbol{H}_{\mathbf{f}}^*(\boldsymbol{H}_{\mathbf{f}}^*)^{-1}\mathbb{E}\left(\frac{u(\boldsymbol{X}_i;\boldsymbol{\beta}^*)\mathbf{f}(\boldsymbol{X}_i)}{1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}\right)\right] = 0.$$

This proves that $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is first order unbiased. □

C.3 Proof of Corollary 2.2

Proof. Recall that even if the propensity score mode is known or pre-specified, the minimum asymptotic variance over the class of regular estimators is given by V_{opt}. In the following, we will verify that with the optimal choice of $\mathbf{f}(\boldsymbol{X})$ our estimator has asymptotic variance V_{opt}.

The asymptotic variance bound V_{opt} can be written as, $V_{\text{opt}} = \Sigma_\mu - \boldsymbol{\alpha}^\top\boldsymbol{\Omega}\boldsymbol{\alpha}$, where

$$\boldsymbol{\Omega} = \mathbb{E}(\boldsymbol{g}_{\boldsymbol{\beta}^o}(T_i,\boldsymbol{X}_i)\boldsymbol{g}_{\boldsymbol{\beta}^o}(T_i,\boldsymbol{X}_i)^\top) = \mathbb{E}\left(\frac{\mathbf{f}(\boldsymbol{X}_i)\mathbf{f}(\boldsymbol{X}_i)^\top}{\pi_i^*(1-\pi_i^*)}\right).$$

We can write the asymptotic variance of our estimator as

$$V = \Sigma_\mu + 2\boldsymbol{H}_y^{*\top}\boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} + \boldsymbol{H}_y^{*\top}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{H}_y^*,$$

where

$$\begin{aligned}
&\boldsymbol{H}_y^* = \mathbb{E}\left(\frac{\partial \mu_{\boldsymbol{\beta}^*}(T_i, Y_i, \boldsymbol{X}_i)}{\partial \boldsymbol{\beta}}\right) = -\mathbb{E}\left(\frac{K(\boldsymbol{X}_i) + (1-\pi_i^*)L(\boldsymbol{X}_i)}{\pi_i^*(1-\pi_i^*)}\frac{\partial \pi_i^*}{\partial \boldsymbol{\beta}}\right),\\
&\boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} = -(\boldsymbol{H}_{\mathbf{f}}^*)^{-1}\operatorname{Cov}(\mu_{\boldsymbol{\beta}^*}(T_i, Y_i, \boldsymbol{X}_i), \boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i, \boldsymbol{X}_i)),\\
&\boldsymbol{H}_{\mathbf{f}}^* = \mathbb{E}\left(\frac{\partial \boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i, \boldsymbol{X}_i)}{\partial \boldsymbol{\beta}}\right) = -\mathbb{E}\left(\frac{\mathbf{f}(\boldsymbol{X}_i)}{\pi_i^*(1-\pi_i^*)}\left(\frac{\partial \pi_i^*}{\partial \boldsymbol{\beta}}\right)^\top\right),\\
&\operatorname{Cov}(\mu_{\boldsymbol{\beta}^*}(T_i, Y_i, \boldsymbol{X}_i), \boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i, \boldsymbol{X}_i)) = \mathbb{E}\left(\frac{K(\boldsymbol{X}) + (1-\pi_i^*)L(\boldsymbol{X}_i)}{\pi_i^*(1-\pi_i^*)}\mathbf{f}(\boldsymbol{X}_i)\right),\\
&\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = (\boldsymbol{H}_{\mathbf{f}}^*)^{-1}\operatorname{Var}(\boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i, \boldsymbol{X}_i))(\boldsymbol{H}_{\mathbf{f}}^{*\top})^{-1},\\
&\operatorname{Var}(\boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i, \boldsymbol{X}_i)) = \mathbb{E}\left(\frac{\mathbf{f}(\boldsymbol{X}_i)\mathbf{f}(\boldsymbol{X}_i)^\top}{\pi_i^*(1-\pi_i^*)}\right).
\end{aligned}$$

If $K(\boldsymbol{X}_i)+(1-\pi_i^*)L(\boldsymbol{X}_i)$ lies in the linear space spanned by $\mathbf{f}(\boldsymbol{X}_i)$, that is, $K(\boldsymbol{X}_i)+(1-\pi_i^*)L(\boldsymbol{X}_i) = \boldsymbol{\alpha}^\top\mathbf{f}(\boldsymbol{X}_i)$, we have

$$\boldsymbol{H}_y^* = -\mathbb{E}\left(\frac{\boldsymbol{\alpha}^\top\mathbf{f}(\boldsymbol{X}_i)}{\pi_i^*(1-\pi_i^*)}\frac{\partial \pi_i^*}{\partial \boldsymbol{\beta}}\right) = (\boldsymbol{\alpha}^\top\boldsymbol{H}_{\mathbf{f}}^*)^\top.$$

So

$$\boldsymbol{H}_y^{*\top}\boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} = -\boldsymbol{\alpha}^\top\boldsymbol{H}_{\mathbf{f}}^*(\boldsymbol{H}_{\mathbf{f}}^*)^{-1}\mathbb{E}\left(\frac{\boldsymbol{\alpha}^\top\mathbf{f}(\boldsymbol{X}_i)\mathbf{f}(\boldsymbol{X}_i)}{\pi_i^*(1-\pi_i^*)}\right) = -\boldsymbol{\alpha}^\top\mathbb{E}\left(\frac{\mathbf{f}(\boldsymbol{X}_i)\mathbf{f}(\boldsymbol{X}_i)^\top}{\pi_i^*(1-\pi_i^*)}\right)\boldsymbol{\alpha},$$

and

$$\boldsymbol{H}_y^{*\top}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{H}_y^* = \boldsymbol{\alpha}^\top\boldsymbol{H}_{\mathbf{f}}^*(\boldsymbol{H}_{\mathbf{f}}^*)^{-1}\mathbb{E}\left(\frac{\mathbf{f}(\boldsymbol{X}_i)\mathbf{f}(\boldsymbol{X}_i)^\top}{\pi_i^*(1-\pi_i^*)}\right)(\boldsymbol{H}_{\mathbf{f}}^{*\top})^{-1}(\boldsymbol{\alpha}^\top\boldsymbol{H}_{\mathbf{f}}^*)^\top = \boldsymbol{\alpha}^\top\mathbb{E}\left(\frac{\mathbf{f}(\boldsymbol{X}_i)\mathbf{f}(\boldsymbol{X}_i)^\top}{\pi_i^*(1-\pi_i^*)}\right)\boldsymbol{\alpha}.$$

It is seen that $\boldsymbol{H}_y^{*\top}\boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} = -\boldsymbol{H}_y^{*\top}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{H}_y^*$. Then we have

$$V = \Sigma_\mu - \boldsymbol{\alpha}^\top\boldsymbol{\Omega}\boldsymbol{\alpha},$$

which corresponds to the minimum asymptotic variance V_{opt}. □

D Proof of Results in Section 3

D.1 Proof of Theorem 3.1

Proof of Theorem 3.1. We first consider the case (1). That is the propensity score model is correctly specified. By Lemma B.2, we have $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^o$. Let

$$r_{\boldsymbol{\beta}}(T, Y, \boldsymbol{X}) = \frac{TY}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X})} - \frac{(1-T)Y}{1-\pi_{\boldsymbol{\beta}}(\boldsymbol{X})}.$$

It is seen that $|r_{\boldsymbol{\beta}}(T, Y, \boldsymbol{X})| \leq 2|Y|/c_0$ and by Assumption 3.1 (6), $\mathbb{E}|Y| < \infty$. Then Lemma B.1 yields $\sup_{\boldsymbol{\beta}\in\Theta} |n^{-1}\sum_{i=1}^n r_{\boldsymbol{\beta}}(T_i, Y_i, \boldsymbol{X}_i) - \mathbb{E}(r_{\boldsymbol{\beta}}(T_i, Y_i, \boldsymbol{X}_i))| = o_p(1)$. In addition, by $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^o$ and the dominated convergence theorem, we obtain that

$$\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} \;=\; \mathbb{E}\Big(\frac{T_iY_i}{\pi_i^o} - \frac{(1-T_i)Y_i}{1-\pi_i^o}\Big) + o_p(1),$$

where $\pi_i^o = \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)$. Since $Y_i = Y_i(1)T_i + Y_i(0)(1-T_i)$ and $Y_i(1), Y_i(0)$ are independent of T_i given $\boldsymbol{X}_i$, we can further simplify the above expression,

$$\begin{aligned}\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} &= \mathbb{E}\Big(\frac{T_iY_i}{\pi_i^o} - \frac{(1-T_i)Y_i}{1-\pi_i^o}\Big) + o_p(1) = \mathbb{E}\Big(\frac{T_iY_i(1)}{\pi_i^o} - \frac{(1-T_i)Y_i(0)}{1-\pi_i^o}\Big) + o_p(1)\\ &= \mathbb{E}\Big(\frac{\mathbb{E}(T_i \mid \boldsymbol{X}_i)\mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i)}{\pi_i^o} - \frac{(1-\mathbb{E}(T_i \mid \boldsymbol{X}_i))\mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i)}{1-\pi_i^o}\Big) + o_p(1).\end{aligned}$$

In addition, if the propensity score model is correctly specified, it further implies

$$\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} \;=\; \mathbb{E}(\mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i) - \mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i)) + o_p(1) = \mathbb{E}(Y_i(1) - Y_i(0)) + o_p(1) = \mu + o_p(1).$$

This completes the proof of consistence of $\widehat{\mu}$ when the propensity score model is correctly specified.

In the following, we consider the case (2). That is $K(\cdot) \in \text{span}\{\mathbf{M}_1\boldsymbol{h}_1(\cdot)\}$ and $L(\cdot) \in \text{span}\{\mathbf{M}_2\boldsymbol{h}_2(\cdot)\}$. By Lemma B.2, we have $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^o$. The first order condition for $\boldsymbol{\beta}^o$ yields $\partial Q(\boldsymbol{\beta}^o)/\partial\boldsymbol{\beta} = 0$, where $Q(\boldsymbol{\beta}) = \mathbb{E}(\boldsymbol{g}_{\boldsymbol{\beta}}^\top)\mathbf{W}^*\mathbb{E}(\boldsymbol{g}_{\boldsymbol{\beta}})$. By Assumption 3.1 (7) and the dominated convergence theorem, we can interchange the differential with integral, and thus $\mathbf{G}^{*\top}\mathbf{W}^*\mathbb{E}(\boldsymbol{g}_{\boldsymbol{\beta}^o}) = 0$. Under the assumption that $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi(\boldsymbol{X}_i) \neq \pi_i^o$, we have

$$\mathbb{E}(\boldsymbol{g}_{1\boldsymbol{\beta}^o}) = \mathbb{E}\Big\{\Big(\frac{\pi(\boldsymbol{X}_i)}{\pi_i^o} - \frac{1-\pi(\boldsymbol{X}_i)}{1-\pi_i^o}\Big)\boldsymbol{h}_1(\boldsymbol{X}_i)\Big\},$$

$$\mathbb{E}(\boldsymbol{g}_{2\boldsymbol{\beta}^o}) = \mathbb{E}\Big\{\Big(\frac{\pi(\boldsymbol{X}_i)}{\pi_i^o} - 1\Big)\boldsymbol{h}_2(\boldsymbol{X}_i)\Big\}.$$

Rewrite $\mathbf{G}^{*\top}\mathbf{W}^* = (\mathbf{M}_1, \mathbf{M}_2)$, where $\mathbf{M}_1 \in \mathbb{R}^{q\times m_1}$ and $\mathbf{M}_1 \in \mathbb{R}^{q\times m_2}$. Then, $\boldsymbol{\beta}^o$ satisfies

$$\mathbb{E}\Big\{\Big(\frac{\pi(\boldsymbol{X}_i)}{\pi_i^o} - \frac{1-\pi(\boldsymbol{X}_i)}{1-\pi_i^o}\Big)\mathbf{M}_1\boldsymbol{h}_1(\boldsymbol{X}_i) + \Big(\frac{\pi(\boldsymbol{X}_i)}{\pi_i^o} - 1\Big)\mathbf{M}_2\boldsymbol{h}_2(\boldsymbol{X}_i)\Big\} = 0. \qquad \text{(D.1)}$$

Following the similar arguments to that in case (1), we can prove that

$$\begin{aligned}\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} &= \mathbb{E}\Big(\frac{T_iY_i}{\pi_i^o} - \frac{(1-T_i)Y_i}{1-\pi_i^o}\Big) + o_p(1)\\ &= \mathbb{E}\Big(\frac{\mathbb{E}(T_i \mid \boldsymbol{X}_i)\mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i)}{\pi_i^o} - \frac{(1-\mathbb{E}(T_i \mid \boldsymbol{X}_i))\mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i)}{1-\pi_i^o}\Big) + o_p(1).\end{aligned}$$

By $\mathbb{E}(T_i \mid \boldsymbol{X}_i) = \pi(\boldsymbol{X}_i)$ and the outcome model, it further implies

$$\begin{aligned}
\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu &= \mathbb{E}\Big\{\frac{\pi(\boldsymbol{X}_i)(K(\boldsymbol{X}_i)+L(\boldsymbol{X}_i))}{\pi_i^o} - \frac{(1-\pi(\boldsymbol{X}_i))K(\boldsymbol{X}_i)}{1-\pi_i^o}\Big\} - \mu + o_p(1) \\
&= \mathbb{E}\Big\{\Big(\frac{\pi(\boldsymbol{X}_i)}{\pi_i^o} - \frac{1-\pi(\boldsymbol{X}_i)}{1-\pi_i^o}\Big)K(\boldsymbol{X}_i)\Big\} + \mathbb{E}\Big\{\frac{\pi(\boldsymbol{X}_i)L(\boldsymbol{X}_i)}{\pi_i^o}\Big\} - \mu + o_p(1) \\
&= \mathbb{E}\Big\{\Big(\frac{\pi(\boldsymbol{X}_i)}{\pi_i^o} - \frac{1-\pi(\boldsymbol{X}_i)}{1-\pi_i^o}\Big)K(\boldsymbol{X}_i)\Big\} + \mathbb{E}\Big\{\Big(\frac{\pi(\boldsymbol{X}_i)}{\pi_i^o} - 1\Big)L(\boldsymbol{X}_i)\Big\} + o_p(1),
\end{aligned}$$

where in the last step we use $\mu = \mathbb{E}(L(\boldsymbol{X}_i))$. By equation (D.1), we obtain $\widehat{\mu} = \mu + o_p(1)$, provided $K(\boldsymbol{X}_i) = \boldsymbol{\alpha}_1^\top \mathbf{M}_1 \boldsymbol{h}_1(\boldsymbol{X}_i)$ and $L(\boldsymbol{X}_i) = \boldsymbol{\alpha}_2^\top \mathbf{M}_2 \boldsymbol{h}_2(\boldsymbol{X}_i)$, where $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are q-dimensional vectors of constants. This completes the whole proof.

□

D.2 Proof of Theorem 3.2

Proof of Theorem 3.2. We first consider the case (1). That is the propensity score model is correctly specified. By the mean value theorem, we have $\widehat{\mu} = \bar{\mu} + \widehat{\mathbf{H}}(\widetilde{\boldsymbol{\beta}})^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)$, where

$$\bar{\mu} = \frac{1}{n}\sum_{i=1}^n \Big(\frac{T_iY_i}{\pi_i^o} - \frac{(1-T_i)Y_i}{1-\pi_i^o}\Big), \quad \widehat{\mathbf{H}}(\widetilde{\boldsymbol{\beta}}) = -\frac{1}{n}\sum_{i=1}^n \Big(\frac{T_iY_i}{\widetilde{\pi}_i^2} + \frac{(1-T_i)Y_i}{(1-\widetilde{\pi}_i)^2}\Big)\frac{\partial \widetilde{\pi}_i}{\partial \boldsymbol{\beta}},$$

where $\pi_i^o = \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)$, $\widetilde{\pi}_i = \pi_{\widetilde{\boldsymbol{\beta}}}(\boldsymbol{X}_i)$ and $\widetilde{\boldsymbol{\beta}}$ is an intermediate value between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^o$. By Assumption 3.2 (2), we can show that the summand in $\widehat{\mathbf{H}}(\widetilde{\boldsymbol{\beta}})$ has a bounded envelop function. By Lemma B.1, we have $\sup_{\boldsymbol{\beta}\in\mathbb{B}_r(\boldsymbol{\beta}^o)} |\widehat{\mathbf{H}}(\boldsymbol{\beta}) - \mathbb{E}(\widehat{\mathbf{H}}(\boldsymbol{\beta}))| = o_p(1)$. Since $\widehat{\boldsymbol{\beta}}$ is consistent, by the dominated convergence theorem we can obtain $\widehat{\mathbf{H}}(\widetilde{\boldsymbol{\beta}}) = \mathbf{H}^* + o_p(1)$, where

$$\begin{aligned}
\mathbf{H}^* &= -\mathbb{E}\Big\{\Big(\frac{T_iY_i}{\pi_i^{o2}} + \frac{(1-T_i)Y_i}{(1-\pi_i^o)^2}\Big)\frac{\partial \pi_i^o}{\partial \boldsymbol{\beta}}\Big\} = -\mathbb{E}\Big\{\Big(\frac{Y_i(1)}{\pi_i^o} + \frac{Y_i(0)}{1-\pi_i^o}\Big)\frac{\partial \pi_i^o}{\partial \boldsymbol{\beta}}\Big\} \\
&= -\mathbb{E}\Big\{\frac{K(\boldsymbol{X}_i) + L(\boldsymbol{X}_i)(1-\pi_i^o)}{\pi_i^o(1-\pi_i^o)}\frac{\partial \pi_i^o}{\partial \boldsymbol{\beta}}\Big\}.
\end{aligned}$$

