High Dimensional Propensity Score Estimation via Covariate Balancing

Yang Ning* Sida Peng† Kosuke Imai‡

May 9, 2017

Abstract

In this paper, we address the problem of estimating the average treatment effect (ATE) and the average treatment effect for the treated (ATT) in observational studies when the number of potential confounders is possibly much greater than the sample size. In particular, we develop a robust method to estimate the propensity score via covariate balancing in high-dimensional settings. Since it is usually impossible to obtain the exact covariate balance in high dimension, we propose to estimate the propensity score by balancing a carefully selected subset of covariates that are predictive of the outcome under the assumption that the outcome model is linear and sparse. The estimated propensity score is, then, used for the Horvitz-Thompson estimator to infer the ATE and ATT. We prove that the proposed methodology has the desired properties such as sample boundedness, root-n consistency, asymptotic normality, and semiparametric efficiency. We then extend these results to the case where the outcome model is a sparse generalized linear model. In addition, we show that the proposed estimator remains root-n consistent and asymptotically normal even when the propensity score model is misspecified. Finally, we conduct simulation studies to evaluate the finite-sample performance of the proposed, and apply the proposed methodology to estimate the causal effects of college attendance on adulthood political participation. Open-source software is available for implementing the proposed methodology.

Key words: average treatment effect, causal inference, covariate balancing propensity score, high dimensional inference, observational studies

1 Introduction

In observational studies, the unbiased estimation of causal effects is difficult because there may exist a large number of confounding variables. To address this confounding bias, numerous methods have been developed and applied in a variety of fields including computer science, statistics, and medical

*Department of Statistical Science, Cornell University, Ithaca, NY 14850, USA; e-mail: yn265@cornell.edu

†Department of Economics, Cornell University, Ithaca, NY 14850, USA; e-mail: sp947@cornell.edu

‡Department of Politics and Center for Statistics and Machine Learning, Princeton University, Princeton, NJ 08544, USA; e-mail: kimai@princeton.edu. URL: http://imai.princeton.edu

and social sciences (e.g., Pearl, 2000; Imbens and Rubin, 2015; Hernán and Robins, 2017). One prominent approach is based on the propensity score of Rosenbaum and Rubin (1983), which is defined as the conditional probability of treatment assignment given the set of confounders (see e.g., Imbens, 2000; Imai and van Dyk, 2004, for extensions to a non-binary treatment). Although the true propensity score is unknown in observational studies, the average treatment effect (ATE) or the average treatment effect for the treated (ATT) can be consistently estimated if a valid estimate of the propensity score is obtained using matching and weighting methods (see e.g., Lunceford and Davidian, 2004; Rubin, 2006). In a number of disciplines, propensity score methods have become part of applied researchers' standard toolkit. As the amount of data available to scientists grows, however, an increasingly important challenge across many fields is the question of how to adjust for a large number of observed pre-treatment covariates that represent potential confounders. For example, Schneeweiss et al. (2009) considers a total of several thousand candidate confounders obtained from the health care claims data.

In this paper, we address the problem of estimating the ATE and ATT in observational studies when the number of potential confounders is possibly much greater than the sample size. In particular, we develop a robust method to estimate the propensity score in a high-dimensional covariate space. The key idea is that we first obtain an initial estimate of the propensity score using a penalized regression model and then calibrate this estimate by balancing a smaller set of observed covariates that are predictive of the outcome under the assumption that the outcome model is linear and sparse. The estimated propensity score is, then, used for the Horvitz-Thompson estimator to infer the ATE and ATT (Horvitz and Thompson, 1952). We prove that the proposed methodology has the desired properties such as sample boundedness, root-n consistency, asymptotic normality, and semiparametric efficiency. We extend these results to the case where the outcome model is a sparse generalized linear model. In addition, we show that the proposed estimator remains root-n consistent and asymptotically normal even when the propensity score model is misspecified.

The proposed methodology, which we call the high-dimensional covariate balancing propensity score (HD-CBPS), builds on two strands of research that have recently emerged in the causal inference literature. First, a number of researchers have recently proposed to estimate the weights used for the estimation of the ATE and ATT by optimizing covariate balance between the treatment and control groups (e.g., Tan, 2010; Hainmueller, 2012; Graham et al., 2012; Imai and Ratkovic, 2014; Chan et al., 2015; Zubizarreta, 2015; Zhao, 2016; Fan et al., 2016). It has been shown, both theoretically and empirically, that these approaches can significantly improve the efficiency and robustness of resulting causal inference. We apply this covariate balancing strategy to the high-dimensional estimation of the propensity score.

In particular, the proposed HD-CBPS methodology extends the covariate balancing propensity score (CBPS) methodology of Imai and Ratkovic (2014) and Fan et al. (2016) to the high-dimensional settings, in which we have a large number of potential confounders. While the original CBPS methodology estimates the propensity score by balancing all covariates, this is impossible in high-dimensional settings especially when the number of covariates is greater than the sample size. Moreover, in such situations, the naive application of penalized estimation is unlikely to achieve the desired degree of covariate balance, resulting in an estimator of the ATE/ATT with the convergence

rate slower than root-n. Instead, we propose to estimate the propensity score model by solving a low-dimensional covariate balancing estimating equations. The proposed HD-CBPS methodology achieves the *weak* covariate balancing property in high-dimensional settings, which is a relaxation of the *strong* covariate balancing property of the original CBPS methodology. We show that the weak covariate balancing property is sufficient to yield a root-n consistent and efficient estimator of the ATE even in high-dimensional settings.

Second, we contribute to the growing literature on the estimation of the ATE/ATT in high-dimensional settings. Belloni et al. (2014) proposed a double selection approach to infer the coefficient of a treatment variable in a partially linear model under the assumption that both the outcome and treatment models are sparse. Farrell (2015), Belloni et al. (2013), and Chernozhukov et al. (2016) extended the augmented inverse probability weighted (AIPW) estimator of Robins et al. (1994) to high dimensional settings. A common characteristic of these methods is to estimate the nuisance parameters (e.g., propensity score model) via the penalized maximum likelihood and then solve the efficient score function to obtain a robust estimator of the ATE/ATT. Different from this line of work, we rely on the covariate balancing strategy for estimating the propensity score model and use the Horvitz-Thompson estimator for inferring the ATE/ATT without employing the efficient score function.

The proposed HD-CBPS estimator improves upon these high-dimensional AIPW estimators in the following respects. The HD-CBPS estimator achieves the sample boundedness property that it lies within the range of the outcome variable (Robins et al., 2007). This property avoids the adverse effects when the estimated propensity score is close to 0 or 1. Since the HD-CBPS methodology relies upon the tuning parameters in its penalized estimation, we also provide the theoretical guarantee for the validity of asymptotic inference with cross-validation. Under mild regularity conditions, the HD-CBPS estimator remains root-n asymptotically normal even if the propensity score model is misspecified. In numerical studies, we find that the HD-CBPS estimator tends to yield smaller mean squared error than the high-dimensional AIPW methods. This result, obtained under high-dimensional settings, agrees with the empirical findings in the aforementioned literature that covariate balancing improves the estimation of ATE/ATT when the dimension is relatively small.

Finally, our methodology is related to the recently proposed approximate residual balancing method (Athey et al., 2016), which does not require the propensity score model to be sparse or even well formulated. We argue that the estimation of the propensity score can help scientists better understand the treatment assignment mechanism (e.g., Rubin, 2008). Nevertheless, the ability of the approximate residual balancing method to avoid modeling the propensity score is an important advantage over the existing estimators including ours: we do show that our HD-CBPS methodology is robust to the misspecification of the propensity score. However, the approximate residual balancing method critically relies upon the linearity assumption of the outcome model and is not directly applicable to binary or categorical outcomes. In contrast, we show that the proposed HD-CBPS estimator can be extended to the cases where the outcome variable follows a generalized linear model. Finally, the HD-CBPS estimator has the sample boundedness property, which the approximate residual balancing method does not possess. Thus, the HD-CBPS estimator also overcomes the same limitation of the original CBPS estimator of Imai and Ratkovic (2014) and

Fan et al. (2016). Open-source software is available for implementing the proposed methodology (Fong et al., 2015).

Organization. The paper is organized as follows. In Section 2, we propose the high-dimensional covariate balancing method under the linearity assumption and study its theoretical properties. In Section 3, the method is further extended to the case that the outcome model follows from the generalized linear model. The simulation results and data analysis are presented in Section 4 and Section 5. The last section contains discussions. We defer the proofs and further technical details to the Appendices.

Notation. Throughout the paper, we use the following notations. For $\mathbf{v} = (v_1, ..., v_d)^T \in \mathbb{R}^d$, and $1 \leq q \leq \infty$, we define $\|\mathbf{v}\|_q = (\sum_{i=1}^d |v_i|^q)^{1/q}$, $\|\mathbf{v}\|_0 = |\text{supp}(\mathbf{v})|$, where $\text{supp}(\mathbf{v}) = \{j : v_j \neq 0\}$ and $|A|$ is the cardinality of a set A. Denote $\|\mathbf{v}\|_\infty = \max_{1\leq i\leq d} |v_i|$ and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^T$. For a matrix $\mathbf{M} = [M_{jk}]$, let $\|\mathbf{M}\|_{\max} = \max_{jk} |M_{jk}|$, $\|\mathbf{M}\|_1 = \sum_{jk} |M_{jk}|$, $\|\mathbf{M}\|_{\ell_\infty} = \max_j \sum_k |M_{jk}|$. If the matrix $\mathbf{M}$ is symmetric, then $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ are the minimal and maximal eigenvalues of $\mathbf{M}$. For $S \subseteq \{1, ..., d\}$, let $\mathbf{v}_S = \{v_j : j \in S\}$ and S^c be the complement of S. For two positive sequences a_n and b_n, we write $a_n \asymp b_n$ if $C \leq a_n/b_n \leq C'$ for some $C, C' > 0$. Similarly, we use $a_n \lesssim b_n$ to denote $a_n \leq Cb_n$ for some constant $C > 0$. A random variable X is sub-exponential if there exists some constant $K_1 > 0$ such that $\mathbb{P}(|X| > t) \leq \exp(1 - t/K_1)$ for all $t \geq 0$. The sub-exponential norm of X is defined as $\|X\|_{\psi_1} = \sup_{p\geq 1} p^{-1}(\mathbb{E}|X|^p)^{1/p}$. A random variable X is sub-Gaussian if there exists some constant $K_2 > 0$ such that $\mathbb{P}(|X| > t) \leq \exp(1 - t^2/K_2^2)$ for all $t \geq 0$. The sub-Gaussian norm of X is defined as $\|X\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p}$.

2 The Proposed Methodology

Suppose that we observe a simple random sample of size n from a population of interest. For each unit i of the sample, we observe a 3-tuple $\{T_i, Y_i, \boldsymbol{X}_i\}$ where $\boldsymbol{X}_i$ is a d-dimensional vector of pre-treatment covariates, Y_i is an outcome variable, and $T_i \in \{0, 1\}$ is a binary treatment variable denoting whether the observation receives the treatment ($T_i = 1$) or not ($T_i = 0$). For each unit, let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes under the treatment and control conditions, respectively, where we make the stable unit treatment value assumption (Rubin, 1990). Then, the observed outcome can be written as $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$. Our goal is to infer the average treatment effect (ATE),

$$\mu^* \;=\; \mathbb{E}(Y_i(1) - Y_i(0)). \tag{2.1}$$

In this section, we first describe how to estimate $\mu_1^* = \mathbb{E}(Y_i(1))$. The estimate of the ATE can be obtained by estimating $\mu_0^* = \mathbb{E}(Y_i(0))$ in a similar manner. We also begin by assuming that the outcome model $K_1(\boldsymbol{X}_i) = \mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i)$ follows a linear model, i.e., $K_1(\boldsymbol{X}_i) = \boldsymbol{\alpha}^{*\top}\boldsymbol{X}_i$ for some unknown parameter $\boldsymbol{\alpha}^* \in \mathbb{R}^d$. The extension to the generalized linear models will be studied in Section 3.2. Finally, throughout this paper, we assume that the treatment assignment is strongly ignorable (Rosenbaum and Rubin, 1983), and the propensity score follows the logistic regression

model. This implies,

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid \boldsymbol{X}_i, \quad \text{and} \quad \mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i) \;=\; \pi(\boldsymbol{X}_i^\top \beta^*) \;=\; \frac{\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*)}{1 + \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*)} \tag{2.2}$$

for all i and $\boldsymbol{\beta}^*$ is an unknown d-dimensional vector with $d \gg n$. That is, we consider the settings where the number of covariates is possibly much greater than the sample size.

2.1 High Dimensional Covariate Balancing Propensity Score under Linearity

In many applications, it is often reasonable to assume that the propensity score model is sparse or approximately sparse. Under this sparsity assumption, the penalized maximum likelihood estimator (PMLE) with ℓ_1 (Tibshirani, 1996) or non-convex penalty function (Fan and Li, 2001) has been proposed to perform estimation and prediction. Unfortunately, the PMLE cannot be directly plugged into the Horvitz-Thompson estimator to infer μ_1^* because the PMLE may incur a large bias due to the shrinkage effect and its limiting distribution is often non-normal (Knight and Fu, 2000). Thus, the resulting estimator may have a slower rate of convergence and an intractable limiting distribution.

To address this problem, we propose to estimate the propensity score such that covariate balance between the treatment and control groups is optimized. To this end, we distinguish the following two types of covariate balancing properties.

Definition 2.1 (Covariate Balancing Properties)**.** Let $\widehat{\pi} = \pi(\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}})$ denote an estimator of the true propensity score $\pi^* = \pi(\boldsymbol{X}^\top \boldsymbol{\beta}^*)$ with $\widehat{\boldsymbol{\beta}}$ being an estimator of $\boldsymbol{\beta}^*$.

(a) We call $\widehat{\pi}$ satisfies the *strong* covariate balancing property if the following equality holds,

$$\sum_{i=1}^{n} \left(\frac{T_i}{\widehat{\pi}_i} - 1 \right) \boldsymbol{X}_i \;=\; 0. \tag{2.3}$$

(b) We call $\widehat{\pi}$ satisfies the *weak* covariate balancing property if the following equality holds,

$$\sum_{i=1}^{n} \left(\frac{T_i}{\widehat{\pi}_i} - 1 \right) \boldsymbol{\alpha}^{*\top} \boldsymbol{X}_i \;=\; 0, \tag{2.4}$$

where $\boldsymbol{\alpha}^*$ is defined by $K_1(\boldsymbol{X}_i) = \boldsymbol{\alpha}^{*\top} \boldsymbol{X}_i$.

It is clear that, if the estimator $\widehat{\pi}_i$ satisfies the strong covariate balancing property then it also satisfies the weak covariate balancing property, but the converse may not hold. The existing covariate balancing methods such as those proposed by Imai and Ratkovic (2014) and Fan et al. (2016) aim to achieve the strong covariate balancing property. However, constructing an estimator $\widehat{\pi}$ with the strong covariate balancing property is difficult in high-dimensional settings. When $d > n$, the solution to equation (2.3) is not unique and therefore not even well defined. In addition, imposing additional penalty or constraint may introduce bias such that the strong covariate balancing property may not hold.

To overcome this difficulty, we propose to estimate the propensity score such that equation (2.4) rather than equation (2.3) approximately holds. We show that the weak covariate balancing property is sufficient to guarantee the root-n consistency and semiparametric efficiency. Here, we introduce the proposed methodology, which we call the high dimensional covariate balancing propensity score (HD-CBPS).

Step 1: Estimate the propensity score via the penalized logistic regression model,

$$\widehat{\boldsymbol{\beta}} \;=\; \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\Big\{T_i(\boldsymbol{\beta}^\top \boldsymbol{X}_i) - \log(1+\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_i))\Big\} + \lambda\|\boldsymbol{\beta}\|_1, \tag{2.5}$$

where $\lambda > 0$ is a tuning parameter. In this step, we may replace the ℓ_1 penalty with a non-convex penalty function (Fan and Li, 2001).

Step 2: Fit the outcome model via the penalized least squares using the treatment group,

$$\widetilde{\boldsymbol{\alpha}} \;=\; \underset{\boldsymbol{\alpha}\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} T_i\{Y_i - \boldsymbol{\alpha}^\top \boldsymbol{X}_i\}^2 + \lambda'\|\boldsymbol{\alpha}\|_1, \tag{2.6}$$

where $\lambda' > 0$ is a tuning parameter. Again, we may use a non-convex penalty function instead of the ℓ_1 penalty.

Step 3: Let $\widetilde{S} = \{j : |\widetilde{\alpha}_j| > 0\}$ denote the support of $\widetilde{\boldsymbol{\alpha}}$ and $\boldsymbol{X}_{\widetilde{S}}$ represent the corresponding subset of $\boldsymbol{X}$. We calibrate the initial estimator $\widehat{\boldsymbol{\beta}}_{\widetilde{S}}$ to balance $\boldsymbol{X}_{\widetilde{S}}$. Specifically, we solve,

$$\widetilde{\boldsymbol{\gamma}} \;=\; \underset{\boldsymbol{\gamma}\in\mathbb{R}^{|\widetilde{S}|}}{\operatorname{argmin}} \|\boldsymbol{g}_n(\boldsymbol{\gamma})\|_2^2 \;\text{ where }\; \boldsymbol{g}_n(\boldsymbol{\gamma}) = n^{-1}\sum_{i=1}^{n}\left(\frac{T_i}{\pi(\boldsymbol{\gamma}^\top \boldsymbol{X}_{i\widetilde{S}} + \widehat{\boldsymbol{\beta}}_{\widetilde{S}^c}^\top \boldsymbol{X}_{i\widetilde{S}^c})} - 1\right)\boldsymbol{X}_{i\widetilde{S}} \tag{2.7}$$

We then set $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}_{\widetilde{S}^c})$ and $\widetilde{\pi}_i = \pi(\widetilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_i)$.

Step 4: Estimate μ_1^* by the Horvitz-Thompson estimator $\widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} T_i Y_i/\widetilde{\pi}_i$.

In Step 1, we obtain an initial estimate of propensity score via the PMLE. However, this initial estimator may be biased because it may fail to select important confounders. In Step 2, we identify a set of covariates that are predictive of the outcome. These covariates may include those that may not have been selected in Step 1. In Step 3, we calibrate the estimated propensity score by balancing a subset of covariates $\boldsymbol{X}_{\widetilde{S}}$. Equation (2.7) implies that the proposed HD-CBPS methodology achieves the strong covariate balancing property only for these covariates $\boldsymbol{X}_{\widetilde{S}}$ but not for the other covariates $\boldsymbol{X}_{\widetilde{S}^c}$. Thus, unlike the original CBPS methodology, the HD-CBPS methodology does not achieve the strong covariate balancing property. Interestingly, however, the HD-CBPS methodology does approximately satisfy the weak covariate balancing property if $\boldsymbol{\alpha}^*$ can be well approximated by $\widetilde{\boldsymbol{\alpha}}$. Specifically,

$$\sum_{i=1}^{n}\left(\frac{T_i}{\widetilde{\pi}_i} - 1\right)\boldsymbol{\alpha}^{*\top}\boldsymbol{X}_i \;\approx\; \sum_{i=1}^{n}\left(\frac{T_i}{\widetilde{\pi}_i} - 1\right)\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X}_i \;=\; \sum_{i=1}^{n}\left(\frac{T_i}{\widetilde{\pi}_i} - 1\right)\widetilde{\boldsymbol{\alpha}}_{\widetilde{S}}^\top \boldsymbol{X}_{i\widetilde{S}} \;=\; 0,$$

where the first equality follows from $\widetilde{\boldsymbol{\alpha}}_{\widetilde{S}^c} = 0$ and the second equality holds due to equation (2.7).

The proposed HD-CBPS methodology requires two tuning parameters λ and λ'. In practice, we choose the values of these parameters via the cross validation method. In Section 2.3, we show that the HD-CBPS estimator with the cross validated tuning parameters remains root-n consistent and asymptotically normal. Our simulation results shown in Section 4 confirm this theoretical finding.

Our procedure is different from the existing methods on high-dimensional regressions; see Zhang and Zhang (2014), Javanmard and Montanari (2013), van de Geer et al. (2014), Belloni et al. (2016), Ning and Liu (2014), and Cai and Guo (2015), among many others. The main idea of these methods is to correct the bias of the Lasso-type estimators or the score function that arises due to regularization. In our problem, we remove the bias of the Lasso or nonconvex estimator $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\alpha}}$ by achieving the strong covariate balancing property for the selected covariates, which then yields the weak covariate balancing property among all covariates. The use of the covariate balancing estimating equations in Step 3 also makes our HD-CBPS method different from those of Farrell (2015) and Belloni et al. (2013), which rely on the efficient score function. The HD-CBPS method also differs from the approximate residual balancing method of Athey et al. (2016). While the HD-CBPS method removes the bias of initial propensity score estimate by balancing covariates, the latter removes the bias of the initial ATE estimate by weighting residuals.

2.2 Theoretical Results

Let us denote $s_1 = \|\boldsymbol{\beta}^*\|_0$ and $s_2 = \|\boldsymbol{\alpha}^*\|_0$. To study the theoretical properties of our HD-CBPS estimator $\widehat{\mu}_1$, we impose the following conditions.

