Improving Covariate Balancing Propensity Score: A Doubly Robust and Efficient Approach [*]

Jianqing Fan[*‡] Kosuke Imai[†‡] Han Liu[*‡] Yang Ning[*] Xiaolin Yang[*†]

Princeton University
[*] Department of Operations Research and Financial Engineering
[†] Department of Politics [‡] Center for Statistics Machine Learning

June 14, 2016

Abstract

Inverse probability of treatment weighting (IPTW) is a popular method for estimating causal effects in many disciplines. However, empirical studies show that the IPTW estimators can be sensitive to the misspecification of propensity score model. To address this problem, several researchers have proposed new methods 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 their theoretical properties. This paper makes two main contributions. First, we conduct a theoretical investigation of one such methodology, the Covariate Balancing Propensity Score (CBPS) recently proposed by Imai and Ratkovic (2014). We characterize the asymptotic bias and efficiency of the CBPS-based IPTW estimator under both arbitrary and local model misspecification as well as correct specification for general balancing functions. Based on this finding, we address an open problem in the literature on how to optimally choose the covariate balancing function for the CBPS methodology. Second, motivated by the form of the optimal covariate balancing function, we further propose a new IPTW estimator by generalizing the CBPS method. We prove that the proposed estimator is consistent if either the propensity score model or the outcome model is correct. In addition to this double robustness property, we also establish that the proposed estimator is semiparametrically efficient when both the propensity score and outcome models are correctly specified. Unlike the standard doubly robust estimators, however, the proposed methodology does not require the estimation of outcome model. To relax the parametric assumptions on the propensity score model and the outcome model, we further consider a sieve estimation approach to estimate the treatment effect. A new "nonparametric double robustness" phenomenon is observed. Our simulations show that the proposed estimator has better finite sample properties than the standard estimators.

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

[*] Supported by NSF grants DMS-1206464 and DMS-1406266 and NIH grants R01-GM072611-12 and R01-GM100474-04.

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 also assume the strong ignorability of the treatment assignment (Rosenbaum and Rubin, 1983):

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

Furthermore, without loss of generality, we assume that the conditional expectation functions of potential outcomes can be written as,

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

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),

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

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 (Robins et al., 1994),

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

Despite its popularity, researchers have found that the estimators based on IPTW are particularly sensitive to the misspecification of propensity score model (e.g., Kang and Schafer, 2007). To overcome this problem, several researchers have recently proposed to estimate the propensity score by optimizing covariate balance rather than maximizing the accuracy of predicting treatment assignment (e.g., Tan, 2010; Hainmueller, 2012; Graham et al., 2012; Imai and Ratkovic, 2014; Chan et al., 2015). In this paper, we focus on the Covariate Balancing Propensity Score (CBPS) methodology proposed by 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) and general treatment regimes (Fong et al., 2015). 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 these related methods.

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

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

for some covariate balancing function $\mathbf{f}(\cdot) : \mathbb{R}^d \to \mathbb{R}^m$, if 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 $\mathbf{f}(\boldsymbol{X}_i) = \pi'_{\boldsymbol{\beta}}(\boldsymbol{X}_i) = \frac{\partial \pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)}{\partial \boldsymbol{\beta}}$. 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 choose $\mathbf{f}(\boldsymbol{X}_i) = (\boldsymbol{X}_i \ \boldsymbol{X}_i^2)$ 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. If $m > q$, then $\widehat{\boldsymbol{\beta}}$ can be estimated by optimizing the covariate balance by the generalized method of moments (GMM) method (Hansen, 1982):

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

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; Fong et al., 2015). Once the estimate of $\boldsymbol{\beta}$ is obtained, we can estimate the ATE using the IPTW estimator in (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. Although some researchers have proposed the nonparametric propensity score estimators to alleviate this problem (e.g., Hirano et al., 2003; Chan et al., 2015), these methods are potentially difficult to apply in practice even when the number of pre-treatment covariates is moderate.

This paper makes two main contributions. First, we conduct a thorough theoretical study of the CBPS-based IPTW estimator with the general balancing function $\mathbf{f}(\cdot)$. We characterize the asymptotic bias of the CBPS-based IPTW estimator under both arbitrary and local misspecification of the propensity score model. We also study the efficiency of the IPTW estimator under correct specification and local model misspecification. 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). In particular, we show that once the covariate balancing function is chosen in this way, the CBPS-based IPTW estimator achieves a double robustness property: the estimator is consistent if either the propensity score model or the outcome model is correct.

However, the optimal choice of $\mathbf{f}(\boldsymbol{X}_i)$ requires the knowledge of the propensity score, which is unknown. Thus, the application of the CBPS method with the optimal $\mathbf{f}(\boldsymbol{X}_i)$ is limited in practice. To address this issue, our second contribution is to propose a new IPTW estimator by generalizing the CBPS method. We show that the IPTW estimator based on the improved CBPS (iCBPS) method retains the double robustness property. We also show that the proposed estimator is semiparametrically efficient when both the propensity score and outcome models are correctly specified (Section 3). Different from the CBPS method with the optimal $\mathbf{f}(\boldsymbol{X}_i)$, the proposed iCBPS method does not require the knowledge of the propensity score and is easy to implement in practice. In addition, unlike the standard doubly robust estimators (Robins et al., 1994), the proposed iCBPS method does not require the estimation of outcome model. Our simulation study shows that the proposed estimator outperforms the standard estimators (Section 3.2).

To relax the parametric assumptions on the propensity score model and the outcome model, we further extend the proposed iCBPS method to the nonparametric/semiparametric settings, by using a sieve estimation approach (Newey, 1997; Chen, 2007). In Section 4, we establish a unified semiparametric efficiency result for the IPTW estimator under many nonparametric/semiparametric settings, including the fully nonparametric model, additive model and partially linear model. Our result provides a more comprehensive theoretical framework than the existing nonparametric literature (e.g., Hirano et al., 2003; Chan et al., 2015), which usually assumes the propensity score model is fully nonparametric and therefore suffers from the curse of dimensionality. In addition, our theoretical results require weaker technical assumptions. For instance, in the fully nonparametric setting, the theory in Hirano et al. (2003) and Chan et al. (2015) requires $s/d > 7$ and $s/d > 13$, respectively, where s is the smoothness parameter of some 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). Similar tools from recent random matrix theory have been used by Hansen (2014); Chen and Christensen (2015); Belloni et al. (2015) to study the optimal rate of convergence for sieve-based least square estimation. However, unlike the sieve-based least square estimator, our sieve estimator of the propensity score does not have a closed form and this leads to extra technical challenges for the establishment of the consistency and rate of convergence. Finally, in this nonparametric setting, we observe an interesting phenomenon that our unified semiparametric efficiency result holds if either the propensity score model or the outcome model is approximated reasonably well. This can be viewed as a nonparametric version of the double robustness property. To the best of our knowledge, this phenomenon does not appear in the existing literature. The last section provides concluding remarks.

2 Consequences of Model Misspecification

Our theoretical investigation starts by examining the consequences of model misspecification for the CBPS-based IPTW estimator. For this, we first derive the asymptotic bias of the IPTW estimator under local misspecification of the propensity score model and show that we can eliminate the bias by carefully choosing the covariate balancing function $\mathbf{f}(\boldsymbol{X}_i)$ such that it spans $K(\boldsymbol{X}_i) + (1 - \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))L(\boldsymbol{X}_i)$, where $\boldsymbol{\beta}^o$ is the limiting value of $\widehat{\boldsymbol{\beta}}$, i.e., $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^o$. In other words, we want to choose the covariate balancing function $\mathbf{f}(\boldsymbol{X}_i)$ such that it spans the weighted average of the two conditional mean functions of potential outcomes, i.e., there exists an $\boldsymbol{\alpha} \in \mathbb{R}^m$ such that $\boldsymbol{\alpha}^\top \mathbf{f}(\boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)\mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i) + (1 - \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))\mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i)$. This result is further extended to arbitrary model misspecification.

The above result implies that when balancing covariates, for any given unit we should give a greater weight to the determinants of the 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 contrast, if a unit is more likely to be assigned to the treatment group, then the covariates that determine the potential outcome under the control condition become more important. We also show that even when the propensity score is correctly specified, this choice of covariate balancing function is optimal, enabling the resulting estimator to attain the semiparametric efficiency bound.

2.1 Bias under Model Misspecification

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 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 ξ represents the magnitude of misspecification. We assume $\xi = o(1)$ and $\xi^{-1} n^{-1/2} = O(1)$, as $n \to \infty$ so that the true propensity score $\pi(\boldsymbol{X}_i)$ is in a local neighborhood of the working model $\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)$. Under this local model misspecification setting, we derive the asymptotic bias and variance of the CBPS-based IPTW estimator in (1.5). The next theorem gives the expression of the asymptotic bias.

Theorem 2.1 (Asymptotic Bias under Local Misspecification)**.** If the propensity score model is locally misspecified as in (2.1), under Assumption A.1 in Appendix A, the bias of the IPTW estimator defined in (1.5) is given by $\mathbb{E}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}) - \mu = B\xi + o(\xi)$, where

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

and $\mathbf{W}^*$ is the limiting value of $\widehat{\mathbf{W}}$ in (1.7),

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

This theorem shows that the estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ has a first order bias term $B\xi$ under the local model misspecification. Although the expression of B looks sophisticated, the next corollary shows how to choose the covariate balancing function to eliminate the first order bias.

