Root-n consistent semiparametric learning with high-dimensional nuisance functions under minimal sparsity

Lin Liu[1]*, Yuhao Wang[2] †

[1]Institute of Natural Sciences, MOE-LSC, School of Mathematical Sciences, CMA-Shanghai, SJTU-Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University;
[2]Institute for Interdisciplinary Information Sciences, Tsinghua University

Preliminary draft‡

May 9, 2023

Abstract

Treatment effect estimation under unconfoundedness is a fundamental task in causal inference. In response to the challenge of analyzing high-dimensional datasets collected in substantive fields such as epidemiology, genetics, economics, and social sciences, many methods for treatment effect estimation with high-dimensional nuisance parameters (the outcome regression and the propensity score) have been developed in recent years. However, it is still unclear what is the necessary and sufficient sparsity condition on the nuisance parameters for the treatment effect to be $\sqrt{n}$-estimable. In this paper, we propose a new Double-Calibration strategy that corrects the estimation bias of the nuisance parameter estimates computed by regularized high-dimensional techniques and demonstrate that the corresponding Doubly-Calibrated estimator achieves $\sqrt{n}$-rate as long as one of the nuisance parameters is sparse with sparsity below $\sqrt{n}/\log p$, where p denotes the ambient dimension of the covariates, whereas the other nuisance parameter can be arbitrarily complex and completely misspecified. The Double-Calibration strategy can also be applied to settings other than treatment effect estimation, e.g. regression coefficient estimation in the presence of diverging number of controls in a semiparametric partially linear model.

Keywords: Causal inference, Multi-calibration, Covariate balancing, High-dimensional statistics, Sparsity, Debiased lasso

1 Introduction

This article concerns the problem of efficient estimation of a parameter of scientific interest, denoted as $\tau \equiv \tau(\theta)$ under the semiparametric framework, where θ denotes the (potentially) high-dimensional nuisance parameters. A typical example of such parameters of interest is $\tau \equiv \mathbb{E}[Y(t)]$, which denotes the mean of a potential outcome $Y(t)$ with the binary treatment T set to $t \in \{0,1\}$ and is identifiable from the observed data under no unmeasured confounding [Imbens and Rubin, 2015, Hernán and Robins, 2023].

*email: linliu@sjtu.edu.cn

†email: yuhaow@tsinghua.edu.cn. LL and YW are ordered alphabetically. Correspondence should be addressed to YW: yuhaow@tsinghua.edu.cn.

‡This is a preliminary draft for YW's presentation at Online Causal Inference Seminar.

Since the average treatment effect (ATE) is just the contrast $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$, to simplify our exposition, we refer to $\tau = \mathbb{E}[Y(1)]$ as ATE and take $t = 1$ throughout. To estimate ATE under the no unmeasured confounding assumption, we need to account for the nuisance parameters θ, which can be decomposed into two components $\theta \equiv (\pi, r)$: the Propensity Score (PS), denoted as $\pi(x) := \mathbb{E}[T|X = x]$, and the Outcome Regression (OR) model, denoted as $r(x) := \mathbb{E}[Y|X = x, T = 1]$. Here X denotes the p-dimensional covariates. Although parameters other than ATE will also be covered in this paper, we mainly focus on ATE in the Introduction.

It is now well-established in modern semiparametric theory that the (nonparametric first-order) influence function [Robins et al., 1994] of τ is of the form

$$\mathsf{IF}(\theta) \equiv H(\theta) - \tau, \text{ where } H(\theta) = \frac{T}{\pi(X)}(Y - r(X)) + r(X), \tag{1}$$

which gives rise to the celebrated Augmented Inverse Probability Weighting (AIPW) estimator [Robins et al., 1994, Hahn, 1998]. The AIPW estimator is Doubly-Robust (DR) [Scharfstein et al., 1999, Robins and Rotnitzky, 2001] and is $\sqrt{n}$-consistent and asymptotically normal when π and r are estimated at sufficiently fast rates. To establish these convergence rates, it is essential to impose certain complexity-reducing assumptions [Robins and Ritov, 1997, Ritov et al., 2014, Liu et al., 2020, 2023] on the nuisance parameters π and r.

During the past decade, in response to the challenge of high-dimensional datasets with (ambient) dimension p potentially much greater than the sample size N, a large body of literature in statistics and econometrics (Farrell [2015], Shortreed and Ertefaie [2017], van der Laan [2017], Chernozhukov et al. [2018], Ju et al. [2020], Chernozhukov et al. [2022], Avagyan and Vansteelandt [2021], Ning et al. [2020], Tan [2020b,a], Tang et al. [2022], Athey et al. [2018], Hirshberg and Wager [2021], Bradic et al. [2019b,a], Wang and Shah [2020], Dukes and Vansteelandt [2021], Sun and Tan [2022] and references therein) has been devoted to solving this problem by positing π and/or r to be sparse Generalized Linear Models (GLMs): $\pi(x) \equiv \phi(\gamma^\top x)$, $r(x) \equiv \psi(\beta^\top x)$ where γ and/or β has only a few non-zero coordinates, and ϕ and ψ are two (possibly nonlinear) link functions. We further denote $s_\pi := \|\gamma\|_0$ and $s_r := \|\beta\|_0$ as the sparsities of π and r.

A natural solution in this context is to first estimate γ and β by high-dimensional regularized regression approaches, and then estimate τ by taking the sample average of $H(\theta)$, with θ replaced by the above estimates of γ and β. This route was indeed taken by some of the earlier works in this direction, such as Farrell [2015] and Chernozhukov et al. [2018][1]. In particular, the seminal Double Machine Learning (DML) estimator proposed in Chernozhukov et al. [2018] splits the entire sample into two non-overlapping subsets, one used to estimate γ, β by ℓ_1-regularized regression, denoted as $\hat{\gamma}, \hat{\beta}$, and the other used to estimate τ by taking the sample average of $H(\hat{\theta})$, denoted as $\hat{\tau}_{\text{DML}}$:

$$\hat{\tau}_{\text{DML}} := \frac{1}{n}\sum_{i=1}^{n} H_i(\hat{\theta}), \text{ where } \hat{\theta} = (\hat{\pi}, \hat{r}) \text{ and } \hat{\pi}(x) = \phi(\hat{\gamma}^\top x), \hat{r}(x) = \psi(\hat{\beta}^\top x). \tag{2}$$

Sample-splitting de-correlates the randomness in the nuisance parameter estimates $\hat{\theta}$ and in $\hat{\tau}_{\text{DML}}$, eliminating the reliance on Donsker-type conditions on θ, which is violated for high-dimensional sparse GLMs [Chernozhukov et al., 2018]. The information loss due to sample-splitting can then be restored by cross-fitting [Schick, 1986, Ayyagari, 2010, Zheng and van der Laan, 2011, Robins et al., 2013]. In this paper, we will also adopt the sample-splitting strategy, and estimate γ and β from a separate training sample (denoted

[1] Chernozhukov et al. [2018] alternatively interpreted this construction as Neyman-Orthogonalization, a terminology also adopted in later works such as Mackey et al. [2018] and Foster and Syrgkanis [2019].

as $\mathcal{D}_{\rm tr}$, see later Section 2.1) using ℓ_1-regularized generalized regression techniques. To simplify exposition, however, we omit the cross-fitting step. We refer readers to Chernozhukov et al. [2018], Jiang et al. [2022] for a more thorough discussion on the potential benefits of cross-fitting.

Since one typically estimates π and r at rates $\sqrt{s_\pi \log p/n}$ and $\sqrt{s_r \log p/n}$ [Bühlmann and van de Geer, 2011], it is straightforward to show that the DML estimator $\hat{\tau}_{\rm DML}$ is $\sqrt{n}$-consistent and asymptotically normal as long as[2]

$$s_\pi \cdot s_r \ll n/\log^2 p. \qquad \text{(DML sparsity)}$$

This condition on $s_\pi \cdot s_r$ could be quite stringent: e.g. if $s_\pi \asymp \sqrt{n}/\log p$, the (DML sparsity) condition implies $s_r \ll \sqrt{n}/\log p$. As a result, statisticians and econometricians have been puzzled for several years by the following question:

Problem*. *What is the sufficient and necessary, and hence minimal, sparsity condition on π and r for the parameter of interest, τ, to be $\sqrt{n}$-estimable; and can we construct a $\sqrt{n}$-consistent estimator of τ under this minimal sparsity condition?*

Literature overview and our contribution

Since Chernozhukov et al. [2018], to address Problem*, various attempts have been made to weaken the (DML sparsity) condition. Chernozhukov et al. [2022] showed that $s_\pi \wedge s_r \ll \sqrt{n}/\log p$ is sufficient for the DML estimator $\hat{\tau}_{\rm DML}$ to be $\sqrt{n}$-consistent, yet with an additional (strong) assumption that $\|\gamma\|_1$ and $\|\beta\|_1$ are bounded,imposing (potentially unnecessary) constraints on the magnitudes of non-zero coordinates of γ and β. Without the bounded ℓ_1-norm condition, by leveraging the idea of covariate balancing [Imai and Ratkovic, 2014, Zubizarreta, 2015, Chan et al., 2016, Athey et al., 2018, Ben-Michael et al., 2021, Bruns-Smith et al., 2023] and certain structures of the linear OR and logistic-linear PS models, Bradic et al. [2019b] relaxed the (DML sparsity) condition to $s_\pi \ll n/\log p, s_r \ll \sqrt{n}/\log p$ or $s_\pi \ll \sqrt{n}/\log p, s_r \ll n^{3/4}/\log p$. Similar result to that of Chernozhukov et al. [2018] was also obtained in Smucler et al. [2019] (the rate-double-robustness part)[3]. But the result of Smucler et al. [2019] is not only derived for ATE, but also for a large class of parameters – often called "DR functionals" [Liu et al., 2023] – with influence functions resembling the form in (1) [Rotnitzky et al., 2021].

Also exploiting the idea of covariate balancing, Athey et al. [2018] constructed a $\sqrt{n}$-consistent and asymptotically normal estimator when the OR model r is linear with sparsity $s_r \ll \sqrt{n}/\log p$, but allowing π to be arbitrarily complex, in which case, to avoid notation clutter, we set $s_\pi = \infty$. Hirshberg and Wager [2021] later generalized the results of Athey et al. [2018] from ATE to a strict subclass of doubly-robust functionals. Complementing the results of Athey et al. [2018], Wang and Shah [2020] obtained a $\sqrt{n}$-consistent and asymptotically normal estimator when the PS model π is logistic-linear[4] with sparsity $s_\pi \ll \sqrt{n}/\log p$, but allowing r to be arbitrarily complex, in which case we set $s_r = \infty$. This is desirable as, in applications such as clinical medicine, one often has a better grasp on the treatment assignment mechanism than on the outcome mechanism. The above results, taken together, suggest that the minimal sparsity condition is likely

$$s_\pi \wedge s_r \ll \sqrt{n}/\log p, s_\pi \vee s_r \text{ is arbitrary.} \qquad \text{(minimal sparsity)}$$

[2] Throughout this paper, we adopt the common asymptotic notation such as $\ll, \gg, \lesssim, \gtrsim, \asymp, o(\cdot), O(\cdot), o_{\shortmid}(\cdot), O_{\shortmid}(\cdot)$.

[3] It is worth noting that Chernozhukov et al. [2018], Smucler et al. [2019] and Bradic et al. [2019a] also considered the slightly more general approximately sparse GLMs. In order not to confuse readers, in the Introduction, we will not explicitly distinguish the approximately sparse GLMs from the exactly sparse GLMs and paraphrase the established results in the literature under exactly sparse GLMs.

[4] In fact, the results in Wang and Shah [2020] can also be generalised to nonlinear link functions other than expit. They have discussed this informally in their Section 5.1.

As an extension of [Athey et al., 2018], Bradic et al. [2019a] constructed an $\sqrt{n}$-consistent ATE estimator that additionally accommodates the setting where both the OR model and the best linear projection of the PS model is sparse. It shall be noted that they also exhibited two separate $\sqrt{n}$-consistent and asymptotically normal estimators for a different parameter of interest: the "regression coefficient", which we will discuss further in Remark 7 of Section 4.1. Bradic et al. [2019a] also established necessary conditions for $\sqrt{n}$-estimability of several parameters including ATE, which is believed to be tight. The guessed minimal sparsity condition is also inspired by this result. The difference is that the necessary condition of Bradic et al. [2019a] is stated in terms of how close the best linear approximation with roughly $\sqrt{n}$ many covariates approximates at least one of the two nuisance parameters.

Tan [2020a], Ning et al. [2020], Avagyan and Vansteelandt [2021], Dukes and Vansteelandt [2021] and Smucler et al. [2019] (the model-double-robustness part) also tackled this problem, striving for $\sqrt{n}$-consistency and asymptotic normal estimators under the (minimal sparsity) condition. However, these results are also not entirely satisfactory (at least theoretically) – although they allow one of two true nuisance parameters to be arbitrarily dense, the limits of their estimates by ℓ_1-regularized regression, $\hat{\pi}$ and $\hat{r}$, still need to be sparse.