Finally, we invoke the central limit theorem and equation (B.1) to obtain that

$$n^{1/2}(\widehat{\mu} - \mu) \xrightarrow{d} N(0, \bar{\mathbf{H}}^{*\top}\boldsymbol{\Sigma}\bar{\mathbf{H}}^*),$$

where $\bar{\mathbf{H}}^* = (1, \mathbf{H}^{*\top})^\top$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = (\mathbf{G}^{*\top}\mathbf{W}^*\mathbf{G}^*)^{-1}\mathbf{G}^{*\top}\mathbf{W}^*\boldsymbol{\Omega}\mathbf{W}^*\mathbf{G}^*(\mathbf{G}^{*\top}\mathbf{W}^*\mathbf{G}^*)^{-1}$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_\mu & \boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}}^\top \\ \boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} & \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \end{pmatrix}.$$

Denote $b_i(T_i, \boldsymbol{X}_i, Y_i(1), Y_i(0)) = T_iY_i(1)/\pi_i^o - (1-T_i)Y_i(0)/(1-\pi_i^o) - \mu$. Here, some simple calculations yield,

$$\Sigma_\mu = \mathbb{E}[b_i^2(T_i, \boldsymbol{X}_i, Y_i(1), Y_i(0))] = \mathbb{E}\Big(\frac{Y_i^2(1)}{\pi_i^o} + \frac{Y_i^2(0)}{1-\pi_i^o}\Big) - \mu^2.$$

In addition, the off diagonal matrix can be written as $\mathbf{\Sigma}_{\mu\boldsymbol{\beta}} = (\mathbf{\Sigma}_{1\mu\boldsymbol{\beta}}^\top, \mathbf{\Sigma}_{2\mu\boldsymbol{\beta}}^\top)^\top$, where

$$\mathbf{\Sigma}_{\mu\boldsymbol{\beta}} = -(\mathbf{G}^{*\top}\mathbf{W}^*\mathbf{G}^*)^{-1}\mathbf{G}^{*\top}\mathbf{W}^*\mathbf{T},$$

where $\mathbf{T} = (\mathbb{E}[\boldsymbol{g}_{1\boldsymbol{\beta}^o}^\top(T_i, \boldsymbol{X}_i)b_i(T_i, \boldsymbol{X}_i, Y_i(1), Y_i(0))], \mathbb{E}[\boldsymbol{g}_{2\boldsymbol{\beta}^o}^\top(T_i, \boldsymbol{X}_i)b_i(T_i, \boldsymbol{X}_i, Y_i(1), Y_i(0))])^\top$ with

$$\boldsymbol{g}_{1\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i) = \Big(\frac{T_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - \frac{1-T_i}{1-\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}\Big)\boldsymbol{h}_1(\boldsymbol{X}_i), \quad \text{and} \quad \boldsymbol{g}_{2\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i) = \Big(\frac{T_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - 1\Big)\boldsymbol{h}_2(\boldsymbol{X}_i).$$

After some algebra, we can show that

$$\mathbf{T} = \Big\{\mathbb{E}\Big(\frac{K(\boldsymbol{X}_i) + (1-\pi_i^o)L(\boldsymbol{X}_i)}{(1-\pi_i^o)\pi_i^o}\boldsymbol{h}_1^\top(\boldsymbol{X}_i)\Big), \mathbb{E}\Big(\frac{K(\boldsymbol{X}_i) + (1-\pi_i^o)L(\boldsymbol{X}_i)}{\pi_i^o}\boldsymbol{h}_2^\top(\boldsymbol{X}_i)\Big)\Big\}^\top.$$

This completes the proof of equation (3.4). Next, we consider the case (2). Recall that $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi(\boldsymbol{X}_i) \neq \pi_{\beta^o}(\boldsymbol{X}_i)$. Following the similar arguments, we can show that

$$\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu = \frac{1}{n}\sum_{i=1}^n D_i + \mathbf{H}^{*\top}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o) + o_p(n^{-1/2}),$$

where

$$D_i = \frac{T_iY_i(1)}{\pi_i^o} - \frac{(1-T_i)Y_i(0)}{1-\pi_i^o} - \mu,$$

and

$$\mathbf{H}^* = -\mathbb{E}\Big\{\Big(\frac{\pi(\boldsymbol{X}_i)(K(\boldsymbol{X}_i) + L(\boldsymbol{X}_i))}{\pi_i^{o2}} + \frac{(1-\pi(\boldsymbol{X}_i))K(\boldsymbol{X}_i)}{(1-\pi_i^o)^2}\Big)\frac{\partial \pi_i^o}{\partial \boldsymbol{\beta}}\Big\}.$$

By equation (B.1) in Lemma B.3, we have that

$$n^{1/2}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu) \xrightarrow{d} N(0, \widetilde{\mathbf{H}}^{*\top}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{H}}^*),$$

where $\widetilde{\mathbf{H}}^* = (1, \mathbf{H}^{*\top})^\top$, $\mathbf{\Sigma}_{\boldsymbol{\beta}} = (\mathbf{G}^{*\top}\mathbf{W}^*\mathbf{G}^*)^{-1}\mathbf{G}^{*\top}\mathbf{W}^*\mathbf{\Omega}\mathbf{W}^*\mathbf{G}^*(\mathbf{G}^{*\top}\mathbf{W}^*\mathbf{G}^*)^{-1}$ and

$$\widetilde{\mathbf{\Sigma}} = \begin{pmatrix} \Sigma_\mu & \widetilde{\mathbf{\Sigma}}_{\mu\boldsymbol{\beta}}^\top \\ \widetilde{\mathbf{\Sigma}}_{\mu\boldsymbol{\beta}} & \widetilde{\mathbf{\Sigma}}_{\boldsymbol{\beta}} \end{pmatrix}.$$

Denote $c_i(T_i, \boldsymbol{X}_i, Y_i(1), Y_i(0)) = T_iY_i(1)/\pi_i^o - (1-T_i)Y_i(0)/(1-\pi_i^o) - \mu$. As shown in the proof of Theorem 3.1, $\mathbb{E}[b_i(T_i, \boldsymbol{X}_i, Y_i(1), Y_i(0))] = 0$. Thus,

$$\begin{aligned}\Sigma_\mu &= \mathbb{E}[c_i^2(T_i, \boldsymbol{X}_i, Y_i(1), Y_i(0))] = \mathbb{E}\Big(\frac{T_iY_i^2(1)}{\pi_i^{o2}} + \frac{(1-T_i)Y_i^2(0)}{(1-\pi_i^o)^2}\Big) - \mu^2 \\ &= \mathbb{E}\Big(\frac{\pi(\boldsymbol{X}_i)Y_i^2(1)}{\pi_i^{o2}} + \frac{(1-\pi(\boldsymbol{X}_i))Y_i^2(0)}{(1-\pi_i^o)^2}\Big) - \mu^2.\end{aligned}$$

Similarly, the off diagonal matrix can be written as $\widetilde{\mathbf{\Sigma}}_{\mu\boldsymbol{\beta}} = (\widetilde{\mathbf{\Sigma}}_{1\mu\boldsymbol{\beta}}^\top, \widetilde{\mathbf{\Sigma}}_{2\mu\boldsymbol{\beta}}^\top)^\top$, where

$$\widetilde{\mathbf{\Sigma}}_{\mu\boldsymbol{\beta}} = -(\mathbf{G}^{*\top}\mathbf{W}^*\mathbf{G}^*)^{-1}\mathbf{G}^{*\top}\mathbf{W}^*\boldsymbol{S},$$

where $\boldsymbol{S} = (\mathbb{E}[\boldsymbol{g}_{1\boldsymbol{\beta}^o}^\top(T_i, \boldsymbol{X}_i)c_i(T_i, \boldsymbol{X}_i, Y_i(1), Y_i(0))], \mathbb{E}[\boldsymbol{g}_{2\boldsymbol{\beta}^o}^\top(T_i, \boldsymbol{X}_i)c_i(T_i, \boldsymbol{X}_i, Y_i(1), Y_i(0))])^\top$ with

$$\boldsymbol{g}_{1\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i) = \Big(\frac{T_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - \frac{1-T_i}{1-\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}\Big)\boldsymbol{h}_1(\boldsymbol{X}_i), \quad \text{and} \quad \boldsymbol{g}_{2\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i) = \Big(\frac{T_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - 1\Big)\boldsymbol{h}_2(\boldsymbol{X}_i). \quad \text{(D.2)}$$

After some tedious algebra, we can show that $\boldsymbol{S} = (\boldsymbol{S}_1^\top, \boldsymbol{S}_2^\top)^\top$, where

$$\boldsymbol{S}_1 = \mathbb{E}\Big\{\Big(\frac{\pi(\boldsymbol{X}_i)(K(\boldsymbol{X}_i) + L(\boldsymbol{X}_i) - \pi_i^o\mu)}{\pi_i^{o2}} + \frac{(1-\pi(\boldsymbol{X}_i))(K(\boldsymbol{X}_i) + (1-\pi_i^o)\mu)}{(1-\pi_i^o)^2}\Big)\boldsymbol{h}_1(\boldsymbol{X}_i)\Big\},$$

$$\boldsymbol{S}_2 = \mathbb{E}\Big\{\Big(\frac{\pi(\boldsymbol{X}_i)[(K(\boldsymbol{X}_i) + L(\boldsymbol{X}_i))(1-\pi_i^o) - \pi_i^o\mu]}{\pi_i^{o2}} + \frac{(1-\pi(\boldsymbol{X}_i))K(\boldsymbol{X}_i) + (1-\pi_i^o)\mu}{1-\pi_i^o}\Big)\boldsymbol{h}_2(\boldsymbol{X}_i)\Big\}.$$

This completes the proof of equation (3.6).

Finally, we start to prove part 3. By (3.4), the asymptotic variance of $\widehat{\mu}$ denoted by V, can be written as

$$V = \Sigma_\mu + 2\mathbf{H}^{*\top}\mathbf{\Sigma}_{\mu\boldsymbol{\beta}} + \mathbf{H}^{*\top}\mathbf{\Sigma}_{\boldsymbol{\beta}}\mathbf{H}^*. \quad \text{(D.3)}$$

Note that by Lemma B.3, we have $\mathbf{\Sigma}_{\boldsymbol{\beta}} = (\mathbf{G}^{*\top}\mathbf{\Omega}^{-1}\mathbf{G}^*)^{-1}$. Under this correctly specified propensity score model, some algebra yields

$$\mathbf{\Omega} = \mathbb{E}[\boldsymbol{g}_{\boldsymbol{\beta}^o}(T_i, \boldsymbol{X}_i)\boldsymbol{g}_{\boldsymbol{\beta}^o}^\top(T_i, \boldsymbol{X}_i)] = \begin{pmatrix} \mathbb{E}(\frac{\boldsymbol{h}_1\boldsymbol{h}_1^\top}{\pi_i^o(1-\pi_i^o)}) & \mathbb{E}(\frac{\boldsymbol{h}_1\boldsymbol{h}_2^\top}{\pi_i^o}) \\ \mathbb{E}(\frac{\boldsymbol{h}_2\boldsymbol{h}_1^\top}{\pi_i^o}) & \mathbb{E}(\frac{\boldsymbol{h}_2\boldsymbol{h}_2^\top(1-\pi_i^o)}{\pi_i^o}) \end{pmatrix},$$

where $\boldsymbol{g}_{\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i) = (\boldsymbol{g}_{1\boldsymbol{\beta}}^\top(T_i, \boldsymbol{X}_i), \boldsymbol{g}_{2\boldsymbol{\beta}}^\top(T_i, \boldsymbol{X}_i))^\top$ and $\boldsymbol{g}_{1\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)$ and $\boldsymbol{g}_{2\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)$ are defined in (D.2). In addition, $\mathbf{G}^* = (\mathbf{G}_1^{*\top}, \mathbf{G}_2^{*\top})^\top$, where

$$\mathbf{G}_1^* = -\mathbb{E}\Big(\frac{\boldsymbol{h}_1(\boldsymbol{X}_i)}{\pi_i^o(1-\pi_i^o)}\Big(\frac{\partial\pi_i^o}{\partial\boldsymbol{\beta}}\Big)^\top\Big), \quad \mathbf{G}_2^* = -\mathbb{E}\Big(\frac{\boldsymbol{h}_2(\boldsymbol{X}_i)}{\pi_i^o}\Big(\frac{\partial\pi_i^o}{\partial\boldsymbol{\beta}}\Big)^\top\Big). \quad \text{(D.4)}$$

Since the functions $\boldsymbol{K}(\cdot)$ and $\boldsymbol{L}(\cdot)$ lie in the linear space spanned by the functions $\mathbf{M}_1\boldsymbol{h}_1(\cdot)$ and $\mathbf{M}_2\boldsymbol{h}_2(\cdot)$ respectively, where $\mathbf{M}_1 \in \mathbb{R}^{q\times m_1}$ and $\mathbf{M}_1 \in \mathbb{R}^{q\times m_2}$ are the partitions of $\mathbf{G}^{*\top}\mathbf{W}^* = (\mathbf{M}_1, \mathbf{M}_2)$. We have $K(\boldsymbol{X}_i) = \boldsymbol{\alpha}_1^\top\mathbf{M}_1\boldsymbol{h}_1(\boldsymbol{X}_i)$ and $L(\boldsymbol{X}_i) = \boldsymbol{\alpha}_2^\top\mathbf{M}_2\boldsymbol{h}_2(\boldsymbol{X}_i)$, where $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are q-dimensional vectors of constants. Thus

$$\begin{aligned}\mathbf{H}^* &= -\mathbb{E}\Big\{\frac{K(\boldsymbol{X}_i) + L(\boldsymbol{X}_i)(1-\pi_i^o)}{\pi_i^o(1-\pi_i^o)}\frac{\partial\pi_i^o}{\partial\boldsymbol{\beta}}\Big\} \\ &= -\mathbb{E}\Big\{\frac{\boldsymbol{\alpha}_1^\top\mathbf{M}_1\boldsymbol{h}_1(\boldsymbol{X}_i) + \boldsymbol{\alpha}_2^\top\mathbf{M}_2\boldsymbol{h}_2(\boldsymbol{X}_i)(1-\pi_i^o)}{\pi_i^o(1-\pi_i^o)}\frac{\partial\pi_i^o}{\partial\boldsymbol{\beta}}\Big\}.\end{aligned}$$

Comparing to the expression of $\mathbf{G}^*$ in (D.4), we can rewrite $\mathbf{H}^*$ as

$$\mathbf{H}^* = \mathbf{G}^{*\top} \begin{pmatrix} \mathbf{M}_1^\top \boldsymbol{\alpha}_1 \\ \mathbf{M}_2^\top \boldsymbol{\alpha}_2 \end{pmatrix}.$$

Following the similar derivations, it is seen that

$$\boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} = -(\mathbf{G}^{*\top}\boldsymbol{\Omega}^{-1}\mathbf{G}^*)^{-1}\mathbf{G}^{*\top}\boldsymbol{\Omega}^{-1} \begin{pmatrix} \mathbb{E}\{\frac{\boldsymbol{\alpha}_1^\top \mathbf{M}_1 \boldsymbol{h}_1(\boldsymbol{X}_i) + \boldsymbol{\alpha}_2^\top \mathbf{M}_2 \boldsymbol{h}_2(\boldsymbol{X}_i)(1-\pi_i^o)}{\pi_i^o(1-\pi_i^o)} \boldsymbol{h}_1(\boldsymbol{X}_i)\} \\ \mathbb{E}\{\frac{\boldsymbol{\alpha}_1^\top \mathbf{M}_1 \boldsymbol{h}_1(\boldsymbol{X}_i) + \boldsymbol{\alpha}_2^\top \mathbf{M}_2 \boldsymbol{h}_2(\boldsymbol{X}_i)(1-\pi_i^o)}{\pi_i^o} \boldsymbol{h}_2(\boldsymbol{X}_i)\} \end{pmatrix},$$

which is equivalent to

$$\boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} = -(\mathbf{G}^{*\top}\boldsymbol{\Omega}^{-1}\mathbf{G}^*)^{-1}\mathbf{G}^{*\top} \begin{pmatrix} \mathbf{M}_1^\top \boldsymbol{\alpha}_1 \\ \mathbf{M}_2^\top \boldsymbol{\alpha}_2 \end{pmatrix}.$$

Hence,

$$\mathbf{H}^{*\top}\boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} = -(\boldsymbol{\alpha}_1^\top \mathbf{M}_1, \boldsymbol{\alpha}_2^\top \mathbf{M}_2)\mathbf{G}^*(\mathbf{G}^{*\top}\boldsymbol{\Omega}^{-1}\mathbf{G}^*)^{-1}\mathbf{G}^{*\top} \begin{pmatrix} \mathbf{M}_1^\top \boldsymbol{\alpha}_1 \\ \mathbf{M}_2^\top \boldsymbol{\alpha}_2 \end{pmatrix} = -\mathbf{H}^{*\top}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\mathbf{H}^*.$$

Together with (D.3), we have

$$V = \Sigma_\mu - (\boldsymbol{\alpha}_1^\top \mathbf{M}_1, \boldsymbol{\alpha}_2^\top \mathbf{M}_2)\mathbf{G}^*(\mathbf{G}^{*\top}\boldsymbol{\Omega}^{-1}\mathbf{G}^*)^{-1}\mathbf{G}^{*\top} \begin{pmatrix} \mathbf{M}_1^\top \boldsymbol{\alpha}_1 \\ \mathbf{M}_2^\top \boldsymbol{\alpha}_2 \end{pmatrix}.$$

This completes of the proof.

