Biometrika (2016), **XX**, XX, *pp*. 1–38

Printed in Great Britain

Advance Access publication on D M 2016

Supplementary material for "Robust Estimation of Causal Effects via High-Dimensional Covariate Balancing Propensity Score"

BY YANG NING

Department of Statistics and Data Science, Cornell University, Ithaca, New York 14853, U.S.A.
yn265@cornell.edu

SIDA PENG

Microsoft Research, Redmond, Washington 98052, U.S.A.
sidpeng@microsoft.com

AND KOSUKE IMAI

Department of Government and Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, U.S.A.
Imai@Harvard.Edu

Throughout this supplementary appendix, we use the following notation. For $v = (v_1, ..., v_d)^{\mathrm{T}} \in \mathbb{R}^d$, and $1 \leq q \leq \infty$, we define $\|v\|_q = (\sum_{i=1}^d |v_i|^q)^{1/q}$, $\|v\|_0 = |\text{supp}(v)|$, where $\text{supp}(v) = \{j : v_j \neq 0\}$ and $|A|$ is the cardinality of a set A. Denote $\|v\|_\infty = \max_{1\leq i\leq d} |v_i|$ and $v^{\otimes 2} = vv^{\mathrm{T}}$. For a matrix M, let $\|M\|_{\max} = \max_{jk} |M_{jk}|$, $\|M\|_1 = \sum_{jk} |M_{jk}|$, $\|M\|_{\ell_\infty} = \max_j \sum_k |M_{jk}|$. If the matrix M is symmetric, then $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ are the minimal and maximal eigenvalues of M. For $S \subseteq \{1, ..., d\}$, let $v_S = \{v_j : j \in S\}$ 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 $\text{pr}(|X| > t) \leq \exp(1 - t/K_1)$ for all $t \geq 0$. The sub-exponential norm of X is defined as $\|X\|_{\psi_1} = \sup_{p\geq 1} p^{-1}(E|X|^p)^{1/p}$. A random variable X is sub-Gaussian if there exists some constant $K_2 > 0$ such that $\text{pr}(|X| > t) \leq \exp(1 - t^2/K_2^2)$ for all $t \geq 0$. The sub-Gaussian norm of X is defined as $\|X\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}(E|X|^p)^{1/p}$.

S1. ESTIMATION OF THE AVERAGE CAUSAL EFFECTS

S1·1. *The Average Treatment Effect*

For clarification purpose, we first present the complete algorithm for the inference on the average treatment effect. Recall that we have constructed the inverse probability weighted estimator $\hat{\mu}_1$ in the main text. Now we focus on how to estimate $\mu_0^* = E\{Y(0)\}$.

Step 1: Define a generalized quasi-likelihood function as

$$Q_n(\beta) = -\frac{1}{n}\sum_{i=1}^n \int_0^{\beta^{\mathrm{T}} X_i} \Big\{\frac{1 - T_i}{1 - \pi(u)} - 1\Big\} w_1(u) du,$$

where $w_1(u)$ is an arbitrary positive weight function. Compute the estimator

$$\hat{\beta} \;=\; \arg\min_{\beta\in\mathbb{R}^d} -Q_n(\beta) + \lambda\|\beta\|_1,$$

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

Step 2: Define a weighted least square loss function using the treatment and control groups as

$$L_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n}(1-T_i)w_2(\hat{\beta}^{\mathrm{T}}X_i)(Y_i - \alpha^{\mathrm{T}}X_i)^2,$$

where $w_2(\cdot)$ is another positive weight function. Compute the estimator

$$\tilde{\alpha} \;=\; \arg\min_{\alpha\in\mathbb{R}^d} L_n(\alpha) + \lambda'\|\alpha\|_1,$$

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

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

$$\tilde{\gamma} \;=\; \arg\min_{\gamma\in\mathbb{R}^{|\tilde{S}|}} \|g_n(\gamma)\|_2^2 \text{ where } g_n(\gamma) = n^{-1}\sum_{i=1}^{n}\Big\{\frac{1-T_i}{1-\pi(\gamma^{\mathrm{T}}X_{i\tilde{S}} + \hat{\beta}_{\tilde{S}^c}^{\mathrm{T}}X_{i\tilde{S}^c})} - 1\Big\}X_{i\tilde{S}}$$

We then set $\tilde{\beta} = (\tilde{\gamma}, \hat{\beta}_{\tilde{S}^c})$ and $\tilde{\pi}_i = \pi(\tilde{\beta}^{\mathrm{T}}X_i)$.

Step 4: Estimate μ_0^* by the Horvitz-Thompson estimator $\hat{\mu}_0 = n^{-1}\sum_{i=1}^{n}(1-T_i)Y_i/(1-\tilde{\pi}_i)$. Then, we estimate the average treatment effect by $\hat{\mu} = \hat{\mu}_1 - \mu_0$.

Recall that $K_1(X_i) = E\{Y_i(1) \mid X_i\} = \alpha_1^{*\mathrm{T}}X$, $K_0(X_i) = E\{Y_i(0) \mid X_i\} = \alpha_0^{*\mathrm{T}}X$, and $\Delta K(X_i) = K_1(X_i) - K_0(X_i)$.

*Assumption S*1. Assume that $\epsilon_0 = Y(0) - \alpha_0^{*\mathrm{T}}X$ and $\|\epsilon_0\|_{\psi_2} \le C_\epsilon$, where C_ϵ is a positive constant.

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

THEOREM S1. *Under the same conditions in Theorem* 1 *and Assumptions S*1*, S*2*, then*

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

where $\max(\|\beta^*\|_0, \|\alpha_1^*\|_0, \|\alpha_0^*\|_0) \le s$*. It implies* $n^{1/2}(\hat{\mu} - \mu^*) \to_d N(0, V)$*, where* V *is the semiparametric asymptotic variance bound, i.e.,*

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

Proof of this theorem is analogous to that of Theorem 1 and hence is omitted. As discussed in Remark 1, the asymptotic variance can be obtained via the plug-in estimator. Similarly, Propositions 1 and 2 hold for $\hat{\mu}$ under either the misspecified propensity score or the misspecified outcome model. We omit the details.

S1·2. *The Average Treatment Effect for the Treated*

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

$T_i = 1\}$. By the law of total probability,

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

Thus, a simple estimator of τ_1^* is

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

To estimate τ_0^*, we notice that

$$\begin{aligned}\tau_0^* &= E[E\{Y_i(0) \mid T_i = 1, X_i\} \mid T_i = 1] = E[E\{Y_i(0) \mid X_i\} \mid T_i = 1] \\ &= \int E\{Y_i(0) \mid X_i\} \frac{\text{pr}(T_i = 1 \mid X_i) f(X_i)}{\text{pr}(T_i = 1)} dX_i = \frac{E[\pi(\beta^{*\mathrm{T}} X_i) E\{Y_i(0) \mid X_i\}]}{\text{pr}(T_i = 1)} \\ &= \frac{1}{\text{pr}(T_i = 1)} E\left\{\frac{\pi(\beta^{*\mathrm{T}} X_i)(1 - T_i) Y_i(0)}{1 - \pi(\beta^{*\mathrm{T}} X_i)}\right\}.\end{aligned}$$

Hence, to accurately estimate τ_0^*, one has to develop an alternative set of the covariate balancing equations. Recall that $\hat{\beta}$ is defined in equation (7). For notational simplicity, we denote the penalized least squared estimator for the control group by $\tilde{\alpha}$ with T_i replaced by $1 - T_i$ in equation (9). Recall that $\tilde{S} = \{j : |\tilde{\alpha}_j| > 0\}$ is the support of $\tilde{\alpha}$. Then, we calibrate the initial estimator $\hat{\beta}_{\tilde{S}}$ to balance $\bar{X}_{i\tilde{S}} = (1, X_{i\tilde{S}}^{\mathrm{T}})^{\mathrm{T}}$. Specifically, we solve $\tilde{\gamma} = \arg\min_{\gamma \in \mathbb{R}^{|\tilde{S}|+1}} \|g_n(\gamma)\|_2^2$, where

$$g_n(\gamma) = n^{-1} \sum_{i=1}^n \left\{T_i - \frac{(1 - T_i)\pi(\gamma^{\mathrm{T}} \bar{X}_{i\tilde{S}} + \hat{\beta}_{\tilde{S}^c}^{\mathrm{T}} X_{i\tilde{S}^c})}{1 - \pi(\gamma^{\mathrm{T}} \bar{X}_{i\tilde{S}} + \hat{\beta}_{\tilde{S}^c}^{\mathrm{T}} X_{i\tilde{S}^c})}\right\} \bar{X}_{i\tilde{S}}. \tag{S1}$$

Then, we set $\tilde{\pi}_i = \pi(\tilde{\beta}^{\mathrm{T}} \bar{X}_i)$ with $\tilde{\beta} = (\tilde{\gamma}, \hat{\beta}_{\tilde{S}^c})$ and estimate τ_0 by

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

where $\tilde{r}_i = \tilde{\pi}_i/(1 - \tilde{\pi}_i)$. The final estimator of the average treatment effect for the treated is $\hat{\tau} = \hat{\tau}_1 - \hat{\tau}_0$. The covariate balancing equations (S1) aim to balance the selected covariate $\bar{X}_{i\tilde{S}}$ reweighted by $\tilde{r}_i$ in the control group with $\bar{X}_{i\tilde{S}}$ in the treatment group. This proposal agrees with the intuition originated from some of the recent work (Hainmueller, 2012; Zubizarreta, 2015). Similar to Theorem 1, the following proposition establishes the asymptotic normality and semiparametric efficiency of the estimator $\hat{\tau}$.

PROPOSITION S1. *Under the same conditions in Theorem* 1*, we have,*

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

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

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

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

One future direction is to consider how to make the inference on average treatment effect for the treated robust to model misspecification.

S2. CONNECTION TO DOUBLY ROBUST ESTIMATOR

Recall that we estimate μ_1^* using the Horvitz-Thompson estimator. In the following, we comment on the connection between the proposed estimator and the other commonly used estimators. First, our estimator can be written as the Horvitz-Thompson estimator with the normalized weights, which is known as the Hajek estimator,

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\tilde{\pi}_i} = \frac{\sum_{i=1}^{n} T_i Y_i/\tilde{\pi}_i}{\sum_{i=1}^{n} T_i/\tilde{\pi}_i},$$

The second equality follows because $\sum_{i=1}^{n}(T_i/\tilde{\pi}_i - 1)/n = 0$ so long as an intercept is included in $X_{i\tilde{S}}$. Busso et al. (2014) showed that the normalized Horvitz-Thompson estimator tends to be more stable than the unnormalized version numerically. Thus, we expect that the proposed estimator has a better finite sample performance than the standard (i.e., unnormalized) Horvitz-Thompson estimator.

Second, our estimator can be also rewritten as an augmented inverse probability weighted estimator with the linear outcome model (Robins et al., 1994),

$$\hat{\mu}_1 \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\tilde{\pi}_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\tilde{\pi}_i} + \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{T_i}{\tilde{\pi}_i}\right)\tilde{\alpha}^{\mathrm{T}} X_i \tag{S2}$$

where the second equality follows from two equalities in (12).

As a side remark, Robins et al. (2007) showed that the ordinary least squared estimator $\hat{\mu}_{OLS} = n^{-1}\sum_{i=1}^{n} X_i\hat{\alpha}$ can be also viewed as an augmented inverse probability weighted estimator, as long as $\hat{\alpha}$ solves the following equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\tilde{\pi}_i}(Y_i - \hat{\alpha}^{\mathrm{T}} X_i) = 0.$$

This is because

$$\hat{\mu}_{OLS} = \frac{1}{n}\sum_{i=1}^{n} X_i\hat{\alpha} + \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\tilde{\pi}_i}(Y_i - \hat{\alpha}^{\mathrm{T}} X_i),$$

where the last expression is exactly the augmented inverse probability weighted estimator.

Finally, we note that when $w_1(u) = 1$ and $w_2(u) = \pi'(u)/\pi^2(u)$, the gradients of $Q_n(\beta)$ and $L_n(\alpha)$ are related to the estimating equations proposed by Robins et al. (2007). Let us consider the fixed dimensional case, and ignore all the penalization in our approach. One propensity score estimator proposed by Robins et al. (2007) is to solve the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i}{\pi(X_i^{\mathrm{T}}\beta)} - 1\right\} X_i \;=\; 0,$$

see section 3 of Robins et al. (2007). This is exactly our covariate balancing estimating equation. Moreover, Robins et al. (2007) considered the extended regression model by adding the covariate $\hat{\pi}_i$ which is the estimated propensity score. Specifically, under the linearity assumption, they assumed $E\{Y_i(1)|X_i\} = \alpha^{\mathrm{T}} X_i + \phi\hat{\pi}_i$ with some extra unknown parameter ϕ. The unknown parameters (α, ϕ) are estimated by the weighted least square estimator $(\hat{\alpha}, \hat{\phi})$, which solves the

following equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\hat{\pi}_i}(Y-\alpha^{\mathrm{T}}X_i-\phi\hat{\pi}_i)W_i \;=\; 0, \tag{S3}$$

where $W_i=(X_i,\hat{\pi}_i)$. To see the connection between (S3) and the gradient of the propensity score adjusted least square loss $L_n(\alpha)$, consider the logistic regression model for the propensity score. That is $\pi(u)=\exp(u)/\{1+\exp(u)\}$. After some simple calculation, we get

$$w_2(u)=\frac{\pi'(u)}{\pi^2(u)}=\frac{1-\pi(u)}{\pi(u)}=\frac{1}{\pi(u)}-1.$$

Thus, the gradient of $L_n(\alpha)$ denoted by $\nabla L_n(\hat{\alpha})$ is

$$\frac{1}{n}\sum_{i=1}^{n}T_i\Big(\frac{1}{\hat{\pi}_i}-1\Big)(Y-\alpha^{\mathrm{T}}X_i)X_i=\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\hat{\pi}_i}(Y-\alpha^{\mathrm{T}}X_i)X_i-\frac{1}{n}\sum_{i=1}^{n}T_i(Y-\alpha^{\mathrm{T}}X_i)X_i.$$

If the outcome model is correctly specified, we would expect that the estimator $\hat{\phi}$ in (S3) is close to 0. Thus, we can ignore the term $\phi\hat{\pi}_i$ in (S3). Then (S3) implies that the estimator $\hat{\alpha}$ satisfies

$$\nabla L_n(\hat{\alpha})\approx 0.$$

The above derivation illustrates the connection between (S3) and $\nabla L_n(\hat{\alpha})$. However, the above derivation relies critically on the particular structure of the logistic regression $\pi(u)$, and (S3) could differ from the gradient of the propensity score adjusted least square loss in more general settings.

S3. Comparison with Zhao (2019)

In another recent work, Zhao (2019) proposed a generalized covariate balancing method based on a class of scoring rules, which is similar to our generalized quasi-likelihood approach. Many existing covariate balancing estimators can be treated as the primal or dual problems of their optimization problem. Zhao (2019) studied the robustness of these estimators to misspecified propensity score models under the constant treatment effect model $E\{Y(1)-Y(0)\mid X\}=\mu^*$ for some constant μ^*. In contrast, our methodology allows for the heterogeneity of causal effects. In addition, while our work mainly focuses on the high-dimensional settings, Zhao (2019) does not provide statistical guarantees in such settings.

Both the proposed generalized quasi-likelihood function and Zhao (2016)'s "tailored loss function" aim to estimate the propensity score model via the covariate balancing idea. However, they differ in some details. To be specific, by equations (8) and (9) in Zhao (2016) his score function has the form

$$\frac{1}{n}\sum_{i=1}^{n}X_i(2T_i-1)\frac{G''\{p_\beta(X_i)\}}{\ell'\{p_\beta(X_i)\}}[T_i\{1-p_\beta(X_i)\}+(1-T_i)p_\beta(X_i)]=0, \tag{S4}$$

where $p_\beta(X_i)$ is the propensity score model. The parameter β in the propensity score model can be estimated by the root of the above equation. When the average treatment effect is of interest, the tailored loss function corresponds to $G''(p)=1/\{p^2(1-p)^2\}$. Moreover, if we assume the logistic regression for the propensity score model, we have $\ell(p)=\mathrm{logit}(p)$ and thus $\ell'(p)=1/\{p(1-p)\}$. Plugging these formulas into equation (S4), we can show that the score

function in Zhao (2016) is

$$\frac{1}{n}\sum_{i=1}^{n} X_i \left\{ \frac{T_i}{p_\beta(X_i)} - \frac{1-T_i}{1-p_\beta(X_i)} \right\} = 0. \tag{S5}$$

Thus, when the average treatment effect is of interest, the proposal in Zhao (2016) is the same as the covariate balancing propensity score method in Imai & Ratkovic (2014) with only the linear term X. As a comparison, our score equation (11) has the form

$$\frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{T_i}{p_\beta(X_i)} - 1 \right\} X_i = 0,$$

where we use the notation $p_\beta(X_i)$ for the propensity score model to match with (S4). These two score functions are different as well as the estimators of β. Indeed, Fan et al. (2016) investigated the properties of the estimator, the root of the estimating equation (S5). They showed that the inverse probability weighted estimator with the propensity score estimated by solving equation (S5) is not consistent for the average treatment effect if the propensity score model is misspecified and the outcome model is linear in X. Thus, the inverse probability weighted estimator is not doubly robust in this case. Thus, we cannot replace the generalized quasi-likelihood function in Step 1 of our algorithm by the tailored loss function in Zhao (2016).

S4. PROOFS

S4·1. *Proof of Theorem* 1

For simplicity of the proof, we focus on the exact sparse case and ignore the approximation errors in the propensity score and outcome models. The treatment of the approximation errors is similar to Belloni et al. (2017) and is not essential for the validity of the proposed method.

To start the proof of Theorem 1, we make the following minor modifications on the estimation of γ in step 3. Define

$$\tilde{\gamma} = \arg\min_{\gamma\in\Omega} \|g_n(\gamma)\|_2^2, \text{ where } \Omega = \{\gamma \in \mathbb{R}^{|\tilde{S}|} : \|\gamma - \hat{\beta}_{\tilde{S}}\|_1 \le \delta/\log n\}$$

for some small constant $\delta > 0$. Here, we introduce a parameter set Ω for γ, which guarantees the existence of a minimizer $\tilde{\gamma}$ 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 $\tilde{\gamma}$ defined in step 3 is still very close to $\hat{\beta}_{\tilde{S}}$. This suggests that the estimator automatically lies within the set Ω. So, this modification has little practical implication.

For notational simplicity, we denote $\mathrm{pr}_n f(X) = n^{-1}\sum_{i=1}^n f(X_i)$, and we use C to denote a generic constant, whose value may change from line to line.

Denote $V_i(u) = T_i\pi'(u)/\pi^2(u)w_1(u) - \{T_i/\pi(u) - 1\}w_1'(u)$. Let S_1 denote the support set of β^*. Define the compatibility factor for $Q_n(\beta)$ in a small neighborhood of the true value as

$$\kappa = \inf_{\{u\in\mathbb{R}^n : |u_i - X_i^{\mathrm{T}}\beta^*| \le \delta\}} \inf_{\{v\in\mathbb{R}^d : \|v_{S_1^c}\|_1 \le 5\|v_{S_1}\|_1\}} \frac{s_1 W(u,v)}{\|v_{S_1}\|_1^2} \tag{S6}$$

where $W(u,v) = n^{-1}\sum_{i=1}^n V_i(u_i)(X_i^{\mathrm{T}}v)^2$ and $\delta = C\lambda s\{\log(d\vee n)\}^{1/2}$ for some positive constant C. Our first lemma shows that κ is lower bounded by a positive constant under the assumptions in Theorem 1.