As one can see, despite the tremendous progress made so far, a fully satisfactory solution to Problem* remains elusive. In this paper, we address Problem* by constructing novel $\sqrt{n}$-consistent and asymptotically normal estimators, under the (minimal sparsity) condition. This new methodology is based on a Double-Calibration (DCal) strategy to correct the bias of the DML estimator $\hat{\tau}_{\text{DML}}$. The basic idea of DCal is also rooted in covariate balancing, which will be elucidated in detail in Section 2. It is noteworthy that the idea of using calibration to correct for the estimation bias of PS and OR has also been used in the regularized calibrated estimator of Tan [2020a]. The difference is that Tan [2020a] estimated the nuisance parameters and calibrated the estimation bias simultaneously, but our approach first proposes initial estimates of PS and OR, and then calibrates these estimates via solving some optimization problems. Our new DCal estimators do not need to know which one of the two nuisance parameters, π or r, is a GLM with sparsity below $\sqrt{n}/\log p$. Furthermore, the DCal estimators are not only applicable to the ATE, but also work for the regression coefficient, answering an open question raised in Bradic et al. [2019a]. To the best of our knowledge, for both ATE and regression coefficient, no alternative estimators have been constructed before, that are $\sqrt{n}$-consistent and asymptotically normal under the (minimal sparsity) condition. Table 1 directly compares some of the aforementioned results with ours. However, we want to point out an important caveat of our theoretical results – we only establish $\sqrt{n}$-consistency and asymptotic normality of the DCal estimator under the (minimal sparsity) condition, but not its semiparametric efficiency [Newey, 1990, Janková and van de Geer, 2018]. This remains an important question for future work.

Notation and organization of the paper

Before moving forwards, we outline some notation that is used throughout. We denote θ as the nuisance parameters. All expectations and probabilities (e.g. $\mathbb{E}$ and $\mathbb{P}$) are understood to be under the true data generating distribution unless stated otherwise (in that case, we will attach θ in the subscripts (e.g. $\mathbb{E}_\theta$ and $\mathbb{P}_\theta$) to emphasize the dependence of the distribution on θ). As will be clear later in Section 2, we may split the entire dataset into three parts, a main dataset $\mathcal{D}$ of size n, an auxiliary dataset $\mathcal{D}_{\text{aux}}$ of size n_{aux}, and a training dataset $\mathcal{D}_{\text{tr}}$ of size n_{tr}. Bold letters will be preserved for to observed data or related quantities over a given dataset. For instance, $\mathbf{T} \equiv (T_1, \cdots, T_n)^\top$ and $\mathbf{X} = (X_1^\top, \cdots, X_n^\top)^\top$, respectively, denote the treatment variables and the covariates over the main dataset. The quantities computed from a given dataset will be attached, in the subscripts, by the index of that dataset. For example, the $n_{\text{tr}} \times p$ covariate matrix from the training dataset will be denoted as $\mathbf{X}_{\text{tr}}$, and similarly for $\mathbf{X}_{\text{aux}}$. Finally, given two vectors v_1, v_2 of the same length, $v_1 \odot v_2$ denotes the component-wise multiplication between v_1 and v_2.

The rest of this paper is organized as follows. In Section 2, we illustrate our new methodology for ATE, with the theoretical analysis detailed in Section 3. Section 4 extends our new approach to (1) the regression coefficient estimation in a high-dimensional partially linear model and (2) the approximately sparse GLMs considered in Belloni et al. [2014], Chernozhukov et al. [2018], Bradic et al. [2019a] and Smucler et al. [2019]. Lastly, Section 5 concludes this article and discusses some future directions. Some of the technical details are deferred to the Appendix.

Paper	condition(s) on s_π and s_r
Chernozhukov et al. [2018]	$s_\pi s_r = o(n/\log^2 p)$
Chernozhukov et al. [2022]	$s_\pi \wedge s_r = o(\sqrt{n}/\log p)$ but $\|\beta\|_1$ and $\|\gamma\|_1$ are bounded
Smucler et al. [2019] rate-double-robustness	$s_\pi s_r = o(n/\log^2 p)$
Bradic et al. [2019b]	$s_\pi = o(n/\log p), s_r = o(\sqrt{n}/\log p)$ or $s_\pi = o(\sqrt{n}/\log p), s_r = o(n^{3/4}/\log p)$
Athey et al. [2018]	$s_r = o(\sqrt{n}/\log p)$, s_π arbitrary
Wang and Shah [2020]	$s_\pi = o(\sqrt{n}/\log p)$, s_r arbitrary
Smucler et al. [2019] model-double-robustness	$s_\pi \wedge s_r = o(\sqrt{n}/\log p)$, $s_\pi \vee s_r$ arbitrary but the estimator of the denser model has sparse limit
Our result	$s_\pi \wedge s_r = o(\sqrt{n}/\log p)$, $s_\pi \vee s_r$ arbitrary

Table 1: A comparison between our theoretical results and other related works on ATE estimation. Note that, for example, "s_π arbitrary" means that s_π can be as large as ∞, namely that π does not necessarily need to follow a GLM. In fact, it can be arbitrarily complex. The results of Bradic et al. [2019a] are difficult to be accommodated into this table due to the limited space and we refer interested readers directly to our literature overview for details.

2 Average treatment effects: The Double-Calibration estimator

In this section, we illustrate our new "Double-Calibration (DCal)" methodology for ATE estimation. We observe data $\{O_i := (X_i, T_i, Y_i)\}_{i=1}^N$ that are i.i.d. draws from some underlying data-generating distribution $\mathbb{P}_{\theta^*}$ where from this section forward $\theta^* = (\pi^*, r^*)$ denotes the true OR and PS models. We assume throughout that, as standard in the literature on treatment effect estimation, $\mathbb{P}_{\theta^*}$ satisfies the unconfoundedness assumption:

Condition 1.

$$Y_i(t) \perp\!\!\!\perp T_i \mid X_i \textit{ almost surely, for } t \in \{0, 1\},$$

and the overlap/positivity assumption:

Condition 2. *There is an absolute constant $c_\pi \in (0, 0.5)$, such that $c_\pi < \pi^*(X) < 1 - c_\pi$ almost surely.*

It is well-known that, under Consistency, Conditions 1 and 2 [Hernán and Robins, 2023], the parameter of interest, ATE, can be identified as $\tau^* \equiv \mathbb{E}[r^*(X)] \equiv \mathbb{E}\left[\frac{TY}{\pi^*(X)}\right] \equiv \mathbb{E}[H(\theta^*)]$. We further impose the following modeling assumption on the nuisance parameters $\theta^* = (\pi^*, r^*)$:

Condition 3. *There exist two (nonlinear) monotonically increasing, twice-differentiable link functions ϕ, ψ with uniformly bounded first and second derivatives, such that* **either** *of the following holds:*

(i) The OR model follows a GLM with (nonlinear) link ψ: $r_i^* \equiv r^*(X_i) := \mathbb{E}(Y_i \mid X_i, T_i = 1) \equiv \psi(X_i^\top \beta^*)$;

(ii) The PS model follows a GLM with (nonlinear) link ϕ: $\pi_i^* \equiv \pi^*(X_i) := \mathbb{P}(T_i = 1 \mid X_i) \equiv \phi(X_i^\top \gamma^*)$.

Since $\pi^* \in (0, 1)$, we also restrict ϕ such that $\phi(t) \to 1$ as $t \to \infty$, and $\phi(t) \to 0$ as $t \to -\infty$.

In words, Condition 3 states that at least one of the two nuisance parameters is truly a GLM, whereas the other one can be arbitrarily complex. We restrict the range of ϕ to respect the fact that π is a probability, but this is not essential for the theoretical results of this paper.

In the rest of this section and Section 3, we consider to have access to three separate datasets: the main dataset $\mathcal{D} := \{(X_i, T_i, Y_i)\}_{i=1}^n$, the auxiliary dataset $\mathcal{D}_{\text{aux}} := \{(X_{\text{aux},i}, T_{\text{aux},i}, Y_{\text{aux},i})\}_{i=1}^{n_{\text{aux}}}$, and the training dataset $\mathcal{D}_{\text{tr}} := \{(X_{\text{tr},i}, T_{\text{tr},i}, Y_{\text{tr},i})\}_{i=1}^{n_{\text{tr}}}$ with $n \asymp n_{\text{aux}} \asymp n_{\text{tr}}$. Such a setup is equivalent to splitting the entire sample with size $N = n + n_{\text{aux}} + n_{\text{tr}}$ into three non-overlapping subsets. As mentioned in the Introduction, cross-fitting can restore the information loss due to sample splitting, but we choose to omit this step to simplify the exposition. Under this simplified observation scheme, we instead consider the target parameter of interest as the ATE averaged over the main dataset $\mathcal{D}$[5]:

$$\bar{\tau}^* := \frac{1}{n} \sum_{i=1}^n r^*(X_i). \tag{3}$$

2.1 Preliminary: Singly-calibrating the OR model

To lay the ground, we start by considering the following estimator that "singly" calibrates the OR model, denoted as $\hat{\tau}_{\text{SCal},r}$:

$$\hat{\tau}_{\text{SCal},r} := \frac{1}{n} \sum_{i=1}^n H_i(\hat{\pi}, \tilde{r}) \equiv \frac{1}{n} \sum_{i=1}^n \left(\frac{T_i(Y_i - \tilde{r}_i)}{\hat{\pi}_i} + \tilde{r}_i \right), \text{ where } \tilde{r}_i := \hat{r}_i + \hat{\mu}_i.$$

Recall that $\hat{r}(x) := \psi(x^\top \hat{\beta})$ is simply the plug-in GLM OR estimate with $\hat{\beta}$ estimated from the training dataset $\mathcal{D}_{\text{tr}}$. For the estimated PS model $\hat{\pi}$, we change the definition slightly and apply a trimming operation on the index $x^\top \hat{\gamma}$:

$$\hat{\pi}(x) := \phi\left(x^\top \hat{\gamma} \mathbb{1}\{|x^\top \hat{\gamma}| \leq M_\gamma\} + \text{sign}(x^\top \hat{\gamma}) M_\gamma \mathbb{1}\{|x^\top \hat{\gamma}| > M_\gamma\} \right),$$

where sign$(\cdot)$ is the function that returns the sign of the input variable. As in the OR case, $\hat{\gamma}$ is also estimated from the dataset $\mathcal{D}_{\text{tr}}$; M_γ is either chosen a priori or is chosen adaptively from $\mathcal{D}_{\text{tr}}$. With such trimming operation, the overlap condition always holds for the estimated PS model. When the PS model is correctly specified and follows some sparsity constraint, one can prove that with probability converging to 1, $\hat{\pi}_i \equiv \phi(X_i^\top \hat{\gamma})$, i.e., such trimming is moot (see e.g. Section 3.2); indeed, the trimming is only helpful when the

[5] It is also legitimate to choose two other alternatives: $\bar{\tau}^{*,1} := n^{-1} \sum_{i=1}^n T_i Y_i / \pi^*(X_i)$ or $\bar{\tau}^{*,2} := n^{-1} \sum_{i=1}^n H_i(\theta^*)$. We decided to go with $\bar{\tau}^*$ only for convenience.

PS model is misspecified. We now turn to $(\hat{\mu}_1, \cdots, \hat{\mu}_n)^\top =: \hat{\boldsymbol{\mu}}$, which is the solution to the following constrained program:

$$\hat{\boldsymbol{\mu}} = \mathrm{argmin}_{\boldsymbol{\mu} \in \mathbb{R}^n} \frac{1}{n} \|\boldsymbol{\mu}\|_2^2 \tag{4}$$

$$\text{s.t.} \quad \left\| \frac{1}{n_{\text{aux}}} \mathbf{X}_{\text{aux}}^\top \mathbf{\Pi}_{\text{aux}} \tilde{\mathbf{R}}_{\text{aux}} - \frac{1}{n} \mathbf{X}^\top \mathbf{\Pi} \boldsymbol{\mu} \right\|_\infty \leq \eta_r, \tag{5}$$

$$\|\boldsymbol{\mu}\|_\infty \leq M_r \tag{6}$$

where $M_r > 0$ is some known constant that depends on the bound on $\hat{r}$ and r^*, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the covariate matrix, $\mathbf{\Pi} \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose (i, i)-th entry is $\phi'(X_i^\top \hat{\gamma}) / \hat{\pi}_i$ and $\tilde{\mathbf{R}}$ is the n-dimensional residual vector whose i-th element is $\tilde{R}_i := T_i / \hat{\pi}_i \cdot (Y_i - \hat{r}_i)$. Recalling from the Notation section, the quantities $\mathbf{X}_{\text{aux}}$ etc. are defined analogously on the auxiliary dataset $\mathcal{D}_{\text{aux}}$.

Remark 1. $\hat{\tau}_{\text{SCal},r}$ is the same as the DML estimator $\hat{\tau}_{\text{DML}}$ except $\hat{r}$ is replaced by the post-calibration OR estimator $\tilde{r} = \hat{r} + \hat{\mu}$. It is in fact a variant of the recently proposed DIPW estimator for ATE [Wang and Shah, 2020]. Constraint (5) balances $\tilde{r}$ such that it is on average close to Y, gauged by covariates X under certain data-dependent weights. The use of the auxiliary dataset $\mathcal{D}_{\text{aux}}$ is to preserve the following moment equation satisfied by the calibrated OR estimator $\tilde{r}$:

$$\frac{1}{n} \sum_{i=1}^n \left(\frac{T_i}{\pi^*(X_i)} - 1 \right) \hat{\mu}_i, \text{ and thus } \frac{1}{n} \sum_{i=1}^n \left(\frac{T_i}{\pi^*(X_i)} - 1 \right) \tilde{r}_i \text{ have mean zero,} \tag{7}$$

because $\hat{\mu}$'s, as constructed, only depends on X, but not on T or Y, from the main dataset. Note that (7) in fact holds with any quantity, that is independent of T_i given X_i from the main dataset $\mathcal{D}$, in place of $\hat{\mu}_i$. The above moment equation plays a key role in the analysis of the statistical property of $\hat{\tau}_{\text{SCal},r}$ in Wang and Shah [2020]. ■

Using the analysis in the proof of Wang and Shah [2020, Theorem 2], since $\hat{\beta}$ and $\hat{\gamma}$, computed using $\mathcal{D}_{\text{tr}}$, are independent from the main sample $\mathcal{D}$, by choosing $\eta_r \asymp \sqrt{\frac{\log p}{n}}$, $\hat{\tau}_{\text{SCal},r}$ can be shown to be a $\sqrt{n}$-consistent estimator of $\bar{\tau}^*$ whenever the PS-sparsity $s_\pi \ll \sqrt{n} / \log p$, under Conditions 1 to 3.

