Optimal Covariate Balancing Conditions in Propensity Score Estimation*

Jianqing Fan† Kosuke Imai‡ Daniel Lee§ Han Liu¶ Yang Ning‖
Xiaolin Yang**

August 2, 2021

Abstract

Inverse probability of treatment weighting (IPTW) is a popular method for estimating the average treatment effect (ATE). However, empirical studies show that the IPTW estimators can be sensitive to the misspecification of the propensity score model. To address this problem, researchers have proposed to estimate propensity score by directly optimizing the balance of pre-treatment covariates. While these methods appear to empirically perform well, little is known about how the choice of balancing conditions affects their theoretical properties. To fill this gap, we first characterize the asymptotic bias and efficiency of the IPTW estimator based on the Covariate Balancing Propensity Score (CBPS) methodology under local model misspecification. Based on this analysis, we show how to optimally choose the covariate balancing functions and propose an optimal CBPS-based IPTW estimator. This estimator is doubly robust; it is consistent for the ATE if either the propensity score model or the outcome model is correct. In addition, the proposed estimator is locally semiparametric efficient when both models are

*Supported by NSF grants DMS-1854637, DMS-1712591, DMS-2053832, CAREER Award DMS-1941945, NIH grant R01-GM072611, and Sloan Grant 2020-13946. An earlier version of this paper was entitled, "Improving Covariate Balancing Propensity Score: A Doubly Robust and Efficient Approach."

†Department of Operations Research and Financial Engineering, Princeton University

‡Department of Government and Department of Statistics, Harvard University

§Department of Statistics and Data Science, Cornell University

¶Department of Electrical Engineering and Computer Science, Northwestern University.

‖Department of Statistics and Data Science, Cornell University

**Amazon

correctly specified. To further relax the parametric assumptions, we extend our method by using a sieve estimation approach. We show that the resulting estimator is globally efficient under a set of much weaker assumptions and has a smaller asymptotic bias than the existing estimators. Finally, we evaluate the finite sample performance of the proposed estimators via simulation and empirical studies. An open-source software package is available for implementing the proposed methods.

Key words: Average treatment effect, causal inference, double robustness, model misspecification, semiparametric efficiency, sieve estimation

1 Introduction

Suppose that we have a random sample of n units from a population of interest. For each unit i, we observe $(T_i, Y_i, \boldsymbol{X}_i)$, where $\boldsymbol{X}_i \in \mathbb{R}^d$ is a d-dimensional vector of pre-treatment covariates, T_i is a binary treatment variable, and Y_i is an outcome variable. In particular, T_i takes 1 if unit i receives the treatment and is equal to 0 if unit i belongs to the control group. The observed outcome can be written as $Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$, where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes under the treatment and control conditions, respectively. This notation implicitly requires the stable unit treatment value assumption (Rubin, 1990). In addition, throughout this paper, we assume the strong ignorability of the treatment assignment (Rosenbaum and Rubin, 1983),

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid \boldsymbol{X}_i \quad \text{and} \quad 0 < \mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) < 1. \tag{1.1}$$

Next, we assume that the conditional mean functions of potential outcomes exist and denote them by,

$$\mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i) = K(\boldsymbol{X}_i) \quad \text{and} \quad \mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i) = K(\boldsymbol{X}_i) + L(\boldsymbol{X}_i), \tag{1.2}$$

for some functions $K(\cdot)$ and $L(\cdot)$, which represent the conditional mean of the potential outcome under the control condition and the conditional average treatment effect, respectively. Under this setting, we are interested in estimating the average treatment effect (ATE),

$$\mu = \mathbb{E}(Y_i(1) - Y_i(0)) = \mathbb{E}(L(\boldsymbol{X}_i)). \tag{1.3}$$

The propensity score is defined as the conditional probability of treatment assignment (Rosenbaum and Rubin, 1983),

$$\pi(\boldsymbol{X}_i) \quad = \quad \mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i). \tag{1.4}$$

In practice, since $\boldsymbol{X}_i$ can be high dimensional, the propensity score is usually parameterized by a model $\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)$ where $\boldsymbol{\beta}$ is a q-dimensional vector of parameters. A popular choice is the logistic regression model, i.e., $\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i) = \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta})/\{1 + \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta})\}$. Once the parameter $\boldsymbol{\beta}$ is estimated (e.g., by the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$), the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), which is based on the inverse probability of treatment weighting (IPTW), can be used to obtain an estimate of the ATE,

$$\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \left(\frac{T_i Y_i}{\pi_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)} - \frac{(1 - T_i) Y_i}{1 - \pi_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)} \right). \tag{1.5}$$

However, it has been shown that the IPTW estimator with the known propensity score does not attain the semiparametric efficiency bound (Hahn, 1998). A variety of efficient ATE estimators have been proposed (see e.g., Robins et al., 1994; Bang and Robins, 2005; Tan, 2006; Qin and Zhang, 2007; Robins et al., 2007; Cao et al., 2009; Tan, 2010; van der Laan, 2010; Rotnitzky et al., 2012; Han and Wang, 2013; Vermeulen and Vansteelandt, 2015, among many others). Despite the popularity of these methods, researchers have found that in practice the estimators can be sensitive to the misspecification of the propensity score model and the outcome model (e.g., Kang and Schafer, 2007). To overcome this problem, several researchers have recently considered the estimation of the propensity score by optimizing covariate balance rather than maximizing the accuracy of predicting treatment assignment (e.g., Hainmueller, 2012; Graham et al., 2012; Imai and Ratkovic, 2014; Chan et al., 2016; Zubizarreta, 2015; Zhao and Percival, 2017; Zhao, 2019). In this paper, we focus on the Covariate Balancing Propensity Score (CBPS) methodology (Imai and Ratkovic, 2014). In spite of its simplicity, several scholars independently found that the CBPS performs well in practice (e.g., Wyss et al., 2014; Frölich et al., 2015). The method can also be extended for the analysis of longitudinal data (Imai and Ratkovic, 2015), general treatment regimes (Fong et al., 2018a) and high-dimensional propensity score (Ning et al., 2018). In this paper, we conduct a theoretical investigation of the CBPS. Given the similarity between the CBPS and some other methods, our theoretical analysis may also provide new insights for understanding other covariate balancing methods.

The CBPS method estimates the parameters of the propensity score model, $\boldsymbol{\beta}$, by solving the following m-dimensional estimating equation,

$$\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) \;=\; \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_{\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i) \;=\; 0 \quad \text{where} \quad \boldsymbol{g}_{\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i) \;=\; \left(\frac{T_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - \frac{1-T_i}{1-\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}\right) \mathbf{f}(\boldsymbol{X}_i), \tag{1.6}$$

for some covariate balancing function $\mathbf{f}(\cdot) : \mathbb{R}^d \to \mathbb{R}^m$ when the number of equations m is equal to the number of parameters q. Imai and Ratkovic (2014) point out that the common practice of fitting a logistic model is equivalent to balancing the score function with $\mathbf{f}(\boldsymbol{X}_i) = \pi'_{\boldsymbol{\beta}}(\boldsymbol{X}_i) = \partial \pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)/\partial \boldsymbol{\beta}$. They find that choosing $\mathbf{f}(\boldsymbol{X}_i) = \boldsymbol{X}_i$, which balances the first moment between the treatment and control groups, significantly reduces the bias of the estimated ATE. Some researchers also include higher moments and/or interactions, e.g., $\mathbf{f}(\boldsymbol{X}_i) = (\boldsymbol{X}_i \;\; \boldsymbol{X}_i^2)$, in their applications. This guarantees that the treatment and control groups have an identical sample mean of $\mathbf{f}(\boldsymbol{X}_i)$ after weighting by the estimated propensity score.

When $m > q$, then $\widehat{\boldsymbol{\beta}}$ can be estimated by optimizing the covariate balance by the generalized method of moments (GMM) method (Hansen, 1982):

$$\widehat{\boldsymbol{\beta}} \;=\; \underset{\boldsymbol{\beta}\in\Theta}{\operatorname{argmin}} \; \bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})^{\top} \; \widehat{\mathbf{W}} \; \bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}), \tag{1.7}$$

where Θ is the parameter space for $\boldsymbol{\beta}$ in $\mathbb{R}^q$ and $\widehat{\mathbf{W}}$ is an $(m \times m)$ positive definite weighting matrix, which we assume in this paper does not depend on $\boldsymbol{\beta}$. Alternatively, the empirical likelihood method can be used (Owen, 2001). Once the estimate of $\boldsymbol{\beta}$ is obtained, we can estimate the ATE using the IPTW estimator in equation (1.5).

The main idea of the CBPS and other related methods is to directly optimize the balance of covariates between the treatment and control groups so that even when the propensity score model is misspecified we still obtain a reasonable balance of the covariates between the treatment and control groups. However, one open question remains in this literature: How shall we choose the covariate balancing function $\mathbf{f}(\boldsymbol{X}_i)$? In particular, if the propensity score model is misspecified, this problem becomes even more important.

This paper makes two main contributions. First, we conduct a thorough theoretical study of the CBPS-based IPTW estimator with an arbitrary covariate balancing function $\mathbf{f}(\cdot)$. We characterize the asymptotic bias and efficiency of this estimator under locally misspecified propensity score models. Based on these findings, we show how to optimally choose the covariate balancing function

$\mathbf{f}(\boldsymbol{X}_i)$ for the CBPS methodology (Section 2).

However, the optimal choice of $\mathbf{f}(\boldsymbol{X}_i)$ requires some initial estimators for the unknown propensity score model and the outcome models. This limits the application of the CBPS method with the optimal $\mathbf{f}(\boldsymbol{X}_i)$ in practice. Our second contribution is to overcome this problem by developing an optimal CBPS method that does not require an initial estimator. We show that the IPTW estimator based on the optimal CBPS (oCBPS) method retains the double robustness property. The proposed estimator is semiparametrically efficient when both the propensity score and outcome models are correctly specified. More importantly, we show that the rate of convergence of the proposed oCBPS estimator is faster than the augmented inverse probability weighted (AIPW) estimator (Robins et al., 1994) under locally misspecified models. (Section 3).

To relax the parametric assumptions on the propensity score model and the outcome model, we further extend the proposed oCBPS method to the nonparametric settings, by using a sieve estimation approach (Newey, 1997; Chen, 2007). In Section 4, we establish the semiparametric efficiency result for the IPTW estimator under the nonparametric setting. Compared to the existing nonparametric propensity score methods (e.g., Hirano et al., 2003; Chan et al., 2016), our theoretical results require weaker smoothness assumptions. For instance, the theories in Hirano et al. (2003), Imbens et al. (2007) and Chan et al. (2016) require $s/d > 7$, $s/d > 9$ and $s/d > 13$, respectively, where s is the smoothness parameter of the corresponding function class and $d = \dim(\boldsymbol{X}_i)$. In comparison, we only require $s/d > 3/4$, which is significantly weaker than the existing conditions. To prove this result, we exploit the matrix Bernstein's concentration inequalities (Tropp, 2015) and a Bernstein-type concentration inequality for U-statistics (Arcones, 1995). Moreover, we show that our estimator has smaller asymptotic bias than the usual nonparametric method (e.g., Hirano et al., 2003). Therefore, the asymptotic normality result is expected to be more accurate in practice (Section 4). The proof of the theoretical results are deferred to the supplementary material.

An open-source R software package CBPS is available for implementing the proposed estimators (Fong et al., 2018b). In Section 5, we conduct simulation studies to evaluate the performance of the proposed methodology and show that the oCBPS methodology indeed performs better than the standard CBPS methodology in a variety of settings. Finally, we conduct an empirical study using a canonical application in labor economics. We show that the oCBPS method is able to yield estimates closer to the experimental benchmark when compared to the standard CBPS method.