Corollary 2.1 (Optimal Choice of Covariate Balancing Function under Local Misspecification)**.** Suppose that we choose the covariate balancing function $\mathbf{f}(\boldsymbol{X})$ such that $\boldsymbol{\alpha}^\top\mathbf{f}(\boldsymbol{X}) = K(\boldsymbol{X}) + (1 - \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}))L(\boldsymbol{X})$ holds, where $\boldsymbol{\alpha}\in\mathbb{R}^m$ is a vector of arbitrary constants. In addition, assume that the number of parameters is equal to the dimension of covariate balancing function $\mathbf{f}(\boldsymbol{X}_i)$, i.e., $m = q$. Then, under the conditions in Theorem 2.1, the IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ given in equation (1.5) is first order unbiased, i.e., $B = 0$.

In the following, we consider the asymptotic bias of the IPTW estimator when the propensity score model is arbitrarily misspecified. We first describe our results in a heuristic way. Assume that the propensity score model is misspecified, i.e., $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi(\boldsymbol{X}_i) \neq \pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)$ for any $\boldsymbol{\beta}\in\Theta$. Let $\boldsymbol{\beta}^o$ represent the limiting value of the CBPS estimator $\widehat{\boldsymbol{\beta}}$, i.e., assuming $\widehat{\boldsymbol{\beta}}\xrightarrow{p}\boldsymbol{\beta}^o$. Then, the asymptotic bias of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is given by $\mathbb{E}(\widehat{\mu}_{\boldsymbol{\beta}^o}) - \mu$ where μ is the true ATE defined in (1.3). The CBPS method ensures that $\boldsymbol{\beta}^o$ satisfies

$$\begin{aligned}
\mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^o}(\boldsymbol{T},\boldsymbol{X})) &= \mathbb{E}\left\{\left(\frac{T_i}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)} - \frac{1-T_i}{1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)}\right)\mathbf{f}(\boldsymbol{X}_i)\right\}\\
&= \mathbb{E}\left\{\frac{\pi(\boldsymbol{X}_i)-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)(1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))}\mathbf{f}(\boldsymbol{X}_i)\right\} = 0. \qquad (2.3)
\end{aligned}$$

Note that $\mathbb{E}(\widehat{\mu}_{\boldsymbol{\beta}^o})$ can be written as

$$\mathbb{E}(\widehat{\mu}_{\boldsymbol{\beta}^o}) = \mathbb{E}\left\{\frac{T_iY_i}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)} - \frac{(1-T_i)Y_i}{1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)}\right\} = \mathbb{E}\left\{\frac{T_i(K(\boldsymbol{X}_i)+L(\boldsymbol{X}_i))}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)} - \frac{(1-T_i)K(\boldsymbol{X}_i)}{1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)}\right\},$$

where the second equality follows from the strong ignorability of treatment assignment and the law of iterated expectation. Therefore, the asymptotic bias of the IPTW estimator is

$$\mathbb{E}(\widehat{\mu}_{\boldsymbol{\beta}^o}) - \mu = \mathbb{E}\left[\left(\frac{T_i}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)} - \frac{1-T_i}{1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)}\right)\left\{K(\boldsymbol{X}_i) + (1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))L(\boldsymbol{X}_i)\right\}\right].$$

Thus, by equation (2.3), we can eliminate the asymptotic bias of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ under arbitrary model misspecification by choosing the covariate balancing function $\mathbf{f}(\boldsymbol{X}_i)$ such that $\boldsymbol{\alpha}^\top\mathbf{f}(\boldsymbol{X}_i) = K(\boldsymbol{X}_i) + (1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))L(\boldsymbol{X}_i)$ holds for some $\boldsymbol{\alpha}\in\mathbb{R}^m$. In other words, $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ remains consistent for μ even if the propensity score model is misspecified. Under regularity conditions in Appendix A, this result can be proved by applying the similar argument in Theorem 3.1 of Section 3. For simplicity, we refer the details to the proof of Theorem 3.1 in Appendix C. Finally, we note that the choice of the covariate balancing function under arbitrary model misspecification is the same as that under local misspecification, and hence extends the result in Corollary 2.1 under local model misspecification to the setting of arbitrary misspecification.

2.2 Efficiency Consideration

We next study how the choice of different covariate balancing functions affects the efficiency of the IPTW estimator. We first consider the case where the propensity score model is correctly specified. We further show that the efficiency result also applies to the case of local misspecification studied above.

Let $\boldsymbol{\beta}^*$ be the true value of the parameter $\boldsymbol{\beta}$ in the propensity score model $\pi_{\boldsymbol{\beta}}(\boldsymbol{X}_i)$. We first derive the asymptotic distribution of the CBPS-based IPTW estimator under correctly specified propensity score model. In this case, the estimator is asymptotically unbiased and follows the normal distribution regardless of the choice of covariate balancing function. The asymptotic variance of the estimator, however, depends on this choice.

Theorem 2.2 (Asymptotic Properties under Correct Specification)**.** Suppose that the propensity score model is correctly specified and $\widehat{\boldsymbol{\beta}}$ is obtained through equation (1.7). Let μ be the true treatment effect and $\mathbf{W}^*$ be the limiting value of $\widehat{\mathbf{W}}$ in (1.7). Let

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

Under Assumption A.1 in Appendix A, $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ satisfies

$$\sqrt{n}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu) \xrightarrow{d} N(0,\ \bar{\boldsymbol{H}}^{*\top}\boldsymbol{\Sigma}\bar{\boldsymbol{H}}^*), \tag{2.4}$$

where $\bar{\boldsymbol{H}}^* = (1, \boldsymbol{H}_y^{*\top})$ and

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

with

$$\Sigma_\mu = \mathrm{Var}\big(\mu_{\boldsymbol{\beta}^*}(T_i, Y_i, \boldsymbol{X}_i)\big) \quad = \quad \mathbb{E}\bigg(\frac{Y_i(1)^2}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)} + \frac{Y_i(0)^2}{1-\pi^*_{\boldsymbol{\beta}}(\boldsymbol{X}_i)} - (\mathbb{E}(Y_i(1)) - \mathbb{E}(Y_i(0)))^2\bigg),$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = (\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\boldsymbol{H}_{\mathbf{f}}^*)^{-1}\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\,\mathrm{Var}(\boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i, \boldsymbol{X}_i))\mathbf{W}^*\boldsymbol{H}_{\mathbf{f}}^*(\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\boldsymbol{H}_{\mathbf{f}}^*)^{-1},$$

$$\boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} = -(\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\boldsymbol{H}_{\mathbf{f}}^*)^{-1}\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\,\mathrm{Cov}(\mu_{\boldsymbol{\beta}^*}(T_i, Y_i, \boldsymbol{X}_i), \boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i, \boldsymbol{X}_i)),$$

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

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

To better understand this result, consider a special case where the dimension of the covariate balancing function is equal to the number of parameters to be estimated, i.e., $m = q$ (Imai and Ratkovic, 2014). In this case, we can solve the optimization problem in (1.7) by setting $\widehat{\mathbf{W}}$ to a diagonal matrix. Then, the asymptotic variance of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is equal to,

$$\begin{aligned}\mathrm{Var}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}) \quad \approx \quad & \mathrm{Var}(\widehat{\mu}_{\boldsymbol{\beta}^*}) + \boldsymbol{H}_y^{*\top}\boldsymbol{H}_{\mathbf{f}}^{*-1}\,\mathrm{Var}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*}(\boldsymbol{T}, \boldsymbol{X}))\boldsymbol{H}_{\mathbf{f}}^{*-1}\boldsymbol{H}_y^* \\ & -2\boldsymbol{H}_y^{*\top}\boldsymbol{H}_{\mathbf{f}}^{*-1}\,\mathrm{Cov}(\widehat{\mu}_{\boldsymbol{\beta}^*}, \bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*}(\boldsymbol{T}, \boldsymbol{X})). \qquad (2.5)\end{aligned}$$

The expression (2.5) contains three parts. The first term $\mathrm{Var}(\widehat{\mu}_{\boldsymbol{\beta}^*})$ is the variance of the estimator under the true value $\boldsymbol{\beta}^*$. The second term is the variance of the balancing equation $\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ under $\boldsymbol{\beta}^*$ scaled by the quadratic term $\boldsymbol{H}_y^{*\top}\boldsymbol{H}_{\mathbf{f}}^{*-1}$. The third term is the covariance between the $\widehat{\mu}_{\boldsymbol{\beta}^*}$ and $\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*}(\boldsymbol{T}, \boldsymbol{X})$ scaled by $\boldsymbol{H}_y^{*\top}\boldsymbol{H}_{\mathbf{f}}^{*-1}$.

Based on the asymptotic variance in equation (2.5), the next corollary shows that the optimal choice of covariate balancing function derived before also results in an IPTW estimator that is semiparametrically efficient.