Nevertheless, to achieve $\sqrt{n}$-rate under the (minimal sparsity) condition, $\hat{\tau}_{\text{SCal},r}$ needs to be modified so that it is also $\sqrt{n}$-consistent if the OR-sparsity $s_r \ll \sqrt{n} / \log p$ but the PS model is arbitrary. In the next Section 2.2, we will complete this remaining piece.

2.2 Double calibration

$\hat{\boldsymbol{\mu}}$ corrects the OR model misspecification in the initial OR estimate $\hat{r}$ and $\tilde{r} = \hat{r} + \hat{\mu}$ can be viewed as the post-calibrated OR estimator. To correct the misspecification of the PS model, we also need to calibrate the initial PS estimate $\hat{\pi}$. To that end, we propose the following Doubly-Calibrated (DCal) estimator for ATE:

$$\hat{\tau}_{\text{DCal}} := \frac{1}{n} \sum_{i=1}^n H_i(\tilde{\pi}, \tilde{r}) \equiv \frac{1}{n} \sum_{i=1}^n \left(\frac{T_i (Y_i - \tilde{r}_i)}{\tilde{\pi}_i} + \tilde{r}_i \right).$$

Apparently, it is the same as the OR-calibrated estimator proposed in the previous section, except that $\hat{\pi}_i$ is replaced by $\tilde{\pi}_i = \phi(\tilde{X}_i^\top \tilde{\gamma})$, where $\tilde{X}_i$ is the same as X_i if $p \geq n$, or $\tilde{X}_i$ is equal to an n-dimensional vector with X_i augmented by some synthetic covariates of dimension $n - p$ if $p < n$. For convenience, here the synthetic covariates are randomly generated from the uniform distribution over the interval $[-1, 1]$ but other strategies are also possible. This covariate augmentation ensures that almost surely there is a vector $\tilde{\gamma}^*$ such that π_i^*, the true PS evaluated at the i-th sample, satisfies $\pi_i^* = \phi(\tilde{X}_i^\top \tilde{\gamma}^*)$, no matter whether or not π^* follows a GLM with link ϕ. Writing $\mathbf{X}_j, \tilde{\mathbf{X}}_j$ as the j-th column of data matrices $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times (n \vee p)}$, similar to $\hat{\boldsymbol{\mu}}$, $\tilde{\gamma}$ solves the following constrained program:

$$
\begin{aligned}
&\tilde{\gamma} \coloneqq \operatorname{argmin}_\gamma \|\gamma - \hat{\gamma}\|_1 && (8)\\
\text{s.t. } &\left| \frac{1}{n} \sum_{i=1}^{n} \left(\frac{T_i}{\pi_i} - 1 \right) \psi'(X_i^\top \hat{\beta}) X_{i,j} \right| \leq \eta_{\pi_1} \cdot \frac{\|\psi'(\mathbf{X}\hat{\beta}) \odot \mathbf{X}_j\|_2}{\sqrt{n}} \quad \forall\, 1 \leq j \leq p && (9)\\
&\left| \frac{1}{n} \sum_{i=1}^{n} \left(\frac{T_i}{\pi_i} - 1 \right) \hat{\mu}_i \right| \leq \eta_{\pi_1} \cdot \frac{\|\hat{\boldsymbol{\mu}}\|_2}{\sqrt{n}} && (10)\\
&\left| \frac{1}{n} \sum_{i=1}^{n} \left(\frac{T_i}{\pi_i} - 1 \right) \tilde{X}_{i,j} \right| \leq \eta_{\pi_2} \cdot \frac{\|\tilde{\mathbf{X}}_j\|_2}{\sqrt{n}} \quad \forall\, 1 \leq j \leq p \vee n && (11)\\
&\max_i T_i/\pi_i \leq M_\pi. && (12)
\end{aligned}
$$

In this program, $\pi_i \equiv \pi(X_i) \coloneqq \phi(\tilde{X}_i^\top \gamma)$ and $M_\pi > 0$ is some known constant that depends on c_π[6].

Without any assumptions on the sparsity of the PS or the OR model, (8) is always guaranteed to have a feasible solution:

Lemma 1. *If Condition 2 holds and almost surely,* $\mathbf{X}$ *is of full rank, then with* $\eta_{\pi_1} \asymp \sqrt{\frac{\log p}{n}}, \eta_{\pi_2} \asymp \sqrt{\frac{\log p \vee n}{n}}$ *large enough, and with* $M_\pi \geq c_\pi^{-1}$*, then with probability converging to* 1*, any* $\tilde{\gamma}^*$ *that satisfies* $\pi_i^* \equiv \phi(\tilde{X}_i^\top \tilde{\gamma}^*)$ *for all* i *is a feasible solution to* (8)*.*

Note that in Lemma 1, we say that "$\mathbf{X}$ is full rank" whenever the rank of $\mathbf{X}$ equals $\min\{n, p\}$, i.e., whenever it is of full row/column rank. When the program is not feasible, one can simply set γ as $\hat{\gamma}$. Noteworthy, Lemma 1 does not necessarily require π^* follow a GLM with link ϕ. In fact, it can be any function satisfying Condition 2.

Remark 2. $\hat{\tau}_{\text{DCal}}$ is the same as $\hat{\tau}_{\text{SCal},r}$ except now $\hat{\pi}$ is also replaced by the post-calibration PS estimator $\tilde{\pi}(x) = \phi(\tilde{\gamma}^\top x)$. Constraints (9) to (11) in the program (8) balances $\tilde{\pi}$ such that $T/\tilde{\pi}$ is on average close to 1, gauged by covariates $\mathbf{X}$ (and the augmented covariates $\tilde{\mathbf{X}}$) and the OR calibrator $\hat{\boldsymbol{\mu}}$ under certain data-dependent weights. There is no need to use the auxiliary dataset in the constrained program (8) because the following moment equation holds as long as $\tilde{\pi}_i$ does not depend on Y_i:

$$
\frac{1}{n} \sum_{i=1}^{n} \frac{T_i}{\tilde{\pi}(X_i)} (Y_i - r_i^*) \text{ has mean zero.} \tag{13}
$$

Obviously, the program (8) is unrelated to Y from the main dataset. The moment equations (13) and (7) in Remark 1 are direct consequence of the double-robustness of the influence function $\mathsf{IF}(\theta)$ defined in (1). ■

3 Average treatment effects: theoretical analysis

In this section, we derive the asymptotic result for the proposed DCal estimator $\hat{\tau}_{\text{DCal}}$. For simplicity of exposition, throughout this section and the next section, we consider an asymptotic regime where the di-

[6] In (9), we use the convention that a function $g : \mathbb{R}^p \mapsto \mathbb{R}$ applied to a matrix $\boldsymbol{M} \in \mathbb{R}^{n \times p}$ (p could be equal to 1) is a n-dimensional vector formed by applying g to each row of the matrix.

mension p grows with n and $n \equiv n_{\text{aux}} \equiv n_{\text{tr}}$. We first prove that $\hat{\tau}_{\text{DCal}}$ is $\sqrt{n}$-consistent when the OR follows a sparse GLM model in Section 3.1, followed by a proof of its $\sqrt{n}$-consistency when the PS follows a sparse GLM model in Section 3.2.

3.1 Theoretical properties under the sparse OR model

Recall that our target of interest is the ATE over the main sample $\bar{\tau}^*$ in (3). In this section, we prove the $\sqrt{n}$-consistency of $\hat{\tau}_{\text{DCal}}$ for estimating $\bar{\tau}^*$ when only the OR model is sparse. Throughout Section 3, we assume the following regularity conditions.

Condition 4. *For some constant $m_Y > 0$, almost surely, $|Y(1)|, |r^*(X)| \leq m_Y$.*

We also invoke the following standard condition on the design matrix $\mathbf{X}$:

Condition 5. *The first component of X is 1, representing an intercept term. Denoting by $Z \in \mathbb{R}^{p-1}$ the remaining components of X, we assume $\mathbb{E}Z = 0$ and almost surely, $\|Z\|_\infty \leq m_Z$. Moreover, $\mathbf{X}$ is full rank and there exists a constant σ_Z such that for any unit vector $u \in \mathbb{R}^{p-1}$ and all $\alpha \in \mathbb{R}$, $\mathbb{E}\{\exp(\alpha u^\top Z)\} \leq \exp(\alpha^2 \sigma_Z^2/2)$.*

As we have mentioned in the comments after Lemma 1, we say that "$\mathbf{X}$ is full rank" if the rank of $\mathbf{X}$ equals $\min\{n, p\}$. In addition, we need the following conditions on the nuisance parameter estimates computed from the training and the auxiliary datasets. Recall that $\hat{\pi}_{\text{aux}}$ and $\hat{r}_{\text{aux}}$ appears in the constraint (5) when calibrating the OR model.

Condition 6. *There exists a constant $m_r > 0$ such that with probability converging to 1, for all i,*

$$|\hat{r}_i| \leq m_r \quad \& \quad |\hat{r}_{\text{aux},i}| \leq m_r.$$

Moreover, we impose the following assumption on the estimated OR coefficient $\hat{\beta}$ computed from the training sample $\mathcal{D}_{\text{tr}}$:

Condition 7. *Condition 3(i) holds true and there exists a constant m_ψ such that for any t, $\psi'(t), \psi''(t) \leq m_\psi$. Moreover, β^* has sparsity s_r and with probability converging to 1,*

$$\|\hat{\beta} - \beta^*\|_1 = O\left(s_r \sqrt{\frac{\log p}{n}}\right) \quad \& \quad \|\hat{\beta} - \beta^*\|_2 = O\left(\sqrt{s_r \frac{\log p}{n}}\right).$$

We then have the following important observation (Lemma 2): when the initial OR estimate $\hat{r}$ is close to the true OR r^*, correspondingly $\|\hat{\boldsymbol{\mu}}\|_2$ is small. This observation tells us that the OR-calibration is harmless even if it is superfluous when $\hat{r}$ is already close to r^*. This is appealing because in practice, we generally have no idea if $\hat{r}$ converges to r^* at a sufficiently fast rate even under the (minimal sparsity) condition.

Lemma 2. *Suppose Conditions 1–2 and 4–7 hold with $s_r = o(\sqrt{n}/\log p)$, then by choosing $\eta_r \asymp \sqrt{\frac{\log p}{n}}$, $M_r, M_\gamma \asymp 1$ sufficiently large, we have that $\|\hat{\boldsymbol{\mu}}\|_2 = o_{\mathbb{P}}(n^{1/4})$.*

The claim of Lemma 2 is expected as $\hat{\boldsymbol{\mu}}$, by design, calibrates the estimation error $r_i^* - \hat{r}_i$. When $\hat{r}_i$ estimates r_i^* with high accuracy, intuitively one should expect the calibrator $\hat{\boldsymbol{\mu}}$ to be negligible. Now equipped with the above result, using Taylor expansion, we can decompose $\hat{\tau}_{\text{DCal}} - \bar{\tau}^*$ as

$$\begin{aligned}&\hat{\tau}_{\text{DCal}} - \bar{\tau}^* \\ &\quad\approx \frac{1}{n}\sum_{i=1}^n \left(\frac{T_i}{\tilde{\pi}_i} - 1\right)\psi'(X_i^\top \hat{\beta})X_i^\top(\hat{\beta} - \beta^*) + \frac{1}{n}\sum_{i=1}^n \left(\frac{T_i}{\tilde{\pi}_i} - 1\right)\hat{\mu}_i + \frac{1}{n}\sum_{i=1}^n \frac{T_i \varepsilon_i(1)}{\tilde{\pi}_i},\end{aligned}$$

where $\varepsilon_i(1) := Y_i(1) - r^*(X_i)$ is the residual noise of the random variable $Y_i(1)$. As explained in Remark 2, by design of the program (8), the last term is mean zero and asymptotically normal. Therefore the first two terms constitute the bias. By applying Hölder's inequality and using the constraint of $\tilde{\pi}_i$ in (9) and (10), we have that the first two terms are, with probability converging to one, of order

$$s_r \frac{\log p}{n} + \frac{\sqrt{\log p}}{n} \|\hat{\boldsymbol{\mu}}\|_2.$$

Now in light of Lemma 2, we further have that the bias is of order

$$s_r \frac{\log p}{n} + \sqrt{s_r} \frac{\log p}{n},$$

which is asymptotically negligible whenever $s_r = o(\sqrt{n}/\log p)$. Putting together the above arguments, we establish Theorem 3 below.