LEMMA S1. *Under the assumptions in Theorem* 1, $\kappa \geq C > 0$ *with probability approaching 1.*

*Proof Proof of Lemma S*1. Denote $V_i = V_i(X_i^{\mathrm{T}}\beta^*)$. We have that

$$\begin{aligned}\frac{s_1 W(u,v)}{\|v_{S_1}\|_1^2} &\geq \frac{W(u,v)}{\|v_{S_1}\|_2^2} \\ &\geq \frac{n^{-1}\sum_{i=1}^n V_i (X_i^{\mathrm{T}} v)^2}{\|v_{S_1}\|_2^2} - \Big| \frac{n^{-1}\sum_{i=1}^n \{V_i(u_i) - V_i\}(X_i^{\mathrm{T}} v)^2}{\|v_{S_1}\|_2^2} \Big| := (i.1) - (i.2).\end{aligned} \tag{135}$$

First, we consider the first term $(i.1)$. Recall that $E(V_i|X_i) = \pi'(X_i^{\mathrm{T}}\beta^*) w_1(X_i^{\mathrm{T}}\beta^*)/\pi(X_i^{\mathrm{T}}\beta^*) \geq C > 0$. Let $E = n^{-1}\sum_{i=1}^n V_i X_i^{\otimes 2} - E(V_i X_i^{\otimes 2})$. Thus,

$$(i.1) \geq \frac{E\{V_i (X_i^{\mathrm{T}} v)^2\}}{\|v_{S_1}\|_2^2} - \frac{|v^{\mathrm{T}} E v|}{\|v_{S_1}\|_2^2} \geq C \frac{E\{(X^{\mathrm{T}} v)^2\}}{\|v_{S_1}\|_2^2} - \frac{|v^{\mathrm{T}} E v|}{\|v_{S_1}\|_2^2} := (i.1.1) - (i.1.2).$$

For the first term (i.1.1), we partition the set S_1^c as the union of K disjoint sets $S_1^c = \cup_{k=1}^K J_k$, where J_k contains the indices that has the m largest (in absolute value) entries of v outside $\cup_{j=1}^{k-1} J_j$. We take $m = s_1 \log n$ or the largest integer no greater than $s_1 \log n$. Then, $|J_k| = m$ and $|J_K| \leq m$. It has been shown that $\|v_{J_{k+1}}\|_2 \leq m^{-1/2}\|v_{J_k}\|_1$ for any $1 \leq k \leq K-1$. Then, it implies

$$\sum_{k=1}^K \|v_{J_k}\|_2 \leq \frac{\|v_{S_1^c}\|_1}{m^{1/2}} \leq \frac{5\|v_{S_1}\|_1}{m^{1/2}} \leq \frac{5 s_1^{1/2}\|v_{S_1}\|_2}{m^{1/2}} = \frac{5\|v_{S_1}\|_2}{(\log n)^{1/2}}.$$

This leads to

$$\{E(X_{S_1^c}^{\mathrm{T}} v_{S_1^c})^2\}^{1/2} = \{E(\sum_{k=1}^K X_{J_k}^{\mathrm{T}} v_{J_k})^2\}^{1/2} \leq \sum_{k=1}^K \{E(X_{J_k}^{\mathrm{T}} v_{J_k})^2\}^{1/2} \leq C \sum_{k=1}^K \|v_{J_k}\|_2 \lesssim \frac{\|v_{S_1}\|_2}{(\log n)^{1/2}},$$

where we use the fact that the largest eigenvalue of a submatrix of Σ with size m is bounded by Assumption 5. Then, the term (i.1.1) can be bounded from below by C times

$$\begin{aligned}\frac{E(X_{S_1}^{\mathrm{T}} v_{S_1} + X_{S_1^c}^{\mathrm{T}} v_{S_1^c})^2}{\|v_{S_1}\|_2^2} &\geq \frac{1}{2}\frac{E(X_{S_1}^{\mathrm{T}} v_{S_1})^2}{\|v_{S_1}\|_2^2} - \frac{E(X_{S_1^c}^{\mathrm{T}} v_{S_1^c})^2}{\|v_{S_1}\|_2^2} \\ &\geq \frac{1}{2}\lambda_{\min}(\Sigma_{S_1 S_1}) - C/\log n \geq C,\end{aligned}$$

as the smallest eigenvalue of a submatrix of Σ with size s_1 is bounded from below by a constant by Assumption 5. For (i.1.2), we have that

$$|v^{\mathrm{T}} E v| \leq \|v\|_1^2 \|E\|_\infty \leq 36\|v_{S_1}\|_1^2 \|E\|_\infty \leq 36 s_1 \|v_{S_1}\|_2^2 \|E\|_\infty.$$

By Assumption 6, we can show that $|V_i| \leq C$ for some constant C. Thus, $V_j X_{ij} X_{ik}$ is sub-exponential with $\|V_j X_{ij} X_{ik}\|_{\psi_1} \leq C$ by Assumption 3. The Bernstein inequality for sub-exponential random variables (Lemma K.2 of Ning & Liu (2017)) and the union bound yield $\|E\|_\infty = O_p((\log d/n)^{1/2})$. Together with the sparsity assumption, we have

$$\sup_{\{v \in \mathbb{R}^d : \|v_{S_1^c}\|_1 \leq 5\|v_{S_1}\|_1\}} (i.1.2) = O_p(s_1\{\log(d \vee n)/n\}^{1/2})$$

Thus, we obtain that $\inf(i.1) \geq C - o_p(1)$, where the inf is taken in the same set. 140

In the following, we focus on $(i.2)$. First, it is easily verified that $V_i(u_i)$ is locally Lipschitz in a small neighborhood of $X_i^{\mathrm{T}}\beta^*$ given Assumption 6. Then

$$\sup_{\{u_i\in\mathbb{R}:|u_i-X_i^{\mathrm{T}}\beta^*|\le\delta\}} |V_i(u_i)-V_i| \le C\delta.$$

Similar to the analysis for the lower bound of $n^{-1}\sum_{i=1}^n (X_i^{\mathrm{T}}v)^2$ in term $(i.1)$, we can also prove that

$$\sup_{\{v\in\mathbb{R}^d:\|v_{S_1^c}\|_1\le 5\|v_{S_1}\|_1\}} \frac{n^{-1}\sum_{i=1}^n (X_i^{\mathrm{T}}v)^2}{\|v_{S_1}\|_2^2} \le 2\lambda_{\max}(\Sigma_{S_1S_1}) + C/\log n \le C.$$

Thus, taking the maximum over u and v for the $(i.2)$ term, we bound (i.2) from above by

$$\sup_{\{v\in\mathbb{R}^d:\|v_{S^c}\|_1\le 5\|v_S\|_1\}} \frac{n^{-1}\sum_{i=1}^n (X_i^{\mathrm{T}}v)^2}{\|v_{S_1}\|_2^2} C\delta \le \{2\lambda_{\max}(\Sigma_{S_1S_1}) + C'/\log n\}C\delta = o_p(1),$$

since $\delta = C\lambda s\{\log(d\vee n)\}^{1/2} = o_p(1)$ by the sparsity Assumption 4. Combining the bounds for $(i.1)$ and $(i.2)$, we complete the proof. □

LEMMA S2. *Under the assumptions in Theorem* 1,

$$\|\hat\beta-\beta^*\|_1 \;=\; O_p\left(s_1\Big\{\frac{\log(d\vee n)}{n}\Big\}^{1/2}\right),\;\; pr_n[X^{\mathrm{T}}(\hat\beta-\beta^*)]^2 = O_p\left(\frac{s_1\log(d\vee n)}{n}\right).$$

*Proof Proof of Lemma S*2. The proof contains two steps. In the first step, we apply a localization trick. Define $B = 20s_1\lambda/(3\kappa)$ and $t = B/(B+\|\hat\beta-\beta^*\|_1)$. Define $\bar\beta = t\hat\beta + (1-t)\beta^*$ to be a convex combination of $\hat\beta$ and β^*. In the first step, we analyze the estimator $\bar\beta$ and will prove that

$$\|\bar\beta-\beta^*\|_1 \;\le\; B/2. \tag{S7}$$

Note that we trivially have $\|\bar\beta-\beta^*\|_1 = B\|\hat\beta-\beta^*\|_1/(B+\|\hat\beta-\beta^*\|_1) \le B$. Since $B = o(1)$ by our sparsity assumption, we can see that by construction $\bar\beta$ is already in a small neighborhood of β^*. Thus, it is a localization step. The Hessian matrix of $-Q_n(\beta)$ is

$$-\nabla^2 Q_n(\beta) = \frac{1}{n}\sum_{i=1}^n \Big\{\frac{T_i\pi'(X_i^{\mathrm{T}}\beta)w_1(X_i^{\mathrm{T}}\beta)}{\pi^2(X_i^{\mathrm{T}}\beta)} - \frac{(T_i-\pi(X_i^{\mathrm{T}}\beta))w_1'(X_i^{\mathrm{T}}\beta)}{\pi(X_i^{\mathrm{T}}\beta)}\Big\}X_i^{\otimes 2},$$

which is semi-positive definite since we only consider the convex loss function, i.e., Q_n is concave. Thus,

$$\begin{aligned} Q_n(\bar\beta)-\lambda\|\bar\beta\|_1 &\ge t\{Q_n(\hat\beta)-\lambda\|\hat\beta\|_1\} + (1-t)\{Q_n(\beta^*)-\lambda\|\beta^*\|_1\} \\ &\ge Q_n(\beta^*)-\lambda\|\beta^*\|_1, \end{aligned}$$

where the last step follows by the definition of $\hat\beta$. By rearranging above inequality, we obtain

$$D(\bar\beta,\beta^*) + \lambda\|\bar\beta\|_1 \le \nabla Q_n(\beta^*)(\bar\beta-\beta^*) + \lambda\|\beta^*\|_1$$

where $D(\bar\beta,\beta^*) = Q_n(\beta^*) - Q_n(\bar\beta) + \nabla Q_n(\beta^*)(\bar\beta-\beta^*)$, which is nonnegative by the concavity of Q_n. This implies that under the event $E_1 := \{\|\nabla Q_n(\beta^*)\|_\infty \le \lambda/3\}$, we have

$$D(\bar\beta,\beta^*) + \lambda\|\bar\beta\|_1 \le \|\nabla Q_n(\beta^*)\|_\infty\|\bar\beta-\beta^*\|_1 + \lambda\|\beta^*\|_1 \le B\lambda/3 + \lambda\|\beta^*\|_1.$$

We note that $\nabla Q_n(\beta^*) = n^{-1}\sum_{i=1}^n W_i$, where $W_i = \{T_i/\pi(X_i^{\mathrm{T}}\beta^*) - 1\}w_1(X_i^{\mathrm{T}}\beta^*)X_i$. Since $\{T_i/\pi(X_i^{\mathrm{T}}\beta^*) - 1\}w_1(X_i^{\mathrm{T}}\beta^*)$ is bounded by a constant, W_i is sub-Gaussian. Applying the Hoeffding inequality and the union bound, we can show that the event E_1 holds with probability at least $1/(d \vee n)$.

Recall that $\|\bar{\beta}\|_1 = \|\bar{\beta}_{S_1}\|_1 + \|\bar{\beta}_{S_1^c}\|_1$ and $\beta^*_{S_1^c} = 0$. We have

$$D(\bar{\beta}, \beta^*) + \lambda\|\bar{\Delta}_{S_1^c}\|_1 \leq B\lambda/3 + \lambda\|\bar{\Delta}_{S_1}\|_1,$$

where $\bar{\Delta} = \bar{\beta} - \beta^*$. We prove (S7) by contradiction. If $\|\bar{\beta} - \beta^*\|_1 > B/2$, (S8) implies

$$D(\bar{\beta}, \beta^*) + \lambda\|\bar{\Delta}_{S_1^c}\|_1 \leq 2\lambda/3\|\bar{\Delta}\|_1 + \lambda\|\bar{\Delta}_{S_1}\|_1.$$

which can be rewritten as

$$D(\bar{\beta}, \beta^*) + (1/3)\lambda\|\bar{\Delta}_{S_1^c}\|_1 \leq (5/3)\lambda\|\bar{\Delta}_{S_1}\|_1. \tag{S8}$$

Since $D(\bar{\beta}, \beta^*) \geq 0$, we obtain $\|\bar{\Delta}_{S_1^c}\|_1 \leq 5\|\bar{\Delta}_{S_1}\|_1$. This gives us the cone condition in the compatibility factor. Note that with high probability $\max_i |X_i^{\mathrm{T}}\bar{\Delta}| \leq \max_i \|X_i\|_\infty\|\bar{\Delta}\|_1 \leq C\{\log(dn)\}^{1/2}B$, where we use the tail bound for sub-Gaussian variables in the last step. Thus, the event $E_2 := \{\max_i |X_i^{\mathrm{T}}\bar{\Delta}| \leq C\{\log(dn)\}^{1/2}B\}$ holds with probability tending to 1. Recall that $W(u, v) = n^{-1}\sum_{i=1}^n V_i(u_i)(X_i^{\mathrm{T}}v)^2$. Under the event E_2, by the mean-value theorem

$$D(\bar{\beta}, \beta^*) \geq \inf_{\{u\in\mathbb{R}^n : |u_i - X_i^{\mathrm{T}}\beta^*| \leq C\{\log(dn)\}^{1/2}B\}} \frac{W(u, \bar{\beta} - \beta^*)}{2} \geq \frac{\kappa\|\bar{\Delta}_{S_1}\|_1^2}{2s_1},$$

where κ is the compatibility factor defined in (S6). Plugging it into (S8), we obtain

$$\frac{\kappa\|\bar{\Delta}_{S_1}\|_1^2}{2s_1} \leq \frac{5}{3}\lambda\|\bar{\Delta}_{S_1}\|_1.$$

This gives us $\|\bar{\Delta}_{S_1}\|_1 \leq 10s_1\lambda/(3\kappa) = B/2$. This leads to the contradiction. Thus, (S7) is true.

In the second step, we link $\bar{\beta}$ back to $\hat{\beta}$ and prove this lemma. Since $\|\bar{\beta} - \beta^*\|_1 = B\|\hat{\beta} - \beta^*\|_1/(B + \|\hat{\beta} - \beta^*\|_1)$, if (S7) is true, this implies $\|\hat{\beta} - \beta^*\|_1 \leq B = 20s_1\lambda/(3\kappa)$. Lemma S1 implies $\kappa \geq C > 0$. With $\lambda = C\{\log(d \vee n)/n\}^{1/2}$, we have

$$\|\hat{\beta} - \beta^*\|_1 \;=\; O_p\Big(s_1\Big\{\frac{\log(d \vee n)}{n}\Big\}^{1/2}\Big).$$

Finally, we repeat the first step with $\bar{\beta}$ replaced by $\hat{\beta}$, we can derive $\|\hat{\beta} - \beta^*\|_2^2 \leq Cs_1\lambda^2$ and $D(\hat{\beta}, \beta^*) \leq 100s_1\lambda^2/(9\kappa)$ by (S8). Following the proof of Lemma S1, we can obtain $D(\hat{\beta}, \beta^*) \geq \{C - o_p(1)\}\mathrm{pr}_n\{X^{\mathrm{T}}(\hat{\beta} - \beta^*)\}^2$. Thus,

$$\mathrm{pr}_n\{X^{\mathrm{T}}(\hat{\beta} - \beta^*)\}^2 = O_p\Big(\frac{s_1\log d}{n}\Big).$$

LEMMA S3. *Under the assumptions in Theorem* 1*,*

$$\|\tilde{\alpha} - \alpha^*\|_1 \;=\; O_p\left((s_1 \vee s_2)\Big\{\frac{\log(d \vee n)}{n}\Big\}^{1/2}\right),\; pr_n\{X^{\mathrm{T}}(\tilde{\alpha} - \alpha^*)\}^2 = O_p\left(\frac{(s_1 \vee s_2)\log(d \vee n)}{n}\right).$$

The proof of this lemma is very similar to Lemma E.5 of Ning & Liu (2017). We omit the details. For notational simplicity, we denote $s = s_1 \vee s_2$.

LEMMA S4. *Under the assumptions in Theorem* 1,

$$\|\tilde{\beta}-\beta^*\|_1 \;=\; O_p\left(\Big\{\frac{s^2\log(d\vee n)}{n}\Big\}^{1/2}\right),\;\; pr_n\{X^{\mathrm{T}}(\tilde{\beta}-\beta^*)\}^2 \;=\; O_p\left(\frac{s\log(d\vee n)}{n}\right).$$

*Proof Proof of Lemma S*4. The proof contains the following three steps:

(i) $|\tilde{S}| \le Cs$, where C is a positive constant.
(ii) $\|\hat{\beta}_{\tilde{S}^c}-\beta^*_{\tilde{S}^c}\|_1 = O_p(s_1\{\log(d\vee n)/n\}^{1/2})$, and $\mathrm{pr}_n\{(\hat{\beta}-\beta^*)^{\mathrm{T}}_{\tilde{S}^c}X_{\tilde{S}^c}\}^2 = O_p(s_1\log(d\vee n)/n)$.
(iii) $\|\tilde{\gamma}-\gamma^*\|_1 \;=\; O_p(\{s^2\log(d\vee n)/n\}^{1/2})$, and

$$\mathrm{pr}_n\{(\hat{\gamma}-\gamma^*)^{\mathrm{T}}X_{\tilde{S}}\}^2 \;=\; O_p\left(\frac{s\log(d\vee n)}{n}\right).$$

We first show (i). By the KKT condition of the penalized least square regression for $\tilde{\alpha}$, we have $\mathrm{pr}_n T\hat{w}_2X_j(Y-X^{\mathrm{T}}\tilde{\alpha}) = -\lambda'\mathrm{sign}(\tilde{\alpha}_j)$ for any $j\in\mathrm{supp}(\tilde{\alpha})$, where $\hat{w}_2 = w_2(\hat{\beta}^{\mathrm{T}}X)$. Then, we have

$$\begin{aligned}\lambda'|\tilde{S}|^{1/2} &= \|\mathrm{pr}_n T\hat{w}_2X_{\tilde{S}}(Y-X^{\mathrm{T}}\tilde{\alpha})\|_2\\ &\le \|\mathrm{pr}_n Tw_2X_{\tilde{S}}(Y-X^{\mathrm{T}}\tilde{\alpha})\|_2 + \|\mathrm{pr}_n T(\hat{w}_2-w_2)X_{\tilde{S}}(Y-X^{\mathrm{T}}\tilde{\alpha})\|_2, \qquad \text{(S9)}\end{aligned}$$

where $w_2 = w_2(\beta^{*\mathrm{T}}X)$. For the first term,

$$\begin{aligned}&\|\mathrm{pr}_n Tw_2X_{\tilde{S}}(Y-X^{\mathrm{T}}\tilde{\alpha})\|_2 \le \|\mathrm{pr}_n Tw_2X_{\tilde{S}}\epsilon_1\|_2 + \|\mathrm{pr}_n Tw_2X_{\tilde{S}}X^{\mathrm{T}}(\tilde{\alpha}-\alpha^*)\|_2\\ &\le |\tilde{S}|^{1/2}\|\mathrm{pr}_n Tw_2X_{\tilde{S}}\epsilon_1\|_\infty + \{\mathrm{pr}_n[X^{\mathrm{T}}(\tilde{\alpha}-\alpha^*)]^2\}^{1/2}\|\mathrm{pr}_n Tw_2X_{\tilde{S}}X^{\mathrm{T}}_{\tilde{S}}\|_2^{1/2}\\ &\le C|\tilde{S}|^{1/2}\{\log(|\tilde{S}|\vee n)/n\}^{1/2} + C\|\mathrm{pr}_n Tw_2X_{\tilde{S}}X^{\mathrm{T}}_{\tilde{S}}\|_2^{1/2}\{(s_1\vee s_2)\log(d\vee n)/n\}^{1/2}, \qquad \text{(S10)}\end{aligned}$$

where the last step follows from Lemma S3 and $\|\mathrm{pr}_n Tw_2X_{\tilde{S}}\epsilon_1\|_\infty \le \max_{|S|\le|\tilde{S}|}\max_{j\in S}\|\mathrm{pr}_n Tw_2X_j\epsilon_1\|_\infty$. Since $Tw_2\le C$, $\|\epsilon_1\|_{\psi_2}\le C_\epsilon$ and $\|X_j\|_{\psi_2}\le C_X$ for any $1\le j\le d$, we have $\|Tw_2X_j\epsilon_1\|_{\psi_1}\le 2CC_\epsilon C_X$. The Bernstein inequality for sub-exponential random variables (Lemma K.2 of Ning & Liu (2017)) yields,

$$\mathrm{pr}\left(\frac{1}{n}\sum_{i=1}^n Tw_2X_{ij}\epsilon_{1i} > t\right) \;\le\; 2\exp\left\{-C''\min\left(\frac{t^2}{4C^2C_\epsilon^2C_X^2}, \frac{t}{2CC_\epsilon C_X}\right)\right\},$$

where C'' is a universal constant. Applying the union bound argument and choose $t = \{\log(|\tilde{S}|\vee n)/n\}^{1/2}$, we can obtain $\|\mathrm{pr}_n Tw_2X_{\tilde{S}}\epsilon_1\|_\infty = O_p(\{\log(|\tilde{S}|\vee n)/n\}^{1/2})$. By equation (S10) and $\lambda'\asymp\{\log(d\vee n)/n\}^{1/2}$, we have

$$|\tilde{S}|^{1/2}\le C\|\mathrm{pr}_n X_{\tilde{S}}X^{\mathrm{T}}_{\tilde{S}}\|_2(s_1\vee s_2)^{1/2}.$$

Since the sparse eigenvalue is sub-linear (e.g., Yang et al. (2018)), there exists a constant $C'>0$ such that $\|\mathrm{pr}_n X_{\tilde{S}}X^{\mathrm{T}}_{\tilde{S}}\|_2\le C'$ with probability tending to one. Thus, equation (S10) implies (i). The same argument can be applied to the second term in (S9). We arrive at the same conclusion.