Corollary 2.2 (Optimal Choice of Covariate Balancing Function under Correct Specification)**.** Choose any covariate balancing function $\mathbf{f}(\boldsymbol{X})$ such that $\boldsymbol{\alpha}^\top\mathbf{f}(\boldsymbol{X}) = K(\boldsymbol{X}) + (1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}))L(\boldsymbol{X})$ holds, for some constant $\boldsymbol{\alpha} \in \mathbb{R}^m$. In addition, assume that the number of parameters is equal to the dimension of covariate balancing function $\mathbf{f}(\boldsymbol{X}_i)$, i.e., $m = q$. Then, under the conditions in Theorem 2.2, the IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ given in equation (1.5) attains the semiparametric asymptotic variance bound in Theorem 1 of Hahn (1998), i.e.,

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

We note that there may exist many choices for $\mathbf{f}(\boldsymbol{X})$, which satisfy $\boldsymbol{\alpha}^\top\mathbf{f}(\boldsymbol{X}) = K(\boldsymbol{X}) + (1 - \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}))L(\boldsymbol{X})$ for some $\boldsymbol{\alpha} \in \mathbb{R}^m$. This corollary implies that, provided $\mathbf{f}(\boldsymbol{X})$ satisfies the above condition, the asymptotic variance of our IPTW estimator dose not depend on the particular choice of $\mathbf{f}(\boldsymbol{X})$ and is the smallest among the class of regular estimators.

Finally, we comment that this efficiency result under correctly specified model also carries over to the locally misspecified case examined earlier. In Appendix B.5, we show that under the locally misspecified propensity score model in (2.1), the IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ satisfies $\sqrt{n}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu) \xrightarrow{d} N(B,\ \bar{\boldsymbol{H}}^{*\top}\boldsymbol{\Sigma}\bar{\boldsymbol{H}}^{*})$, where B is the first order bias given in equation (2.2) of Theorem 2.1. Thus, together with Corollary 2.1, the aforementioned choice of covariate balancing function, i.e., $\boldsymbol{\alpha}^\top \mathbf{f}(\boldsymbol{X}_i) = K(\boldsymbol{X}_i) + (1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))L(\boldsymbol{X}_i)$, yields an asymptotically unbiased and efficient estimator of the ATE under local misspecification.

In summary, the theoretical analysis presented in this section has shown that the optimal covariate balancing function $\mathbf{f}(\boldsymbol{X}_i)$ for the CBPS methodology needs to satisfy $\boldsymbol{\alpha}^\top \mathbf{f}(\boldsymbol{X}_i) = K(\boldsymbol{X}_i) + (1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))L(\boldsymbol{X}_i)$, which leads to an asymptotically unbiased and efficient estimator of the ATE under various scenarios. Recall that $K(\boldsymbol{X}_i)$ is the conditional mean of the potential outcome under the control group and $L(\boldsymbol{X}_i)$ is the conditional average treatment effect. Both $K(\cdot)$ and $L(\cdot)$ can be estimated by imposing additional parametric assumptions. However, the optimal choice of $\mathbf{f}(\boldsymbol{X}_i)$ also requires the knowledge of the propensity score model $\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)$, where the limiting value $\boldsymbol{\beta}^o$ of $\widehat{\boldsymbol{\beta}}$ depends on the choice of $\mathbf{f}(\boldsymbol{X}_i)$ itself. Thus, to construct the optimal covariate balancing function $\mathbf{f}(\boldsymbol{X}_i)$, one needs some prior knowledge of $\mathbf{f}(\boldsymbol{X}_i)$ to estimate the propensity score model in (1.6). This "chicken-and-egg" relationship between the optimal covariate balancing function $\mathbf{f}(\boldsymbol{X}_i)$ and propensity score model $\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)$ makes the implementation of the existing CBPS method with the optimal $\mathbf{f}(\boldsymbol{X}_i)$ difficult in practice. To address this issue, in the following section we propose a new IPTW estimator by generalizing the CBPS method, such that the optimal covariate balancing function does not depend on the knowledge of the propensity score model $\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)$.

3 The Improved CBPS Methodology

Recall that the optimal covariate balancing function $\mathbf{f}(\boldsymbol{X}_i)$ for the CBPS methodology needs to satisfy $\boldsymbol{\alpha}^\top \mathbf{f}(\boldsymbol{X}_i) = K(\boldsymbol{X}_i) + (1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))L(\boldsymbol{X}_i)$. Notice that the asymptotic bias of the IPTW estimator with the optimal covariate balancing function $\mathbf{f}(\boldsymbol{X}_i)$ can be decomposed into two terms in the following manner,

$$\begin{aligned}
\mathbb{E}(\widehat{\mu}_{\boldsymbol{\beta}^o}) - \mu &= \mathbb{E}\left[\left(\frac{T_i}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)} - \frac{1-T_i}{1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)}\right)\left\{K(\boldsymbol{X}_i) + (1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))L(\boldsymbol{X}_i)\right\}\right] \\
&= \mathbb{E}\left[\left(\frac{T_i}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)} - \frac{1-T_i}{1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)}\right)K(\boldsymbol{X}_i) + \left(\frac{T_i}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)} - 1\right)L(\boldsymbol{X}_i)\right]. \qquad (3.1)
\end{aligned}$$

Our main idea is to minimize the magnitudes of both terms to eliminate the bias of the IPTW estimator. Motivated by this observation, we propose to balance the first term and second term separately. This leads to 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 $\bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) = n^{-1}\sum_{i=1}^n \boldsymbol{g}_{1\boldsymbol{\beta}}(\boldsymbol{T}_i, \boldsymbol{X}_i)$ and $\bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) = n^{-1}\sum_{i=1}^n \boldsymbol{g}_{2\boldsymbol{\beta}}(\boldsymbol{T}_i, \boldsymbol{X}_i)$ with

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

for some 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 $m_1 + m_2 = m$. Note that, as seen in the asymptotic bias of the IPTW estimator in (3.1), $\boldsymbol{h}_1(\boldsymbol{X}_i)$ in $\bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ aims to recover the conditional mean function of the potential outcome under the control condition, i.e., $K(\boldsymbol{X}_i)$, whereas $\boldsymbol{h}_2(\boldsymbol{X}_i)$ in $\bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ aims to recover the conditional mean function of the treatment effect, i.e., $L(\boldsymbol{X}_i)$. It is easily seen that $\bar{\boldsymbol{g}}_{1\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ 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. More importantly, unlike the existing CBPS method, we introduce a new 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 "covariate-imbalance" estimating function $\bar{\boldsymbol{g}}_{2\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$ is the key to remove the dependence of the optimal covariate balancing function $\mathbf{f}(\boldsymbol{X}_i)$ on the propensity score model $\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)$. Thus, our method is an improved version of the CBPS method (iCBPS). Given the estimating functions in (3.2), we can estimate $\boldsymbol{\beta}$ by the GMM estimator $\widehat{\boldsymbol{\beta}}$ in (1.7). Similarly, the ATE is estimated by the IPTW estimator $\hat{\mu}_{\widehat{\boldsymbol{\beta}}}$ in (1.5). The implementation of the proposed iCBPS method (e.g., the choice of $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$) will be discussed in later sections.

3.1 Theoretical Properties

We now derive the theoretical properties of the IPTW estimator given in (1.5) based on the proposed iCBPS method. In particular, we will show that the proposed estimator is doubly robust and semiparametrically 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. 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. $\boldsymbol{\beta}^o \in \mathrm{int}(\Theta)$, where Θ is a compact set.
4. $\pi_{\boldsymbol{\beta}}(\boldsymbol{X})$ is continuous in $\boldsymbol{\beta}$.
5. There exists a constant $0 < c_0 < 1/2$ such that with probability tending to one, $c_0 \le \pi_{\boldsymbol{\beta}}(\boldsymbol{X}) \le 1 - c_0$, for any $\boldsymbol{\beta} \in \mathrm{int}(\Theta)$.
6. $\mathbb{E}|Y(1)|^2 < \infty$ and $\mathbb{E}|Y(0)|^2 < \infty$.
7. $\mathbf{G}^* := \mathbb{E}(\partial \boldsymbol{g}(\boldsymbol{\beta}^o)/\partial \boldsymbol{\beta})$ 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 $\sup_{\boldsymbol{\beta}\in\mathbb{B}_r(\boldsymbol{\beta}^o)} |\partial \pi_{\boldsymbol{\beta}}(\boldsymbol{X})/\partial \beta_k| \le C_k(\boldsymbol{X})$ for $1 \le k \le q$, and $\mathbb{E}(|h_{1j}(\boldsymbol{X})| C_k(\boldsymbol{X})) < \infty$ for $1 \le j \le m_1$, $1 \le k \le q$ and $\mathbb{E}(|h_{2j}(\boldsymbol{X})| C_k(\boldsymbol{X})) < \infty$ for $1 \le j \le m_2$, $1 \le k \le q$, 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). Condition 5 is commonly used in the missing data and causal inference literature, which essentially says the propensity score cannot be too close to 0 and 1 (Robins et al., 1994, 1995). Conditions 6-7 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 $C(\boldsymbol{X})$ 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 iCBPS-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., $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)$;

2. The functions $K(\cdot)$ and $L(\cdot)$ 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 $\mathbf{M}_2 \in \mathbb{R}^{q\times m_2}$ are the partitions of $\mathbf{G}^{*\top}\mathbf{W}^* = (\mathbf{M}_1, \mathbf{M}_2)$. That is $K(\cdot) \in \operatorname{span}\{\mathbf{M}_1\boldsymbol{h}_1(\cdot)\}$ and $L(\cdot) \in \operatorname{span}\{\mathbf{M}_2\boldsymbol{h}_2(\cdot)\}$.