Theorem 3. *Suppose Conditions 1–2 and 4–7 hold with $s_r = o(\sqrt{n}/\log p)$, by choosing $\eta_r, \eta_{\pi_1} \asymp \sqrt{\frac{\log p}{n}}, \eta_{\pi_2} \asymp \sqrt{\frac{\log p \vee n}{n}}, M_r, M_\pi, M_\gamma \asymp 1$ sufficiently large, we have the following representation*

$$\sqrt{n}(\hat{\tau}_{DCal} - \bar{\tau}^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{T_i \varepsilon_i(1)}{\tilde{\pi}_i} + o_{\mathbb{P}}(1).$$

Let $\bar{\sigma}_r^2 := n^{-1} \sum_{i=1}^{n} T_i \mathrm{var}(\varepsilon_i(1) \mid X_i)/\tilde{\pi}_i^2$. Informally, Theorem 3 states that with sufficiently large n, with high probability, we can obtain by conditioning on $\{(\mathbf{X}, \mathbf{T}), \mathcal{D}_{\text{aux}}, \mathcal{D}_{\text{tr}}\}$ that

$$\sqrt{n}(\hat{\tau}_{\text{DCal}} - \bar{\tau}^*) \mid \{(\mathbf{X}, \mathbf{T}), \mathcal{D}_{\text{aux}}, \mathcal{D}_{\text{tr}}\} \approx \mathcal{N}(0, \bar{\sigma}_r^2).$$

Once we further have the assumption that almost surely, $\mathrm{var}(\varepsilon_i(1) \mid X_i)$ is bounded above by some constant, we have that with probability converging to 1, $\bar{\sigma}_r$ is $O(1)$. This means that $\hat{\tau}_{\text{DCal}}$ is a $\sqrt{n}$-consistent estimator for $\bar{\tau}^*$. In fact, our convergence guarantee is quite similar to the one given by Athey et al. [2018], except that they replaced $1/\tilde{\pi}_i$ by some weights learned by a convex program.

Remark 3. As mentioned in the Introduction, one missing piece of Theorem 3 is if the asymptotic linear representation can achieve the semiparametric efficiency bound (with respect to the main dataset $\mathcal{D}$), which requires the representation to be of the following form:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{T_i}{\pi_i^*} \varepsilon_i(1) + r_i^* - \tau^* \right) + o_{\mathbb{P}}(1).$$

We conjecture that it should be the case under the additional assumption that π^* is consistently estimable by $\tilde{\pi}$ when only the OR is assumed to be sparse. ■

3.2 Theoretical properties under the sparse PS model

In this section, we consider the other way around, namely when the PS model satisfies a sparse GLM. Specifically, we impose the following assumption on the estimated PS coefficient $\hat{\gamma}$ computed from the training sample $\mathcal{D}_{\text{tr}}$:

Condition 8. *Condition 3(ii) holds true and there exists a constant m_ϕ such that for any t, $\phi'(t), \phi''(t) \leq m_\phi$. Moreover, γ^* has sparsity s_π and with probability converging to 1, the estimated $\hat{\gamma}$ satisfies that*

$$\|\hat{\gamma} - \gamma^*\|_1 = O\left(s_\pi \sqrt{\frac{\log p}{n}}\right) \quad \& \quad \|\hat{\gamma} - \gamma^*\|_2 = O\left(\sqrt{s_\pi \frac{\log p}{n}}\right).$$

Once we further have that $s_\pi = o(\sqrt{n}/\log p)$ (which we will assume in the rest of this section), we have that with probability converging to 1,

$$\max_i |X_i^\top(\hat{\gamma} - \gamma^*)| = \max_i \|X_i\|_\infty \|\hat{\gamma} - \gamma^*\|_1 = o_{\mathbb{P}}(1).$$

This, together with Condition 2, implies that with probability converging 1, by choosing $M_\gamma \asymp 1$ large enough, $|X_i^\top \hat{\gamma}|, |X_{\text{aux},i}^\top \hat{\gamma}| \leq M_\gamma$, i.e., that with probability converging to 1, $\bar{\pi}_i \equiv \phi(X_i^\top \hat{\gamma})$ and $\bar{\pi}_{\text{aux},i} \equiv \phi(X_{\text{aux},i}^\top \hat{\gamma})$. This is helpful to derive the $\sqrt{n}$-consistency of $\hat{\tau}_{\text{DCal}}$.

We also impose the following constraint on the "lower bound" of ϕ':

Condition 9. *For all $u > 0$, there exists a constant $c_{\phi,u}$ depending on u such that*

$$\min_{|w| \leq u} |\phi'(w)| \geq c_{\phi,u}.$$

Condition 9 is satisfied by e.g. logistic and probit regression models. With these additional constraints on the sparsity and link functions, the calibrated $\tilde{\gamma}$ satisfies that

Lemma 4. *Suppose Conditions 1–2, 4–6 and Conditions 8–9 hold; suppose moreover that $s_\pi = o(\sqrt{n}/\log p)$. By choosing the tuning parameters as in Theorem 3, we have that with probability converging to 1,*

$$\|\tilde{\gamma} - \hat{\gamma}\|_1 = O\left(s_\pi \sqrt{\frac{\log p}{n}}\right) \quad \& \quad \frac{1}{n}\sum_{i=1}^n (\tilde{X}_i^\top(\tilde{\gamma} - \hat{\gamma}))^2 = O\left(s_\pi \frac{\sqrt{\log p \log p \vee n}}{n}\right).$$

Armed with this lemma, and recall that we already have $\hat{\tau}_{\text{SCal},r}$ is $\sqrt{n}$-consistent, it remains to analyze the difference between $\hat{\tau}_{\text{DCal}}$ and $\hat{\tau}_{\text{SCal},r}$. We now apply the following decomposition

$$\begin{aligned}
\hat{\tau}_{\text{DCal}} - \hat{\tau}_{\text{SCal},r} &= \frac{1}{n}\sum_{i=1}^n \left(\frac{T_i(Y_i - \hat{r}_i - \hat{\mu}_i)}{\tilde{\pi}_i} - \frac{T_i(Y_i - \hat{r}_i - \hat{\mu}_i)}{\hat{\pi}_i}\right) \\
&= \frac{1}{n}\sum_{i=1}^n \left(\frac{T_i(Y_i - \hat{r}_i)}{\tilde{\pi}_i\hat{\pi}_i} - \frac{\hat{\mu}_i}{\hat{\pi}_i}\right)(\hat{\pi}_i - \tilde{\pi}_i) + \frac{1}{n}\sum_{i=1}^n (T_i - \tilde{\pi}_i)\frac{\hat{\mu}_i}{\hat{\pi}_i\tilde{\pi}_i}(\hat{\pi}_i - \tilde{\pi}_i).
\end{aligned}$$

Now by Taylor expansion, and using that with probability converging to 1, $\hat{\pi}_i \equiv \phi(X_i^\top \hat{\gamma})$, we are able to bound the above display as follows:

$$\begin{aligned}
\hat{\tau}_{\text{DCal}} - \hat{\tau}_{\text{SCal},r} \lesssim &\left|\frac{1}{n}\sum_{i=1}^n \left(\frac{T_i(Y_i - \hat{r}_i)}{\tilde{\pi}_i\hat{\pi}_i} - \frac{\hat{\mu}_i}{\hat{\pi}_i}\right)\phi'(X_i^\top\hat{\gamma})\tilde{X}_i^\top(\tilde{\gamma} - \hat{\gamma})\right| \\
&+ \left|\frac{1}{n}\sum_{i=1}^n \left(\frac{T_i}{\tilde{\pi}_i} - 1\right)\frac{\phi'(X_i^\top\hat{\gamma})\hat{\mu}_i}{\hat{\pi}_i}\tilde{X}_i^\top(\tilde{\gamma} - \hat{\gamma})\right| + \frac{1}{n}\sum_{i=1}^n (X_i^\top(\tilde{\gamma} - \gamma^*))^2.
\end{aligned} \tag{14}$$

Using an analogous analysis as the bias term in Wang and Shah [2020] and that $\|\hat{\gamma} - \tilde{\gamma}\|_1 \lesssim s_\pi \sqrt{\frac{\log p}{n}}$ from Lemma 4, we are able to bound the first term to be of order $s_\pi \frac{\sqrt{\log p \log p \vee n}}{n}$. The third term can be controlled using Lemma 4 as well. Therefore the only missing piece is to control the second term of (14). To this end, we utilize the following identity $\frac{T_i}{\tilde{\pi}_i} - 1 = \frac{T_i}{\pi_i^*} - 1 + \frac{T_i}{\tilde{\pi}_i} - \frac{T_i}{\pi_i^*}$ to bound the second term as below:

$$\left| \frac{1}{n} \sum_{i=1}^{n} \left(\frac{T_i}{\pi_i^*} - 1 \right) \frac{\phi'(X_i^\top \hat{\gamma}) \hat{\mu}_i}{\hat{\pi}_i} \tilde{X}_i^\top (\tilde{\gamma} - \hat{\gamma}) \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \left(\frac{T_i}{\tilde{\pi}_i} - \frac{T_i}{\pi_i^*} \right) \frac{\phi'(X_i^\top \hat{\gamma}) \hat{\mu}_i}{\hat{\pi}_i} \tilde{X}_i^\top (\tilde{\gamma} - \hat{\gamma}) \right|$$
$$\lesssim \left\| \frac{1}{n} \sum_{i=1}^{n} \left(\frac{T_i}{\pi_i^*} - 1 \right) \frac{\phi'(X_i^\top \hat{\gamma})}{\hat{\pi}_i} \hat{\mu}_i \tilde{X}_i \right\|_\infty \|\tilde{\gamma} - \hat{\gamma}\|_1 + \frac{1}{n} \sum_{i=1}^{n} (X_i^\top (\tilde{\gamma} - \gamma^*))^2$$

where the last inequality follows from Hölder's inequality and Cauchy-Schwarz inequality; for a detailed derivation, we refer readers to Appendix A.5. Then observing that $\{T_i/\pi_i^* - 1, i = 1, \cdots, n\}$ are mean-zero sub-Gaussian random variables and are conditionally uncorrelated with $\tilde{X}_i$'s, the first term of the above display can be shown to be of order $s_\pi \frac{\sqrt{\log p \log p \vee n}}{\sqrt{n}}$. Based on the above arguments, we obtain Theorem 5 below.

Theorem 5. *Suppose Conditions 1–2, 4–6 and Conditions 8–9 hold with $s_\pi = o(\sqrt{n}/\sqrt{\log p \log p \vee n})$, then by choosing we the tuning parameters in the same way as in Theorem 3, we have the following representation*

$$\sqrt{n}(\hat{\tau}_{DCal} - \bar{\tau}^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{T_i \varepsilon_i(1)}{\pi_i^*} + \frac{(T_i - \pi_i^*)(r_i^* - \tilde{r}_i)}{\pi_i^*} \right) + o_{\mathbb{P}}(1).$$

We now write

$$\sigma_\mu^2 := \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - \pi_i^*)(r_i^* - \tilde{r}_i)^2}{\pi_i^*}, \quad \bar{\sigma}_\pi^2 := \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{var}(\varepsilon_i(1)^2 \mid X_i)}{\pi_i^*}.$$

Informally, Theorem 5 says that with sufficiently large n, with high probability, we can obtain by conditioning on $\{\mathbf{X}, \mathcal{D}_{\mathrm{aux}}, \mathcal{D}_{\mathrm{tr}}\}$ that

$$\sqrt{n}(\hat{\tau}_{\mathrm{DCal}} - \bar{\tau}^*) \mid \{\mathbf{X}, \mathcal{D}_{\mathrm{aux}}, \mathcal{D}_{\mathrm{tr}}\} \approx \mathcal{N}(0, \bar{\sigma}_\pi^2 + \sigma_\mu^2).$$

As also discussed in Theorem 3, under some standard regularity conditions, one can expect $\bar{\sigma}_\pi$ to be $O(1)$. For σ_μ^2, using the overlap condition as in Condition 2 and recall the ℓ_∞ constraint (6), one can expect σ_μ^2 to be of order $O(1)$. This proves the $\sqrt{n}$-consistency of our estimator under a sparse PS model. This asymptotic distribution is similar to the one in Wang and Shah [2020].

Remark 4. Similar to the comment in Remark 3, one missing piece of Theorem 5 is if the asymptotic linear representation can achieve the semiparametric efficiency bound (with respect to the main dataset $\mathcal{D}$). We again conjecture that it should be the case under the additional assumption that r^* is consistently estimable by $\tilde{r}$ when only the PS is assumed to be sparse. ■

To conclude, Theorem 3 and 5, taken together, address Problem* when $p > n$ as the sparsity constraint $s_\pi \ll \sqrt{n}/\sqrt{\log p \log p \vee n}$ reduces to $s_\pi \ll \sqrt{n}/\log p$. When $p \le n$, the DCal estimator achieves $\sqrt{n}$-consistency under the (minimal sparsity) condition up to a minor $\sqrt{\log n/\log p}$ factor, which is a constant if we take $p = n^\delta$ for some fixed $\delta \in (0, 1)$. We conjecture that this extra log factor can be removed by calibrating the PS in a similar fashion to the OR calibration program (4), which we leave to future work.

For the sake of completeness, a corresponding lower bound proof can be found in Appendix B but we acknowledge that the lower bound proof of Bradic et al. [2019a] for approximately sparse linear models can be adapted to prove the necessity of the (minimal sparsity) condition for the $\sqrt{n}$-estimability of $\bar{\tau}^*$.

Remark 5. The current analysis still requires that the observed data $O = (X, T, Y)$ be light-tailed. It is of great interest to investigate if the proposed approach can be extended into settings in which (part of) O has heavier tails [Kuchibhotla and Chakrabortty, 2022] or has even diverging higher-order moments (e.g. transformed covariates by B-splines or Cohen-Daubechies-Vial wavelets [Cohen et al., 1993, Belloni et al., 2015, Liu et al., 2017, Mukherjee and Sen, 2018, Liu et al., 2021])[7]. ■

4 Some extensions of the Double-Calibration approach

In this section, we extend the Double-Calibration approach for ATE estimation in two directions.