To show (ii), note that $\|\hat{\beta}_{\tilde{S}^c}-\beta^*_{\tilde{S}^c}\|_1\le\|\hat{\beta}-\beta^*\|_1 = O_p(s_1\{\log(d\vee n)/n\}^{1/2})$, where the last step follows from Lemma S2. In addition, $\lambda_{\max}(\mathrm{pr}_n X_{\tilde{S}}X^{\mathrm{T}}_{\tilde{S}}) = O_p(1)$. To see this, by Weyl's

inequality,

$$|\lambda_{\max}(\mathrm{pr}_n X_{\tilde{S}} X_{\tilde{S}}^{\mathrm{T}}) - \lambda_{\max}(\mathrm{pr} X_{\tilde{S}} X_{\tilde{S}}^{\mathrm{T}})| \leq \|(\mathrm{pr}_n - \mathrm{pr}) X_{\tilde{S}} X_{\tilde{S}}^{\mathrm{T}}\|_2 \leq \max_{|S| \leq C s_2} \|(\mathrm{pr}_n - \mathrm{pr}) X_S X_S^{\mathrm{T}}\|_2$$
$$\leq C s_2 \|(\mathrm{pr}_n - \mathrm{pr}) X_S X_S^{\mathrm{T}}\|_{\max} = O_p(s_2 (\log s_2 / n)^{1/2}),$$

where in the last step we use the Bernstein inequality for sub-exponential random variable and the standard union bound argument. Since $\max_{|S| \leq C s} \lambda_{\max}(\mathrm{pr} X_S X_S^{\mathrm{T}}) \leq 1/C$ by the eigenvalue assumption, we obtain that $\lambda_{\max}(\mathrm{pr}_n X_{\tilde{S}} X_{\tilde{S}}^{\mathrm{T}}) = O_p(1)$. For the second result in (ii), we have

$$\mathrm{pr}_n \{(\hat{\beta} - \beta^*)_{\tilde{S}^c}^{\mathrm{T}} X_{\tilde{S}^c}\}^2 \leq 2\mathrm{pr}_n \{(\hat{\beta} - \beta^*)^{\mathrm{T}} X\}^2 + 2\mathrm{pr}_n \{(\hat{\beta} - \beta^*)_{\tilde{S}}^{\mathrm{T}} X_{\tilde{S}}\}^2 = O_p(s_1 \log(d \vee n)/n)$$

where the last step follows from Lemma S2, and

$$\mathrm{pr}_n \{(\hat{\beta} - \beta^*)_{\tilde{S}}^{\mathrm{T}} X_{\tilde{S}}\}^2 \leq \|(\hat{\beta} - \beta^*)_{\tilde{S}}^{\mathrm{T}}\|_2^2 \lambda_{\max}(\mathrm{pr}_n X_{\tilde{S}} X_{\tilde{S}}^{\mathrm{T}}) = O_p(s_1 \log(d \vee n)/n).$$

This completes the proof of (ii).

In the following, we aim to show (iii). For notational simplicity, let $\pi = \pi(\gamma^{*\mathrm{T}} X_{\tilde{S}} + \hat{\beta}_{\tilde{S}^c}^{\mathrm{T}} X_{\tilde{S}^c})$ and $\tilde{\pi} = \pi(\tilde{\gamma}^{\mathrm{T}} X_{\tilde{S}} + \hat{\beta}_{\tilde{S}^c}^{\mathrm{T}} X_{\tilde{S}^c})$. By the definition of $\tilde{\gamma}$, we have

$$\left\| \mathrm{pr}_n \left(\frac{T}{\pi} - 1 \right) X_{\tilde{S}} \right\|_2^2 \geq \left\| \mathrm{pr}_n \left(\frac{T}{\tilde{\pi}} - 1 \right) X_{\tilde{S}} \right\|_2^2$$
$$= \left\| \mathrm{pr}_n \left(\frac{T}{\pi} - 1 \right) X_{\tilde{S}} \right\|_2^2 + \left\| \mathrm{pr}_n \left(\frac{T}{\tilde{\pi}} - \frac{T}{\pi} \right) X_{\tilde{S}} \right\|_2^2 + 2\mathrm{pr}_n \left(\frac{T}{\pi} - 1 \right) X_{\tilde{S}} \mathrm{pr}_n \left(\frac{T}{\tilde{\pi}} - \frac{T}{\pi} \right) X_{\tilde{S}}. \quad (190)$$

The first inequality comes from $\gamma^* \in \Omega$, since $\|\hat{\beta} - \beta^*\|_1 = O_p(s_1 \{\log(d \vee n)/n\}^{1/2})$. This yields

$$\left\| \mathrm{pr}_n \left(\frac{T}{\tilde{\pi}} - \frac{T}{\pi} \right) X_{\tilde{S}} \right\|_2^2 \leq -2\mathrm{pr}_n \left(\frac{T}{\pi} - 1 \right) X_{\tilde{S}} \mathrm{pr}_n \left(\frac{T}{\tilde{\pi}} - \frac{T}{\pi} \right) X_{\tilde{S}}. \quad \text{(S11)}$$

Let π' denote the derivative of π evaluated at an intermediate value between $\gamma^{*\mathrm{T}} X_{\tilde{S}} + \hat{\beta}_{\tilde{S}^c}^{\mathrm{T}} X_{\tilde{S}^c}$ and $\tilde{\gamma}^{\mathrm{T}} X_{\tilde{S}} + \hat{\beta}_{\tilde{S}^c}^{\mathrm{T}} X_{\tilde{S}^c}$. Then

$$\left\| \mathrm{pr}_n \left(\frac{T}{\tilde{\pi}} - \frac{T}{\pi} \right) X_{\tilde{S}} \right\|_2^2 = \left\| \mathrm{pr}_n \frac{T \pi'}{\tilde{\pi} \pi} X_{\tilde{S}}^{\otimes 2} (\tilde{\gamma} - \gamma^*) \right\|_2^2$$
$$\geq \|\tilde{\gamma} - \gamma^*\|_2^2 \lambda_{\min} \left(\mathrm{pr}_n \frac{T \pi'}{\tilde{\pi} \pi} X_{\tilde{S}}^{\otimes 2} \right) \geq C \|\tilde{\gamma} - \gamma^*\|_2^2 \lambda_{\min} \left(\mathrm{pr}_n T X_{\tilde{S}}^{\otimes 2} \right),$$

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

$$\max_{1 \leq i \leq n} |\tilde{\pi}_i' - \pi_i^{*\prime}| \leq \max_{1 \leq i \leq n} \{\|\tilde{\gamma} - \gamma^*\|_1 \|X_{i\tilde{S}}\|_\infty + \|\hat{\beta}_{\tilde{S}^c} - \beta_{\tilde{S}^c}^*\|_1 \|X_{i\tilde{S}^c}\|_\infty\} = o_p(1),$$

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

$$|\lambda_{\min}(\mathrm{pr}_n T X_{\tilde{S}}^{\otimes 2}) - \lambda_{\min}(\mathrm{pr} T X_{\tilde{S}}^{\otimes 2})| \leq C s \|(\mathrm{pr}_n - \mathrm{pr}) T X_{\tilde{S}}^{\otimes 2}\|_{\max} = O_p(s (\log s / n)^{1/2}).$$

Since $\min_{|S| \leq C s} \lambda_{\min}(\mathrm{pr} T X_S^{\otimes 2}) \geq c_0 C$, we obtain that $\lambda_{\min}(\mathrm{pr}_n T X_{\tilde{S}}^{\otimes 2})$ is lower bounded by a positive constant. This implies

$$\left\| \mathrm{pr}_n \left(\frac{T}{\tilde{\pi}} - \frac{T}{\pi} \right) X_{\tilde{S}} \right\|_2^2 \geq C \|\tilde{\gamma} - \gamma^*\|_2^2. \quad \text{(S12)}$$

Following the similar argument, we can show that the right hand side of equation (S11) is bounded above by $2\|\Delta_n\|_2\|A_n\|_2\|\tilde{\gamma}-\gamma^*\|_2$, where

$$\Delta_n = \mathrm{pr}_n\left(\frac{T}{\pi}-1\right)X_{\tilde{S}},\ \ A_n = \mathrm{pr}_n\frac{T\pi'}{\tilde{\pi}\pi}X_{\tilde{S}}^{\otimes 2}.$$

Similarly, we can show that $\|A_n\|_2 \le C\lambda_{\max}(\mathrm{pr}_n X_{\tilde{S}}^{\otimes 2}) \le C'$ for some constants $C, C' > 0$. In addition, we decompose $\Delta_n = I_n + II_n$, where

$$I_n = \mathrm{pr}_n\left(\frac{T}{\pi^*}-1\right)X_{\tilde{S}},\ \ II_n = -\mathrm{pr}_n\frac{T\pi'}{\pi\pi^*}X_{\tilde{S}}X_{\tilde{S}^c}^{\mathrm{T}}(\hat{\beta}-\beta^*)_{\tilde{S}^c},$$

where similarly π' is the derivative of π evaluated at some intermediate value. Thus, by the Bernstein inequality and the union bound argument,

$$\|I_n\|_2 \le |\tilde{S}|^{1/2}\left\|\mathrm{pr}_n\left(\frac{T}{\pi^*}-1\right)X_{\tilde{S}}\right\|_\infty = O_p(\{s\log(s\vee n)/n\}^{1/2}).$$

In addition,

$$\begin{aligned}
\|II_n\|_2 &\le \sup_{\|v\|_2=1}\left|\mathrm{pr}_n\frac{T\pi'}{\pi\pi^*}v^{\mathrm{T}}X_{\tilde{S}}X_{\tilde{S}^c}^{\mathrm{T}}(\hat{\beta}-\beta^*)_{\tilde{S}^c}\right| \\
&\le \sup_{\|v\|_2=1}\left|\mathrm{pr}_n\frac{T\pi'}{\pi\pi^*}(v^{\mathrm{T}}X_{\tilde{S}})^2\right|^{1/2}\left|\mathrm{pr}_n\frac{T\pi'}{\pi\pi^*}\{X_{\tilde{S}^c}^{\mathrm{T}}(\hat{\beta}-\beta^*)_{\tilde{S}^c}\}^2\right|^{1/2} \\
&\le C\lambda_{\max}^{1/2}(\mathrm{pr}_n X_{\tilde{S}}X_{\tilde{S}}^{\mathrm{T}})\cdot\left|\mathrm{pr}_n\{X_{\tilde{S}^c}^{\mathrm{T}}(\hat{\beta}-\beta^*)_{\tilde{S}^c}\}^2\right|^{1/2} = O_p(\{s_1\log(d\vee n)/n\}^{1/2}).
\end{aligned}$$

This implies $\|\Delta_n\|_2 = O_p(\{s\log(d\vee n)/n\}^{1/2})$. Combining with equations (S12) and (S11), we obtain that

$$\|\tilde{\gamma}-\gamma^*\|_2 = O_p(\{(s\log(d\vee n))/n\}^{1/2}),$$

and

$$\|\tilde{\gamma}-\gamma^*\|_1 \le Cs^{1/2}\|\tilde{\gamma}-\gamma^*\|_2 = O_p(\{(s^2\log(d\vee n))/n\}^{1/2}).$$

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

$$\|\tilde{\beta}-\beta^*\|_1 = \|\tilde{\gamma}-\gamma^*\|_1 + \|(\tilde{\beta}-\beta^*)_{\tilde{S}^c}\|_1 = O_p\left(\left\{\frac{s^2\log(d\vee n)}{n}\right\}^{1/2}\right),$$

and

$$\begin{aligned}
\mathrm{pr}_n\{X^{\mathrm{T}}(\tilde{\beta}-\beta^*)\}^2 &\le 2\mathrm{pr}_n\{(\hat{\beta}-\beta^*)_{\tilde{S}^c}^{\mathrm{T}}X_{\tilde{S}^c}\}^2 + 2\mathrm{pr}_n\{(\hat{\gamma}-\gamma^*)^{\mathrm{T}}X_{\tilde{S}}\}^2 \\
&= O_p\left(\frac{s\log(d\vee n)}{n}\right).
\end{aligned}$$

Finally, we start the proof of Theorem 1.

Proof Proof of Theorem 1. By the rearrangement of terms, we have

$$\hat{\mu}_1-\mu_1^* = \mathrm{pr}_n\left[\frac{T}{\pi^*}\{Y(1)-K_1(X)\}+K_1(X)-\mu_1^*\right]+I_1+I_2,$$

where

$$I_1 = \mathrm{pr}_n\left(\frac{T}{\tilde{\pi}} - \frac{T}{\pi^*}\right)\{Y(1) - K_1(X)\}, \;\; I_2 = \mathrm{pr}_n\left(\frac{T}{\tilde{\pi}} - 1\right)K_1(X).$$

By (iii) in the proof of Lemma S4, we have that with probability tending to one $\tilde{\gamma}$ belongs to the interior of Ω. The KKT condition implies $\{\partial g_n(\tilde{\gamma})/\partial\gamma\} g_n(\tilde{\gamma}) = 0$. As seen in the proof of Lemma S4, $\partial g_n(\tilde{\gamma})/\partial\gamma$ is invertible with probability tending to one. Thus, we have $g_n(\tilde{\gamma}) = 0$, and therefore

$$\mathrm{pr}_n\left(\frac{T}{\tilde{\pi}} - 1\right)\tilde{\alpha}_{\tilde{S}}^{\mathrm{T}} X_{\tilde{S}} = \mathrm{pr}_n\left(\frac{T}{\tilde{\pi}} - 1\right)\tilde{\alpha}^{\mathrm{T}} X = 0.$$

Once we can show that

$$\mathrm{pr}_n\left(\frac{T}{\tilde{\pi}} - 1\right)X^{\mathrm{T}}(\tilde{\alpha} - \alpha^*) = o_p(n^{-1/2}), \tag{S13}$$

it yields $I_2 = o_p(n^{-1/2})$. To show equation (S13), by the Taylor theorem,

$$\begin{aligned}
&\left|\mathrm{pr}_n\left(\frac{T}{\tilde{\pi}} - 1\right)X^{\mathrm{T}}(\tilde{\alpha} - \alpha^*)\right| = \left|\mathrm{pr}_n\left(\frac{T}{\pi^*} - 1\right)X^{\mathrm{T}}(\tilde{\alpha} - \alpha^*)\right| + \left|\mathrm{pr}_n\frac{T\pi'(t)}{\pi^{*2}}(\tilde{\beta} - \beta^*)^{\mathrm{T}} X X^{\mathrm{T}}(\tilde{\alpha} - \alpha^*)\right| \\
&\leq \left\|\mathrm{pr}_n\left(\frac{T}{\pi^*} - 1\right)X^{\mathrm{T}}\right\|_\infty \|\tilde{\alpha} - \alpha^*\|_1 + \frac{1}{c_0^2}\left[\mathrm{pr}_n\{X^{\mathrm{T}}(\tilde{\beta} - \beta^*)\}^2\right]^{1/2}\left[\mathrm{pr}_n\{X^{\mathrm{T}}(\tilde{\alpha} - \alpha^*)\}^2\right]^{1/2},
\end{aligned} \tag{S14}$$

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

$$\left\|\mathrm{pr}_n\left(\frac{T}{\pi^*} - 1\right)X^{\mathrm{T}}\right\|_\infty = O_p\left(\left(\frac{\log d}{n}\right)^{1/2}\right).$$

Combining Lemmas S3 and S4 with equation (S14), we obtain that

$$\mathrm{pr}_n\left(\frac{T}{\tilde{\pi}} - 1\right)X^{\mathrm{T}}(\tilde{\alpha} - \alpha^*) = O_p\left(\frac{s\log(d \vee n)}{n} + \frac{s\log(d \vee n)}{n}\right) = o_p(n^{-1/2}),$$

where the last step follows from $\max(s_1, s_2)\log(d \vee n)/n^{1/2} = o(1)$. This completes the proof of equation (S13).

In the following, we will bound I_1 by using the empirical process theory. First, we note that we can show that $|\mathrm{supp}(\hat{\beta})| \leq Cs_1$ by applying the same argument in the proof of Lemma S3. This further implies $|\mathrm{supp}(\tilde{\beta})| \leq Cs$. Furthermore, Lemma S4 implies

$$\|\tilde{\beta} - \beta^*\|_2^2 \leq \mathrm{pr}_n\{X^{\mathrm{T}}(\tilde{\beta} - \beta^*)\}^2 / \inf_S \lambda_{\min}(\mathrm{pr}_n X_S^{\otimes 2}),$$

where S is a subset of $\{1, ..., d\}$ with cardinality no larger than Cs. As shown in the proof of Lemma S4 that $\min_{|S| \leq Cs}\lambda_{\min}(\mathrm{pr}_n X_S^{\otimes 2}) \geq C$, this implies $\|\tilde{\beta} - \beta^*\|_2^2 = O_p(s\log(d \vee n)/n)$. Define the set $\Omega = \{\beta \in \mathbb{R}^d : \|\beta\|_0 \leq Cs, \|\beta - \beta^*\|_2^2 \leq Cs\log(d \vee n)/n\}$. Obviously, $\tilde{\beta}$ be-

longs to Ω with probability tending to 1. Let $\epsilon_1 = Y(1) - K_1(X)$ and $\pi_\beta = \pi(X_i^{\mathrm{T}}\beta)$. Then

$$|I_1| \le \sup_{\beta\in\Omega}\Big|\mathrm{pr}_n \frac{T(\pi_\beta - \pi^*)\epsilon_1}{\pi_\beta\pi^*}\Big| \lesssim n^{-1/2} E\sup_{\beta\in\Omega}\Big|\mathbb{G}_n \frac{T(\pi_\beta - \pi^*)\epsilon_1}{\pi_\beta\pi^*}\Big|, \tag{S15}$$

where $\mathbb{G}_n f = n^{1/2}(\mathrm{pr}_n f - \mathrm{pr} f)$. Let $f_\beta = T(\pi_\beta - \pi^*)\epsilon_1/\pi_\beta\pi^*$. The envelop function is $F = \sup_{\beta\in\Omega}|f_\beta| \le C|\epsilon_1|$ by the lower and upper bounds of π. Since ϵ_1^2 is sub-exponential, the property of the sub-exponential norm implies $\mathrm{pr}\max_i F(T_i, X_i, Y_i)^2 \le C\mathrm{pr}\max_i \epsilon_{1i}^2 \le C\log n$. Similarly, we can show that

$$\begin{aligned}\sup_{\beta\in\Omega}\mathrm{pr} f^2 &\le C\sup_{\beta\in\Omega}\mathrm{pr} T^2\{(\beta-\beta^*)^{\mathrm{T}}X\}^2\epsilon_1^2\\ &\le C\sup_{\beta\in\Omega}\mathrm{pr}\{(\beta-\beta^*)^{\mathrm{T}}X\}^2\\ &\le C\sup_{\beta\in\Omega}\|\beta-\beta^*\|_2^2 \max_{|S|\le Cs}\lambda_{\max}(\Sigma_{SS}) \le C\frac{s\log(d\vee n)}{n}.\end{aligned}$$

Let $\mathcal{F} = \{f_\beta : \beta\in\Omega\}$. In fact when the support of β is fixed, the VC-dimension of $\{X^{\mathrm{T}}(\beta-\beta^*) : \beta\in\mathbb{R}^{Cs}\} = Cs+2$. Thus, $\sup N(\epsilon, \{X^{\mathrm{T}}(\beta-\beta^*) : \beta\in\Omega\}, \|\cdot\|_{Q,2}) \le sd^{Cs}(C/\epsilon)^{Cs}$. By the Lipschitz property of π, we have $\sup N(\epsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \le \sup N(\epsilon, \{X^{\mathrm{T}}(\beta-\beta^*) : \beta\in\Omega\}, \|\cdot\|_{Q,2}) \le sd^{Cs}(C/\epsilon)^{Cs}$. Then, applying the maximal inequality (e.g., Lemma C1 in Belloni et al. (2017)), we get

$$E\sup_{\beta\in\Omega}\Big|\mathbb{G}_n\frac{T(\pi_\beta-\pi^*)\epsilon_1}{\pi_\beta\pi^*}\Big| \lesssim \Big\{\frac{s^2\log d}{n}\log\Big(\frac{dn}{s\log d}\Big)\Big\}^{1/2} + \frac{s(\log n)^{1/2}}{n^{1/2}}\log\Big(\frac{dn}{s\log d}\Big).$$

Plugging into (S15), we have $|I_1| \le Cs\log(d\vee n)/n = o_p(n^{-1/2})$. Thus, we have $\hat\mu_1 - \mu_1^* = n^{-1}\sum_{i=1}^n S_i + \Delta$, where $S_i = T_i/\pi^*\{Y_i(1) - K_1(X_i)\} + K_1(X_i) - \mu_1^*$ and $|\Delta| = o_p(n^{-1/2})$. Following the similar derivation in (Hahn, 1998), it is easy to verify that $T/\pi^*\{Y(1) - K_1(X)\} + K_1(X) - \mu_1^*$ corresponds to the efficient score function for μ_1^*. This implies the semi-parametric efficiency of $\hat\mu_1$. Finally, after some tedious algebra, we can verify that the Lindberg condition holds for S_i under the assumption $E(\alpha^{*\mathrm{T}}X)^4 = O(s_2^2)$, and therefore the central limit theorem holds. This completes the proof.