Assumption 2.1 (Sub-Gaussian condition)**.** Assume that $\epsilon_1 = Y(1) - \boldsymbol{\alpha}^{*\top}\boldsymbol{X}$ and X_j satisfy $\|\epsilon_1\|_{\psi_2} \leq C_\epsilon$ and $\|X_j\|_{\psi_2} \leq C_X$ for any $1 \leq j \leq d$, where C_X and C_ϵ are two positive constants.

Assumption 2.2 (Overlap)**.** There exists a constant $c_0 > 0$ such that $\pi_i^* \geq c_0$ for $1 \leq i \leq n$.

Assumption 2.3 (Eigenvalue condition)**.** Denote $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{X}^{\otimes 2})$. There exists a constant $C > 0$ such that $C \leq \lambda_{\min}(\boldsymbol{\Sigma}_{SS}) \leq \lambda_{\max}(\boldsymbol{\Sigma}_{SS}) \leq 1/C$ for any $S \subset \{1, ..., d\}$ with $|S| \lesssim s_2$.

Assumption 2.1 controls the tail behavior of the error ϵ_1 and the covariate X_j, which facilitates the use of many existing exponential inequalities (Vershynin, 2010). Assumption 2.2, known as the overlap condition, is a standard assumption in the causal inference literature; see e.g., Assumption 3 of Athey et al. (2016). Finally, Assumption 2.3 implies the sparse eigenvalue condition for the design matrix (Bickel et al., 2009; Zhang, 2010), which is commonly used in the existing high-dimensional statistics literature. Given these assumptions, the following theorem establishes the asymptotic normality and semiparametric efficiency of the proposed HD-CBPS estimator $\widehat{\mu}_1$.

Theorem 2.1 (Asymptotic Normality and Semiparametric Efficiency)**.** Suppose that Assumptions 2.1, 2.2, and 2.3 hold. In addition, assume $\max(s_1, s_2) \log d \sqrt{\log n / n} = o(1)$. If we take $\lambda \asymp \lambda' \asymp \sqrt{\log d / n}$, then we have,

$$\widehat{\mu}_1 - \mu_1^* = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{T_i}{\pi_i^*} (Y_i(1) - \boldsymbol{\alpha}^{*\top} \boldsymbol{X}_i) + \boldsymbol{\alpha}^{*\top} \boldsymbol{X}_i - \mu_1^* \right] + o_p(n^{-1/2}).$$

Let V be the semiparametric asymptotic variance bound, i.e.,

$$V \;=\; \mathbb{E}\left[\frac{1}{\pi^*}\mathbb{E}(\epsilon_1^2 \mid \boldsymbol{X}) + (\boldsymbol{\alpha}^{*\top}\boldsymbol{X} - \mu_1^*)^2\right],$$

and $\Phi(t)$ denote the CDF of the standard normal distribution. In addition, assume that $\mathbb{E}(\epsilon_1^2) \geq c$ for some constant $c > 0$. Then,

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\widehat{\mu}_1 - \mu_1^*}{\sqrt{V/n}} \leq t\right) - \Phi(t)\right| \;\leq\; C\left(\frac{1 + \max(s_1, s_2)\log d\sqrt{\log n}}{\sqrt{n}} + \frac{1}{d}\right), \tag{2.8}$$

for some constant $C > 0$.

Proof is given in Appendix A.1. The theorem shows that $\widehat{\mu}_1 - \mu_1^*$ is asymptotically equivalent to the average of efficient score functions and hence $\widehat{\mu}_1$ is efficient. In addition, the right hand side of equation (2.8) specifies the rate of convergence for the normal approximation of $(\widehat{\mu}_1 - \mu_1^*)/\sqrt{V/n}$.

In this theorem, we assume that both propensity score and outcome models are sparse such that $\max(s_1, s_2)\log d\sqrt{\log n/n} = o(1)$ holds. This condition is stronger than Athey et al. (2016) who only require the sparsity of the outcome model. As shown in Section 2.5, however, the HD-CBPS estimator is robust to the misspecification of the propensity score model. In addition, the estimation of the propensity score can help scientists better understand the treatment assignment mechanisms (e.g., Rubin, 2008). Indeed, as a byproduct of Theorem 2.1, our estimated propensity score is uniformly consistent, i.e.,

$$\max_{1\leq i\leq n}|\widetilde{\pi}_i - \pi_i^*| \;=\; O_p\left(\frac{\max(s_1, s_2)\log d\sqrt{\log n}}{\sqrt{n}}\right).$$

Finally, the approximate residual balancing method of Athey et al. (2016) critically depends on the assumed linearity of the outcome model and is not readily applicable if the outcome model is nonlinear. In contrast, our method is more widely applicable. As an illustration of this generality, we consider the extension to the generalized linear models in Section 3.2.

Remark 2.1. To construct a confidence interval for μ_1^*, we estimate V by

$$\widehat{V} \;=\; \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\widehat{\sigma}_1^2}{\widetilde{\pi}_i} + (\widetilde{\boldsymbol{\alpha}}^\top\boldsymbol{X}_i - \widehat{\mu}_1)^2\right\} \quad \text{where} \quad \widehat{\sigma}_1^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\widetilde{\pi}_i}(Y_i - \widetilde{\boldsymbol{\alpha}}^\top\boldsymbol{X}_i)^2. \tag{2.9}$$

Here, for simplicity, we assume the homoskedastic error, i.e., $\mathbb{E}(\epsilon_1^2 \mid \boldsymbol{X}) = \sigma_1^2$. It is possible to construct a confidence interval for the heteroskedastic error by using a kernel estimator under certain smoothness assumption on $\mathbb{E}(\epsilon_1^2 \mid \boldsymbol{X})$. The following Lemma shows the consistency of $\widehat{V}$ under the homoskedasticity assumption.

Lemma 2.1. Under the assumptions maintained for Theorem 2.1, we have

$$|\widehat{V} - V| \;=\; O_p\left(\frac{\max(s_1, s_2)\sqrt{\log d\log n}}{\sqrt{n}}\right).$$

Proof is given in Appendix A.2. Together with the Slutsky's theorem, Lemma 2.1 and Theorem 2.1 imply,

$$|\mathbb{P}(\mu_1^* \in \mathcal{I}) - (1-\eta)| \;\leq\; C\left(\frac{1+\max(s_1,s_2)\log d\sqrt{\log n}}{\sqrt{n}} + \frac{1}{d}\right), \tag{2.10}$$

where

$$\mathcal{I} \;=\; \left[\widehat{\mu}_1 - z_{1-\eta/2}\sqrt{\widehat{V}/n},\; \widehat{\mu}_1 + z_{1-\eta/2}\sqrt{\widehat{V}/n}\right],$$

and $z_{1-\eta/2}$ is the $(1-\eta/2)$ quantile of the standard normal distribution. In fact, the confidence interval $\mathcal{I}$ is honest in the sense that equation (2.10) holds uniformly over all probability distributions that satisfy Assumptions 2.1, 2.2, and 2.3.

Remark 2.2. The proposed HD-CBPS estimator $\widehat{\mu}_1$ satisfies the so-called *sample boundedness property* (Robins et al., 2007), which states that $\widehat{\mu}_1$ lies in the range of $\{Y_i : T_i = 1, i = 1, ..., n\}$. This is because, by construction, the covariate balancing equation satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\widetilde{\pi}_i} - 1\right) \;=\; 0, \tag{2.11}$$

so long as an intercept is included in $\boldsymbol{X}_{i\widetilde{S}}$. Equation (2.11) implies that the estimated propensity score $\widetilde{\pi}_i$ must be greater than or equal to $1/n$ for any treated observation. This makes the our estimator $\widehat{\mu}_1$ more stable especially in high-dimensional settings, since the estimated propensity score cannot be too close to 0. To see why the sample boundedness property holds, we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\widetilde{\pi}_i} \;\geq\; \frac{\min_{i:T_i=1} Y_i}{n}\sum_{i=1}^{n}\frac{T_i}{\widetilde{\pi}_i} \;=\; \min_{i:T_i=1} Y_i,$$

where the last equality follows from equation (2.11). Similarly, we can easily show that $\widehat{\mu}_1 \leq \max_{i:T_i=1} Y_i$. However, this sample boundedness property does not hold for the existing estimators such as those proposed by Farrell (2015), Belloni et al. (2013), Chernozhukov et al. (2016), and Athey et al. (2016).

2.3 Data Driven Choice of Tuning Parameters

While Theorem 2.1 specifies the rate of tuning parameters, λ and λ', in practice we must know the values of the constants a and a' in $\lambda = a\sqrt{\log d/n}$ and $\lambda' = a'\sqrt{\log d/n}$ to implement the method. We show here that K-fold cross validation can be used to determine the values of these tuning parameters. Specifically, we randomly divide the sample into K disjoint sets $I_1, ..., I_K$ with roughly equal size. For concreteness, we focus on the K-fold cross validation for λ' in Step 2 of the HD-CBPS methodology described in Section 2.1. Let $L^{(I_k)}(\boldsymbol{\alpha}) = \frac{1}{|I_k|}\sum_{i\in I_k} T_i(Y_i - \boldsymbol{\alpha}^\top \boldsymbol{X}_i)^2$ and $\widetilde{\boldsymbol{\alpha}}^{(-I_k)}(\lambda') = \operatorname{argmin} L^{(-I_k)}(\boldsymbol{\alpha}) + \lambda'\|\boldsymbol{\alpha}\|_1$ where $-I_k = \{1, ..., n\}\backslash I_k$. Then, the optimal value of λ'

that minimizes the cross validation error is given by,

$$\widehat{\lambda}' \;=\; \underset{\lambda\in\Lambda}{\operatorname{argmin}}\, CV(\lambda) \;\text{ where }\; CV(\lambda) \;=\; \sum_{k=1}^{K} L^{(I_k)}(\widetilde{\boldsymbol{\alpha}}^{(-I_k)}(\lambda)),$$

and $\Lambda = \{a_1, ..., a_M\} \times \sqrt{\log d/n}$ is a grid for λ with some pre-specified constants $a_1, ..., a_M$. For the theoretical purpose, we assume M is fixed. Thus, the estimator of $\boldsymbol{\alpha}$ based on K-fold cross validation can be denoted by $\widetilde{\boldsymbol{\alpha}}(\widehat{\lambda}')$.

Similarly, we search for the optimal value of λ in Step 1 of the HD-CBPS methodology described in Section 2.1 using a grid and denote it by $\widehat{\lambda}$. The resulting estimator is $\widetilde{\boldsymbol{\beta}}(\widehat{\lambda})$. Once we plug $\widetilde{\boldsymbol{\alpha}}(\widehat{\lambda}')$ and $\widetilde{\boldsymbol{\beta}}(\widehat{\lambda})$ into the estimation procedure for μ_1^*, we obtain the estimator $\widehat{\mu}_1(\widehat{\lambda}, \widehat{\lambda}')$. This estimator has the advantage of being fully data driven so that users need not specify the values of tuning parameters. However, since the tuning parameters depend on the data and therefore are random, Theorem 2.1 is not applicable. Fortunately, the following corollary shows that the normal approximation result still holds for the data driven estimator $\widehat{\mu}_1(\widehat{\lambda}, \widehat{\lambda}')$.

Corollary 2.1 (Validity of Normal Approximation with Cross Validation)**.** Under the assumptions maintained in Theorem 2.1, we have,

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\widehat{\mu}_1(\widehat{\lambda}, \widehat{\lambda}') - \mu_1^*}{\sqrt{V/n}} \le t \right) - \Phi(t) \right| \;\le\; C\left(\frac{1 + \max(s_1, s_2) \log d \sqrt{\log n}}{\sqrt{n}} + \frac{1}{d} \right)$$

for some constant $C > 0$.

Proof is given in Appendix A.3. The corollary provides a formal theoretical justification for the use of K-fold cross validation when constructing confidence intervals and conducting hypothesis tests.

2.4 Estimation of the Average Causal Effects

So far, we have focused on the estimator of μ_1^* and its theoretical properties. Here, we show how to estimate the ATE and the ATT using the proposed HD-CBPS methodology.

2.4.1 The Average Treatment Effect (ATE)

To estimate the ATE using the HD-CBPS methodology, we simply apply a similar procedure to estimate μ_0^* using the control group and then obtain the estimator of the ATE as,

$$\widehat{\mu} \;=\; \widehat{\mu}_1 - \widehat{\mu}_0, \tag{2.12}$$

where $\widehat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{n}(1 - T_i)Y_i/\bar{\pi}_i$. We estimate $\bar{\pi}_i$ via the HD-CBPS methodology in the same way as $\widetilde{\pi}_i$ except that in Steps 2 and 3 of the procedure described in Section 2.1 T_i is replaced with $1 - T_i$.

The exact algorithm for estimating the ATE and its theoretical properties are described in Appendix C. In particular, we show,

$$\widehat{\mu} - \mu^* \;=\; \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i}{\pi_i^*}(Y_i(1) - K_1(\boldsymbol{X}_i)) - \frac{1-T_i}{1-\pi_i^*}(Y_i(0) - K_0(\boldsymbol{X}_i)) + \Delta K(\boldsymbol{X}_i) - \mu^*\right] + o_p(n^{-1/2}),$$

where $K_1(\boldsymbol{X}_i) = \mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i)$, $K_0(\boldsymbol{X}_i) = \mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i)$, and $\Delta K(\boldsymbol{X}_i) = K_1(\boldsymbol{X}_i) - K_0(\boldsymbol{X}_i)$. Therefore, the estimator $\widehat{\mu}$ attains the following semiparametric asymptotic variance bound for estimating the ATE,

$$\mathbb{E}\left[\frac{1}{\pi^*}\mathbb{E}(\epsilon_1^2 \mid \boldsymbol{X}) + \frac{1}{1-\pi^*}\mathbb{E}(\epsilon_0^2 \mid \boldsymbol{X}) + (\Delta K(\boldsymbol{X}) - \mu^*)^2\right],$$

where $\epsilon_0 = Y(0) - K_0(\boldsymbol{X})$.

As discussed in Remark 2.1, the asymptotic variance can be obtained via the plug-in estimator under the homoskedasticity assumption. To construct confidence intervals for the ATE μ^*, we estimate V by

$$\widehat{V} \;=\; \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\widehat{\sigma}_1^2}{\widetilde{\pi}_i} + \frac{\widehat{\sigma}_0^2}{1-\bar{\pi}_i} + (\widetilde{\boldsymbol{\alpha}}_1^\top \boldsymbol{X}_i - \widetilde{\boldsymbol{\alpha}}_0^\top \boldsymbol{X}_i - \widehat{\mu}_1)^2\right\},$$

where

$$\widehat{\sigma}_0^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{1-T_i}{1-\bar{\pi}_i}(Y_i - \widetilde{\boldsymbol{\alpha}}_0^\top \boldsymbol{X}_i)^2 \quad \text{and} \quad \widehat{\sigma}_1^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\widetilde{\pi}_i}(Y_i - \widetilde{\boldsymbol{\alpha}}_1^\top \boldsymbol{X}_i)^2.$$

2.4.2 The Average Treatment Effect for the Treated (ATT)

Next, we consider the estimation of the average treatment effect for the treated (ATT), which is defined as $\tau^* = \mathbb{E}(Y_i(1) - Y_i(0) \mid T_i = 1)$. Let $\tau_1^* = \mathbb{E}(Y_i(1) \mid T_i = 1)$ and $\tau_0^* = \mathbb{E}(Y_i(0) \mid T_i = 1)$. By the law of total probability,

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

Thus, a simple estimator of τ_1^* is

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

To estimate τ_0^*, we notice that

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

Hence, to accurately estimate τ_0^*, one has to develop an alternative set of the covariate balancing equations. Recall that $\widehat{\boldsymbol{\beta}}$ is defined in equation (2.5). For notational simplicity, we denote the penalized least squared estimator for the control group by $\widetilde{\boldsymbol{\alpha}}$ with T_i replaced by $1-T_i$ in equation (2.6). Recall that $\widetilde{S} = \{j : |\widetilde{\alpha}_j| > 0\}$ is the support of $\widetilde{\boldsymbol{\alpha}}$. Then, we calibrate the initial estimator $\widehat{\boldsymbol{\beta}}_{\widetilde{S}}$ to balance $\bar{\boldsymbol{X}}_{i\widetilde{S}} = (1, \boldsymbol{X}_{i\widetilde{S}}^\top)^\top$. Specifically, we solve $\widetilde{\boldsymbol{\gamma}} = \operatorname{argmin}_{\boldsymbol{\gamma}\in\mathbb{R}^{|\widetilde{S}|+1}} \|\boldsymbol{g}_n(\boldsymbol{\gamma})\|_2^2$, where

$$\boldsymbol{g}_n(\boldsymbol{\gamma}) = n^{-1}\sum_{i=1}^n \left(T_i - \frac{(1-T_i)\pi(\boldsymbol{\gamma}^\top \bar{\boldsymbol{X}}_{i\widetilde{S}} + \widehat{\boldsymbol{\beta}}_{\widetilde{S}^c}^\top \boldsymbol{X}_{i\widetilde{S}^c})}{1-\pi(\boldsymbol{\gamma}^\top \bar{\boldsymbol{X}}_{i\widetilde{S}} + \widehat{\boldsymbol{\beta}}_{\widetilde{S}^c}^\top \boldsymbol{X}_{i\widetilde{S}^c})} \right) \bar{\boldsymbol{X}}_{i\widetilde{S}}. \tag{2.13}$$

Then, we set $\widetilde{\pi}_i = \pi(\widetilde{\boldsymbol{\beta}}^\top \bar{\boldsymbol{X}}_i)$ with $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}_{\widetilde{S}^c})$ and estimate τ_0 by

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

where $\widetilde{r}_i = \widetilde{\pi}_i/(1-\widetilde{\pi}_i)$. The final estimator of the ATT is $\widehat{\tau} = \widehat{\tau}_1 - \widehat{\tau}_0$. The covariate balancing equations (2.13) aim to balance the selected covariate $\bar{\boldsymbol{X}}_{i\widetilde{S}}$ reweighted by $\widetilde{r}_i$ in the control group with $\bar{\boldsymbol{X}}_{i\widetilde{S}}$ in the treatment group. This proposal agrees with the intuition originated from some of the recent work (Hainmueller, 2012; Zubizarreta, 2015). Similar to Theorem 2.1, the following proposition establishes the asymptotic normality and semiparametric efficiency of the estimator $\widehat{\tau}$, whose proof is given in Appendix A.5.

Proposition 2.1 (Asymptotic Normality and Semiparametric Efficiency for $\widehat{\tau}$)**.** Suppose that Assumptions 2.1, 2.2, and 2.3 hold. In addition, assume $\max(s_1, s_2)\log d\sqrt{\log n/n} = o(1)$, where $s_1 = \|\boldsymbol{\beta}^*\|_0$ and $s_2 = \|\boldsymbol{\alpha}^*\|_0$. If we take $\lambda \asymp \lambda' \asymp \sqrt{\log d/n}$, then we have,

$$\widehat{\tau} - \tau^* = \frac{1}{n}\sum_{i=1}^n \frac{1}{p}\left[T_i\epsilon_{1i} - (1-T_i)r_i^*\epsilon_{0i} + T_i(\Delta K(\boldsymbol{X}_i) - \tau^*)\right] + o_p(n^{-1/2}),$$

where $p = \mathbb{P}(T_i = 1)$ and $r_i^* = \pi_i^*/(1-\pi_i^*)$. This implies $n^{1/2}(\widehat{\tau} - \tau^*) \to_d N(0, W)$, where

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

is the semiparametric asymptotic variance bound for τ (Hahn, 1998).

2.5 Robustness to the Misspecification of Propensity Score Model

We next investigate the robustness of the proposed HD-CBPS methodology to the misspecification of propensity score model. We show that even if the propensity score model is misspecified, Theorem 2.1, i.e., the asymptotic normality of the HD-CBPS, still holds under mild conditions so long as the outcome model is corrected specified. Specifically, suppose that the true propensity score $\pi^* = \mathbb{P}(T = 1 \mid \boldsymbol{X})$ does not belong to the assumed parametric class $\{\pi(\boldsymbol{X}^\top\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{R}^d\}$ where $\pi(x) = \exp(x)/\{1+\exp(x)\}$. To characterize the limiting behavior of $\widetilde{\boldsymbol{\beta}}$, we further assume that

there exists $\boldsymbol{\beta}^o$ such that

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 \;=\; O_p\left((s_1+s_2)\sqrt{\frac{\log d}{n}}\right)$$

and

$$\frac{1}{n}\sum_{i=1}^{n}[\boldsymbol{X}_i^\top(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)]^2 \;=\; O_p\left(\frac{(s_1+s_2)\log d}{n}\right),$$

where $s_1 = \|\boldsymbol{\beta}^o\|_0 \ll n$. The existence of a sparse estimand $\boldsymbol{\beta}^o$ is typically made when misspecified models are studied in high-dimensional settings; see e.g., Section 5 of Ning and Liu (2014) and Assumption (A5) in Bühlmann and van de Geer (2015). Given $\boldsymbol{\beta}^o$ is sparse, the proof for the rate of convergence of $\widetilde{\boldsymbol{\beta}}$ is similar to that of Lemma A.2 in Appendix A.1. Therefore, we omit the details.