4.1 Regression coefficient

As mentioned in the Introduction, our Double-Calibration strategy can be extended to target parameters other than ATE. In this section, we demonstrate the broad applicability of our approach via another important parameter in statistics and econometrics. Specifically, we consider the following semiparametric partially linear model [Robinson, 1988]:

$$Y = T\tau^* + r^*(X) + \varepsilon, \quad \& \quad T = \pi^*(X) + e, \tag{15}$$

for which the parameter of interest is τ^* and the nuisance parameters, with a slight abuse of notations, are again denoted as $\theta^* = (r^*, \pi^*)$. For simplicity of exposition we assume that ε and e are noises with homoscedastic variances $\sigma_\varepsilon^2, \sigma_e^2$ respectively; moreover, we assume that the noise distributions do not change with the sample size n. Since τ^* is the parameter of interest, we also regard it as a fixed constant.

Given initial estimates $\hat{\tau}, \hat{r}$ and $\tilde{\pi}$ computed from the training dataset $\mathcal{D}_{\text{tr}}$, just as in the ATE case, we write $\hat{r}_i := \hat{r}(X_i), \tilde{\pi}_i := \tilde{\pi}(X_i)$. We propose the following estimator for the regression coefficient τ^*:

$$\hat{\tau}_{\text{DCal}} := \hat{\tau} + \frac{1}{n}\sum_{i=1}^{n} \frac{(T_i - \tilde{\pi}_i)(Y_i - T_i\hat{\tau} - \hat{r}_i - \hat{\mu}_i)}{\tilde{\sigma}_e^2},$$

where $\hat{\mu}$ is estimated by solving the same constrained quadratic program as in (5) for ATE, except that we replace the diagonal elements of $\mathbf{\Pi}$ with $\phi'(X_i\hat{\gamma})$ and $\tilde{R}_i$ with $Y_i - T_i\hat{\tau} - \hat{r}_i$; $\tilde{\sigma}_e^2$ is an estimated variance using $\tilde{\pi}$, i.e.,

$$\tilde{\sigma}_e^2 := \frac{1}{n}\sum_{i=1}^{n}(T_i - \tilde{\pi}_i)^2.$$

Remark 6. Here we deliberately choose not to write $\tilde{r}$ in place of $\hat{r} + \hat{\mu}$ as the purpose of $\hat{\mu}$ is to correct the estimation error of $T\hat{\tau} + \hat{r}$ as an estimator of $\pi^*\tau^* + r^*$. This is slightly different from the setting of ATE in Section 2 and 3. ■

We now turn to the nuisance function $\tilde{\pi}$, which is constructed by modifying the estimator in (8) slightly. Specifically, we keep the objective function and the first three constraints the same as before, except that

[7] Based on personal communications [Mukherjee, 2023], R. Mukherjee and collaborators have established nontrivial (and possibly tight) lower bounds for certain doubly-robust functionals in this context in an unpublished technical report.

the term $T_i/\pi_i - 1$ becomes $T_i - \pi_i$. As for the last constraint (12), we break it into the following two constraints:

$$\begin{aligned} &\max_i |\pi_i| \leq M_\pi, \quad \frac{1}{n}\sum_{i=1}^n (T_i - \pi_i)^2 \geq M_\pi^{-1}, \\ &\left| \frac{1}{n}\sum_{i=1}^n (T_i - \pi_i)\pi_i \right| \leq \eta_{\pi_1} \cdot \frac{\|\boldsymbol{\pi}\|_2}{\sqrt{n}}. \end{aligned} \tag{16}$$

To understand the theoretical property of the DCal estimator of the regression coefficient τ^*, we need to further modify the assumptions in Section 3. In particular, we adapt Conditions 4 and 6 to the following:

Condition 10. *We have that*

(i) For some constant $m > 0$, almost surely the absolute value of random variables $(\varepsilon, e, r^*(X), \pi^*(X))$ are bounded below by m;

(ii) There exists a constant $\hat{m} > 0$ such that with probability converging to 1, $|\hat{\tau}| \leq \hat{m}$ and that for all i,

$$\max\{|\hat{r}_i|, |\hat{\pi}_i|, |\hat{r}_{\text{aux},i}|, |\hat{\pi}_{\text{aux},i}|\} \leq \hat{m}.$$

With the above preparation, by mimicking the proofs of Theorems 3 and 5 (see Appendix A.6), we have the following results on the asymptotic statistical property of $\hat{\tau}_{\text{DCal}}$ for estimating the τ^* in the semiparametric partially linear model (15).

Theorem 6. *Under Conditions 5 and 10, by choosing $\eta_r, \eta_{\pi_1} \asymp \sqrt{\frac{\log p}{n}}, \eta_{\pi_2} \asymp \sqrt{\frac{\log p \vee n}{n}}, M_r, M_\pi \asymp 1$ sufficiently large, we have the following:*

(i) If Condition 7 holds, with the additional assumption that $|\hat{\tau} - \tau^*| = O(\sqrt{s_r \log p/n})$, then we have

$$\sqrt{n}(\hat{\tau}_{DCal} - \tau^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{(T_i - \tilde{\pi}_i)\varepsilon_i}{\tilde{\sigma}_e^2} + o_{\mathbb{P}}(1).$$

(ii) If Conditions 8 and 9 holds, then we have

$$\sqrt{n}(\hat{\tau}_{DCal} - \tau^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{e_i\varepsilon_i}{\sigma_e^2} + \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{e_i((\tau^* - \hat{\tau})\pi_i^* + r_i^* - \hat{r}_i - \hat{\mu}_i)}{\sigma_e^2} + o_{\mathbb{P}}(1).$$

We first discuss the intuitions of Theorem 6(i). With a slight abuse of notation, we write $\bar{\sigma}_r^2 := \frac{\text{var}(\varepsilon)}{\hat{\sigma}_\varepsilon^2}$. Analogous to the discussions below Theorem 3, informally, it says that

$$\sqrt{n}(\hat{\tau}_{\text{DCal}} - \tau^*) \mid \{(\mathbf{T}, \mathbf{X}), \mathcal{D}_{\text{aux}}, \mathcal{D}_{\text{tr}}\} \approx \mathcal{N}(0, \bar{\sigma}_r^2),$$

where the program for $\tilde{\gamma}$ ensures that with probability converging to 1, $\bar{\sigma}_r^2$ is $O(1)$. For Theorem 6(ii), using analogous discussions to those below Theorem 5, we have that

$$\sqrt{n}(\hat{\tau}_{\text{DCal}} - \tau^*) \mid \{(\mathbf{T}, \mathbf{X}), \mathcal{D}_{\text{aux}}, \mathcal{D}_{\text{tr}}\} \approx \mathcal{N}(0, \bar{\sigma}_\pi^2 + \sigma_\mu^2),$$

where with a slight abuse of notation, we redefine

$$\sigma_\mu^2 \coloneqq \frac{1}{n}\sum_{i=1}^{n}((\tau^* - \hat{\tau})\pi_i^* + r_i^* - \hat{r}_i - \hat{\mu}_i)^2 \quad \& \quad \bar{\sigma}_\pi^2 \coloneqq \sigma_\varepsilon^2.$$

Since all the summands in σ_μ^2 are bounded with probability converging to 1, σ_μ^2 is with probability converging to 1 of order $O(1)$, which in turn implies that $\hat{\tau}_{\text{DCal}}$ is a $\sqrt{n}$-consistent estimator of τ^*.

Remark 7. As alluded to in the Introduction, Bradic et al. [2019a] also studied the problem of estimating the regression coefficient in a semiparametric partially linear model with the nuisance parameters approximated by sparse linear models. However, Bradic et al. [2019a] constructed two separate $\sqrt{n}$-consistent estimators of τ^*, one under the assumption r^* is sparse, and the other under the assumption π^* is sparse. In the penultimate paragraph of the main text of Bradic et al. [2019a], they listed as an open question to construct one estimator that is agnostic to the identity of the sparse nuisance parameters. Here we demonstrate that our new DCal estimator solves this question at least in the exact sparsity case.

Our result is also related to the vast literature on debiased lasso [Zhang and Zhang, 2014, van de Geer et al., 2014, Javanmard and Montanari, 2014, 2018, Cai and Guo, 2017, Zhu and Bradic, 2018, Cattaneo et al., 2018, 2019]. We refer readers who are interested in a comparison between our result and the above references to Section 1 of Bradic et al. [2019a], which has comprehensively surveyed the conditions under which $\sqrt{n}$-consistent estimator of τ^* were developed in some of these referenced works. ■

Remark 8. Last but not least, we briefly comment on the computational issue of the above constrained program for calibrating regression coefficient estimation. Since the constraint (16) is non-convex, the time complexity of the resulted optimization problem is generally NP-hard. Note also that the above computational issue is also related to ATE estimation in Section 2 and 3. Depending on the link functions, the resulting optimization problem for Double-Calibration can also be non-convex. Smucler et al. [2019] also discussed extensively on the potential drawback of solving non-convex program, such as the lack of theoretical convergence guarantees. This is a very important and potentially difficult problem that does not catch enough attention in the causal inference literature. But in order not to lead readers astray from the main message of the paper, we leave this problem to future work. ■

4.1.1 Further explanation on the DCal estimator for the regression coefficient

We expect that it may be helpful to explain the DCal estimator for regression coefficient estimation from the perspective of the influence function of DR functionals [Rotnitzky et al., 2021]. Generically, the influence function of a DR functional τ with nuisance parameters $\theta = (a, b)$ has the following form:

$$\mathsf{IF} = H(\theta) - \tau, \text{ with } H(\theta) = S_1 a(Z) b(Z) + m_a(O, a) + m_b(O, b) + S_0,$$

where the observable is $O = (Z, W)$, and $a \mapsto m_a(O, a)$ and $b \mapsto m_b(O, b)$ are two linear maps that satisfy some extra regularity and moment conditions (which are to ensure IF is doubly robust). We refer readers to Rotnitzky et al. [2021] and Liu et al. [2023] for more details.

To see how this connects with the regression coefficient τ of T in (15), we first decompose $O = (Z, Y)$ with $Z = (T, X)$. Then an influence function of τ is:

$$\mathsf{IF} = H(\theta) - \tau, \text{ with } H(\theta) = -a(Z)b(Z) + a(Z)Y + \frac{\partial}{\partial T} b(Z), \text{ and } \frac{\partial}{\partial T} b(Z) \equiv \tau \text{ by (15)},$$

where $a(z) \coloneqq (t - \pi(x))/\mathbb{E}[(T - \pi(X))^2]$, $b(z) \coloneqq r(x) + t \cdot \tau$. Hence as the ATE, the regression coefficient τ is also a special case of the DR functionals, as obviously $a(Z)Y$ and $\partial b(Z)/\partial T$ are linear maps of a and

b respectively. Similarly, we denote $\theta^* = (a^*, b^*)$ as the true nuisance parameters. Similar to ATE, a natural estimator for τ is the DML estimator:

$$\hat{\tau}_{\text{DML}} \coloneqq \hat{\tau} + \frac{1}{n}\sum_{i=1}^{n} \hat{a}(Z_i)\left(Y_i - \hat{b}(Z_i)\right)$$

where $\hat{b}(z) \equiv \hat{r}(x) + \hat{\tau}t \equiv \psi(x^\top \hat{\beta}) + \hat{\tau}t$ is simply some variant of the lasso estimate with the coefficients $(\hat{\tau}, \hat{\beta}^\top)^\top$ computed from the training dataset $\mathcal{D}_{\text{tr}}$. As for $\hat{a}(z)$, as is done in this paper, one can simply first estimate $\hat{\gamma}$ to obtain $\hat{\pi}$ and then estimate the denominator $\mathbb{E}[(T - \pi^*(X))^2]$ by its DML estimator $n^{-1}\sum_{i=1}^{n}(T_i - \phi(X_i^\top \hat{\gamma}))^2$. To achieve $\sqrt{n}$-rate under $s_r \wedge s_\pi \ll \sqrt{n}/\log p$, one uses the Double-Calibration program to bias-correcting both $\hat{a}$ and $\hat{b}$ by using $\tilde{a}(Z) = (T - \phi(X^\top \tilde{\gamma})))/n^{-1}\sum_{i=1}^{n}(T_i - \phi(X_i^\top \tilde{\gamma}))^2$ and $\tilde{b} = \hat{b}(Z) + \hat{\mu}$ instead. By design of the Double-Calibration program, $\tilde{a}$ and $\tilde{b}$ again satisfy the following moment equations, similar to those in Remark 1 and 2:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial T_i}\tilde{b}_i - a_i^* \tilde{b}_i\right) \text{ and } \frac{1}{n}\sum_{i=1}^{n}\tilde{a}_i(Y_i - b_i^*) \text{ have mean zero.}$$

4.2 Extension to approximately sparse GLMs

In previous sections, we have assumed that at least one of the nuisance parameters, π^* or r^*, exactly follows a sparse GLM, with sparsity below $\sqrt{n}/\log p$. Belloni et al. [2014], Chernozhukov et al. [2018], Bradic et al. [2019a] and Smucler et al. [2019] considered the so-called approximately sparse GLMs to model the nuisance parameters. Similar to Condition 3, the condition below imposes that at least one of the two nuisance parameters, r^* or π^*, follows the approximately sparse GLM.