□

D.3 Proof of Corollary 3.1

Proof of Corollary 3.1. By Theorem 3.2, it suffices to show that

$$(\bar{\boldsymbol{\alpha}}_1^\top \bar{\mathbf{M}}_1, \bar{\boldsymbol{\alpha}}_2^\top \bar{\mathbf{M}}_2)\bar{\mathbf{G}}^*\bar{\mathbf{C}}\bar{\mathbf{G}}^{*\top} \begin{pmatrix} \bar{\mathbf{M}}_1^\top \bar{\boldsymbol{\alpha}}_1 \\ \bar{\mathbf{M}}_2^\top \bar{\boldsymbol{\alpha}}_2 \end{pmatrix} \leq (\boldsymbol{\alpha}_1^\top \mathbf{M}_1, \boldsymbol{\alpha}_2^\top \mathbf{M}_2)\mathbf{G}^*\mathbf{C}\mathbf{G}^{*\top} \begin{pmatrix} \mathbf{M}_1^\top \boldsymbol{\alpha}_1 \\ \mathbf{M}_2^\top \boldsymbol{\alpha}_2 \end{pmatrix}, \tag{D.5}$$

where $\mathbf{C} = (\mathbf{G}^{*\top}\boldsymbol{\Omega}^{-1}\mathbf{G}^*)^{-1}$ and $\bar{\boldsymbol{\alpha}}_1$ and $\bar{\mathbf{M}}_1$ among others are the corresponding quantities with $\bar{\boldsymbol{h}}_1(\boldsymbol{X})$ and $\bar{\boldsymbol{h}}_2(\boldsymbol{X})$. Assume that $\bar{\boldsymbol{h}}_1(\boldsymbol{X}) \in \mathbb{R}^{m_1+a_1}$ and $\bar{\boldsymbol{h}}_2(\boldsymbol{X}) \in \mathbb{R}^{m_2+a_2}$. Since $K(\boldsymbol{X}_i) = \boldsymbol{\alpha}_1^\top \mathbf{M}_1 \boldsymbol{h}_1(\boldsymbol{X}_i)$ and $L(\boldsymbol{X}_i) = \boldsymbol{\alpha}_2^\top \mathbf{M}_2 \boldsymbol{h}_2(\boldsymbol{X}_i)$, we find that $(\bar{\boldsymbol{\alpha}}_1^\top \bar{\mathbf{M}}_1, \bar{\boldsymbol{\alpha}}_2^\top \bar{\mathbf{M}}_2) = (\boldsymbol{\alpha}_1^\top \mathbf{M}_1, 0, \boldsymbol{\alpha}_2^\top \mathbf{M}_2, 0)$, which is a vector in $\mathbb{R}^{m+a}$ with $a = a_1 + a_2$. Because some components of $(\bar{\boldsymbol{\alpha}}_1^\top \bar{\mathbf{M}}_1, \bar{\boldsymbol{\alpha}}_2^\top \bar{\mathbf{M}}_2)$ are 0, by the matrix algebra, (D.5) holds if $\mathbf{C} - \bar{\mathbf{C}}$ is positive semidefinite. Without loss of generality, we rearrange orders and write the $(m+a) \times q$ matrix $\bar{\mathbf{G}}^*$ and the $(m+a) \times (m+a)$ matrix $\bar{\boldsymbol{\Omega}}^*$ as

$$\bar{\mathbf{G}}^* = \begin{pmatrix} \mathbf{G}^* \\ \mathbf{A} \end{pmatrix}, \quad \text{and} \quad \bar{\boldsymbol{\Omega}} = \begin{pmatrix} \boldsymbol{\Omega} & \boldsymbol{\Omega}_1 \\ \boldsymbol{\Omega}_1 & \boldsymbol{\Omega}_2 \end{pmatrix}.$$

For simplicity, we use the following notation: two matrices satisfy $\mathbf{O}_1 \geq \mathbf{O}_2$ if $\mathbf{O}_1 - \mathbf{O}_2$ is positive semidefinite. To show $\mathbf{C} \geq \bar{\mathbf{C}}$, we have the following derivation

$$\begin{aligned} \bar{\mathbf{G}}^{*\top} \bar{\boldsymbol{\Omega}}^{-1} \bar{\mathbf{G}}^* &= (\mathbf{G}^{*\top}, \mathbf{A}^\top) \begin{pmatrix} \boldsymbol{\Omega} & \boldsymbol{\Omega}_1 \\ \boldsymbol{\Omega}_1 & \boldsymbol{\Omega}_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{G}^* \\ \mathbf{A} \end{pmatrix} \\ &\geq (\mathbf{G}^{*\top}, \mathbf{A}^\top) \begin{pmatrix} \boldsymbol{\Omega}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{G}^* \\ \mathbf{A} \end{pmatrix} = \mathbf{G}^{*\top} \boldsymbol{\Omega}^{-1} \mathbf{G}^*. \end{aligned}$$

This completes the proof of (D.5), and therefore the corollary holds.

□

D.4 Proof of Corollary 3.2

Proof of Corollary 3.2. The proof of the double robustness property mainly follows from Theorem 3.1. In this case, we only need to verify that $\text{span}\{\boldsymbol{h}_1(\cdot)\} = \text{span}\{\mathbf{M}_1 \boldsymbol{h}_1(\cdot)\}$ and $\text{span}\{\boldsymbol{h}_2(\cdot)\} = \text{span}\{\mathbf{M}_2 \boldsymbol{h}_2(\cdot)\}$, where $\mathbf{M}_1 \in \mathbb{R}^{q \times m_1}$ and $\mathbf{M}_1 \in \mathbb{R}^{q \times m_2}$ are the partitions of $\mathbf{G}^{*\top} \mathbf{W}^* = (\mathbf{M}_1, \mathbf{M}_2)$. Apparently, we have $\text{span}\{\mathbf{M}_1 \boldsymbol{h}_1(\cdot)\} \subseteq \text{span}\{\boldsymbol{h}_1(\cdot)\}$, since the former can always be written as a linear combination of $\boldsymbol{h}_1(\cdot)$. To show $\text{span}\{\boldsymbol{h}_1(\cdot)\} \subseteq \text{span}\{\mathbf{M}_1 \boldsymbol{h}_1(\cdot)\}$, note that the $m_1 \times m_1$ principal submatrix $\mathbf{M}_{11}$ of $\mathbf{M}_1$ is invertible. Thus, $\text{span}\{\boldsymbol{h}_1(\cdot)\} = \text{span}\{\mathbf{M}_{11} \boldsymbol{h}_1(\cdot)\} \subseteq \text{span}\{\mathbf{M}_1 \boldsymbol{h}_1(\cdot)\}$. This is because the m_1 dimensional functions $\mathbf{M}_{11} \boldsymbol{h}_1(\cdot)$ are identical to the first m_1 coordinates of $\mathbf{M}_1 \boldsymbol{h}_1(\cdot)$. This completes the proof of double robustness property. The efficiency property follows from Theorem 3.2. We do not replicate the details. □

E Regularity Conditions in Section 4

Assumption E.1. The following regularity conditions are assumed.

1. The minimizer $\boldsymbol{\beta}^o = \text{argmin}_{\boldsymbol{\beta} \in \Theta} \|\mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}))\|_2^2$ is unique.

2. $\boldsymbol{\beta}^o \in \text{int}(\Theta)$, where Θ is a compact set.

3. There exist constants $0 < c_0 < 1/2$, $c_1 > 0$ and $c_2 > 0$ such that $c_0 \leq J(v) \leq 1 - c_0$ and $0 < c_1 \leq \partial J(v)/\partial v \leq c_2$, for any $v = \boldsymbol{\beta}^\top \boldsymbol{B}(\boldsymbol{x})$ with $\boldsymbol{\beta} \in \text{int}(\Theta)$. There exists a small neighborhood of $v^* = \boldsymbol{\beta}^{*\top} \boldsymbol{B}(\boldsymbol{x})$, say $\mathcal{B}$ such that for any $v \in \mathcal{B}$ it holds that $|\partial^2 J(v)/\partial v^2| \leq c_3$ for some constant $c_3 > 0$.

4. $\mathbb{E}|Y(1)|^2 < \infty$ and $\mathbb{E}|Y(0)|^2 < \infty$.

5. Let $\mathbf{G}^* := \mathbb{E}[\boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top \boldsymbol{\Delta}_i(\psi^*(\boldsymbol{X}_i))]$, where $\boldsymbol{\Delta}_i(\psi(\boldsymbol{X}_i)) = \text{diag}(\xi_i(\psi(\boldsymbol{X}_i))\mathbf{1}_{m_1}, \phi_i(\psi(\boldsymbol{X}_i))\mathbf{1}_{m_2})$ is a $\kappa \times \kappa$ diagonal matrix with

$$\xi_i(\psi(\boldsymbol{X}_i)) = -\Big(\frac{T_i}{J^2(\psi(\boldsymbol{X}_i))} + \frac{1 - T_i}{(1 - J(\psi(\boldsymbol{X}_i)))^2}\Big)\frac{\partial J(\psi(\boldsymbol{X}_i))}{\partial \psi},$$
$$\phi_i(\psi(\boldsymbol{X}_i)) = -\frac{T_i}{J^2(\psi(\boldsymbol{X}_i))}\frac{\partial J(\psi(\boldsymbol{X}_i))}{\partial \psi}.$$

Here, $\mathbf{1}_{m_1}$ is a vector of 1's with length m_1. Assume that there exists a constant $C_1 > 0$, such that $\lambda_{\min}(\mathbf{G}^{*\top}\mathbf{G}^*) \geq C_1$, where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of a matrix.

6. For some constant C, it holds $\|\mathbb{E}[\boldsymbol{h}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top]\|_2 \leq C$ and $\|\mathbb{E}[\boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{B}(\boldsymbol{X}_i)^\top]\|_2 \leq C$, where $\|\mathbf{A}\|_2$ denotes the spectral norm of the matrix $\mathbf{A}$. In addition, $\sup_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{h}(\boldsymbol{x})\|_2 \leq C\kappa^{1/2}$, and $\sup_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{B}(\boldsymbol{x})\|_2 \leq C\kappa^{1/2}$.

7. Let $m^*(\cdot) \in \mathcal{M}$ and $K(\cdot), L(\cdot) \in \mathcal{H}$, where $\mathcal{M}$ and $\mathcal{H}$ are two sets of smooth functions. Assume that $\log N_{[\,]}(\epsilon, \mathcal{M}, L_2(P)) \leq C(1/\epsilon)^{1/k_1}$ and $\log N_{[\,]}(\epsilon, \mathcal{H}, L_2(P)) \leq C(1/\epsilon)^{1/k_2}$, where C is a positive constant and $k_1, k_2 > 1/2$. Here, $N_{[\,]}(\epsilon, \mathcal{M}, L_2(P))$ denotes the minimum number of ϵ-brackets needed to cover $\mathcal{M}$; see Definition 2.1.6 of van der Vaart and Wellner (1996).

Note that the first five conditions are similar to Assumptions 3.1 and 3.2. In particular, Condition 5 is the natural extension of Condition 1 of Assumption 3.2, when the dimension of the matrix $\mathbf{G}^*$ grows with the sample size n. Condition 6 is a mild technical condition on the basis functions $\boldsymbol{h}(\boldsymbol{x})$ and $\boldsymbol{B}(\boldsymbol{x})$, which is implied by Assumption 2 of Newey (1997). In particular, this condition is satisfied by many bases such as the regression spline, trigonometric polynomial, wavelet bases; see Newey (1997); Horowitz et al. (2004); Chen (2007); Belloni et al. (2015). Finally, Condition 7 is a technical condition on the complexity of the function classes $\mathcal{M}$ and $\mathcal{H}$. Specifically, it requires that the bracketing number $N_{[\,]}(\epsilon, \cdot, L_2(P))$ of $\mathcal{M}$ and $\mathcal{H}$ cannot increase too fast as ϵ approaches to 0. This condition holds for many commonly used function classes. For instance, if $\mathcal{M}$ corresponds to the Hölder class with smoothness parameter s defined on a bounded convex subset of $\mathbb{R}^d$, then $\log N_{[\,]}(\epsilon, \mathcal{M}, L_2(P)) \leq C(1/\epsilon)^{d/s}$ by Corollary 2.6.2 of van der Vaart and Wellner (1996). Hence, this condition simply requires $s/d > 1/2$. Given Assumption E.1, the following theorem establishes the asymptotic normality and semiparametric efficiency of the estimator $\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}}$.

F Proof of Results in Section 4

For notational simplicity, we denote $\pi^*(\boldsymbol{x}) = J(m^*(\boldsymbol{x}))$, $J^*(\boldsymbol{x}) = J(\boldsymbol{\beta}^{*\top}\boldsymbol{B}(\boldsymbol{x}))$, and $\widetilde{J}(\boldsymbol{x}) = J(\widetilde{\boldsymbol{\beta}}^\top \boldsymbol{B}(\boldsymbol{x}))$. Define $Q_n(\boldsymbol{\beta}) = \|\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})\|_2^2$ and $Q(\boldsymbol{\beta}) = \|\mathbb{E}\boldsymbol{g}_{\boldsymbol{\beta}}(\boldsymbol{T}_i, \boldsymbol{X}_i)\|_2^2$. In the following proof, we use C, C' and C'' to denote generic positive constants, whose values may change from line to line. In this section, denote $K = \kappa$ and $\psi(\boldsymbol{X}) = m(\boldsymbol{X})$.

Lemma F.1 (Bernstein's inequality for U-statistics (Arcones, 1995))**.** Given i.i.d. random variables $Z_1, \ldots Z_n$ taking values in a measurable space $(\mathbb{S}, \mathcal{B})$ and a symmetric and measurable kernel function $h\colon \mathbb{S}^m \to R$, we define the U-statistics with kernel h as $U := \binom{n}{m}^{-1} \sum_{i_1<\ldots<i_m} h(Z_{i_1}, \ldots, Z_{i_m})$. Suppose that $\mathbb{E}h(Z_{i_1}, \ldots, Z_{i_m}) = 0$, $\mathbb{E}\big\{\mathbb{E}[h(Z_{i_1}, \ldots, Z_{i_m}) \mid Z_{i_1}]\big\}^2 = \sigma^2$ and $\|h\|_\infty \leq b$. There exists a constant $K(m) > 0$ depending on m such that

$$\mathbb{P}(|U| > t) \leq 4\exp\big\{-nt^2/[2m^2\sigma^2 + K(m)bt]\big\},\ \forall t > 0.$$

Lemma F.2. Under the conditions in Theorem 4.1, it holds that

$$\sup_{\boldsymbol{\beta}\in\Theta}\Big|Q_n(\boldsymbol{\beta}) - Q(\boldsymbol{\beta})\Big| = O_p\Big(\sqrt{\frac{K^2\log K}{n}}\Big).$$

Proof of Lemma F.2. Let $\boldsymbol{\xi}(\boldsymbol{\beta}) = (\xi_1(\boldsymbol{\beta}), ..., \xi_n(\boldsymbol{\beta}))^\top$ and $\boldsymbol{\phi}(\boldsymbol{\beta}) = (\phi_1(\boldsymbol{\beta}), ..., \phi_n(\boldsymbol{\beta}))^\top$, where

$$\xi_i(\boldsymbol{\beta}) = \frac{T_i}{J(\boldsymbol{\beta}^\top\boldsymbol{B}(\boldsymbol{X}_i))} - \frac{1-T_i}{1-J(\boldsymbol{\beta}^\top\boldsymbol{B}(\boldsymbol{X}_i))},\quad \phi_i(\boldsymbol{\beta}) = \frac{T_i}{J(\boldsymbol{\beta}^\top\boldsymbol{B}(\boldsymbol{X}_i))} - 1.$$

Then we have

$$\begin{aligned} Q_n(\boldsymbol{\beta}) &= n^{-2}\sum_{i=1}^n\sum_{j=1}^n \big[\xi_i(\boldsymbol{\beta})\xi_j(\boldsymbol{\beta})\boldsymbol{h}_1(\boldsymbol{X}_i)^\top\boldsymbol{h}_1(\boldsymbol{X}_j) + \phi_i(\boldsymbol{\beta})\phi_j(\boldsymbol{\beta})\boldsymbol{h}_2(\boldsymbol{X}_i)^\top\boldsymbol{h}_2(\boldsymbol{X}_j)\big] \\ &= n^{-2}\sum_{1\leq i\neq j\leq n}\big[\xi_i(\boldsymbol{\beta})\xi_j(\boldsymbol{\beta})\boldsymbol{h}_1(\boldsymbol{X}_i)^\top\boldsymbol{h}_1(\boldsymbol{X}_j) + \phi_i(\boldsymbol{\beta})\phi_j(\boldsymbol{\beta})\boldsymbol{h}_2(\boldsymbol{X}_i)^\top\boldsymbol{h}_2(\boldsymbol{X}_j)\big] + A_n(\boldsymbol{\beta}), \end{aligned}$$

where $A_n(\boldsymbol{\beta}) = n^{-2}\sum_{i=1}^n\big[\xi_i(\boldsymbol{\beta})^2\|\boldsymbol{h}_1(\boldsymbol{X}_i)\|_2^2 + \phi_i(\boldsymbol{\beta})^2\|\boldsymbol{h}_2(\boldsymbol{X}_i)\|_2^2\big]$. Since there exists a constant $c_0 > 0$ such that $c_0 \leq |J(\boldsymbol{\beta}^\top\boldsymbol{B}(\boldsymbol{x}))| \leq 1 - c_0$ for any $\boldsymbol{\beta} \in \Theta$ and $T_i \in \{0, 1\}$, it implies that

$\sup_{\boldsymbol{\beta}\in\Theta}\max_{1\le i\le n}|\xi_i(\boldsymbol{\beta})|\le C$ and $\sup_{\boldsymbol{\beta}\in\Theta}\max_{1\le i\le n}|\phi_i(\boldsymbol{\beta})|\le C$ for some constant $C>0$. Then we can show that

$$\mathbb{E}\Big(\sup_{\boldsymbol{\beta}\in\Theta}|A_n(\boldsymbol{\beta})|\Big)\le\frac{C}{n}\mathbb{E}(\|\boldsymbol{h}(\boldsymbol{X}_i)\|_2^2)=O(K/n).$$