2 CBPS under Locally Misspecified Propensity Score Models

Our theoretical investigation starts by examining the consequences of model misspecification for the CBPS-based IPTW estimator. While researchers can avoid gross model misspecification through careful model fitting, in practice it is often difficult to nail down the exact specification. The prominent simulation study of Kang and Schafer (2007), for example, is designed to illustrate this phenomenon. We therefore consider the consequences of local misspecification of propensity score model in the general framework of Copas and Eguchi (2005). In particular, we assume that the true propensity score $\pi(\boldsymbol{X}_i)$ is related to the working model $\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)$ through the exponential tilt for some $\boldsymbol{\beta}^*$,

$$\pi(\boldsymbol{X}_i) \quad = \quad \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i) \exp(\xi \; u(\boldsymbol{X}_i; \boldsymbol{\beta}^*)), \tag{2.1}$$

where $u(\boldsymbol{X}_i; \boldsymbol{\beta}^*)$ is a function determining the direction of misspecification and $\xi \in \mathbb{R}$ represents the magnitude of misspecification. We assume $\xi = o(1)$ as $n \to \infty$ so that the true propensity score $\pi(\boldsymbol{X}_i)$ is in a local neighborhood of the working model $\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)$. Intuitively, since $\pi(\boldsymbol{X}_i) \approx \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)$ holds, we can interpret $\boldsymbol{\beta}^*$ as the approximate true value of $\boldsymbol{\beta}$. The main advantage of this exponential tilt approach is that $\pi(\boldsymbol{X})$ is always nonnegative, while it does not guarantee $\pi(\boldsymbol{X}) \leq 1$. However, with $\xi = o(1)$ and Assumption B.1 (i.e., $|u(\boldsymbol{X}; \boldsymbol{\beta}^*)| \leq C$ almost surely for some constant $C > 0$), we can show that $\pi(\boldsymbol{X}) \leq 1$ holds with probability tending to 1. Finally, we note that under suitable regularity conditions model (2.1) can be approximated by $\pi(\boldsymbol{X}) = \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}) + \xi \bar{u}(\boldsymbol{X}; \boldsymbol{\beta}^*) + O_p(\xi^2)$, for some $\bar{u}(\boldsymbol{X}; \boldsymbol{\beta}^*)$. This provides an asymptotically equivalent specification of the locally missepecified model. To keep our presentation focused, in this section we assume model (2.1) holds.

In the following, we will establish the asymptotic normality of the CBPS-based IPTW estimator in (1.5) under this local model misspecification framework.

To derive the asymptotic bias and variance, let us define some necessary quantities,

$$\begin{aligned} B \quad = \quad & \left\{ \mathbb{E}\left[\frac{u(\boldsymbol{X}_i; \boldsymbol{\beta}^*)\{K(\boldsymbol{X}_i) + L(\boldsymbol{X}_i)(1 - \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))\}}{1 - \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)} \right] \right. \\ & \left. + \boldsymbol{H}_y^* (\boldsymbol{H}_{\mathbf{f}}^{*\top} \mathbf{W}^* \boldsymbol{H}_{\mathbf{f}}^*)^{-1} \boldsymbol{H}_{\mathbf{f}}^{*\top} \mathbf{W}^* \mathbb{E}\left(\frac{u(\boldsymbol{X}_i; \boldsymbol{\beta}^*) \mathbf{f}(\boldsymbol{X}_i)}{1 - \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)} \right) \right\}, \end{aligned} \tag{2.2}$$

where $K(\boldsymbol{X}_i)$ and $L(\boldsymbol{X}_i)$ are defined in (1.2), $\mathbf{W}^*$ is the limiting value of $\widehat{\mathbf{W}}$ in (1.7), and

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

Furthermore, denote $\mu_{\boldsymbol{\beta}^*}(T_i,Y_i,\boldsymbol{X}_i)=\frac{T_iY_i}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}-\frac{(1-T_i)Y_i}{1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}$,

$$\bar{\boldsymbol{H}}^*=(1,\boldsymbol{H}_y^{*\top}) \quad\text{and}\quad \boldsymbol{\Sigma}=\begin{pmatrix}\Sigma_\mu & \boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}}^\top\\ \boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} & \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\end{pmatrix}, \tag{2.3}$$

where

$$\begin{aligned}
\Sigma_\mu=\mathrm{Var}\big(\mu_{\boldsymbol{\beta}^*}(T_i,Y_i,\boldsymbol{X}_i)\big) &= \mathbb{E}\bigg(\frac{Y_i(1)^2}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}+\frac{Y_i(0)^2}{1-\pi^*_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}-(\mathbb{E}(Y_i(1))-\mathbb{E}(Y_i(0)))^2\bigg),\\
\boldsymbol{\Sigma}_{\boldsymbol{\beta}} &= (\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\boldsymbol{H}_{\mathbf{f}}^*)^{-1}\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\,\mathrm{Var}(\boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i,\boldsymbol{X}_i))\mathbf{W}^*\boldsymbol{H}_{\mathbf{f}}^*(\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\boldsymbol{H}_{\mathbf{f}}^*)^{-1},\\
\boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} &= -(\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\boldsymbol{H}_{\mathbf{f}}^*)^{-1}\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\,\mathrm{Cov}(\mu_{\boldsymbol{\beta}^*}(T_i,Y_i,\boldsymbol{X}_i),\boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i,\boldsymbol{X}_i)),
\end{aligned}$$

in which $\boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i,\boldsymbol{X}_i)$ is defined in (1.6). Under the model in equation (1.2), we have

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i,\boldsymbol{X}_i)) &= \mathbb{E}\left(\frac{\mathbf{f}(\boldsymbol{X}_i)\mathbf{f}(\boldsymbol{X}_i)^\top}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)(1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))}\right),\\
\mathrm{Cov}(\mu_{\boldsymbol{\beta}^*}(T_i,Y_i,\boldsymbol{X}_i),\boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i,\boldsymbol{X}_i)) &= \mathbb{E}\left[\frac{\{K(\boldsymbol{X}_i)+(1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))L(\boldsymbol{X}_i)\}\mathbf{f}(\boldsymbol{X}_i)}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)(1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))}\right].
\end{aligned}$$

The following theorem establishes the asymptotic normality of the CBPS-based IPTW estimator under the local misspecification of the propensity score model.

Theorem 2.1 (*Asymptotic Distribution under Local Misspecification of the Propensity Score Model*)**.** If the propensity score model is locally misspecified as in (2.1) with $\xi=n^{-1/2}$ and Assumption B.1 in Appendix B holds, the estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ in (1.5), where $\widehat{\boldsymbol{\beta}}$ is obtained by GMM (1.7), has the following asymptotic distribution

$$\sqrt{n}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}-\mu) \quad\xrightarrow{d}\quad N(B,\ \bar{\boldsymbol{H}}^{*\top}\boldsymbol{\Sigma}\bar{\boldsymbol{H}}^*), \tag{2.4}$$

where B is the asymptotic bias given in equation (2.2) and the asymptotic variance $\bar{\boldsymbol{H}}^{*\top}\boldsymbol{\Sigma}\bar{\boldsymbol{H}}^*$ is obtained from (2.3).

The theorem shows that the first order asymptotic bias of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is given by B under local model misspecification. In particular, this bias term implicitly depends on the covariate balancing function $\mathbf{f}(\cdot)$. Thus, we consider how to choose $\mathbf{f}(\cdot)$ such that the first order bias $|B|$ is minimized. While at the first glance the expression of B appears to be mathematically intractable, the next corollary shows that any $\mathbf{f}(\boldsymbol{X})$ satisfying (2.5) can eliminate the first order bias, $B = 0$.

Corollary 2.1. Suppose that the covariate balancing function $\mathbf{f}(\boldsymbol{X})$ satisfies the following condition: there exits some $\boldsymbol{\alpha} \in \mathbb{R}^m$ such that

$$\boldsymbol{\alpha}^\top \mathbf{f}(\boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)\mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i) + (1 - \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))\mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i). \tag{2.5}$$

In addition, assume that the dimension of $\mathbf{f}(\boldsymbol{X}_i)$ is equal to the number of parameters, i.e., $m = q$. Then, under the conditions in Theorem 2.1, the asymptotic bias of the IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is 0, i.e., $B = 0$.

Intuitively, the above result can be viewed as a "local" version of robustness of IPTW with respect to the misspecification of the propensity score model. The form of $\mathbf{f}(\boldsymbol{X}_i)$ in (2.5) implies that when balancing covariates, for any given unit we should give a greater weight to the determinants of the mean potential outcome that is less likely to be realized. For example, if a unit is less likely to be treated, then it is more important to balance the covariates that influence the mean potential outcome under the treatment condition. In the following, we focus on the asymptotic variance of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ in Theorem 2.1. Interestingly, we can show that the same choice of $\mathbf{f}(\boldsymbol{X}_i)$ in (2.5) minimizes the asymptotic variance.

Corollary 2.2. Under the same conditions in Corollary 2.1, the asymptotic variance of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is minimized by any covariate balancing function $\mathbf{f}(\boldsymbol{X}_i)$ which satisfies (2.5). In this case, the CBPS-based IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ attains the semiparametric asymptotic variance bound in Theorem 1 of Hahn (1998), i.e.,

$$V_{\text{opt}} = \mathbb{E}\left[\frac{\text{Var}(Y_i(1) \mid \boldsymbol{X}_i)}{\pi(\boldsymbol{X}_i)} + \frac{\text{Var}(Y_i(0) \mid \boldsymbol{X}_i)}{1 - \pi(\boldsymbol{X}_i)} + \{L(\boldsymbol{X}_i) - \mu\}^2\right]. \tag{2.6}$$

Based on Theorem 2.1, we can define the asymptotic mean squared error (AMSE) of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ as $AMSE = B^2 + \bar{\boldsymbol{H}}^{*\top}\boldsymbol{\Sigma}\bar{\boldsymbol{H}}^*$. Corollaries 2.1 and 2.2 together imply that $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ with $\mathbf{f}(\boldsymbol{X})$ satisfying (2.5) attains the minimum AMSE over all possible covariate balancing estimators. Thus, we refer to (2.5) as the optimality condition for the covariate balancing function. We note that there may exist

many choices of $\mathbf{f}(\boldsymbol{X})$ which satisfy (2.5). For instance, we can choose $f_1(\boldsymbol{X}) = \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)\mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i) + (1 - \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))\mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i)$ and $f_2, ..., f_m$ in an arbitrary way, as long as the estimating equation $\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) = 0$ is not degenerate. In this case, to implement $f_1(\boldsymbol{X})$, we need to further estimate $\boldsymbol{\beta}^*$ by some initial estimator, e.g., the maximum likelihood estimator, and estimate the conditional mean $\mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i)$ and $\mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i)$ by some parametric/nonparametric models. While Corollaries 2.1 and 2.2 hold with this choice of $\mathbf{f}(\boldsymbol{X})$, the empirical performance of the resulting estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is often unstable due to the estimation error of the initial estimators. To overcome this problem, we will next construct the optimal CBPS estimator that does not require any initial estimator.

3 The Optimal CBPS Methodology

Recall that the optimal covariate balancing function $\mathbf{f}(\boldsymbol{X})$ is given by (2.5). Plugging $\mathbf{f}(\boldsymbol{X})$ into the estimating function $\boldsymbol{g}_{\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)$ in (1.6), we obtain that

$$\begin{aligned}\boldsymbol{\alpha}^\top \boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i, \boldsymbol{X}_i) &= \left(\frac{T_i}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)} - \frac{1 - T_i}{1 - \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}\right)\left[\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)K(\boldsymbol{X}_i) + (1 - \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))(K(\boldsymbol{X}_i) + L(\boldsymbol{X}_i))\right] \\ &= \left(\frac{T_i}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)} - \frac{1 - T_i}{1 - \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}\right) K(\boldsymbol{X}_i) + \left(\frac{T_i}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)} - 1\right) L(\boldsymbol{X}_i). \qquad (3.1)\end{aligned}$$

In other words, the optimality condition (2.5) holds if and only if some linear combination of estimating function $\boldsymbol{g}_{\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)$ satisfies (3.1). Motivated by this observation, we construct the following set of estimating functions,

$$\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) = \begin{pmatrix} \bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) \\ \bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) \end{pmatrix}, \qquad (3.2)$$

where $\bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) = n^{-1}\sum_{i=1}^n \boldsymbol{g}_{1\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)$ and $\bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) = n^{-1}\sum_{i=1}^n \boldsymbol{g}_{2\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)$ with

$$\boldsymbol{g}_{1\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i) = \left(\frac{T_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - \frac{1 - T_i}{1 - \pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}\right)\boldsymbol{h}_1(\boldsymbol{X}_i), \boldsymbol{g}_{2\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i) = \left(\frac{T_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - 1\right)\boldsymbol{h}_2(\boldsymbol{X}_i), \qquad (3.3)$$

for some pre-specified functions $\boldsymbol{h}_1(\cdot) : \mathbb{R}^d \to \mathbb{R}^{m_1}$ and $\boldsymbol{h}_2(\cdot) : \mathbb{R}^d \to \mathbb{R}^{m_2}$ with $m_1 + m_2 = m$. It is easy to see that if the functions $K(\cdot)$ and $L(\cdot)$ lie in the linear space spanned by the functions $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$ respectively, then there exists a vector $\boldsymbol{\alpha} \in \mathbb{R}^m$ such that (3.1) holds for $(\boldsymbol{g}_{1\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i), \boldsymbol{g}_{2\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i))$, further implying that the optimality condition (2.5) is met.