S4·2. *Proof of Corollary* 1

We first show that

$$|\hat\xi^2 - \xi^2| = O_p\left(\Big\{\frac{s\log(d\vee n)}{n}\Big\}^{1/2}\right), \tag{S16}$$

where

$$\hat\xi^2 = \frac{1}{n}\sum_{i=1}^n \frac{T_i}{\tilde\pi_i^2}(Y_i - \tilde\alpha^{\mathrm{T}}X_i)^2$$

and $\xi = E(\epsilon^2/\pi_i^*)$. Then $|\hat\xi^2 - \xi^2| \le I_1 + I_2$, where

$$I_1 = \left|\frac{1}{n}\sum_{i=1}^n \frac{T_i}{(\pi_i^*)^2}(Y_i - \tilde\alpha^{\mathrm{T}}X_i)^2 - \xi^2\right|, \quad I_2 = \left|\frac{1}{n}\sum_{i=1}^n \frac{T_i\{\tilde\pi_i^2 - (\pi_i^*)^2\}}{(\pi_i^*\tilde\pi_i)^2}(Y_i - \tilde\alpha^{\mathrm{T}}X_i)^2\right|.$$

We can further decompose I_1 as follows

$$I_1 \le \left|\frac{1}{n}\sum_{i=1}^n \frac{T_i}{(\pi_i^*)^2}\epsilon_{1i}^2 - \xi^2\right| + \left|\frac{2}{n}\sum_{i=1}^n \frac{T_i}{(\pi_i^*)^2}\epsilon_{1i}X_i^{\mathrm{T}}(\tilde{\alpha}-\alpha^*)\right| + \left|\frac{1}{n}\sum_{i=1}^n \frac{T_i}{(\pi_i^*)^2}\{X_i^{\mathrm{T}}(\tilde{\alpha}-\alpha^*)\}^2\right| \quad (240)$$

$$\le \left|\frac{1}{n}\sum_{i=1}^n \frac{T_i}{(\pi_i^*)^2}\epsilon_{1i}^2 - \xi^2\right| + \left\|\frac{2}{n}\sum_{i=1}^n \frac{T_i}{(\pi_i^*)^2}\epsilon_{1i}X_i^{\mathrm{T}}\right\|_\infty \|\tilde{\alpha}-\alpha^*\|_1 + \left|\frac{1}{n}\sum_{i=1}^n \frac{1}{c_0}\{X_i^{\mathrm{T}}(\tilde{\alpha}-\alpha^*)\}^2\right|.$$

Applying the Bernstein inequality for the first two terms and Lemma S3, we obtain that $I_1 = O_p(s\log(d\vee n)/n + 1/n^{1/2})$. For I_2, following the similar argument we can show that

$$I_2 \lesssim [\mathrm{pr}_n\{X^{\mathrm{T}}(\tilde{\beta}-\beta^*)\}^2]^{1/2} \times [\mathrm{pr}_n(Y-\alpha^{*\mathrm{T}}X)^4]^{1/2}$$
$$\le C[\mathrm{pr}_n\{X^{\mathrm{T}}(\tilde{\beta}-\beta^*)\}^2]^{1/2} = O_p\left(\Big\{\frac{s\log(d\vee n)}{n}\Big\}\right). \quad (245)$$

The upper bounds for I_1 and I_2 imply that equation (S16) hold.

Now we consider the last term in $\hat{V}-V$, by doing a similar decomposition,

$$\left|\frac{1}{n}\sum_{i=1}^n(\tilde{\alpha}^{\mathrm{T}}X_i-\hat{\mu}_1)^2 - E(\alpha^{*\mathrm{T}}X_i-\mu_1^*)^2\right|$$
$$\le \Big|B_n - EB_n\Big| + 2A_n + 2(\hat{\mu}_1-\mu_1^*)^2 + 2A_n^{1/2}B_n^{1/2} + 2B_n^{1/2}(\hat{\mu}_1-\mu_1^*),$$

where $A_n = n^{-1}\sum_{i=1}^n\{(\tilde{\alpha}-\alpha^*)^{\mathrm{T}}X_i\}^2$ and $B_n = n^{-1}\sum_{i=1}^n(\alpha^{*\mathrm{T}}X_i-\mu_1^*)^2$. Recall that $E(\alpha^{*\mathrm{T}}X_i)^4 = O(s_2^2)$. By Lemma S3 $A_n = O_p(s\log(d\vee n)/n)$. By the Markov inequality and Cauchy inequality, we have $B_n \lesssim E(\alpha^{*\mathrm{T}}X_i)^2 \le \{E(\alpha^{*\mathrm{T}}X_i)^4\}^{1/2} = O_p(s_2)$. Similarly, by Markov inequality $|\hat{\mu}_1-\mu_1^*| \lesssim (V/n)^{1/2} = O_p(s_2^{1/2}/n^{1/2})$. Finally, $|B_n - EB_n| \lesssim \{E(\alpha^{*\mathrm{T}}X_i)^4\}^{1/2}/n^{1/2} = O_p(s_2/n^{1/2})$. Thus

$$\left|\frac{1}{n}\sum_{i=1}^n(\tilde{\alpha}^{\mathrm{T}}X_i-\hat{\mu}_1)^2 - E(\alpha^{*\mathrm{T}}X_i-\mu_1^*)^2\right| = O_p\left((s_1\vee s_2)\Big\{\frac{\log(d\vee n)}{n}\Big\}^{1/2}\right).$$

Together with equations (S16), we complete the proof. (250)

S4·3. *Proof of Proposition* 1

We first prove the asymptotic normality of $\hat{\mu}_1$ when $w_1(u)=1$. Recall that when the propensity score model is misspecified, the estimand of $\hat{\beta}$ in (7) is defined as

$$\beta^o = \arg\max E\Big[\int_0^{\beta^{\mathrm{T}}X_i}\Big\{\frac{T_i}{\pi(u)}-1\Big\}du\Big]. \quad (S17)$$

Since π_i^o is bounded from below by a positive constant, the dominated convergence theorem implies that β^o is the solution of the equation $E\{T_i/\pi(\beta^{\mathrm{T}}X_i)-1\}X_i = 0$. By replicating the proof of Lemma S2, we can show that

$$\|\hat{\beta}-\beta^o\|_1 \;=\; O_p\left(s_1\Big\{\frac{\log(d\vee n)}{n}\Big\}^{1/2}\right),\;\; \mathrm{pr}_n[X^{\mathrm{T}}(\hat{\beta}-\beta^o)]^2 = O_p\left(\frac{s_1\log(d\vee n)}{n}\right).$$

Furthermore, Lemma S3 and S4 imply

$$\|\tilde{\alpha}-\alpha^*\|_1 \;=\; O_p\left((s_1\vee s_2)\Big\{\frac{\log(d\vee n)}{n}\Big\}^{1/2}\right),\;\; \mathrm{pr}_n\{X^{\mathrm{T}}(\tilde{\alpha}-\alpha^*)\}^2 = O_p\left(\frac{(s_1\vee s_2)\log(d\vee n)}{n}\right),$$

$$\|\tilde{\beta}-\beta^o\|_1 \;=\; O_p\left(\Big\{\frac{s^2\log(d\vee n)}{n}\Big\}^{1/2}\right),\;\; \mathrm{pr}_n\{X^{\mathrm{T}}(\tilde{\beta}-\beta^o)\}^2 \;=\; O_p\left(\frac{s\log(d\vee n)}{n}\right),$$

where $s = s_1 \vee s_2$. Similar to that of Theorem 1, we have

$$\hat{\mu}_1-\mu_1^* \;=\; \mathrm{pr}_n\left[\frac{T}{\pi^o}\{Y(1)-K_1(X)\}+K_1(X)-\mu_1^*\right]+I_1+I_2,$$

where

$$I_1 \;=\; \mathrm{pr}_n\left(\frac{T}{\tilde{\pi}}-\frac{T}{\pi^o}\right)\{Y(1)-K_1(X)\},\;\; I_2 \;=\; \mathrm{pr}_n\left(\frac{T}{\tilde{\pi}}-1\right)K_1(X).$$

The first term I_1 can be bounded by applying the same maximal inequalities for empirical processes. The same rate $|I_1| \le Cs\log(d\vee n)/n = o_p(n^{-1/2})$ can be obtained. The difficulty comes from the second term I_2. First, the covariate balancing equation in step 3 forces $\mathrm{pr}_n(T/\tilde{\pi}-1)X^{\mathrm{T}}\tilde{\alpha}=0$, which removes the bias induced by plugging in the estimator $\tilde{\beta}$ and $\tilde{\alpha}$. Second, for I_2, following our previous proof, we have

$$\begin{aligned}
|I_2| &= \left|\mathrm{pr}_n\left(\frac{T}{\tilde{\pi}}-1\right)X^{\mathrm{T}}(\tilde{\alpha}-\alpha^*)\right| \\
&\le \left|\mathrm{pr}_n\left(\frac{T}{\pi^o}-1\right)X^{\mathrm{T}}(\tilde{\alpha}-\alpha^*)\right| + C\left|\mathrm{pr}_n(\tilde{\beta}-\beta^*)^{\mathrm{T}}XX^{\mathrm{T}}(\tilde{\alpha}-\alpha^*)\right| \\
&\le \left\|\mathrm{pr}_n\left(\frac{T}{\pi^o}-1\right)X^{\mathrm{T}}\right\|_\infty \|\tilde{\alpha}-\alpha^*\|_1 + C\left[\mathrm{pr}_n\{X^{\mathrm{T}}(\tilde{\beta}-\beta^*)\}^2\right]^{1/2}\left[\mathrm{pr}_n\{X^{\mathrm{T}}(\tilde{\alpha}-\alpha^*)\}^2\right]^{1/2}.
\end{aligned}$$

The last term has the fast convergence rate $O_p(s\log(d\vee n)/n)$. For the first term, we have $E\{(T/\pi^o-1)X\}=0$. Thus, we can apply the Bernstein equality to show $\|\mathrm{pr}_n(T/\pi^o-1)X\|_\infty = O_p((\log d/n)^{1/2})$. Together with the convergence rate of $\tilde{\alpha}$, we obtain $|I_2| \le Cs\log(d\vee n)/n = o_p(n^{-1/2})$. The asymptotic normality of $\hat{\mu}_1$ follows directly from the central limit theorem. We note that in general $\|E(T/\pi^o-1)X\|_\infty$ can be viewed as a bias term, which typically has the order of $O_p(1)$. We get a sharper rate because $E\{(T/\pi^o-1)X\}=0$, which benefits from the definition of the least false parameter in (S17). For instance, if we choose $w_1(u)\neq 1$ in (S17), the bias term has the order of $\|\mathrm{pr}_n(T/\pi^o-1)X\|_\infty = O_p(1)$. This leads to a slower rate $|I_2| \le Cs\{\log(d\vee n)/n\}^{1/2}$. Indeed, we can get a sharper bound for I_2 by applying the Cauchy inequality

$$\left|\mathrm{pr}_n\left(\frac{T}{\pi^o}-1\right)X^{\mathrm{T}}(\tilde{\alpha}-\alpha^*)\right| \le \left|\mathrm{pr}_n\left(\frac{T}{\pi^o}-1\right)^2\right|^{1/2}|\mathrm{pr}_n\{X^{\mathrm{T}}(\tilde{\alpha}-\alpha^*)\}^2|^{1/2} = O_p\left(\Big\{\frac{s\log(d\vee n)}{n}\Big\}^{1/2}\right).$$

Thus $\hat{\mu}_1-\mu_1^* = O_p(\{s\log(d\vee n)/n\}^{1/2})$ for $w_1(u)\neq 1$.

S4·4. *Proof of Proposition* 2

Recall that when the outcome model is misspecified, the least false parameter is defined as

$$\alpha^o = \arg\min E\Big\{T_i w_2(\beta^{*\mathrm{T}}X_i)(Y_i-\alpha^{\mathrm{T}}X_i)^2\Big\}.$$

Again, the dominated convergence theorem implies $E\{T_i w_2(\beta^{*\mathrm{T}}X_i)(Y_i-\alpha^{o\mathrm{T}}X_i)X_i\}=0$. Similarly, by the proof of Lemma S3, we have

$$\|\tilde{\alpha}-\alpha^o\|_1 \;=\; O_p\left((s_1\vee s_2)\Big\{\frac{\log(d\vee n)}{n}\Big\}^{1/2}\right),\;\; \mathrm{pr}_n[X^{\mathrm{T}}(\tilde{\alpha}-\alpha^o)]^2 = O_p\left(\frac{(s_1\vee s_2)\log(d\vee n)}{n}\right).$$

To prove this proposition, we note the similar decomposition holds

$$\hat{\mu}_1 - \mu_1^* \;=\; \mathrm{pr}_n\left[\frac{T}{\pi^*}\{Y(1) - \alpha^{o\mathrm{T}}X\} + \alpha^{o\mathrm{T}}X_i - \mu_1^*\right] + I_1 + I_2,$$

where

$$I_1 \;=\; \mathrm{pr}_n\left(\frac{T}{\tilde{\pi}} - \frac{T}{\pi^*}\right)\{Y(1) - \alpha^{o\mathrm{T}}X\},\;\; I_2 \;=\; \mathrm{pr}_n\left(\frac{T}{\tilde{\pi}} - 1\right)\alpha^{o\mathrm{T}}X.$$

By adding the equations $\mathrm{pr}_n(T/\tilde{\pi} - 1)X^{\mathrm{T}}\tilde{\alpha} = 0$ to I_2, we can apply the same steps in the proof of Proposition 1 and note that the propensity score is correctly specified in this case. This gives us $|I_2| \le Cs\log(d \vee n)/n = o_p(n^{-1/2})$, which holds regardless of the choice of the weight functions $w_1(u)$ and $w_2(u)$. In this case, the difficulty comes from the first term I_1. By the Taylor expansion,

$$|I_1| = \left|\mathrm{pr}_n\frac{T\pi'}{\pi^*\tilde{\pi}}\epsilon^o X^{\mathrm{T}}(\tilde{\beta} - \beta^*)\right| \le \left\|\mathrm{pr}_n\frac{T\pi'}{\pi^*\tilde{\pi}}\epsilon^o X\right\|_\infty \|\tilde{\beta} - \beta^*\|_1,$$

where π' is the derivative of $\pi(u)$ evaluated at some intermediate value between $X_i^{\mathrm{T}}\beta^*$ and $X_i^{\mathrm{T}}\tilde{\beta}$ and $\epsilon^o = Y(1) - \alpha^{o\mathrm{T}}X$. Since $E(\epsilon^o|X) \neq 0$ in general, we no longer have $E\{T(\pi^*)'/(\pi^*)^2\epsilon^o X\} = 0$ as when the outcome model is correctly specified. In this sense, we can treat $E\{T(\pi^*)'/(\pi^*)^2\epsilon^o X\}$ as the bias term under the misspecified outcome model. So, in general, we can bound I_1 as follows

$$\begin{aligned}|I_1| = \left|\mathrm{pr}_n\frac{T\pi'}{\pi^*\tilde{\pi}}\epsilon^o X^{\mathrm{T}}(\tilde{\beta} - \beta^*)\right| &\le \left|\mathrm{pr}_n\frac{(T\pi')^2}{(\pi^*\tilde{\pi})^2}\epsilon^{o2}\right|^{1/2}[\mathrm{pr}_n\{X^{\mathrm{T}}(\tilde{\beta} - \beta^*)\}^2]^{1/2}\\ &\le C|\mathrm{pr}_n T\epsilon^{o2}|^{1/2}[\mathrm{pr}_n\{X^{\mathrm{T}}(\tilde{\beta} - \beta^*)\}^2]^{1/2} = O_p(\{s\log(d \vee n)/n\}^{1/2}).\end{aligned}$$

Thus, we prove that $\hat{\mu}_1 - \mu_1^* = O_p(\{s\log(d \vee n)/n\}^{1/2})$. Finally, we consider the case that $w_2(u) = \pi'(u)/\pi^2(u)$. With this particular choice of w_2, we have $E\{T(\pi^*)'/(\pi^*)^2\epsilon^o X\} = 0$, i.e., the bias term becomes 0. In the following, we will prove it formally $|I_1| \le Cs\log(d \vee n)/n = o_p(n^{-1/2})$ when $w_2(u) = \pi'(u)/\pi^2(u)$. Once it is proved, the asymptotic normality of $\hat{\mu}_1$ follows directly.