By the Markov inequality, we have $\sup_{\boldsymbol{\beta}\in\Theta}|A_n(\boldsymbol{\beta})|=O_p(K/n)=o_p(1)$. Following the similar arguments, it can be easily shown that $\sup_{\boldsymbol{\beta}\in\Theta}|Q(\boldsymbol{\beta})|/n=O(K/n)$. Thus, it holds that

$$\sup_{\boldsymbol{\beta}\in\Theta}|Q_n(\boldsymbol{\beta})-Q(\boldsymbol{\beta})|=\sup_{\boldsymbol{\beta}\in\Theta}\Big|\frac{2}{n(n-1)}\sum_{1\le i<j\le n}u_{ij}(\boldsymbol{\beta})\Big|+O_p(K/n), \tag{F.1}$$

where $u_{ij}(\boldsymbol{\beta})=u_{1ij}(\boldsymbol{\beta})+u_{2ij}(\boldsymbol{\beta})$ is a kernel function of a U-statistic with

$$\begin{aligned}u_{1ij}(\boldsymbol{\beta})&=\xi_i(\boldsymbol{\beta})\xi_j(\boldsymbol{\beta})\boldsymbol{h}_1(\boldsymbol{X}_i)^\top\boldsymbol{h}_1(\boldsymbol{X}_j)-\mathbb{E}[\xi_i(\boldsymbol{\beta})\xi_j(\boldsymbol{\beta})\boldsymbol{h}_1(\boldsymbol{X}_i)^\top\boldsymbol{h}_1(\boldsymbol{X}_j)],\\ u_{2ij}(\boldsymbol{\beta})&=\phi_i(\boldsymbol{\beta})\phi_j(\boldsymbol{\beta})\boldsymbol{h}_2(\boldsymbol{X}_i)^\top\boldsymbol{h}_2(\boldsymbol{X}_j)-\mathbb{E}[\phi_i(\boldsymbol{\beta})\phi_j(\boldsymbol{\beta})\boldsymbol{h}_2(\boldsymbol{X}_i)^\top\boldsymbol{h}_2(\boldsymbol{X}_j)].\end{aligned}$$

Since Θ is a compact set in $\mathbb{R}^K$, by the covering number theory, there exists a constant C such that $M=(C/r)^K$ balls with the radius r can cover Θ. Namely, $\Theta\subseteq\cup_{1\le m\le M}\Theta_m$, where $\Theta_m=\{\boldsymbol{\beta}\in\mathbb{R}^K:\|\boldsymbol{\beta}-\boldsymbol{\beta}_m\|_2\le r\}$ for some $\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_M$. Thus, for any given $\epsilon>0$,

$$\begin{aligned}\mathbb{P}\Big(\sup_{\boldsymbol{\beta}\in\Theta}\Big|\frac{2}{n(n-1)}\sum_{1\le i<j\le n}u_{1ij}(\boldsymbol{\beta})\Big|>\epsilon\Big)&\le\sum_{m=1}^M\mathbb{P}\Big(\sup_{\boldsymbol{\beta}\in\Theta_m}\Big|\frac{2}{n(n-1)}\sum_{1\le i<j\le n}u_{1ij}(\boldsymbol{\beta})\Big|>\epsilon\Big)\\ &\le\sum_{m=1}^M\Big[\mathbb{P}\Big(\Big|\frac{2}{n(n-1)}\sum_{1\le i<j\le n}u_{1ij}(\boldsymbol{\beta}_m)\Big|>\epsilon/2\Big)\\ &\quad+\mathbb{P}\Big(\sup_{\boldsymbol{\beta}\in\Theta_m}\frac{2}{n(n-1)}\sum_{1\le i<j\le n}\Big|u_{1ij}(\boldsymbol{\beta})-u_{1ij}(\boldsymbol{\beta}_m)\Big|>\epsilon/2\Big)\Big].\end{aligned} \tag{F.2}$$

By the Cauchy-Schwarz inequality, $|\boldsymbol{h}_1(\boldsymbol{X}_i)^\top\boldsymbol{h}_1(\boldsymbol{X}_j)|\le\|\boldsymbol{h}_1(\boldsymbol{X}_i)\|_2\|\boldsymbol{h}_1(\boldsymbol{X}_j)\|_2\le CK$, and thus $|u_{1ij}(\boldsymbol{\beta}_m)|\le CK$. In addition, for any $\boldsymbol{\beta}$,

$$\begin{aligned}&\mathbb{E}\big\{\xi_i(\boldsymbol{\beta})\boldsymbol{h}_1(\boldsymbol{X}_i)^\top\mathbb{E}[\xi_j(\boldsymbol{\beta})\boldsymbol{h}_1(\boldsymbol{X}_j)]-\mathbb{E}[\xi_i(\boldsymbol{\beta})\xi_j(\boldsymbol{\beta})\boldsymbol{h}_1(\boldsymbol{X}_i)^\top\boldsymbol{h}_1(\boldsymbol{X}_j)]\big\}^2\\ &\le\mathbb{E}\big\{\xi_i(\boldsymbol{\beta})\boldsymbol{h}_1(\boldsymbol{X}_i)^\top\mathbb{E}[\xi_j(\boldsymbol{\beta})\boldsymbol{h}_1(\boldsymbol{X}_j)]\big\}^2\le\|\mathbb{E}\xi_i^2(\boldsymbol{\beta})\boldsymbol{h}_1(\boldsymbol{X}_i)\boldsymbol{h}_1(\boldsymbol{X}_i)^\top\|_2\cdot\|\mathbb{E}\xi_j(\boldsymbol{\beta})\boldsymbol{h}_1(\boldsymbol{X}_j)\|_2^2\le CK,\end{aligned}$$

for some constant $C>0$. Here, in the last step we use that fact that

$$\|\mathbb{E}\xi_j(\boldsymbol{\beta})\boldsymbol{h}_1(\boldsymbol{X}_j)\|_2^2\le\mathbb{E}\|\xi_j(\boldsymbol{\beta})\boldsymbol{h}_1(\boldsymbol{X}_j)\|_2^2\le C\cdot\mathbb{E}\|\boldsymbol{h}_1(\boldsymbol{X}_j)\|_2^2\le CK,$$

and $\|\mathbb{E}\xi_i^2(\boldsymbol{\beta})\boldsymbol{h}_1(\boldsymbol{X}_i)\boldsymbol{h}_1(\boldsymbol{X}_i)^\top\|_2$ is bounded because $\|\mathbb{E}\boldsymbol{h}_1(\boldsymbol{X}_j)\boldsymbol{h}_1(\boldsymbol{X}_j)^\top\|_2$ is bounded by assumption. Thus, we can apply the Bernstein's inequality in Lemma F.1 to the U-statistic with kernel function $u_{1ij}(\boldsymbol{\beta}_m)$,

$$\mathbb{P}\Big(\Big|\frac{2}{n(n-1)}\sum_{1\le i<j\le n} u_{1ij}(\boldsymbol{\beta}_m)\Big| > \epsilon/2\Big) \le 2\exp\big(-Cn\epsilon^2/[K+K\epsilon]\big), \tag{F.3}$$

for some constant $C>0$. Since $|\partial J(v)/\partial v|$ is upper bounded by a constant for any $v=\boldsymbol{\beta}^\top\boldsymbol{B}(\boldsymbol{x})$, it is easily seen that for any $\boldsymbol{\beta}\in\Theta_m$, $|\xi_i(\boldsymbol{\beta})-\xi_i(\boldsymbol{\beta}_m)|\le C|(\boldsymbol{\beta}-\boldsymbol{\beta}_m)^\top\boldsymbol{B}(\boldsymbol{X}_i)|\le CrK^{1/2}$, where the last step follows from the Cauchy-Schwarz inequalty. This further implies $|\xi_i(\boldsymbol{\beta})\xi_j(\boldsymbol{\beta})-\xi_i(\boldsymbol{\beta}_m)\xi_j(\boldsymbol{\beta}_m)|\le CrK^{1/2}$ for some constant $C>0$ by performing a standard perturbation analysis. Thus,

$$|u_{1ij}(\boldsymbol{\beta})-u_{1ij}(\boldsymbol{\beta}_m)|\le CrK^{1/2}|\boldsymbol{h}_1(\boldsymbol{X}_i)^\top\boldsymbol{h}_1(\boldsymbol{X}_j)|\le CrK^{3/2},$$

and note that with $r=K^{-2}$, then $CrK^{1/2}\mathbb{E}|\boldsymbol{h}_1(\boldsymbol{X}_i)^\top\boldsymbol{h}_1(\boldsymbol{X}_j)|\le\epsilon/4$ for n large enough. Thus

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{\boldsymbol{\beta}\in\Theta_m}\frac{2}{n(n-1)}\sum_{1\le i<j\le n}\Big|u_{1ij}(\boldsymbol{\beta})-u_{1ij}(\boldsymbol{\beta}_m)\Big|>\epsilon/2\Big)\\
&\le\mathbb{P}\Big(\frac{2CrK^{1/2}}{n(n-1)}\sum_{1\le i<j\le n}|\boldsymbol{h}_1(\boldsymbol{X}_i)^\top\boldsymbol{h}_1(\boldsymbol{X}_j)|>\epsilon/2\Big)\\
&\le\mathbb{P}\Big(\frac{2CrK^{1/2}}{n(n-1)}\sum_{1\le i<j\le n}\big[|\boldsymbol{h}_1(\boldsymbol{X}_i)^\top\boldsymbol{h}_1(\boldsymbol{X}_j)|-\mathbb{E}|\boldsymbol{h}_1(\boldsymbol{X}_i)^\top\boldsymbol{h}_1(\boldsymbol{X}_j)|\big]>\epsilon/4\Big)\\
&\le 2\exp(-CnK\epsilon^2),
\end{aligned}\tag{F.4}$$

where the last step follows from the Hoeffding inequality for U-statistic. Thus, combining (F.2), (F.3) and (F.4), we have for some constants $C_1,C_2,C_3>0$, as n goes to infinity,

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{\boldsymbol{\beta}\in\Theta}\Big|\frac{2}{n(n-1)}\sum_{1\le i<j\le n}u_{1ij}(\boldsymbol{\beta})\Big|>\epsilon\Big)\\
&\qquad\le\exp(C_1K\log K-C_2n\epsilon^2/[K+K\epsilon])+\exp(C_1K\log K-C_3n\epsilon^2K)\to 0,
\end{aligned}$$

where we take $\epsilon=C\sqrt{K^2\log K/n}$ for some constant C sufficiently large. This implies

$$\sup_{\boldsymbol{\beta}\in\Theta}\Big|\frac{2}{n(n-1)}\sum_{1\le i<j\le n}u_{1ij}(\boldsymbol{\beta})\Big|=O_p\Big(\sqrt{\frac{K^2\log K}{n}}\Big).$$

Following the same arguments, we can show that with the same choice of ϵ,

$$\sup_{\boldsymbol{\beta}\in\Theta}\Big|\frac{2}{n(n-1)}\sum_{1\le i<j\le n}u_{2ij}(\boldsymbol{\beta})\Big|=O_p\Big(\sqrt{\frac{K^2\log K}{n}}\Big).$$

Plugging these results into (F.1), we complete the proof. □

Lemma F.3 (Bernstein's inequality for random matrices (Tropp, 2015))**.** Let $\{\mathbf{Z}_k\}$ be a sequence of independent random matrices with dimensions $d_1 \times d_2$. Assume that $\mathbb{E}\mathbf{Z}_k = \mathbf{0}$ and $\|\mathbf{Z}_k\|_2 \leq R_n$ almost sure. Define

$$\sigma_n^2 = \max\Big\{\Big\|\sum_{k=1}^{n}\mathbb{E}(\mathbf{Z}_k\mathbf{Z}_k^\top)\Big\|_2, \Big\|\sum_{k=1}^{n}\mathbb{E}(\mathbf{Z}_k^\top\mathbf{Z}_k)\Big\|_2\Big\}.$$

Then, for all $t \geq 0$,

$$\mathbb{P}\Big(\Big\|\sum_{k=1}^{n}\mathbf{Z}_k\Big\|_2 \geq t\Big) \leq (d_1+d_2)\exp\Big(-\frac{t^2/2}{\sigma_n^2 + R_n t/3}\Big).$$

Lemma F.4. Let $\mathbf{H} = (\boldsymbol{h}(\boldsymbol{X}_1), ..., \boldsymbol{h}(\boldsymbol{X}_n))^\top$ and $\mathbf{B} = (\boldsymbol{B}(\boldsymbol{X}_1), ..., \boldsymbol{B}(\boldsymbol{X}_n))^\top$ be two $n \times K$ matrices. Under the conditions in Theorem 4.1, then

$$\|\mathbf{H}^\top\mathbf{H}/n - \mathbb{E}[\boldsymbol{h}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top]\|_2 = O_p(\sqrt{K\log K/n}) \tag{F.5}$$

and

$$\|\mathbf{B}^\top\mathbf{B}/n - \mathbb{E}[\boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{B}(\boldsymbol{X}_i)^\top]\|_2 = O_p(\sqrt{K\log K/n}). \tag{F.6}$$

Proof of Lemma F.4. We prove this result by applying Lemma F.3. In particular, to prove (F.5), we take $\mathbf{Z}_i = n^{-1}[\boldsymbol{h}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top - \mathbb{E}(\boldsymbol{h}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top)]$. It is easily seen that

$$\|\mathbf{Z}_i\|_2 \leq n^{-1}[\mathrm{tr}(\boldsymbol{h}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top) + \|\mathbb{E}(\boldsymbol{h}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top)\|_2] \leq (CK + C)/n,$$

where C is some positive constant. Moreover,

$$\begin{aligned}\Big\|\sum_{i=1}^{n}\mathbb{E}(\mathbf{Z}_i\mathbf{Z}_i^\top)\Big\|_2 &\leq n^{-1}\Big(\|\mathbb{E}\boldsymbol{h}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top\boldsymbol{h}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top\|_2 + \|\mathbb{E}(\boldsymbol{h}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top)\|_2^2\Big) \\ &\leq n^{-1}(CK \cdot \|\mathbb{E}(\boldsymbol{h}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top)\|_2 + C^2) \leq n^{-1}(C^2K + C^2).\end{aligned}$$

Note that $\sqrt{K\log K/n} = o(1)$. Now, if we take $t = C\sqrt{K\log K/n}$ in Lemma F.3 for some constant C sufficiently large, then we have $\mathbb{P}(\|\sum_{k=1}^{n}\mathbf{Z}_k\|_2 \geq t) \leq 2K\exp(-C'\log K)$ for some $C' > 1$. Then, the right hand side converges to 0, as $K \to \infty$. This completes the proof of (F.5). The proof of (F.6) follows from the same arguments and is omitted for simplicity. □

Lemma F.5. Under the conditions in Theorem 4.1, the following results hold.

1 Let $\bar{\boldsymbol{U}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{U}_i$, $\boldsymbol{U}_i = (\boldsymbol{U}_{i1}^\top, \boldsymbol{U}_{i2}^\top)^\top$, with

$$\boldsymbol{U}_{i1} = \Big(\frac{T_i}{\pi_i^*} - \frac{1-T_i}{1-\pi_i^*}\Big)\boldsymbol{h}_1(\boldsymbol{X}_i), \quad \boldsymbol{U}_{i2} = \Big(\frac{T_i}{\pi_i^*} - 1\Big)\boldsymbol{h}_2(\boldsymbol{X}_i).$$

Then $\|\bar{\boldsymbol{U}}\|_2 = O_p(K^{1/2}/n^{1/2})$.

2 Let $\mathbb{B}(r) = \{\boldsymbol{\beta} \in \mathbb{R}^K : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq r\}$, and $r = O(K^{1/2}/n^{1/2} + K^{-r_b})$. Then

$$\sup_{\boldsymbol{\beta}\in\mathbb{B}(r)} \Big\| \frac{\partial \bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X})}{\partial \boldsymbol{\beta}} - \mathbf{G}^* \Big\|_2 = O_p\Big(K^{1/2} r + \sqrt{\frac{K \log K}{n}}\Big).$$

3 Let $J_i = J(\boldsymbol{\beta}^\top \boldsymbol{B}(\boldsymbol{X}_i))$, $\dot{J}_i = \partial J(v)/\partial v|_{v=\boldsymbol{\beta}^\top \boldsymbol{B}(\boldsymbol{X}_i)}$, and

$$\mathbf{T}^* = \mathbb{E}\Big\{\Big[\frac{\mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i)}{\pi_i^*} - \frac{\mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i)}{1-\pi_i^*}\Big] \dot{J}_i^* \boldsymbol{B}(\boldsymbol{X}_i)\Big\}.$$

Then

$$\sup_{\boldsymbol{\beta}\in\mathbb{B}(r)} \Big\| \frac{1}{n}\sum_{i=1}^n \Big[\frac{T_i Y_i(1)}{J_i^2} + \frac{(1-T_i)Y_i(0)}{(1-J_i)^2}\Big] \dot{J}_i \boldsymbol{B}(\boldsymbol{X}_i) + \mathbf{G}^{*\top}\boldsymbol{\alpha}^* \Big\|_2 = O_p\Big(K^{1/2} r + K^{-r_h}\Big).$$

Proof of Lemma F.5. We start from the proof of the first result. Note that $\mathbb{E}(\boldsymbol{U}_i) = 0$. Then $\mathbb{E}\|\bar{\boldsymbol{U}}\|_2^2 = \mathbb{E}(\boldsymbol{U}_i^\top \boldsymbol{U}_i)/n$ and then there exists some constant $C > 0$,

$$\begin{aligned}\mathbb{E}\|\bar{\boldsymbol{U}}\|_2^2 &= \mathbb{E}\Big[n^{-1}\sum_{k=1}^K \Big(\frac{T_i}{\pi_i^*} - \frac{1-T_i}{1-\pi_i^*}\Big)^2 h_k(\boldsymbol{X}_i)^2 I(k \leq m_1) + \Big(\frac{T_i}{\pi_i^*} - 1\Big)^2 h_k(\boldsymbol{X}_i)^2 I(k > m_1)\Big] \\ &\leq C \sum_{k=1}^K \mathbb{E}\{h_k(\boldsymbol{X}_i)^2\}/n = O(K/n).\end{aligned}$$