Theorem 3.1 implies that the proposed estimator is consistent if either the propensity score model or the outcome model is correctly specified. In particular, the second condition can be rewritten as $K(\boldsymbol{X}_i) = \boldsymbol{\alpha}_1^\top\mathbf{M}_1\boldsymbol{h}_1(\boldsymbol{X}_i)$ and $L(\boldsymbol{X}_i) = \boldsymbol{\alpha}_2^\top\mathbf{M}_2\boldsymbol{h}_2(\boldsymbol{X}_i)$, for some vectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathbb{R}^q$. Hence, the functions $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$ play very different roles in the proposed iCBPS methodology. Specifically, $\mathbf{M}_1\boldsymbol{h}_1(\cdot)$ serves as the basis functions for the conditional baseline effect $K(\cdot)$ while $\mathbf{M}_2\boldsymbol{h}_2(\cdot)$ represents the basis functions for the conditional treatment effect $L(\cdot)$.

Next, we establish the asymptotic normality of the proposed estimator if either the propensity score model or the outcome model is correctly specified. For this result, we require an additional set of regularity conditions.

Assumption 3.2. The following regularity conditions are assumed.

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 $C(\boldsymbol{X})$ 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 iCBPS-based IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ has the following asymptotic distribution:

$$\sqrt{n}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu) \xrightarrow{d} N\left(0,\ \bar{\mathbf{H}}^{*\top}\boldsymbol{\Sigma}\bar{\mathbf{H}}^*\right), \tag{3.3}$$

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

$$\begin{aligned}
\mathbf{H}^* &= -\mathbb{E}\left(\frac{K(\boldsymbol{X}_i) + (1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))L(\boldsymbol{X}_i)}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)(1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))} \cdot \frac{\partial \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}{\partial \boldsymbol{\beta}}\right), \\
\boldsymbol{\Sigma} &= \begin{pmatrix} \Sigma_\mu & \boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}}^\top \\ \boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} & \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \end{pmatrix}, \quad \text{with } \Sigma_\mu = \mathbb{E}\left(\frac{Y_i^2(1)}{\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)} + \frac{Y_i^2(0)}{1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}\right) - \mu^2. \qquad (3.4)
\end{aligned}$$

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

$$\boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} = -(\mathbf{G}^{*\top}\mathbf{W}^*\mathbf{G}^*)^{-1}\mathbf{G}^{*\top}\mathbf{W}^*\mathbb{E}\left(\frac{K(\boldsymbol{X}_i) + (1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))L(\boldsymbol{X}_i)}{(1-\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i))\pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)}\boldsymbol{h}(\boldsymbol{X}_i)\right),$$

where $\boldsymbol{h}(\boldsymbol{X}_i) = (\boldsymbol{h}_1^\top(\boldsymbol{X}_i), \boldsymbol{h}_2^\top(\boldsymbol{X}_i))^\top$.

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

$$\sqrt{n}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu) \xrightarrow{d} N\left(0,\ \bar{\mathbf{H}}^{*\top}\boldsymbol{\Sigma}\bar{\mathbf{H}}^*\right), \qquad (3.5)$$

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

$$\mathbf{H}^* = -\mathbb{E}\left[\left\{\frac{\pi(\boldsymbol{X}_i)(K(\boldsymbol{X}_i) + L(\boldsymbol{X}_i))}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)^2} + \frac{(1-\pi(\boldsymbol{X}_i))K(\boldsymbol{X}_i)}{(1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))^2}\right\}\frac{\partial \pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)}{\partial \boldsymbol{\beta}^o}\right],$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_\mu & \boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}}^\top \\ \boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} & \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \end{pmatrix} \text{ with } \Sigma_\mu = \mathbb{E}\left(\frac{\pi(\boldsymbol{X}_i)Y_i^2(1)}{\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i)^2} + \frac{(1-\pi(\boldsymbol{X}_i))Y_i^2(0)}{(1-\pi_{\boldsymbol{\beta}^o}(\boldsymbol{X}_i))^2}\right) - \mu^2.$$

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

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

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

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

3. If both Conditions 1 and 2 of Theorem 3.1 hold and $\mathbf{W}^* = \boldsymbol{\Omega}^{-1}$, then the proposed iCBPS-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

$$V = \Sigma_\mu - (\boldsymbol{\alpha}_1^\top \mathbf{M}_1, \boldsymbol{\alpha}_2^\top \mathbf{M}_2)\mathbf{G}^*(\mathbf{G}^{*\top}\boldsymbol{\Omega}^{-1}\mathbf{G}^*)^{-1}\mathbf{G}^{*\top}\begin{pmatrix} \mathbf{M}_1^\top \boldsymbol{\alpha}_1 \\ \mathbf{M}_2^\top \boldsymbol{\alpha}_2 \end{pmatrix} \tag{3.6}$$

and Σ_μ is defined in (3.4).

The asymptotic variance V in (3.6) 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 its true value $\boldsymbol{\beta}^*$. The second term can be interpreted as the effect of estimating $\boldsymbol{\beta}^*$ via covariate balancing 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., $V \le \Sigma_\mu$. In particular, Henmi and Eguchi (2004) offer 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 additional covariate balancing conditions into $\boldsymbol{h}_1(\cdot)$ and $\boldsymbol{h}_2(\cdot)$. However, the following corollary shows that when both the propensity score and outcome models are correctly specified, 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 balancing 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 iCBPS-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 Condition 3 of Theorem 3.1, we have $V \le \bar{V}$, where V is defined in (3.6).

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 if $\text{rank}(\mathbf{M}_1) = m_1$ and $\text{rank}(\mathbf{M}_2) = m_2$, Condition 2 of Theorem 3.1 can be written as $K(\cdot) \in \text{span}\{\boldsymbol{h}_1(\cdot)\}$ and $L(\cdot) \in \text{span}\{\boldsymbol{h}_2(\cdot)\}$. 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

$$V_{\text{opt}} \;=\; \Sigma_\mu - (\boldsymbol{\alpha}_1^\top \mathbf{M}_1, \boldsymbol{\alpha}_2^\top \mathbf{M}_2)\boldsymbol{\Omega}\begin{pmatrix} \mathbf{M}_1^\top \boldsymbol{\alpha}_1 \\ \mathbf{M}_2^\top \boldsymbol{\alpha}_2 \end{pmatrix}. \tag{3.7}$$

Comparing this expression with (3.6), we see that the proposed estimator is semiparametrically efficient if $\mathbf{G}^*$ is a square matrix (i.e., $m = q$) and invertible. In this special case, the dimension of $\boldsymbol{\beta}$ must be identical to that of the covariate balancing functions $\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X})$. We can accomplish this by over-parameterizing the propensity score model (e.g., including higher-order terms and interactions). 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 $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)$.
2. 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. That is $K(\cdot) \in \operatorname{span}\{\boldsymbol{h}_1(\cdot)\}$ and $L(\cdot) \in \operatorname{span}\{\boldsymbol{h}_2(\cdot)\}$.

In addition, under Assumption 3.2, if both conditions hold, then the proposed estimator is semiparametrically efficient with the asymptotic variance given in (3.7).

This corollary shows that the proposed estimator has two advantages over the existing CBPS estimator in Imai and Ratkovic (2014) with balancing first (and second) moment of $\boldsymbol{X}_i$ and/or the score function of the propensity score model. First, the proposed estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is doubly robust to model misspecification, whereas the existing CBPS estimator does not have this property. Second, the estimator can be more efficient than the existing CBPS estimator.

The result in Corollary 3.2 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 ensure $m = q$ as required in Corollary 3.2. This observation further motivates us to consider a nonparametric/semiparametric approach to improve the iCBPS method, which will be shown in Section 4.

Remark 3.1. Robins et al. (1994) propose the following estimator of the ATE with the double robustness and semiparametric efficiency properties

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

where $K(\boldsymbol{X}_i,\boldsymbol{\alpha})$ and $L(\boldsymbol{X}_i,\boldsymbol{\gamma})$ are some parametric models with finite dimensional parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. Unlike the projection approach behind this classical doubly robust estimator $\widehat{\mu}_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}}$ (Tsiatis, 2007), the proposed iCBPS-based IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is constructed by a new covariate balancing method. From a practical perspective, one needs to plug consistent estimators of $(\boldsymbol{\alpha},\boldsymbol{\gamma})$ and $\boldsymbol{\beta}$ (e.g., usually MLE) into $\widehat{\mu}_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}}$ to estimate the treatment effect. In contrast, the proposed method is easier to implement, which avoids the estimation of the parameters in the conditional mean models $K(\boldsymbol{X}_i,\boldsymbol{\alpha})$ and $L(\boldsymbol{X}_i,\boldsymbol{\gamma})$.

Remark 3.2 (Implementation of iCBPS). 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 such that $m = q$ holds. The resulting iCBPS-based IPTW estimator may reduce bias under model misspecification and attain high efficiency.

Table 3.1: Correct Outcome Model with Correct Propensity Score Model.