Condition 11. *There exist two (nonlinear) monotonically increasing, twice-differentiable link functions ϕ, ψ with uniformly bounded first and second derivatives, such that the following hold: for any positive integer k*

(i) The OR model can be approximated by a GLM with (nonlinear) link ψ using s covariates with approximation error decaying with rate $s^{-\xi_r}$ for some $\xi_r > 0$: there exists β_s^* such that $\|\beta_s^*\|_0 \leq s$ and $\mathbb{E}[(r^*(X) - \psi(\beta_s^{*\top} X))^2] \ll s^{-2\xi_r}$;

(ii) The PS model can be approximated by a GLM with (nonlinear) link ψ using s covariates with approximation error decaying with rate $s^{-\xi_r}$ for some $\xi_\pi > 0$: there exists γ_s^* such that $\|\gamma_s^*\|_0 \leq s$ and $\mathbb{E}[(\pi^*(X) - \phi(\gamma_s^{*\top} X))^2] \ll s^{-2\xi_\pi}$.

Since $\pi^* \in (0,1)$, we also restrict ϕ such that $\phi(t) \to 1$ as $t \to \infty$, and $\phi(t) \to 0$ as $t \to -\infty$.

We silence the dependence on s by denoting β^* and γ^* when $s = p$; in particular, without loss of generality, we take β^* and γ^* to minimize $\mathbb{E}[(r^*(X) - \psi(\beta^\top X))^2]$ and $\mathbb{E}[(\pi^*(X) - \phi(\gamma^\top X))^2]$. Bradic et al. [2019a] proved the following necessary condition for the existence of $\sqrt{n}$-consistent estimator of the ATE τ when $p \gtrsim n$:

$$\xi_r \vee \xi_\pi > 1/2 \text{ by choosing } s \asymp \sqrt{n/\log p}. \tag{17}$$

For simplicity, we take $p \asymp n^2$. Although it is possible to relax the lower bound requirement on p by imposing conditions similar to Assumption 9 of Bradic et al. [2019a], we decide not to further pursue this in this paper. As a corollary of the results in Section 3, the DCal estimator achieves $\sqrt{n}$-consistency under (17). To state this result, we need to modify the regularity conditions on the OR and PS coefficient estimates computed from the training dataset $\mathcal{D}_{\text{tr}}$ as follows. First, we need to modify Condition 7 as

Condition 12. *Condition 11(i) holds true and there exists a constant m_ψ such that for any t, $\psi'(t), \psi''(t) \leq m_\psi$. Moreover, when $\xi_r > 1/2$ and $\xi_\pi > 0$, with probability converging to 1, there exists $\hat{\beta}$ computed from $\mathcal{D}_{\mathrm{tr}}$ such that*

$$\|\hat{\beta} - \beta^*\|_1 = O\left(\left(\sqrt{\frac{n}{\log p}}\right)^{-\frac{2\xi_r - 1}{2\xi_r + 1}}\right) \quad \& \quad \|\hat{\beta} - \beta^*\|_2 = O\left(\left(\sqrt{\frac{n}{\log p}}\right)^{-\frac{2\xi_r}{2\xi_r + 1}}\right).$$

We then need to modify Condition 8 as

Condition 13. *Condition 11(ii) holds true and there exists a constant m_ϕ such that for any t, $\phi'(t), \phi''(t) \leq m_\phi$. Moreover, when $\xi_\pi > 1/2$ and $\xi_r > 0$, with probability converging to 1, there exists $\hat{\gamma}$ computed from $\mathcal{D}_{\mathrm{tr}}$ such that*

$$\|\hat{\gamma} - \gamma^*\|_1 = O\left(\left(\sqrt{\frac{n}{\log p}}\right)^{-\frac{2\xi_\pi - 1}{2\xi_\pi + 1}}\right) \quad \& \quad \|\hat{\gamma} - \gamma^*\|_2 = O\left(\left(\sqrt{\frac{n}{\log p}}\right)^{-\frac{2\xi_\pi}{2\xi_\pi + 1}}\right).$$

Remark 9. The existence of $\hat{\beta}$ and $\hat{\gamma}$ satisfying Condition 12 and Condition 13 can follow similarly from the proof in Appendix B of Bradic et al. [2019a] or from Smucler et al. [2019]. In fact, Condition 12 and Condition 13 closely resemble Condition 7 and Condition 8, respectively, which are standard results of lasso. In the latter case, the ℓ_1-norm convergence rates are multiples of the ℓ_2-norm convergence rates by a factor of the square root of the sparsity $\sqrt{s_r}$ or $\sqrt{s_\pi}$, while in the former case, the ℓ_1-norm convergence rates are multiples of the ℓ_2-norm convergence rates by a factor of $(\sqrt{n/\log p})^{1/(2\xi_r+1)}$ or $(\sqrt{n/\log p})^{1/(2\xi_\pi+1)}$, which can be interpreted as the square root of the "effective sparsity" in the approximately sparse models (also see Section 1 of Bradic et al. [2019a]). ■

Finally, we state the following corollary of Theorem 3 and 5 in Section 3 that demonstrates the $\sqrt{n}$-consistency of $\hat{\tau}_{\mathrm{DCal}}$ under (17). The proof sketch is deferred to Appendix C.

Corollary 7. *We have the following:*

(i) The statement of Theorem 3 still holds, with Condition 7 replaced by Condition 12, and $s_r = o(\sqrt{n}/\log p)$ replaced by $\xi_r > 1/2$;

(ii) The statement of Theorem 5 still holds, with Condition 8 replaced by Condition 13, and $s_\pi = o(\sqrt{n}/\log p)$ replaced by $\xi_\pi > 1/2$.

5 Concluding remarks

In this paper, we present a novel methodology called Double-Calibration, which produces $\sqrt{n}$-consistent and asymptotic normal estimators for ATE and regression coefficient in high-dimensional semiparametric partially linear models under the (minimal sparsity) condition. There are several problems that are worth considering in future works. First, as mentioned in the end of the Introduction, we only construct $\sqrt{n}$-consistent and asymptotic normal estimators for ATE. Semiparametric efficient estimator under the (minimal sparsity) condition remains an open question. Second, here we only considered sparse GLMs. It is an open question to bridge the following four regimes in high-dimensional statistics: the sparse regime, the dense but $1 \ll p \ll n$ regime [Liu et al., 2017, Liu and Li, 2023, Su et al., 2023], the dense but proportional asymptotic regime ($p/n \to c \in (0, \infty)$) [Yadlowsky, 2022, Jiang et al., 2022], and beyond ($p \gg n$)

[Robins et al., 2008, Liu et al., 2021]. Verzelen and Gassiat [2018] have made some attempt at this problem for quadratic functionals. Finally, we point out another interesting observation that warrants further study. When the nuisance parameters are modeled as sparse high-dimensional GLMs, at least for ATE, we have seen that $\sqrt{n}$-consistent estimation can be achieved by the proposed Double-Calibration strategy. However, in the context of classical nonparametric-type models (e.g. Hölder class) for the nuisance parameters, the only known $\sqrt{n}$-consistent estimators for ATE under minimal condition is based on higher-order influence functions [Robins et al., 2008, Liu et al., 2017, Liu and Li, 2023]. It remains to be seen if similar idea to that developed in this paper can also be applied to the nonparametric setting.

References

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.

Vahe Avagyan and Stijn Vansteelandt. High-dimensional inference for the average treatment effect under model misspecification using penalized bias-reduced double-robust estimation. *Biostatistics & Epidemiology*, pages 1–18, 2021.

Rajeev Ayyagari. *Applications of influence functions to semiparametric regression models*. PhD thesis, Harvard University, 2010.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788, 2014.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366, 2015.

Eli Ben-Michael, Avi Feller, David A Hirshberg, and José R Zubizarreta. The balancing act in causal inference. *arXiv preprint arXiv:2110.14831*, 2021.

Jelena Bradic, Victor Chernozhukov, Whitney K Newey, and Yinchu Zhu. Minimax semiparametric learning with approximate sparsity. *arXiv preprint arXiv:1912.12213*, 2019a.

Jelena Bradic, Stefan Wager, and Yinchu Zhu. Sparsity double robust inference of average treatment effects. *arXiv preprint arXiv:1905.00744*, 2019b.

David Bruns-Smith, Oliver Dukes, Avi Feller, and Elizabeth L Ogburn. Augmented balancing weights as linear regression. *arXiv preprint arXiv:2304.14545*, 2023.

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

T Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017.

T Tony Cai and Zijian Guo. Accuracy assessment for high-dimensional linear regression. *The Annals of Statistics*, 46(4):1807–1836, 2018.

T Tony Cai, Zijian Guo, and Rong Ma. Statistical inference for high-dimensional generalized linear models with binary outcomes. *Journal of the American Statistical Association*, pages 1–14, 2021.

Matias D Cattaneo, Michael Jansson, and Whitney K Newey. Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, 113(523):1350–1361, 2018.

Matias D Cattaneo, Michael Jansson, and Xinwei Ma. Two-step estimation and inference with possibly many included covariates. *The Review of Economic Studies*, 86(3):1095–1122, 2019.

Kwun Chuen Gary Chan, Sheung Chi Phillip Yam, and Zheng Zhang. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):673–700, 2016.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Debiased machine learning of global and local parameters using regularized Riesz representers. *The Econometrics Journal*, 25(3):576–601, 2022.

Albert Cohen, Ingrid Daubechies, and Pierre Vial. Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis*, 1(1):54–81, 1993.

Oliver Dukes and Stijn Vansteelandt. Inference for treatment effect parameters in potentially misspecified high-dimensional models. *Biometrika*, 108(2):321–334, 2021.

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

Cees M Fortuin, Pieter W Kasteleyn, and Jean Ginibre. Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, 22:89–103, 1971.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

Miguel A Hernán and James M Robins. *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC, 2023.

David A Hirshberg and Stefan Wager. Augmented minimax linear estimation. *The Annals of Statistics*, 49 (6):3206–3227, 2021.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press, 2015.

Jana Janková and Sara van de Geer. Semiparametric efficiency bounds for high-dimensional models. *The Annals of Statistics*, 46(5):2336–2359, 2018.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Adel Javanmard and Andrea Montanari. Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622, 2018.

Kuanhao Jiang, Rajarshi Mukherjee, Subhabrata Sen, and Pragya Sur. A new central limit theorem for the augmented IPW estimator: Variance inflation, cross-fit covariance and beyond. *arXiv preprint arXiv:2205.10198*, 2022.

Cheng Ju, David Benkeser, and Mark J van der Laan. Robust inference on the average treatment effect using the outcome highly adaptive lasso. *Biometrics*, 76(1):109–118, 2020.

Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11(4):1389–1456, 2022.

Lin Liu and Chang Li. New $\sqrt{n}$-consistent, numerically stable empirical higher-order influence function estimators. *arXiv preprint arXiv:2302.08097*, 2023.

Lin Liu, Rajarshi Mukherjee, Whitney K Newey, and James M Robins. Semiparametric efficient empirical higher order influence function estimators. *arXiv preprint arXiv:1705.07577*, 2017.

Lin Liu, Rajarshi Mukherjee, and James M Robins. On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning. *Statistical Science*, 35(3):518–539, 2020.

Lin Liu, Rajarsh Mukherjee, James M Robins, and Eric Tchetgen Tchetgen. Adaptive estimation of non-parametric functionals. *Journal of Machine Learning Research*, 22(99):1–66, 2021.

Lin Liu, Rajarshi Mukherjee, and James M Robins. Can we falsify the validity of published Wald confidence intervals for doubly-robust functionals, without assumptions? Technical report, Shanghai Jiao Tong University, 2023. URL https://linliu-stats.github.io/files/JoE_revision.pdf.

Lester Mackey, Vasilis Syrgkanis, and Ilias Zadik. Orthogonal machine learning: Power and limitations. In *International Conference on Machine Learning*, pages 3375–3383. PMLR, 2018.

Rajarshi Mukherjee. Personal communications, 2023.

Rajarshi Mukherjee and Subhabrata Sen. Optimal adaptive inference in random design binary regression. *Bernoulli*, 24(1):699–739, 2018.

Whitney K Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2):99–135, 1990.

Yang Ning, Sida Peng, and Kosuke Imai. Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika*, 107(3):533–554, 2020.

Ya'acov Ritov, Peter J Bickel, Anthony C Gamst, and Bastiaan Jan Korneel Kleijn. The Bayesian analysis of complex, high-dimensional models: Can it be CODA? *Statistical Science*, 29(4):619–639, 2014.

James Robins, Lingling Li, Eric Tchetgen Tchetgen, and Aad van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008.

James Robins, Eric Tchetgen Tchetgen, Lingling Li, and Aad van der Vaart. Semiparametric minimax rates. *Electronic Journal of Statistics*, 3:1305–1321, 2009.

James M Robins and Ya'acov Ritov. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16(3):285–319, 1997.

James M Robins and Andrea Rotnitzky. Comments on "Inference for semiparametric models: some questions and an answer". *Statistica Sinica*, 11(4):920–936, 2001.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

James M Robins, Peng Zhang, Rajeev Ayyagari, Roger Logan, Eric Tchetgen Tchetgen, Lingling Li, Thomas Lumley, and Aad van der Vaart. New statistical approaches to semiparametric regression with application to air pollution research. Research Report 175, Health Effects Institute, Boston, MA, 2013.

Peter M Robinson. Root-N-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

Andrea Rotnitzky, Ezequiel Smucler, and James M Robins. Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238, 2021.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Rejoinder. *Journal of the American Statistical Association*, 94(448):1135–1146, 1999.

Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, 14(3):1139–1151, 1986.

Susan M Shortreed and Ashkan Ertefaie. Outcome-adaptive lasso: variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017.

Ezequiel Smucler, Andrea Rotnitzky, and James M Robins. A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. *arXiv preprint arXiv:1904.03737*, 2019.