By the Markov inequality, this implies $\|\bar{\boldsymbol{U}}\|_2 = O_p(K^{1/2}/n^{1/2})$, which completes the proof of the first result. In the following, we prove the second result. Denote

$$\begin{aligned}\xi_i(m(\boldsymbol{X}_i)) &= -\Big(\frac{T_i}{J^2(m(\boldsymbol{X}_i))} + \frac{1-T_i}{(1-J(m(\boldsymbol{X}_i)))^2}\Big)\dot{J}(m(\boldsymbol{X}_i)) \\ \phi_i(m(\boldsymbol{X}_i)) &= -\frac{T_i}{J^2(m(\boldsymbol{X}_i))}\dot{J}(m(\boldsymbol{X}_i)),\end{aligned}$$

and $\boldsymbol{\Delta}_i(m(\boldsymbol{X}_i)) = \mathrm{diag}(\xi_i(m(\boldsymbol{X}_i))\mathbf{1}_{m_1}, \phi_i(m(\boldsymbol{X}_i))\mathbf{1}_{m_2})$ is a $K \times K$ diagonal matrix, where $\mathbf{1}_{m_1}$ is a vector of 1 with length m_1. Then, note that

$$\frac{\partial \bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T},\boldsymbol{X})}{\partial \boldsymbol{\beta}} - \mathbf{G}^* = \frac{1}{n}\sum_{i=1}^n \boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top \boldsymbol{\Delta}_i(\boldsymbol{\beta}^\top \boldsymbol{B}(\boldsymbol{X}_i)) - \mathbb{E}[\boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top \boldsymbol{\Delta}_i(m^*(\boldsymbol{X}_i))],$$

which can be decomposed into the two terms $I_{\boldsymbol{\beta}} + II$, where

$$\begin{aligned}I_{\boldsymbol{\beta}} &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top [\boldsymbol{\Delta}_i(\boldsymbol{\beta}^\top \boldsymbol{B}(\boldsymbol{X}_i)) - \boldsymbol{\Delta}_i(m^*(\boldsymbol{X}_i))], \quad II = \sum_{i=1}^n \mathbf{Z}_i, \\ \mathbf{Z}_i &= n^{-1}\Big\{\boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top \boldsymbol{\Delta}_i(m^*(\boldsymbol{X}_i)) - \mathbb{E}[\boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top \boldsymbol{\Delta}_i(m^*(\boldsymbol{X}_i))]\Big\}.\end{aligned}$$

We first consider the term II. It can be easily verified that $\|\boldsymbol{\Delta}_i(m^*(\boldsymbol{X}_i))\|_2 \le C$ for some constant $C > 0$. In addition, $\|\boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top\|_2 \le \|\boldsymbol{B}(\boldsymbol{X}_i)\|_2 \cdot \|\boldsymbol{h}(\boldsymbol{X}_i)\|_2 \le CK$. Thus, $\|\mathbf{Z}_i\|_2 \le CK/n$. Following the similar argument in the proof of Lemma F.4,

$$\Big\|\sum_{i=1}^n \mathbb{E}(\mathbf{Z}_i\mathbf{Z}_i^\top)\Big\|_2 \le n^{-1}\|\mathbb{E}\boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top\boldsymbol{\Delta}_i(m^*(\boldsymbol{X}_i))\boldsymbol{\Delta}_i(m^*(\boldsymbol{X}_i))\boldsymbol{h}(\boldsymbol{X}_i)\boldsymbol{B}(\boldsymbol{X}_i)^\top\|_2 \\ + n^{-1}\|\mathbb{E}\boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top\boldsymbol{\Delta}_i(m^*(\boldsymbol{X}_i))\|_2^2.$$

We now consider the last two terms separately. Note that

$$\begin{aligned}&\|\mathbb{E}\boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top\boldsymbol{\Delta}_i(m^*(\boldsymbol{X}_i))\|_2^2 = \sup_{\|\mathbf{u}\|_2=1,\|\mathbf{v}\|_2=1}|\mathbb{E}\mathbf{u}^\top\boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top\boldsymbol{\Delta}_i(m^*(\boldsymbol{X}_i))\mathbf{v}|^2\\ &\le \sup_{\|\mathbf{u}\|_2=1}|\mathbb{E}\mathbf{u}^\top\boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{B}(\boldsymbol{X}_i)^\top\mathbf{u}|\cdot\sup_{\|\mathbf{v}\|_2=1}|\mathbb{E}\mathbf{v}^\top\boldsymbol{\Delta}_i(m^*(\boldsymbol{X}_i))\boldsymbol{h}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top\boldsymbol{\Delta}_i(m^*(\boldsymbol{X}_i))\mathbf{v}|\\ &\le \|\mathbb{E}(\boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{B}(\boldsymbol{X}_i)^\top)\|_2\cdot C\|\mathbb{E}(\boldsymbol{h}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top)\|_2 \le C', \end{aligned} \tag{F.7}$$

where C, C' are some positive constants. Following the similar arguments to (F.7),

$$\begin{aligned}&\|\mathbb{E}\boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top\boldsymbol{\Delta}_i(m^*(\boldsymbol{X}_i))\boldsymbol{\Delta}_i(m^*(\boldsymbol{X}_i))\boldsymbol{h}(\boldsymbol{X}_i)\boldsymbol{B}(\boldsymbol{X}_i)^\top\|_2\\ &\le CK\cdot\sup_{\|\mathbf{u}\|_2=1}|\mathbb{E}\mathbf{u}^\top\boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{B}(\boldsymbol{X}_i)^\top\mathbf{u}| \le CK\cdot\|\mathbb{E}\boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{B}(\boldsymbol{X}_i)^\top\|_2 \le C'K,\end{aligned}$$

for some constants $C, C' > 0$. This implies $\|\sum_{i=1}^n\mathbb{E}(\mathbf{Z}_i\mathbf{Z}_i^\top)\|_2 \le CK/n$. Thus, Lemma F.3 implies $\|II\|_2 = O_p(\sqrt{K\log K/n})$. Next, we consider the term $I_{\boldsymbol{\beta}}$. Following the similar arguments to (F.7), we can show that

$$\begin{aligned}\sup_{\boldsymbol{\beta}\in\mathbb{B}(r)}\|I_{\boldsymbol{\beta}}\|_2 &= \sup_{\boldsymbol{\beta}\in\mathbb{B}(r)}\sup_{\|\mathbf{u}\|_2=1,\|\mathbf{v}\|_2=1}\Big|\frac{1}{n}\sum_{i=1}^n\mathbf{u}^\top\boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top[\boldsymbol{\Delta}_i(\boldsymbol{\beta}^\top\boldsymbol{B}(\boldsymbol{X}_i)) - \boldsymbol{\Delta}_i(m^*(\boldsymbol{X}_i))]\mathbf{v}\Big|\\ &\le \Big\|\frac{1}{n}\sum_{i=1}^n\boldsymbol{B}(\boldsymbol{X}_i)\boldsymbol{B}(\boldsymbol{X}_i)^\top\Big\|_2^{1/2}\cdot\Big\|\frac{1}{n}\sum_{i=1}^n\boldsymbol{h}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top\Big\|_2^{1/2}\\ &\quad\cdot\sup_{\boldsymbol{\beta}\in\mathbb{B}(r)}\max_{1\le i\le n}\|\boldsymbol{\Delta}_i(\boldsymbol{\beta}^\top\boldsymbol{B}(\boldsymbol{X}_i)) - \boldsymbol{\Delta}_i(m^*(\boldsymbol{X}_i))\|_2\\ &\le C\sup_{\boldsymbol{\beta}\in\mathbb{B}(r)}\sup_{\boldsymbol{x}\in\mathcal{X}}|(\boldsymbol{\beta}^*-\boldsymbol{\beta})^\top\boldsymbol{B}(\boldsymbol{x})| + C\sup_{\boldsymbol{x}\in\mathcal{X}}|m^*(\boldsymbol{x}) - \boldsymbol{\beta}^{*\top}\boldsymbol{B}(\boldsymbol{x})|\\ &\le C'(K^{1/2}r + K^{-r_b}) \le C''K^{1/2}r,\end{aligned}$$

for some $C, C', C'' > 0$, where the second inequality follows from Lemma F.4 and the Lipschitz property of $\xi_i(\cdot)$ and $\phi_i(\cdot)$, and the third inequality is due to the Cauchy-Schwarz inequality and approximation assumption of the sieve estimator. This completes the proof of the second result. For the third result, let

$$\eta_i(m(\boldsymbol{X}_i)) = \Big(\frac{T_i Y_i(1)}{J^2(m(\boldsymbol{X}_i))} + \frac{(1-T_i)Y_i(0)}{(1-J(m(\boldsymbol{X}_i)))^2}\Big)\dot{J}(m(\boldsymbol{X}_i)).$$

Thus, the following decomposition holds,

$$\frac{1}{n}\sum_{i=1}^{n}\eta_i(\boldsymbol{\beta}^\top \boldsymbol{B}(\boldsymbol{X}_i))\boldsymbol{B}(\boldsymbol{X}_i) + \mathbf{G}^{*\top}\boldsymbol{\alpha}^* = T_{1\boldsymbol{\beta}} + T_2 + T_3,$$

where

$$\begin{aligned} T_{1\boldsymbol{\beta}} &= \frac{1}{n}\sum_{i=1}^{n}[\eta_i(\boldsymbol{\beta}^\top \boldsymbol{B}(\boldsymbol{X}_i)) - \eta_i(m^*(\boldsymbol{B}(\boldsymbol{X}_i)))]\boldsymbol{B}(\boldsymbol{X}_i) \\ T_2 &= \frac{1}{n}\sum_{i=1}^{n}\Big[\eta_i(m^*(\boldsymbol{B}(\boldsymbol{X}_i)))\boldsymbol{B}(\boldsymbol{X}_i) - \mathbb{E}\eta_i(m^*(\boldsymbol{B}(\boldsymbol{X}_i)))\boldsymbol{B}(\boldsymbol{X}_i)\Big] \\ T_3 &= \mathbb{E}\eta_i(m^*(\boldsymbol{B}(\boldsymbol{X}_i)))\boldsymbol{B}(\boldsymbol{X}_i) + \mathbf{G}^{*\top}\boldsymbol{\alpha}^*. \end{aligned}$$

Similar to the proof for $\sup_{\boldsymbol{\beta}\in\mathbb{B}(r)}\|I_{\boldsymbol{\beta}}\|_2$ previously, we can easily show that $\sup_{\boldsymbol{\beta}\in\mathbb{B}(r)}\|T_{1\boldsymbol{\beta}}\|_2 = O_p(K^{1/2}r)$. Again, the key step is to use the results from Lemma F.4. For the second term T_2, we can use the similar arguments in the proof of the first result to show that $\mathbb{E}\|T_2\|_2^2 \le CK\cdot\mathbb{E}[\eta_i(m^*(\boldsymbol{B}(\boldsymbol{X}_i))^2]/n = O(K/n)$. The Markov inequality implies $\|T_2\|_2 = O_p(K^{1/2}/n^{1/2})$. For the third term T_3, after some algebra, we can show that

$$\|T_3\|_2 \le C\Big(\sup_{\boldsymbol{x}\in\mathcal{X}}|K(\boldsymbol{x}) - \boldsymbol{\alpha}_1^{*\top}\boldsymbol{h}_1(\boldsymbol{x})| + \sup_{\boldsymbol{x}\in\mathcal{X}}|L(\boldsymbol{x}) - \boldsymbol{\alpha}_2^{*\top}\boldsymbol{h}_2(\boldsymbol{x})|\Big) = O_p(K^{-r_h}).$$

Combining the L_2 error bound for $T_{1\boldsymbol{\beta}}$, T_2 and T_3, we obtain the last result. This completes the whole proof. □

Lemma F.6. Under the conditions in Theorem 4.1, it holds that

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = o_p(1).$$

Proof of Lemma F.6. Recall that $\boldsymbol{\beta}^o$ is the minimizer of $Q(\boldsymbol{\beta})$. We now decompose $Q(\widetilde{\boldsymbol{\beta}}) - Q(\boldsymbol{\beta}^o)$ as

$$Q(\widetilde{\boldsymbol{\beta}}) - Q(\boldsymbol{\beta}^o) = \underbrace{[Q(\widetilde{\boldsymbol{\beta}}) - Q_n(\widetilde{\boldsymbol{\beta}})]}_{I} + \underbrace{[Q_n(\widetilde{\boldsymbol{\beta}}) - Q_n(\boldsymbol{\beta}^o)]}_{II} + \underbrace{[Q_n(\boldsymbol{\beta}^o) - Q(\boldsymbol{\beta}^o)]}_{III}. \tag{F.8}$$

In the following, we study the terms I, II and III one by one. For the term I, Lemma F.2 implies $|Q(\widetilde{\boldsymbol{\beta}}) - Q_n(\widetilde{\boldsymbol{\beta}})| \le \sup_{\boldsymbol{\beta}\in\Theta}\Big|Q_n(\boldsymbol{\beta}) - Q(\boldsymbol{\beta})\Big| = o_p(1)$. This shows that $|I| = o_p(1)$ and the same argument yields $|III| = o_p(1)$. For the term II, by the definition of $\widetilde{\boldsymbol{\beta}}$, it is easy to see that $II \le 0$. Thus, combining with (F.8), we have for any constant $\eta > 0$ to be chosen later, $Q(\widetilde{\boldsymbol{\beta}}) - Q(\boldsymbol{\beta}^o) < \eta$ with probability tending to one. For any $\epsilon > 0$, define $E_\epsilon = \Theta \cap \{\|\boldsymbol{\beta} - \boldsymbol{\beta}^o\|_2 \ge \epsilon\}$. By the uniqueness of $\boldsymbol{\beta}^o$, for any $\boldsymbol{\beta} \in E_\epsilon$, we have $Q(\boldsymbol{\beta}) > Q(\boldsymbol{\beta}^o)$. Since E_ϵ is a compact set, we have $\inf_{\boldsymbol{\beta} \in E_\epsilon} Q(\boldsymbol{\beta}) > Q(\boldsymbol{\beta}^o)$. This implies that for any $\epsilon > 0$, there exists $\eta' > 0$ such that $Q(\boldsymbol{\beta}) > Q(\boldsymbol{\beta}^o) + \eta'$ for any $\boldsymbol{\beta} \in E_\epsilon$. If $\widetilde{\boldsymbol{\beta}} \in E_\epsilon$, then $Q(\boldsymbol{\beta}^o) + \eta > Q(\widetilde{\boldsymbol{\beta}}) > Q(\boldsymbol{\beta}^o) + \eta'$ with probability tending to one. Apparently, this does not holds if we take $\eta < \eta'$. Thus, we have proved that $\widetilde{\boldsymbol{\beta}} \notin E_\epsilon$, that is $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_2 \le \epsilon$ for any $\epsilon > 0$. Thus, we have $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_2 = o_p(1)$.

Next, we shall show that $\|\boldsymbol{\beta}^o - \boldsymbol{\beta}^*\|_2 = o_p(1)$. It is easily seen that these together lead to the desired consistency result

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le \|\boldsymbol{\beta}^o - \boldsymbol{\beta}^*\|_2 + \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_2 = o_p(1).$$

To show $\|\boldsymbol{\beta}^o - \boldsymbol{\beta}^*\|_2 = o_p(1)$, we use the similar strategy. That is we want to show that for any constant $\eta > 0$, $Q(\boldsymbol{\beta}^*) - Q(\boldsymbol{\beta}^o) < \eta$. In the following, we prove that $Q(\boldsymbol{\beta}^*) = O(K^{1-2r_b})$. Note that

$$Q(\boldsymbol{\beta}^*) \le C^2 K^{-2r_b} \sum_{j=1}^{K} \mathbb{E}|\boldsymbol{h}_j(\boldsymbol{X})|^2 = O(K^{1-2r_b}),$$

where the first inequality follows from the Cauchy-Schwarz inequality and the last step uses the assumption that $\sup_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{h}(\boldsymbol{x})\|_2 = O(K^{1/2})$. In addition, it holds that $Q(\boldsymbol{\beta}^o) \le Q(\boldsymbol{\beta}^*) = O(K^{1-2r_b})$. As $K \to \infty$, it yields $Q(\boldsymbol{\beta}^*) - Q(\boldsymbol{\beta}^o) < \eta$, for any constant $\eta > 0$. The same arguments yield $\|\boldsymbol{\beta}^o - \boldsymbol{\beta}^*\|_2 = o_p(1)$. This completes the proof of the consistency result. □

Lemma F.7. Under the conditions in Theorem 4.1, there exists a global minimizer $\widetilde{\boldsymbol{\beta}}$ (if $Q_n(\boldsymbol{\beta})$ has multiple minimizers), such that

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p(K^{1/2}/n^{1/2} + K^{-r_b}). \tag{F.9}$$