As discussed in Section 2, the choice of the optimal covariate balancing function is not unique. Unlike the one mentioned after Corollary 2.2, the estimating function in (3.2) does not require any

initial estimators for $\boldsymbol{\beta}$ or the conditional mean models, and is more convenient for implementation. Given the estimating functions in (3.2), we can estimate $\boldsymbol{\beta}$ by the GMM estimator $\widehat{\boldsymbol{\beta}}$ in (1.7). We call this method as the optimal CBPS method (oCBPS). Similarly, the ATE is estimated by the IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ in (1.5). The implementation of the proposed oCBPS method (e.g., the choice of $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$) will be discussed in later sections.

It is worthwhile to note that $\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ has the following interpretation. The first set of functions $\bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ is the same as the existing covariate balancing moment function in (1.6), which balances the covariates $\boldsymbol{h}_1(\boldsymbol{X}_i)$ between the treatment and control groups. However, unlike the original CBPS method, we introduce another set of functions $\bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ which matches the weighted covariates $\boldsymbol{h}_2(\boldsymbol{X}_i)$ in the treatment group to the unweighted covariates $\boldsymbol{h}_2(\boldsymbol{X}_i)$ in the control group, because $\bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) = 0$ can be rewritten as

$$\sum_{T_i=1} \frac{1 - \pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} \boldsymbol{h}_2(\boldsymbol{X}_i) \;=\; \sum_{T_i=0} \boldsymbol{h}_2(\boldsymbol{X}_i).$$

As seen in the derivation of (3.1), the auxiliary estimating function $\bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ is required in order to satisfy the optimality condition.

3.1 Theoretical Properties

We now derive the theoretical properties of the IPTW estimator (1.5) based on the proposed oCBPS method. In particular, we will show that the proposed estimator is doubly robust and locally efficient. The following set of assumptions are imposed for the establishment of double robustness.

Assumption 3.1. The following regularity conditions are assumed.

1. There exists a positive definite matrix $\mathbf{W}^*$ such that $\widehat{\mathbf{W}} \xrightarrow{p} \mathbf{W}^*$.
2. For any $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$ in (3.3), the minimizer $\boldsymbol{\beta}^o = \operatorname{argmin}_{\boldsymbol{\beta} \in \Theta} \mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}))^\top \mathbf{W}^* \mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}))$ is unique.
3. $\boldsymbol{\beta}^o \in \operatorname{int}(\Theta)$, 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 $\boldsymbol{\beta} \in \text{int}(\Theta)$.

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

7. For any $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$ in (3.3) and $\mathbf{W}^*$ in part 1, $\mathbf{G}^* := \mathbb{E}(\partial \boldsymbol{g}(\boldsymbol{\beta}^o)/\partial \boldsymbol{\beta})$ exists where $\boldsymbol{g}(\boldsymbol{\beta}) = (\boldsymbol{g}_{1\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})^\top, \boldsymbol{g}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})^\top)^\top$ and there is a q-dimensional function $C(\boldsymbol{X})$ and a small constant $r > 0$ such that $\sup_{\boldsymbol{\beta} \in \mathbb{B}_r(\boldsymbol{\beta}^o)} |\partial \pi_{\boldsymbol{\beta}}(\boldsymbol{X})/\partial \beta_k| \leq C_k(\boldsymbol{X})$ for $1 \leq k \leq q$, and $\mathbb{E}(|h_{1j}(\boldsymbol{X})| C_k(\boldsymbol{X})) < \infty$ for $1 \leq j \leq m_1$, $1 \leq k \leq q$ and $\mathbb{E}(|h_{2j}(\boldsymbol{X})| C_k(\boldsymbol{X})) < \infty$ for $1 \leq j \leq m_2$, $1 \leq k \leq q$, where $\mathbb{B}_r(\boldsymbol{\beta}^o)$ is a ball in $\mathbb{R}^q$ with radius r and center $\boldsymbol{\beta}^o$.

Conditions 1–4 of Assumption 3.1 are the standard conditions for consistency of the GMM estimator (Newey and McFadden, 1994). Note that we allow the propensity score model to be misspecified, so that we use the notation $\boldsymbol{\beta}^o$ in Condition 2 to distinguish it from $\boldsymbol{\beta}^*$ used in the previous section. Condition 5 is the positivity assumption commonly used in the causal inference literature (Robins et al., 1994, 1995). Conditions 6 and 7 are technical conditions that enable us to apply the dominated convergence theorem. Note that, $\sup_{\boldsymbol{\beta} \in \mathbb{B}_r(\boldsymbol{\beta}^o)} |\partial \pi_{\boldsymbol{\beta}}(\boldsymbol{X})/\partial \beta_k| \leq C_k(\boldsymbol{X})$ in Condition 7 is a local condition in the sense that it only requires the existence of an envelop function $C_k(\boldsymbol{X})$ around a small neighborhood of $\boldsymbol{\beta}^o$.

We now establish the double robustness of the proposed estimator under Assumption 3.1.

Theorem 3.1 (*Double Robustness*)**.** Under Assumption 3.1, the proposed oCBPS-based IPTW estimator $\widehat{\mu}_{\widehat{\beta}}$ is doubly robust. That is, $\widehat{\mu}_{\widehat{\beta}} \xrightarrow{p} \mu$ if at least one of the following two conditions holds:

1. The propensity score model is correctly specified, i.e., $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)$;

2. The functions $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$ in (3.3) and $\mathbf{W}^*$ in Assumption 3.1 satisfy the following condition. There exist some vectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathbb{R}^q$ such that $K(\boldsymbol{X}_i) = \boldsymbol{\alpha}_1^\top \mathbf{M}_1 \boldsymbol{h}_1(\boldsymbol{X}_i)$ and $L(\boldsymbol{X}_i) = \boldsymbol{\alpha}_2^\top \mathbf{M}_2 \boldsymbol{h}_2(\boldsymbol{X}_i)$, where $\mathbf{M}_1 \in \mathbb{R}^{q \times m_1}$ and $\mathbf{M}_2 \in \mathbb{R}^{q \times m_2}$ are the partitions of $\mathbf{G}^{*\top} \mathbf{W}^* = (\mathbf{M}_1, \mathbf{M}_2)$.

Next, we establish the asymptotic normality of the proposed estimator if either the propensity score model (Condition 1 in Theorem 3.1) or the outcome model is correctly specified (Condition 2 in Theorem 3.1) . For this result, we require an additional set of regularity conditions.

Assumption 3.2. The following regularity conditions are assumed.

1. For any $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$ in (3.3) and $\mathbf{W}^*$ in Assumption 3.1, $\mathbf{G}^{*\top}\mathbf{W}^*\mathbf{G}^*$ and $\boldsymbol{\Omega} = \mathbb{E}(\boldsymbol{g}_{\boldsymbol{\beta}^o}(T_i, \boldsymbol{X}_i)\boldsymbol{g}_{\boldsymbol{\beta}^o}(T_i, \boldsymbol{X}_i)^\top)$ are nonsingular.

2. The function $C_k(\boldsymbol{X})$ defined in Condition 7 of Assumption 3.1 satisfies $\mathbb{E}(|Y(0)|C_k(\boldsymbol{X})) < \infty$ and $\mathbb{E}(|Y(1)|C_k(\boldsymbol{X})) < \infty$ for $1 \leq k \leq q$.

Condition 1 of Assumption 3.2 ensures the non-singularity of the asymptotic variance matrix and Condition 2 is a mild technical condition required for the dominated convergence theorem.

Theorem 3.2 (*Asymptotic Normality*)**.** Suppose that Assumptions 3.1 and 3.2 hold.

1. If Condition 1 of Theorem 3.1 holds, then the proposed oCBPS-based IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ has the following asymptotic distribution:

$$\sqrt{n}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu) \xrightarrow{d} N\left(0,\ \bar{\mathbf{H}}^{*\top}\boldsymbol{\Sigma}\bar{\mathbf{H}}^*\right), \tag{3.4}$$

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

$$\begin{aligned}
\mathbf{H}^* &= -\mathbb{E}\left(\frac{K(\boldsymbol{X}_i) + (1 - \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))L(\boldsymbol{X}_i)}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)(1 - \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))} \cdot \frac{\partial \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)}{\partial \boldsymbol{\beta}}\right), \\
\boldsymbol{\Sigma} &= \begin{pmatrix} \Sigma_\mu & \boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}}^\top \\ \boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} & \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \end{pmatrix}, \quad \text{with} \quad \Sigma_\mu = \mathbb{E}\left(\frac{Y_i^2(1)}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)} + \frac{Y_i^2(0)}{1 - \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)}\right) - \mu^2.
\end{aligned} \tag{3.5}$$

In addition, $\boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}}$ is given by

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} = -(\mathbf{G}^{*\top}\mathbf{W}^*\mathbf{G}^*)^{-1}\mathbf{G}^{*\top}\mathbf{W}^*\Big\{&\mathbb{E}\Big(\frac{K(\boldsymbol{X}_i) + (1 - \pi_i^o)L(\boldsymbol{X}_i)}{(1 - \pi_i^o)\pi_i^o}\boldsymbol{h}_1^\top(\boldsymbol{X}_i)\Big), \\
&\mathbb{E}\Big(\frac{K(\boldsymbol{X}_i) + (1 - \pi_i^o)L(\boldsymbol{X}_i)}{\pi_i^o}\boldsymbol{h}_2^\top(\boldsymbol{X}_i)\Big)\Big\}^\top.
\end{aligned}$$

2. If Condition 2 of Theorem 3.1 holds, then the proposed oCBPS-based IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ has the following asymptotic distribution:

$$\sqrt{n}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu) \xrightarrow{d} N\left(0,\ \widetilde{\mathbf{H}}^{*\top}\widetilde{\boldsymbol{\Sigma}}\widetilde{\mathbf{H}}^*\right), \tag{3.6}$$

where $\widetilde{\mathbf{H}}^* = (1, \check{\mathbf{H}}^{*\top})^\top$,

$$\begin{aligned}
\check{\mathbf{H}}^* &= -\mathbb{E}\left[\left\{\frac{\pi(\boldsymbol{X}_i)(K(\boldsymbol{X}_i) + L(\boldsymbol{X}_i))}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)^2} + \frac{(1 - \pi(\boldsymbol{X}_i))K(\boldsymbol{X}_i)}{(1 - \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))^2}\right\}\frac{\partial \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)}{\partial \boldsymbol{\beta}^o}\right], \\
\widetilde{\boldsymbol{\Sigma}} &= \begin{pmatrix} \widetilde{\Sigma}_\mu & \widetilde{\boldsymbol{\Sigma}}_{\mu\boldsymbol{\beta}}^\top \\ \widetilde{\boldsymbol{\Sigma}}_{\mu\boldsymbol{\beta}} & \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \end{pmatrix} \quad \text{with} \quad \widetilde{\Sigma}_\mu = \mathbb{E}\left(\frac{\pi(\boldsymbol{X}_i)Y_i^2(1)}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)^2} + \frac{(1 - \pi(\boldsymbol{X}_i))Y_i^2(0)}{(1 - \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))^2}\right) - \mu^2.
\end{aligned}$$

In addition, $\widetilde{\boldsymbol{\Sigma}}_{\mu\boldsymbol{\beta}}$ is given by