Recall that $\widetilde{S} = \{j : |\widetilde{\alpha}_j| > 0\}$ and $S = \{j : |\alpha_j^*| > 0\}$. The following theorem establishes the asymptotic properties of the proposed HD-CBPS estimator $\widehat{\mu}_1$ described in Section 2.1 under misspecified propensity score models.

Proposition 2.2 (Asymptotic Properties under Misspecified Propensity Score Models)**.** Suppose that Assumptions 2.1 and 2.3 hold. Furthermore, Assumption 2.2 holds for π_i^o where $\pi_i^o = \pi(\boldsymbol{X}_i^\top\boldsymbol{\beta}^o)$. In addition, the sure screening property holds for $\widetilde{S}$, i.e., $S \subseteq \widetilde{S}$ with probability tending to one. If $\lambda \asymp \lambda' \asymp \sqrt{\log d/n}$ and $\max(s_1,s_2)\log d\sqrt{\log n} = o(n^{1/2})$, then

$$\widehat{\mu}_1 - \mu_1^* \;=\; \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i}{\pi_i^o}(Y_i(1) - \boldsymbol{\alpha}^{*\top}\boldsymbol{X}_i) + \boldsymbol{\alpha}^{*\top}\boldsymbol{X}_i - \mu_1^*\right] + o_p(n^{-1/2}).$$

This implies $n^{1/2}(\widehat{\mu}_1 - \mu_1^*) \to_d N(0, V_{mis})$ where

$$V_{mis} \;=\; \mathbb{E}\left[\frac{\pi^*}{\pi^{o2}}\mathbb{E}(\epsilon_1^2 \mid \boldsymbol{X}) + (\boldsymbol{\alpha}^{*\top}\boldsymbol{X} - \mu_1^*)^2\right].$$

Proof is given in Appendix A.6. If the propensity score model is correctly specified, i.e., $\pi_i^o = \pi_i^*$, then the asymptotic variance V_{mis} in Proposition 2.2 agrees with the asymptotic variance V in Theorem 2.1. Since Proposition 2.2 allows for misspecified propensity score models, we require a stronger condition for the outcome model. In particular, we assume the sure screening property holds for $\widetilde{S}$ (Fan and Lv, 2008): the screening property for the Lasso estimator has been studied by Wasserman and Roeder (2009) and Meinshausen et al. (2009). Bühlmann (2013) showed that sure screening requires weaker conditions than variable selection and often holds in practice.

Remark 2.3. Unlike the correctly specified case, the validity of Proposition 2.2 hinges on the screening property rather than the consistency of $\widetilde{\boldsymbol{\alpha}}$. This has the following two implications. First, our Lasso procedure for $\boldsymbol{\alpha}$ can be generalized to many different high dimensional estimation methods, provided the sure screening property holds. In particular, Fan and Lv (2008) and Xu and Chen (2014) showed that such a property is attained by the marginal regression and a modified Lasso estimator, respectively. A robust screening method is considered by Li et al. (2012). We refer to the review paper by Liu et al. (2015) for an overview of recent advances on screening methods.

Second, the screening property, together with the covariate balancing estimating equations, is crucial for eliminating the misspecification bias of the propensity score and guaranteeing the root-n consistency of $\widehat{\mu}_1$. In contrast, due to the estimation error of the penalized estimator $\widetilde{\boldsymbol{\alpha}}$, the AIPW methods such as those proposed by Farrell (2015), Belloni et al. (2013), and Chernozhukov et al. (2016) cannot guarantee the root-n consistency of $\widehat{\mu}_1$ under similar conditions. A heuristic analysis of the AIPW estimator is presented in Appendix B.

Remark 2.4. At first glance, the sure screening property $S \subseteq \widetilde{S}$ may appear to be strong, as the signals (i.e., nonzero values of $\boldsymbol{\alpha}^*$) need to be large enough. As can be seen in the proof of Proposition 2.2, the sure screening property can be relaxed to the condition $\|\boldsymbol{\alpha}^*_{S\setminus\widetilde{S}}\|_2 = o(n^{-1/2})$, which allows the presence of weak signals. For example, consider the case in which s_2 is a constant. Then, the condition $\|\boldsymbol{\alpha}^*_{S\setminus\widetilde{S}}\|_2 = o(n^{-1/2})$ holds, if the support set can be written as $S = S_{\text{strong}} \cup S_{\text{weak}}$ where $S_{\text{strong}} = \{j : |\alpha_j| \gtrsim \sqrt{\log d/n}\}$ contains strong signals and $S_{\text{weak}} = \{j : |\alpha_j| = o(n^{-1/2})\}$ represents weak signals. This can be seen from the fact that under the mutual coherence condition, the Lasso estimator can identify strong signals, i.e., $S_{\text{strong}} \subseteq \widetilde{S}$ (see Theorem 1 of Lounici et al., 2008). Thus, by the definition of S_{weak}, $\|\boldsymbol{\alpha}^*_{S\setminus\widetilde{S}}\|_2 \leq \|\boldsymbol{\alpha}^*_{S_{\text{weak}}}\|_2 = o(n^{-1/2})$ holds. As illustrated by this example, the sure screening assumption can be relaxed so that both strong and weak signals for $\boldsymbol{\alpha}^*$ are permitted.

3 Generalization

One advantage of the original CBPS methodology is that it can be extended to a number of situations (e.g., Imai and Ratkovic, 2014, 2015; Fong et al., 2016). The proposed HD-CBPS methodology can also be generalized to various settings. For illustration, we show that the HD-CBPS methodology can be generalized to the settings where the treatment variable takes more than two values and the outcome variable follows a generalized linear model.

3.1 Multi-valued Treatment

Suppose that the treatment variable T_i takes values in $\{0, 1, ..., K-1\}$ for some integer $K \geq 2$. For simplicity, we assume that the generalized propensity score follows the multinomial logistic regression model,

$$\mathbb{P}(T_i = k \mid \boldsymbol{X}_i) \;=\; \frac{\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}_k^*)}{1 + \sum_{k=1}^{K-1} \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}_k^*)}$$

where $\boldsymbol{\beta}_k^*$ is a vector of unknown parameters for $1 \leq k \leq K-1$. We can easily modify the algorithm described in Section 2.1. Specifically, in Step 1, we estimate $\boldsymbol{\beta}_k^*$ by the penalized multinomial regression. In Step 2, we estimate the parameter $\boldsymbol{\alpha}_k^*$ in the treatment group $T_i = k$ where $\boldsymbol{\alpha}_k^*$ is defined by $\mathbb{E}(Y(k) \mid \boldsymbol{X}) = \boldsymbol{\alpha}_k^{*\top} \boldsymbol{X}$. In Step 3, we re-estimate $\boldsymbol{\beta}_k^*$ by solving equation (2.7) in which T_i replaced with the indicator function $\mathbb{I}\{T_i = k\}$. Finally, we can construct the Horvitz-Thompson estimator for $\mathbb{E}(Y(k))$ as $\sum_{i=1}^n Y_i \mathbb{I}\{T_i = k\}/\widetilde{\pi}_{ik}$ where $\widetilde{\pi}_{ik}$ is the estimated generalized propensity score for the treatment $T_i = k$ (Imbens, 2000). Under the conditions similar to those of

Theorem 2.1, it can be shown that the resulting estimator is asymptotically normal and efficient in high-dimensional settings.

3.2 HD-CBPS under Generalized Linear Models

In this section, we extend the HD-CBPS methodology to the setting in which the outcome follows a generalize linear model. The validity of many existing methods such as those proposed by Imai and Ratkovic (2014), Fan et al. (2016), and Athey et al. (2016) critically rely on the assumption that the outcome follows a linear model with covariates $\boldsymbol{X}_i$ or some transformations (e.g., spline basis) of $\boldsymbol{X}_i$. Thus, generalizing the HD-CBPS to non-linear models is an important extension.

We begin by assuming that the conditional distribution of $Y_i(1)$ given $\boldsymbol{X}_i$ belongs to the exponential family,

$$p(y \mid \boldsymbol{x}) \;=\; h(y, \phi) \exp\left\{ \frac{y \boldsymbol{\alpha}^{*\top} \boldsymbol{x} - b(\boldsymbol{\alpha}^{*\top} \boldsymbol{x})}{a(\phi)} \right\}$$

where $h(\cdot, \cdot)$, $a(\cdot)$ and $b(\cdot)$ are known functions, ϕ is the dispersion parameter, and $\boldsymbol{\alpha}^*$ is a d-dimensional vector of unknown regression parameters. For simplicity, we assume that the dispersion parameter ϕ is known. Given this setup, we propose the following modification of the original HD-CBPS methodology.

Step 1: This step is identical to Step 1 described in Section 2.1, yielding the initial estimator $\widehat{\boldsymbol{\beta}}$.

Step 2: Fit the generalized linear model via the penalized maximum likelihood estimation using the treatment group,

$$\widetilde{\boldsymbol{\alpha}} \;=\; \underset{\boldsymbol{\alpha} \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \frac{T_i}{a(\phi)} \{ Y_i \boldsymbol{\alpha}^\top \boldsymbol{X}_i - b(\boldsymbol{\alpha}^\top \boldsymbol{X}_i) \} + \lambda' \|\boldsymbol{\alpha}\|_1 \right\}$$

where $\lambda' > 0$ is a tuning parameter.

Step 3: Let $\widetilde{S} = \{ j : |\widetilde{\alpha}_j| > 0 \}$ denote the support of $\widetilde{\boldsymbol{\alpha}}$. Define $\mathbf{f}(\boldsymbol{X}) = (b'(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X}), b''(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X}) \boldsymbol{X}_{\widetilde{S}}^\top)^\top$. Compute,

$$\widetilde{\boldsymbol{\gamma}} \;=\; \underset{\boldsymbol{\gamma} \in \mathbb{R}^{|\widetilde{S}|+1}}{\operatorname{argmin}} \|\boldsymbol{g}_n(\boldsymbol{\gamma})\|_2^2 \quad \text{where} \quad \boldsymbol{g}_n(\boldsymbol{\gamma}) \;=\; \frac{1}{n} \sum_{i=1}^{n} \left(\frac{T_i}{\pi(\boldsymbol{\gamma}^\top \bar{\boldsymbol{X}}_{i\widetilde{S}} + \widehat{\boldsymbol{\beta}}_{\widetilde{S}^c}^\top \boldsymbol{X}_{i\widetilde{S}^c})} - 1 \right) \mathbf{f}(\boldsymbol{X}_i). \quad (3.1)$$

Set $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}_{\widetilde{S}^c})$ and $\widetilde{\pi}_i = \pi(\widetilde{\boldsymbol{\beta}}^\top \bar{\boldsymbol{X}}_i)$ where $\bar{\boldsymbol{X}}_{i\widetilde{S}} = (1, \boldsymbol{X}_{i\widetilde{S}}^\top)^\top$.

Step 4: Estimate μ_1^* by the Horvitz-Thompson estimator $\widehat{\mu}_1 = \frac{1}{n} \sum_{i=1}^{n} T_i Y_i / \widetilde{\pi}_i$.

When compared with the algorithm described in Section 2.1, the main difference is that in Step 3 we balance $b'(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X}_i)$ and $b''(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X}_i) \boldsymbol{X}_{i\widehat{S}}$ instead of $\boldsymbol{X}_{i\widehat{S}}$ in equation (2.7). The reason is that to achieve a similar weak covariate balancing property, one must balance a vector of functions $\mathbf{f}(\boldsymbol{X})$ such that $b'(\boldsymbol{\alpha}^{*\top} \boldsymbol{X}) \in \operatorname{span}\{\mathbf{f}(\boldsymbol{X})\}$ where $\operatorname{span}\{\mathbf{f}(\boldsymbol{X})\}$ represents the linear space generated by the basis functions $\mathbf{f}(\boldsymbol{X})$. Since $b'(\boldsymbol{\alpha}^{*\top} \boldsymbol{X})$ is unknown in practice, we approximate $b'(\boldsymbol{\alpha}^{*\top} \boldsymbol{X})$

by a local linear estimator $b'(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X}) + b''(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X})(\boldsymbol{\alpha}^* - \widetilde{\boldsymbol{\alpha}})\boldsymbol{X}$. In low dimensional settings, one may use the following covariate balancing estimating equations,

$$\boldsymbol{g}_n(\boldsymbol{\beta}) \;=\; \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\pi(\boldsymbol{\beta}^\top \bar{\boldsymbol{X}}_i)} - 1\right)\mathbf{f}(\boldsymbol{X}_i), \tag{3.2}$$

where $\mathbf{f}(\boldsymbol{X}) = (b'(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X}), b''(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X})\boldsymbol{X}^\top)^\top$. To the best of our knowledge, the covariate balancing estimating equations for generalized linear models have never been derived even in low dimensional settings.

In high-dimensional settings, it is usually impossible to simultaneously balance these $d+1$ functions. In fact, due to the sparsity of $\boldsymbol{\alpha}^*$, most functions in $\mathbf{f}(\boldsymbol{X}) = (b'(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X}), b''(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X})\boldsymbol{X}^\top)^\top$ are not useful to approximate $b'(\boldsymbol{\alpha}^{*\top}\boldsymbol{X})$ and hence can be removed from $\mathbf{f}(\boldsymbol{X})$. To see this, let S denote the support set for $\boldsymbol{\alpha}$, i.e., $S = \{j : |\alpha_j^*| > 0\}$. Since $b'(\boldsymbol{\alpha}^{*\top}\boldsymbol{X}) = b'(\boldsymbol{\alpha}_S^{*\top}\boldsymbol{X}_S)$, applying a similar local linear approximation, we can choose $\mathbf{f}(\boldsymbol{X}) = (b'(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X}), b''(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X})\boldsymbol{X}_S^\top)^\top$, yielding a total of $|S|+1 \ll d+1$ equations. Furthermore, if we replace S by an estimator $\widetilde{S}$ defined in Step 3, then we recover our choice of $\mathbf{f}(\boldsymbol{X})$ given in equation (3.1). This illustrates the intuition behind the choice of the covariate balancing function $\mathbf{f}(\boldsymbol{X})$ in Step 3.

To derive the theoretical properties of the proposed HD-CBPS estimator, we impose the following conditions.

Assumption 3.1 (Sub-Exponential Condition)**.** Assume that $\epsilon_1 = Y(1) - b'(\boldsymbol{\alpha}^{*\top}\boldsymbol{X})$ and X_j satisfy $\|\epsilon_1\|_{\psi_1} \le C_\epsilon$ and $|X_j| \le C_X$ for any $1 \le j \le d$, where C_X and C_ϵ are two positive constants.

Assumption 3.2 (Bounded Mean Effect)**.** There exists an interval $[K_1, K_2]$ such that $K_1 \le \boldsymbol{\alpha}^{*\top}\boldsymbol{X}_i \le K_2$. Assume that $b(t)$ is third order continuously differentiable.

Assumption 3.3 (Singular value condition)**.** Let $\boldsymbol{\Delta} = (b'(\boldsymbol{\alpha}^{*\top}\boldsymbol{X})\mathbf{1}_s, b''(\boldsymbol{\alpha}^{*\top}\boldsymbol{X})\mathbf{1}_s^{\otimes 2})$ where $s = |S|$ and $\mathbf{1}_s$ is a vector of 1 with length s. Denote $\boldsymbol{\Gamma}_{SS} = \mathbb{E}(\boldsymbol{\Delta} \circ \bar{\boldsymbol{X}}_S^{\otimes 2})$, where $\circ$ represents the Hadamard product. There exists a constant $C > 0$ such that $\min_{|S| \lesssim s_2} \sigma_{\min}(\boldsymbol{\Gamma}_{SS}) \ge C$, where $\sigma_{\min}(\cdot)$ denotes the minimum singular value.

When compared to the sub-Gaussian condition in Assumption 2.1, we allow the error ϵ_1 to be sub-exponential in Assumption 3.1. This extension is necessary because many examples of generalized linear models (e.g., exponential regression and Poisson regression) satisfy the sub-exponential condition but not sub-Gaussian condition. In addition, Assumption 3.2 is a mild condition, stating that the mean function of the outcome model is bounded. Finally, Assumption 3.3 is a technical condition which is used to show that with probability tending to one the gradient of the estimating function $\boldsymbol{g}_n(\boldsymbol{\gamma})$ is invertible. Such a condition is commonly used in the estimating equation literature (see e.g., Theorem 5.21 of Van der Vaart, 1998).

The following theorem, which is a counterpart of Theorem 2.1 for generalized linear models, establishes the asymptotic normality of $\widehat{\mu}_1$ when the outcome variable follows a generalized linear model.

Theorem 3.1. Suppose that Assumptions 2.2, 2.3, 3.1, 3.2, and 3.3 hold. In addition, assume $\max(s_1, s_2) \log d \log n / n^{1/2} = o(1)$. Then, if we take $\lambda \asymp \lambda' \asymp \sqrt{\log d / n}$, we have,

$$\widehat{\mu}_1 - \mu_1^* \;=\; \mathbb{P}_n \left[\frac{T}{\pi^*} (Y(1) - b'(\boldsymbol{\alpha}^{*\top} \boldsymbol{X})) + b'(\boldsymbol{\alpha}^{*\top} \boldsymbol{X}) - \mu_1^* \right] + o_p(n^{-1/2}).$$

Therefore, $\widehat{\mu}_1$ achieves the following semiparametric efficiency bound,

$$\mathbb{E} \left[\frac{1}{\pi^*} \mathbb{E}(\epsilon_1^2 \mid \boldsymbol{X}) + \{ b'(\boldsymbol{\alpha}^{*\top} \boldsymbol{X}) - \mu_1^* \}^2 \right].$$

Proof is given in Appendix A.4. The scaling condition, i.e., $\max(s_1, s_2) \log d \log n / n^{1/2} = o(1)$, of this theorem is identical to that of Theorem 2.1 up to a logarithmic factor of n, due to the sub-exponential condition on ϵ_1. Although not described here, the estimation of the ATE and ATT can be done by following the approaches similar to those proposed in Section 2.4.

4 Simulation Studies

In this section, we conduct simulation studies to evaluate the finite sample performance of the proposed HD-CBPS methodology. We consider the following data generating processes. First, we generate the d dimensional covariate $\boldsymbol{X}_i \sim N(0, \boldsymbol{\Sigma})$ where $\Sigma_{jk} = \rho^{|j-k|}$. We generate the binary treatment T_i using the logistic regression model of the form, $\pi(\boldsymbol{X}_i) = 1 - 1/(1 + \exp(-X_{i1} + 0.5X_{i2} - 0.25X_{i3} - 0.1X_{i4}))$. For the potential outcomes, we consider both linear and logistic regression models as specified below. The observed outcome is given by $Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$.

In the simulations, we vary the sample size n from 100 to 1,000 and dimension d from 10 to 2,000. The simulation is repeated 500 times under each setting. Throughout the simulation studies, whenever possible, we compare our method (HD-CBPS) to the approximate residual balancing (RB) method (Athey et al., 2016) as well as the regularized AIPW (AIPW) method (Farrell, 2015). For the sake of comparison, we use the Lasso penalty in both HD-CBPS and AIPW methods, and all tuning parameters are determined by the 5 fold cross-validation. For the RB method, we use their default values of the tuning parameters in the R package balanceHD.