Proof of Lemma F.7. We first prove that there exists a local minimizer $\widetilde{\boldsymbol{\Delta}}$ of $Q_n(\boldsymbol{\beta}^* + \boldsymbol{\Delta})$, such that $\widetilde{\Delta} \in \mathcal{C}$, where $\mathcal{C} = \{\boldsymbol{\Delta} \in \mathbb{R}^K : \|\boldsymbol{\Delta}\|_2 \le r\}$, and $r = C(K^{1/2}/n^{1/2} + K^{-r_b})$ for some constant C large enough. To this end, it suffices to show that

$$\mathbb{P}\Big\{ \inf_{\boldsymbol{\Delta} \in \partial \mathcal{C}} Q_n(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - Q_n(\boldsymbol{\beta}^*) > 0 \Big\} \to 1, \quad \text{as } n \to \infty, \tag{F.10}$$

where $\partial\mathcal{C} = \{\boldsymbol{\Delta} \in \mathbb{R}^K : \|\boldsymbol{\Delta}\|_2 = r\}$. Applying the mean value theorem to each component of $\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*+\boldsymbol{\Delta}}(\boldsymbol{T}, \boldsymbol{X})$,

$$\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*+\boldsymbol{\Delta}}(\boldsymbol{T}, \boldsymbol{X}) = \bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*}(\boldsymbol{T}, \boldsymbol{X}) + \widetilde{\mathbf{G}}\boldsymbol{\Delta},$$

where $\widetilde{\mathbf{G}} = \frac{\partial \bar{\boldsymbol{g}}_{\bar{\boldsymbol{\beta}}}(\boldsymbol{T},\boldsymbol{X})}{\partial \boldsymbol{\beta}}$ and for notational simplicity we assume there exists a common $\bar{\boldsymbol{\beta}} = v\boldsymbol{\beta}^* + (1-v)\widetilde{\boldsymbol{\beta}}$ for some $0 \le v \le 1$ lies between $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^* + \boldsymbol{\Delta}$ (Rigorously speaking, we need different $\bar{\boldsymbol{\beta}}$ for different component of $\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*+\boldsymbol{\Delta}}(\boldsymbol{T}, \boldsymbol{X})$). Thus, for any $\boldsymbol{\Delta} \in \partial\mathcal{C}$,

$$\begin{aligned} Q_n(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - Q_n(\boldsymbol{\beta}^*) &= 2\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*}(\boldsymbol{T}, \boldsymbol{X})\widetilde{\mathbf{G}}\boldsymbol{\Delta} + \boldsymbol{\Delta}^\top(\widetilde{\mathbf{G}}^\top\widetilde{\mathbf{G}})\boldsymbol{\Delta} \\ &\ge -2\|\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*}(\boldsymbol{T}, \boldsymbol{X})\|_2 \cdot \|\widetilde{\mathbf{G}}\|_2 \cdot \|\boldsymbol{\Delta}\|_2 + \|\boldsymbol{\Delta}\|_2^2 \cdot \lambda_{\min}(\widetilde{\mathbf{G}}^\top\widetilde{\mathbf{G}}) \\ &\ge -C(K^{1/2}/n^{1/2} + K^{-r_b}) \cdot r + C \cdot r^2, \end{aligned} \tag{F.11}$$

for some constant $C > 0$. In the last step, we first use the results that $\|\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*}(\boldsymbol{T}, \boldsymbol{X})\|_2 = O_p(K^{1/2}/n^{1/2} + K^{-r_b})$, which is derived by combining Lemma F.5 with the arguments similar to (F.14) in the proof of Lemma F.8. In addition, $\|\widetilde{\mathbf{G}}\|_2 \le \|\widetilde{\mathbf{G}} - \mathbf{G}^*\|_2 + \|\mathbf{G}^*\|_2 \le C$, since $\|\mathbf{G}^*\|_2$ is bounded by a constant and $\|\widetilde{\mathbf{G}} - \mathbf{G}^*\|_2 = o_p(1)$ by Lemma F.5. By the Weyl inequality and Lemma F.5,

$$\begin{aligned} \lambda_{\min}(\widetilde{\mathbf{G}}^\top\widetilde{\mathbf{G}}) &\ge \lambda_{\min}(\mathbf{G}^{*\top}\mathbf{G}^*) - \|\widetilde{\mathbf{G}}^\top\widetilde{\mathbf{G}} - \mathbf{G}^{*\top}\mathbf{G}^*\|_2 \\ &\ge C - \|\widetilde{\mathbf{G}} - \mathbf{G}^*\|_2 \cdot \|\widetilde{\mathbf{G}}\|_2 - \|\widetilde{\mathbf{G}} - \mathbf{G}^*\|_2 \cdot \|\mathbf{G}^*\|_2 \ge C/2, \end{aligned}$$

for n sufficiently large. By (F.11), if $r = C(K^{1/2}/n^{1/2} + K^{-r_b})$ for some constant C large enough, the right hand side is positive for n large enough. This establishes (F.10). Next, we show that $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + \widetilde{\boldsymbol{\Delta}}$ is a global minimizer of $Q_n(\boldsymbol{\beta})$. This is true because the first order condition implies

$$\Big(\frac{\partial \bar{\boldsymbol{g}}_{\widetilde{\boldsymbol{\beta}}}(\boldsymbol{T}, \boldsymbol{X})}{\partial \boldsymbol{\beta}}\Big)\bar{\boldsymbol{g}}_{\widetilde{\boldsymbol{\beta}}}(\boldsymbol{T}, \boldsymbol{X}) = 0, \quad \Longrightarrow \quad \bar{\boldsymbol{g}}_{\widetilde{\boldsymbol{\beta}}}(\boldsymbol{T}, \boldsymbol{X}) = 0,$$

provided $\partial\bar{\boldsymbol{g}}_{\widetilde{\boldsymbol{\beta}}}(\boldsymbol{T}, \boldsymbol{X})/\partial\boldsymbol{\beta}$ is invertible. Following the similar arguments by applying the Weyl inequality, $\partial\bar{\boldsymbol{g}}_{\widetilde{\boldsymbol{\beta}}}(\boldsymbol{T}, \boldsymbol{X})/\partial\boldsymbol{\beta}$ is invertible with probability tending to one. Since $\bar{\boldsymbol{g}}_{\widetilde{\boldsymbol{\beta}}}(\boldsymbol{T}, \boldsymbol{X}) = 0$, it implies $Q_n(\widetilde{\boldsymbol{\beta}}) = 0$. Noting that $Q_n(\boldsymbol{\beta}) \ge 0$ for any $\boldsymbol{\beta}$, we obtain that $\widetilde{\boldsymbol{\beta}}$ is indeed a global minimizer of $Q_n(\boldsymbol{\beta})$. □

Lemma F.8. Under the conditions in Theorem 4.1, $\widetilde{\boldsymbol{\beta}}$ satisfies the following asymptotic expansion

$$\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = -\mathbf{G}^{-1}\bar{\boldsymbol{U}} + \boldsymbol{\Delta}_n, \tag{F.12}$$

where $\bar{\boldsymbol{U}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{U}_i$, $\boldsymbol{U}_i = (\boldsymbol{U}_{i1}^\top, \boldsymbol{U}_{i2}^\top)^\top$, with

$$\boldsymbol{U}_{i1} = \left(\frac{T_i}{\pi_i^*} - \frac{1-T_i}{1-\pi_i^*}\right)\boldsymbol{h}_1(\boldsymbol{X}_i), \quad \boldsymbol{U}_{i2} = \left(\frac{T_i}{\pi_i^*} - 1\right)\boldsymbol{h}_2(\boldsymbol{X}_i),$$

and

$$\|\boldsymbol{\Delta}_n\|_2 = O_p\Big(K^{1/2}\cdot\Big(\frac{K^{1/2}}{n^{1/2}} + \frac{1}{K^{r_b}}\Big)^2 + \sqrt{\frac{K\log K}{n}}\cdot\Big(\frac{K^{1/2}}{n^{1/2}} + \frac{1}{K^{r_b}}\Big)\Big).$$

Proof of Lemma F.8. Similar to the proof of Lemma F.7, we apply the mean value theorem to each component of $\bar{\boldsymbol{g}}_{\widetilde{\boldsymbol{\beta}}}(\boldsymbol{T}, \boldsymbol{X})$,

$$\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*}(\boldsymbol{T}, \boldsymbol{X}) + \Big(\frac{\partial \bar{\boldsymbol{g}}_{\bar{\boldsymbol{\beta}}}(\boldsymbol{T}, \boldsymbol{X})}{\partial \boldsymbol{\beta}}\Big)(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = 0,$$

where for notational simplicity we assume there exists a common $\bar{\boldsymbol{\beta}} = v\boldsymbol{\beta}^* + (1-v)\widetilde{\boldsymbol{\beta}}$ for some $0 \le v \le 1$ lies between $\boldsymbol{\beta}^*$ and $\widetilde{\boldsymbol{\beta}}$. After rearrangement, we derive

$$\begin{aligned}\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* &= -\mathbf{G}^{*-1}\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*}(\boldsymbol{T}, \boldsymbol{X}) + \Big[\mathbf{G}^{*-1} - \Big(\frac{\partial \bar{\boldsymbol{g}}_{\bar{\boldsymbol{\beta}}}(\boldsymbol{T}, \boldsymbol{X})}{\partial \boldsymbol{\beta}}\Big)^{-1}\Big]\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*}(\boldsymbol{T}, \boldsymbol{X}) \\ &= -\mathbf{G}^{*-1}\bar{\boldsymbol{U}} + \boldsymbol{\Delta}_{n1} + \boldsymbol{\Delta}_{n2} + \boldsymbol{\Delta}_{n3}, \end{aligned} \tag{F.13}$$

where

$$\boldsymbol{\Delta}_{n1} = \mathbf{G}^{*-1}[\bar{\boldsymbol{U}} - \bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*}(\boldsymbol{T}, \boldsymbol{X})], \quad \boldsymbol{\Delta}_{n2} = \Big[\mathbf{G}^{*-1} - \Big(\frac{\partial \bar{\boldsymbol{g}}_{\bar{\boldsymbol{\beta}}}(\boldsymbol{T}, \boldsymbol{X})}{\partial \boldsymbol{\beta}}\Big)^{-1}\Big]\bar{\boldsymbol{U}}$$

and

$$\boldsymbol{\Delta}_{n3} = \Big[\mathbf{G}^{*-1} - \Big(\frac{\partial \bar{\boldsymbol{g}}_{\bar{\boldsymbol{\beta}}}(\boldsymbol{T}, \boldsymbol{X})}{\partial \boldsymbol{\beta}}\Big)^{-1}\Big]\cdot[\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*}(\boldsymbol{T}, \boldsymbol{X}) - \bar{\boldsymbol{U}}].$$

We first consider $\boldsymbol{\Delta}_{n1}$ in (F.13). Let $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^\top$, where

$$\xi_i = T_i\Big(\frac{1}{\pi_i^*} - \frac{1}{J_i^*}\Big) - (1-T_i)\Big(\frac{1}{1-\pi_i^*} - \frac{1}{1-J_i^*}\Big), \quad \text{for } 1 \le i \le m_1,$$

and

$$\xi_i = T_i\Big(\frac{1}{\pi_i^*} - \frac{1}{J_i^*}\Big), \quad \text{for } m_1 + 1 \le i \le K.$$

Let $\mathbf{H} = (\boldsymbol{h}(X_1), ..., \boldsymbol{h}(X_n))^\top$ be a $n \times K$ matrix. Then, for some constants $C, C' > 0$,

$$\begin{aligned}\|\boldsymbol{\Delta}_{n1}\|_2^2 &= n^{-2}\boldsymbol{\xi}^\top\mathbf{H}\mathbf{G}^{*-1}\mathbf{G}^{*-1}\mathbf{H}^\top\boldsymbol{\xi} \le n^{-2}\|\boldsymbol{\xi}\|_2^2\cdot\|\mathbf{H}\mathbf{G}^{*-1}\mathbf{G}^{*-1}\mathbf{H}^\top\|_2 \\ &\le Cn^{-1}\|\boldsymbol{\xi}\|_2^2\cdot\|\mathbf{H}^\top\mathbf{H}/n\|_2 \le C'n^{-1}\|\boldsymbol{\xi}\|_2^2, \end{aligned} \tag{F.14}$$

where the third step follows from the fact that $\|\mathbf{G}^{*-1}\|_2$ is bounded and the last step follows from Lemma F.4 and the maximum eigenvalue of $\mathbb{E}[\boldsymbol{h}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)^\top]$ is bounded. Since $|\partial J(v)/\partial v|$ is upper bounded by a constant for any $v \leq \sup_{\boldsymbol{x}\in\mathcal{X}} |m^*(\boldsymbol{x})|$, then there exist some constants $C, C' > 0$, suc that for any $m_1 + 1 \leq i \leq K$,

$$|\xi_i| \leq C|\pi_i^* - J_i^*| \leq C' \sup_{\boldsymbol{x}\in\mathcal{X}} |m^*(\boldsymbol{x}) - \boldsymbol{\beta}^{*\top}\boldsymbol{B}(\boldsymbol{x})| \leq C'K^{-r_b}.$$

Similarly, $|\xi_i| \leq 2C'K^{-r_b}$ for any $1 \leq i \leq m_1$. Thus, it yields $n^{-1}\|\boldsymbol{\xi}\|_2^2 = O_p(K^{-2r_b})$. Combining with (F.14), we conclude that $\|\boldsymbol{\Delta}_{n1}\|_2 = O_p(K^{-r_b})$.

Next, we consider $\boldsymbol{\Delta}_{n2}$. Since $\|\mathbf{G}^{*-1}\|_2$ is bounded, we have

$$\begin{aligned}\|\boldsymbol{\Delta}_{n2}\|_2 &\leq \|\mathbf{G}^{*-1}\|_2 \cdot \Big\|\Big(\frac{\partial \bar{\boldsymbol{g}}_{\bar{\boldsymbol{\beta}}}(\boldsymbol{T},\boldsymbol{X})}{\partial\boldsymbol{\beta}}\Big)^{-1}\Big\|_2 \cdot \Big\|\mathbf{G}^* - \frac{\partial \bar{\boldsymbol{g}}_{\bar{\boldsymbol{\beta}}}(\boldsymbol{T},\boldsymbol{X})}{\partial\boldsymbol{\beta}}\Big\|_2 \cdot \|\bar{\boldsymbol{U}}\|_2 \\ &\leq C\Big(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 K^{1/2} + \sqrt{\frac{K\log K}{n}}\Big) \cdot \sqrt{\frac{K}{n}},\end{aligned}$$

where the last step follows from Lemma F.5.

Finally, we consider $\boldsymbol{\Delta}_{n3}$. By the same arguments in the control of terms $\boldsymbol{\Delta}_{n1}$ and $\boldsymbol{\Delta}_{n2}$, we can prove that

$$\begin{aligned}\|\boldsymbol{\Delta}_{n3}\|_2 &\leq \Big\|\mathbf{G}^{*-1} - \Big(\frac{\partial \bar{\boldsymbol{g}}_{\bar{\boldsymbol{\beta}}}(\boldsymbol{T},\boldsymbol{X})}{\partial\boldsymbol{\beta}}\Big)^{-1}\Big\|_2 \cdot \|\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*}(\boldsymbol{T},\boldsymbol{X}) - \bar{\boldsymbol{U}}\|_2 \\ &\leq C\Big(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 K^{1/2} + \sqrt{\frac{K\log K}{n}}\Big) \cdot K^{-r_b}.\end{aligned}$$

Combining the rates of $\|\boldsymbol{\Delta}_{n1}\|_2$, $\|\boldsymbol{\Delta}_{n2}\|_2$ and $\|\boldsymbol{\Delta}_{n3}\|_2$ with (F.13), by Lemma F.5, we obtain

$$\begin{aligned}\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 &\leq \|\mathbf{G}^{*-1}\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*}(\boldsymbol{T},\boldsymbol{X})\|_2 + \|\boldsymbol{\Delta}_{n1}\|_2 + \|\boldsymbol{\Delta}_{n2}\|_2 + \|\boldsymbol{\Delta}_{n3}\|_2 \\ &\leq C\Big(\frac{K^{1/2}}{n^{1/2}} + \frac{1}{K^{r_b}}\Big) + C'\Big(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 K^{1/2} + \sqrt{\frac{K\log K}{n}}\Big) \cdot \Big(\frac{K^{1/2}}{n^{1/2}} + \frac{1}{K^{r_b}}\Big),\end{aligned}$$

for some constants $C, C' > 0$. Therefore, (F.12) holds with $\boldsymbol{\Delta}_n = \boldsymbol{\Delta}_{n1} + \boldsymbol{\Delta}_{n2} + \boldsymbol{\Delta}_{n3}$, where

$$\|\boldsymbol{\Delta}_n\|_2 = O_p\Big(K^{1/2} \cdot \Big(\frac{K^{1/2}}{n^{1/2}} + \frac{1}{K^{r_b}}\Big)^2 + \sqrt{\frac{K\log K}{n}} \cdot \Big(\frac{K^{1/2}}{n^{1/2}} + \frac{1}{K^{r_b}}\Big)\Big).$$

This completes the proof. □

Proof of Theorem 4.1. We now consider the following decomposition of $\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}} - \mu$,

$$\begin{aligned}
\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}} - \mu = & \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{T_i(Y_i(1) - K(\boldsymbol{X}_i) - L(\boldsymbol{X}_i))}{\widetilde{J}_i} - \frac{(1-T_i)(Y_i(0) - K(\boldsymbol{X}_i))}{1-\widetilde{J}_i}\Big] \\
& + \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{T_i}{\widetilde{J}_i} - \frac{1-T_i}{1-\widetilde{J}_i}\Big)K(\boldsymbol{X}_i) + \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{T_i}{\widetilde{J}_i} - 1\Big)L(\boldsymbol{X}_i) + \frac{1}{n}\sum_{i=1}^{n}L(\boldsymbol{X}_i) - \mu \\
= & \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{T_i(Y_i(1) - K(\boldsymbol{X}_i) - L(\boldsymbol{X}_i))}{\widetilde{J}_i} - \frac{(1-T_i)(Y_i(0) - K(\boldsymbol{X}_i))}{1-\widetilde{J}_i}\Big] \\
& + \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{T_i}{\widetilde{J}_i} - \frac{1-T_i}{1-\widetilde{J}_i}\Big)\Delta_K(\boldsymbol{X}_i) + \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{T_i}{\widetilde{J}_i} - 1\Big)\Delta_L(\boldsymbol{X}_i) + \frac{1}{n}\sum_{i=1}^{n}L(\boldsymbol{X}_i) - \mu,
\end{aligned}$$

where $\widetilde{J}_i = J(\widetilde{\boldsymbol{\beta}}^\top \boldsymbol{B}(X_i))$, $\Delta_K(\boldsymbol{X}_i) = K(\boldsymbol{X}_i) - \boldsymbol{\alpha}_1^{*\top}\boldsymbol{h}_1(\boldsymbol{X}_i)$ and $\Delta_L(\boldsymbol{X}_i) = L(\boldsymbol{X}_i) - \boldsymbol{\alpha}_2^{*\top}\boldsymbol{h}_2(\boldsymbol{X}_i)$. Here, the second equality holds by the definition of $\widetilde{\boldsymbol{\beta}}$. Thus, we have

$$\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}} - \mu = \frac{1}{n}\sum_{i=1}^{n} S_i + R_0 + R_1 + R_2 + R_3$$

where

$$S_i = \frac{T_i}{\pi_i^*}\big[Y_i(1) - K(\boldsymbol{X}_i) - L(\boldsymbol{X}_i)\big] - \frac{1-T_i}{1-\pi_i^*}\big[Y_i(0) - K(\boldsymbol{X}_i)\big] + L(\boldsymbol{X}_i) - \mu,$$

$$R_0 = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i(Y_i(1) - K(\boldsymbol{X}_i) - L(\boldsymbol{X}_i))}{\widetilde{J}_i\pi_i^*}(\pi_i^* - \widetilde{J}_i),$$

$$R_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)(Y_i(0) - K(\boldsymbol{X}_i))}{(1-\widetilde{J}_i)(1-\pi_i^*)}(\pi_i^* - \widetilde{J}_i),$$

$$R_2 = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{T_i}{\widetilde{J}_i} - \frac{1-T_i}{1-\widetilde{J}_i}\Big)\Delta_K(\boldsymbol{X}_i), \quad R_3 = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{T_i}{\widetilde{J}_i} - 1\Big)\Delta_L(\boldsymbol{X}_i).$$

In the following, we will show that $R_j = o_p(n^{-1/2})$ for $0 \le j \le 3$. Thus, the asymptotic normality of $n^{1/2}(\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}} - \mu)$ follows from the previous decomposition. In addition, S_i agrees with the efficient score function for estimating μ (Hahn, 1998). Thus, the proposed estimator $\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}}$ is also semiparametrically efficient.