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

where $\boldsymbol{S} = (\boldsymbol{S}_1^\top, \boldsymbol{S}_2^\top)^\top$ and

$$\begin{aligned}
\boldsymbol{S}_1 \;&=\; \mathbb{E}\left[\left\{\frac{\pi(\boldsymbol{X}_i)(K(\boldsymbol{X}_i)+L(\boldsymbol{X}_i)-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)\mu)}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)^2}\right.\right. \\
&\qquad\left.\left.+\frac{(1-\pi(\boldsymbol{X}_i))(K(\boldsymbol{X}_i)+(1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))\mu)}{(1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))^2}\right\}\boldsymbol{h}_1(\boldsymbol{X}_i)\right], \\
\boldsymbol{S}_2 \;&=\; \mathbb{E}\left[\left\{\frac{\pi(\boldsymbol{X}_i)[(K(\boldsymbol{X}_i)+L(\boldsymbol{X}_i))(1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)\mu]}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)^2}\right.\right. \\
&\qquad\left.\left.+\frac{(1-\pi(\boldsymbol{X}_i))K(\boldsymbol{X}_i)+(1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))\mu}{1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)}\right\}\boldsymbol{h}_2(\boldsymbol{X}_i)\right].
\end{aligned}$$

3. If both Conditions 1 and 2 of Theorem 3.1 hold and $\mathbf{W}^* = \boldsymbol{\Omega}^{-1}$, then the proposed oCBPS-based IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ has the following asymptotic distribution:

$$\sqrt{n}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu) \;\xrightarrow{d}\; N(0, V),$$

where

$$V = \Sigma_\mu - (\boldsymbol{\alpha}_1^\top\mathbf{M}_1, \boldsymbol{\alpha}_2^\top\mathbf{M}_2)\mathbf{G}^*(\mathbf{G}^{*\top}\boldsymbol{\Omega}^{-1}\mathbf{G}^*)^{-1}\mathbf{G}^{*\top}\begin{pmatrix}\mathbf{M}_1^\top\boldsymbol{\alpha}_1\\ \mathbf{M}_2^\top\boldsymbol{\alpha}_2\end{pmatrix} \tag{3.7}$$

and Σ_μ is defined in (3.5).

The asymptotic variance V in (3.7) contains two terms. The first term Σ_μ represents the variance of each summand in the estimator defined in equation (1.5) with $\widehat{\boldsymbol{\beta}}$ replaced by $\boldsymbol{\beta}^o$. The second term can be interpreted as the effect of estimating $\boldsymbol{\beta}$ via covariate balance conditions. Since this second term is nonnegative, the proposed estimator is more efficient than the standard IPTW estimator with the true propensity score model, i.e., $V \leq \Sigma_\mu$. In particular, Henmi and Eguchi (2004) offered a theoretical analysis of such efficiency gain due to the estimation of nuisance parameters under a general estimating equation framework.

Since the choice of $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$ can be arbitrary, it might be tempting to incorporate more covariate balancing conditions into $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$. However, the following corollary shows that under Conditions 1 and 2 of Theorem 3.1 one cannot improve the efficiency of the proposed estimator by increasing the number of functions $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$ or equivalently, the dimensionality of covariate balance conditions, i.e., $\bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ and $\bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$.

Corollary 3.1. Define $\bar{\boldsymbol{h}}_1(\boldsymbol{X}) = (\boldsymbol{h}_1^\top(\boldsymbol{X}), \boldsymbol{a}_1^\top(\boldsymbol{X}))^\top$ and $\bar{\boldsymbol{h}}_2(\boldsymbol{X}) = (\boldsymbol{h}_2^\top(\boldsymbol{X}), \boldsymbol{a}_2^\top(\boldsymbol{X}))^\top$, where $\boldsymbol{a}_1(\cdot)$ and $\boldsymbol{a}_2(\cdot)$ are some additional covariate balancing functions. Similarly, let $\bar{\boldsymbol{g}}_1(\boldsymbol{X})$ and $\bar{\boldsymbol{g}}_2(\boldsymbol{X})$ denote the corresponding estimating equations defined by $\bar{\boldsymbol{h}}_1(\boldsymbol{X})$ and $\bar{\boldsymbol{h}}_2(\boldsymbol{X})$. The resulting oCBPS-based IPTW estimator is denoted by $\bar{\mu}_{\widehat{\boldsymbol{\beta}}}$ where $\widehat{\boldsymbol{\beta}}$ is in (1.7) and its asymptotic variance is denoted by $\bar{V}$. Under Conditions 1 and 2 of Theorem 3.1, we have $V \leq \bar{V}$, where V is defined in (3.7).

The above corollary shows a potential trade-off between robustness and efficiency when choosing $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$. Recall that Condition 2 of Theorem 3.1 implies $K(\boldsymbol{X}_i) = \boldsymbol{\alpha}_1^\top \mathbf{M}_1 \boldsymbol{h}_1(\boldsymbol{X}_i)$ and $L(\boldsymbol{X}_i) = \boldsymbol{\alpha}_2^\top \mathbf{M}_2 \boldsymbol{h}_2(\boldsymbol{X}_i)$. Therefore, we can make the proposed estimator more robust by incorporating more basis functions into $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$, such that this condition is more likely to hold. However, Corollary 3.1 shows that doing so may inflate the variance of the proposed estimator.

In the following, we focus on the efficiency of the estimator. Using the notations in this section, we can rewrite the semiparametric asymptotic variance bound V_{opt} in (2.6) as

$$V_{\text{opt}} \;=\; \Sigma_\mu - (\boldsymbol{\alpha}_1^\top \mathbf{M}_1, \boldsymbol{\alpha}_2^\top \mathbf{M}_2) \boldsymbol{\Omega} \begin{pmatrix} \mathbf{M}_1^\top \boldsymbol{\alpha}_1 \\ \mathbf{M}_2^\top \boldsymbol{\alpha}_2 \end{pmatrix}. \tag{3.8}$$

Comparing this expression with (3.7), we see that the proposed estimator is semiparametrically efficient if $\mathbf{G}^*$ is a square matrix (i.e., $m = q$) and invertible. This important result is summarized as the following corollary.

Corollary 3.2. Assume $m = q$ and $\mathbf{G}^*$ is invertible. Under Assumption 3.1, the proposed estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ in (1.5) is doubly robust in the sense that $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} \xrightarrow{p} \mu$ if either of the following conditions holds:

1. The propensity score model is correctly specified. That is $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)$.
2. There exist some vectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathbb{R}^q$ such that $K(\boldsymbol{X}_i) = \boldsymbol{\alpha}_1^\top \boldsymbol{h}_1(\boldsymbol{X}_i)$ and $L(\boldsymbol{X}_i) = \boldsymbol{\alpha}_2^\top \boldsymbol{h}_2(\boldsymbol{X}_i)$.

In addition, under Assumption 3.2, if both conditions hold, then the proposed estimator has the asymptotic variance given in (3.8). Thus, our estimator is a locally semiparametric efficient estimator in the sense of Robins et al. (1994).

The corollary shows that the proposed oCBPS method has two advantages over the original CBPS method (Imai and Ratkovic, 2014) with balancing first and second moments of $\boldsymbol{X}_i$ and/or the score function of the propensity score model. First, the proposed estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is robust to

model misspecification, whereas the original CBPS estimator does not have that property. Second, the proposed oCBPS estimator can be more efficient than the original CBPS estimator.

Corollary 3.2 also implies that the asymptotic variance of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is identical to the semiparametric variance bound V_{opt}, even if we incorporate additional covariate balancing functions into $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$. Namely, under the conditions in Corollary 3.2, we have $V = \bar{V} = V_{\text{opt}}$ in the context of Corollary 3.1. Thus, in this setting, we can improve the robustness of the estimator without sacrificing the efficiency by increasing the number of functions $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$. Meanwhile, this also makes the propensity score model more flexible, since we need to increase the number of parameters $\boldsymbol{\beta}$ to match $m = q$ as required in Corollary 3.2. This observation further motivates us to consider a sieve estimation approach to improve the oCBPS method, as shown in Section 4.

Remark 3.1 (Implementation of the oCBPS method). Based on Corollary 3.2, $\boldsymbol{h}_1(\cdot)$ serves as the basis functions for the baseline conditional mean function $K(\cdot)$, while $\boldsymbol{h}_2(\cdot)$ represents the basis functions for the conditional average treatment effect function $L(\cdot)$. Thus, in practice, researchers can choose a set of basis functions for the baseline conditional mean function and the conditional average treatment effect function when determining the specification for $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$. Once these functions are selected, they can over-parameterize the propensity score model by including some higher order terms or interactions such that $m = q$ holds. The resulting oCBPS-based IPTW estimator may reduce bias under model misspecification and attain high efficiency.

Remark 3.2. We also extend the oCBPS method to the estimation of the average treatment effect for the treated (ATT). Given the space limitation, we defer the details to the supplementary material.

3.2 Comparison with Related Estimators

Next, we compare the proposed estimator with some related estimators from the literature. We begin with the following standard AIPW estimator of Robins et al. (1994),

$$\widehat{\mu}^{AIPW}_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_iY_i}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - \frac{(1-T_i)Y_i}{1-\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - (T_i - \pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i))\left(\frac{K(\boldsymbol{X}_i,\boldsymbol{\alpha}) + L(\boldsymbol{X}_i,\boldsymbol{\gamma})}{\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} + \frac{K(\boldsymbol{X}_i,\boldsymbol{\alpha})}{1-\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}\right)\right\},$$

where $K(\boldsymbol{X}_i,\boldsymbol{\alpha})$ and $L(\boldsymbol{X}_i,\boldsymbol{\gamma})$ are the conditional mean models indexed by finite dimensional parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. Assume the linear outcome models: $K(\boldsymbol{X}_i,\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T\boldsymbol{h}_1(\boldsymbol{X}_i)$ and $L(\boldsymbol{X}_i,\boldsymbol{\gamma}) =$

$\boldsymbol{\gamma}^T \boldsymbol{h}_2(\boldsymbol{X}_i)$. It is interesting to note that our IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ in Corollary 3.2 can be rewritten as the AIPW estimator $\widehat{\mu}^{AIPW}_{\widehat{\boldsymbol{\beta}},\boldsymbol{\alpha},\boldsymbol{\gamma}}$ (for any $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$), since we have,

$$\frac{1}{n}\sum_{i=1}^{n}(T_i - \pi_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{X}_i))\left(\frac{K(\boldsymbol{X}_i,\boldsymbol{\alpha}) + L(\boldsymbol{X}_i,\boldsymbol{\gamma})}{\pi_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)} + \frac{K(\boldsymbol{X}_i,\boldsymbol{\alpha})}{1-\pi_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{X}_i)}\right) \;=\; 0,$$

by the definition of the covariate balancing estimating equations in (3.2).

It is well known that the AIPW estimator is consistent provided that either the propensity score model or the outcome model is correctly specified. Since both the AIPW estimator and our estimator are doubly robust and locally efficient, in the following we conduct a theoretical investigation of these two estimators under the scenario that *both* propensity score and outcome models are misspecified. Indeed, this scenario corresponds to the simulation settings used in the influential study of Kang and Schafer (2007).

To make the comparison mathematically tractable, we focus on the case that both of these two models are locally misspecified. Similar to Section 2, we assume that the true treatment assignment satisfies, $\pi(\boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)\exp(\xi\, u(\boldsymbol{X}_i;\boldsymbol{\beta}^*))$ in (2.1), while the true regression functions $K(\boldsymbol{X}_i)$ and $L(\boldsymbol{X}_i)$ in (1.2) satisfy

$$K(\boldsymbol{X}_i) = \boldsymbol{\alpha}^{*\top}\boldsymbol{h}_1(\boldsymbol{X}_i) + \delta r_1(\boldsymbol{X}_i), \quad L(\boldsymbol{X}_i) = \boldsymbol{\gamma}^{*\top}\boldsymbol{h}_2(\boldsymbol{X}_i) + \delta r_2(\boldsymbol{X}_i), \tag{3.9}$$

where $\boldsymbol{\alpha}^*$ and $\boldsymbol{\gamma}^*$ can be viewed as the approximate true values of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, the functions $r_1(\boldsymbol{X}_i)$ and $r_2(\boldsymbol{X}_i)$ determine the direction of misspecification, and $\delta \in \mathbb{R}$ represents the magnitude of misspecification.