4.1 Linear Outcome Model

We first consider the setting in which the potential outcomes are generated from the linear regression models:

$$\begin{aligned} Y_i(1) &= 2 + 0.137X_{i7} + 0.137X_{i8} + 0.137X_{i9} + \epsilon_{1i}, \\ Y_i(0) &= 1 + 0.291X_{i5} + 0.291X_{i6} + 0.291X_{i7} + 0.291X_{i8} + 0.291X_{i9} + \epsilon_{0i}, \end{aligned}$$

where ϵ_{1i} and ϵ_{0i} are independent standard normal random variables. Under this setting, we consider the following three scenarios. In the first scenario, we assume that the propensity score model and outcome models are both correctly specified. In the second scenario, the outcome models are correctly specified but the propensity score is misspecified. The misspecification is due to the

	$n = 100$			$n = 500$			$n = 1000$		
d	HD-CBPS	RB	AIPW	HD-CBPS	RB	AIPW	HD-CBPS	RB	AIPW
(1) Both models are correct									
10	**0.2163**	0.2301	0.2209	0.1163	**0.1067**	0.1470	**0.0707**	0.0720	0.0950
100	**0.2272**	0.2421	0.2273	0.0965	0.1008	**0.0928**	0.0765	0.0787	**0.0692**
500	0.3262	**0.3132**	0.3859	**0.0924**	0.0997	0.0929	**0.0729**	0.0810	0.1009
1000	0.2083	0.2413	**0.2072**	0.1072	0.1240	**0.0991**	**0.0817**	0.0894	0.0992
2000	0.2327	0.2212	**0.2058**	**0.1053**	0.1199	0.1201	**0.0716**	0.0760	0.0903
(2) Propensity score model is correct									
10	0.2230	0.2348	**0.2174**	0.0969	**0.0948**	0.1317	**0.0714**	0.0734	0.0934
100	0.2163	0.2185	**0.2104**	**0.1007**	0.1084	0.1116	0.0720	0.0784	**0.0688**
500	**0.3255**	0.3256	0.3742	**0.0884**	0.1000	0.0900	**0.0731**	0.0852	0.1063
1000	0.2273	0.2462	**0.2239**	0.1004	0.1264	**0.0912**	**0.0892**	0.0981	0.1074
2000	0.2232	0.2348	**0.2061**	**0.1082**	0.1260	0.1266	**0.0764**	0.0844	0.0924
(3) Both models are misspecified									
10	**0.2386**	0.2462	0.2614	**0.1006**	0.1016	0.1120	**0.0655**	0.0661	0.0724
100	**0.2385**	0.2556	0.2422	0.1002	0.1052	**0.0988**	**0.0647**	0.0680	0.0693
500	**0.3324**	0.3400	0.3696	**0.0897**	0.1012	0.0985	0.0790	0.0861	**0.0765**
1000	0.2226	0.2369	**0.2209**	**0.1327**	0.1629	0.1625	**0.0725**	0.0822	0.0765
2000	0.2108	0.2157	**0.1974**	**0.1105**	0.1231	0.1370	**0.0843**	0.0878	0.0972

Table 1: Standardized root-mean-squared error for the estimation of the average treatment effect. Three methods – high-dimensional CBPS (HD-CBPS), approximate residual balancing (RB), and regularized augmented inverse probability weighting (AIPW) – are compared where d in the first column represents the number of dimensions. The bold letters indicate the best performance within a given simulation setting. Three scenarios are considered: (1) both propensity and outcome models are correct, (2) propensity score model is correct but outcome model is misspecified, and (3) both models are misspecified.

use of transformed variables $\boldsymbol{X}_{mis} = (\exp(X_1/2), X_2(1+\exp(X_1))^{-1}+10, (X_1X_3/25+0.6)^3, (X_2+X_4+20)^2, X_5, \cdots, X_d)$ rather than the original covariates $\boldsymbol{X}_i$. Finally, we consider a scenario where both the outcome and propensity score models are misspecified using the transformed covariates, $\boldsymbol{X}_{mis} = (\exp(X_1/2), X_2(1+\exp(X_1))^{-1}+10, (X_1X_3/25+0.6)^3, (X_2+X_4+20)^2, X_6, \exp(X_6+X_7), X_9^2, X_7^3-20, X_9, \cdots X_d)$. The way in which model misspecification is induced in our simulation studies follow the work of Kang and Schafer (2007) who evaluate the empirical performance of the regularized AIPW estimator in low dimensional settings.

Table 1 shows the standardized root mean squared error (RMSE), $\sqrt{\mathbb{E}(\widehat{\mu}-\mu)^2}/\mu$, for the estimation of the ATE under the three scenarios. We find that the HD-CBPS method outperforms the competing methods especially for the largest sample size. This pattern is observed consistently across the three simulation settings. The fact that the HD-CBPS performs well under misspecification provides empirical support for the robustness property established in Proposition 2.2. Finally,

d	$n = 100$		$n = 500$		$n = 1000$	
10	0.9450	(0.8550)	0.9500	(0.3674)	0.9550	(0.2766)
100	0.9000	(0.7649)	0.9450	(0.3767)	0.9700	(0.2747)
500	0.9000	(0.8027)	0.9300	(0.3800)	0.9650	(0.2617)
1000	0.9350	(0.7268)	0.9700	(0.3718)	0.9350	(0.2697)
2000	0.9050	(0.7882)	0.9450	(0.4000)	0.9350	(0.2609)

Table 2: Empirical coverage probability of the proposed 95% confidence interval for the average treatment effect when both propensity score model and outcome models are correctly specified. Numbers in parentheses are the averaged length of confidence intervals, while d in the first column represents the number of dimensions.

we examine the accuracy of the proposed confidence interval in Table 2. While the coverage probability ranges from 89% to 97%, the results are generally reasonable. As expected, the coverage probability tends to be more accurate when the sample size is greater.

In summary, the proposed HD-CBPS estimator tends to have a smaller mean squared error, is more robust to model misspecification, and exhibits accurate coverage probability in finite samples. Our results are consistent with the empirical findings of Imai and Ratkovic (2014) and Fan et al. (2016) that covariate balancing tends to outperform the regularized AIPW estimator in low dimensional settings. Our simulation studies indicate that the same conclusion appears to hold in high dimensional settings as well.

4.2 Logistic Outcome Model

Next, we consider the binary outcome and assume that the potential outcomes are generated by the following logistic regression models,

$$\begin{aligned}
\mathbb{P}(Y_i(1) = 1 \mid \boldsymbol{X}_i) &= 1 - 1/(1 + \exp(2 + 0.137X_{i7} + 0.137X_{i8} + 0.137X_{i9})), \\
\mathbb{P}(Y_i(0) = 1 \mid \boldsymbol{X}_i) &= 1 - 1/(1 + \exp(1 + 0.291X_{i5} + 0.291X_{i6} + 0.291X_{i7} + 0.291X_{i8} + 0.291X_{i9})).
\end{aligned}$$

When the outcome variable is binary, the approximate residual balancing method is not directly applicable. Thus, we only compare the HD-CBPS method with the AIPW method. For the logistic outcome model, we focus on the standardized mean squared error $\sqrt{\mathbb{E}(\widehat{\mu} - \mu)^2}/\mu$ when both the propensity score model and outcome models are correctly specified. The results under misspecified models demonstrate the patterns similar to those found under the linear outcome model and therefore are omitted. Table 3 shows that the proposed HD-CBPS method outperforms the regularized AIPW method especially when the sample size is large. Finally, Table 4 confirms that the proposed confidence intervals under the logistic outcome model have accurate coverage probability in a finite sample.

	$n = 100$		$n = 500$		$n = 1000$	
d	HD-CBPS	AIPW	HD-CBPS	AIPW	HD-CBPS	AIPW
10	0.0947	**0.0908**	**0.0398**	0.0441	**0.0252**	0.0297
100	**0.0745**	0.0759	**0.0352**	0.0367	**0.0239**	0.0252
500	**0.1075**	0.1082	**0.0351**	0.0354	**0.0303**	0.0367
1000	**0.0729**	0.0730	**0.0350**	0.0358	**0.0294**	0.0320
2000	**0.2113**	0.2144	**0.0357**	0.0378	**0.0232**	0.0255

Table 3: Standardized root-mean-squared error for the estimation of the ATE under the logistic outcome model. Both the propensity score and outcome models are correctly specified. The first column d represents the number of dimensions.

d	$n = 100$		$n = 500$		$n = 1000$	
10	0.9500	(0.3060)	0.9300	(0.1458)	0.9550	(0.1076)
100	0.9300	(0.3171)	0.9600	(0.1488)	0.9400	(0.0998)
500	0.9100	(0.2842)	0.9350	(0.1397)	0.9750	(0.1002)
1000	0.9300	(0.3618)	0.9350	(0.1408)	0.9450	(0.1037)
2000	0.9550	(0.3262)	0.9350	(0.1430)	0.9550	(0.0982)

Table 4: Empirical coverage probability of the proposed 95% confidence interval for the ATE under the logistic outcome model. Numbers in parentheses are the averaged length of confidence intervals.

5 Empirical Illustration

For empirical illustration, we consider a dataset obtained from the first two waves of Jennings' and Niemi's Political Socialization Panel Study, which is originally analyzed by Kam and Palmer (2008). One purpose of this study is to understand the effect of higher education on political participation. The dataset consists of $1,051$ randomly selected high school seniors in the class of 1965. The information about each sample is collected via in-person interviews in the first wave of the study, which we treat as pre-treatment covariates. The second wave of the study conducted in 1973 collects the outcome variable, political participation, as well as the dichotomous treatment variable, college attendance.

For the purpose of comparison, we follow the original study (Kam and Palmer, 2008) and use 81 pre-treatment covariates, which include gender, race, club participation, and academic performance. Since many of the covariates are categorical variables with more than two levels, we create an indicator variable that represents each level. Therefore, a total of 204 pre-treatment variables are used in the propensity score and outcome models. The outcome variable represents an index of adult political participation, which is equal to the sum of eight acts including the turnout in the 1972 presidential election, attending campaign rallies, making a donation to a campaign, and displaying a campaign button and bumper sticker. Since this variable takes an integer value ranging

	HD-CBPS	CBPS	AIPW
Overall (ATE)	0.8293	1.0163	0.8796
	(0.1247)	(0.2380)	(0.1043)
Overall (ATT)	0.8439	1.1232	
	(0.1420)	(0.3094)	
Whites (ATE)	0.8445		0.8977
	(0.1279)		(0.1089)

Table 5: The estimated average effects of college attendance on political participation. The estimates based on the proposed HD-CBPS methodology are compared with those of the original CBPS estimator and the regularized augmented inverse probability weighted estimator (AIPW). Standard errors appear in parentheses.

from zero to eight, we use the binomial logistic regression for the outcome model. The propensity score model is assumed to be the logistic regression. We then estimate both the ATE and ATT of college attendace on political participation (the number of treated observations is 675).

We apply three methods to analyze this dataset, the proposed HD-CBPS methodology, the regularized AIPW method (Farrell, 2015) and the original CBPS method (Imai and Ratkovic, 2014). The estimation procedures for the first two methods are identical to those described in the simulation studies. For the original CBPS methodology, we note that it is designed for the linear outcome model, which does not provide an ideal fit to the outcome variable of interest. In addition, we use the bootstrap method to approximate the standard error of the estimator based on the CBPS method.

The results are shown in Table 5. All three methods suggest that the overall ATE of college education on political participation is positive and statistically significant. The ATE estimates and their associated standard errors based on the regularized methods (i.e., HD-CBPS and AIPW) are quite similar to each other and somewhat smaller than that of the CBPS. More importantly, for both the ATE and the ATT, the regularized estimates have much smaller standard errors than the original CBPS. There are at least two reasons for this difference in standard errors. First, as shown in Section 3.2, the HD-CBPS methodology uses a different covariate balancing estimating equation when the outcome model is nonlinear, achieving the semiparametric efficiency bound. Second, the original CBPS methodology tends to be unstable when balancing a large number of covariates (204 in this case). Thus, the proposed HD-CBPS method improves the existing covariate balancing methods when the outcome model belongs to the class of generalized linear models and the number of covariates is large.

We also apply the HD-CBPS and AIPW methods to the subsample of whites separately. Among a total of 1,051 respondents, there are 966 white respondents. The results appear at the last row of Table 5. Again, the estimates of the two methods are similar so are the standard errors. The HD-CBPS method selects 27 covariates, which includes a total of 26 covariates selected by the regularized AIPW estimator.

6 Conclusion

In this paper, we introduce the weak covariate balancing property and propose a post-penalization covariate balancing propensity score method to infer the ATE in the high dimensional setting. The proposed method is further extended to handle the binary or count data. We show that, in high dimensions, the resulting estimator of ATE is sample bounded, root-n consistent, asymptotically normal and semiparametrically efficient. Moreover, the estimator is robust under misspecified propensity score models in high dimensions. Through extensive numerical studies, we show that our method tends to have smaller mean squared errors than the existing methods.

A Proofs

A.1 Proof of Theorem 2.1

To start the proof of Theorem 2.1, we make the following minor modifications on the estimation of $\boldsymbol{\gamma}$ in step 3. Define

$$\widetilde{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}\in\Omega} \|\boldsymbol{g}_n(\boldsymbol{\gamma})\|_2^2, \quad \text{where} \quad \Omega = \{\boldsymbol{\gamma} \in \mathbb{R}^{|\widetilde{S}|} : \|\boldsymbol{\gamma} - \widehat{\boldsymbol{\beta}}_{\widetilde{S}}\|_1 \le \delta/\log n\}$$

for some small constant $\delta > 0$. Here, we introduce a parameter set Ω for $\boldsymbol{\gamma}$, which guarantees the existence of a minimizer $\widetilde{\boldsymbol{\gamma}}$ within the interior of Ω with probability tending to one. By using this modification, we can avoid to impose further unnecessary technical assumptions. In practice, we find that without the modification the estimator $\widetilde{\boldsymbol{\gamma}}$ defined in step 3 is still very close to $\widehat{\boldsymbol{\beta}}_{\widetilde{S}}$. This suggests that the estimator automatically lies within the set Ω. So, this modification has little practical implication.

Lemma A.1. Under the assumptions in Theorem 2.1,

$$\|\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 = O_p\Big(s_2\sqrt{\frac{\log d}{n}}\Big), \quad \mathbb{P}_n[\boldsymbol{X}^\top(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)]^2 = O_p\Big(\frac{s_2 \log d}{n}\Big).$$

Noting that $T_i(Y_i - \boldsymbol{\alpha}^{*\top}\boldsymbol{X}_i) = T_i(Y_i(1) - \boldsymbol{\alpha}^{*\top}\boldsymbol{X}_i)$, by the ignorability of treatment assignment, we have

$$\mathbb{E}(T_i(Y_i - \boldsymbol{\alpha}^{*\top}\boldsymbol{X}_i)) = \mathbb{E}(\pi_i^*\mathbb{E}(Y_i(1) - \boldsymbol{\alpha}^{*\top}\boldsymbol{X}_i \mid \boldsymbol{X}_i)) = 0.$$

By the overlap Assumption 2.2, $\pi_i^* \ge c_0 > 0$, together with Assumption 2.3, we can show that $\lambda_{\min}(\mathbb{E}(\pi_i^*\boldsymbol{X}_i^{\otimes 2})) \ge Cc_0 > 0$. It is easily seen that the remaining proof follows from Lemma D.3 of Ning and Liu (2014). We omit the details.

Lemma A.2. Under the assumptions in Theorem 2.1,

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O_p\Big(\sqrt{\frac{(s_1+s_2)s_1\log d + s_2^2\log s_2}{n}}\Big),$$

and

$$\mathbb{P}_n[\boldsymbol{X}^\top(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)]^2 = O_p\Big(\frac{s_1\log d + s_2\log s_2}{n}\Big).$$

Proof of Lemma A.2. The proof contains the following three steps:

(i) $|\widetilde{S}| \le Cs_2$, where C is a positive constant.

(ii) $\|\widehat{\boldsymbol{\beta}}_{\widetilde{S}^c} - \boldsymbol{\beta}^*_{\widetilde{S}^c}\|_1 = O_p(s_1\sqrt{\log d/n})$, and $\mathbb{P}_n[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top_{\widetilde{S}^c}\boldsymbol{X}_{\widetilde{S}^c}]^2 = O_p(s_1\log d/n)$.

(iii) $\|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_1 = O_p(\sqrt{(s_2^2\log s_2 + s_2 s_1\log d)/n})$, and

$$\mathbb{P}_n[(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)^\top\boldsymbol{X}_{\widetilde{S}}]^2 = O_p\Big(\frac{s_1\log d + s_2\log s_2}{n}\Big).$$

We first show (i). By the KKT condition of the penalized least square regression for $\widetilde{\boldsymbol{\alpha}}$, we have $\mathbb{P}_n X_j(Y - \boldsymbol{X}^\top \widetilde{\boldsymbol{\alpha}}) = -\lambda' \mathrm{sign}(\widetilde{\alpha}_j)$ for any $j \in \mathrm{supp}(\widetilde{\boldsymbol{\alpha}})$. Then, we have

$$\begin{aligned}\lambda' |\widetilde{S}|^{1/2} &= \|\mathbb{P}_n \boldsymbol{X}_{\widetilde{S}}(Y - \boldsymbol{X}^\top \widetilde{\boldsymbol{\alpha}})\|_2 \le \|\mathbb{P}_n \boldsymbol{X}_{\widetilde{S}} \epsilon_1\|_2 + \|\mathbb{P}_n \boldsymbol{X}_{\widetilde{S}} \boldsymbol{X}^\top (\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\|_2 \\ &\le |\widetilde{S}|^{1/2} \|\mathbb{P}_n \boldsymbol{X}_{\widetilde{S}} \epsilon_1\|_\infty + \{\mathbb{P}_n [\boldsymbol{X}^\top (\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)]^2\}^{1/2} \|\mathbb{P}_n \boldsymbol{X}_{\widetilde{S}} \boldsymbol{X}_{\widetilde{S}}^\top\|_2^{1/2} \\ &\le C |\widetilde{S}|^{1/2} \sqrt{\log |\widetilde{S}|/n} + C \|\mathbb{P}_n \boldsymbol{X}_{\widetilde{S}} \boldsymbol{X}_{\widetilde{S}}^\top\|_2^{1/2} \sqrt{s_2 \log d/n}, \end{aligned} \tag{A.1}$$

where the last step follows from Lemma A.1 and $\|\mathbb{P}_n \boldsymbol{X}_{\widetilde{S}} \epsilon_1\|_\infty \le \max_{|S| \le |\widetilde{S}|} \max_{j \in S} \|\mathbb{P}_n X_j \epsilon_1\|_\infty$. Since $\|\epsilon_1\|_{\psi_2} \le C_\epsilon$ and $\|X_j\|_{\psi_2} \le C_X$ for any $1 \le j \le d$, we have $\|X_j \epsilon_1\|_{\psi_1} \le 2 C_\epsilon C_X$. The Bernstein inequality for sub-exponential random variables (Lemma K.2 of Ning and Liu (2014)) yields,

$$\mathbb{P}\Big(\frac{1}{n} \sum_{i=1}^n X_{ij} \epsilon_{1i} > t\Big) \le 2 \exp\Big[- C'' \min\Big(\frac{t^2}{4 C_\epsilon^2 C_X^2}, \frac{t}{2 C_\epsilon C_X}\Big)\Big],$$

where C'' is a universal constant. Applying the union bound argument and choose $t = \sqrt{\log |\widetilde{S}|/n}$, we can obtain $\|\mathbb{P}_n \boldsymbol{X}_{\widetilde{S}} \epsilon_1\|_\infty = O_p(\sqrt{\log |\widetilde{S}|/n})$. By equation (A.1) and $\lambda' \asymp \sqrt{\log d/n}$, we have

$$|\widetilde{S}|^{1/2} \le C \|\mathbb{P}_n \boldsymbol{X}_{\widetilde{S}} \boldsymbol{X}_{\widetilde{S}}^\top\|_2 s_2^{1/2}.$$

Since the sparse eigenvalue is sub-linear (e.g., Yang et al. (2014)), there exists a constant $C' > 0$ such that $\|\mathbb{P}_n \boldsymbol{X}_{\widetilde{S}} \boldsymbol{X}_{\widetilde{S}}^\top\|_2 \le C'$. Thus, equation (A.1) implies (i).

To show (ii), note that $\|\widehat{\boldsymbol{\beta}}_{\widetilde{S}^c} - \boldsymbol{\beta}^*_{\widetilde{S}^c}\|_1 \le \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O_p(s_1 \sqrt{\log d/n})$, where the last step follows from Lemma E.1 of Ning and Liu (2014). In addition, $\lambda_{\max}(\mathbb{P}_n \boldsymbol{X}_{\widetilde{S}} \boldsymbol{X}_{\widetilde{S}}^\top) = O_p(1)$. To see this, by Weyl's inequality,

$$\begin{aligned}|\lambda_{\max}(\mathbb{P}_n \boldsymbol{X}_{\widetilde{S}} \boldsymbol{X}_{\widetilde{S}}^\top) - \lambda_{\max}(\mathbb{P} \boldsymbol{X}_{\widetilde{S}} \boldsymbol{X}_{\widetilde{S}}^\top)| &\le \|(\mathbb{P}_n - \mathbb{P}) \boldsymbol{X}_{\widetilde{S}} \boldsymbol{X}_{\widetilde{S}}^\top\|_2 \le \max_{|S| \le C s_2} \|(\mathbb{P}_n - \mathbb{P}) \boldsymbol{X}_S \boldsymbol{X}_S^\top\|_2 \\ &\le C s_2 \|(\mathbb{P}_n - \mathbb{P}) \boldsymbol{X}_S \boldsymbol{X}_S^\top\|_{\max} = O_p(s_2 \sqrt{\log s_2/n}),\end{aligned}$$

where in the last step we use the Bernstein inequality for sub-exponential random variable and the standard union bound argument. Since $\max_{|S| \le C s_2} \lambda_{\max}(\mathbb{P} \boldsymbol{X}_S \boldsymbol{X}_S^\top) \le 1/C$ by assumption, we obtain that $\lambda_{\max}(\mathbb{P}_n \boldsymbol{X}_{\widetilde{S}} \boldsymbol{X}_{\widetilde{S}}^\top) = O_p(1)$. For the second result in (ii), we have

$$\mathbb{P}_n[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\widetilde{S}^c}^\top \boldsymbol{X}_{\widetilde{S}^c}]^2 \le 2 \mathbb{P}_n[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \boldsymbol{X}]^2 + 2 \mathbb{P}_n[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\widetilde{S}}^\top \boldsymbol{X}_{\widetilde{S}}]^2 = O_p(s_1 \log d/n)$$

where the last step follows from Lemma E.1 of Ning and Liu (2014), and

$$\mathbb{P}_n[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\widetilde{S}}^\top \boldsymbol{X}_{\widetilde{S}}]^2 \le \|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\widetilde{S}}^\top\|_2^2 \lambda_{\max}(\mathbb{P}_n \boldsymbol{X}_{\widetilde{S}} \boldsymbol{X}_{\widetilde{S}}^\top) = O_p(s_1 \log d/n).$$

This completes the proof of (ii).