Now, we first focus on R_0. Consider the following empirical process $\mathbb{G}_n(f_0) = n^{1/2}(\mathbb{P}_n - \mathbb{P})f_0(T, Y(1), \boldsymbol{X})$, where $\mathbb{P}_n$ stands for the empirical measure and $\mathbb{P}$ stands for the expectation, and

$$f_0(T, Y(1), \boldsymbol{X}) = \frac{T(Y(1) - K(\boldsymbol{X}) - L(\boldsymbol{X}))}{J(m(\boldsymbol{X}))\pi^*(\boldsymbol{X})}[\pi^*(\boldsymbol{X}) - J(m(\boldsymbol{X}))].$$

By Lemma F.7, we can easily show that

$$\sup_{\boldsymbol{x}\in\mathcal{X}}|J(\widetilde{\boldsymbol{\beta}}^\top \boldsymbol{B}(\boldsymbol{x})) - \pi^*(\boldsymbol{x})| \lesssim \sup_{\boldsymbol{x}\in\mathcal{X}} |\widetilde{\boldsymbol{\beta}}^\top \boldsymbol{B}(\boldsymbol{x}) - \boldsymbol{\beta}^{*\top}\boldsymbol{B}(\boldsymbol{x})| \\ + \sup_{\boldsymbol{x}\in\mathcal{X}} |m^*(\boldsymbol{x}) - \boldsymbol{\beta}^{*\top}\boldsymbol{B}(\boldsymbol{x})| = O_p(K/n^{1/2} + K^{1/2-r_b}) = o_p(1).$$

For notational simplicity, we denote $\|f\|_\infty = \sup_{\boldsymbol{x}\in\mathcal{X}}|f(\boldsymbol{x})|$. Define the set of functions $\mathcal{F} = \{f_0 : \|m - m^*\|_\infty \le \delta\}$, where $\delta = C(K/n^{1/2} + K^{1/2-r_b})$ for some constant $C > 0$. By the strong ignorability of the treatment assignment, we have that $\mathbb{P}f_0(T, Y(1), \boldsymbol{X}) = 0$. By the Markov inequality and the maximal inequality in Corollary 19.35 of Van der Vaart (2000),

$$n^{1/2}R_0 \le \sup_{f_0\in\mathcal{F}} \mathbb{G}_n(f_0) \lesssim \mathbb{E}\sup_{f_0\in\mathcal{F}} \mathbb{G}_n(f_0) \lesssim J_{[\,]}(\|F_0\|_{P,2}, \mathcal{F}, L_2(P)),$$

where $J_{[\,]}(\|F_0\|_{P,2}, \mathcal{F}, L_2(P))$ is the bracketing integral, and F_0 is the envelop function. Since J is bounded away from 0, we have $|f_0(T, Y(1), \boldsymbol{X})| \lesssim \delta|Y(1) - K(\boldsymbol{X}) - L(\boldsymbol{X})| := F_0$. Then $\|F_0\|_{P,2} \le \delta\{\mathbb{E}|Y(1)|^2\}^{1/2} \lesssim \delta$. Next, we consider $N_{[\,]}(\epsilon, \mathcal{F}, L_2(P))$. Define $\mathcal{F}_0 = \{f_0 : \|m - m^*\|_\infty \le C\}$ for some constant $C > 0$. Thus, it is easily seen that $\log N_{[\,]}(\epsilon, \mathcal{F}, L_2(P)) \lesssim \log N_{[\,]}(\epsilon, \mathcal{F}_0\delta, L_2(P)) = \log N_{[\,]}(\epsilon/\delta, \mathcal{F}_0, L_2(P)) \lesssim \log N_{[\,]}(\epsilon/\delta, \mathcal{M}, L_2(P)) \lesssim (\delta/\epsilon)^{1/k_1}$, where we use the fact that J is bounded away from 0 and J is Lipschitz. The last step follows from the assumption on the bracketing number of $\mathcal{M}$. Then

$$J_{[\,]}(\|F_0\|_{P,2}, \mathcal{F}, L_2(P)) \lesssim \int_0^\delta \sqrt{\log N_{[\,]}(\epsilon, \mathcal{F}, L_2(P))}d\epsilon \lesssim \int_0^\delta (\delta/\epsilon)^{1/(2k_1)}d\epsilon,$$

which goes to 0, as $\delta \to 0$, because $2k_1 > 1$ by assumption and thus the integral converges. Thus, this shows that $n^{1/2}R_0 = o_p(1)$. By the similar argument, we can show that $n^{1/2}R_1 = o_p(1)$.

Next, we consider R_2. Define the following empirical process $\mathbb{G}_n(f_2) = n^{1/2}(\mathbb{P}_n - \mathbb{P})f_2(T, \boldsymbol{X})$, where

$$f_2(T, \boldsymbol{X}) = \frac{T - J(m(\boldsymbol{X}))}{J(m(\boldsymbol{X}))(1 - J(m(\boldsymbol{X})))}\Delta_K(\boldsymbol{X}).$$

By the assumption on the approximation property of the basis functions, we have $\|\Delta_K\|_\infty \lesssim K^{-r_h}$. In addition,

$$\|J(\widetilde{\boldsymbol{\beta}}^\top \boldsymbol{B}(\boldsymbol{X})) - \pi^*(\boldsymbol{X})\|_{P,2} \le \|J(\widetilde{\boldsymbol{\beta}}^\top \boldsymbol{B}(\boldsymbol{X})) - J(\boldsymbol{\beta}^{*\top}\boldsymbol{B}(\boldsymbol{X}))\|_{P,2} + \|J(\boldsymbol{\beta}^{*\top}\boldsymbol{B}(\boldsymbol{X})) - \pi^*(\boldsymbol{X})\|_{P,2} \\ \lesssim \|\widetilde{\boldsymbol{\beta}}^\top \boldsymbol{B}(\boldsymbol{X}) - \boldsymbol{\beta}^{*\top}\boldsymbol{B}(\boldsymbol{X})\|_{P,2} + \sup_{\boldsymbol{x}\in\mathcal{X}} |m^*(\boldsymbol{x}) - \boldsymbol{\beta}^{*\top}\boldsymbol{B}(\boldsymbol{x})| \\ = O_p(K^{1/2}/n^{1/2} + K^{-r_b}),$$

where the last step follows from Lemma F.7.

Define the set of functions $\mathcal{F} = \{f_2 : \|m - m^*\|_{P,2} \le \delta_1, \|\Delta\|_\infty \le \delta_2\}$, where $\delta_1 = C(K^{1/2}/n^{1/2} + K^{-r_b})$ and $\delta_2 = CK^{-r_h}$ for some constant $C > 0$. Thus,

$$n^{1/2} R_2 \le \sup_{f_2 \in \mathcal{F}} \mathbb{G}_n(f_2) + n^{1/2} \sup_{f_2 \in \mathcal{F}} \mathbb{P} f_2.$$

We first consider the second term $n^{1/2} \sup_{f_2 \in \mathcal{F}} \mathbb{P} f_2$. Let $\mathcal{G}_1 = \{m \in \mathcal{M} : \|m - m^*\|_{P,2} \le \delta_1\}$ and $\mathcal{G}_2 = \{\Delta \in \mathcal{H} - \boldsymbol{\alpha}_1^{*\top} \boldsymbol{h}_1 : \|\Delta\|_\infty \le \delta_2\}$. By the definition of the propensity score and Cauchy inequality,

$$\begin{aligned} n^{1/2} \sup_{f_2 \in \mathcal{F}} \mathbb{P} f_2 &= n^{1/2} \sup_{m \in \mathcal{G}_1, \Delta \in \mathcal{G}_2} \mathbb{E} \frac{\pi^*(\boldsymbol{X}) - J(m(\boldsymbol{X}))}{J(m(\boldsymbol{X}))(1 - J(m(\boldsymbol{X})))} \Delta(\boldsymbol{X}) \\ &\lesssim n^{1/2} \sup_{m \in \mathcal{G}_1} \|\pi^* - J(m)\|_{P,2} \sup_{\Delta \in \mathcal{G}_2} \|\Delta\|_{P,2} \\ &\lesssim n^{1/2} \delta_1 \delta_2 \lesssim n^{1/2} (K^{1/2}/n^{1/2} + K^{-r_b}) K^{-r_h} = o(1), \end{aligned}$$

where the last step follows from $r_h > 1/2$ and the scaling assumption $n^{1/2} \lesssim K^{r_b + r_h}$ in this theorem. Next, we need to control the maximum of the empirical process $\sup_{f_2 \in \mathcal{F}} \mathbb{G}_n(f_2)$. Following the similar argument to that for R_0, we only need to upper bound the bracketing integral $J_{[\,]}(\|F_2\|_{P,2}, \mathcal{F}, L_2(P))$. Since J is bounded away from 0 and 1, we can set the envelop function to be $F_2 := C\delta_2$ for some constant $C > 0$ and thus $\|F_2\|_{P,2} \lesssim \delta_2$. Define $\mathcal{F}_0 = \{f_2 : \|m - m^*\|_{P,2} \le C, \|\Delta\|_{P,2} \le 1\}$ for some constant $C > 0$, $\mathcal{G}_{10} = \{m \in \mathcal{M} + m^* : \|m\|_{P,2} \le C\}$ and $\mathcal{G}_{20} = \{\Delta \in \mathcal{H} - \boldsymbol{\alpha}_1^{*\top} \boldsymbol{h}_1 : \|\Delta\|_{P,2} \le 1\}$. Similarly, we have

$$\begin{aligned} \log N_{[\,]}(\epsilon, \mathcal{F}, L_2(P)) &\lesssim \log N_{[\,]}(\epsilon/\delta_2, \mathcal{F}_0, L_2(P)) \\ &\lesssim \log N_{[\,]}(\epsilon/\delta_2, \mathcal{G}_{10}, L_2(P)) + \log N_{[\,]}(\epsilon/\delta_2, \mathcal{G}_{20}, L_2(P)) \\ &\lesssim \log N_{[\,]}(\epsilon/\delta_2, \mathcal{M}, L_2(P)) + \log N_{[\,]}(\epsilon/\delta_2, \mathcal{H}, L_2(P)) \\ &\lesssim (\delta_2/\epsilon)^{1/k_1} + (\delta_2/\epsilon)^{1/k_2}, \end{aligned}$$

where the second step follows from the boundness assumption on J and its Lipschitz property, the third step is due to $\mathcal{G}_{10} - m^* \subset \mathcal{M}$ and $\mathcal{G}_{20} + \boldsymbol{\alpha}_1^{*\top} \boldsymbol{h}_1 \subset \mathcal{H}$ and the last step is by the bracketing number condition in our assumption. Since $2k_1 > 1$ and $2k_2 > 1$, it is easily seen that the bracketing integral $J_{[\,]}(\|F_2\|_{P,2}, \mathcal{F}, L_2(P)) = o(1)$. This shows that $\sup_{f_2 \in \mathcal{F}} \mathbb{G}_n(f_2) = o_p(1)$. Thus, we conclude that $n^{1/2} R_2 = o_p(1)$. By the similar argument, we can show that $n^{1/2} R_3 = o_p(1)$. This completes the whole proof. □

G Discussion on the Results in Section 4

Under the conditions in Theorem 4.1, it is well known that the convergence rate for estimating $K(\boldsymbol{x})$ (and also $L(\boldsymbol{x})$, $\psi^*(\boldsymbol{x})$) in the $L_2(P)$ norm (i.e, $\int(\widehat{K}(\boldsymbol{x}) - K(\boldsymbol{x}))^2 P(d\boldsymbol{x})$) is $O_p(\kappa^{-2r_h} + \kappa/n)$; see Newey (1997). Thus, the optimal choice of κ that minimizes the rate is $\kappa \asymp n^{1/(2r_h+1)}$. Assume that $r_b = r_h$. With $\kappa \asymp n^{1/(2r_h+1)}$, the conditions $\kappa = o(n^{1/3})$ and $n^{\frac{1}{2(r_b+r_h)}} = o(\kappa)$ always hold as long as $r_h > 1$. Recall that from the previous discussion $r_h = s/d$, where s is the smoothness parameter and d is the dimension of $\boldsymbol{X}$. Thus under very mild conditions $s > d$, we do not need to under-smooth the estimator.

Remark G.1. By the proof of Theorem 4.1, we find that when $\kappa = o(n^{1/(2r_b+1)})$ and $\kappa = o(n^{1/(2r_h+1)})$ hold, the asymptotic bias of the estimator $\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}}$ is of order $O_p(K^{-(r_b+r_h)})$, which is the product of the approximation errors for $\psi^*(\boldsymbol{x})$ and $K(\boldsymbol{x})$ (also $L(\boldsymbol{x})$). Thus, to make the bias of the estimator $\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}}$ asymptotically ignorable, we can require either r_b or r_h sufficiently large (not necessarily both). This phenomenon can be viewed as the double robustness property in the nonparametric context, which holds for the kernel based doubly robust estimator (Rothe and Firpo, 2013) and the targeted maximum likelihood estimator (Benkeser et al., 2017). In addition, our estimator has smaller asymptotic bias than the usual nonparametric method. For simplicity, assume $r_b = r_h = r$. The asymptotic bias of the IPTW estimator in Hirano et al. (2003) is generally of order $O_p(\kappa^{-r})$, whereas our estimator has a smaller bias of order $O_p(\kappa^{-2r})$.

H Estimation of ATT

We consider the estimation of the average treatment effect for the treated (ATT)

$$\tau^* = \mathbb{E}(Y_i(1) - Y_i(0)|T_i = 1).$$

Let $\tau_1^* = \mathbb{E}(Y_i(1) \mid T_i = 1)$ and $\tau_0^* = \mathbb{E}(Y_i(0) \mid T_i = 1)$. By the law of total probability,

$$\tau_1^* = \mathbb{E}(T_i Y_i(1) \mid T_i = 1) = \mathbb{E}(T_i Y_i(1))/\mathbb{P}(T_i = 1).$$

Thus, a simple estimator of τ_1^* is

$$\widehat{\tau}_1 = \frac{\sum_{i=1}^n T_i Y_i}{\sum_{i=1}^n T_i}.$$

To estimate τ_0^*, we notice that

$$\begin{aligned}
\tau_0^* &= \mathbb{E}[\mathbb{E}(Y_i(0) \mid T_i = 1, \boldsymbol{X}_i) \mid T_i = 1] = \mathbb{E}[\mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i) \mid T_i = 1] \\
&= \frac{\mathbb{E}[T_i \mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i)]}{\mathbb{P}(T_i = 1)} = \frac{\mathbb{E}(\pi(\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i)\mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i))}{\mathbb{P}(T_i = 1)} \\
&= \frac{1}{\mathbb{P}(T_i = 1)} \mathbb{E}\left\{ \frac{\pi(\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i)(1 - T_i)Y_i(0)}{1 - \pi(\boldsymbol{\beta}^{*\top}\boldsymbol{X}_i)} \right\}.
\end{aligned}$$

Similar to the bias and variance calculation for the ATE, we can estimate $\boldsymbol{\beta}$ by the solving the following estimating equations

$$n^{-1} \sum_{i=1}^n \left(T_i - \frac{(1 - T_i)\pi(\boldsymbol{\beta}^\top \boldsymbol{X})}{1 - \pi(\boldsymbol{\beta}^\top \boldsymbol{X})} \right) \mathbf{f}(\boldsymbol{X}) = 0.$$

Then, we set $\widehat{\pi}_i = \pi(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i)$ and estimate τ_0 by

$$\widehat{\tau}_0 = \frac{\sum_{i=1}^n (1 - T_i)\widehat{r}_i Y_i}{\sum_{i=1}^n (1 - T_i)\widehat{r}_i},$$

where $\widehat{r}_i = \widehat{\pi}_i/(1 - \widehat{\pi}_i)$. The final estimator of the ATT is $\widehat{\tau} = \widehat{\tau}_1 - \widehat{\tau}_0$. Similar to the proof of the main results on ATE, we can show that when both models are correct, $n^{1/2}(\widehat{\tau} - \tau^*) \to_d N(0, W)$, where

$$W \;=\; p^{-2}\mathbb{E}\left[\pi^* \mathbb{E}(\epsilon_1^2 \mid \boldsymbol{X}) + \frac{\pi^{*2}}{1 - \pi_i^*} \mathbb{E}(\epsilon_0^2 \mid \boldsymbol{X}) + \pi^*(L(\boldsymbol{X}_i) - \tau^*)^2 \right].$$