Fangzhou Su, Wenlong Mou, Peng Ding, and Martin Wainwright. When is the estimated propensity score better? high-dimensional analysis and bias correction. *arXiv preprint arXiv:2303.17102*, 2023.

BaoLuo Sun and Zhiqiang Tan. High-dimensional model-assisted inference for local average treatment effects with instrumental variables. *Journal of Business & Economic Statistics*, 40(4):1732–1744, 2022.

Zhiqiang Tan. Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *The Annals of Statistics*, 48(2):811–837, 2020a.

Zhiqiang Tan. Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika*, 107(1):137–158, 2020b.

Dingke Tang, Dehan Kong, Wenliang Pan, and Linbo Wang. Ultra-high dimensional variable selection for doubly robust causal inference. *Biometrics*, 2022.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Mark van der Laan. A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. *The International Journal of Biostatistics*, 13(2), 2017.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Nicholas Verzelen and Elisabeth Gassiat. Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli*, 24(4B):3683–3710, 2018.

Nicolas Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90, 2012.

Yuhao Wang and Rajen D Shah. Debiased inverse propensity score weighting for estimation of average treatment effects with high-dimensional confounders. *arXiv preprint arXiv:2011.08661*, 2020.

Steve Yadlowsky. Explaining practical differences between treatment effect estimators with high dimensional asymptotics. *arXiv preprint arXiv:2203.12538*, 2022.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76 (1):217–242, 2014.

Wenjing Zheng and Mark J van der Laan. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pages 459–474. Springer, 2011.

Yinchu Zhu and Jelena Bradic. Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, 113(524):1583–1600, 2018.

José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

A Proof of the main results

A.1 Proof of Lemma 1

Proof of Lemma 1. Let $W_i := T_i/\pi_i^* - 1$. Then by conditioning on $\{\mathbf{X}, \mathcal{D}_{\text{aux}}, \mathcal{D}_{\text{tr}}\}$, the W_i's are independent sub-Gaussian random variables with variance proxy bounded below by a constant $(1 - c_\pi)/c_\pi$. From this, as a direct consequence of the Chernoff bound, we have by conditioning on $\{\mathbf{X}, \mathcal{D}_{\text{aux}}, \mathcal{D}_{\text{tr}}\}$, with probability converging to 1, the constraints in (8) are satisfied with π_i replaced by π_i^* and by choosing $\eta_\pi \asymp \sqrt{\log p/n}$ sufficiently large.

From Condition 2, we have that all the π_i^*'s are in the domain space of the link function $\phi(\cdot)$, which means that the vector $\phi^{-1}(\boldsymbol{\pi}^*)$ is well defined. Moreover, from the construction of $\tilde{\mathbf{X}}$ and Condition 5, it is immediate that the matrix $\tilde{\mathbf{X}}$ is almost surely full row rank, so that there is at least one γ such that almost surely the following equality holds,

$$\tilde{\mathbf{X}}\gamma = \phi^{-1}(\boldsymbol{\pi}^*).$$

This proves the desired result. □

A.2 Proof of Lemma 2

Let

$$\hat{\mu}_{\mathtt{ora},i} := \frac{\pi_i^*(r_i^* - \hat{r}_i)}{\hat{\pi}_i}.$$

We then have the following Lemma.

Lemma 8. *Under the same setup as Lemma 2, we have that with probability converging to* 1*,* $\hat{\boldsymbol{\mu}}_{ora}$ *is a feasible solution to* (5)*.*

Proof. This follows directly from the proof of Wang and Shah [2020, Theorem 19]. □

Proof of Lemma 2. Suppose that we are under the event where Lemma 8 holds true, then it is easy to see that with probability converging to 1,

$$\|\hat{\boldsymbol{\mu}}\|_2 \leq \|\hat{\boldsymbol{\mu}}_{\mathtt{ora}}\|_2 \lesssim \sqrt{\sum_{i=1}^n (X_i^\top(\hat{\beta} - \beta^*))^2},$$

where the last inequality follows from (i) the definition of $\hat{\pi}$ and (ii) that $\psi'(\cdot)$ is uniformly bounded since we are under Condition 7. Now applying a Bernstein's inequality [Vershynin, 2018, Lemma 2.7.7], with probability converging to 1, we have

$$\|\hat{\boldsymbol{\mu}}\|_2 \lesssim \sqrt{n}\|\hat{\beta} - \beta^*\|_2.$$

The desired result is then a direct consequence of Condition 7 and the sparsity constraint $s_r \ll \sqrt{n}/\log p$. □

A.3 Proof of Theorem 3

Proof. First, we have that for some $\eta_i \in [0, 1]$ for $i = 1, \cdots, n$, the following decomposition holds

$$\begin{aligned}
&\hat{\tau}_{\text{DCal}} - \bar{\tau}^* \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^n \left(\frac{T_i}{\tilde{\pi}_i} - 1\right) \psi'(X_i^\top \hat{\beta}) X_i^\top(\hat{\beta} - \beta^*)}_{\text{I}} \\
&\quad + \underbrace{\frac{1}{n}\sum_{i=1}^n \left(\frac{T_i}{\tilde{\pi}_i} - 1\right) \psi''(\eta_i X_i^\top \hat{\beta} + (1-\eta_i) X_i^\top \beta^*) \left(X_i^\top(\hat{\beta} - \beta^*)\right)^2}_{\text{II}} \\
&\quad + \underbrace{\frac{1}{n}\sum_{i=1}^n \left(\frac{T_i}{\tilde{\pi}_i} - 1\right) \hat{\mu}_i}_{\text{III}} + \frac{1}{n}\sum_{i=1}^n \frac{T_i \varepsilon_i(1)}{\tilde{\pi}_i}.
\end{aligned}$$

As explained in Remark 2, the last term in the above decomposition has mean zero by design of the program (8). Invoking constraint (9) and Condition 7, we have that with probability converging to 1,

$$\begin{aligned}|\mathrm{I}| &\leq \left\|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\tilde{\pi}_i}-1\right)\psi'(X_i^\top\hat{\beta})X_i\right\|_\infty \|\hat{\beta}-\beta^*\|_1 \\ &\lesssim s_r\frac{\log p}{n}\cdot\max_j\frac{\|\psi'(\mathbf{X}\hat{\beta})\odot\mathbf{X}_j\|_2}{\sqrt{n}} \lesssim s_r\frac{\log p}{n}\cdot\max_j\frac{\|\mathbf{X}_j\|_2}{\sqrt{n}},\end{aligned}$$

where for the last inequality we use the assumption that ψ' is uniformly bounded since we are under Condition 7. Under Condition 5, we further have that almost surely, $\max_j\|\mathbf{X}_j\|_2/\sqrt{n}\lesssim 1$. Putting these together and recalling that $s_r = o(\sqrt{n}/\log p)$, we can conclude that $|\mathrm{I}| = o_{\mathbb{P}}(1)$.

For II, since Conditions 6 and 7 hold, together with constraint (12), we have that, with probability converging to 1,

$$\mathrm{II}\lesssim\frac{1}{n}\sum_{i=1}^{n}\left(X_i^\top(\hat{\beta}-\beta^*)\right)^2.$$

In light of this upper bound and the Bernstein's inequality [Vershynin, 2018, Lemma 2.7.7], with probability converging to 1, we have that $\mathrm{II}\lesssim\|\hat{\beta}-\beta^*\|_2^2$. Then it immediately follows that $\mathrm{II} = o_{\mathbb{P}}(1)$ using again Condition 7 and that $s_r = o(\sqrt{n}/\log p)$.

For III, under constraint (10) and the conclusion of Lemma 2, we have that with probability converging to 1,

$$|\mathrm{III}|\lesssim\sqrt{s_r}\frac{\log p}{n}\lesssim s_r\frac{\log p}{n}.$$

Given the above control of I, II and III, the desired result follows. □

A.4 Proof of Lemma 4

Proof. From Lemma 1, we have that with probability converging to 1, γ^* is a feasible solution to (8), so that

$$\|\tilde{\gamma}-\hat{\gamma}\|_1\lesssim\|\gamma^*-\hat{\gamma}\|_1\lesssim s_\pi\sqrt{\frac{\log p}{n}},$$

where for the last inequality we invoke Condition 8. Moreover,

$$\|\tilde{\gamma}-\gamma^*\|_1\leq\|\tilde{\gamma}-\hat{\gamma}\|_1+\|\hat{\gamma}-\gamma^*\|_1\lesssim s_\pi\sqrt{\frac{\log p}{n}}.$$

We now focus on proving the second result. Using that we are under constraint (11), we have that

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(T_i/\tilde{\pi}_i-1)\tilde{X}_i^\top\right\|_\infty\lesssim\sqrt{\frac{\log p\vee n}{n}}.$$

Using the same analysis as in Lemma 1, we have that with probability converging to 1, the above inequality holds, but with $\tilde{\pi}_i$ replaced by π_i^*. Putting these together yields

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(T_i/\tilde{\pi}_i-T_i/\pi_i^*)\tilde{X}_i^\top\right\|_\infty\lesssim\sqrt{\frac{\log p\vee n}{n}}.$$

From above, and by applying Hölder's inequality, we have that

$$\left|\frac{1}{n}\sum_{i=1}^{n}(T_i/\tilde{\pi}_i - T_i/\pi_i^*)\tilde{X}_i^\top(\tilde{\gamma}-\gamma^*)\right| \lesssim s_\pi \frac{\sqrt{\log p \log p \vee n}}{n}.$$

Using Taylor expansion we have that there exists $\iota_1, \cdots, \iota_n \in [0,1]$ such that the left hand side in the above inequality is equal to

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\tilde{\pi}_i\pi_i^*}\phi'(\iota_i X_i^\top\gamma^* + (1-\iota)X_i^\top\tilde{\gamma})(X_i^\top(\tilde{\gamma}-\gamma^*))^2.$$

From Condition 5, we have that for all i, $|X_i^\top(\tilde{\gamma}-\gamma^*)| \le \|X_i\|_\infty\|\tilde{\gamma}-\gamma^*\|_1 \lesssim 1$. Using Condition 2, we have that $|X_i^\top\gamma^*| \lesssim 1$ as well. From this, and the lower bound in Condition 9, we have that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\tilde{\pi}_i\pi_i^*}(X_i^\top(\tilde{\gamma}-\gamma^*))^2 \lesssim \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\tilde{\pi}_i\pi_i^*}\phi'(\iota_i X_i^\top\gamma^* + (1-\iota)X_i^\top\tilde{\gamma})(X_i^\top(\tilde{\gamma}-\gamma^*))^2 \lesssim s_\pi\frac{\sqrt{\log p \log p \vee n}}{n}.$$

Now in light of the constraint (12), we further have with probability converging to 1,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\pi_i^*}(X_i^\top(\tilde{\gamma}-\gamma^*))^2 \lesssim s_\pi\frac{\sqrt{\log p \log p \vee n}}{n}.$$

Finally, applying Bernstein's inequality and Condition 2, we have that with probability converging to 1,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\pi_i^*}(X_i^\top(\hat{\gamma}-\gamma^*))^2 \lesssim \frac{1}{n}\sum_{i=1}^{n}(X_i^\top(\hat{\gamma}-\gamma^*))^2 \lesssim s_\pi\frac{\log p}{n}. \tag{18}$$

Putting the above results together yields

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\pi_i^*}(X_i^\top(\tilde{\gamma}-\hat{\gamma}))^2 \lesssim s_\pi\frac{\sqrt{\log p \log p \vee n}}{n}.$$

For the final missing piece, we first observe the following:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\tilde{X}_i^\top(\tilde{\gamma}-\gamma^*))^2 &= \frac{1}{n}\sum_{i=1}^{n}T_i/\pi_i^*(\tilde{X}_i^\top(\tilde{\gamma}-\gamma^*))^2 + (\tilde{\gamma}-\gamma^*)^\top\left(\frac{\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}}{n} - \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\pi_i^*}\tilde{X}_i\tilde{X}_i^\top\right)(\tilde{\gamma}-\gamma^*)\\ &\lesssim s_\pi\frac{\sqrt{\log p \log p \vee n}}{n} + \left\|\frac{\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}}{n} - \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\pi_i^*}\tilde{X}_i\tilde{X}_i^\top\right\|_\infty\|\tilde{\gamma}-\gamma^*\|_1^2.\end{aligned}$$

By Hoeffding's inequality, with probability converging to 1,

$$\left\|\frac{\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}}{n} - \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\pi_i^*}\tilde{X}_i\tilde{X}_i^\top\right\|_\infty \lesssim \sqrt{\frac{\log p \vee n}{n}}.$$

From above, we conclude that

$$\frac{1}{n}\sum_{i=1}^{n}(\tilde{X}_i^\top(\tilde{\gamma}-\gamma^*))^2 \lesssim s_\pi\frac{\sqrt{\log p \log p \vee n}}{n} + s_\pi^2\frac{\log p}{n}\sqrt{\frac{\log p \vee n}{n}} \lesssim s_\pi\frac{\sqrt{\log p \log p \vee n}}{n},$$

where the last inequality follows because $s_\pi = o(\sqrt{n}/\log p)$. In light of this and (18), the stated conclusion follows. □

A.5 Proof of Theorem 5

We begin with the following lemma on the negligible effect of trimming in $\hat\pi$ when $s_\pi \ll \sqrt{n}/\log p$:

Lemma 9. *Under the conditions of Theorem 5, we have that with probability converging to 1, for all i,*

$$\hat\pi_i \equiv \phi(X_i^\top\hat\gamma) \quad \& \quad \hat\pi_{\text{aux},i} \equiv \phi(X_{\text{aux},i}^\top\hat\gamma)$$