In the following, we aim to show (iii). For notational simplicity, let $\pi = \pi(\boldsymbol{\gamma}^{*\top} \boldsymbol{X}_{\widetilde{S}} + \widehat{\boldsymbol{\beta}}_{\widetilde{S}^c}^\top \boldsymbol{X}_{\widetilde{S}^c})$

and $\widetilde{\pi} = \pi(\widetilde{\boldsymbol{\gamma}}^\top \boldsymbol{X}_{\widetilde{S}} + \widehat{\boldsymbol{\beta}}_{\widetilde{S}^c}^\top \boldsymbol{X}_{\widetilde{S}^c})$. By the definition of $\widetilde{\boldsymbol{\gamma}}$, we have

$$\begin{aligned}
&\Big\|\mathbb{P}_n\Big(\frac{T}{\pi} - 1\Big)\boldsymbol{X}_{\widetilde{S}}\Big\|_2^2 \geq \Big\|\mathbb{P}_n\Big(\frac{T}{\widetilde{\pi}} - 1\Big)\boldsymbol{X}_{\widetilde{S}}\Big\|_2^2 \\
&= \Big\|\mathbb{P}_n\Big(\frac{T}{\pi} - 1\Big)\boldsymbol{X}_{\widetilde{S}}\Big\|_2^2 + \Big\|\mathbb{P}_n\Big(\frac{T}{\widetilde{\pi}} - \frac{T}{\pi}\Big)\boldsymbol{X}_{\widetilde{S}}\Big\|_2^2 + 2\mathbb{P}_n\Big(\frac{T}{\pi} - 1\Big)\boldsymbol{X}_{\widetilde{S}}\mathbb{P}_n\Big(\frac{T}{\widetilde{\pi}} - \frac{T}{\pi}\Big)\boldsymbol{X}_{\widetilde{S}}.
\end{aligned}$$

The first inequality comes from $\boldsymbol{\gamma}^* \in \Omega$, since $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O_p(s_1\sqrt{\log d/n})$. This yields

$$\Big\|\mathbb{P}_n\Big(\frac{T}{\widetilde{\pi}} - \frac{T}{\pi}\Big)\boldsymbol{X}_{\widetilde{S}}\Big\|_2^2 \leq -2\mathbb{P}_n\Big(\frac{T}{\pi} - 1\Big)\boldsymbol{X}_{\widetilde{S}}\mathbb{P}_n\Big(\frac{T}{\widetilde{\pi}} - \frac{T}{\pi}\Big)\boldsymbol{X}_{\widetilde{S}}. \tag{A.2}$$

Let π' denote the derivative of π evaluated at an intermediate value between $\boldsymbol{\gamma}^{*\top}\boldsymbol{X}_{\widetilde{S}} + \widehat{\boldsymbol{\beta}}_{\widetilde{S}^c}^\top \boldsymbol{X}_{\widetilde{S}^c}$ and $\widetilde{\boldsymbol{\gamma}}^\top \boldsymbol{X}_{\widetilde{S}} + \widehat{\boldsymbol{\beta}}_{\widetilde{S}^c}^\top \boldsymbol{X}_{\widetilde{S}^c}$. Then

$$\begin{aligned}
\Big\|\mathbb{P}_n\Big(\frac{T}{\widetilde{\pi}} - \frac{T}{\pi}\Big)\boldsymbol{X}_{\widetilde{S}}\Big\|_2^2 &= \Big\|\mathbb{P}_n \frac{T\pi'}{\widetilde{\pi}\pi}\boldsymbol{X}_{\widetilde{S}}^{\otimes 2}(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)\Big\|_2^2 \\
&\geq \|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2^2 \lambda_{\min}\Big(\mathbb{P}_n \frac{T\pi'}{\widetilde{\pi}\pi}\boldsymbol{X}_{\widetilde{S}}^{\otimes 2}\Big) \geq C\|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2^2 \lambda_{\min}\Big(\mathbb{P}_n T\boldsymbol{X}_{\widetilde{S}}^{\otimes 2}\Big),
\end{aligned}$$

for some constant $C > 0$. The last step follows from $\widetilde{\pi}_i \leq 1$ and $\pi_i \leq 1$ and $\pi_i' \geq C$, since

$$\max_{1\leq i\leq n} |\widetilde{\pi}_i' - \pi_i^{*'}| \leq \max_{1\leq i\leq n}\{\|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_1 \|\boldsymbol{X}_{i\widetilde{S}}\|_\infty + \|\widehat{\boldsymbol{\beta}}_{\widetilde{S}^c} - \boldsymbol{\beta}^*_{\widetilde{S}^c}\|_1 \|\boldsymbol{X}_{i\widetilde{S}^c}\|_\infty\} = o_p(1),$$

by the definition of Ω and the convergence rate of $\widehat{\boldsymbol{\beta}}$. It is easily seen that

$$|\lambda_{\min}(\mathbb{P}_n T\boldsymbol{X}_{\widetilde{S}}^{\otimes 2}) - \lambda_{\min}(\mathbb{P}T\boldsymbol{X}_{\widetilde{S}}^{\otimes 2})| \leq Cs_2\|(\mathbb{P}_n - \mathbb{P})T\boldsymbol{X}_{\widetilde{S}}^{\otimes 2}\|_{\max} = O_p(s_2\sqrt{\log s_2/n}).$$

Since $\max_{|S|\leq Cs_2} \lambda_{\min}(\mathbb{P}T\boldsymbol{X}_S^{\otimes 2}) \geq c_0 C$, we obtain that $\lambda_{\min}(\mathbb{P}_n T\boldsymbol{X}_{\widetilde{S}}^{\otimes 2})$ is lower bounded by a positive constant. This implies

$$\Big\|\mathbb{P}_n\Big(\frac{T}{\widetilde{\pi}} - \frac{T}{\pi}\Big)\boldsymbol{X}_{\widetilde{S}}\Big\|_2^2 \geq C\|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2^2. \tag{A.3}$$

Following the similar argument, we can show that the right hand side of equation (A.2) is bounded above by $2\|\boldsymbol{\Delta}_n\|_2\|A_n\|_2\|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2$, where

$$\boldsymbol{\Delta}_n = \mathbb{P}_n\Big(\frac{T}{\pi} - 1\Big)\boldsymbol{X}_{\widetilde{S}}, \quad A_n = \mathbb{P}_n \frac{T\pi'}{\widetilde{\pi}\pi}\boldsymbol{X}_{\widetilde{S}}^{\otimes 2}.$$

Similarly, we can show that $\|A_n\|_2 \leq C\lambda_{\max}(\mathbb{P}_n\boldsymbol{X}_{\widetilde{S}}^{\otimes 2}) \leq C'$ for some constants $C, C' > 0$. In addition, we decompose $\boldsymbol{\Delta}_n = I_n + II_n$, where

$$I_n = \mathbb{P}_n\Big(\frac{T}{\pi^*} - 1\Big)\boldsymbol{X}_{\widetilde{S}}, \quad II_n = -\mathbb{P}_n \frac{T\pi'}{\pi\pi^*}\boldsymbol{X}_{\widetilde{S}}\boldsymbol{X}_{\widetilde{S}^c}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\widetilde{S}^c},$$

where similarly π' is the derivative of π evaluated at some intermediate value. Thus, by the

Bernstein inequality and the union bound argument,

$$\|I_n\|_2 \le |\widetilde{S}|^{1/2}\Big\|\mathbb{P}_n\Big(\frac{T}{\pi^*}-1\Big)\boldsymbol{X}_{\widetilde{S}}\Big\|_\infty = O_p(\sqrt{s_2\log s_2/n}).$$

In addition,

$$\begin{aligned}
\|II_n\|_2 &\le \sup_{\|\boldsymbol{v}\|_2=1}\Big|\mathbb{P}_n\frac{T\pi'}{\pi\pi^*}\boldsymbol{v}^\top\boldsymbol{X}_{\widetilde{S}}\boldsymbol{X}_{\widetilde{S}^c}^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_{\widetilde{S}^c}\Big| \\
&\le \sup_{\|\boldsymbol{v}\|_2=1}\Big|\mathbb{P}_n\frac{T\pi'}{\pi\pi^*}[\boldsymbol{v}^\top\boldsymbol{X}_{\widetilde{S}}]^2\Big|^{1/2}\Big|\mathbb{P}_n\frac{T\pi'}{\pi\pi^*}[\boldsymbol{X}_{\widetilde{S}^c}^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_{\widetilde{S}^c}]^2\Big|^{1/2} \\
&\le C\lambda_{\max}^{1/2}(\mathbb{P}_n\boldsymbol{X}_{\widetilde{S}}\boldsymbol{X}_{\widetilde{S}}^\top)\cdot\Big|\mathbb{P}_n[\boldsymbol{X}_{\widetilde{S}^c}^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_{\widetilde{S}^c}]^2\Big|^{1/2} = O_p(\sqrt{s_1\log d/n}).
\end{aligned}$$

This implies $\|\boldsymbol{\Delta}\|_2 = O_p(\sqrt{(s_2\log s_2 + s_1\log d)/n})$. Combining with equations (A.3) and (A.2), we obtain that

$$\|\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2 = O_p(\sqrt{(s_2\log s_2+s_1\log d)/n}),$$

and

$$\|\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1 \le Cs_2^{1/2}\|\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2 = O_p(\sqrt{(s_2^2\log s_2 + s_2s_1\log d)/n}).$$

This completes the proof of (iii). Finally, we combine the results in (i), (ii) and (iii) to show the desired result:

$$\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1 = \|\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1 + \|(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_{\widetilde{S}^c}\|_1 = O_p\Big(\sqrt{\frac{(s_1+s_2)s_1\log d + s_2^2\log s_2}{n}}\Big),$$

and

$$\begin{aligned}
\mathbb{P}_n[\boldsymbol{X}^\top(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2 &\le 2\mathbb{P}_n[(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_{\widetilde{S}^c}^\top\boldsymbol{X}_{\widetilde{S}^c}]^2 + 2\mathbb{P}_n[(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)^\top\boldsymbol{X}_{\widetilde{S}}]^2 \\
&= O_p\Big(\frac{s_1\log d + s_2\log s_2}{n}\Big).
\end{aligned}$$

□

Finally, we start the proof of Theorem 2.1.

Proof of Theorem 2.1. By the rearrangement of terms, we have

$$\widehat{\mu}_1-\mu_1^* = \mathbb{P}_n\Big[\frac{T}{\pi^*}(Y(1)-K_1(\boldsymbol{X}))+K_1(\boldsymbol{X})-\mu_1^*\Big] + I_1 + I_2,$$

where

$$I_1 = \mathbb{P}_n\Big[\frac{T}{\widetilde{\pi}}-\frac{T}{\pi^*}\Big](Y(1)-K_1(\boldsymbol{X})),\quad I_2 = \mathbb{P}_n\Big[\frac{T}{\widetilde{\pi}}-1\Big]K_1(\boldsymbol{X}).$$

By (iii) in the proof of Lemma A.2, we have that with probability tending to one $\widetilde{\boldsymbol{\gamma}}$ belongs to the interior of Ω. The KKT condition implies $(\partial\boldsymbol{g}_n(\widetilde{\boldsymbol{\gamma}})/\partial\boldsymbol{\gamma})\boldsymbol{g}_n(\widetilde{\boldsymbol{\gamma}}) = 0$. As seen in the proof of Lemma A.2, $\partial\boldsymbol{g}_n(\widetilde{\boldsymbol{\gamma}})/\partial\boldsymbol{\gamma}$ is invertible with probability tending to one. Thus, we have $\boldsymbol{g}_n(\widetilde{\boldsymbol{\gamma}}) = 0$,

and therefore

$$\mathbb{P}_n\Big[\frac{T}{\widetilde{\pi}}-1\Big]\widetilde{\boldsymbol{\alpha}}_{\widetilde{S}}^{\top}\boldsymbol{X}_{\widetilde{S}}=\mathbb{P}_n\Big[\frac{T}{\widetilde{\pi}}-1\Big]\widetilde{\boldsymbol{\alpha}}^{\top}\boldsymbol{X}=0.$$

Once we can show that

$$\mathbb{P}_n\Big[\frac{T}{\widetilde{\pi}}-1\Big]\boldsymbol{X}^{\top}(\widetilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)=o_p(n^{-1/2}), \tag{A.4}$$

it yields $I_2=o_p(n^{-1/2})$. To show equation (A.4), by the Taylor theorem,

$$\begin{aligned}&\Big|\mathbb{P}_n\Big[\frac{T}{\widetilde{\pi}}-1\Big]\boldsymbol{X}^{\top}(\widetilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)\Big|=\Big|\mathbb{P}_n\Big[\frac{T}{\pi^*}-1\Big]\boldsymbol{X}^{\top}(\widetilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)\Big|+\Big|\mathbb{P}_n\frac{T\pi'(t)}{\pi^{*2}}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)^{\top}\boldsymbol{X}\boldsymbol{X}^{\top}(\widetilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)\Big|\\&\le\Big\|\mathbb{P}_n\Big[\frac{T}{\pi^*}-1\Big]\boldsymbol{X}^{\top}\Big\|_{\infty}\|\widetilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*\|_1+\frac{1}{c_0^2}\Big\{\mathbb{P}_n[\boldsymbol{X}^{\top}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2\Big\}^{1/2}\Big\{\mathbb{P}_n[\boldsymbol{X}^{\top}(\widetilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)]^2\Big\}^{1/2},\end{aligned} \tag{A.5}$$

where $\pi'(\cdot)$ is the derivative of $\pi(\cdot)$ and evaluated at some intermediate value, and it is easily seen that $|\pi'(t)|\le 1$, and the last step follows from the Cauchy inequality. Since $|T/\pi^*-1|\le 1/c_0$ and X_j is a sub-Gaussian random variable, we have that $[\frac{T}{\pi^*}-1]X_j$ is a sub-exponential random variable with $\|[\frac{T}{\pi^*}-1]X_j\|_{\psi_1}\le 2C_X/c_0$. By the Bernstein inequality and the union bound argument, we have

$$\Big\|\mathbb{P}_n\Big[\frac{T}{\pi^*}-1\Big]\boldsymbol{X}^{\top}\Big\|_{\infty}=O_p\Big(\sqrt{\frac{\log d}{n}}\Big).$$

Combining Lemmas A.1 and A.2 with equation (A.5), we obtain that

$$\mathbb{P}_n\Big[\frac{T}{\widetilde{\pi}}-1\Big]\boldsymbol{X}^{\top}(\widetilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)=O_p\Big(\frac{s_2\log d}{n}+\frac{s_1^{1/2}s_2^{1/2}\log d}{n}\Big)=o_p(n^{-1/2}),$$

where the last step follows from $\max(s_1,s_2)\log d/n^{1/2}=o(1)$. This completes the proof of equation (A.4).

In the following, we will focus on I_1. Let $\epsilon_1=Y(1)-K_1(\boldsymbol{X})$. Again, by the Taylor theorem,

$$\begin{aligned}|I_1|&\le\Big|\mathbb{P}_n\frac{T\pi'}{\pi^{*2}}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)^{\top}\boldsymbol{X}\epsilon_1\Big|+\Big(\frac{1}{c_0^2}+\frac{2}{c_0^3}\Big)\Big|\mathbb{P}_n[(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)^{\top}\boldsymbol{X}]^2\Big|\max_{1\le i\le n}|\epsilon_{1i}|\\&\le\Big\|\mathbb{P}_n\frac{T\pi'}{\pi^{*2}}\epsilon_1\boldsymbol{X}\Big\|_{\infty}\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1+\Big(\frac{1}{c_0^2}+\frac{2}{c_0^3}\Big)\Big|\mathbb{P}_n[(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)^{\top}\boldsymbol{X}]^2\Big|\max_{1\le i\le n}|\epsilon_{1i}|.\end{aligned} \tag{A.6}$$

Here, π' is the derivative of π evaluated at the truth. Since $\frac{T\pi'}{\pi^{*2}}\le 1/c_0^2$, ϵ_1 and X_j are both sub-Gaussian random variable, this implies that $|\frac{T\pi'}{\pi^{*2}}\epsilon_1X_j|$ is a sub-exponential random variable with $\|\frac{T\pi'}{\pi^{*2}}\epsilon_1X_j\|_{\psi_1}\le 2C_XC_\epsilon/c_0^2$. By the Bernstein inequality and the union bound argument, it is easy to show that

$$\Big\|\mathbb{P}_n\frac{T\pi'}{\pi^{*2}}\epsilon_1\boldsymbol{X}\Big\|_{\infty}=O_p\Big(\sqrt{\frac{\log d}{n}}\Big).$$

In addition, it is easy to show that $\max_{1\le i\le n}|\epsilon_{1i}|$ (i.e., the maximum of independent sub-Gaussian random variables) is of order $\sqrt{\log n}$. Together with Lemma A.2, we have shown that

$$|I_1|=O_p\Big(\frac{s_2\log d+s_2^{1/2}s_1^{1/2}\log d}{n}+\frac{s_1\log d\sqrt{\log n}+s_2\log s_2\sqrt{\log n}}{n}\Big)=o_p(n^{-1/2}).$$

Thus, we have $\widehat{\mu}_1 - \mu_1^* = \frac{1}{n}\sum_{i=1}^n S_i + \Delta$, where $S_i = \frac{T_i}{\pi^*}(Y_i(1) - K_1(\boldsymbol{X}_i)) + K_1(\boldsymbol{X}_i) - \mu_1^*$ and $|\Delta| = o_p(n^{-1/2})$. Following the similar derivation in (Hahn, 1998), it is easy to verify that $\frac{T}{\pi^*}(Y(1) - K_1(\boldsymbol{X})) + K_1(\boldsymbol{X}) - \mu_1^*$ corresponds to the efficient score function for μ_1^*. This implies the semiparametric efficiency of $\widehat{\mu}_1$.