		$n = 300$				$n = 1000$			
	β_1	0	0.33	0.67	1	0	0.33	0.67	1
Bias	True	0.19	2.48	−6.14	−1.97	−0.42	1.77	−1.71	3.39
	GLM	0.77	1.13	−4.80	−15.61	0.64	0.46	0.08	−2.61
	CBPS	0.70	0.59	−1.17	−5.28	0.68	0.65	0.03	−2.02
	iCBPS	0.61	0.63	−0.86	−4.50	0.70	0.73	0.23	−1.28
Std Dev	True	27.17	39.61	84.09	213.13	15.53	21.58	45.50	228.89
	GLM	5.52	17.69	65.45	163.59	2.44	8.06	34.77	108.03
	CBPS	3.36	3.66	4.54	7.40	1.79	1.96	2.36	3.84
	iCBPS	3.20	3.26	3.38	5.12	1.72	1.80	1.85	2.50
RMSE	True	27.17	39.69	84.31	213.14	15.54	21.65	45.53	228.91
	GLM	5.58	17.73	65.63	164.34	2.52	8.07	34.77	108.07
	CBPS	3.43	3.71	4.69	9.09	1.91	2.06	2.36	4.34
	iCBPS	3.26	3.32	3.49	6.82	1.86	1.94	1.87	2.80

3.2 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

$$\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{\exp(-\beta_1 x_{i1} + 0.5x_{i2} - 0.25x_{i3} - 0.1x_{i4})}{1 + \exp(-\beta_1 x_{i1} + 0.5x_{i2} - 0.25x_{i3} - 0.1x_{i4})}, \tag{3.8}$$

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. Each covariate is generated independently from $N(3, 2)$. Each set of results is based on 1,000 Monte Carlo simulations.

We examine the performance of the IPTW estimator when the propensity score model is fitted using the maximum likelihood (GLM), the standard CBPS with balancing the first moment (CBPS), and the proposed improvement of CBPS (iCBPS) as well as the case where the true propensity score (True), i.e., $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, is used for the IPTW estimator. In the first set of simulations, we use correctly specified propensity score model. Table 3.1 shows the standard deviation, bias, root mean square error (RMSE) of these estimators when the sample size is $n = 300$ and $n = 1000$. We find that CBPS and iCBPS substantially outperform both True and GLM in terms of efficiency. In addition, iCBPS is more efficient than CBPS in all cases though both have similar biases in most cases. These findings are consistent with Corollary 3.2.

Table 3.2: Correct Outcome Model with Misspecified Propensity Score Model.

		$n = 300$				$n = 1000$			
	β_1	0	0.33	0.67	1	0	0.33	0.67	1
Bias	True	0.41	−0.38	−2.28	−2.57	0.46	0.02	1.03	−6.59
	GLM	1.76	−24.12	−57.51	−72.73	0.93	−25.34	−53.49	−74.57
	CBPS	1.77	−0.94	−4.73	−6.69	0.74	−2.51	−6.41	−8.32
	iCBPS	0.50	0.41	−0.63	−1.95	0.63	0.30	−0.60	−1.87
Std Dev	True	47.60	41.63	70.40	310.96	25.80	21.12	54.64	90.80
	GLM	9.72	26.53	161.18	48.66	2.67	13.65	23.74	30.76
	CBPS	4.14	4.15	4.84	5.36	1.79	2.08	2.76	3.45
	iCBPS	3.42	3.11	3.18	3.20	1.75	1.76	1.82	1.74
RMSE	True	47.60	41.64	70.44	310.97	25.80	21.12	54.65	91.04
	GLM	9.88	35.86	171.13	87.50	2.83	28.78	58.52	80.67
	CBPS	4.50	4.26	6.77	8.57	1.94	3.26	6.97	9.01
	iCBPS	3.45	3.14	3.25	3.75	1.86	1.78	1.92	2.55

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 (3.8) 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 3.2 shows the results for this case. As expected from the double robustness property shown in Theorem 3.1, we find that the bias for iCBPS becomes significantly smaller than CBPS and GLM. iCBPS also dominates the other estimators in terms of efficiency.

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

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

whereas the propensity score model is the same as the one in (3.8) with β_1 varying from 0 to 0.4. Table 3.3 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 though iCBPS has the least bias. CBPS and iCBPS again dominate GLM in terms of efficiency with iCBPS having the smallest standard deviation. Finally, we consider the case where both the outcome and propensity score models are misspecified. Table 3.4 shows the results. We find that iCBPS outperforms all the other estimators in both bias and efficiency.

Table 3.3: 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	1.11	4.65	1.19	0.75	2.33	3.46	2.36	1.91
	GLM	2.09	2.76	2.02	-0.17	2.72	2.35	2.69	2.21
	CBPS	2.18	2.46	1.27	-0.98	2.64	2.30	2.38	1.57
	iCBPS	2.07	2.22	1.08	-0.70	2.57	2.28	2.38	1.56
Std Dev	True	38.57	42.90	47.86	57.09	21.35	22.57	27.24	31.66
	GLM	14.36	17.72	23.99	32.35	7.41	8.67	12.53	17.82
	CBPS	12.12	13.68	15.11	15.76	6.52	6.94	8.32	9.44
	iCBPS	11.81	12.64	13.50	13.64	6.37	6.42	7.27	7.97
RMSE	True	38.58	43.15	47.88	57.10	21.48	22.84	27.34	31.72
	GLM	14.51	17.94	24.08	32.35	7.89	8.98	12.82	17.95
	CBPS	12.32	13.90	15.16	15.79	7.03	7.31	8.65	9.57
	iCBPS	11.99	12.83	13.54	13.66	6.87	6.82	7.65	8.12

Table 3.4: 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	1.23	1.63	1.19	0.13	0.19	1.51	0.47	1.63
	GLM	3.73	1.54	-1.58	-7.16	1.63	-0.01	-3.26	-8.46
	CBPS	8.32	6.35	4.23	1.36	1.89	0.75	-0.70	-3.17
	iCBPS	4.11	2.28	1.18	-0.32	1.45	0.50	-0.32	-1.53
Std Dev	True	60.98	57.09	52.36	50.10	33.47	30.57	27.66	28.41
	GLM	23.20	19.45	19.53	20.52	7.71	8.05	8.46	9.70
	CBPS	18.29	16.66	14.76	14.64	6.85	7.04	6.83	6.67
	iCBPS	14.05	12.27	11.95	11.64	6.69	6.77	6.49	6.14
RMSE	True	60.99	57.11	52.37	50.10	33.47	30.61	27.67	28.46
	GLM	23.50	19.51	19.59	21.73	7.88	8.05	9.07	12.87
	CBPS	20.10	17.83	15.35	14.70	7.11	7.08	6.87	7.38
	iCBPS	14.64	12.48	12.01	11.64	6.84	6.79	6.50	6.33

4 Nonparametric/Semiparametric iCBPS Method

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 $L(\cdot)$ can be well approximated by the linear spans of $\boldsymbol{h}_1(\cdot)$ and

$\boldsymbol{h}_2(\cdot)$, respectively. Therefore, the correct specification of $K(\cdot)$ and $L(\cdot)$ can be attained. However, the parametric assumption on the propensity score model $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)$ imposed in Corollary 3.2 can be restrictive and prone to be violated. Once the propensity score model is misspecified, the proposed iCBPS-based IPTW estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is inefficient and even inconsistent. To relax the strong parametric assumptions imposed in the previous sections, in what follows, we propose a flexible nonparametric/semiparametric approach for modeling the propensity score and the conditional mean functions. The main advantage of this nonparametric approach is that, the resulting iCBPS-based IPTW estimator is semiparametrically efficient under a much broader class of propensity score models and the conditional mean models than those in Corollary 3.2.

More specifically, we assume $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = J(m^*(\boldsymbol{X}_i))$, where $J(\cdot)$ is a known monotonic link function (e.g., $J(\cdot) = \exp(\cdot)/(1+\exp(\cdot))$), and $m^*(\cdot)$ is an unknown smooth function. One practical way to estimate $m^*(\cdot)$ is to approximate it by the linear combination of K basis functions, where K is allowed to grow with n. This approach is known as the sieve estimation (Andrews, 1991; Newey, 1997). In detail, let $\boldsymbol{B}(\boldsymbol{x}) = \{b_1(\boldsymbol{x}), ..., b_K(\boldsymbol{x})\}$ denote a collection of K basis functions, whose mathematical requirement is given in Assumption 4.1. Intuitively, we would like to approximate $m^*(\boldsymbol{x})$ by $\boldsymbol{\beta}^{*\top}\boldsymbol{B}(\boldsymbol{x})$, for some coefficient $\boldsymbol{\beta}^* \in \mathbb{R}^K$.

To estimate $\boldsymbol{\beta}^*$, similar to the parametric case, we define $\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}) = n^{-1}\sum_{i=1}^n \boldsymbol{g}_{\boldsymbol{\beta}}(\boldsymbol{T}_i, \boldsymbol{X}_i)$, 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 $m = K$, 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

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

The following assumptions are imposed to establish the large sample properties of $\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}}$.

Assumption 4.1. 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 $c_0 \le J(v) \le 1 - c_0$ and $0 < c_1 \le \partial J(v)/\partial v \le c_2$, 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 $|\partial^2 J(v)/\partial v^2| \le c_3$ for some constant $c_3 > 0$.

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

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

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

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 smallest 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, $\sup_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{h}(\boldsymbol{x})\|_2 \leq CK^{1/2}$, and $\sup_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{B}(\boldsymbol{x})\|_2 \leq CK^{1/2}$.