Here, $\epsilon_0 = Y(0) - K(\boldsymbol{X})$, $\epsilon_1 = Y(1) - K(\boldsymbol{X}) - L(\boldsymbol{X})$ and $p = \mathbb{P}(Y = 1)$

I Derivation of (3.10) and (3.11)

In this appendix, we only provide a sketch of the proof of (3.10) and (3.11), because the detail is very similar to the proof of Theorem 2.1. Recall that as in Section 2, $\boldsymbol{\beta}^o$ which satisfies $\mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^o}(\boldsymbol{T}, \boldsymbol{X})) = 0$ is the limiting value of $\widehat{\boldsymbol{\beta}}$ as in Lemma B.2. In addition, denote $K^o(\boldsymbol{X}_i) = \boldsymbol{\alpha}^{*T}\boldsymbol{h}_1(\boldsymbol{X}_i) + \delta \boldsymbol{A}_1 \boldsymbol{h}_1(\boldsymbol{X}_i)$ and $L^o(\boldsymbol{X}_i) = \boldsymbol{\gamma}^{*T}\boldsymbol{h}_2(\boldsymbol{X}_i) + \delta \boldsymbol{A}_2 \boldsymbol{h}_2(\boldsymbol{X}_i)$, where the vectors $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are to be determined. We have the following decomposition

$$\begin{aligned}
\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu &= \frac{1}{n} \sum_{i=1}^n \Big[\frac{T_i}{\pi_{\beta^o}(\boldsymbol{X}_i)} \{Y_i(1) - K^o(\boldsymbol{X}_i) - L^o(\boldsymbol{X}_i)\} - \frac{1 - T_i}{1 - \pi_{\beta^o}(\boldsymbol{X}_i)} \{Y_i(0) - K^o(\boldsymbol{X}_i)\} + L^o(\boldsymbol{X}_i) - \mu \Big] \\
&\quad + \frac{1}{n} \sum_{i=1}^n \left\{ \frac{T_i}{\pi_{\widehat{\beta}}(\boldsymbol{X}_i)} - \frac{T_i}{\pi_{\beta^o}(\boldsymbol{X}_i)} \right\} \{Y_i(1) - K^o(\boldsymbol{X}_i) - L^o(\boldsymbol{X}_i)\} \\
&\quad - \frac{1}{n} \sum_{i=1}^n \left\{ \frac{1 - T_i}{1 - \pi_{\widehat{\beta}}(\boldsymbol{X}_i)} - \frac{1 - T_i}{1 - \pi_{\beta^o}(\boldsymbol{X}_i)} \right\} \{Y_i(0) - K^o(\boldsymbol{X}_i)\} := I_1 + I_2 + I_3.
\end{aligned}$$

We first consider I_2. The mean value theorem implies

$$I_2 = -\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\pi_{\widetilde{\beta}}^2(\boldsymbol{X}_i)}\frac{\partial\pi_{\widetilde{\beta}}(\boldsymbol{X}_i)}{\partial\beta}\{Y_i(1) - K^o(\boldsymbol{X}_i) - L^o(\boldsymbol{X}_i)\}(\widehat{\beta} - \beta^o),$$

where $\widetilde{\beta}$ is an intermediate value between $\widehat{\beta}$ and β^o. Under Assumptions similar to B.1, the dominated convergence theorem implies

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\pi_{\widetilde{\beta}}^2(\boldsymbol{X}_i)}\frac{\partial\pi_{\widetilde{\beta}}(\boldsymbol{X}_i)}{\partial\beta}\{Y_i(1) - K^o(\boldsymbol{X}_i) - L^o(\boldsymbol{X}_i)\} = O_p(\delta).$$

Similar to Lemma B.3, we can show that $\widehat{\beta} - \beta^o = O_p(n^{-1/2})$. The Slutsky theorem yields $I_2 = O_p(\delta n^{-1/2})$. The same argument implies that $I_3 = O_p(\delta n^{-1/2})$. Finally, we focus on I_1. Note that

$$I_1 - \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{T_i}{\pi(\boldsymbol{X}_i)}\{Y_i(1) - K(\boldsymbol{X}_i) - L(\boldsymbol{X}_i)\} - \frac{1-T_i}{1-\pi(\boldsymbol{X}_i)}\{Y_i(0) - K(\boldsymbol{X}_i)\} + L(\boldsymbol{X}_i) - \mu\Big] = \frac{1}{n}\sum_{i=1}^{n}\Delta_i,$$

where

$$\begin{aligned}\Delta_i = &\Big\{\frac{T_i}{\pi_{\beta^o}(\boldsymbol{X}_i)} - \frac{T_i}{\pi(\boldsymbol{X}_i)}\Big\}\{Y_i(1) - K^o(\boldsymbol{X}_i) - L^o(\boldsymbol{X}_i)\}\\ &- \Big\{\frac{1-T_i}{1-\pi_{\beta^o}(\boldsymbol{X}_i)} - \frac{1-T_i}{1-\pi(\boldsymbol{X}_i)}\Big\}\{Y_i(0) - K^o(\boldsymbol{X}_i)\}\\ &- \frac{T_i}{\pi(\boldsymbol{X}_i)}\{K^o(\boldsymbol{X}_i) + L^o(\boldsymbol{X}_i) - K(\boldsymbol{X}_i) - L(\boldsymbol{X}_i)\}\\ &+ \frac{1-T_i}{1-\pi(\boldsymbol{X}_i)}\{K^o(\boldsymbol{X}_i) - K(\boldsymbol{X}_i)\} + L^o(\boldsymbol{X}_i) - L(\boldsymbol{X}_i).\end{aligned}$$

The central limit theorem implies $n^{1/2}(\frac{1}{n}\sum_{i=1}^{n}\Delta_i - \mathbb{E}\Delta_i)/sd(\Delta_i) \to N(0,1)$. In order to derive the order of $\frac{1}{n}\sum_{i=1}^{n}\Delta_i$, it suffices to compute the $\mathbb{E}(\Delta_i)$ and $sd(\Delta_i)$. As in the derivation of (C.1), after some algebra, we similarly obtain

$$\beta^o - \beta^* = \xi\mathbf{T}^{-1}\mathbf{M} + O(\xi^2),$$

where

$$\mathbf{M} = \begin{pmatrix}\mathbb{E}(\frac{1}{1-\pi_{\beta^*}(\boldsymbol{X}_i)}u_i^*\boldsymbol{h}_1(\boldsymbol{X}_i))\\ \mathbb{E}(u_i^*\boldsymbol{h}_2(\boldsymbol{X}_i))\end{pmatrix}$$

and $\mathbf{T} = [\mathbb{E}(\frac{1}{\pi_{\beta^*}(\boldsymbol{X}_i)(1-\pi_{\beta^*}(\boldsymbol{X}_i))}\frac{\partial\pi_{\beta^*}(\boldsymbol{X}_i)}{\partial\beta}\boldsymbol{h}_1^T(\boldsymbol{X}_i)), \mathbb{E}(\frac{1}{\pi_{\beta^*}(\boldsymbol{X}_i)}\frac{\partial\pi_{\beta^*}(\boldsymbol{X}_i)}{\partial\beta}\boldsymbol{h}_2^T(\boldsymbol{X}_i))]^T$.

Denote $\widetilde{r}_1(\boldsymbol{X}_i) = r_1(\boldsymbol{X}_i) - \boldsymbol{A}_1\boldsymbol{h}_1(\boldsymbol{X}_i)$ and $\widetilde{r}_2(\boldsymbol{X}_i) = r_2(\boldsymbol{X}_i) - \boldsymbol{A}_2\boldsymbol{h}_2(\boldsymbol{X}_i)$. Note that

$$\begin{aligned}
\mathbb{E}(\Delta_i) &= \mathbb{E}\Big\{\frac{\pi(\boldsymbol{X}_i)}{\pi_{\beta^o}(\boldsymbol{X}_i)}\delta(\widetilde{r}_1(\boldsymbol{X}_i) + \widetilde{r}_2(\boldsymbol{X}_i)) - \frac{1-\pi(\boldsymbol{X}_i)}{1-\pi_{\beta^o}(\boldsymbol{X}_i)}\delta\widetilde{r}_1(\boldsymbol{X}_i) - \delta\widetilde{r}_2(\boldsymbol{X}_i)\Big\} \\
&= \mathbb{E}\Big\{\{1 + \xi u_i^* - \frac{1}{\pi_{\beta^*}(\boldsymbol{X}_i)}\frac{\partial \pi_{\beta^*}(\boldsymbol{X}_i)}{\partial \beta}(\beta^o - \beta^*)\}\delta(\widetilde{r}_1(\boldsymbol{X}_i) + \widetilde{r}_2(\boldsymbol{X}_i)) \\
&\quad - \{1 - \frac{\pi_{\beta^*}(\boldsymbol{X}_i)}{1-\pi_{\beta^*}(\boldsymbol{X}_i)}\xi u_i^* + \frac{1}{1-\pi_{\beta^*}(\boldsymbol{X}_i)}\frac{\partial \pi_{\beta^*}(\boldsymbol{X}_i)}{\partial \beta}(\beta^o - \beta^*)\}\delta\widetilde{r}_1(\boldsymbol{X}_i) - \delta\widetilde{r}_2(\boldsymbol{X}_i))\Big\} + O(\xi^2\delta) \\
&= \xi\delta\mathbb{E}\Big[\{\frac{1}{1-\pi_{\beta^*}(\boldsymbol{X}_i)}u_i^* - \frac{1}{(1-\pi_{\beta^*}(\boldsymbol{X}_i))\pi_{\beta^*}(\boldsymbol{X}_i)}\frac{\partial \pi_{\beta^*}(\boldsymbol{X}_i)}{\partial \beta}\mathbf{T}^{-1}\mathbf{M}\}\widetilde{r}_1(\boldsymbol{X}_i)\Big] \\
&\quad + \xi\delta\mathbb{E}\Big[\{u_i^* - \frac{1}{\pi_{\beta^*}(\boldsymbol{X}_i)}\frac{\partial \pi_{\beta^*}(\boldsymbol{X}_i)}{\partial \beta}\mathbf{T}^{-1}\mathbf{M}\}\widetilde{r}_2(\boldsymbol{X}_i)\Big] + O(\xi^2\delta).
\end{aligned}$$

Assume that at least one entry of $\mathbb{E}\Big[\{\frac{1}{1-\pi_{\beta^*}(\boldsymbol{X}_i)}u_i^* - \frac{1}{(1-\pi_{\beta^*}(\boldsymbol{X}_i))\pi_{\beta^*}(\boldsymbol{X}_i)}\frac{\partial \pi_{\beta^*}(\boldsymbol{X}_i)}{\partial \beta}\mathbf{T}^{-1}\mathbf{M}\}\boldsymbol{h}_1(\boldsymbol{X}_i)\Big]$ is nonzero. Then, there exists $\boldsymbol{A}_1$ such that

$$\begin{aligned}
&\boldsymbol{A}_1\mathbb{E}\Big[\{\frac{1}{1-\pi_{\beta^*}(\boldsymbol{X}_i)}u_i^* - \frac{1}{(1-\pi_{\beta^*}(\boldsymbol{X}_i))\pi_{\beta^*}(\boldsymbol{X}_i)}\frac{\partial \pi_{\beta^*}(\boldsymbol{X}_i)}{\partial \beta}\mathbf{T}^{-1}\mathbf{M}\}\boldsymbol{h}_1(\boldsymbol{X}_i)\Big] \\
&= \mathbb{E}\Big[\{\frac{1}{1-\pi_{\beta^*}(\boldsymbol{X}_i)}u_i^* - \frac{1}{(1-\pi_{\beta^*}(\boldsymbol{X}_i))\pi_{\beta^*}(\boldsymbol{X}_i)}\frac{\partial \pi_{\beta^*}(\boldsymbol{X}_i)}{\partial \beta}\mathbf{T}^{-1}\mathbf{M}\}r_1(\boldsymbol{X}_i)\Big],
\end{aligned}$$

which implies

$$\mathbb{E}\Big[\{\frac{1}{1-\pi_{\beta^*}(\boldsymbol{X}_i)}u_i^* - \frac{1}{(1-\pi_{\beta^*}(\boldsymbol{X}_i))\pi_{\beta^*}(\boldsymbol{X}_i)}\frac{\partial \pi_{\beta^*}(\boldsymbol{X}_i)}{\partial \beta}\mathbf{T}^{-1}\mathbf{M}\}\widetilde{r}_1(\boldsymbol{X}_i)\Big] = 0.$$

Similarly, by choosing a proper $\boldsymbol{A}_2$, we have

$$\mathbb{E}\Big[\{u_i^* - \frac{1}{\pi_{\beta^*}(\boldsymbol{X}_i)}\frac{\partial \pi_{\beta^*}(\boldsymbol{X}_i)}{\partial \beta}\mathbf{T}^{-1}\mathbf{M}\}\widetilde{r}_2(\boldsymbol{X}_i)\Big] = 0.$$

As a result, we obtain $\mathbb{E}(\Delta_i) = O(\xi^2\delta)$. Finally, after some tedious calculation, we can show that $sd(\Delta_i) = O(\xi + \delta)$. This implies $\frac{1}{n}\sum_{i=1}^n \Delta_i = O_p(\xi^2\delta + \xi n^{-1/2} + \delta n^{-1/2})$. This completes the proof of (3.10). The proof of (3.11) follows from the similar argument and we omit the details.

J Asymptotic Variance Formulas Used for Simulations

In this appendix, we present the asymptotic variance formulas used for constructing the 95% confidence intervals for calculating the coverage probabilities in the simulations in Section 5.1. In particular, for a generic estimator $\widehat{\mu}$, the 95% confidence interval is $(\widehat{\mu} - 1.96 * \widehat{\sigma}, \widehat{\mu} + 1.96 * \widehat{\sigma})$, where $\widehat{\sigma}^2$ is the estimate of the asymptotic variance of $\sqrt{n}(\widehat{\mu} - \mu)$.

For the True estimator, the asymptotic variance formula is similar to the one given in Section 2 and is as follows:

$$\Sigma_{\mu_0} = \mathrm{Var}\big(\mu_{\boldsymbol{\beta}_0}(T_i, Y_i, \boldsymbol{X}_i)\big) \;=\; \mathbb{E}\bigg(\tfrac{Y_i(1)^2}{\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X}_i)} + \tfrac{Y_i(0)^2}{1-\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X}_i)} - (\mathbb{E}(Y_i(1)) - \mathbb{E}(Y_i(0)))^2\bigg).$$

For the GLM estimator, the asymptotic variance formula is as follows:

$$\Sigma_{\mathrm{GLM}} = \Sigma_{\mu_0} - \boldsymbol{H}_y^\top \boldsymbol{I}^{-1} \boldsymbol{H}_y$$

where Σ_{μ_0} is defined like before, $\boldsymbol{I}$ is the Fisher Information Matrix, and

$$\boldsymbol{H}_y \;=\; -\mathbb{E}\left(\frac{K(\boldsymbol{X}_i) + (1-\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X}_i))L(\boldsymbol{X}_i)}{\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X}_i)(1-\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X}_i))} \cdot \frac{\partial \pi_{\boldsymbol{\beta}_0}(\boldsymbol{X}_i)}{\partial \boldsymbol{\beta}}\right).$$

Since the second term is positive definite, $\Sigma_{GLM} < \Sigma_\mu$ and thus the variance decreases.

The GAM estimator achieves the semiparametric efficiency bound (Hirano et al., 2003) and so we can use V_{opt} given in (2.6) as the asymptotic variance formula. The CBPS estimator has the following asymptotic variance formula:

$$\begin{aligned}\Sigma_{\mathrm{CBPS}} \;=\; \Sigma_{\mu_0} \;&+\; \boldsymbol{H}_y^\top (\boldsymbol{H}_{\mathbf{f}}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{H}_{\mathbf{f}})^{-1} \boldsymbol{H}_y \\ &-\; 2\boldsymbol{H}_y^\top (\boldsymbol{H}_{\mathbf{f}}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{H}_{\mathbf{f}})^{-1} \boldsymbol{H}_{\mathbf{f}}^\top \boldsymbol{\Omega}^{-1}\, \mathrm{Cov}(\mu_{\boldsymbol{\beta}_0}(T_i, Y_i, \boldsymbol{X}_i), \boldsymbol{g}_{\boldsymbol{\beta}_0}(T_i, \boldsymbol{X}_i))\end{aligned}$$

where Σ_{μ_0} and $\boldsymbol{H}_y$ are defined like before, and we have:

$$\begin{aligned}\boldsymbol{H}_{\mathbf{f}} &= -\mathbb{E}\left(\frac{\mathbf{f}(\boldsymbol{X}_i)}{\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X}_i)(1-\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X}_i))}\left(\frac{\partial \pi_{\boldsymbol{\beta}_0}(\boldsymbol{X}_i)}{\partial \boldsymbol{\beta}}\right)^\top\right) \\ \boldsymbol{\Omega} &= \mathrm{Var}(\boldsymbol{g}_{\boldsymbol{\beta}_0}(T_i, \boldsymbol{X}_i)) \\ \boldsymbol{g}_{\boldsymbol{\beta}_0}(T_i, \boldsymbol{X}_i) &= \left(\frac{T_i}{\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X}_i)} - \frac{1-T_i}{1-\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X}_i)}\right)\mathbf{f}(\boldsymbol{X}_i) \\ \mu_{\boldsymbol{\beta}_0}(T_i, Y_i, \boldsymbol{X}_i) &= \frac{T_i Y_i}{\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X}_i)} - \frac{(1-T_i)Y_i}{1-\pi_{\boldsymbol{\beta}_0}(\boldsymbol{X}_i)}.\end{aligned}$$

The asymptotic variance for the DR estimator is automatically computed in the R package drtmle and the confidence interval was constructed accordingly.

Finally, we note that when we estimate the asymptotic variances, we simply replace the quantities π_{β_0} and $K(\boldsymbol{X})$ and $L(\boldsymbol{X})$ with their estimates and replace the expectation with the sample average. To save space, we do not repeat the formulas of the estimated variances.