While the Lemmas A.1 and A.2 are presented in an asymptotic manner, it is easy to show the following non-asymptotic result: $|\Delta| \le C\max(s_1, s_2)\log d\sqrt{\log n}/n$ with probability at least $1 - d^{-1} - n^{-1}$. Then, for any $t > 0$, we have

$$\begin{aligned}\Big|\mathbb{P}(n^{1/2}(\widehat{\mu}_1 - \mu_1^*)V^{-1/2} \le t) - \Phi(t)\Big| &= \Big|\mathbb{P}(n^{-1/2}\sum_{i=1}^n S_i V^{-1/2} \le t - n^{1/2}\Delta V^{-1/2}) - \Phi(t)\Big| \\ &\le \max\Big(\Big|\mathbb{P}(n^{-1/2}\sum_{i=1}^n S_i V^{-1/2} \le t - \delta_n) - \Phi(t)\Big|, \\ &\qquad \Big|\mathbb{P}(n^{-1/2}\sum_{i=1}^n S_i V^{-1/2} \le t + \delta_n) - \Phi(t)\Big|\Big) + d^{-1} + n^{-1}\end{aligned}$$

where $\delta_n = Cc^{-1/2}\max(s_1, s_2)\log d\sqrt{\log n}/n^{1/2}$. In addition, the third moment of S_i is bounded and Berry-Esseen theorem implies

$$\begin{aligned}&\Big|\mathbb{P}(n^{-1/2}\sum_{i=1}^n S_i V^{-1/2} \le t - \delta_n) - \Phi(t)\Big| \\ &\le \Big|\mathbb{P}(n^{-1/2}\sum_{i=1}^n S_i V^{-1/2} \le t - \delta_n) - \Phi(t - \delta_n)\Big| + |\Phi(t) - \Phi(t - \delta_n)| \le \frac{C}{n^{1/2}} + C\delta_n,\end{aligned}$$

for some constant $C > 0$. Similarly, we can show that

$$\Big|\mathbb{P}(n^{-1/2}\sum_{i=1}^n S_i V^{-1/2} \le t + \delta_n) - \Phi(t)\Big| \le \frac{C}{n^{1/2}} + C\delta_n.$$

Thus, we have

$$\sup_{t\in\mathbb{R}}\Big|\mathbb{P}(n^{1/2}(\widehat{\mu}_1 - \mu_1^*)V^{-1/2} \le t) - \Phi(t)\Big| \le C\Big(\frac{1 + \max(s_1, s_2)\log d\sqrt{\log n}}{n^{1/2}} + \frac{1}{d}\Big),$$

for some constant $C > 0$. □

A.2 Proof of Lemma 2.1

We first show that

$$|\widehat{\sigma}_1^2 - \sigma_1^2| = O_p\Big(\frac{\max(s_1, s_2)\sqrt{\log d\log n}}{n^{1/2}}\Big). \tag{A.7}$$

Recall that $\widehat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^n \frac{T_i}{\widetilde{\pi}_i}(Y_i - \widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X}_i)^2$. Then $|\widehat{\sigma}_1^2 - \sigma_1^2| \le I_1 + I_2$, where

$$I_1 = \Big|\frac{1}{n}\sum_{i=1}^n \frac{T_i}{\pi_i^*}(Y_i - \widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X}_i)^2 - \sigma_1^2\Big|, \quad I_2 = \Big|\frac{1}{n}\sum_{i=1}^n \frac{T_i(\widetilde{\pi}_i - \pi_i^*)}{\pi_i^*\widetilde{\pi}_i}(Y_i - \widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X}_i)^2\Big|.$$

We can further decompose I_1 as follows

$$\begin{aligned} I_1 &\le \Big|\frac{1}{n}\sum_{i=1}^n \frac{T_i}{\pi_i^*}\epsilon_{1i}^2 - \sigma_1^2\Big| + \Big|\frac{2}{n}\sum_{i=1}^n \frac{T_i}{\pi_i^*}\epsilon_{1i}\boldsymbol{X}_i^\top(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\Big| + \Big|\frac{1}{n}\sum_{i=1}^n \frac{T_i}{\pi_i^*}\{\boldsymbol{X}_i^\top(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\}^2\Big| \\ &\le \Big|\frac{1}{n}\sum_{i=1}^n \frac{T_i}{\pi_i^*}\epsilon_{1i}^2 - \sigma_1^2\Big| + \Big\|\frac{2}{n}\sum_{i=1}^n \frac{T_i}{\pi_i^*}\epsilon_{1i}\boldsymbol{X}_i^\top\Big\|_\infty \|\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 + \Big|\frac{1}{n}\sum_{i=1}^n \frac{1}{c_0}\{\boldsymbol{X}_i^\top(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\}^2\Big|. \end{aligned}$$

Applying the Bernstein inequality for the first two terms and Lemma A.1, we obtain that $I_1 = O_p(s_2 \log d/n + 1/n^{1/2})$. For I_2, we can easily show that there exists a constant $C > 0$ such that

$$I_2 \le C\Big\|\frac{1}{n}\sum_{i=1}^n \epsilon_{1i}^2\boldsymbol{X}_i\Big\|_\infty \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \le C\sqrt{\log n}\Big\|\frac{1}{n}\sum_{i=1}^n |\epsilon_{1i}\boldsymbol{X}_i|\Big\|_\infty \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1,$$

where $|\boldsymbol{X}_i| = (|X_{i1}|, ..., |X_{in}|)$. Since $\epsilon_{1i}X_{ij}$ is sub-exponential and $\mathbb{E}(|\epsilon_{1i}X_{ij}|)$ is uniformly bounded, we can again apply the Bernstein inequality and the union bound. Together with Lemma A.2, we obtain $I_2 = O_p(\max(s_1, s_2)\sqrt{\log d \log n}/n^{1/2})$. The upper bounds for I_1 and I_2 imply that equation (A.7) hold.

Next, we would like to show

$$\Big|\frac{1}{n}\sum_{i=1}^n \frac{\widehat{\sigma}_1^2}{\widetilde{\pi}_i} - \mathbb{E}\Big(\frac{\sigma_1^2}{\pi^*}\Big)\Big| = O_p\Big(\frac{\max(s_1, s_2)\sqrt{\log d \log n}}{n^{1/2}}\Big). \tag{A.8}$$

To this end, consider the following decomposition,

$$\Big|\frac{1}{n}\sum_{i=1}^n \frac{\widehat{\sigma}_1^2}{\widetilde{\pi}_i} - \mathbb{E}\Big(\frac{\sigma_1^2}{\pi^*}\Big)\Big| \le \Big|\frac{1}{n}\sum_{i=1}^n \frac{\sigma_1^2}{\pi_i^*} - \mathbb{E}\Big(\frac{\sigma_1^2}{\pi^*}\Big)\Big| + \Big|\frac{1}{n}\sum_{i=1}^n \frac{\widehat{\sigma}_1^2 - \sigma_1^2}{\widetilde{\pi}_i}\Big| + \sigma_1^2\Big|\frac{1}{n}\sum_{i=1}^n \frac{\pi_i^* - \widetilde{\pi}_i}{\pi_i^*\widetilde{\pi}_i}\Big|.$$

For simplicity, we denote the three terms in the right hand side by J_1, J_2, J_3, respectively. Since σ_1^2 and $1/\pi_i^*$ are bounded, the Hoeffding inequality implies $J_1 = O_p(n^{-1/2})$. In addition, equation (A.7) implies $J_2 = O_p(\max(s_1, s_2)\sqrt{\log d \log n}/n^{1/2})$. Finally, similar to the upper bound argument for I_2, we can verify that $J_3 = O_p(\max(s_1, s_2)\sqrt{\log d/n})$. This proves equation (A.8).

Following the similar argument, we can also verify that

$$\Big|\frac{1}{n}\sum_{i=1}^n (\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X}_i - \widehat{\mu}_1)^2 - \mathbb{E}(\boldsymbol{\alpha}^{*\top}\boldsymbol{X}_i - \mu_1^*)^2\Big| = O_p\Big(\frac{\max(s_1, s_2)\sqrt{\log d \log n}}{n^{1/2}}\Big).$$

Together with equations (A.7) and (A.8), we complete the proof. Since our proof is based on the finite sample concentration inequalities, one can strengthen our asymptotic result to the following

non-asymptotic statement: $|\widehat{V} - V| \le C \max(s_1, s_2)\sqrt{\log d \log n}/n^{1/2}$ with probability at least $1 - d^{-1} - n^{-1}$ for some constant $C > 0$.

A.3 Proof of Corollary 2.1

By the proof of Theorem 2.1, for each $\lambda = a\sqrt{\log d/n}$ and $\lambda' = a'\sqrt{\log d/n}$, where $a, a' \in \{a_1, ..., a_M\}$, we have

$$\mathbb{P}\Big(\Big|\widehat{\mu}_1(\lambda, \lambda') - \mu_1^* - \frac{1}{n}\sum_{i=1}^n S_i\Big| \ge C(a, a') \max(s_1, s_2) \log d\sqrt{\log n}/n\Big) \le d^{-1} + n^{-1},$$

where $S_i = \frac{T_i}{\pi^*}(Y_i(1) - K_1(\boldsymbol{X}_i)) + K_1(\boldsymbol{X}_i) - \mu_1^*$. Since $a_1, ..., a_M$ are constants and M is bounded, the union bound yields

$$\mathbb{P}\Big(\max_{\lambda,\lambda' \in \Lambda}\Big|\widehat{\mu}_1(\lambda, \lambda') - \mu_1^* - \frac{1}{n}\sum_{i=1}^n S_i\Big| \ge C' \max(s_1, s_2) \log d\sqrt{\log n}/n\Big) \le M(d^{-1} + n^{-1}),$$

where $C' = \max\{C(a, a') : a, a' \in \{a_1, ..., a_M\}\}$. Thus,

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\widehat{\mu}_1(\widehat{\lambda}, \widehat{\lambda}') - \mu_1^* - \frac{1}{n}\sum_{i=1}^n S_i\Big| \ge C' \max(s_1, s_2) \log d\sqrt{\log n}/n\Big) \\
&\le \mathbb{P}\Big(\max_{\lambda,\lambda' \in \Lambda}\Big|\widehat{\mu}_1(\lambda, \lambda') - \mu_1^* - \frac{1}{n}\sum_{i=1}^n S_i\Big| \ge C' \max(s_1, s_2) \log d\sqrt{\log n}/n\Big) \le M(d^{-1} + n^{-1}).
\end{aligned}$$

This shows that $\widehat{\mu}_1(\widehat{\lambda}, \widehat{\lambda}') - \mu_1^* = \frac{1}{n}\sum_{i=1}^n S_i + o_p(n^{-1/2})$. Following the similar argument in the proof of Theorem 2.1, we complete the proof.

A.4 Proof of Theorem 3.1

Without loss of generality, we let $a(\phi) = 1$. Similar to the previous appendix, we make the modifications on the estimation of $\boldsymbol{\gamma}$ in step 3. Specifically, let $\widetilde{\boldsymbol{\gamma}} = \operatorname{argmin}_{\boldsymbol{\gamma} \in \Omega} \|\boldsymbol{g}_n(\boldsymbol{\gamma})\|_2^2$, where $\Omega = \{\boldsymbol{\gamma} : \|\boldsymbol{\gamma} - \widehat{\boldsymbol{\beta}}_{\widetilde{S}}\|_1 \le \delta/\log n\}$ for some small constant $\delta > 0$.

Lemma A.3. Under the assumptions in Theorem 3.1,

$$\|\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 = O_p\Big(s_2\sqrt{\frac{\log d}{n}}\Big), \quad \mathbb{P}_n[\boldsymbol{X}^\top(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)]^2 = O_p\Big(\frac{s_2 \log d}{n}\Big).$$

Following the similar argument after Lemma A.1, we can easily see that this lemma can be proved by modifying Lemma E.1 of Ning and Liu (2014). We omit the details.

Lemma A.4. Under the assumptions in Theorem 3.1,

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O_p\Big((s_1 + s_2)\sqrt{\frac{\log d}{n}}\Big),$$

and

$$\mathbb{P}_n[\boldsymbol{X}^\top(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2 = O_p\Big(\frac{s_1\log d + s_2\log d}{n}\Big).$$

Proof of Lemma A.4. Similar to the proof of Lemma A.2, we need the following three steps:

(i) $|\widetilde{S}| \le Cs_2$, where C is a positive constant.

(ii) $\|\widehat{\boldsymbol{\beta}}_{\widetilde{S}^c} - \boldsymbol{\beta}^*_{\widetilde{S}^c}\|_1 = O_p(s_1\sqrt{\log d/n})$, and $\mathbb{P}_n[(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)^\top_{\widetilde{S}^c}\boldsymbol{X}_{\widetilde{S}^c}]^2 = O_p(s_1\log d/n)$.

(iii) $\|\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1 = O_p(\sqrt{((s_2+s_1)\log d)/n})$, and

$$\mathbb{P}_n[(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)^\top\boldsymbol{X}_{\widetilde{S}}]^2 = O_p\Big(\frac{s_1\log d + s_2\log d}{n}\Big).$$

We first show (i). For notational simplicity, we set $a(\phi)=1$. By the KKT condition of the penalized log-likelihood for $\widetilde{\boldsymbol{\alpha}}$, we have $\mathbb{P}_n X_j(Y(1)-b'(\boldsymbol{X}^\top\widetilde{\boldsymbol{\alpha}})) = \lambda'\mathrm{sign}(\widetilde{\alpha}_j)$ for any $j\in\mathrm{supp}(\widetilde{\boldsymbol{\alpha}})$. Then,

$$\begin{aligned}
\lambda'|\widetilde{S}|^{1/2} &= \|\mathbb{P}_n\boldsymbol{X}_{\widetilde{S}}(Y(1)-b'(\boldsymbol{X}^\top\widetilde{\boldsymbol{\alpha}}))\|_2 \le \|\mathbb{P}_n\boldsymbol{X}_{\widetilde{S}}\epsilon_1\|_2 + \|\mathbb{P}_n\boldsymbol{X}_{\widetilde{S}}(b'(\boldsymbol{X}^\top\widetilde{\boldsymbol{\alpha}})-b'(\boldsymbol{X}^\top\boldsymbol{\alpha}^*))\|_2\\
&= \|\mathbb{P}_n\boldsymbol{X}_{\widetilde{S}}\epsilon_1\|_2 + \|\mathbb{P}_n\boldsymbol{X}_{\widetilde{S}}b''(t)\boldsymbol{X}^\top(\widetilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)\|_2\\
&\le |\widetilde{S}|^{1/2}\|\mathbb{P}_n\boldsymbol{X}_{\widetilde{S}}\epsilon_1\|_\infty + C\{\mathbb{P}_n[\boldsymbol{X}^\top(\widetilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)]^2\}^{1/2}\|\mathbb{P}_n\boldsymbol{X}_{\widetilde{S}}\boldsymbol{X}_{\widetilde{S}}^\top\|_2^{1/2}\\
&\le C|\widetilde{S}|^{1/2}\sqrt{\log|\widetilde{S}|/n} + C\|\mathbb{P}_n\boldsymbol{X}_{\widetilde{S}}\boldsymbol{X}_{\widetilde{S}}^\top\|_2^{1/2}\sqrt{s_2\log d/n}.
\end{aligned}$$

In the second step, we first use the mean value theorem to get that for some t between $\boldsymbol{X}^\top\widetilde{\boldsymbol{\alpha}}$ and $\boldsymbol{X}^\top\boldsymbol{\alpha}^*$, $b'(\boldsymbol{X}^\top\widetilde{\boldsymbol{\alpha}})-b'(\boldsymbol{X}^\top\boldsymbol{\alpha}^*) = b''(t)\boldsymbol{X}^\top(\widetilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)$. In the third step, we have $|b''(t)|\le C$ for some constant $C>0$ since $\max_{1\le i\le n}|\boldsymbol{X}_i^\top(\widetilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)| = o_p(1)$ and then we apply the Cauchy inequality. Similarly, in the last step we use the Bernstein inequality for sub-exponential random variables, the union bound argument as well as Lemma A.3. The remaining proof of (i) is identical to the proof of Lemma A.2. We omit the details. The proof of (ii) is also identical to the proof of Lemma A.2. In the following, we focus on (iii).

In the following, we aim to show (iii). We let $\pi = \pi(\boldsymbol{\gamma}^{*\top}\bar{\boldsymbol{X}}_{\widetilde{S}} + \widehat{\boldsymbol{\beta}}^\top_{\widetilde{S}^c}\boldsymbol{X}_{\widetilde{S}^c})$ and $\widetilde{\pi} = \pi(\widetilde{\boldsymbol{\gamma}}^\top\bar{\boldsymbol{X}}_{\widetilde{S}} + \widehat{\boldsymbol{\beta}}^\top_{\widetilde{S}^c}\boldsymbol{X}_{\widetilde{S}^c})$. We have that

$$\Big\|\mathbb{P}_n\Big(\frac{T}{\widetilde{\pi}}-\frac{T}{\pi}\Big)\mathbf{f}(\boldsymbol{X})\Big\|_2^2 \le -2\mathbb{P}_n\Big(\frac{T}{\pi}-1\Big)\mathbf{f}(\boldsymbol{X})\mathbb{P}_n\Big(\frac{T}{\widetilde{\pi}}-\frac{T}{\pi}\Big)\mathbf{f}(\boldsymbol{X}). \tag{A.9}$$

Let π' denote the derivative of π evaluated at an intermediate value between $\boldsymbol{\gamma}^{*\top}\boldsymbol{X}_{\widetilde{S}} + \widehat{\boldsymbol{\beta}}^\top_{\widetilde{S}^c}\boldsymbol{X}_{\widetilde{S}^c}$ and $\widetilde{\boldsymbol{\gamma}}^\top\boldsymbol{X}_{\widetilde{S}} + \widehat{\boldsymbol{\beta}}^\top_{\widetilde{S}^c}\boldsymbol{X}_{\widetilde{S}^c}$. Then

$$\begin{aligned}
\Big\|\mathbb{P}_n\Big(\frac{T}{\widetilde{\pi}}-\frac{T}{\pi}\Big)\mathbf{f}(\boldsymbol{X})\Big\|_2^2 &= \Big\|\mathbb{P}_n\frac{T\pi'}{\widetilde{\pi}\pi}\mathbf{f}(\boldsymbol{X})\bar{\boldsymbol{X}}^\top_{\widetilde{S}}(\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\Big\|_2^2\\
&\ge \|\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2^2\sigma_{\min}\Big(\mathbb{P}_n\frac{T\pi'}{\widetilde{\pi}\pi}\mathbf{f}(\boldsymbol{X})\bar{\boldsymbol{X}}^\top_{\widetilde{S}}\Big) \ge C\|\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2^2\sigma_{\min}\Big(\mathbb{P}_n T\mathbf{f}(\boldsymbol{X})\bar{\boldsymbol{X}}^\top_{\widetilde{S}}\Big),
\end{aligned}$$

for some constant $C>0$, where $\sigma_{\min}$ is the minimum singular value. For notational simplicity, denote $\mathbf{f}^*(\boldsymbol{X}) = (b'(\boldsymbol{\alpha}^{*\top}\boldsymbol{X}), b''(\boldsymbol{\alpha}^{*\top}\boldsymbol{X})\boldsymbol{X}^\top_{\widetilde{S}})^\top$. By the Weyl's inequality for singular values and (i),

we have

$$|\sigma_{\min}(\mathbb{P}_n T\mathbf{f}(\boldsymbol{X})\bar{\boldsymbol{X}}_{\widetilde{S}}^\top) - \sigma_{\min}(\mathbb{P}T\mathbf{f}^*(\boldsymbol{X})\bar{\boldsymbol{X}}_{\widetilde{S}}^\top)| \le \max_{|S|\le Cs_2} \|\mathbb{P}_n T\mathbf{f}(\boldsymbol{X})\bar{\boldsymbol{X}}_S^\top - \mathbb{P}T\mathbf{f}^*(\boldsymbol{X})\bar{\boldsymbol{X}}_S^\top\|_2 \le J_1 + J_2,$$

where

$$J_1 = \max_{|S|\le Cs_2} \|\mathbb{P}_n T(\mathbf{f}(\boldsymbol{X}) - \mathbf{f}^*(\boldsymbol{X}))\bar{\boldsymbol{X}}_S^\top\|_2, \quad J_2 = \max_{|S|\le Cs_2} \|(\mathbb{P}_n - \mathbb{P})T\mathbf{f}^*(\boldsymbol{X})\bar{\boldsymbol{X}}_S^\top\|_2.$$

For J_2, we have $J_2 \le Cs_2 \max_{|S|\le Cs_2} \|(\mathbb{P}_n - \mathbb{P})T\mathbf{f}^*(\boldsymbol{X})\bar{\boldsymbol{X}}_S^\top\|_{\max}$. It is easily seen that each element of $\mathbf{f}^*(\boldsymbol{X})\bar{\boldsymbol{X}}_S^\top$ is bounded by a constant. Applying the Hoeffding inequality and the union bound, we have $J_2 = O_p(s_2\sqrt{\log s_2/n})$. In the following, we focus on J_1. Let $\widetilde{\mathbf{A}} = (b''(t_1)\mathbf{1}_s, b'''(t_2)\mathbf{1}_s^{\otimes 2})$, where t_1 and t_2 are some intermediate values between $\boldsymbol{\alpha}^{*\top}\boldsymbol{X}$ and $\widetilde{\boldsymbol{\alpha}}^\top\boldsymbol{X}$. Applying the mean value theorem, it is easily seen that

$$\begin{aligned} J_1 &= \max_{|S|\le Cs_2} \sup_{\|\boldsymbol{u}\|_2=1, \|\boldsymbol{v}\|_2=1} \left|\mathbb{P}_n(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)^\top \boldsymbol{X}T\boldsymbol{u}^\top \widetilde{\mathbf{A}} \circ \boldsymbol{X}_S^{\otimes 2}\boldsymbol{v}\right| \\ &\le \left|\mathbb{P}_n[(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)^\top \boldsymbol{X}]^2\right|^{1/2} \max_{|S|\le Cs_2} \sup_{\|\boldsymbol{u}\|_2=1} \left|\mathbb{P}_n T\boldsymbol{u}^\top[\widetilde{\mathbf{A}} \circ \boldsymbol{X}_S^{\otimes 2}]^{\otimes 2}\boldsymbol{u}\right|^{1/2}. \end{aligned} \tag{A.10}$$

Recall that we have $|b''(t_1)| \le C$ and and $|b'''(t_2)| \le C$ for some constant C. Let $\boldsymbol{u} = (u_1, \boldsymbol{u}_2)$. After some algebra similar to equation (A.12), we can obtain

$$\boldsymbol{u}^\top[\widetilde{\mathbf{A}} \circ \boldsymbol{X}_S^{\otimes 2}]^{\otimes 2}\boldsymbol{u} \le C[\|\boldsymbol{X}_S\|_2^2 + \|\boldsymbol{X}_S\|_2^2(\boldsymbol{u}_2^\top \boldsymbol{X}_S)^2] \le C[s_2 + s_2(\boldsymbol{u}_2^\top \boldsymbol{X}_S)^2],$$

where the last step follows from $\|\boldsymbol{X}\|_\infty \le C$ for some constant C. If we plug above inequality into equation (A.10), we obtain

$$\max_{|S|\le Cs_2} \sup_{\|\boldsymbol{u}\|_2=1} \left|\mathbb{P}_n T\boldsymbol{u}^\top[\widetilde{\mathbf{A}} \circ \boldsymbol{X}_S^{\otimes 2}]^{\otimes 2}\boldsymbol{u}\right|^{1/2} \le \left[Cs_2 + Cs_2 \max_{|S|\le Cs_2} \lambda_{\max}(\mathbb{P}_n \boldsymbol{X}_S^{\otimes 2})\right]^{1/2}.$$