Assume further that the models are locally misspecified, i.e., $\xi, \delta = o(1)$. Under regularity conditions similar to Section 2, we can show that the proposed estimator satisfies,

$$\begin{aligned}\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{T_i}{\pi(\boldsymbol{X}_i)}\{Y_i(1) - K(\boldsymbol{X}_i) - L(\boldsymbol{X}_i)\} - \frac{1-T_i}{1-\pi(\boldsymbol{X}_i)}\{Y_i(0) - K(\boldsymbol{X}_i)\} + L(\boldsymbol{X}_i) - \mu\Big] \\ + O_p(\xi^2\delta + \delta n^{-1/2} + \xi n^{-1/2}),\end{aligned} \tag{3.10}$$

whereas the AIPW estimator satisfies,

$$\begin{aligned}\widehat{\mu}^{AIPW}_{\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\alpha}},\widetilde{\boldsymbol{\gamma}}} - \mu = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{T_i}{\pi(\boldsymbol{X}_i)}\{Y_i(1) - K(\boldsymbol{X}_i) - L(\boldsymbol{X}_i)\} - \frac{1-T_i}{1-\pi(\boldsymbol{X}_i)}\{Y_i(0) - K(\boldsymbol{X}_i)\} + L(\boldsymbol{X}_i) - \mu\Big] \\ + O_p(\xi\delta + \delta n^{-1/2} + \xi n^{-1/2}),\end{aligned} \tag{3.11}$$

where $\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}$ and $\widetilde{\boldsymbol{\gamma}}$ are the corresponding maximum likelihood and least square estimators. The derivation of (3.10) and (3.11) is shown in Appendix I.

The leading terms in the asymptotic expansions of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu$ and $\widehat{\mu}^{AIPW}_{\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\alpha}},\widetilde{\boldsymbol{\gamma}}} - \mu$ are identical and are known as the efficient influence function for μ. However, the remainder terms in (3.10) and (3.11) may have different order. Consider the following two scenarios. First, if $\xi\delta \gg n^{-1/2}$, then we have $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu = O_p(\xi^2\delta + n^{-1/2})$ and $\widehat{\mu}^{AIPW}_{\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\alpha}},\widetilde{\boldsymbol{\gamma}}} - \mu = O_p(\xi\delta)$. Thus, the proposed estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ converges in probability to the ATE at a faster rate than $\widehat{\mu}^{AIPW}_{\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\alpha}},\widetilde{\boldsymbol{\gamma}}}$. Second, if $\xi\delta = o(n^{-1/2})$, the two estimators have the same limiting distribution, i.e., $\sqrt{n}(\widehat{\mu}-\mu) \xrightarrow{d} N(0, V_{\text{opt}})$, where $\widehat{\mu}$ can be either $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ or $\widehat{\mu}^{AIPW}_{\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\alpha}},\widetilde{\boldsymbol{\gamma}}}$. However, the rates of convergence of the Gaussian approximation determined by the remainder terms in (3.10) and (3.11) are different. For instance, assume that $\xi = \delta = n^{-(1/4+\epsilon)}$ for some small positive $\epsilon < 1/4$. We observe that the remainder term in (3.10) is of order $O_p(n^{-(3/4+\epsilon)})$ and is smaller in magnitude than the corresponding term in (3.11), which is of order $O_p(n^{-(1/2+2\epsilon)})$. As a result, the proposed estimator converges in distribution to $N(0, V_{\text{opt}})$ at a faster rate than the AIPW estimator. The above analysis justifies the theoretical advantage of the proposed oCBPS estimator over the standard AIPW estimator.

Furthermore, the proposed estimator is related to the class of bias-reduced doubly robust estimators (Vermeulen and Vansteelandt, 2015), see also Robins et al. (2007). To see this, we consider the derivative of $\widehat{\mu}^{AIPW}_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}}$ with respect to the nuisance parameters $\boldsymbol{\alpha}, \boldsymbol{\gamma}$. In particular, under the linear outcome models, it is easily shown that $\partial\widehat{\mu}^{AIPW}_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}}/\partial\boldsymbol{\alpha} = \bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ and $\partial\widehat{\mu}^{AIPW}_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}}/\partial\boldsymbol{\gamma} = \bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$, where $\bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ and $\bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ are our covariate balancing functions in (3.2). This provides an alternative justification for the proposed method: the oCBPS estimator $\widehat{\boldsymbol{\beta}}$, which satisfies $\bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) = 0$ and $\bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) = 0$, removes the local effect of the estimated nuisance parameters, i.e., $\partial\widehat{\mu}^{AIPW}_{\widehat{\boldsymbol{\beta}},\boldsymbol{\alpha},\boldsymbol{\gamma}}/\partial\boldsymbol{\alpha} = 0$ and $\partial\widehat{\mu}^{AIPW}_{\widehat{\boldsymbol{\beta}},\boldsymbol{\alpha},\boldsymbol{\gamma}}/\partial\boldsymbol{\gamma} = 0$. This property would not hold if we replace $\widehat{\boldsymbol{\beta}}$ by the maximum likelihood estimator or other convenient estimators of $\boldsymbol{\beta}$. Vermeulen and Vansteelandt (2015) defined the class of bias-reduced doubly robust estimator as $\widehat{\mu}^{AIPW}_{\bar{\boldsymbol{\beta}},\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{\gamma}}}$, where $(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\gamma}})$ are the estimators corresponding to the estimating equations $\partial\widehat{\mu}^{AIPW}_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}}/\partial\boldsymbol{\alpha} = 0, \partial\widehat{\mu}^{AIPW}_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}}/\partial\boldsymbol{\gamma} = 0, \partial\widehat{\mu}^{AIPW}_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}}/\partial\boldsymbol{\beta} = 0$. The first two sets of estimating equations are identical to the covariate balancing estimating equations in (3.2), whereas the last set of estimating equations $\partial\widehat{\mu}^{AIPW}_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}}/\partial\boldsymbol{\beta} = 0$ (leading to the estimators $\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\gamma}}$) is unnecessary in our setting because $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} = \widehat{\mu}^{AIPW}_{\widehat{\boldsymbol{\beta}},\boldsymbol{\alpha},\boldsymbol{\gamma}}$ does not rely on how $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are estimated. As expected, all the theoretical

properties of the bias-reduced doubly robust estimator in Section 3 of Vermeulen and Vansteelandt (2015) hold for our estimator.

Recently, a variety of empirical likelihood based estimators are proposed to match the moment of covariates in treatment and control groups (e.g., Tan, 2006, 2010; Hainmueller, 2012; Graham et al., 2012; Han and Wang, 2013; Chan et al., 2016; Zubizarreta, 2015; Zhao and Percival, 2017). Usually, these methods aim to estimate $\mathbb{E}(Y_i(1))$ and $\mathbb{E}(Y_i(0))$ (or $\mathbb{E}(Y_i(1) \mid T_i = 0)$ and $\mathbb{E}(Y_i(0) \mid T_i = 1)$) separately and combine then to estimate the ATE. Our approach directly estimates the propensity score and ATE by jointly solving the potentially over-identified estimating functions (3.2). In addition, our asymptotic results and the discussion rely on the GMM theory for over-identified estimating functions which is different from these methods. Another recent paper by Zhao (2019) studied the robustness of a general class of loss function based covariate balancing methods. When the goal is to estimate the ATE, his score function reduces to our first set of estimating functions $\bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ in (3.2). In this case, his estimator is robust to the misspecification of the propensity score model under the constant treatment effect model, i.e., $L(\boldsymbol{X}) = \tau^*$ for some constant τ^*. In comparison, our methodology and theoretical results cover a broader case that allows for heterogeneous treatment effects.

4 Nonparametric oCBPS Methodology

In this section, we extend our theoretical results of the oCBPS methodology to nonparametric estimation. As seen in Corollary 3.2, the proposed estimator is efficient if both the propensity score $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i)$ and the conditional mean functions $K(\cdot)$ and $L(\cdot)$ are correctly specified. To avoid model misspecification, we can choose a large number of basis functions $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$, such that the conditional mean functions $K(\cdot)$ and $L(\cdot)$ satisfy the condition 2 in Corollary 3.2.

However, the parametric assumption for the propensity score model $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)$ imposed in Corollary 3.2 may be too restrictive. Once the propensity score model is misspecified, the proposed oCBPS-based IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is inefficient and could even become inconsistent. To relax the strong parametric assumptions imposed in the previous sections, we propose a flexible nonparametric approach for modeling the propensity score and the conditional mean functions. The main advantage of this nonparametric approach is that, the resulting oCBPS-based IPTW estimator is semiparametrically efficient under a much broader class of propensity score models

and the conditional mean models than those of Corollary 3.2.

Specifically, we assume $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = J(\psi^*(\boldsymbol{X}_i))$, where $J(\cdot)$ is a known monotonic link function (e.g., $J(\cdot) = \exp(\cdot)/(1+\exp(\cdot))$), and $\psi^*(\cdot)$ is an unknown smooth function. One practical way to estimate $\psi^*(\cdot)$ is to approximate it by the linear combination of κ basis functions, where κ is allowed to grow with n. This approach is known as the sieve estimation (Andrews, 1991; Newey, 1997). In detail, let $\boldsymbol{B}(\boldsymbol{x}) = \{b_1(\boldsymbol{x}), ..., b_\kappa(\boldsymbol{x})\}$ denote a collection of κ basis functions, whose mathematical requirement is given in Assumption E.1. Intuitively, we would like to approximate $\psi^*(\boldsymbol{x})$ by $\boldsymbol{\beta}^{*\top}\boldsymbol{B}(\boldsymbol{x})$, for some coefficient $\boldsymbol{\beta}^* \in \mathbb{R}^\kappa$.

To estimate $\boldsymbol{\beta}^*$, similar to the parametric case, we define $\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) = \sum_{i=1}^n \boldsymbol{g}_{\boldsymbol{\beta}}(\boldsymbol{T}_i, \boldsymbol{X}_i)/n$, where $\boldsymbol{g}_{\boldsymbol{\beta}}(\boldsymbol{T}_i, \boldsymbol{X}_i) = (\boldsymbol{g}_{1\boldsymbol{\beta}}^\top(\boldsymbol{T}_i, \boldsymbol{X}_i), \boldsymbol{g}_{2\boldsymbol{\beta}}^\top(\boldsymbol{T}_i, \boldsymbol{X}_i))^\top$ with,

$$\begin{aligned}
\boldsymbol{g}_{1\boldsymbol{\beta}}(\boldsymbol{T}_i, \boldsymbol{X}_i) &= \left(\frac{T_i}{J(\boldsymbol{\beta}^\top\boldsymbol{B}(\boldsymbol{X}_i))} - \frac{1-T_i}{1-J(\boldsymbol{\beta}^\top\boldsymbol{B}(\boldsymbol{X}_i))}\right)\boldsymbol{h}_1(\boldsymbol{X}_i), \\
\boldsymbol{g}_{2\boldsymbol{\beta}}(\boldsymbol{T}_i, \boldsymbol{X}_i) &= \left(\frac{T_i}{J(\boldsymbol{\beta}^\top\boldsymbol{B}(\boldsymbol{X}_i))} - 1\right)\boldsymbol{h}_2(\boldsymbol{X}_i).
\end{aligned}$$

Recall that $\boldsymbol{h}_1(\boldsymbol{X}) \in \mathbb{R}^{m_1}$ and $\boldsymbol{h}_2(\boldsymbol{X}) \in \mathbb{R}^{m_2}$ are interpreted as the basis functions for $K(\boldsymbol{X})$ and $L(\boldsymbol{X})$. Let $m_1 + m_2 = m$ and $\boldsymbol{h}(\boldsymbol{X}) = (\boldsymbol{h}_1(\boldsymbol{X})^\top, \boldsymbol{h}_2(\boldsymbol{X})^\top)^\top$. Here, we assume $m = \kappa$, so that the number of equations in $\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ is identical to the dimension of the parameter $\boldsymbol{\beta}$. Then define $\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\Theta} \|\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})\|_2^2$, where Θ is the parameter space for $\boldsymbol{\beta}$ and $\|\boldsymbol{v}\|_2$ represents the L_2 norm of the vector $\boldsymbol{v}$. The resulting IPTW estimator is,

$$\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}} = \frac{1}{n}\sum_{i=1}^n \left(\frac{T_i Y_i}{J(\widetilde{\boldsymbol{\beta}}^\top\boldsymbol{B}(\boldsymbol{X}_i))} - \frac{(1-T_i)Y_i}{1-J(\widetilde{\boldsymbol{\beta}}^\top\boldsymbol{B}(\boldsymbol{X}_i))}\right).$$

To establish the large sample properties of $\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}}$, we require a few regularity conditions. Due to the space constraint, we defer the regularity conditions to the supplementary material. The following theorem establishes the asymptotic normality and semiparametric efficiency of the estimator $\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}}$.