By adding and subtracting $\mathrm{pr}_n\{T(\pi^*)'/(\pi^*)^2\epsilon^o X^{\mathrm{T}}(\tilde{\beta} - \beta^*)\}$, we have

$$I_1 = \mathrm{pr}_n\Big\{\frac{\tilde{\pi} - \pi^*}{\pi^*\tilde{\pi}} - \frac{(\pi^*)'X^{\mathrm{T}}(\tilde{\beta} - \beta^*)}{(\pi^*)^2}\Big\}T\epsilon^o + \mathrm{pr}_n\frac{(\pi^*)'X^{\mathrm{T}}(\tilde{\beta} - \beta^*)}{(\pi^*)^2}T\epsilon^o := I_{11} + I_{12}.$$

For I_{12},

$$|I_{12}| \le \|\tilde{\beta} - \beta^*\|_1\Big\|\mathrm{pr}_n\frac{(\pi^*)'X}{(\pi^*)^2}T\epsilon^o\Big\|_\infty = O_p\Big(\frac{s\log(d \vee n)}{n}\Big), \tag{S18}$$

where we plug in the rate of convergence of $\tilde{\beta}$ and apply the Bernstein inequality and union bounds for $(\pi^*)'X_j/(\pi^*)^2T\epsilon^o$, since $|(\pi^*)'X_j/(\pi^*)^2T\epsilon^o| \le C|X_j\epsilon^o|$ is sub-exponential. Now, we consider the term I_{11}. Similarly, we define the set $\Omega = \{\beta \in \mathbb{R}^d : \|\beta\|_0 \le Cs, \|\beta - \beta^*\|_2^2 \le Cs\log(d \vee n)/n\}$, so that $\tilde{\beta}$ belongs to Ω with probability tending to 1. Let $\pi_\beta = \pi(X_i^{\mathrm{T}}\beta)$. Then

$$|I_{11}| \le \sup_{\beta\in\Omega}|\mathrm{pr}_n f_\beta| \le \sup_{\beta\in\Omega}|(\mathrm{pr}_n - \mathrm{pr})f_\beta| + \sup_{\beta\in\Omega}|\mathrm{pr} f_\beta|, \tag{S19}$$

where

$$f_\beta = \Big\{ \frac{\pi_\beta - \pi^*}{\pi^* \pi_\beta} - \frac{(\pi^*)' X^{\mathrm{T}} (\beta - \beta^*)}{(\pi^*)^2} \Big\} T \epsilon^o.$$

We use the maximal inequality to bound the first term. For any $\beta \in \Omega$, we have $\|\beta - \beta^*\|_1 \leq |\mathrm{supp}(\beta - \beta^*)|^{1/2} \|\beta - \beta^*\|_2 \leq Cs\{\log(d \vee n)/n\}^{1/2}$. Thus, with probability tending to 1, $|X^{\mathrm{T}}(\beta - \beta^*)| \leq 1$, and $|f_\beta| \leq C|\epsilon^o|$. We take the envelop function as $F = C|\epsilon^o|$. Similarly, $\mathrm{pr} \max_i F(T_i, X_i, Y_i)^2 \leq C \mathrm{pr} \max_i \epsilon_{1i}^2 \leq C \log n$. We now look at $\sup_{\beta \in \Omega} \mathrm{pr} f_\beta^2$. Note that applying the mean-value theorem, we have

$$\begin{aligned} \sup_{\beta \in \Omega} \mathrm{pr} f_\beta^2 &= \sup_{\beta \in \Omega} \mathrm{pr} T^2 \epsilon^{o2} \{(\beta - \beta^*)^{\mathrm{T}} X\}^2 \Big(\frac{\bar{\pi}'}{\pi_\beta \pi^*} - \frac{(\pi^*)'}{(\pi^*)^2} \Big)^2 \\ &\leq C \sup_{\beta \in \Omega} \mathrm{pr} \{(\beta - \beta^*)^{\mathrm{T}} X\}^2 \leq C \frac{s \log(d \vee n)}{n}, \end{aligned}$$

where $\bar{\pi}' = \pi' \{ t X^{\mathrm{T}} \beta^o + (1-t) X^{\mathrm{T}} \beta \}$ for some $t \in (0,1)$ which is bounded by Assumption 6. Let $\mathcal{F} = \{ f_\beta : \beta \in \Omega \}$. The uniform covering number of $\mathcal{F}$ satisfies $\sup N(\epsilon \|F\|_{Q,2}, \mathcal{F}, \| \cdot \|_{Q,2}) \leq s d^s (C/\epsilon)^{Cs}$. Then, applying the maximal inequality (e.g., Lemma C1 in Belloni et al. (2017)), we get

$$E \sup_{\beta \in \Omega} \Big| \mathbb{G}_n f_\beta \Big| \lesssim \Big\{ \frac{s^2 \log d}{n} \log \Big(\frac{dn}{s \log d} \Big) \Big\}^{1/2} + \frac{s (\log n)^{1/2}}{n^{1/2}} \log \Big(\frac{dn}{s \log d} \Big).$$

$$\sup_{\beta \in \Omega} |(\mathrm{pr}_n - \mathrm{pr}) f_\beta| \lesssim n^{-1/2} E \sup_{\beta \in \Omega} \Big| \mathbb{G}_n f_\beta \Big| = O_p \Big(\frac{s \log(d \vee n)}{n} \Big).$$

Next, we will bound the second term in (S19). To get a sharp bound, we apply the same mean-value theorem and the Cauchy inequality,

$$\begin{aligned} \sup_{\beta \in \Omega} |\mathrm{pr} f_\beta| &\leq \sup_{\beta \in \Omega} \Big\{ \mathrm{pr} \frac{T^2 \epsilon^{o2}}{(\pi^*)^2} \Big(\frac{\bar{\pi}'}{\pi_\beta} - \frac{(\pi^*)'}{\pi^*} \Big)^2 \Big\}^{1/2} [\mathrm{pr} \{(\beta - \beta^*)^{\mathrm{T}} X\}^2]^{1/2} \\ &\leq C \sup_{\beta \in \Omega} \{ \mathrm{pr} (\bar{\pi}' \pi^* - \pi_\beta (\pi^*)')^2 \}^{1/2} [\mathrm{pr} \{(\beta - \beta^*)^{\mathrm{T}} X\}^2]^{1/2} \\ &\leq C \frac{s \log(d \vee n)}{n}, \end{aligned}$$

where we apply the standard perturbation analysis to the last term, i.e.,

$$\begin{aligned} \{ \bar{\pi}' \pi^* - \pi_\beta (\pi^*)' \}^2 &= [\{ \bar{\pi}' - (\pi^*)' \} \pi^* + (\pi^* - \pi_\beta)(\pi^*)']^2 \\ &\leq 2 \{ \bar{\pi}' - (\pi^*)' \}^2 (\pi^*)^2 + 2 (\pi^* - \pi_\beta)^2 \{ (\pi^*)' \}^2 \\ &\leq 2C [X^{\mathrm{T}} (\beta - \beta^*)^2 (\pi^*)^2 + X^{\mathrm{T}} (\beta - \beta^*)^2 \{ (\pi^*)' \}^2], \end{aligned}$$

where the last step follows from the Lipschitz condition in Assumption 6. Plugging into (S19), we have $|I_{11}| \leq Cs \log(d \vee n)/n = o_p(n^{-1/2})$. Combining with (S18), we finally prove that $|I_1| \leq Cs \log(d \vee n)/n = o_p(n^{-1/2})$. This completes the proof.

S4·5. *Proof of Theorem* 2

Without loss of generality, we let $a(\phi) = 1$. Similar to the previous appendix, we make the modifications on the estimation of γ in step 3. Specifically, let $\tilde{\gamma} = \arg\min_{\gamma \in \Omega} \|g_n(\gamma)\|_2^2$, where

$\Omega = \{\gamma : \|\gamma - \hat{\beta}_{\tilde{S}}\|_1 \leq \delta/\log n\}$ for some small constant $\delta > 0$. We first prove part (1) in Theorem 2.

LEMMA S5. *Under the assumptions in Theorem 2,*

$$\|\hat{\alpha} - \alpha^*\|_1 = O_p\left(C_n s_2 \Big\{\frac{\log(d \vee n)}{n}\Big\}^{1/2}\right), \ pr_n[X^{\mathrm{T}}(\hat{\alpha} - \alpha^*)]^2 = O_p\left(\frac{C_n^2 s_2 \log(d \vee n)}{n}\right).$$

The proof of this lemma is similar to Lemma S2. The main difference is to show the following concentration inequality

$$\mathrm{pr}\Big(\|\mathrm{pr}_n T X_i \epsilon_1\|_\infty \geq C C_n \Big\{\frac{\log(d \vee n)}{n}\Big\}^{1/2}\Big) \leq \frac{1}{d \vee n}.$$

This is true by noting that $\|T X_{ij}\epsilon_i\|_{\psi_1} \leq C C_n$ and we can apply the Bernstein inequality and the union bound argument. Recall that we denote $s = s_1 \vee s_2$.

LEMMA S6. *Under the assumptions in Theorem 2,*

$$\|\hat{\beta} - \beta^*\|_1 \ = \ O_p\left(C_n s \Big\{\frac{\log(d \vee n)}{n}\Big\}^{1/2}\right), \ pr_n[X^{\mathrm{T}}(\hat{\beta} - \beta^*)]^2 = O_p\left(\frac{C_n^2 s \log(d \vee n)}{n}\right).$$

The proof is similar to Lemma S2. We only highlight the differences. Denote

$$Q_n(\beta) = \frac{1}{n}\sum_{i=1}^n \int_0^{\beta^{\mathrm{T}} X_i} \Big\{\frac{T_i}{\pi(u)} - 1\Big\} w_1(\alpha^{*\mathrm{T}} X_i, u) du,$$

$$\hat{Q}_n(\beta) = \frac{1}{n}\sum_{i=1}^n \int_0^{\beta^{\mathrm{T}} X_i} \Big\{\frac{T_i}{\pi(u)} - 1\Big\} w_1(\hat{\alpha}^{\mathrm{T}} X_i, u) du,$$

As shown in in Lemma S2, we have

$$\hat{D}(\bar{\beta}, \beta^*) + \lambda\|\bar{\beta}\|_1 \leq \nabla\hat{Q}_n(\beta^*)(\bar{\beta} - \beta^*) + \lambda\|\beta^*\|_1$$

where $\hat{D}(\bar{\beta}, \beta^*) = \hat{Q}_n(\beta^*) - \hat{Q}_n(\bar{\beta}) + \nabla\hat{Q}_n(\beta^*)(\bar{\beta} - \beta^*)$. Note that

$$\begin{aligned}
\nabla\hat{Q}_n(\beta^*)(\bar{\beta} - \beta^*) &= \nabla Q_n(\beta^*)(\bar{\beta} - \beta^*) + [\{\nabla\hat{Q}_n(\beta^*) - \nabla Q_n(\beta^*)\}(\bar{\beta} - \beta^*)] \\
&\leq \nabla Q_n(\beta^*)(\bar{\beta} - \beta^*) + C C_n \Big\{\frac{s_2 \log(d \vee n)}{n}\Big\}^{1/2} \delta_n \\
&\leq B\lambda/3 + C C_n \Big\{\frac{s_2 \log(d \vee n)}{n}\Big\}^{1/2} \delta_n,
\end{aligned}$$

where $\delta_n^2 = n^{-1}\sum_{i=1}^n \{X_i^{\mathrm{T}}(\bar{\beta} - \beta^*)\}^2$. In the second step, we use the Lipschitz property and the Cauchy inequality. The last step holds under the event $\|\nabla Q_n(\beta^*)(\bar{\beta} - \beta^*)\|_\infty \leq \lambda/3$. Again, by choosing $\lambda = C C_n\{\log(d \vee n)/n\}^{1/2}$, the Bernstein inequality implies that this event holds with high probability. In addition, by Assumption 9 after some algebra, we can show

$$\begin{aligned}
\hat{D}(\bar{\beta}, \beta^*) &\geq C\delta_n^2 - \frac{1}{n}\sum_{i=1}^n (T_i - \pi_i^*)\{X_i^{\mathrm{T}}(\bar{\beta} - \beta^*)\}^2 \\
&\geq C\delta_n^2 - C C_n \|\bar{\Delta}\|_1^2 \Big\{\frac{\log(d \vee n)}{n}\Big\}^{1/2},
\end{aligned}$$

with high probability. If $\|\bar{\Delta}\|_1 > B/2$, then

$$C\delta_n^2 + \frac{1}{3}\lambda\|\bar{\Delta}_{S^c}\|_1 \le \frac{5}{3}\lambda\|\bar{\Delta}_S\|_1 + CC_n\Big\{\frac{s_2\log(d\vee n)}{n}\Big\}^{1/2}\delta_n + CC_n\|\bar{\Delta}\|_1^2\Big\{\frac{\log(d\vee n)}{n}\Big\}^{1/2}.$$

Since $\|\bar{\Delta}\|_1 \le B \lesssim 1$, this implies

$$C\delta_n^2 + \frac{1}{6}\lambda\|\bar{\Delta}_{S^c}\|_1 \le \frac{11}{6}\lambda\|\bar{\Delta}_S\|_1 + CC_n\Big\{\frac{s_2\log(d\vee n)}{n}\Big\}^{1/2}\delta_n.$$

We want to first show $\delta_n \lesssim s^{1/2}\lambda$. Otherwise, $C\delta_n^2 - CC_n\{s_2\log(d\vee n)/n\}^{1/2}\delta_n > 0$, and then the cone condition $\|\bar{\Delta}_{S^c}\|_1 \le 11\|\bar{\Delta}_S\|_1$ holds. By the compatibility factor condition, we have $\|\bar{\Delta}_S\|_1 \lesssim s_1^{1/2}\delta_n$. Thus, $C\delta_n^2 \lesssim s_1^{1/2}\lambda\delta_n + C_n\{s_2\log(d\vee n)/n\}^{1/2}\delta_n$. This proves $\delta_n \lesssim s^{1/2}\lambda$. We further show that $\|\bar{\Delta}\|_1 \lesssim s\lambda$. For some constant t to be chosen later, if $\|\bar{\Delta}_{S^c}\|_1 \le t\|\bar{\Delta}_S\|_1$ holds, $\|\bar{\Delta}\|_1 \le (1+t)\|\bar{\Delta}_S\|_1 \le (1+t)s_1^{1/2}\delta_n \lesssim s\lambda$. However if $\|\bar{\Delta}_{S^c}\|_1 > t\|\bar{\Delta}_S\|_1$ holds, then $\frac{t-11}{6}\lambda\|\bar{\Delta}_S\|_1 \le CC_n\{s_2\log(d\vee n)/n\}^{1/2}\delta_n$. Choosing $t = 12$, it yields $\|\bar{\Delta}_S\|_1 \le 6Cs_2^{1/2}\delta_n \lesssim s\lambda$. Similarly, we can show that $\|\bar{\Delta}_{S^c}\|_1 \lesssim s\lambda$ and therefore $\|\bar{\Delta}\|_1 \lesssim s\lambda$. By choosing $B = Cs\lambda$ for some C large enough, we obtain that $\|\bar{\Delta}\|_1 \le B/2$. The remaining proof follows the same argument as in the proof of Lemma S2 and is omitted.

LEMMA S7. *Under the assumptions in Theorem* 2*,*

$$\|\tilde{\alpha} - \alpha^*\|_1 = O_p\left(C_n s\Big\{\frac{\log(d\vee n)}{n}\Big\}^{1/2}\right),\ \ pr_n\{X^{\mathrm{T}}(\tilde{\alpha}-\alpha^*)\}^2 = O_p\left(\frac{C_n^2 s\log(d\vee n)}{n}\right).$$

Similar to Lemma S3, the proof essentially follows from Lemma E.1 of Ning & Liu (2017). We omit the details.

LEMMA S8. *Under the assumptions in Theorem* 2*,*

$$\|\tilde{\beta} - \beta^*\|_1 = O_p\left(C_n s\Big\{\frac{\log(d\vee n)}{n}\Big\}^{1/2}\right),\ \ pr_n\{X^{\mathrm{T}}(\tilde{\beta}-\beta^*)\}^2 = O_p\left(\frac{C_n^2 s\log(d\vee n)}{n}\right).$$

The proof is similar to the proof of Lemma S4, we omit the details.

Proof Proof of Theorem 2. Like the proof of Theorem 1, we start with the decomposition:

$$\hat{\mu}_1 - \mu_1^* = \mathrm{pr}_n\left[\frac{T}{\pi^*}\{Y(1) - K_1(X)\} + K_1(X) - \mu_1^*\right] + I_1 + I_2, \tag{S20}$$

where

$$I_1 = \mathrm{pr}_n\left(\frac{T}{\tilde{\pi}} - \frac{T}{\pi^*}\right)\{Y(1) - K_1(X)\},\ \ I_2 = \mathrm{pr}_n\left(\frac{T}{\tilde{\pi}} - 1\right)\{K_1(X) - \hat{K}_1(X)\}.$$

We first consider I_2. Recall that $K_1(X) = b'(\alpha^{*\mathrm{T}}X), \hat{K}_1(X) = b'(\tilde{\alpha}^{\mathrm{T}}X)$. Note that

$$\begin{aligned} I_2 &= \mathrm{pr}_n\left(\frac{T}{\tilde{\pi}} - 1\right)\{b'(\alpha^{*T}X) - b'(\tilde{\alpha}^{\mathrm{T}}X) - b''(\tilde{\alpha}^{\mathrm{T}}X)(\alpha^* - \tilde{\alpha})_{\tilde{S}}^{\mathrm{T}}X_{\tilde{S}}\} \\ &= \mathrm{pr}_n\left(\frac{T}{\tilde{\pi}} - 1\right)[b''(\tilde{\alpha}^{\mathrm{T}}X)(\alpha^* - \tilde{\alpha})_{\tilde{S}^c}^{\mathrm{T}}X_{\tilde{S}^c} + b'''(t)\{(\alpha^* - \tilde{\alpha})^{\mathrm{T}}X\}^2], \end{aligned}$$

where t is an intermediate value between $\alpha^{*\mathrm{T}}X$ and $\tilde{\alpha}^{\mathrm{T}}X$. We denote

$$I_{21} = \mathrm{pr}_n\left(\frac{T}{\tilde{\pi}} - 1\right)b''(\tilde{\alpha}^{\mathrm{T}}X)(\alpha^* - \tilde{\alpha})_{\tilde{S}^c}^{\mathrm{T}}X_{\tilde{S}^c},\ \ I_{22} = \mathrm{pr}_n\left(\frac{T}{\tilde{\pi}} - 1\right)b'''(t)\{(\alpha^* - \tilde{\alpha})^{\mathrm{T}}X\}^2.$$

For I_{21}, we further apply the multivariate Taylor theorem to expand $\tilde{\pi}$ and $b''(\tilde{\alpha}^{\mathrm{T}}X)$,

$$I_{21} = \mathrm{pr}_n \left(\frac{T}{\pi^*} - 1 \right) b''(\alpha^{*T}X)(\alpha^* - \tilde{\alpha})^{\mathrm{T}}_{\tilde{S}^c} X_{\tilde{S}^c} - \mathrm{pr}_n \frac{T\pi'(t_1)}{\pi^{*2}} b''(t_2)(\tilde{\beta} - \beta^*)^{\mathrm{T}} \bar{X} (\alpha^* - \tilde{\alpha})^{\mathrm{T}}_{\tilde{S}^c} X_{\tilde{S}^c}$$
$$+ \mathrm{pr}_n \left\{ \frac{T}{\pi(t_1)} - 1 \right\} b'''(t_2)(\tilde{\alpha} - \alpha^*)^{\mathrm{T}} X (\alpha^* - \tilde{\alpha})^{\mathrm{T}}_{\tilde{S}^c} X_{\tilde{S}^c}, \tag{S21}$$

where t_1 and t_2 are the intermediate values between $\beta^{*T}\bar{X}$ and $\tilde{\beta}^{\mathrm{T}}\bar{X}$, and between $\alpha^{*T}X$ and $\tilde{\alpha}^{\mathrm{T}}X$. For the first term in equation (S21), we have $|T/\pi^* - 1| \leq 1/c_0$, and $b''(\alpha^{*T}X)$ and X_j are both bounded random variables, we can apply the Hoeffding inequality and the union bound argument, which gives us

$$\left\| \mathrm{pr}_n \left(\frac{T}{\pi^*} - 1 \right) b''(\alpha^{*T}X) X_{\tilde{S}^c} \right\|_\infty = O_p \left(C_n \left\{ \frac{\log(d \vee n)}{n} \right\}^{1/2} \right).$$

Together with Lemma S7, we have

$$\left| \mathrm{pr}_n \left(\frac{T}{\pi^*} - 1 \right) b''(\alpha^{*T}X)(\alpha^* - \tilde{\alpha})^{\mathrm{T}}_{\tilde{S}^c} X_{\tilde{S}^c} \right| = O_p \left(\frac{C_n^2 s \log(d \vee n)}{n} \right).$$

Similar to the derivation in equation (S14), for the second term in equation (S21), by Lemma S8,

$$\left| \mathrm{pr}_n \frac{T\pi'(t_1)}{\pi^{*2}} b''(t_2)(\tilde{\beta} - \beta^*)^{\mathrm{T}} \bar{X} (\alpha^* - \tilde{\alpha})^{\mathrm{T}}_{\tilde{S}^c} X_{\tilde{S}^c} \right| = O_p \left(\frac{C_n^2 s \log(d \vee n)}{n} \right).$$

For the last term in equation (S21), first we note that $b'''(\cdot)$ is continuous by assumption, and $|b'''(t_2)|$ is bounded. In addition, $\pi(t_1)$ is also bounded away from 0. We apply the Cauchy inequality,

$$\left| \mathrm{pr}_n \left\{ \frac{T}{\pi(t_1)} - 1 \right\} b'''(t_2)(\tilde{\alpha} - \alpha^*)^{\mathrm{T}} X (\alpha^* - \tilde{\alpha})^{\mathrm{T}}_{\tilde{S}^c} X_{\tilde{S}^c} \right|$$
$$\leq \left[\mathrm{pr}_n \left\{ \frac{T}{\pi(t_1)} - 1 \right\}^2 b'''(t_2)^2 \{(\tilde{\alpha} - \alpha^*)^{\mathrm{T}} X\}^2 \right]^{1/2} \left[\mathrm{pr}_n \{(\alpha^* - \tilde{\alpha})^{\mathrm{T}}_{\tilde{S}^c} X_{\tilde{S}^c}\}^2 \right]^{1/2}$$
$$= O_p \left(\frac{C_n^2 s \log(d \vee n)}{n} \right).$$

Combining these results with equation (S21), we obtain

$$|I_{21}| = O_p \left(\frac{C_n^2 s \log(d \vee n)}{n} \right).$$

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

$$|I_{22}| \leq C \mathrm{pr}_n \{(\alpha^* - \tilde{\alpha})^{\mathrm{T}} X\}^2 = O_p \left(\frac{C_n^2 s \log(d \vee n)}{n} \right). \tag{S22}$$

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

$$|I_2| = O_p \left(\frac{C_n^2 s \log(d \vee n)}{n} \right).$$

For I_1, recall that $\epsilon_1 = Y(1) - K_1(X)$ is sub-exponential. Applying the same empirical process argument as in the proof of Theorem 1, we can show that $|I_1| \lesssim C_n^2 s \log(d \vee n)/n$. Thus, we obtain the part (1) in this theorem.