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 in 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 4.1, the following theorem establishes the asymptotic normality and semiparametric efficiency of the estimator $\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}}$.

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

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

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

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

Assume $K = o(n^{1/3})$ and $n^{\frac{1}{2(r_b+r_h)}} = o(K)$. 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 extension of Corollary 3.2. It shows that one can construct an efficient estimator of the treatment effect without imposing strong parametric assumptions on the propensity score model and the outcome model.

In the following, we comment on the technical assumptions in this theorem. We assume $m^*(\boldsymbol{x})$ 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 $\boldsymbol{h}_2(\boldsymbol{x})$), respectively. It is well known that the uniform rate of convergence is related to the smoothness of the functions $m^*(\boldsymbol{x})$ and $K(\boldsymbol{x})$ (also $L(\boldsymbol{x})$) and the dimension of $\boldsymbol{X}$. For instance, if the function class $\mathcal{M}$ for $m^*(\boldsymbol{x})$ and $\mathcal{H}$ for $K(\boldsymbol{x})$ (also $L(\boldsymbol{x})$) correspond to the Hölder class with smoothness parameter s on the domain $\mathcal{X} = [0,1]^d$, then the uniform convergence holds for the spline basis and wavelet basis with $r_b = r_h = s/d$; see Newey (1997); Chen (2007) for details. Thus, Theorem 4.1 is directly applicable to this setting in which no additional structure assumptions are imposed on $m^*(\boldsymbol{x})$ and $K(\boldsymbol{x})$ (also $L(\boldsymbol{x})$). Compared to the nonparametric result in Hirano et al. (2003) which requires $s/d > 7$ and Chan et al. (2015) which requires $s/d > 13$, our theorem needs a much weaker condition, i.e., $s/d > 3/4$. To achieve this improved result, our proof exploits the matrix Bernstein's concentration inequalities (Tropp, 2015) and a Bernstein-type concentration inequality for U-statistics (Arcones, 1995).

When d is large, to handle the curse of dimensionality and meanwhile retain the flexibility of the model, we may assume the following additive structure $m^*(\boldsymbol{x}) = \sum_{j=1}^{d} m_j^*(x_j)$; see Stone (1985); Hastie and Tibshirani (1990); Horowitz et al. (2004); Fan and Jiang (2005), among others. One major advantage of the sieve estimation approach is that it is convenient to incorporate such structural assumption and the results in Theorem 4.1 are still applicable. In this case, the uniform convergence holds with $r_b = s$, where s is a common smoothness parameter of the Hölder class for all $\{m_j^*(x)\}_{j=1,\ldots,d}$. In particular, if $m^*(\boldsymbol{x})$ and $K(\boldsymbol{x})$ (also $L(\boldsymbol{x})$) all have the additive structures and each unknown component lies in the Hölder class with $s = 1$, then Theorem 4.1 holds provided $K = o(n^{1/3})$ and $n^{1/4} = o(K)$. Similar to the additive model, our method can be easily extended to the partially linear model for $m^*(\boldsymbol{x})$ and $K(\boldsymbol{x})$ (also $L(\boldsymbol{x})$), by redefining the basis functions $\boldsymbol{B}(\boldsymbol{x})$ and $\boldsymbol{h}(\boldsymbol{x})$. As discussed in Section 5 of Newey (1997), the large sample results (e.g., our Theorem 4.1) of the sieve estimator are directly applicable to the partially linear case. Thus, our Theorem 4.1 provides a unified semiparametric efficiency result under many nonparametric/semiparametric settings, which is more comprehensive than the existing fully nonparametric propensity score methods in Hirano et al. (2003); Chan et al. (2015).

Interestingly, we find that the asymptotic bias of the estimator $\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}}$ is proportional to $K^{-(r_b+r_h)}$, which is the product of the approximation errors for $m^*(\boldsymbol{x})$ and $K(\boldsymbol{x})$ (also $L(\boldsymbol{x})$). Thus, to make the bias of the estimator $\widetilde{\mu}_{\widetilde{\boldsymbol{\beta}}}$ asymptotically ignorable, we may require either r_b or r_h sufficiently large, due to the multiplicative form of the asymptotic bias. This phenomenon can be viewed as a nonparametric version of the double robustness property, in the sense that the asymptotic normality and semiparametric efficiency results hold if either the propensity score model or the outcome model is approximated reasonably well by basis functions. To the best of our knowledge, this property does not appear in the existing literature. In fact, in this theorem, it is not necessary to assume both the propensity score model and the outcome models are approximated with very high accuracy (i.e., both r_b and r_h are large).

5 Concluding Remarks

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. We then show that this choice is also optimal under arbitrary misspecification. Furthermore, it turns out that this covariate balancing function is also optimal even under correct specification, attaining the semiparametric efficiency bound.

Given these theoretical insights, we propose an improvement to the standard CBPS methodology by carefully choosing the covariate balancing estimating functions. We prove that the proposed iCBPS-based IPTW estimator is consistent if either the outcome or propensity score model is correct. In addition to this double robustness property, the proposed estimator is semiparametrically efficient. The advantage of the proposed methodology, over the standard doubly robust estimators, is that it does not require the separate estimation of an outcome model. Our simulation results confirm the theoretical results, demonstrating the advantages of the proposed iCBPS methodology.

To relax the parametric assumptions and improve the double robustness property, we further extend the iCBPS method to the nonparametric/semiparametric settings. We show that the proposed estimator can achieve the semiparametric efficiency bound without imposing parametric assumptions on the propensity score model and the outcome model. A new nonparametric version of the double robustness property is discovered. Finally, we note that our theoretical results require much weaker technical conditions and cover a broader class of semiparametric models than the existing nonparametric propensity score methods.

References

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.

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.

Chan, K., Yam, S. and Zhang, Z. (2015). Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society, Series B, Methodological* Forthcoming.

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

Chen, X. and Christensen, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics* **188** 447–465.

Fan, J. and Jiang, J. (2005). Nonparametric inferences for additive models. *Journal of the American Statistical Association* **100** 890–907.

Fong, C., Hazlett, C. and Imai, K. (2015). Covariate balancing propensity score for general treatment regimes. Tech. rep., Department of Politics, Princeton University.

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.

Hansen, B. E. (2014). A unified asymptotic distribution theory for parametric and non-parametric least squares. Tech. rep., Working paper, University of Wisconsin.

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

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, vol. 43. CRC Press.

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.

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.

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.

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

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.

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.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The annals of Statistics* 689–705.

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* .

Tsiatis, A. (2007). *Semiparametric theory and missing data.* Springer Science & Business Media.

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.

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.

A 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 A.1. (Regularity Conditions for CBPS in Section 2)

1. There exists a positive definite matrix $\mathbf{W}^*$ such that $\widehat{\mathbf{W}} \xrightarrow{p} \mathbf{W}^*$.
2. The minimizer $\boldsymbol{\beta}^o = \mathrm{argmin}_{\boldsymbol{\beta}} \mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}))^\top \mathbf{W}^* \mathbb{E}(\bar{\boldsymbol{g}}_{\boldsymbol{\beta}}(\boldsymbol{T}, \boldsymbol{X}))$ is unique.
3. $\boldsymbol{\beta}^o \in \mathrm{int}(\Theta)$, where Θ is a compact set.
4. $\pi_{\boldsymbol{\beta}}(\boldsymbol{X})$ is continuous in $\boldsymbol{\beta}$.
5. There exists a constant $0 < c_0 < 1/2$ such that with probability tending to one, $c_0 \leq \pi_{\boldsymbol{\beta}}(\boldsymbol{X}) \leq 1 - c_0$, for any $\boldsymbol{\beta} \in \mathrm{int}(\Theta)$.
6. $\mathbb{E}|f_j(\boldsymbol{X})| < \infty$ for $1 \leq j \leq m$ and $\mathbb{E}|Y(1)|^2 < \infty$, $\mathbb{E}|Y(0)|^2 < \infty$.
7. $\mathbf{G}^* := \mathbb{E}(\partial \boldsymbol{g}(\boldsymbol{\beta}^o)/\partial \boldsymbol{\beta})$ exists and there is a q-dimensional function $C(\boldsymbol{X})$ and a small constant $r > 0$ such that $\sup_{\boldsymbol{\beta} \in \mathbb{B}_r(\boldsymbol{\beta}^o)} |\partial \pi_{\boldsymbol{\beta}}(\boldsymbol{X})/\partial \boldsymbol{\beta}| \leq C(\boldsymbol{X})$ and $\mathbb{E}(|f_j(\boldsymbol{X})| C(\boldsymbol{X})) < \infty$ for $1 \leq j \leq m$, where $\mathbb{B}_r(\boldsymbol{\beta}^o)$ is a ball in $\mathbb{R}^q$ with radius r and center $\boldsymbol{\beta}^o$. In addition, $\mathbb{E}(|Y| C(\boldsymbol{X})) < \infty$.
8. $\mathbf{G}^{*\top} \mathbf{W}^* \mathbf{G}^*$ and $\mathbb{E}(\boldsymbol{g}_{\boldsymbol{\beta}^o}(T_i, \boldsymbol{X}_i) \boldsymbol{g}_{\boldsymbol{\beta}^o}(T_i, \boldsymbol{X}_i)^\top)$ are nonsingular.
9. In the locally misspecified model (2.1), assume $|u(\boldsymbol{X}; \boldsymbol{\beta}^*)| \leq 1$ almost surely.