Theorem 4.1 (*Efficiency under nonparametric models).* Assume that Assumption E.1 in the supplementary material holds, and there exist $r_b, r_h > 1/2$, $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}_1^{*\top}, \boldsymbol{\alpha}_2^{*\top})^\top \in \mathbb{R}^\kappa$, such that the propensity score model satisfies

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |\psi^*(\boldsymbol{x}) - \boldsymbol{\beta}^{*\top}\boldsymbol{B}(\boldsymbol{x})| = O(\kappa^{-r_b}), \tag{4.1}$$

and the outcome models $K(\cdot)$ and $L(\cdot)$ satisfy

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |K(\boldsymbol{x}) - \boldsymbol{\alpha}_1^{*\top}\boldsymbol{h}_1(\boldsymbol{x})| = O(\kappa^{-r_h}), \quad \sup_{\boldsymbol{x}\in\mathcal{X}} |L(\boldsymbol{x}) - \boldsymbol{\alpha}_2^{*\top}\boldsymbol{h}_2(\boldsymbol{x})| = O(\kappa^{-r_h}). \tag{4.2}$$

Assume $\kappa = o(n^{1/3})$ and $n^{\frac{1}{2(r_b+r_h)}} = o(\kappa)$. Then

$$n^{1/2}(\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}} - \mu) \xrightarrow{d} N(0, V_{\text{opt}}),$$

where V_{opt} is the asymptotic variance bound in (2.6). Thus, $\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}}$ is semiparametrically efficient.

This theorem can be viewed as a nonparametric version of Corollary 3.2. It shows that one can construct a globally efficient estimator of the treatment effect without imposing strong parametric assumptions on the propensity score model and the outcome model. Since the estimator is asymptotically equivalent to the sample average of the efficient influence function, it is also adaptive in the sense of Bickel et al. (1998).

In the following, we comment on the technical assumptions of Theorem 4.1. We assume $\psi^*(\boldsymbol{x})$ and $K(\boldsymbol{x})$ (also $L(\boldsymbol{x})$) can be uniformly approximated by the basis functions $\boldsymbol{B}(\boldsymbol{x})$ and $\boldsymbol{h}_1(\boldsymbol{x})$ (also $\boldsymbol{h}_2(\boldsymbol{x})$) in (4.1) and (4.2), respectively. It is well known that the uniform rate of convergence is related to the smoothness of the functions $\psi^*(\boldsymbol{x})$ and $K(\boldsymbol{x})$ (also $L(\boldsymbol{x})$) and the dimension of $\boldsymbol{X}$. For instance, if the function class $\mathcal{M}$ for $\psi^*(\boldsymbol{x})$ and $\mathcal{H}$ for $K(\boldsymbol{x})$ (also $L(\boldsymbol{x})$) correspond to the Hölder class with smoothness parameter s on the domain $\mathcal{X} = [0,1]^d$, under the assumption that $m_1 \asymp m_2 \asymp \kappa$, (4.1) and (4.2) hold for the spline basis and wavelet basis with $r_b = r_h = s/d$; see Newey (1997); Chen (2007) for details. In the same setting, Hirano et al. (2003) considered a nonparametric IPTW estimator, which is globally efficient under the condition $s/d > 7$. Imbens et al. (2007) established the asymptotic equivalence between a regression based estimator and Hirano et al. (2003)'s estimator under $s/d > 9$. Recently, Chan et al. (2016) proposed a sieve based calibration estimator under the condition $s/d > 13$. Compared to these existing results, our theorem needs a much weaker condition, i.e., $s/d > 3/4$. We refer to the supplementary material for further technical discussion of our nonparametric estimator.

5 Simulation and Empirical Studies

5.1 Simulation Studies

In this section, we conduct a set of simulation studies to examine the performance of the proposed methodology. We consider the following linear model for the potential outcomes,

$$\begin{aligned} Y_i(1) &= 200 + 27.4X_{i1} + 13.7X_{i2} + 13.7X_{i3} + 13.7X_{i4} + \varepsilon_i, \\ Y_i(0) &= 200 + 13.7X_{i2} + 13.7X_{i3} + 13.7X_{i4} + \varepsilon_i. \end{aligned}$$

where $\varepsilon_i \sim N(0,1)$, independent of $\boldsymbol{X}_i$, and consider the following true propensity score model

$$\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{\exp(-\beta_1 x_{i1} + 0.5x_{i2} - 0.25x_{i3} - 0.1x_{i4})}{1 + \exp(-\beta_1 x_{i1} + 0.5x_{i2} - 0.25x_{i3} - 0.1x_{i4})}, \tag{5.1}$$

where β_1 varies from 0 to 1. When implementing the proposed methodology, we set $\boldsymbol{h}_1(\boldsymbol{x}_i) = (1, x_{i2}, x_{i3}, x_{i4})$ and $\boldsymbol{h}_2(\boldsymbol{x}_i) = x_{i1}$ so that the number of equations is equal to the number of parameters to be estimated. Covariate X_{i1} is generated independently from $N(3,2)$ and X_{i2}, X_{i3} and X_{i4} are generated from $N(0,1)$. Each set of results is based on 500 Monte Carlo simulations.

We examine the performance of the IPTW estimator when the propensity score model is fitted using maximum likelihood (GLM), the standard CBPS with balancing the first moment (CBPS), and the proposed optimal CBPS (oCBPS) as well as the case where the true propensity score (True), i.e., $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, is used for the IPTW estimator. In addition, we include the IPTW estimator when the propensity score model is estimated by logistic series (Hirano et al., 2003). Since a fully nonparametric logistic series approach is impractical to implement due to the curse of dimensionality, instead we consider a generalized additive model (GAM) and apply the logistic series approach to each of the covariate separately. Finally, we also include the targeted maximum likelihood estimator, a doubly robust estimator (DR, Benkeser et al. (2017)), using the R package drtmle.

In the first set of simulations, we use the correctly specified propensity score and outcome models. Table 5.1 shows the standard deviation, bias, root mean square error (RMSE), and the coverage probability of the constructed 95% confidence intervals of these estimators when the sample size is $n = 300$ and $n = 1000$. The confidence intervals are constructed using estimates of the asymptotic variances of the estimators. The exact formulas can be found in the supplementary material. We find that CBPS and oCBPS substantially outperform True, GLM, and GAM in terms of efficiency, and in most cases outperform DR as well. In addition, oCBPS is more efficient than

Table 5.1: The bias, standard deviation, root mean squared error (RMSE), and the coverage probability of the constructed 95% C.I. of the IPTW estimator with known propensity score (True), the IPTW estimator when the propensity score is fitted using the maximum likelihood (GLM), the IPTW estimator when the propensity score is fitted using the generalized additive model (GAM), the targeted maximum likelihood estimator (DR), the standard CBPS estimator balancing the first moment (CBPS), and the proposed optimal CBPS estimator (oCBPS) under the scenario that both the outcome model and the propensity score model are correctly specified. We vary the value of β_1 in the data generating model (5.1).

		$n = 300$				$n = 1000$			
	β_1	0	0.33	0.67	1	0	0.33	0.67	1
Bias	True	−0.43	−0.01	1.15	−5.19	0.00	0.09	−2.43	9.99
	GLM	−0.18	−0.86	0.15	−4.32	−0.04	0.02	0.32	11.15
	GAM	−0.74	−4.60	−15.55	−35.38	−0.19	−1.16	−2.85	−6.86
	DR	0.08	−1.04	−3.41	−8.32	0.18	−0.56	−2.14	−4.50
	CBPS	−0.05	−0.09	0.54	−0.27	0.04	0.04	0.20	0.45
	oCBPS	−0.04	0.03	0.07	0.06	0.04	0.06	0.16	0.08
Std Dev	True	29.52	39.46	72.56	138.33	15.73	22.36	38.18	88.33
	GLM	4.45	12.31	63.35	144.25	2.21	5.49	22.93	114.45
	GAM	4.31	14.91	43.08	100.16	2.06	5.22	21.27	51.96
	DR	2.39	2.57	4.25	8.06	1.20	1.29	1.76	3.32
	CBPS	2.39	2.35	2.66	15.94	1.24	1.26	1.27	1.45
	oCBPS	2.26	2.16	2.27	2.39	1.20	1.20	1.18	1.22
RMSE	True	29.52	39.46	72.57	138.43	15.73	22.36	38.26	88.89
	GLM	4.46	12.34	63.35	144.32	2.21	5.49	22.93	114.99
	GAM	4.37	15.60	45.81	106.23	2.07	5.35	21.46	52.41
	DR	2.39	2.77	5.45	11.58	1.21	1.41	2.77	5.59
	CBPS	2.39	2.35	2.72	15.94	1.24	1.26	1.29	1.52
	oCBPS	2.26	2.16	2.27	2.39	1.20	1.20	1.19	1.23
Coverage Probability (of the 95% C.I.)	True	0.936	0.938	0.922	0.948	0.962	0.942	0.926	0.948
	GLM	0.946	0.946	0.946	0.946	0.944	0.954	0.954	0.958
	GAM	0.704	0.310	0.090	0.028	0.754	0.382	0.108	0.048
	DR	0.928	0.876	0.576	0.278	0.960	0.906	0.562	0.268
	CBPS	0.944	0.944	0.944	0.944	0.960	0.958	0.958	0.968
	oCBPS	0.950	0.964	0.962	0.982	0.956	0.954	0.962	0.966

CBPS in all the cases as well. The efficiency improvement is consistent with Corollary 3.2. The coverage probabilities of True, GLM, CBPS and oCBPS are close to the nominal level. However, GAM yields much lower coverage probability, partly because the estimates of the propensity score from logistic series are unstable. The pattern becomes more evident as β_1 increases, corresponding to the setting that the propensity score can be close to 0 or 1. Similarly, the coverage probability of DR also deteriorates as β_1 increases.

Table 5.2: Correct Outcome Model with a Misspecified Propensity Score Model.

		$n = 300$				$n = 1000$			
	β_1	0	0.33	0.67	1	0	0.33	0.67	1
Bias	True	0.00	2.13	0.08	4.79	−1.28	−0.36	1.83	3.62
	GLM	0.41	−6.67	−18.84	−32.15	0.19	−6.33	−19.21	−32.96
	GAM	15.61	3.11	−7.16	−20.76	4.07	0.28	−4.98	−14.11
	DR	−0.29	−0.68	−1.89	−3.60	−0.21	−0.39	−1.23	−2.75
	CBPS	0.84	−0.05	−2.06	−2.44	0.06	−0.79	−2.74	−3.28
	oCBPS	−0.20	−0.02	−0.13	0.07	−0.04	0.03	0.01	−0.05
Std Dev	True	45.43	36.03	39.77	77.26	26.32	19.36	39.15	88.45
	GLM	11.23	12.66	15.73	26.82	2.17	5.32	8.61	10.92
	GAM	19.91	9.40	8.81	16.18	4.29	2.87	4.14	8.52
	DR	3.35	2.57	2.52	3.16	1.42	1.27	1.28	1.57
	CBPS	3.21	2.74	3.18	3.61	1.25	1.41	1.74	2.04
	oCBPS	2.26	2.30	2.28	2.34	1.24	1.26	1.24	1.29
RMSE	True	45.43	36.10	39.77	77.40	26.36	19.36	39.20	88.52
	GLM	11.24	14.31	24.55	41.86	2.18	8.27	21.05	34.72
	GAM	25.30	9.90	11.35	26.32	5.91	2.89	6.47	16.48
	DR	3.37	2.65	3.15	4.79	1.44	1.33	1.78	3.16
	CBPS	3.32	2.74	3.79	4.36	1.26	1.62	3.24	3.86
	oCBPS	2.27	2.30	2.29	2.34	1.24	1.26	1.24	1.29
Coverage Probability (of the 95% C.I.)	True	0.952	0.936	0.964	0.972	0.946	0.950	0.960	0.988
	GLM	0.964	0.898	0.740	0.834	0.948	0.714	0.300	0.346
	GAM	0.236	0.434	0.286	0.066	0.356	0.648	0.178	0.042
	DR	0.882	0.904	0.822	0.596	0.908	0.938	0.788	0.392
	CBPS	0.956	0.978	0.924	0.914	0.944	0.928	0.742	0.654
	oCBPS	0.946	0.944	0.952	0.944	0.950	0.950	0.954	0.954