Next, we will prove part (2). Since the outcome model is correct, Lemma S5 holds. The same argument in Lemma S6 implies

$$\|\hat{\beta}-\beta^o\|_1 \;=\; O_p\left(C_n s\Big\{\frac{\log(d\vee n)}{n}\Big\}^{1/2}\right),\;\; \mathrm{pr}_n\{X^{\mathrm{T}}(\hat{\beta}-\beta^o)\}^2 = O_p\left(\frac{C_n^2 s\log(d\vee n)}{n}\right). \tag{S23}$$

Given the above convergence rate of $\hat{\beta}$, Lemma S7 holds regardless of the choice of $w_2(u)$. The rate in (S23) also holds for $\tilde{\beta}$ as in Lemma S8. With π^* replaced by π^o, we have the similar decomposition in (S20), where

$$I_1 = \mathrm{pr}_n\left(\frac{T}{\tilde{\pi}}-\frac{T}{\pi^o}\right)\{Y(1)-K_1(X)\},\;\; I_2 = \mathrm{pr}_n\left(\frac{T}{\tilde{\pi}}-1\right)\{K_1(X)-\hat{K}_1(X)\}.$$

The bound for I_1 is the same as before. For I_2, it can be further decomposed as the sum of I_{21} and I_{22}. As shown in (S21), the second and third term of I_{21} is the same and the first term is controlled by the Hoeffding inequality and the union bound argument, i.e.,

$$\left\|\mathrm{pr}_n\left(\frac{T}{\pi^o}-1\right)b''(\alpha^{*T}X)X_{\tilde{S}^c}\right\|_\infty = O_p\left(C_n\Big\{\frac{\log(d\vee n)}{n}\Big\}^{1/2}\right),$$

where we use the fact that β^o is defined as $E\{(T/\pi^o-1)b''(\alpha^{*T}X)X\}=0$ when we choose $w_1(u,v)=b''(u)$. The same bound (S22) holds for I_{22}. To sum things up, we have $|I_2|\lesssim C_n^2 s\log(d\vee n)/n$. This implies part (2) in this theorem.

Finally, we will prove the part (3). Given the decomposition in (S20), we have $|I_2|\lesssim C_n^2 s\log(d\vee n)/n$ by applying the same argument, which holds regardless of the choice of $w_1(u,v)$ and $w_2(u)$. We now consider I_1. Let $\epsilon^o = Y-b'(X^{\mathrm{T}}\alpha^o)$. When $w_2(u)=\pi'(u)/\pi^2(u)$, by adding and subtracting $\mathrm{pr}_n\{T(\pi^*)'/(\pi^*)^2\epsilon^o X^{\mathrm{T}}(\tilde{\beta}-\beta^*)\}$, we have

$$I_1 = \mathrm{pr}_n\Big\{\frac{\tilde{\pi}-\pi^*}{\pi^*\tilde{\pi}}-\frac{(\pi^*)'X^{\mathrm{T}}(\tilde{\beta}-\beta^*)}{(\pi^*)^2}\Big\}T\epsilon^o + \mathrm{pr}_n\frac{(\pi^*)'X^{\mathrm{T}}(\tilde{\beta}-\beta^*)}{(\pi^*)^2}T\epsilon^o := I_{11}+I_{12}.$$

For I_{12},

$$|I_{12}|\le\|\tilde{\beta}-\beta^*\|_1\Big\|\mathrm{pr}_n\frac{(\pi^*)'X}{(\pi^*)^2}T\epsilon^o\Big\|_\infty = O_p\Big(\frac{C_n^2 s\log(d\vee n)}{n}\Big), \tag{S24}$$

where we plug in the rate of convergence of $\tilde{\beta}$ and apply the Bernstein inequality and union bounds for $(\pi^*)'X_j/(\pi^*)^2T\epsilon^o$, since it is mean 0 by the definition of α^o and ψ_1 norm CC_n. The same empirical process theory as in the proof of Proposition 2 can be applied to derive $|I_{11}|\lesssim C_n^2 s\log(d\vee n)/n = o_p(n^{-1/2})$. Thus the same bound holds for I_1. This completes the proof. □

S5. MODIFIED ALGORITHM BASED ON SAMPLE SPLITTING

For simplicity, assume that we divide the data randomly into three folds I_1, I_2, I_3, where $I_1\cup I_2\cup I_3=[n]$. We consider the following algorithm.

Step 1: For a given $k\in\{1,2,3\}$, define a generalized quasi-likelihood function for data I_k as

$$Q_n(\beta;I_k) = \frac{1}{|I_k|}\sum_{i\in I_k}\int_0^{\beta^{\mathrm{T}}X_i}\Big\{\frac{T_i}{\pi(u)}-1\Big\}w_1(u)du,$$

where $w_1(u)$ is an arbitrary positive weight function. Compute the estimator

$$\hat{\beta}_{I_k} \;=\; \arg\min_{\beta\in\mathbb{R}^d} -Q_n(\beta; I_k) + \lambda\|\beta\|_1,$$

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

Step 2: For a given $k' \in \{1,2,3\}$ and $k' \neq k$, define a weighted least square loss function using the treatment group as

$$L_n(\alpha; I_{k'}) = \frac{1}{|I_{k'}|}\sum_{i\in I_{k'}} T_i w_2(\hat{\beta}_{I_k}^{\mathrm{T}} X_i)(Y_i - \alpha^{\mathrm{T}} X_i)^2,$$

where $w_2(\cdot)$ is another positive weight function. Compute the estimator

$$\hat{\alpha}_{I_{k'}} \;=\; \arg\min_{\alpha\in\mathbb{R}^d} L_n(\alpha; I_{k'}) + \lambda'\|\alpha\|_1,$$

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

Step 3: Compute the augmented inverse probability weighted estimator based on the sample $k'' \in \{1,2,3\}$ and $k'' \neq k$ and $k'' \neq k$

$$\hat{\mu}_1^{(k,k',k'')} = \frac{1}{|I_{k''}|}\sum_{i\in I_{k''}} \frac{T_i Y_i}{\pi_i(\hat{\beta}_{I_k}^{\mathrm{T}} X_i)} - \frac{1}{|I_{k''}|}\sum_{i\in I_{k''}} \{\frac{T_i}{\pi_i(\hat{\beta}_{I_k}^{\mathrm{T}} X_i)} - 1\}(\hat{\alpha}_{I_{k'}}^{\mathrm{T}} X_i).$$

Step 4: Define the final estimator as

$$\hat{\mu}_1 = \frac{1}{6}\sum_{\{k,k',k''\}=\{1,2,3\}} \hat{\mu}_1^{(k,k',k'')}.$$

It can be easily verified that

$$\|\tilde{\alpha}_{I_k} - \alpha^*\|_1 = O_p\left(s_2\Big\{\frac{\log(d\vee n)}{n}\Big\}^{1/2}\right),\;\; \mathrm{pr}_n\{X^{\mathrm{T}}(\hat{\alpha}_{I_k} - \alpha^*)\}^2 = O_p\left(\frac{s_2\log(d\vee n)}{n}\right).$$

$$\|\hat{\beta}_{I_k} - \beta^*\|_1 \;=\; O_p\left(s_1\Big\{\frac{\log(d\vee n)}{n}\Big\}^{1/2}\right),\;\; \mathrm{pr}_n\{X^{\mathrm{T}}(\hat{\beta}_{I_k} - \beta^*)\}^2 = O_p\left(\frac{s_1\log(d\vee n)}{n}\right).$$

Similar to Chernozhukov et al. (2018), we can show that

$$\hat{\mu}_1^{(k,k',k'')} - \mu_1^* = \frac{1}{|I_{k''}|}\sum_{i\in I_{k''}}\left[\frac{T_i}{\pi_i^*}\{Y_i(1) - \alpha^{*\mathrm{T}} X_i\} + \alpha^{*\mathrm{T}} X_i - \mu_1^*\right] + O_p\Big(\frac{(s_1 s_2)^{1/2}\log(d\vee n)}{n}\Big). \tag{S25}$$

The proof follows from the proof of Theorem 1, where we have the residual terms

$$R_1 = \mathrm{pr}_n\left(\frac{T}{\hat{\pi}} - \frac{T}{\pi^*}\right)\{Y(1) - K_1(X)\},\;\; R_2 = \mathrm{pr}_n\left(\frac{T}{\hat{\pi}} - 1\right)(\hat{\alpha} - \alpha^*)^{\mathrm{T}} X.$$

Here, for simplicity of notation, we use $\hat{\beta}$ and $\hat{\alpha}$ to denote $\hat{\beta}_{I_k}$ and $\hat{\alpha}_{I_k}$. Due to the use of independent samples, $\hat{\alpha}$ and $\hat{\beta}$ are independent of (T_i, X_i, Y_i) in the terms R_1 and R_2. We further decompose R_2 into

$$R_2 = \mathrm{pr}_n\left(\frac{T}{\pi^*} - 1\right)(\hat{\alpha} - \alpha^*)^{\mathrm{T}} X + \mathrm{pr}_n\left(\frac{T}{\hat{\pi}} - \frac{T}{\pi^*}\right)(\hat{\alpha} - \alpha^*)^{\mathrm{T}} X.$$

The second term can be bounded similar to the proof of Theorem 1, which yields

$$\left|\mathrm{pr}_n\left(\frac{T}{\hat{\pi}}-\frac{T}{\pi^*}\right)(\hat{\alpha}-\alpha^*)^{\mathrm{T}}X\right| \lesssim \frac{(s_1 s_2)^{1/2}\log(d\vee n)}{n}.$$

For the first term, we note that

$$E\left(\frac{T}{\pi^*}-1\right)(\hat{\alpha}-\alpha^*)^{\mathrm{T}}X = E\left\{E\left(\frac{T}{\pi^*}-1\right)(\hat{\alpha}-\alpha^*)^{\mathrm{T}}X|\hat{\alpha},X,T\right\}=0$$

holds, due to the independence of $\hat{\alpha}$ and (T_i, X_i, Y_i). Unlike the proof of Theorem 1, we directly apply the Hoeffding inequality. For some t to be chosen later, we have

$$\begin{aligned}\mathrm{pr}\Big\{\mathrm{pr}_n\left(\frac{T}{\pi^*}-1\right)(\hat{\alpha}-\alpha^*)^{\mathrm{T}}X>t\Big\} &= E\Big\{\mathrm{pr}\Big(\mathrm{pr}_n\left(\frac{T}{\pi^*}-1\right)(\hat{\alpha}-\alpha^*)^{\mathrm{T}}X>t|\hat{\alpha}\Big)\Big\}\\ &\le 2E\Big\{\exp\Big(-\frac{Ct^2n}{\Delta_n^2}\Big)\Big\}\end{aligned}$$

where $\Delta_n^2=\mathrm{pr}_n\{X^{\mathrm{T}}(\hat{\alpha}_{I_k}-\alpha^*)\}^2$. Since given any $\epsilon>0$, $\mathrm{pr}(\Delta_n^2\ge C's_1\log(d\vee n)/n)\le\epsilon$ for C' sufficiently large, by taking $t^2=C''C's_1\log(d\vee n)/n$ we have

$$\begin{aligned}E\Big\{\exp\Big(-\frac{Ct^2n}{\Delta_n^2}\Big)\Big\} &= E\Big[\exp\Big(-\frac{Ct^2n}{\Delta_n^2}\Big)I\{\Delta_n^2\ge C'\frac{s_1\log(d\vee n)}{n}\}\Big]\\ &\quad+E\Big[\exp\Big(-\frac{Ct^2n}{\Delta_n^2}\Big)I\{\Delta_n^2< C'\frac{s_1\log(d\vee n)}{n}\}\Big]\\ &\le \mathrm{pr}\{\Delta_n^2\ge C'\frac{s_1\log(d\vee n)}{n}\}+E\Big\{\exp\Big(-CC''\Big)\Big\}\le 2\epsilon,\end{aligned}$$

for C'' sufficiently large. This implies

$$\mathrm{pr}_n\left(\frac{T}{\pi^*}-1\right)(\hat{\alpha}-\alpha^*)^{\mathrm{T}}X\lesssim\Big\{\frac{1}{n}\frac{s_1\log(d\vee n)}{n}\Big\}^{1/2}.$$

Similarly, we can show that $R_1\lesssim s_2^{1/2}\log(d\vee n)/n$. This proves (S25). Finally, when we aggregate the estimators from different subsamples, by the Bahadur representation in (S25), we can easily show that Theorem 1 still holds for the final estimator.

S6. CONVERGENCE RATE OF THE AUGMENTED INVERSE PROBABILITY WEIGHTED ESTIMATORS UNDER MISSPECIFIED PROPENSITY SCORE MODELS

In this appendix, we prove the rate of convergence of the augmented inverse probability weighted estimator under misspecified propensity score models. Recall that, given estimators $\hat{\beta}$ and $\hat{\alpha}$, the augmented inverse probability weighted is defined as

$$\hat{\mu}_{AIPW} \;=\; \frac{1}{n}\sum_{i=1}^n\frac{T_i}{\hat{\pi}_i}Y_i-\frac{1}{n}\sum_{i=1}^n\left(\frac{T_i}{\hat{\pi}_i}-1\right)\hat{\alpha}^{\mathrm{T}}X_i,$$

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

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

$$I_1 \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\hat{\pi}_i}Y_i - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\hat{\pi}_i}-1\right)\alpha^{*\mathrm{T}}X_i,\;\; I_2 \;=\; \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\hat{\pi}_i}-1\right)(\hat{\alpha}-\alpha^*)^{\mathrm{T}}X_i.$$

After rearrangement of I_1, we have

$$I_1 \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\hat{\pi}_i}\epsilon_{1i} - \frac{1}{n}\sum_{i=1}^{n}\alpha^{*\mathrm{T}}X_i \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\pi_i^o}\epsilon_{1i} - \frac{1}{n}\sum_{i=1}^{n}\alpha^{*\mathrm{T}}X_i + o_p(n^{-1/2}),$$

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

The problem of augmented inverse probability weighted comes from the I_2 term. For instance, if $\hat{\alpha}$ is the Lasso estimator, then

$$I_2 \;\le\; \|\hat{\alpha}-\alpha^*\|_1\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\pi_i^o}-1\right)X_i\right\|_\infty + o_p(1)\right\} \;\lesssim\; \|\hat{\alpha}-\alpha^*\|_1 \;=\; O_p\left(s\Big(\frac{\log d}{n}\Big)^{1/2}\right),$$

where in the first step we use the rate of convergence of $\hat{\beta}$, the second step follows from $E\{(T_i/\pi_i^o - 1)X_{ij}\} \neq 0$ 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 $\hat{\mu}_{AIPW} \;=\; \mu_1^* + O_p(s(\log d/n)^{1/2})$. By applying the Cauchy inequality, we can similarly show $I_2 = O_p(s(\log d/n)^{1/2})$ and therefore we can obtain a slightly stronger result $\hat{\mu}_{AIPW} \;=\; \mu_1^* + O_p(s(\log d/n)^{1/2})$. Clearly, the bias term I_2 cannot converge 0 with rate faster than $n^{-1/2}$ even if the double machine learning (i.e, sample splitting or cross-fitting) in Chernozhukov et al. (2018) is applied. In sum, $\hat{\mu}_{AIPW}$ has a slower rate $O_p(s(\log d/n)^{1/2})$ under misspecified propensity score models.

In an independent work, Tan (2018) also realized the importance of (16) when studying the decomposition of $\hat{\mu}_{AIPW} - \mu_1^*$ as above. He proposed an alternative estimation procedure for β under the logistic propensity score model, rather than the covariate balancing estimator. In addition, our estimator $\hat{\mu}_1$ is an inverse probability weighted estimator, whereas he considered the augmented inverse probability weighted estimator $\hat{\mu}_{AIPW}$ in his approach. Finally, our theoretical results are more comprehensive. For instance, our Theorem 1, and Propositions 1 and 2 show that there exists a large class of estimators that is asymptotically normal under possible model misspecification. Second, our theory holds for generalized linear models as shown in Theorem 2, whereas his method is not applicable if the propensity score model is misspecified

S7. Additional Simulation Results

S7·1. *Summary*

Due to the space constraint, we include more extensive numerical results in this section. Section S7·3 contains simulations under different data generating processes. Section S7·4 contains simulations under logistic outcome models. Section S7·5 contains simulations for non-sparse models. Section S7·6 contains simulations under moderate dimensions. Section S7·7 contains the comparison with the normalized inverse probability weighted estimator. Finally, Section S7·8 contains sensitivity analysis with respect to the choice of tuning parameters.

Table 1: Bias, standard error (Std Err), standardized root-mean-squared error (RMSE), coverage probability of 95% confidence intervals (Coverage), and length of 95% confidence intervals (CI length) for the estimation of the average treatment effect. Four methods – the proposed method (HD-CBPS), approximate residual balancing (RB), regularized augmented inverse probability weighted estimator (AIPW), and double selection (D-SELECT) – are compared.