Lemma A.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 A.2. Under Assumption A.1 (or Assumptions 3.1), we have $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^o$. Moreover, if $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi_i^*(\boldsymbol{X}_i)$, then $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^*$.

Proof of Lemma A.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 A.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 A.1 (6). Moreover, if $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)$, in the following we show that $\boldsymbol{\beta}^* = \boldsymbol{\beta}^o$. This is because by $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)$, we have $\mathbb{E}(\boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i, \boldsymbol{X}_i)) = 0$ and $\mathbb{E}(\boldsymbol{g}_{\boldsymbol{\beta}^*}^\top(T_i, \boldsymbol{X}_i)) \mathbf{W}^* \mathbb{E}(\boldsymbol{g}_{\boldsymbol{\beta}^*}(T_i, \boldsymbol{X}_i)) = 0$. Since $\mathbf{W}^*$ is positive definite, we can see that $\mathbb{E}(\boldsymbol{g}_{\boldsymbol{\beta}}^\top(T_i, \boldsymbol{X}_i)) \mathbf{W}^* \mathbb{E}(\boldsymbol{g}_{\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i)) \geq 0$. Hence $\boldsymbol{\beta}^*$ is the minimizer of $\mathbb{E}(\boldsymbol{g}_{\boldsymbol{\beta}}^\top(T_i, \boldsymbol{X}_i)) \mathbf{W}^* \mathbb{E}(\boldsymbol{g}_{\boldsymbol{\beta}}(T_i, \boldsymbol{X}_i))$

and by the uniqueness of the minimizer we prove that $\boldsymbol{\beta}^* = \boldsymbol{\beta}^o$. In addition, for the proof of Theorem 3.1, we similarly verify the following conditions to prove this lemma for the iCBPS 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 $|g_{1\boldsymbol{\beta}j}(T_i, \boldsymbol{X}_i)| \le 2|h_{1j}(\boldsymbol{X}_i)|/c_0$ 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 A.3. Under Assumption A.1 (or Assumptions 3.1 and 3.2), we have

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

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

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

Proof. The proof of (A.1) and (A.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 A.1 (3), (4), (2) and Assumption A.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 A.1. To see this, by Assumption A.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 A.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 (A.2). As shown in Lemma A.2, if $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)$ holds, then $\boldsymbol{\beta}^o = \boldsymbol{\beta}^*$. Thus, the asymptotic normality of $n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ follows from (A.2). The proof is complete. □

B Proof of Results in Section 2

B.1 Proof of Theorem 2.1

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

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

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

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

By the dominated convergence theorem, the bias of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ can be represented by

$$\mathbb{E}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}) - \mu = \mathbb{E}(D_i) + o(n^{-1/2}).$$

By ignorability of the treatment assignment and the definition of the true propensity score,

$$\mathbb{E}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}) - \mu = \mathbb{E}\Big\{\frac{\pi_i^*(1 + \xi u_i^*)Y_i(1)}{\pi_i^o} - \frac{(1 - \pi_i^* - \xi \pi_i^* u_i^*)Y_i(0)}{1 - \pi_i^o}\Big\} - \mu + o(\xi).$$

Similarly, we expand π_i^o around π_i^*,

$$\mathbb{E}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}) - \mu = \mathbb{E}\Big\{Y_i(1) - Y_i(0) + \xi(u_i^* Y_i(1) + \frac{u_i^* \pi_i^*}{1 - \pi_i^*} Y_i(0)) + \boldsymbol{H}_y^*(\boldsymbol{\beta}^o - \boldsymbol{\beta}^*)\Big\} - \mu + o(\xi), \tag{B.2}$$

where

$$\boldsymbol{H}_y^* = -\mathbb{E}\Big\{((1-\pi_i^*)Y_i(1) + \pi_i^* Y_i(0))\mathbf{f}(\boldsymbol{X}_i)\Big\}.$$

Plugging (B.1) into above equation, the bias of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is

$$\begin{aligned}\mathbb{E}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}) - \mu = \xi\Big\{&\mathbb{E}(u_i^*Y_i(1) + \frac{u_i^*\pi_i^*}{1-\pi_i^*}Y_i(0)) \\ &- \boldsymbol{H}_y^*(\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\boldsymbol{H}_{\mathbf{f}}^*)^{-1}\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\mathbb{E}\Big\{\Big(\frac{u_i^*}{1-\pi_i^*}\Big)\mathbf{f}(\boldsymbol{X}_i)\Big\}\Big\} + o(\xi) \\ = \xi\Big\{&\mathbb{E}\Big(\frac{u_i^*(K(\boldsymbol{X}_i)+L(\boldsymbol{X}_i)(1-\pi_i^*))}{1-\pi_i^*}\Big) \\ &+ \boldsymbol{H}_y^*(\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\boldsymbol{H}_{\mathbf{f}}^*)^{-1}\boldsymbol{H}_{\mathbf{f}}^{*\top}\mathbf{W}^*\mathbb{E}\Big(\frac{u_i^*\mathbf{f}(\boldsymbol{X}_i)}{1-\pi_i^*}\Big)\Big\} + o(\xi).\end{aligned}$$

This completes the proof. □

B.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. □

B.3 Proof of Theorem 2.2

Proof. The proof is similar to that of Theorem 3.2. We omit the details. □

B.4 Proof of Corollary 2.2

Proof. 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}. □

B.5 Asymptotic Distribution under Local Misspecification

Theorem B.1. Under the locally misspecified propensity score model in (2.1) with $\xi = n^{-1/2}$ and Assumption A.1 in Appendix A, the estimator $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ in (1.5), where $\widehat{\boldsymbol{\beta}}$ is obtained by (1.7), has the following asymptotic distribution

$$\sqrt{n}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu) \quad \xrightarrow{d} \quad N(B,\ \bar{\boldsymbol{H}}^{*\top}\boldsymbol{\Sigma}\bar{\boldsymbol{H}}^*), \tag{B.3}$$

where B is the first order bias given in equation (2.2) of Theorem 2.1.

Proof. Recall that in the proof of Theorem 2.1 we obtain that

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

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, we have shown 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. The detailed calculation of the asymptotic variance is similar to Theorem 2.2, which is omitted for simplicity. □

C Proof of Results in Section 3

C.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 A.2, we have $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^*$. 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 A.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 $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^*$ and the dominated convergence theorem, we obtain that

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

where $\pi_i^* = \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)$. 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,

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

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 $K(\cdot) \in \operatorname{span}\{\mathbf{M}_1 \boldsymbol{h}_1(\cdot)\}$ and $L(\cdot) \in \operatorname{span}\{\mathbf{M}_2 \boldsymbol{h}_2(\cdot)\}$. By Lemma A.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

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

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 (C.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.

□

C.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 $\widehat{\mu} = \bar{\mu} + \widehat{\mathbf{H}}(\widetilde{\boldsymbol{\beta}})^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$, where

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

where $\pi_i^* = \pi_{\boldsymbol{\beta}^*}(\boldsymbol{X}_i)$, $\widetilde{\pi}_i = \pi_{\widetilde{\boldsymbol{\beta}}}(\boldsymbol{X}_i)$ and $\widetilde{\boldsymbol{\beta}}$ is an intermediate value between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$. 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 A.1, we have $\sup_{\boldsymbol{\beta}\in\mathbb{B}_r(\boldsymbol{\beta}^*)} |\widehat{\mathbf{H}}(\boldsymbol{\beta}) - \mathbb{E}(\widehat{\mathbf{H}}(\boldsymbol{\beta}))| = o_p(1)$. Since $\widehat{\boldsymbol{\beta}}$ is consistent, by the dominated convergence theorem we can obtain $\widehat{\mathbf{H}}(\widetilde{\boldsymbol{\beta}}) = \mathbf{H}^* + o_p(1)$, where

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

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

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

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

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

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

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

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

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

After some algebra, we can show that

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

This completes the proof of equation (3.3). Next, we consider the case (2). Recall that $\mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) = \pi(\boldsymbol{X}_i) \neq \pi_{\beta^o}(\boldsymbol{X}_i)$. Following the similar arguments, we can show that

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

where

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

and

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

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

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

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

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

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

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

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

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

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 (3.5).

Finally, we start to prove part 3. By (3.3), the asymptotic variance of $\widehat{\mu}$ denoted by V, can be written as

$$V = \Sigma_\mu + 2\mathbf{H}^{*\top}\boldsymbol{\Sigma}_{\mu\boldsymbol{\beta}} + \mathbf{H}^{*\top}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\mathbf{H}^*. \tag{C.3}$$

Note that by Lemma A.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

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

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 (C.2). In addition, $\mathbf{G}^* = (\mathbf{G}_1^{*\top}, \mathbf{G}_2^{*\top})^\top$, where

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

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

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

Comparing to the expression of $\mathbf{G}^*$ in (C.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

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

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 (C.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.

□

C.3 Proof of Corollary 3.1

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

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

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, (C.5) holds if $\mathbf{C} - \bar{\mathbf{C}}$ is positive semidefinite. Without loss of generality, we rearrange orders and write the $(m+a) \times q$ matrix $\bar{\mathbf{G}}^*$ and the $(m+a) \times (m+a)$ matrix $\bar{\boldsymbol{\Omega}}^*$ as

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

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

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

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

□

C.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. □

D 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.