By the proof of Lemma A.2, $\max_{|S|\le Cs_2} \lambda_{\max}(\mathbb{P}_n \boldsymbol{X}_S^{\otimes 2}) = O_p(1)$. Combining with the results $\mathbb{P}_n[(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)^\top \boldsymbol{X}]^2 = O_p(s_2 \log d/n)$, by equation (A.10), we have that $J_1 = O_p(s_2\sqrt{\log d/n})$. Thus, $J_1 + J_2 = o_p(1)$ and by our assumption, $\sigma_{\min}(\mathbb{P}\mathbf{f}^*(\boldsymbol{X})\bar{\boldsymbol{X}}_{\widetilde{S}}^\top)$ is bounded away from 0 by a constant. This implies $\sigma_{\min}(\mathbb{P}_n T\mathbf{f}(\boldsymbol{X})\bar{\boldsymbol{X}}_{\widetilde{S}}^\top)$ is lower bounded by a positive constant, and therefore

$$\left\|\mathbb{P}_n\Big(\frac{T}{\widetilde{\pi}} - \frac{T}{\pi}\Big)\mathbf{f}(\boldsymbol{X})\right\|_2^2 \ge C\|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2^2. \tag{A.11}$$

Following the similar argument, we can show that the right hand side of equation (A.9) is bounded above by $2\|\boldsymbol{\Delta}_n\|_2\|A_n\|_2^{1/2}|K_n|^{1/2}$, where

$$\boldsymbol{\Delta}_n = \mathbb{P}_n\Big(\frac{T}{\pi} - 1\Big)\mathbf{f}(\boldsymbol{X}), \quad A_n = \mathbb{P}_n\Big(\frac{T\pi'}{\widetilde{\pi}\pi}\Big)^2\mathbf{f}(\boldsymbol{X})^{\otimes 2}, \quad K_n = \mathbb{P}_n[\bar{\boldsymbol{X}}_{\widetilde{S}}^\top(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)]^2$$

We can show that $\|A_n\|_2 \le C\lambda_{\max}(\mathbb{P}_n \mathbf{f}(\boldsymbol{X})^{\otimes 2})$ for some constants $C > 0$. For simplicity, let $\widetilde{b}' = b'(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X})$, $\widetilde{b}'' = b''(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X})$. Note that

$$\begin{aligned}\lambda_{\max}(\mathbb{P}_n \mathbf{f}(\boldsymbol{X})^{\otimes 2}) &= \sup_{\|\boldsymbol{v}\|_2=1} \mathbb{P}_n[\boldsymbol{v}^\top \mathbf{f}(\boldsymbol{X})]^2 \le 2 \sup_{\|\boldsymbol{v}\|_2=1} \mathbb{P}_n[(\widetilde{b}' v_1)^2 + (\widetilde{b}'' \boldsymbol{v}_2^\top \boldsymbol{X}_{\widetilde{S}})^2] \\ &\le C + C \sup_{\|\boldsymbol{v}\|_2=1} \mathbb{P}_n(\boldsymbol{v}_2^\top \boldsymbol{X}_{\widetilde{S}})^2 \le C + C\lambda_{\max}(\mathbb{P}_n \boldsymbol{X}_{\widetilde{S}}^{\otimes 2}) \le C'. \end{aligned} \tag{A.12}$$

Here, the second inequality follows from $\max_{1\le i\le n} \widetilde{b}_i' \le C + C\|\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 \max_{1\le i\le n} |\boldsymbol{X}_i| \le C'$ for some constants $C, C' > 0$ and the same bound applies to $\max_{1\le i\le n} \widetilde{b}_i''$ as well; the last step follows from the similar argument in the proof of Lemma A.2 by applying the Hoeffding inequality. We omit the details here.

Now, we consider $\boldsymbol{\Delta}$. We decompose $\boldsymbol{\Delta}_n = I_n + II_n$, where

$$I_n = \mathbb{P}_n\Big(\frac{T}{\pi^*} - 1\Big)\mathbf{f}(\boldsymbol{X}), \quad II_n = -\mathbb{P}_n \frac{T\pi'}{\pi\pi^*} \mathbf{f}(\boldsymbol{X}) \boldsymbol{X}_{\widetilde{S}^c}^\top (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\widetilde{S}^c},$$

where similarly π' is the derivative of π evaluated at some intermediate value. Here, we further decompose I_n as $I_n = I_{n1} + I_{n2}$, where

$$I_{n1} = \mathbb{P}_n\Big(\frac{T}{\pi^*} - 1\Big)\mathbf{f}^*(\boldsymbol{X}), \quad I_{n2} = \mathbb{P}_n\Big(\frac{T}{\pi^*} - 1\Big)(\mathbf{f}(\boldsymbol{X}) - \mathbf{f}^*(\boldsymbol{X})),$$

and $\mathbf{f}^*(\boldsymbol{X}) = (b'(\boldsymbol{\alpha}^{*\top}\boldsymbol{X}), b''(\boldsymbol{\alpha}^{*\top}\boldsymbol{X})\boldsymbol{X}_{\widetilde{S}}^\top)^\top$. Thus, by the Bernstein inequality and the union bound argument,

$$\|I_{n1}\|_2 \le |\widetilde{S}|^{1/2} \Big\| \mathbb{P}_n\Big(\frac{T}{\pi^*} - 1\Big)\mathbf{f}^*(\boldsymbol{X}) \Big\|_\infty = O_p(\sqrt{s_2 \log s_2 / n}).$$

For I_{n2}, applying the mean value theorem, for some intermediate value t, we have

$$\begin{aligned}\|I_{n2}\|_2 &= \sup_{\|\boldsymbol{v}\|_2=1} \Big| \mathbb{P}_n\Big(\frac{T}{\pi^*} - 1\Big) C(t)(\boldsymbol{v}^\top \bar{\boldsymbol{X}}_{\widetilde{S}}) \boldsymbol{X}^\top (\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) \Big| \\ &\le \sup_{\|\boldsymbol{v}\|_2=1} \Big| \mathbb{P}_n\Big(\frac{T}{\pi^*} - 1\Big)^2 C^2(t)(\boldsymbol{v}^\top \bar{\boldsymbol{X}}_{\widetilde{S}})^2 \Big|^{1/2} |\mathbb{P}_n[\boldsymbol{X}^\top (\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)]^2|^{1/2} \\ &\le C\lambda_{\min}^{1/2}(\mathbb{P}_n \bar{\boldsymbol{X}}_{\widetilde{S}}^{\otimes 2}) \cdot \sqrt{s_2 \log d / n} = O_p(\sqrt{s_2 \log d / n}),\end{aligned}$$

where $C(t) = \max(b''(t), b'''(t)) \le C$. In addition,

$$\begin{aligned}\|II_n\|_2 &\le \sup_{\|\boldsymbol{v}\|_2=1} \Big| \mathbb{P}_n \frac{T\pi'}{\pi\pi^*} \boldsymbol{v}^\top \mathbf{f}(\boldsymbol{X}) \boldsymbol{X}_{\widetilde{S}^c}^\top (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\widetilde{S}^c} \Big| \\ &\le \sup_{\|\boldsymbol{v}\|_2=1} \Big| \mathbb{P}_n \frac{T\pi'}{\pi\pi^*} [\boldsymbol{v}^\top \mathbf{f}(\boldsymbol{X})]^2 \Big|^{1/2} \Big| \mathbb{P}_n \frac{T\pi'}{\pi\pi^*} [\boldsymbol{X}_{\widetilde{S}^c}^\top (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\widetilde{S}^c}]^2 \Big|^{1/2} \\ &\le C\lambda_{\max}^{1/2}(\mathbb{P}_n \mathbf{f}(\boldsymbol{X})^{\otimes 2}) \cdot \Big| \mathbb{P}_n [\boldsymbol{X}_{\widetilde{S}^c}^\top (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\widetilde{S}^c}]^2 \Big|^{1/2} = O_p(\sqrt{s_1 \log d / n}).\end{aligned}$$

This implies $\|\boldsymbol{\Delta}\|_2 = O_p(\sqrt{(s_2 \log d + s_1 \log d)/n})$. For K_n, by the similar argument in equa-

tion (A.12), we have

$$K_n \le \lambda_{\max}(\mathbb{P}_n \bar{\boldsymbol{X}}_{\widetilde{S}}^{\otimes 2}) \|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2^2 \le C\|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2^2,$$

for some constant $C > 0$. Combining with equations (A.11) and (A.9), we obtain that

$$\|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 = O_p(\sqrt{((s_2 + s_1)\log d)/n}),$$

and

$$\|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_1 \le C s_2^{1/2} \|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 = O_p(\sqrt{(s_2^2 \log d + s_2 s_1 \log d)/n}).$$

This completes the proof of (iii). Finally, we combine the results in (i), (ii) and (iii) to show the desired result:

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_1 + \|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\widetilde{S}^c}\|_1 = O_p\Big((s_1 + s_2)\sqrt{\frac{\log d}{n}}\Big),$$

and

$$\begin{aligned}\mathbb{P}_n[\bar{\boldsymbol{X}}^\top(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)]^2 &\le 2\mathbb{P}_n[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\widetilde{S}^c}^\top \boldsymbol{X}_{\widetilde{S}^c}]^2 + 2\mathbb{P}_n[(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)^\top \bar{\boldsymbol{X}}_{\widetilde{S}}]^2 \\ &= O_p\Big(\frac{s_1 \log d + s_2 \log d}{n}\Big).\end{aligned}$$

This completes the proof. □

Proof of Theorem 3.1. Same as the proof of Theorem 2.1, we start with the decomposition:

$$\widehat{\mu}_1 - \mu_1^* = \mathbb{P}_n\Big[\frac{T}{\pi^*}(Y(1) - K_1(\boldsymbol{X})) + K_1(\boldsymbol{X}) - \mu_1^*\Big] + I_1 + I_2,$$

where

$$I_1 = \mathbb{P}_n\Big[\frac{T}{\widetilde{\pi}} - \frac{T}{\pi^*}\Big](Y(1) - K_1(\boldsymbol{X})), \quad I_2 = \mathbb{P}_n\Big[\frac{T}{\widetilde{\pi}} - 1\Big]K_1(\boldsymbol{X}).$$

We first consider I_2. Recall that $K_1(\boldsymbol{X}) = b'(\boldsymbol{\alpha}^{*T}\boldsymbol{X})$. Note that

$$\begin{aligned}I_2 &= \mathbb{P}_n\Big[\frac{T}{\widetilde{\pi}} - 1\Big](b'(\boldsymbol{\alpha}^{*T}\boldsymbol{X}) - b'(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X}) - b''(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X})(\boldsymbol{\alpha}^* - \widetilde{\boldsymbol{\alpha}})_{\widetilde{S}}^\top \boldsymbol{X}_{\widetilde{S}}) \\ &= \mathbb{P}_n\Big[\frac{T}{\widetilde{\pi}} - 1\Big](b''(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X})(\boldsymbol{\alpha}^* - \widetilde{\boldsymbol{\alpha}})_{\widetilde{S}^c}^\top \boldsymbol{X}_{\widetilde{S}^c} + b'''(t)[(\boldsymbol{\alpha}^* - \widetilde{\boldsymbol{\alpha}})^\top \boldsymbol{X}]^2),\end{aligned}$$

where t is an intermediate value between $\boldsymbol{\alpha}^{*\top}\boldsymbol{X}$ and $\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X}$. We denote

$$I_{21} = \mathbb{P}_n\Big[\frac{T}{\widetilde{\pi}} - 1\Big]b''(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X})(\boldsymbol{\alpha}^* - \widetilde{\boldsymbol{\alpha}})_{\widetilde{S}^c}^\top \boldsymbol{X}_{\widetilde{S}^c}, \quad I_{22} = \mathbb{P}_n\Big[\frac{T}{\widetilde{\pi}} - 1\Big]b'''(t)[(\boldsymbol{\alpha}^* - \widetilde{\boldsymbol{\alpha}})^\top \boldsymbol{X}]^2.$$

For I_{21}, we further apply the multivariate Taylor theorem to expand $\widetilde{\pi}$ and $b''(\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X})$,

$$I_{21} = \mathbb{P}_n\Big[\frac{T}{\pi^*} - 1\Big] b''(\boldsymbol{\alpha}^{*T}\boldsymbol{X})(\boldsymbol{\alpha}^* - \widetilde{\boldsymbol{\alpha}})_{\widetilde{S}^c}^\top \boldsymbol{X}_{\widetilde{S}^c} - \mathbb{P}_n \frac{T\pi'(t_1)}{\pi^{*2}} b''(t_2)(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \bar{\boldsymbol{X}}(\boldsymbol{\alpha}^* - \widetilde{\boldsymbol{\alpha}})_{\widetilde{S}^c}^\top \boldsymbol{X}_{\widetilde{S}^c}$$
$$+ \mathbb{P}_n\Big[\frac{T}{\pi(t_1)} - 1\Big] b'''(t_2)(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)^\top \boldsymbol{X}(\boldsymbol{\alpha}^* - \widetilde{\boldsymbol{\alpha}})_{\widetilde{S}^c}^\top \boldsymbol{X}_{\widetilde{S}^c}, \tag{A.13}$$

where t_1 and t_2 are the intermediate values between $\boldsymbol{\beta}^{*T}\bar{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{\beta}}^\top \bar{\boldsymbol{X}}$, and between $\boldsymbol{\alpha}^{*T}\boldsymbol{X}$ and $\widetilde{\boldsymbol{\alpha}}^\top \boldsymbol{X}$. For the first term in equation (A.13), we have $|T/\pi^* - 1| \le 1/c_0$, and $b''(\boldsymbol{\alpha}^{*T}\boldsymbol{X})$ and X_j are both bounded random variables, we can apply the Hoeffding inequality and the union bound argument, which gives us

$$\Big\| \mathbb{P}_n\Big[\frac{T}{\pi^*} - 1\Big] b''(\boldsymbol{\alpha}^{*T}\boldsymbol{X}) \boldsymbol{X}_{\widetilde{S}^c} \Big\|_\infty = O_p\Big(\sqrt{\frac{\log d}{n}}\Big).$$

Together with Lemma A.3, we have

$$\Big| \mathbb{P}_n\Big[\frac{T}{\pi^*} - 1\Big] b''(\boldsymbol{\alpha}^{*T}\boldsymbol{X})(\boldsymbol{\alpha}^* - \widetilde{\boldsymbol{\alpha}})_{\widetilde{S}^c}^\top \boldsymbol{X}_{\widetilde{S}^c} \Big| = O_p\Big(\frac{s_2 \log d}{n}\Big).$$

Similar to the derivation in equation (A.5), for the second term in equation (A.13), by Lemma A.4,

$$\Big| \mathbb{P}_n \frac{T\pi'(t_1)}{\pi^{*2}} b''(t_2)(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \bar{\boldsymbol{X}}(\boldsymbol{\alpha}^* - \widetilde{\boldsymbol{\alpha}})_{\widetilde{S}^c}^\top \boldsymbol{X}_{\widetilde{S}^c} \Big| = O_p\Big(\frac{(s_1 + s_2)\log d}{n}\Big).$$

For the last term in equation (A.13), first we note that $b'''(\cdot)$ is continuous by assumption, and thus $|b'''(t_2)|$ is bounded since $|\boldsymbol{\alpha}^{*\top}\boldsymbol{X}_i| \le K$. In addition, $\pi(t_1)$ is also bounded away from 0. We apply the Cauchy inequality,

$$\Big| \mathbb{P}_n\Big[\frac{T}{\pi(t_1)} - 1\Big] b'''(t_2)(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)^\top \boldsymbol{X}(\boldsymbol{\alpha}^* - \widetilde{\boldsymbol{\alpha}})_{\widetilde{S}^c}^\top \boldsymbol{X}_{\widetilde{S}^c} \Big|$$
$$\le \Big[\mathbb{P}_n\Big[\frac{T}{\pi(t_1)} - 1\Big]^2 b'''(t_2)^2 [(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)^\top \boldsymbol{X}]^2 \Big]^{1/2} \Big[\mathbb{P}_n [(\boldsymbol{\alpha}^* - \widetilde{\boldsymbol{\alpha}})_{\widetilde{S}^c}^\top \boldsymbol{X}_{\widetilde{S}^c}]^2 \Big]^{1/2} = O_p\Big(\frac{s_2 \log d}{n}\Big).$$

Combining these results with equation (A.13), we obtain

$$|I_{21}| = O_p\Big(\frac{s_2 \log d}{n} + \frac{(s_1 + s_2)\log d}{n}\Big).$$

The same argument above can be used to control the magnitude of I_{22}, which yields

$$|I_{22}| \le C \mathbb{P}_n [(\boldsymbol{\alpha}^* - \widetilde{\boldsymbol{\alpha}})^\top \boldsymbol{X}]^2 = O_p\Big(\frac{s_2 \log d}{n}\Big),$$

for some constant $C > 0$. This together implies the rate of convergence of I_2

$$|I_2| = O_p\Big(\frac{s_2 \log d}{n} + \frac{(s_1 + s_2)\log d}{n}\Big).$$

For I_1, recall that $\epsilon_1 = Y(1) - K_1(\boldsymbol{X})$. Again, by the Taylor theorem,

$$|I_1| \leq \Big\|\mathbb{P}_n \frac{T\pi'(t)}{\pi^{*2}} \epsilon \bar{\boldsymbol{X}}\Big\|_\infty \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \Big(\frac{1}{c_0^2} + \frac{2}{c_0^3}\Big)\Big|\mathbb{P}_n[(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \bar{\boldsymbol{X}}]^2\Big| \max_{1\leq i\leq n} |\epsilon_{1i}|,$$

where t is an intermediate value. For the first terms, we still have $\frac{T\pi'(t)}{\pi^{*2}} \leq 1/c_0^2$, but now ϵ_1 is a sub-exponential random variable and X_j is uniformly bounded. This implies that $|\frac{T\pi'(t)}{\pi^{*2}}\epsilon_1 X_j|$ is a sub-exponential random variable with $\|\frac{T\pi'(t)}{\pi^{*2}}\epsilon_1 X_j\|_{\psi_1} \leq C_X C_\epsilon / c_0^2$. By the Bernstein inequality and the union bound argument, we have

$$\Big\|\mathbb{P}_n \frac{T\pi'(t)}{\pi^{*2}} \epsilon \bar{\boldsymbol{X}}\Big\|_\infty = O_p\Big(\sqrt{\frac{\log d}{n}}\Big).$$

For the second term, since $\mathbb{P}(|\epsilon_{1i}| > t) \leq C_1 \exp(-C_2 t)$ for some constants $C_1, C_2 > 0$, this implies $\mathbb{P}(\max_{1\leq i\leq n} |\epsilon_{1i}| > t) \leq nC_1 \exp(-C_2 t)$. With $t = (2/C_2)\log n$,

$$\mathbb{P}(\max_{1\leq i\leq n} |\epsilon_{1i}| > C \log n) \leq C_1/n = o(1).$$

Together with Lemma A.4, we have shown that

$$|I_1| = O_p\Big(\frac{(s_1+s_2)\log d}{n} + \frac{(s_1+s_2)\log d \log n}{n}\Big) = o_p(n^{-1/2}).$$

Thus, we have

$$\widehat{\mu}_1 - \mu_1^* = \mathbb{P}_n\Big[\frac{T}{\pi^*}(Y(1) - K_1(\boldsymbol{X})) + K_1(\boldsymbol{X}) - \mu_1^*\Big] + o_p(n^{-1/2}).$$

Note that $\frac{T}{\pi^*}(Y(1) - K_1(\boldsymbol{X})) + K_1(\boldsymbol{X}) - \mu_1^*$ corresponds to the efficient score function for μ_1^* (Hahn, 1998). This implies the semiparametric efficiency of $\widehat{\mu}_1$. □

A.5 Proof of Proposition 2.1

The proof is similar to that of Theorem 2.1. So, we only provide the key step. First notice that

$$\begin{aligned}\widehat{\tau}_1 - \tau_1^* &= \frac{\sum_{i=1}^n T_i \epsilon_{1i}}{\sum_{i=1}^n T_i} + \frac{\sum_{i=1}^n T_i (K_1(\boldsymbol{X}_i) - \tau_1^*)}{\sum_{i=1}^n T_i} \\ &= \frac{\sum_{i=1}^n T_i \epsilon_{1i}}{np} + \frac{\sum_{i=1}^n T_i (K_1(\boldsymbol{X}_i) - \tau_1^*)}{np} + o_p(n^{-1/2}). \end{aligned} \tag{A.14}$$

Then, we focus on $\widehat{\tau}_0$:

$$\widehat{\tau}_0 - \tau_0^* = \frac{\sum_{i=1}^n (1-T_i)\widetilde{r}_i \epsilon_{0i}}{\sum_{i=1}^n (1-T_i)\widetilde{r}_i} + \frac{\sum_{i=1}^n (1-T_i)\widetilde{r}_i (K_0(\boldsymbol{X}_i) - \tau_0^*)}{\sum_{i=1}^n (1-T_i)\widetilde{r}_i}.$$