$n = 1000$	$d = 1000$				$d = 2000$			
	HD-CBPS	RB	AIPW	D-SELECT	HD-CBPS	RB	AIPW	D-SELECT
(1) Both models are correct								
Bias	-0·0233	-0·0234	-0·0814	-0·0476	0·0199	0·0186	-0·0056	0·0249
Std Err	0·0669	0·0777	0·0647	0·0690	0·0659	0·07476	0·0654	0·0757
RMSE	0·0695	0·0729	0·0678	0·0839	0·0689	0·0769	0·0657	0·0797
Coverage	0·955	0·940	0·905	0·920	0·940	0·935	0·950	0·955
CI length	0·2828	0·3010	0·2700	0·2978	0·2746	0·2979	0·2697	0·3187
(2) Propensity score model is misspecified								
Bias	-0·0297	-0·0455	-0·0931	-0·0362	0·0164	0·0135	-0·0137	0·0116
Std Err	0·0607	0·0694	0·0605	0·0665	0·0662	0·0758	0·0659	0·0846
RMSE	0·0671	0·0842	0·1105	0·0757	0·0682	0·0770	0·0673	0·0854
Coverage	0·970	0·930	0·855	0·930	0·940	0·955	0·955	0·935
CI length	0·2801	0·3040	0·2694	0·2960	0·2746	0·2987	0·2697	0·3381
(3) Outcome model is misspecified								
Bias	-0·0222	-0·0229	-0·0821	-0·0436	-0·0062	-0·0026	-0·0517	0·0262
Std Err	0·0670	0·0699	0·0653	0·0671	0·0653	0·0709	0·0630	0·0669
RMSE	0·0706	0·0735	0·1049	0·0800	0·0656	0·0709	0·0815	0·0718
Coverage	0·960	0·960	0·890	0·955	0·975	0·970	0·930	0·965
CI length	0·2842	0·3058	0·2709	0·3002	0·2848	0·3139	0·2726	0·2920
(4) Both models are misspecified								
Bias	-0·0157	-0·0072	-0·0504	-0·0366	0·0150	0·0009	-0·0635	0·0076
Std Err	0·0701	0·0822	0·0721	0·0687	0·0613	0·0765	0·0598	0·0792
RMSE	0·0718	0·0825	0·0880	0·0779	0·0631	0·0765	0·0872	0·0796
Coverage	0·945	0·960	0·905	0·925	0·990	0·960	0·905	0·950
CI length	0·2872	0·3117	0·2774	0·3046	0·2882	0·3281	0·2739	0·3426

S7·2. *Simulation results for* $n = 1000$

Table 1 shows the bias, standard error, standardized root mean squared error $\{\mathbb{E}(\hat{\mu} - \mu)^2\}^{1/2}/\mu$, coverage probability of 95% confidence intervals, and their length for the estimation of the average treatment effect under the same data generating process with $n = 1000$.

S7·3. *Simulation Results under Different Data Generating Processes*

In Table 2, we generate the d dimensional covariate $X_i \sim N(0, \Sigma)$ where $\Sigma_{jk} = \rho^{|j-k|}$ with $\rho = 0$. We generate the binary treatment T_i using the logistic regression model of the form, $\pi(X_i) = 1 - 1/\{1 + \exp(-X_{i1} + X_{i2}/2 - X_{i3}/4 - X_{i4}/10 - X_{i5}/10 + X_{i6}/10)\}$. The potential outcomes are generated from the linear regression models:

$$\begin{aligned} Y_i(1) &= 2 + 0 \cdot 137(X_{i5} + X_{i6} + X_{i7} + X_{i8}) + \epsilon_{1i}, \\ Y_i(0) &= 1 + 0 \cdot 291(X_{i5} + X_{i6} + X_{i7} + X_{i8} + X_{i9} + X_{i10}) + \epsilon_{0i}, \end{aligned}$$

where ϵ_{1i} and ϵ_{0i} are independent standard normal random variables. We use the same way as in the main paper to generate misspecified models. Four different estimators are compared in Table 2.

Table 2: Bias, standard error (Std Err), root-mean-squared error (RMSE), coverage probability of 95% confidence intervals (Coverage), length of 95% confidence intervals (CI length) for the estimation of the average treatment effect. Four methods – the proposed method (HD-CBPS), approximate residual balancing (RB), regularized augmented inverse probability weighting (AIPW), and double selection (D-SELECT) – are compared ($\rho = 0$).

$n = 500$	$d = 1000$				$d = 2000$			
	HD-CBPS	RB	AIPW	D-SELECT	HD-CBPS	RB	AIPW	D-SELECT
(1) Both models are correct								
Bias	-0·0935	-0·1186	-0·1642	-0·1047	-0·0252	-0·0290	-0·0087	-0·0511
Std Err	0·0986	0·1087	0·0945	0·0993	0·0898	0·0982	0·0863	0·0927
RMSE	0·1359	0·1609	0·1894	0·1443	0·0966	0·1064	0·0871	0·1059
Coverage	0·915	0·970	0·920	0·890	0·925	0·970	0·950	0·950
CI length	0·3876	0·4397	0·3719	0·4286	0·3483	0·4301	0·3350	0·4263
(2) Propensity score model is misspecified								
Bias	0·0233	0·0312	-0·0338	0·0923	0·0232	0·0184	0·0251	-0·0473
Std Err	0·0931	0·1014	0·0932	0·1064	0·0906	0·1055	0·0913	0·0926
RMSE	0·0987	0·1105	0·1112	0·1117	0·0963	0·1086	0·0979	0·1040
Coverage	0·930	0·975	0·895	0·925	0·930	0·945	0·920	0·960
CI length	0·3530	0·4447	0·3422	0·7331	0·3419	0·4271	0·3379	0·4083
(3) Outcome model is misspecified								
Bias	0·0197	0·0326	0·0508	-0·0234	-0·0560	-0·0509	-0·0617	-0·0246
Std Err	0·0999	0·1022	0·0895	0·1052	0·0963	0·1047	0·0925	0·1024
RMSE	0·1036	0·1120	0·1147	0·1077	0·1246	0·1269	0·1271	0·1053
Coverage	0·920	0·965	0·895	0·965	0·910	0·935	0·855	0·950
CI length	0·3555	0·4466	0·3388	0·4263	0·3455	0·4411	0·3360	0·4219
(4) Both models are misspecified								
Bias	0·0206	0·0416	0·0437	-0·0850	-0·0077	-0·0049	-0·0124	-0·0586
Std Err	0·0891	0·1062	0·0902	0·1095	0·0897	0·1049	0·0941	0·0972
RMSE	0·0937	0·1213	0·1093	0·1386	0·0903	0·1050	0·0956	0·1134
Coverage	0·950	0·945	0·920	0·870	0·970	0·975	0·945	0·930
CI length	0·3608	0·4517	0·3524	0·4340	0·3706	0·4679	0·3630	0·4345

In Tables 3 and 4, we consider the following data generating model. We generate the binary treatment T_i using the logistic regression model of the form, $\pi(X_i) = 1 - 1/\{1 + \exp(-X_{i1} + X_{i2}/2 - X_{i3}/4 - X_{i4}/10)\}$. The potential outcomes are generated from the linear regression models:

$$\begin{aligned} Y_i(1) &= 2 + 0 \cdot 137(X_{i5} + X_{i6} + X_{i7} + X_{i8}) + \epsilon_{1i}, \\ Y_i(0) &= 1 + 0 \cdot 291(X_{i5} + X_{i6} + X_{i7} + X_{i8} + X_{i9} + X_{i10}) + \epsilon_{0i}. \end{aligned}$$

In this data generating model, we assume there are no confounding variables. We consider $n = 500, 1000$ in Table 3 and Table 4, respectively.

Table 3: Simulation results for data generating processes without confounding variables under $n = 500$.

$n = 500$	$d = 1000$				$d = 2000$			
	HD-CBPS	RB	AIPW	D-SELECT	HD-CBPS	RB	AIPW	D-SELECT
(1) Both models are correct								
Bias	-0·0294	-0·0350	-0·1100	-0·0950	-0·0693	-0·0772	-0·1293	-0·0579
Std Err	0·0904	0·0977	0·0824	0·0970	0·1031	0·1131	0·0961	0·1140
RMSE	0·0951	0·1037	0·1374	0·1358	0·1242	0·1369	0·1611	0·1279
Coverage	0·955	0·960	0·855	0·890	0·880	0·895	0·795	0·930
CI length	0·3990	0·4401	0·3823	0·4295	0·3869	0·4349	0·3699	0·4600
(2) Propensity score model is misspecified								
Bias	-0·0683	-0·1090	-0·1564	-0·0581	-0·0454	-0·0718	-0·1263	0·0456
Std Err	0·1016	0·1157	0·0983	0·1614	0·0993	0·1131	0·0979	0·1142
RMSE	0·1225	0·1590	0·1847	0·1715	0·1092	0·1340	0·1560	0·1229
Coverage	0·905	0·840	0·575	0·960	0·900	0·890	0·710	0·930
CI length	0·3897	0·4423	0·3757	0·6469	0·3840	0·4386	0·3712	0·4806
(3) Outcome model is misspecified								
Bias	-0·0097	-0·0190	-0·1039	-0·0382	0·0315	-0·0383	-0·0765	-0·0174
Std Err	0·1002	0·1101	0·0962	0·1193	0·0959	0·1007	0·0927	0·1113
RMSE	0·1007	0·1118	0·1416	0·1252	0·1010	0·1077	0·1202	0·1127
Coverage	0·965	0·965	0·825	0·980	0·940	0·960	0·885	0·960
CI length	0·4033	0·4458	0·3867	0·4860	0·3927	0·4590	0·3769	0·4791
(4) Both models are misspecified								
Bias	-0·0294	-0·1079	-0·1857	-0·0850	-0·0184	-0·0555	-0·1383	-0·0586
Std Err	0·1222	0·1106	0·0952	0·1095	0·0994	0·1110	0·0950	0·0972
RMSE	0·1257	0·1545	0·2087	0·1386	0·1011	0·1241	0·1678	0·1135
Coverage	0·895	0·865	0·490	0·870	0·950	0·940	0·710	0·930
CI length	0·4024	0·4601	0·3804	0·4340	0·3922	0·4539	0·3769	0·4345

S7·4. *Simulation Results for Logistic Outcome Model*

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

$$\begin{aligned}\text{pr}\{Y_i(1) = 1 \mid X_i\} &= 1 - 1/\{1 + \exp(2 + 0.137X_{i1} + 0.137X_{i2} + 0.137X_{i3} + 0.137X_{i7} \\ &\qquad + 0.137X_{i8} + 0.137X_{i9})\}, \\ \text{pr}\{Y_i(0) = 1 \mid X_i\} &= 1 - 1/\{1 + \exp(1 + 0.291X_{i1} + 0.291X_{i2} + 0.291X_{i3} + 0.291X_{i5} \\ &\qquad + 0.291X_{i6} + 0.291X_{i7} + 0.291X_{i8} + 0.291X_{i9})\}.\end{aligned}$$

When the outcome variable is binary, the approximate residual balancing method is not directly applicable. Thus, we only compare our method with the regularized augmented inverse probability weighted and double selection methods. For the logistic outcome model, we report the bias, standard error, standardized root mean squared error, coverage probability of 95% confidence intervals, and length of 95% confidence intervals for estimating the average treatment effect. Similarly, the misspecified models are considered. We consider $n = 500, 1000$ in Table 5 and Table 6, respectively. It is seen that the proposed method outperforms the regularized augmented inverse probability weighted and double selection methods in almost all scenarios. In contrast to augmented inverse probability weighted estimator, the proposed confidence inter-

Table 4: Simulation results for data generating processes without confounding variables under $n = 1000$.

$n = 1000$	$d = 1000$				$d = 2000$			
	HD-CBPS	RB	AIPW	D-SELECT	HD-CBPS	RB	AIPW	D-SELECT
(1) Both models are correct								
Bias	-0·0135	-0·0059	-0·0423	-0·0050	0·0121	0·0188	-0·0146	0·0119
Std Err	0·0669	0·0777	0·0647	0·0713	0·0728	0·0806	0·0727	0·0820
RMSE	0·0682	0·0779	0·0772	0·0714	0·0739	0·0815	0·0742	0·0828
Coverage	0·960	0·950	0·925	0·960	0·925	0·940	0·925	0·935
CI length	0·2786	0·2998	0·2694	0·2916	0·2736	0·2952	0·2682	0·3107
(2) Propensity score model is misspecified								
Bias	-0·0145	-0·0186	-0·0476	-0·0116	0·0103	0·0071	-0·0231	0·0185
Std Err	0·0607	0·0694	0·0605	0·0650	0·0621	0·0725	0·0619	0·0669
RMSE	0·0624	0·0718	0·0770	0·0660	0·0629	0·0728	0·0660	0·0694
Coverage	0·975	0·975	0·920	0·970	0·970	0·955	0·955	0·950
CI length	0·2763	0·2996	0·2695	0·2895	0·2730	0·2979	0·2682	0·2876
(3) Outcome model is misspecified								
Bias	-0·0088	0·0082	-0·0480	-0·0428	-0·0081	-0·0017	-0·0517	0·0396
Std Err	0·0661	0·0721	0·0643	0·0670	0·0694	0·0747	0·0670	0·0692
RMSE	0·0667	0·0726	0·0803	0·0795	0·0699	0·0747	0·0846	0·0797
Coverage	0·965	0·955	0·925	0·945	0·965	0·970	0·885	0·935
CI length	0·2852	0·3100	0·2760	0·3002	0·2848	0·3139	0·2726	0·3010
(4) Both models are misspecified								
Bias	-0·0155	-0·0088	-0·0504	-0·0056	0·0108	-0·0007	-0·0514	0·0058
Std Err	0·0626	0·0696	0·0615	0·0693	0·0597	0·0737	0·0603	0·0681
RMSE	0·0645	0·0702	0·0795	0·0695	0·0607	0·0737	0·0792	0·0683
Coverage	0·980	0·975	0·915	0·970	0·990	0·970	0·940	0·970
CI length	0·2869	0·3103	0·2768	0·2983	0·2835	0·3144	0·2749	0·3049

vals under the logistic outcome model have accurate coverage probabilities under misspecified models, which is consistent with our theoretical results.

S7·5. *Simulation Results for Non-sparse Models*

In this section, we consider the following non-sparse data generating process. In the first scenario, we generate the treatment variable and outcome variables from

$$\text{pr}(T_i = 0|X_i) = \{1 + \exp(-X_{i1} + 0.5X_{i2} - 0.25X_{i3} - 0.1X_{i4} - 0.01X_{i5} + 0.01X_{i6}\}^{-1},$$

$$Y_i(1) = 2 + \sum_{j=3}^{d} j^{-\ell} X_{ij} + \epsilon_{1i}, \;\; Y_i(0) = 1 + 2\sum_{j=3}^{d} j^{-\ell} X_{ij} + \epsilon_{0i},$$

where $\ell = 2$. Apparently, the outcome models depend on almost all variables, and we let the coefficients shrink to 0 for large j. The coefficients for large j represent the weak signals in the model which may not be selected by the penalized estimators. Thus, in the setting the variable selection consistency is impossible for the outcome models. In the second scenario, we further

Table 5: Simulation results for logistic outcome models with $n = 500$.

$n = 500$	$d = 1000$			$d = 2000$		
	AIPW	D-SELECT	HD-CBPS	AIPW	D-SELECT	HD-CBPS
(1) Both models are correct						
Bias	0·0339	0·0298	0·0295	0·0484	0·0296	0·0226
Std Err	0·0480	0·0428	0·0381	0·0504	0·0437	0·0394
RMSE	0·0587	0·0522	0·0482	0·0698	0·0528	0·0455
Coverage	0·830	0·915	0·960	0·795	0·890	0·960
CI length	0·1645	0·1765	0·1908	0·1633	0·1777	0·1805
(2) Propensity score model is misspecified						
Bias	0·0771	0·0311	0·0274	0·0582	0·0211	0·0220
Std Err	0·0454	0·0478	0·0329	0·0480	0·0466	0·0365
RMSE	0·0895	0·0570	0·0428	0·0754	0·0511	0·0426
Coverage	0·355	0·905	0·920	0·560	0·905	0·890
CI length	0·1272	0·1908	0·1420	0·1262	0·1719	0·1363
(3) Outcome model is misspecified						
Bias	0·0219	-0·0070	-0·0070	0·0560	0·0290	0·0228
Std Err	0·0402	0·0427	0·0310	0·0455	0·0422	0·0330
RMSE	0·0458	0·0433	0·0318	0·0722	0·0512	0·0402
Coverage	0·850	0·980	0·975	0·550	0·910	0·895
CI length	0·1273	0·1892	0·1461	0·1244	0·1760	0·1308
(4) Both models are misspecified						
Bias	0·0427	0·0241	0·0152	0·0226	0·0259	0·0074
Std Err	0·0431	0·0504	0·0360	0·0407	0·0397	0·0350
RMSE	0·0607	0·0558	0·0391	0·0466	0·0474	0·0357
Coverage	0·665	0·895	0·925	0·860	0·920	0·950
CI length	0·1281	0·1911	0·1417	0·1287	0·1673	0·1396

allow the propensity score model to be also non-sparse:

$$\text{pr}(T_i = 0|X_i) = \{1 + \exp(-X_{i1} + 0.5X_{i2} - 0.25X_{i3} - 0.1X_{i4} - 0.01X_{i5} + 0.01X_{i6} + \sum_{j=7}^{d} j^{-\ell} X_{ij})\}^{-1},$$

where $\ell = 2$. The same non-sparse outcome models are used. Thus, in this setting, both propensity score and outcome models are non-sparse. The results in Table 7 show that the proposed method can still accurately estimate the treatment effect when there exist weak signals. Again, this empirical result is consistent with our theoretical result which says that the asymptotic inference based on the covariate balancing approach does not require variable selection consistency as a priori. In addition, it also implies that our method is robust to minor violation of the sparsity assumption. We also consider a more challenging non-sparse case with $\ell = 1$, that is the coefficients decay to 0 more slowly. The results are shown in Table 8. We observe the same phenomenon.

S7·6. *Simulation Results Under Moderate Dimension*

In this section, we compare eight methods under moderate dimension: our method, approximate residual balancing method (Athey et al., 2018), regularized augmented inverse probability

Table 6: Simulation results for logistic outcome models with $n = 1000$.

$n = 1000$	$d = 1000$			$d = 2000$		
	AIPW	D-SELECT	HD-CBPS	AIPW	D-SELECT	HD-CBPS
(1) Both models are correct						
Bias	-0·0251	0·0043	0·0081	-0·0307	0·0098	0·0092
Std Err	0·0300	0·0302	0·0291	0·0282	0·0303	0·0275
RMSE	0·0391	0·0305	0·0302	0·0417	0·0318	0·0290
Coverage	0·890	0·950	0·960	0·871	0·929	0·971
CI length	0·1233	0·1147	0·1329	0·1219	0·1213	0·1348
(2) Propensity score model is misspecified						
Bias	0·0170	0·0018	0·0039	0·0198	0·0014	0·0034
Std Err	0·0288	0·0264	0·0256	0·0284	0·0373	0·0259
RMSE	0·0335	0·0265	0·0259	0·0346	0·0373	0·0261
Coverage	0·850	0·965	0·935	0·810	0·942	0·956
CI length	0·0980	0·1129	0·1017	0·0969	0·1491	0·1036
(3) Outcome model is misspecified						
Bias	0·0482	0·0034	0·0051	0·00253	0·0053	0·0035
Std Err	0·0315	0·0395	0·0241	0·0293	0·0311	0·0245
RMSE	0·0576	0·0397	0·0246	0·0387	0·0315	0·0248
Coverage	0·480	0·975	0·960	0·745	0·960	0·955
CI length	0·0942	0·1575	0·1041	0·0934	0·1219	0·0983
(4) Both models are misspecified						
Bias	0·0199	-0·0015	-0·0028	0·0126	-0·0289	-0·0113
Std Err	0·0302	0·0279	0·0252	0·0323	0·0936	0·0282
RMSE	0·0362	0·0280	0·0254	0·0346	0·0980	0·0304
Coverage	0·780	0·975	0·960	0·845	0·985	0·945
CI length	0·0966	0·1211	0·1104	0·0982	0·2793	0·1152

weighted method (Farrell, 2015; Belloni et al., 2017), double selection (Belloni et al., 2014), calibrated likelihood (Tan, 2010), targeted maximum likelihood estimator (Van der Laan & Rose, 2011), covariate balancing propensity score (Imai & Ratkovic, 2014), inverse propensity score weighted estimator with the maximum likelihood estimator. The first four methods are implemented in the same way as before. The calibrated likelihood, targeted maximum likelihood estimator and covariate balancing propensity score are computed using the R packages iWeigReg, tmle, CBPS, respectively. Since these existing softwares cannot directly incorporate the penalty in the estimation procedures, we mainly focus on the comparison under moderate dimension (in the sense that d^2 is comparable to n). In particular, we consider two scenarios $(n, d) = (500, 10)$ and $(n, d) = (500, 30)$. The data generating processes are the same as the ones used in the main paper. The only difference is that we reduce d from thousands to $d = 10, 30$. The results are shown in Tables 9 and 10. It shows that our method is comparable to these well established methods for low dimensional problems, such as augmented inverse probability weighted, calibrated likelihood and targeted maximum likelihood methods. Recall that for the regularized augmented inverse probability weighted estimator $\hat{\mu}_1$, we show that

$$\hat{\mu}_1 - \mu_1^* = O_p\Big(\Big\{\frac{(s_1 \vee s_2)\log(d \vee n)}{n}\Big\}^{1/2}\Big),$$

which is slower than the rate of our estimator. However, when d is relatively small and fixed, $\hat{\mu}_1$ reduces to the root-n rate up to a logarithmic factor of n, which is identical to our estimator. Thus,

Table 7: Simulation results for non-sparse models with $n = 500, d = 1000$ and $\ell = 2$.