We further evaluate our method by considering different cases of misspecification for the outcome and propensity score models. We begin with the case where the outcome model is linear like before but the propensity score is misspecified. While we use the model given in equation (5.1) when estimating the propensity score, the actual treatment is generated according to the following different model,

$$\mathbb{P}(T_i = 1 \mid \boldsymbol{X} = \boldsymbol{x}_i) \quad = \quad \frac{\exp(-\beta_1 x^*_{i1} + 0.5x^*_{i2} - 0.25x^*_{i3} - 0.1x^*_{i4})}{1 + \exp(-\beta_1 x^*_{i1} + 0.5x^*_{i2} - 0.25x^*_{i3} - 0.1x^*_{i4})},$$

with $x^*_{i1} = \exp(x_{i1}/3)$, $x^*_{i2} = x_{i2}/\{1+\exp(x_{i1})\}+10$, $x^*_{i3} = x_{i1}x_{i3}/25+0.6$, and $x^*_{i4} = x_{i1}+x_{i4}+20$ where β_1 again varies from 0 to 1. In other words, the model misspecification is introduced using nonlinear transformations. Table 5.2 shows the results for this case. As expected from the double robustness property shown in Theorem 3.1, we find that the bias for the oCBPS becomes significantly smaller than all the other estimators. The oCBPS also dominates the other estimators in terms of efficiency and maintains the desired coverage probability.

We also consider the case when the propensity score is locally misspecified with the equation (2.1). In the case, we use (5.1) as the working model $\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)$, set $\xi = n^{-1/2}$ as in Theorem 2.1 and choose the function $u(\boldsymbol{X}_i; \boldsymbol{\beta}) = X^2_{i1}$ as the direction of misspecification. We compute the true propensity score from the model (2.1) and use it to generate the treatment variables. We note that sometimes the true propensity score may exceed 1. In this case we simply replace its value with 0.95. The results are given in Table 5.3. oCBPS dominates all the other estimators in terms of bias, standard deviation and root mean square error, but CBPS and DR are also noticeably better than True, GLM, and GAM.

We next examine the cases where the outcome model is misspecified. We do this by generating potential outcomes from the following quadratic model

$$\begin{aligned}
\mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i = \boldsymbol{x}_i) \quad &= \quad 200 + 27.4x^2_{i1} + 13.7x^2_{i2} + 13.7x^2_{i3} + 13.7x^2_{i4}, \\
\mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}_i) \quad &= \quad 200 + 13.7x^2_{i2} + 13.7x^2_{i3} + 13.7x^2_{i4},
\end{aligned}$$

whereas the propensity score model is the same as the one in (5.1) with β_1 varying from 0 to 0.4. Table 5.4 shows the results when the outcome model is misspecified but the propensity score model is correct. We find that the magnitude of bias is similar across all estimators with the exception of GAM and DR, which seem to have a significantly larger bias. The DR dominates in terms of

Table 5.3: Correctly Specified Outcome with a Locally Misspecified Propensity Score Model.

		$n = 300$				$n = 1000$			
	β_1	0	0.33	0.67	1	0	0.33	0.67	1
Bias	True	−1.96	0.69	0.80	4.87	0.04	0.87	−0.42	3.07
	GLM	−16.73	8.43	5.85	19.96	8.55	0.84	4.65	21.07
	GAM	−8.19	7.68	−4.35	−10.79	4.62	−0.25	−0.63	2.95
	DR	0.43	0.34	−0.83	−3.67	0.38	0.08	−1.39	−3.50
	CBPS	−0.76	−2.15	0.56	1.34	−1.92	−0.34	0.22	0.37
	oCBPS	−0.41	0.05	0.10	0.06	−0.05	0.02	−0.01	−0.02
Std Dev	True	41.03	33.16	41.86	82.09	20.65	18.39	28.44	59.63
	GLM	67.79	9.55	23.67	72.99	9.43	3.23	13.86	81.20
	GAM	46.08	8.92	21.56	52.34	11.06	2.91	11.78	52.31
	DR	3.10	2.51	2.87	5.74	1.37	1.29	1.59	2.60
	CBPS	3.26	2.56	2.44	2.77	1.58	1.28	1.33	1.43
	oCBPS	2.47	2.24	2.25	2.26	1.29	1.22	1.26	1.29
RMSE	True	41.07	33.17	41.87	82.24	20.65	18.41	28.44	59.70
	GLM	69.82	12.74	24.39	75.67	12.73	3.34	14.62	83.89
	GAM	46.80	11.77	21.99	53.44	11.98	2.92	11.80	52.39
	DR	3.13	2.53	2.99	6.81	1.42	1.29	2.11	4.36
	CBPS	3.35	3.34	2.51	3.07	2.49	1.32	1.34	1.48
	oCBPS	2.50	2.24	2.26	2.27	1.29	1.22	1.26	1.29
Coverage Probability (of the 95% C.I.)	True	0.962	0.948	0.962	0.938	0.934	0.946	0.954	0.942
	GLM	0.804	0.788	0.888	0.916	0.652	0.936	0.918	0.910
	GAM	0.132	0.294	0.238	0.076	0.154	0.612	0.144	0.052
	DR	0.856	0.922	0.866	0.556	0.916	0.936	0.736	0.332
	CBPS	0.912	0.914	0.926	0.958	0.752	0.954	0.954	0.952
	oCBPS	0.916	0.946	0.936	0.954	0.950	0.948	0.958	0.954

Table 5.4: Misspecified Outcome Model with Correct Propensity Score Model.

		$n = 300$				$n = 1000$			
	β_1	0	0.13	0.27	0.4	0	0.13	0.27	0.4
Bias	True	−4.37	−0.03	−4.24	1.51	0.80	−1.00	2.31	2.67
	GLM	0.38	−0.64	−2.67	−1.33	0.11	−0.44	0.05	0.75
	GAM	−2.03	−5.49	−10.43	−13.66	−0.65	−1.72	−1.95	−3.04
	DR	−2.77	−5.06	−9.92	−14.36	−2.98	−4.98	−7.43	−10.11
	CBPS	0.07	−0.69	−2.59	−3.94	0.05	−0.55	−0.71	−1.63
	oCBPS	−0.56	−0.97	−3.05	−4.37	−0.03	−0.68	−0.84	−1.70
Std Dev	True	49.87	58.75	74.32	100.35	27.61	33.62	44.75	53.58
	GLM	18.12	24.87	34.83	56.17	9.68	12.37	18.45	31.16
	GAM	17.59	23.19	34.72	49.87	9.07	11.36	16.85	26.50
	DR	14.02	14.65	15.58	16.65	7.95	8.26	8.21	8.40
	CBPS	15.51	17.60	18.83	20.66	8.74	9.47	10.64	12.05
	oCBPS	14.74	16.15	17.13	18.55	8.44	9.03	9.68	10.87
RMSE	True	50.06	58.75	74.45	100.36	27.62	33.64	44.81	53.60
	GLM	18.13	24.88	34.93	56.18	9.68	12.37	18.45	31.17
	GAM	17.71	23.83	36.25	51.71	9.09	11.49	16.96	26.67
	DR	14.29	15.50	18.47	21.99	8.49	9.65	11.07	13.15
	CBPS	15.51	17.62	19.01	21.03	8.74	9.49	10.66	12.16
	oCBPS	14.75	16.18	17.40	19.06	8.44	9.06	9.72	11.00
Coverage Probability (of the 95% C.I.)	True	0.948	0.954	0.946	0.920	0.938	0.950	0.910	0.922
	GLM	0.896	0.852	0.870	0.868	0.908	0.862	0.816	0.802
	GAM	0.912	0.832	0.676	0.476	0.932	0.846	0.690	0.516
	DR	0.930	0.910	0.838	0.716	0.924	0.874	0.794	0.688
	CBPS	0.920	0.870	0.790	0.676	0.914	0.862	0.776	0.668
	oCBPS	0.950	0.930	0.908	0.904	0.954	0.920	0.902	0.862

standard deviation, but oCBPS closely follows. In terms of the root mean square error, oCBPS is on par with DR.

Table 5.5: Misspecified Outcome with Misspecified Propensity Score Models.

		$n = 300$				$n = 1000$			
	β_1	0	0.13	0.27	0.4	0	0.13	0.27	0.4
Bias	True	0.54	−1.74	1.71	−3.56	−2.66	−2.52	−2.06	−0.36
	GLM	2.94	−1.70	−8.47	−20.25	−0.18	−2.07	−8.89	−18.79
	GAM	20.74	12.05	3.42	−8.06	4.95	2.35	−1.03	−5.01
	DR	9.16	6.66	4.52	0.46	6.55	4.91	2.80	0.36
	CBPS	9.57	4.10	0.37	−7.62	0.46	−0.81	−4.94	−11.18
	oCBPS	2.51	−0.24	−1.62	−4.82	0.04	−0.61	−2.29	−4.54
Std Dev	True	59.12	55.64	54.16	58.35	34.79	31.31	28.41	31.62
	GLM	25.00	19.44	22.49	26.01	9.67	9.79	11.17	12.44
	GAM	30.85	23.01	19.46	21.72	10.23	9.53	9.19	9.25
	DR	15.18	15.14	13.71	13.60	7.86	7.85	7.69	7.70
	CBPS	26.74	18.65	19.74	18.92	9.16	9.11	9.36	9.63
	oCBPS	16.28	15.38	15.08	14.42	8.93	8.60	8.32	8.27
RMSE	True	59.12	55.66	54.19	58.45	34.89	31.42	28.48	31.62
	GLM	25.18	19.51	24.03	32.96	9.67	10.00	14.28	22.53
	GAM	37.17	25.97	19.76	23.17	11.37	9.81	9.25	10.52
	DR	17.73	16.54	14.43	13.60	10.24	9.26	8.19	7.71
	CBPS	28.40	19.10	19.75	20.40	9.17	9.15	10.59	14.76
	oCBPS	16.47	15.38	15.17	15.20	8.93	8.62	8.63	9.43
Coverage Probability (of the 95% C.I.)	True	0.952	0.940	0.936	0.952	0.936	0.940	0.952	0.916
	GLM	0.854	0.902	0.866	0.788	0.890	0.878	0.772	0.540
	GAM	0.714	0.810	0.860	0.832	0.868	0.902	0.916	0.834
	DR	0.878	0.920	0.934	0.946	0.876	0.906	0.936	0.946
	CBPS	0.866	0.892	0.890	0.866	0.894	0.888	0.852	0.670
	oCBPS	0.940	0.964	0.926	0.934	0.944	0.942	0.926	0.894

Finally, when both the outcome and propensity score models are misspecified, we observe that DR and oCBPS dominate all other estimators with respect to all three criteria. In particular, oCBPS performs much better than CBPS in all scenarios. The results are organized in Table 5.5.

In summary, the proposed oCBPS method outperforms the CBPS method with respect to root

mean square error (RMSE) under all five scenarios we examined. In addition, the oCBPS method often yields better or at least comparable results relative to all the other estimators.

5.2 An Empirical Application

We next apply the oCBPS methodology to a well-known study where the experimental benchmark estimate is available. Specifically, LaLonde (1986) conducted a study, in which after the randomized evaluation study was implemented, the experimental control group is replaced with a set of untreated individuals taken from the Panel Study of Income Dynamics. This created an artificial observational study with 297 treated observations and 2,490 control observations. Ever since the original study, this data set has been used for evaluating whether a new statistical methodology can recover the experimental benchmark estimate (see e.g., Dehejia and Wahba, 1999; Smith and Todd, 2005). In the original CBPS article, Imai and Ratkovic (2014) use this data set to show that the propensity score matching estimator based on the CBPS method outperforms the matching estimator based on the standard logistic regression. In the following, we evaluate whether the proposed oCBPS method can further improve the CBPS methodology.