Lemma D.1 (Bernstein's inequality for U-statistics (Arcones, 1995))**.** Given i.i.d. random variables $Z_1, \dots Z_n$ 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 $U \coloneqq \binom{n}{m}^{-1} \sum_{i_1 < \ldots < i_m} h(Z_{i_1}, \dots, Z_{i_m})$. Suppose that $\mathbb{E} h(Z_{i_1}, \dots, Z_{i_m}) = 0$, $\mathbb{E}\big\{\mathbb{E}[h(Z_{i_1}, \dots, Z_{i_m}) \mid Z_{i_1}]\big\}^2 = \sigma^2$ and $\|h\|_\infty \le b$. There exists a constant $K(m) > 0$ depending on m such that

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

Lemma D.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 D.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

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

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 $c_0 \le |J(\boldsymbol{\beta}^\top \boldsymbol{B}(\boldsymbol{x}))| \le 1 - c_0$ for any $\boldsymbol{\beta} \in \Theta$ and $T_i \in \{0, 1\}$, it implies $\sup_{\boldsymbol{\beta} \in \Theta} \max_{1 \le i \le n} |\xi_i(\boldsymbol{\beta})| \le C$ 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

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

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 $\Theta_m = \{\boldsymbol{\beta} \in$

$\mathbb{R}^K : \|\boldsymbol{\beta} - \boldsymbol{\beta}_m\|_2 \leq r\}$ for some $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_M$. Thus, for any given $\epsilon > 0$,

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

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

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

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

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

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 D.1 to the U-statistic with kernel function $u_{1ij}(\boldsymbol{\beta}_m)$,

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

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 $\boldsymbol{\beta} \in \Theta_m$, $|\xi_i(\boldsymbol{\beta}) - \xi_i(\boldsymbol{\beta}_m)| \leq C|(\boldsymbol{\beta} - \boldsymbol{\beta}_m)^\top\boldsymbol{B}(\boldsymbol{X}_i)| \leq CrK^{1/2}$, where the last step follows from the Cauchy-Schwarz inequalty. This further implies $|\xi_i(\boldsymbol{\beta})\xi_j(\boldsymbol{\beta}) - \xi_i(\boldsymbol{\beta}_m)\xi_j(\boldsymbol{\beta}_m)| \leq CrK^{1/2}$ for some constant $C > 0$ by performing a standard perturbation analysis. Thus,

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

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

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

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

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

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 (D.1), we complete the proof. □

Lemma D.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 $t \ge 0$,

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

Lemma D.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

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

and

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

Proof of Lemma D.4. We prove this result by applying Lemma D.3. In particular, to prove (D.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,

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

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

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

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

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

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

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

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

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

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

Then

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

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

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

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

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

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

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

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

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

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 D.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

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

where C, C' are some positive constants. Following the similar arguments to (D.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 D.3 implies $\|II\|_2 = O_p(\sqrt{K\log K/n})$. Next, we consider the term $I_{\boldsymbol{\beta}}$. Following the similar arguments to (D.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 D.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 D.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 D.6. Under the conditions in Theorem 4.1, it holds that

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

Proof of Lemma D.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

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

In the following, we study the terms I, II and III one by one. For the term I, Lemma D.2 implies $|Q(\widetilde{\boldsymbol{\beta}}) - Q_n(\widetilde{\boldsymbol{\beta}})| \le \sup_{\boldsymbol{\beta}\in\Theta}\Big|Q_n(\boldsymbol{\beta}) - Q(\boldsymbol{\beta})\Big| = o_p(1)$. This shows that $|I| = o_p(1)$ and the same argument yields $|III| = o_p(1)$. For the term II, by the definition of $\widetilde{\boldsymbol{\beta}}$, it is easy to see that $II \le 0$. Thus, combining with (D.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 D.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

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

Proof of Lemma D.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

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

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 $0 \leq v \leq 1$ 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 $\boldsymbol{\Delta} \in \partial\mathcal{C}$,

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

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 D.5 with the arguments similar to (D.14) in the proof of Lemma D.8. In addition, $\|\widetilde{\mathbf{G}}\|_2 \leq \|\widetilde{\mathbf{G}} - \mathbf{G}^*\|_2 + \|\mathbf{G}^*\|_2 \leq C$, since $\|\mathbf{G}^*\|_2$ is bounded by a constant and $\|\widetilde{\mathbf{G}} - \mathbf{G}^*\|_2 = o_p(1)$ by Lemma D.5. By the Weyl inequality and Lemma D.5,

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

for n sufficiently large. By (D.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 (D.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 $Q_n(\boldsymbol{\beta}) \geq 0$ for any $\boldsymbol{\beta}$, we obtain that $\widetilde{\boldsymbol{\beta}}$ is indeed a global minimizer of $Q_n(\boldsymbol{\beta})$. □

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

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

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 D.8. Similar to the proof of Lemma D.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 $0 \leq v \leq 1$ lies between $\boldsymbol{\beta}^*$ and $\widetilde{\boldsymbol{\beta}}$. After rearrangement, we derive

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

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 (D.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 $C, C' > 0$,

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

where the third step follows from the fact that $\|\mathbf{G}^{*-1}\|_2$ is bounded and the last step follows from Lemma D.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 (D.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 D.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 $\|\mathbf{\Delta}_{n1}\|_2$, $\|\mathbf{\Delta}_{n2}\|_2$ and $\|\mathbf{\Delta}_{n3}\|_2$ with (D.13), by Lemma D.5, we obtain

$$
\begin{aligned}
\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2 &\leq \|\mathbf{G}^{*-1}\bar{\boldsymbol{g}}_{\boldsymbol{\beta}^*}(\boldsymbol{T},\boldsymbol{X})\|_2+\|\mathbf{\Delta}_{n1}\|_2+\|\mathbf{\Delta}_{n2}\|_2+\|\mathbf{\Delta}_{n3}\|_2 \\
&\leq C\Big(\frac{K^{1/2}}{n^{1/2}}+\frac{1}{K^{r_b}}\Big)+C'\Big(\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2K^{1/2}+\sqrt{\frac{K\log K}{n}}\Big)\cdot\Big(\frac{K^{1/2}}{n^{1/2}}+\frac{1}{K^{r_b}}\Big),
\end{aligned}
$$

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

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

This completes the proof. □

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

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

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

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

where

$$
\begin{aligned}
S_i &= \frac{T_i}{\pi_i^*}\big[Y_i(1)-K(\boldsymbol{X}_i)-L(\boldsymbol{X}_i)\big]-\frac{1-T_i}{1-\pi_i^*}\big[Y_i(0)-K(\boldsymbol{X}_i)\big]+L(\boldsymbol{X}_i)-\mu, \\
R_0 &= \frac{1}{n}\sum_{i=1}^n\frac{T_i(Y_i(1)-K(\boldsymbol{X}_i)-L(\boldsymbol{X}_i))}{\widetilde{J}_i\pi_i^*}(\pi_i^*-\widetilde{J}_i), \\
R_1 &= \frac{1}{n}\sum_{i=1}^n\frac{(1-T_i)(Y_i(0)-K(\boldsymbol{X}_i))}{(1-\widetilde{J}_i)(1-\pi_i^*)}(\pi_i^*-\widetilde{J}_i), \\
R_2 &= \frac{1}{n}\sum_{i=1}^n\Big(\frac{T_i}{\widetilde{J}_i}-\frac{1-T_i}{1-\widetilde{J}_i}\Big)\Delta_K(\boldsymbol{X}_i),\quad R_3=\frac{1}{n}\sum_{i=1}^n\Big(\frac{T_i}{\widetilde{J}_i}-1\Big)\Delta_L(\boldsymbol{X}_i).
\end{aligned}
$$

In the following, we will show that $R_j=o_p(n^{-1/2})$ for $0\leq j\leq 3$. 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 D.7, we can easily show that

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

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

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

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

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

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

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

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

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

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

where the last step follows from Lemma D.7.

Define the set of functions $\mathcal{F} = \{f_2 : \|m - m^*\|_{P,2} \leq \delta_1, \|\Delta\|_\infty \leq \delta_2\}$, 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,

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

We first consider the second term $n^{1/2} \sup_{f_2 \in \mathcal{F}} \mathbb{P} f_2$. Let $\mathcal{G}_1 = \{m \in \mathcal{M} : \|m - m^*\|_{P,2} \leq \delta_1\}$ and $\mathcal{G}_2 = \{\Delta \in \mathcal{H} - \boldsymbol{\alpha}_1^{*\top} \boldsymbol{h}_1 : \|\Delta\|_\infty \leq \delta_2\}$. 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 $\mathcal{F}_0 = \{f_2 : \|m - m^*\|_{P,2} \leq C, \|\Delta\|_{P,2} \leq 1\}$ for some constant $C > 0$, $\mathcal{G}_{10} = \{m \in \mathcal{M} + m^* : \|m\|_{P,2} \leq C\}$ and $\mathcal{G}_{20} = \{\Delta \in \mathcal{H} - \boldsymbol{\alpha}_1^{*\top} \boldsymbol{h}_1 : \|\Delta\|_{P,2} \leq 1\}$. 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. □