Applying the Taylor theorem, similar to equation (A.6), we can easily verify that

$$\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^{n}(1-T_i)\widetilde{r}_i\epsilon_{0i} - \frac{1}{n}\sum_{i=1}^{n}(1-T_i)r_i^*\epsilon_{0i}\Big| \\
&\le \Big|\frac{1}{n}\sum_{i=1}^{n}(1-T_i)r_i'\epsilon_{0i}\boldsymbol{X}_i^\top(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\Big| + \frac{1}{2}\Big|\frac{1}{n}\sum_{i=1}^{n}(1-T_i)r_i''(t)\epsilon_{0i}[\boldsymbol{X}_i^\top(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2\Big| \lesssim n^{-1/2},
\end{aligned}$$

where $r_i''(t)$ is the second order derivative of $r_i(t) = \pi_i(t)/(1-\pi_i(t))$ with respect to t evaluated at some intermediate value. The proof of the last step relies on a similar version of Lemma A.2, which says

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O_p\Big(\sqrt{\frac{(s_1+s_2)^2\log d}{n}}\Big),$$

and

$$\mathbb{P}_n[\boldsymbol{X}^\top(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2 = O_p\Big(\frac{(s_1+s_2)\log d}{n}\Big).$$

Then we decompose

$$\frac{1}{n}\sum_{i=1}^{n}(1-T_i)\widetilde{r}_i(K_0(\boldsymbol{X}_i)-\tau_0^*) = \frac{1}{n}\sum_{i=1}^{n}T_i(K_0(\boldsymbol{X}_i)-\tau_0^*) + \frac{1}{n}\sum_{i=1}^{n}[(1-T_i)\widetilde{r}_i - T_i](K_0(\boldsymbol{X}_i)-\tau_0^*),$$

where the last term is denoted by I_1. By the covariate balancing equation (2.13), we have

$$\begin{aligned}
I_1 &= \frac{1}{n}\sum_{i=1}^{n}[(1-T_i)\widetilde{r}_i - T_i]K_0(\boldsymbol{X}_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}[(1-T_i)\widetilde{r}_i - T_i]\widehat{K}_0(\boldsymbol{X}_i) + \frac{1}{n}\sum_{i=1}^{n}[(1-T_i)\widetilde{r}_i - T_i][K_0(\boldsymbol{X}_i) - \widehat{K}_0(\boldsymbol{X}_i)] \\
&= \frac{1}{n}\sum_{i=1}^{n}[(1-T_i)\widetilde{r}_i - T_i]\boldsymbol{X}_i^\top(\widetilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*).
\end{aligned}$$

Similar to the proof of equation (A.4), we can show that the last term is of order $o(n^{-1/2})$. Hence,

$$\begin{aligned}
\widehat{\tau}_0 - \tau_0^* &= \frac{\sum_{i=1}^n(1-T_i)r_i^*\epsilon_{0i}}{\sum_{i=1}^n(1-T_i)\widetilde{r}_i} + \frac{\sum_{i=1}^n T_i(K_0(\boldsymbol{X}_i)-\tau_0^*)}{\sum_{i=1}^n(1-T_i)\widetilde{r}_i} + o_p(n^{-1/2}) \\
&= \frac{\sum_{i=1}^n(1-T_i)r_i^*\epsilon_{0i}}{np} + \frac{\sum_{i=1}^n T_i(K_0(\boldsymbol{X}_i)-\tau_0^*)}{np} + o_p(n^{-1/2}). \qquad \text{(A.15)}
\end{aligned}$$

Combining equations (A.14) and (A.15), we obtain

$$\widehat{\tau} - \tau^* = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p}\left[T_i\epsilon_{1i} - (1-T_i)r_i^*\epsilon_{0i} + T_i(\Delta K(\boldsymbol{X}_i) - \tau^*)\right] + o_p(n^{-1/2}).$$

Then, we can apply the central limit theorem and the asymptotic variance follows by the standard moment calculation. We can show that it agrees with the semiparametric asymptotic variance bound in Hahn (1998).

A.6 Proof of Proposition 2.2

The proof is similar to that of Theorem 2.1. So, we only provide a sketch. By the rearrangement of terms, we have

$$\widehat{\mu}_1 - \mu_1^* = \mathbb{P}_n\Big[\frac{T}{\pi^o}(Y(1) - K_1(\boldsymbol{X})) + K_1(\boldsymbol{X}) - \mu_1^*\Big] + I_1 + I_2,$$

where

$$I_1 = \mathbb{P}_n\Big[\frac{T}{\widetilde{\pi}} - \frac{T}{\pi^o}\Big](Y(1) - K_1(\boldsymbol{X})), \quad I_2 = \mathbb{P}_n\Big[\frac{T}{\widetilde{\pi}} - 1\Big]K_1(\boldsymbol{X}).$$

The key of the proof is to notice the following fact. Since $S \subset \widetilde{S}$, we have $I_2 = \boldsymbol{\alpha}_S^{*\top}\mathbb{P}_n[\frac{T}{\widetilde{\pi}} - 1]\boldsymbol{X}_S = 0$, where the last equality follows from the definition of the covariate balancing estimating equations. By the Taylor theorem,

$$\begin{aligned}
|I_1| &\le \Big|\mathbb{P}_n \frac{T\pi'}{\pi^{*2}}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)^\top \boldsymbol{X}\epsilon_1\Big| + \Big(\frac{1}{c_0^2} + \frac{2}{c_0^3}\Big)\Big|\mathbb{P}_n[(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)^\top \boldsymbol{X}]^2\Big| \max_{1\le i\le n}|\epsilon_{1i}| \\
&\le \Big\|\mathbb{P}_n \frac{T\pi'}{\pi^{*2}}\epsilon_1 \boldsymbol{X}\Big\|_\infty \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \Big(\frac{1}{c_0^2} + \frac{2}{c_0^3}\Big)\Big|\mathbb{P}_n[(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)^\top \boldsymbol{X}]^2\Big| \max_{1\le i\le n}|\epsilon_{1i}|.
\end{aligned}$$

Since

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 = O_p\Big((s_1 + s_2)\sqrt{\frac{\log d}{n}}\Big),$$

and

$$\frac{1}{n}\sum_{i=1}^n [\boldsymbol{X}_i^\top(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)]^2 = O_p\Big(\frac{(s_1 + s_2)\log d}{n}\Big),$$

we can replicate the proof of Theorem 2.1. For simplicity, we omit the details.

B Heuristic Analysis of the AIPW Estimators under Misspecified Propensity Score Models

In this appendix, we provide a heuristic analysis of the AIPW estimator under misspecified propensity score models. Recall that, given estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$, the AIPW is defined as

$$\widehat{\mu}_{AIPW} = \frac{1}{n}\sum_{i=1}^n \frac{T_i}{\widehat{\pi}_i} Y_i - \frac{1}{n}\sum_{i=1}^n \Big(\frac{T_i}{\widehat{\pi}_i} - 1\Big)\widehat{\boldsymbol{\alpha}}^\top \boldsymbol{X}_i,$$

where $\widehat{\pi}_i = \pi(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})$. Our goal is to show that the rate of convergence of $\widehat{\mu}_{AIPW}$ could be slower than root-n under misspecified propensity score models.

We can show that $\mu_{AIPW} = I_1 + I_2$, where

$$I_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\widehat{\pi}_i}Y_i - \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{T_i}{\widehat{\pi}_i} - 1\Big)\boldsymbol{\alpha}^{*\top}\boldsymbol{X}_i, \quad I_2 = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{T_i}{\widehat{\pi}_i} - 1\Big)(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)^\top\boldsymbol{X}_i.$$

After rearrangement of I_1, we have

$$I_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\widehat{\pi}_i}\epsilon_{1i} - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\alpha}^{*\top}\boldsymbol{X}_i = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\pi_i^o}\epsilon_{1i} - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\alpha}^{*\top}\boldsymbol{X}_i + o_p(n^{-1/2}),$$

where the last step is identical to the proof of Proposition 2.2. Same as the proof of Theorem 2.1, we can show that $I_1 = \mu_1^* + O_p(n^{-1/2})$.

The problem of AIPW is due to the I_2 term. For instance, if $\widehat{\alpha}$ is the Lasso estimator, then

$$I_2 \le \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1\Big\{\Big\|\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{T_i}{\pi_i^o} - 1\Big)\boldsymbol{X}_i\Big\|_\infty + o_p(1)\Big\} \lesssim \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_1 = O_p\Big(s_2\sqrt{\frac{\log d}{n}}\Big),$$

where in the first step we use the rate of convergence of $\widehat{\boldsymbol{\beta}}$, the second step follows from $\mathbb{E}\{(T_i/\pi_i^o - 1)X_{ij}\} \neq 0$ but is bounded and the Bernstein inequality, and the last step follows from the rate of convergence of the Lasso estimator. Putting together the order of I_1 and I_2, we have

$$\widehat{\mu}_{AIPW} = \mu_1^* + O_p(s_2\sqrt{\log d/n}).$$

Thus, $\widehat{\mu}_{AIPW}$ has a slower rate than root-n under misspecified propensity score models.

One may argue that the rate of $\widehat{\mu}_{AIPW}$ can be improved given the screening property. In the following, we impose a even stronger assumption. That is the Lasso estimator of $\boldsymbol{\alpha}$ attains model selection consistency, i.e., $\widetilde{S} = S$. We pick $\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{\alpha}}_S, 0)$ as the oracle estimator, where $\widehat{\boldsymbol{\alpha}}_S$ is the least squared estimator on the support S. Then $\widehat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}^* = (\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{iS}^{\otimes 2})^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{iS}\epsilon_{1i}$. As a result,

$$\begin{aligned} I_2 &= \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{T_i}{\pi_i^o} - 1\Big)(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)^\top\boldsymbol{X}_i + o_p(n^{-1/2}) \\ &= \Big\{\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{T_i}{\pi_i^o} - 1\Big)\boldsymbol{X}_{iS}\Big\}^\top\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{iS}^{\otimes 2}\Big)^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{iS}\epsilon_{1i} + o_p(n^{-1/2}). \end{aligned}$$

Denote the term in the second line by A. It is easy to see that $\mathbb{E}(A) = 0$, and

$$\mathbb{E}(A^2) = n^{-1}\mathbb{E}\Big[\Big\{\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\pi_i^*}{\pi_i^o} - 1\Big)\boldsymbol{X}_{iS}\Big\}^\top\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{iS}^{\otimes 2}\Big)^{-1}\Big\{\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\pi_i^*}{\pi_i^o} - 1\Big)\boldsymbol{X}_{iS}\Big\}\Big].$$

Under standard conditions on the design matrix, $\mathbb{E}(A^2) \asymp n^{-1}\mathbb{E}\{\|\frac{1}{n}\sum_{i=1}^{n}(\frac{\pi_i^*}{\pi_i^o} - 1)\boldsymbol{X}_{iS}\|_2^2\} \lesssim s_2/n$.

Thus, the Markov inequality implies $I_2 = O_p(\sqrt{s_2/n})$, and

$$\widehat{\mu}_{AIPW} = \mu_1^* + O_p(\sqrt{s_2/n}),$$

even if we assume $\widetilde{S} = S$. It seems unclear how to remove $\sqrt{s_2}$ in the second term. Unless s_2 is a constant, again $\widehat{\mu}_{AIPW}$ has a slower rate than root-n.

C The Algorithm and Its Theoretical Properties for Estimating the ATE

For clarification purpose, we first present the complete algorithm for the inference on ATE.

Step 1: Estimate the propensity score model by the penalized logistic regression

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\Big\{T_i(\boldsymbol{\beta}^\top \boldsymbol{X}_i) - \log(1+\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_i))\Big\} + \lambda\|\boldsymbol{\beta}\|_1,$$

where $\lambda > 0$ is a tuning parameter.

Step 2: Fit the penalized least squared estimation in the treatment and control groups, respectively,

$$\widetilde{\boldsymbol{\alpha}}_1 = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} T_i\{Y_i - \boldsymbol{\alpha}^\top \boldsymbol{X}_i\}^2 + \lambda_1'\|\boldsymbol{\alpha}\|_1,$$

$$\widetilde{\boldsymbol{\alpha}}_0 = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(1-T_i)\{Y_i - \boldsymbol{\alpha}^\top \boldsymbol{X}_i\}^2 + \lambda_2'\|\boldsymbol{\alpha}\|_1,$$

where $\lambda_1', \lambda_2' > 0$ are tuning parameters.

Step 3: Let $\widetilde{S}_1 = \{j : |\widetilde{\alpha}_{1j}| > 0\}$ and $\widetilde{S}_0 = \{j : |\widetilde{\alpha}_{0j}| > 0\}$ denote the support sets of $\widetilde{\boldsymbol{\alpha}}_1$ and $\widetilde{\boldsymbol{\alpha}}_0$. Solve

$$\widetilde{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^{|\widetilde{S}_1|}} \|\boldsymbol{g}_n(\boldsymbol{\gamma})\|_2^2, \quad \text{where} \quad \boldsymbol{g}_n(\boldsymbol{\gamma}) = n^{-1}\sum_{i=1}^{n}\Big(\frac{T_i}{\pi(\boldsymbol{\gamma}^\top \boldsymbol{X}_{i\widetilde{S}_1} + \widehat{\boldsymbol{\beta}}_{\widetilde{S}_1^c}^\top \boldsymbol{X}_{i\widetilde{S}_1^c})} - 1\Big)\boldsymbol{X}_{i\widetilde{S}_1}.$$

Let us denote $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}_{\widetilde{S}_1^c})$. Thus, we re-estimate the propensity score by $\widetilde{\pi}_i = \pi(\widetilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_i)$. In addition, solve

$$\bar{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^{|\widetilde{S}_0|}} \|\boldsymbol{g}_n(\boldsymbol{\gamma})\|_2^2, \quad \text{where} \quad \boldsymbol{g}_n(\boldsymbol{\gamma}) = n^{-1}\sum_{i=1}^{n}\Big(\frac{1-T_i}{1-\pi(\boldsymbol{\gamma}^\top \boldsymbol{X}_{i\widetilde{S}_0} + \widehat{\boldsymbol{\beta}}_{\widetilde{S}_0^c}^\top \boldsymbol{X}_{i\widetilde{S}_0^c})} - 1\Big)\boldsymbol{X}_{i\widetilde{S}_0}.$$

Let us denote $\bar{\boldsymbol{\beta}} = (\bar{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}_{\widetilde{S}_0^c})$. Thus, we re-estimate the propensity score by $\bar{\pi}_i = \pi(\bar{\boldsymbol{\beta}}^\top \boldsymbol{X}_i)$.

Step 4: Estimate ATE by the Horvitz-Thompson estimator

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{T_i Y_i}{\widetilde{\pi}_i} - \frac{(1-T_i)Y_i}{1-\bar{\pi}_i}\Big).$$

Assumption C.1. Assume that $\epsilon_1 = Y(1) - \boldsymbol{\alpha}_1^{*\top}\boldsymbol{X}$, $\epsilon_0 = Y(0) - \boldsymbol{\alpha}_0^{*\top}\boldsymbol{X}$ and X_j satisfy $\|\epsilon_1\|_{\psi_2} \le C_\epsilon$, $\|\epsilon_0\|_{\psi_2} \le C_\epsilon$ and $\|X_j\|_{\psi_2} \le C_X$ for any $1 \le j \le d$, where C_X and C_ϵ are two positive constants.

Assumption C.2. There exists a constant $0 < c_0 < 1/2$ such that $c_0 \le \pi_i^* \le 1 - c_0$ for any $1 \le i \le n$.

Assumption C.3. Denote $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{X}^{\otimes 2})$. There exists a constant $C > 0$ such that $C \le \lambda_{\min}(\boldsymbol{\Sigma}_{SS}) \le \lambda_{\max}(\boldsymbol{\Sigma}_{SS}) \le 1/C$ for any $S \subset \{1, ..., d\}$ with $|S| \asymp s_2$, where $s_2 = \|\boldsymbol{\alpha}^*\|_0$.

Recall that $K_1(\boldsymbol{X}_i) = \mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i) = \boldsymbol{\alpha}_1^{*\top}\boldsymbol{X}$, $K_0(\boldsymbol{X}_i) = \mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i) = \boldsymbol{\alpha}_0^{*\top}\boldsymbol{X}$, and $\Delta K(\boldsymbol{X}_i) = K_1(\boldsymbol{X}_i) - K_0(\boldsymbol{X}_i)$.

Theorem C.1. Under Assumptions C.1, C.2, C.3, and $s \log d\sqrt{\log n}/n^{1/2} = o(1)$, if we take $\lambda \asymp \lambda_1' \asymp \lambda_0' \asymp \sqrt{\log d/n}$, then

$$\widehat{\mu} - \mu^* = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{T_i}{\pi_i^*}(Y_i(1) - K_1(\boldsymbol{X}_i)) - \frac{1-T_i}{1-\pi_i^*}(Y_i(0) - K_0(\boldsymbol{X}_i)) + \Delta K(\boldsymbol{X}_i) - \mu^*\Big] + o_p(n^{-1/2}),$$

where $\max(\|\boldsymbol{\beta}^*\|_0, \|\boldsymbol{\alpha}_1^*\|_0, \|\boldsymbol{\alpha}_0^*\|_0) \le s$. It implies $n^{1/2}(\widehat{\mu} - \mu^*) \to_d N(0, V)$, where V is the semi-parametric asymptotic variance bound, i.e.,

$$V = \mathbb{E}\Big[\frac{1}{\pi^*}\mathbb{E}(\epsilon_1^2 \mid \boldsymbol{X}) + \frac{1}{1-\pi^*}\mathbb{E}(\epsilon_0^2 \mid \boldsymbol{X}) + (\Delta K(\boldsymbol{X}) - \mu^*)^2\Big].$$

Proof of this theorem is analogous to that of Theorem 2.1 and hence is omitted.

References

Athey, S., Imbens, G. W. and Wager, S. (2016). Efficient inference of average treatment effects in high dimensions via approximate residual balancing. *arXiv preprint arXiv:1604.07125* .

Belloni, A., Chernozhukov, V., Fernández-Val, I. and Hansen, C. (2013). Program evaluation with high-dimensional data. *arXiv preprint arXiv:1311.2645* .

Belloni, A., Chernozhukov, V. and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* **81** 608–650.

Belloni, A., Chernozhukov, V. and Wei, Y. (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics* **16** 606–619.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* **37** 1705–1732.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242.

Bühlmann, P. and van de Geer, S. (2015). High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics* **9** 1449–1473.

Cai, T. T. and Guo, Z. (2015). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *arXiv preprint arXiv:1506.05539* .

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.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. and Newey, W. (2016). Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060, 2016* .

Fan, J., Imai, K., Liu, H., Ning, Y. and Yang, X. (2016). Improving covariate balancing propensity score: A doubly robust and efficient approach. *Technical Report* .

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 849–911.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics* **189** 1–23.

Fong, C., Hazlett, C. and Imai, K. (2016). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. Tech. rep., Department of Politics, Princeton University.

Fong, C., Ratkovic, M., Hazlett, C. and Imai, K. (2015). CBPS: R package for covariate balancing propensity score. available at the Comprehensive R Archive Network (CRAN). http://CRAN.R-project.org/package=CBPS.

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.

Hernán, M. A. and Robins, J. M. (2017). *Causal inference*. Chapman & Hall/CRC, Boca Raton. Forthcoming.

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. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **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.

Imai, K. and van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* **99** 854–866.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87** 706–710.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Javanmard, A. and Montanari, A. (2013). Confidence intervals and hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1306.3171* .

Kam, C. D. and Palmer, C. L. (2008). Reconsidering the effects of education on political participation. *The Journal of Politics* **70** 612–631.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science* 523–539.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics* 1356–1378.

Li, G., Peng, H., Zhang, J. and Zhu, L. (2012). Robust rank correlation based screening. *The Annals of Statistics* 1846–1877.

Liu, J., Zhong, W. and Li, R. (2015). A selective overview of feature screening for ultrahigh-dimensional data. *Science China Mathematics* **58** 1–22.

Lounici, K. et al. (2008). Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of statistics* **2** 90–102.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* **23** 2937–2960.

Meinshausen, N., Meier, L. and Bühlmann, P. (2009). p-values for high-dimensional regression. *Journal of the American Statistical Association* **104** 1671–1681.

Ning, Y. and Liu, H. (2014). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *arXiv preprint arXiv:1412.8765* .

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York.

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

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

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.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge University Press, Cambridge.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics* **2** 808–840.

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H. and Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* **20** 512–522.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* 1166–1202.

Van der Vaart, A. V. (1998). *Asymptotic statistics*. Cambridge University Press, Cambridge, UK.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .

Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *The Annals of Statistics* 2178–2201.

Xu, C. and Chen, J. (2014). The sparse mle for ultrahigh-dimensional feature screening. *Journal of the American Statistical Association* **109** 1257–1269.

Yang, Z., Ning, Y. and Liu, H. (2014). On semiparametric exponential family graphical models. *arXiv preprint arXiv:1412.8697* .

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 217–242.

Zhang, T. (2010). Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research* **11** 1081–1107.

Zhao, Q. (2016). Covariate balancing propensity score by tailored loss functions. *arXiv preprint arXiv:1601.05890* .

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