	Outcome models are non-sparse				Outcome and propensity score models are non-sparse			
	HD-CBPS	RB	AIPW	D-SELECT	HD-CBPS	RB	AIPW	D-SELECT
(1) Both models are correct								
Bias	-0·0213	-0·0073	-0·0400	-0·0025	-0·0275	-0·0095	-0·0352	0·0117
Std Err	0·0899	0·0997	0·0870	0·1183	0·0916	0·1047	0·0909	0·1014
RMSE	0·0924	0·1000	0·0958	0·1184	0·0957	0·1051	0·0975	0·1021
Coverage	0·955	0·965	0·920	0·960	0·920	0·965	0·920	0·965
CI length	0·3559	0·4215	0·3365	0·4510	0·3459	0·4100	0·3384	0·4172
(2) Propensity score model is misspecified								
Bias	-0·0085	-0·0148	-0·0476	-0·0025	-0·0066	-0·0004	-0·0283	-0·0068
Std Err	0·0877	0·1002	0·0832	0·1168	0·0951	0·1058	0·0949	0·0978
RMSE	0·0881	0·1013	0·0959	0·1168	0·0954	0·1058	0·0990	0·0980
Coverage	0·950	0·970	0·910	0·975	0·945	0·940	0·895	0·960
CI length	0·3563	0·4261	0·3361	0·5075	0·3476	0·4091	0·3400	0·3705
(3) Outcome model is misspecified								
Bias	-0·0040	-0·0020	-0·0407	-0·0146	-0·0003	0·0105	-0·0125	0·0136
Std Err	0·0969	0·1046	0·0919	0·1137	0·0905	0·1045	0·0924	0·1091
RMSE	0·0970	0·1046	0·1005	0·1146	0·0905	0·1050	0·0933	0·1100
Coverage	0·925	0·965	0·905	0·955	0·960	0·950	0·940	0·960
CI length	0·3545	0·4215	0·3344	0·4626	0·3446	0·4085	0·3389	0·4161
(4) Both models are misspecified								
Bias	0·0009	-0·0199	-0·0476	0·0067	-0·0015	-0·0038	-0·0043	0·0186
Std Err	0·0992	0·1118	0·0955	0·1271	0·0916	0·1084	0·0902	0·0929
RMSE	0·0993	0·1136	0·1067	0·1272	0·0916	0·1085	0·0903	0·0948
Coverage	0·920	0·945	0·885	0·955	0·945	0·950	0·935	0·960
CI length	0·3602	0·42327	0·3360	0·4887	0·3504	0·4093	0·3372	0·3739

the improvement of the convergence rate does not occur in low dimensional case. We believe this is the main reason why our method has similar performance to these well established methods. The inverse probability weighted estimator demonstrates larger bias and has wider confidence intervals when the propensity score model is misspecified. All the other methods lead to fairly accurate coverage probability provided either the propensity score model or outcome models is correctly specified.

S7·7. *Comparison with Normalized Horvitz-Thompson Estimator*

Table 11 is to compare the proposed method with the so-called normalized inverse probability weighted. The normalized inverse probability weighted estimator is defined as

$$\hat{\mu}_{NIPW} = \frac{\sum_{i=1}^{n} T_i/\hat{\pi}_i Y_i}{\sum_{i=1}^{n} T_i/\hat{\pi}_i} - \frac{\sum_{i=1}^{n} (1-T_i)/(1-\hat{\pi}_i) Y_i}{\sum_{i=1}^{n} (1-T_i)/(1-\hat{\pi}_i)},$$

where $\hat{\pi}_i$ is the estimated propensity score for the ith sample based on the penalized maximum likelihood estimation. The confidence interval of $\hat{\mu}_{NIPW}$ is obtained by the bootstrap method.

We generate the binary treatment T_i using the logistic regression model of the form, $\pi(X_i) = 1 - 1/\{1 + \exp(-X_{i1} + X_{i2}/2 - X_{i3}/4 - X_{i4}/10 - X_{i5}/10 + X_{i6}/10)\}$. The potential out-

Table 8: Simulation results for non-sparse models with $n = 500, d = 1000$ and $\ell = 1$.

	Outcome models are non-sparse				Outcome and propensity score models are non-sparse			
	HD-CBPS	RB	AIPW	D-SELECT	HD-CBPS	RB	AIPW	D-SELECT
(1) Both models are correct								
Bias	-0.0051	-0.0022	-0.0047	-0.0033	0.0079	0.0123	0.0064	-0.0151
Std Err	0.0898	0.0970	0.0894	0.1176	0.0951	0.1040	0.0951	0.1568
RMSE	0.0900	0.0970	0.0895	0.1176	0.0955	0.1048	0.0953	0.1575
Coverage	0.9500	0.9600	0.9300	0.9650	0.9192	0.9394	0.9040	0.9747
CI length	0.3494	0.4079	0.3335	0.4552	0.3467	0.4080	0.3309	0.6406
(2) Propensity score model is misspecified								
Bias	-0.0041	0.0020	-0.0057	-0.0056	0.0083	0.0172	0.0030	0.0041
Std Err	0.0930	0.0960	0.0938	0.0963	0.0936	0.1051	0.0942	0.0941
RMSE	0.0930	0.0960	0.0940	0.0965	0.0940	0.1065	0.0942	0.0942
Coverage	0.9350	0.9600	0.9200	0.9450	0.9343	0.9197	0.9343	0.9489
CI length	0.3425	0.3980	0.3337	0.3681	0.3491	0.4041	0.3404	0.3723
(3) Outcome model is misspecified								
Bias	-0.0121	-0.0203	-0.0133	-0.0266	0.0053	-0.0021	0.0058	-0.0039
Std Err	0.1005	0.1080	0.0947	0.1324	0.0952	0.1043	0.0899	0.1263
RMSE	0.1012	0.1099	0.0956	0.1350	0.0954	0.1043	0.0901	0.1263
Coverage	0.9300	0.9400	0.9100	0.9450	0.9750	0.9750	0.9600	0.9550
CI length	0.3635	0.4192	0.3310	0.4964	0.3622	0.4204	0.3323	0.4951
(4) Both models are misspecified								
Bias	0.0098	-0.0018	0.0109	0.0099	0.0010	-0.0057	0.0006	-0.0017
Std Err	0.0945	0.1074	0.0987	0.1027	0.0871	0.1074	0.0920	0.0970
RMSE	0.0950	0.1074	0.0993	0.1031	0.0871	0.1075	0.0920	0.0970
Coverage	0.9300	0.9750	0.9100	0.9350	0.9350	0.9350	0.9200	0.9400
CI length	0.3443	0.4118	0.3336	0.3791	0.3448	0.4137	0.3339	0.3831

comes are generated from the linear regression models:

$$\begin{aligned} Y_i(1) &= 2 + 0 \cdot 137(X_{i5} + X_{i6} + X_{i7} + X_{i8}) + \epsilon_{1i}, \\ Y_i(0) &= 1 + 0 \cdot 291(X_{i5} + X_{i6} + X_{i7} + X_{i8} + X_{i9} + X_{i10}) + \epsilon_{0i}, \end{aligned}$$

where ϵ_{1i} and ϵ_{0i} are independent standard normal random variables. We use the same way as in the main paper to generate misspecified models. We find that the normalized inverse probability weighted estimator has smaller mean squared error when the dimension d is small, say $d = 10$. Given the large sample size $n = 1000$, we do not expect our method to outperform the classical proposals, including the normalized inverse probability weighted, in this low-dimensional case. Indeed, this phenomenon is consistent with the more extensive simulation studies under moderate dimension conducted in Section S7·6. However, as the dimension d grows to 30, the proposed method starts to have smaller or at least comparable mean squared error in the moderate dimensional case. Finally, in the very high dimensional case ($d = 1000, 2000$), the proposed method clearly shows smaller mean squared error and more accurate coverage probability.

S7·8. *Sensitivity Analysis for Tuning Parameters*

In this section, we conduct sensitivity analysis to examine whether our numerical results are sensitive to the tuning parameters. We generate the binary treatment T_i using the logistic regression model of the form, $\pi(X_i) = 1 - 1/\{1 + \exp(-X_{i1} + X_{i2}/2 - X_{i3}/4 - X_{i4}/10 -$

Table 9: Simulation results under moderate dimension, $n = 500, d = 10$. The proposed method (HD-CBPS) is compared with the approximate residual balancing (RB), regularized augmented inverse probability weighted (AIPW), double selection (D-SELECT), calibrated likelihood (CL), targeted maximum likelihood estimator (TMLE), covariate balancing propensity score (CBPS), and inverse propensity score weighted estimator with the maximum likelihood estimator (IPW).

	HD-CBPS	RB	AIPW	D-SELECT	CL	TMLE	CBPS	IPW
(1) Both models are correct								
Bias	0·0878	0·0884	0·0269	0·0889	0·0871	0·0923	-0·1638	0·1225
Std Err	0·1031	0·1009	0·0983	0·1019	0·1034	0·1052	0·0976	0·1032
RMSE	0·1354	0·1341	0·1019	0·1352	0·1352	0·1400	0·1907	0·1602
Coverage	0·955	0·955	0·940	0·965	0·975	0·945	0·915	1·000
CI length	0·4060	0·3949	0·3837	0·4078	0·4130	0·3896	0·5885	0·5070
(2) Propensity score model is misspecified								
Bias	0·0889	0·0907	0·0276	0·0895	0·0908	0·1028	0·0634	0·0883
Std Err	0·1042	0·1074	0·0971	0·1008	0·1096	0·1106	0·1002	0·1026
RMSE	0·1370	0·1406	0·1009	0·1348	0·1423	0·1510	0·1186	0·1354
Coverage	0·965	0·970	0·945	0·980	0·975	0·925	0·960	0·970
CI length	0·4066	0·4149	0·3829	0·4056	0·4143	0·3986	0·4819	0·5050
(3) Outcome model is misspecified								
Bias	0·0940	0·0899	0·0272	0·0954	0·0901	0·0939	-0·1343	0·1373
Std Err	0·0969	0·0953	0·0935	0·0953	0·0964	0·0994	0·0908	0·0968
RMSE	0·1350	0·1310	0·0973	0·1348	0·1320	0·1367	0·1621	0·1680
Coverage	0·955	0·980	0·940	0·965	0·960	0·940	0·950	1·000
CI length	0·4086	0·4005	0·3845	0·4091	0·4146	0·3873	0·5882	0·5797
(4) Both models are misspecified								
Bias	0·0215	-0·0384	-0·0093	0·0414	0·0452	0·0643	0·0162	0·0302
Std Err	0·1023	0·0887	0·0833	0·0846	0·0919	0·0926	0·0850	0·0876
RMSE	0·1045	0·0966	0·0838	0·0942	0·1024	0·1127	0·0865	0·0927
Coverage	0·950	0·970	0·990	0·960	0·955	0·920	0·995	1·000
CI length	0·4125	0·4220	0·3889	0·4099	0·4204	0·4079	0·5245	0·5452

$X_{i5}/10 + X_{i6}/10)\}$. The potential outcomes are generated from the linear regression models:

$$\begin{aligned} Y_i(1) &= 2 + 0 \cdot 137(X_{i5} + X_{i6} + X_{i7} + X_{i8}) + \epsilon_{1i}, \\ Y_i(0) &= 1 + 0 \cdot 291(X_{i5} + X_{i6} + X_{i7} + X_{i8} + X_{i9} + X_{i10}) + \epsilon_{0i}, \end{aligned}$$

where ϵ_{1i} and ϵ_{0i} are independent standard normal random variables. We consider $n = 500$ and $d = 1000$. The proposed method contains two tuning parameters λ in the propensity score model, λ' in the outcome model. In our method, we recommend the cross-validation method for choosing tuning parameters. In this example, the cross-validation method leads to tuning parameters $\lambda_{CV} = 0.0496$ and $\lambda'_{CV} = 0.1674$ on average. In Table 12, we consider two cases. In case (1): we estimate the treatment effect by changing λ in the propensity score model from the cross-validation value by a small number and fixing λ' at the cross-validation value. In case (2): we estimate the treatment effect by changing λ' in the outcome model and fixing λ at the cross-validation value. We use the grid $[\lambda_{CV} - 0 \cdot 02, \lambda_{CV} + 0 \cdot 02]$, because by further increasing λ_{CV} all variables in the propensity score model are shrunk to 0 and similarly by further decreasing λ_{CV} most variables are kept in the propensity score model leading to slow convergence of the gradient descent algorithms for solving the penalized optimization problem. This seems to

Table 10: Simulation results under moderate dimension, $n = 500, d = 30$.

	HD-CBPS	RB	AIPW	D-SELECT	CL	TMLE	CBPS	IPW
(1) Both models are correct								
Bias	-0·0002	-0·0107	0·0160	0·0061	-0·0099	-0·0182	-0·0467	-0·0290
Std Err	0·0703	0·0740	0·0680	0·0753	0·0776	0·0761	0·0706	0·0749
RMSE	0·0703	0·0748	0·0699	0·0755	0·0782	0·0783	0·0847	0·0803
Coverage	0·950	0·965	0·915	0·950	0·945	0·945	0·955	0·975
CI length	0·2726	0·2919	0·2521	0·3012	0·2959	0·2802	0·3304	0·3724
(2) Propensity score model is misspecified								
Bias	-0·0036	-0·0139	0·0160	0·0018	-0·0169	-0·0156	-0·0202	0·0813
Std Err	0·0660	0·0716	0·0635	0·0714	0·0731	0·0726	0·0660	0·0748
RMSE	0·0661	0·0730	0·0655	0·0714	0·0750	0·0742	0·0690	0·1104
Coverage	0·960	0·965	0·965	0·970	0·945	0·950	0·995	1·000
CI length	0·2695	0·2629	0·2526	0·3205	0·3001	0·3027	0·3255	0·4227
(3) Outcome model is misspecified								
Bias	0·0042	-0·0008	0·0157	0·0043	0·0063	-0·0025	-0·0187	-0·0032
Std Err	0·0714	0·0751	0·0687	0·0767	0·0768	0·0770	0·0725	0·0746
RMSE	0·0715	0·0751	0·0705	0·0768	0·0770	0·0770	0·0748	0·0746
Coverage	0·950	0·945	0·935	0·955	0·955	0·935	0·985	0·990
CI length	0·2849	0·2967	0·2606	0·3067	0·3014	0·2944	0·3522	0·3983
(4) Both models are misspecified								
Bias	0·0262	0·0128	0·0255	0·0161	0·0126	0·0119	0·0844	0·1170
Std Err	0·0662	0·0676	0·0620	0·0765	0·0728	0·0730	0·0707	0·0766
RMSE	0·0712	0·0688	0·0670	0·0782	0·0739	0·0740	0·1100	0·1398
Coverage	0·955	0·970	0·940	0·955	0·960	0·960	0·900	0·965
CI length	0·4125	0·4220	0·3889	0·4099	0·4204	0·4079	0·5245	0·5452

be a reasonable choice. Table 12 shows that the mean squared error and the coverage probability of our method are not very sensitive to the tuning parameters.

REFERENCES

ATHEY, S., IMBENS, G. W. & WAGER, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 597–623.

BELLONI, A., CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. & HANSEN, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica* **85**, 233–298.

BELLONI, A., CHERNOZHUKOV, V. & HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* **81**, 608–650.

BUSSO, M., DINARDO, J. & MCCRARY, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics* **96**, 885–897.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. & ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21**, 733–750.

FAN, J., IMAI, K., LIU, H., NING, Y. & YANG, X. (2016). Improving covariate balancing propensity score: A doubly robust and efficient approach. Tech. rep., Princeton University.

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

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: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* **20**, 25–46.

Table 11: Comparison with normalized inverse probability weighted (N-IPW) under $n = 1000$

	$d = 10$		$d = 30$		$d = 1000$		$d = 2000$	
	HD-CBPS	N-IPW	HD-CBPS	N-IPW	HD-CBPS	N-IPW	HD-CBPS	N-IPW
(1) Both models are correct								
Bias	0·0501	0·0288	-0·0409	-0·0505	0·0038	-0·0124	-0·0146	-0·1008
Std Err	0·0637	0·0669	0·0645	0·0660	0·0719	0·0731	0·0667	0·0664
RMSE	0·0811	0·0728	0·0763	0·0831	0·0720	0·0741	0·0682	0·1207
Coverage	0·915	0·960	0·935	0·950	0·930	0·955	0·960	0·800
CI length	0·2868	0·3180	0·2850	0·3096	0·2636	0·3046	0·2785	0·3233
(2) Propensity score model is misspecified								
Bias	0·0535	0·0165	-0·0428	-0·0511	0·0007	-0·0285	-0·0149	-0·1164
Std Err	0·0597	0·0610	0·0670	0·0706	0·0620	0·0634	0·0638	0·0657
RMSE	0·0802	0·0632	0·0795	0·0872	0·0620	0·0695	0·0655	0·1337
Coverage	0·945	0·990	0·955	0·945	0·975	0·960	0·965	0·740
CI length	0·2932	0·2993	0·2912	0·3138	0·2649	0·2983	0·2799	0·3172
(3) Outcome model is misspecified								
Bias	0·0490	0·0229	-0·0423	-0·0574	0·0258	-0·0690	0·0056	0·1109
Std Err	0·0653	0·0695	0·0705	0·0740	0·0670	0·0678	0·0594	0·0583
RMSE	0·0817	0·0732	0·0823	0·0936	0·0718	0·0965	0·0596	0·1109
Coverage	0·900	0·980	0·940	0·910	0·965	0·905	0·990	0·860
CI length	0·2870	0·3172	0·2848	0·3146	0·2880	0·3003	0·2807	0·3241
(4) Both models are misspecified								
Bias	0·0104	0·0068	-0·0236	-0·0522	-0·0134	-0·0651	-0·0601	-0·1264
Std Err	0·0707	0·0698	0·0842	0·0710	0·0808	0·0714	0·0728	0·0703
RMSE	0·0714	0·0702	0·0874	0·0881	0·0819	0·0966	0·0944	0·1446
Coverage	0·970	0·965	0·930	0·930	0·940	0·960	0·860	0·675
CI length	0·3005	0·3049	0·3028	0·3166	0·2940	0·3635	0·2845	0·3263

Table 12: Sensitivity analysis for the tuning parameters.

	-0.02	-0.01	CV	+0.01	+0.02
(1) Change of the tuning parameter in the propensity score model					
Bias	-0·0165	-0·0173	-0·0149	-0·0156	-0·0143
Std Err	0·1035	0·1000	0·0978	0·0995	0·0972
RMSE	0·1048	0·1016	0·0989	0·1007	0·0983
Coverage	0·975	0·975	0·965	0·975	0·960
CI length	0·4627	0·4425	0·4345	0·4311	0·4284
(2) Change of the tuning parameter in the outcome model					
Bias	-0·0186	-0·0153	-0·0145	-0·0153	-0·0144
Std Err	0·1038	0·1048	0·1071	0·1090	0·1050
RMSE	0·1054	0·1059	0·1081	0·1100	0·1060
Coverage	0·960	0·970	0·960	0·950	0·980
CI length	0·4324	0·4332	0·4328	0·4312	0·4352

IMAI, K. & RATKOVIC, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 243–263.

NING, Y. & LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics* **45**, 158–195.

ROBINS, J., SUED, M., LEI-GOMEZ, Q. & 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. & ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* **89**, 846–866.

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

TAN, Z. (2018). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *arXiv preprint arXiv:1801.09817* .

VAN DER LAAN, M. J. & ROSE, S. (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

YANG, Z., NING, Y. & LIU, H. (2018). On semiparametric exponential family graphical models. *The Journal of Machine Learning Research* **19**, 2314–2372.

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

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

[*Received M Y. Revised M Y*]