We begin by replicating the original results of Imai and Ratkovic (2014) and then compare those results with those of the proposed oCBPS methodology. To do this, we focus on the estimation of the average treatment effect for the treated (ATT). The response of interest is earnings in 1978 and the treatment variable is whether or not the individual participates the job training program. The original randomized experiment yields the ATT estimate \$886, which is used as a benchmark for the later comparison. Imai and Ratkovic (2014) consider the propensity score estimation based on the standard logistic regression (GLM), the just-identified CBPS with moment balance condition only (CBPS1) and the over-identified CBPS with score equation and moment balance condition (CBPS2). Based on each set of these estimated propensity scores, we estimate the ATT using the 1-to-1 nearest neighbor matching with replacement. The estimates of standard errors are based on the results in Abadie and Imbens (2006). We then add the estimated propensity score based on the proposed oCBPS methodology. Since the quantity of interest is the ATT, we use a slightly modified oCBPS estimator described in Appendix H.

We follow the propensity score model specifications examined in Imai and Ratkovic (2014). The covariates we adjust include age, education, race (white, black or Hispanic), marriage status,

Table 5.6: The bias and standard errors (shown in parentheses) of estimates of the average treatment effect for the treated in the LaLonde's Study. We use the benchmark $886 as the true value.

	GLM	CBPS1	CBPS2	oCBPS
Linear	-1190.92	-462.7	-702.33	-306.01
	(1437.02)	(1295.19)	(1240.79)	(1662.02)
Quadratic	-1808.16	-646.54	207.13	-370.03
	(1382.38)	(1284.13)	(1567.33)	(1773.03)
Smith & Todd	-1620.49	-1154.07	-462.24	-383.12
	(1424.57)	(1711.66)	(1404.15)	(1748.87)

high school degree, earnings in 1974 and earnings in 1975 as pretreatment variables. We consider three different specification of balance conditions: the first moment of covariates (Linear), the first and second moment of covariates (Quadratic), and the Quadratic specification with some interactions selected by Smith and Todd (2005) (Smith & Todd). We compare the performance of each methodology across these three specifications.

The results are shown in Table 5.6. We find that although the standard error is relatively large as in any evaluation study based on the LaLonde data, the proposed oCBPS method yields much smaller bias than GLM and CBPS1 under all three specifications. The oCBPS also improves the CBPS2 under the linear and Smith & Todd's specifications of covariates. We note that the standard error of the oCBPS method appears to be larger than the competing methods. This may be due to the fact that the uncertainty of the estimated propensity score is ignored when we calculate the standard error of the matching estimators (i.e., GLM, CBPS1 and CBPS2). In summary, consistent with the theoretical results, the proposed oCBPS method yields more accurate estimates of ATT than the original CBPS estimator or the standard logistic regression. Finally, it is important to note that these results are only suggestive since we do not know whether the assumptions of propensity score methods hold in this study.

6 Conclusion

This paper presents a theoretical investigation of the covariate balancing propensity score methodology that others have found work well in practice (e.g., Wyss et al., 2014; Frölich et al., 2015). We derive the optimal choice of the covariate balancing function so that the resulting IPTW estimator is first order unbiased under local misspecification of the propensity score model. Furthermore, it turns out that the CBPS-based IPTW estimator with the same covariate balancing function attains the semiparametric efficiency bound.

Given these theoretical insights, we propose an optimal CBPS methodology by carefully choosing the covariate balancing estimating functions. We prove that the proposed oCBPS-based IPTW estimator is doubly robust and locally efficient. More importantly, we show that the rate of convergence of the proposed estimator is faster than the standard AIPW estimator under locally misspecified models. To relax the parametric assumptions and improve the double robustness property, we further extend the oCBPS method to the nonparametric setting. We show that the proposed estimator can achieve the semiparametric efficiency bound without imposing parametric assumptions on the propensity score and outcome models. The theoretical results require weaker technical conditions than existing methods and the estimator has smaller asymptotic bias. Our simulation and empirical studies confirm the theoretical results, demonstrating the advantages of the proposed oCBPS methodology.

In this work, we mainly focus on the theoretical development of the IPTW estimator with the propensity score estimated by the optimal CBPS approach. It is a very interesting research problem to establish the theoretical results for the matching estimators combined with the optimal CBPS approach or some variants. While the asymptotic theory (i.e., consistency and asymptotic normality) for the estimated propensity score via the optimal CBPS approach can be derived from the current results (by the Delta method), the full development is beyond the scope of this work. We leave it for a future study.

Supplementary Material

The supplementary material contains the Appendix of this paper which collects the proofs and further technical details.

References

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *econometrica* **74** 235–267.

Andrews, D. W. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica* 307–345.

Arcones, M. A. (1995). A bernstein-type inequality for u-statistics and u-processes. *Statistics & probability letters* **22** 239–247.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–973.

Belloni, A., Chernozhukov, V., Chetverikov, D. and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics* **186** 345–366.

Benkeser, D., Carone, M., van der Laan, M. J. and Gilbert, P. B. (2017). Doubly-robust nonparametric inference on the average treatment effect. *Biometrika* **104** 863–880.

Bickel, P. J., Klaassen, C. A., Ritov, Y., Wellner, J. A. et al. (1998). *Efficient and adaptive estimation for semiparametric models*. Springer-Verlag.

Cao, W., Tsiatis, A. A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* asp033.

Chan, K. C. G., Yam, S. C. P. and Zhang, Z. (2016). Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society, Series B, Methodological* **78** 673–700.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics* **6** 5549–5632.

Copas, J. and Eguchi, S. (2005). Local model uncertainty and incomplete-data bias. *Journal of the Royal Statistical Society, Series B (Methodological)* **67** 459–513.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association* **94** 1053–1062.

Fong, C., Hazlett, C. and Imai, K. (2018a). Covariate balancing propensity score for a con-

tinuous treatment: Application to the efficacy of political advertisements. *Annals of Applied Statistics* **12** 156–177.

Fong, C., Ratkovic, M. and Imai, K. (2018b). CBPS: R package for covariate balancing propensity score. available at the Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=CBPS.

Frölich, M., Huber, M. and Wiesenfarth, M. (2015). The finite sample performance of semi- and nonparametric estimators for treatment effects and policy evaluation. Tech. rep., IZA Discussion Paper No. 8756.

Graham, B. S., Pinto, C. and Egel, D. (2012). Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies* **79** 1053–1079.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 315–331.

Hainmueller, J. (2012). Entropy balancing for causal effects: Multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* **20** 25–46.

Han, P. and Wang, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika* ass087.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054.

Henmi, M. and Eguchi, S. (2004). A paradox concerning nuisance parameters and projected estimating functions. *Biometrika* **91** 929–941.

Hirano, K., Imbens, G. and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1307–1338.

Horowitz, J. L., Mammen, E. et al. (2004). Nonparametric estimation of an additive model with a link function. *The Annals of Statistics* **32** 2412–2443.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47** 663–685.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **76** 243–263.

Imai, K. and Ratkovic, M. (2015). Robust estimation of inverse probability weights for marginal structural models. *Journal of the American Statistical Association* **110** 1013–1023.

Imbens, G. W., Newey, W. K. and Ridder, G. (2007). Mean-square-error calculations for average treatment effects. *Technical Report* .

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 574–580.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review* 604–620.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* **79** 147–168.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics* **4** 2111–2245.

Ning, Y., Peng, S. and Imai, K. (2018). Robust estimation of causal effects via high-dimensional covariate balancing propensity score. *arXiv preprint arXiv:1812.08683* .

Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC, New York.

Qin, J. and Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 101–122.

Robins, J., Sued, M., Lei-Gomez, Q. and Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when inverse probability weights are highly variable. *Statistical Science* **22** 544–559.

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

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90** 106–121.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.

Rothe, C. and Firpo, S. (2013). Semiparametric estimation and inference using doubly robust moment conditions. *Technical Report* .

Rotnitzky, A., Lei, Q., Sued, M. and Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika* **99** 439–456.

Rubin, D. B. (1990). Comments on "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. *Statistical Science* **5** 472–480.

Smith, J. A. and Todd, P. E. (2005). Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of econometrics* **125** 305–353.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association* **101** 1619–1637.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* **97** 661–682.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571* .

van der Laan, M. J. (2010). Targeted maximum likelihood based causal inference: Part i. *The International Journal of Biostatistics* **6**.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, vol. 3. Cambridge university press.

Vermeulen, K. and Vansteelandt, S. (2015). Bias-reduced doubly robust estimation. *Journal of the American Statistical Association* **110** 1024–1036.

Wyss, R., Ellis, A. R., Brookhart, M. A., Girman, C. J., Funk, M. J., LoCasale, R. and Stürmer, T. (2014). The role of prediction modeling in propensity score estimation: An evaluation of logistic regression, bCART, and the covariate-balancing propensity score. *American Journal of Epidemiology* **180** 645–655.

Zhao, Q. (2019). Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics* **47** 965–993.

Zhao, Q. and Percival, D. (2017). Primal-dual covariate balance and minimal double robustness via entropy balancing. *Journal of Causal Inference* **5**.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association* **110** 910–922.

Supplementary Material

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 $\boldsymbol{\beta}^o = \text{argmin}_{\boldsymbol{\beta}} \mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}))^\top \mathbf{W}^* \mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}))$ is unique.
3. $\boldsymbol{\beta}^o \in \text{int}(\Theta)$, 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 $\boldsymbol{\beta} \in \text{int}(\Theta)$.

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 $|r_{\boldsymbol{\beta}}(T,Y,\boldsymbol{X})| \le 2|Y|/c_0$ 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

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

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

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

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

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

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 $b_i(T_i, \boldsymbol{X}_i, Y_i(1), Y_i(0)) = T_i Y_i(1)/\pi_i^o - (1-T_i)Y_i(0)/(1-\pi_i^o) - \mu$. 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 $\boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} = (\boldsymbol{\Sigma}_{1\mu\boldsymbol{\beta}}^\top, \boldsymbol{\Sigma}_{2\mu\boldsymbol{\beta}}^\top)^\top$, where

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

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

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

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

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

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

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

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

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

Denote $c_i(T_i, \boldsymbol{X}_i, Y_i(1), Y_i(0)) = T_i Y_i(1)/\pi_i^o - (1-T_i)Y_i(0)/(1-\pi_i^o) - \mu$. 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,

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

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

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

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

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

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

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

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 $\bar{\mathbf{\Omega}}^*$ as

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

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

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

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 $\|\mathbf{Z}_k\|_2 \le R_n$ 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 $t \ge 0$,

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

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

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

where C is some positive constant. Moreover,

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

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 $\mathbb{P}(\|\sum_{k=1}^n \mathbf{Z}_k\|_2 \ge t) \le 2K\exp(-C'\log K)$ 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 $\mathbb{B}(r) = \{\boldsymbol{\beta} \in \mathbb{R}^K : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \le r\}$, 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 $C > 0$,

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

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

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

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 $II \leq 0$. 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 $E_\epsilon = \Theta \cap \{\|\boldsymbol{\beta} - \boldsymbol{\beta}^o\|_2 \geq \epsilon\}$. 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 $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_2 \leq \epsilon$ 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

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

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

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

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 $Q(\boldsymbol{\beta}^o) \leq Q(\boldsymbol{\beta}^*) = O(K^{1-2r_b})$. 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 $\mathcal{C} = \{\boldsymbol{\Delta} \in \mathbb{R}^K : \|\boldsymbol{\Delta}\|_2 \leq r\}$, 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 $v \le \sup_{\boldsymbol{x}\in\mathcal{X}} |m^*(\boldsymbol{x})|$, then there exist some constants $C, C' > 0$, suc that for any $m_1 + 1 \le i \le K$,

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

Similarly, $|\xi_i| \le 2C'K^{-r_b}$ for any $1 \le i \le m_1$. 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

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

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

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

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

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

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 $\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}} - \mu$,

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

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 $\delta\to0$, 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,

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

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, $\int(\widehat{K}(\boldsymbol{x})-K(\boldsymbol{x}))^2P(d\boldsymbol{x})$) 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,

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

Thus, a simple estimator of τ_1^* is

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

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

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

We first consider I_2. The mean value theorem implies

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

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

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

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

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

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.