Proof. Without loss of generality, we only prove the case for $\hat\pi_i$. From Condition 2 and using the monotonicity of the link function ϕ, we have that almost surely, there exists a M_γ^* such that

$$|X_i^\top\gamma^*| \le M_\gamma^*.$$

Using Hölder's inequality, we have

$$\max_i |X_i^\top\hat\gamma - X_i^\top\gamma^*| \le \max_i \|X_i\|_\infty \|\hat\gamma - \gamma^*\|_1 = o_{\mathbb{P}}(1),$$

where for the last inequality we use Condition 8 and that $s_\pi = o(\sqrt{n}/\log p)$. Thus the statement of this lemma is proved. □

Armed with the above lemma, we can see that in the large sample limit, $\hat\tau_{\text{SCal},r}$ is numerically equivalent to the estimator by redefining $\hat\pi := \phi(X^\top\hat\gamma)$, which in turn makes $\hat\tau_{\text{SCal},r}$ numerically equivalent to a variant of the DIPW estimator [Wang and Shah, 2020] in asymptotic sense. This allows us to obtain the following intermediate result on the asymptotic property of $\hat\tau_{\text{SCal},r}$:

Theorem 10. *[Wang and Shah, 2020, Theorem 2] Under the same conditions of Theorem 5, we have that $\sqrt{n}(\hat\tau_{SCal,r} - \bar\tau^*)$ has the same asymptotic representation of $\sqrt{n}(\hat\tau_{DCal} - \bar\tau^*)$ as stated in Theorem 5.*

Speculating the proof in Theorem 2 of Wang and Shah [2020], it boils down to controlling the difference between $\hat\tau_{\text{DCal}}$ and $\hat\tau_{\text{SCal},r}$:

$$\hat\tau_{\text{DCal}} - \hat\tau_{\text{SCal},r} = \frac{1}{n}\sum_{i=1}^n \left(\frac{T_i(Y_i - \hat r_i - \hat\mu_i)}{\tilde\pi_i} - \frac{T_i(Y_i - \hat r_i - \hat\mu_i)}{\hat\pi_i}\right)$$
$$= \underbrace{\frac{1}{n}\sum_{i=1}^n \left(\frac{T_i(Y_i - \hat r_i)}{\hat\pi_i\tilde\pi_i} - \frac{\hat\mu_i}{\hat\pi_i}\right)(\hat\pi_i - \tilde\pi_i)}_{\text{I}} - \underbrace{\frac{1}{n}\sum_{i=1}^n \left(\frac{T_i}{\pi_i^*} - 1\right)\frac{\hat\mu_i}{\hat\pi_i}(\hat\pi_i - \tilde\pi_i)}_{\text{II}} - \frac{1}{n}\sum_{i=1}^n \left(\frac{T_i}{\tilde\pi_i} - \frac{T_i}{\pi_i^*}\right)\frac{\hat\mu_i}{\hat\pi_i}(\hat\pi_i - \tilde\pi_i). \tag{19}$$

We first consider the third term. From Lemma 9 and using mean value theorem and that $\phi'(\cdot)$ is uniformly bounded, we get with probability converging to 1, for all i,

$$|\hat\pi_i - \pi_i^*| \lesssim \left|X_i^\top(\hat\gamma - \gamma^*)\right|.$$

Treating $|\tilde\pi_i - \pi_i^*|$ analogously, and using that $\frac{T_i}{\pi_i^*}, \frac{T_i}{\tilde\pi_i}, \frac{T_i}{\hat\pi_i}, \hat\mu_i$ are, with probability converging to 1, all bounded, we can bound the third term via the following chain of inequalities:

$$\frac{1}{n}\sum_{i=1}^n \left(\frac{T_i}{\tilde\pi_i} - \frac{T_i}{\pi_i^*}\right)\frac{\hat\mu_i}{\hat\pi_i}(\hat\pi_i - \tilde\pi_i) \lesssim \frac{1}{n}\sum_{i=1}^n \left|X_i^\top(\hat\gamma - \gamma^*)\tilde X_i^\top(\tilde\gamma - \gamma^*)\right|$$
$$\le \frac{1}{n}\|\mathbf{X}(\hat\gamma - \gamma^*)\|_2^2 + \frac{1}{n}\|\mathbf{X}(\hat\gamma - \gamma^*)\|_2 \left\|\tilde{\mathbf{X}}(\tilde\gamma - \hat\gamma)\right\|_2,$$

where for the last inequality we decompose $\tilde{X}_i^\top(\tilde{\gamma}-\gamma^*) = X_i^\top(\hat{\gamma}-\gamma^*) + \tilde{X}_i^\top(\tilde{\gamma}-\hat{\gamma})$ and employ Cauchy-Schwarz inequality. Then in light of Lemma 4, Condition 8 and $s_\pi = o(\sqrt{n}/\sqrt{\log p \log p \vee n})$, we can show the third term to be of order $o_{\mathbb{P}}(1)$.

We now focus on the first term. Using Taylor's theorem and Lemma 9, we have that for some $\iota_i, \iota_i' \in [0,1]$ for $i = 1, \dots, n$,

$$\begin{aligned}
\mathrm{I} &= \frac{1}{n}\sum_{i=1}^n \left(\frac{T_i(Y_i-\hat{r}_i)}{\hat{\pi}_i^2} - \frac{\hat{\mu}_i}{\hat{\pi}_i}\right)\phi'(X_i^\top\hat{\gamma})\tilde{X}_i^\top(\hat{\gamma}-\tilde{\gamma}) \\
&\quad + \frac{1}{n}\sum_{i=1}^n \left(\frac{T_i(Y_i-\hat{r}_i)}{\hat{\pi}_i^2} - \frac{\hat{\mu}_i}{\hat{\pi}_i}\right)\phi''(\iota_i X_i^\top\hat{\gamma} + (1-\iota_i)\tilde{X}_i^\top\tilde{\gamma})(\tilde{X}_i^\top(\hat{\gamma}-\tilde{\gamma}))^2 + o_{\mathbb{P}}(1)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{II} &= \frac{1}{n}\sum_{i=1}^n \left(\frac{T_i}{\pi_i^*} - 1\right)\frac{\hat{\mu}_i}{\hat{\pi}_i}\phi'(X_i^\top\hat{\gamma})\tilde{X}_i^\top(\hat{\gamma}-\tilde{\gamma}) \\
&\quad + \frac{1}{n}\sum_{i=1}^n \left(\frac{T_i}{\pi_i^*} - 1\right)\frac{\hat{\mu}_i}{\hat{\pi}_i}\phi''(\iota_i' X_i^\top\hat{\gamma} + (1-\iota_i')\tilde{X}_i^\top\tilde{\gamma})(\tilde{X}_i^\top(\hat{\gamma}-\tilde{\gamma}))^2 + o_{\mathbb{P}}(1).
\end{aligned}$$

Applying an analogous analysis as above, we can control the second terms in the above two decompositions to be of order $o_{\mathbb{P}}(1/\sqrt{n})$. Thus, to prove the desired result, it remains to show that the first terms in the above two decompositions are of order $o_{\mathbb{P}}(1/\sqrt{n})$ as well. Applying Hölder's inequality, this is a direct consequence of Lemma 4 and the following lemma.

Lemma 11. *Under the conditions of Theorem 5, with probability converging to* 1*, we have the following:*

$$\left\|\frac{1}{n}\sum_{i=1}^n \left(\frac{T_i}{\pi_i^*} - 1\right)\frac{\hat{\mu}_i}{\hat{\pi}_i}\phi'(X_i^\top\hat{\gamma})\tilde{X}_i\right\|_\infty \lesssim \sqrt{\frac{\log p \vee n}{n}}$$

and

$$\left\|\frac{1}{n}\sum_{i=1}^n \left(\frac{T_i(Y_i-\hat{r}_i)}{\hat{\pi}_i^2} - \frac{\hat{\mu}_i}{\hat{\pi}_i}\right)\phi'(X_i^\top\hat{\gamma})\tilde{X}_i\right\|_\infty \lesssim \sqrt{\frac{\log p \vee n}{n}}. \tag{20}$$

Proof. By conditioning on $(\mathbf{X}, \mathcal{D}_{\text{aux}}, \mathcal{D}_{\text{tr}})$, $\{\frac{T_i}{\pi_i^*} - 1, i = 1, \cdots, n\}$ of the main dataset $\mathcal{D}$ are independent mean-zero and bounded random variables. Then as a direct consequence of Hoeffding's inequality, we obtain the first inequality. For the second inequality, following the same proof as in Wang and Shah [2020, Lemma 14], we get, with probability converging to 1,

$$\left\|\frac{1}{n}\sum_{i=1}^n \left(\frac{T_i(Y_i-\hat{r}_i)}{\hat{\pi}_i^2} - \frac{\hat{\mu}_i}{\hat{\pi}_i}\right)\phi'(X_i^\top\hat{\gamma})X_i\right\|_\infty \lesssim \sqrt{\frac{\log p}{n}}.$$

Finally, using that the augmented entries of $\tilde{X}_i$ are i.i.d. random variables drawn uniformly from $[-1,1]$ and are independent of $(\mathbf{X}, \mathcal{D}_{\text{aux}}, \mathcal{D}_{\text{tr}})$, the second inequality is again a direct consequence of Hoeffding's inequality. □

A.6 Proof of Theorem 6

A.6.1 Proof of Theorem 6(i)

Using the same analysis as in Lemma 1, we have that with probability converging to 1, the $\tilde{\gamma}$ such that $\pi_i^* \equiv \phi(\tilde{X}_i^\top \tilde{\gamma})$ is a feasible solution to the convex program. Now we have the decomposition

$$\hat{\tau}_{\text{DCal}} - \tau^* = \underbrace{(\hat{\tau} - \tau^*)\left(1 - \frac{\sum_{i=1}^n T_i(T_i - \tilde{\pi}_i)}{\sum_{i=1}^n (T_i - \tilde{\pi}_i)^2}\right)}_{=:\text{I}} + \underbrace{\frac{\sum_{i=1}^n (T_i - \tilde{\pi}(X_i))(r(X_i) - \hat{r}(X_i))}{\sum_{i=1}^n (T_i - \tilde{\pi}(X_i))^2}}_{=:\text{II}}$$

$$- \underbrace{\frac{\sum_{i=1}^n (T_i - \tilde{\pi}(X_i))\hat{\mu}_i}{\sum_{i=1}^n (T_i - \tilde{\pi}(X_i))^2}}_{=:\text{III}} + \frac{\sum_{i=1}^n (T_i - \tilde{\pi}(X_i))\varepsilon_i}{\sum_{i=1}^n (T_i - \tilde{\pi}(X_i))^2}.$$

The first term can be bounded, with probability converging to 1, by

$$\text{I} \le \left|\frac{\sum_{i=1}^n \tilde{\pi}(X_i)(T_i - \tilde{\pi}(X_i))}{\sum_{i=1}^n (T_i - \tilde{\pi}(X_i))^2}\right| |\hat{\tau} - \tau^*| \lesssim \sqrt{s_r}\frac{\log p}{n},$$

where for the last inequality we apply that $|\hat{\tau} - \tau^*| \lesssim \sqrt{s_r \frac{\log p}{n}}$ and the constraint (16). For the second term, using an analogous analysis as in the proof of Theorem 3, we get that with probability converging to 1,

$$\text{II} \lesssim \left\|\frac{1}{n}\sum_{i=1}^n (T_i - \tilde{\pi}(X_i))\psi'(X_i^\top \hat{\beta})X_i\right\|_\infty \|\hat{\beta} - \beta^*\|_1 + \frac{1}{n}\sum_{i=1}^n (X_i^\top(\hat{\beta} - \beta^*))^2 \lesssim s_r \frac{\log p}{n}.$$

For the third term, using the constraint in the calibration program, we have that with probability converging to 1,

$$\text{III} \lesssim \frac{\sqrt{\log p}}{n}\|\hat{\boldsymbol{\mu}}\|_2.$$

Then the desired result is a direct consequence of the following lemma.

Lemma 12. *Under the conditions of Theorem 6(i), with probability converging to* 1*, we have that* $\|\hat{\boldsymbol{\mu}}\|_2 = O(\sqrt{s_r \log p})$.

Proof. Following an analogous analysis as in the proof of Lemma 8, we have that

$$\hat{\mu}_{\texttt{ora},i} := \pi_i^*(\tau^* - \hat{\tau}) + r_i^* - \hat{r}_i$$

is, with probability converging to 1, a feasible solution to the program used to construct $\hat{\boldsymbol{\mu}}$ by choosing $\eta_r \asymp \sqrt{\frac{\log p}{n}}$, so that the desired result follows from an analogous analysis as Lemma 2. □

A.6.2 Proof of Theorem 6(ii)

We start by presenting the following lemmas.

Lemma 13. *Under the conditions of Theorem 6(ii), the same conclusion in Lemma 4 still holds.*

Proof. This follows from the same analysis as in Lemma 4. □

Lemma 14. *Under the conditions of Theorem 6(ii), we have that*

$$\left|\tilde{\sigma}_e^2 - \bar{\sigma}_e^2\right| = o_{\mathbb{P}}(1/\sqrt{n}),$$

where $\bar{\sigma}_e^2 := n^{-1}\sum_{i=1}^n e_i^2$.

Proof. We have the decomposition

$$\tilde{\sigma}_e^2 - \frac{1}{n}\sum_{i=1}^n e_i^2 = \frac{1}{n}\sum_{i=1}^n(\pi_i^* - \tilde{\pi}_i)^2 + \frac{2}{n}\sum_{i=1}^n e_i(\pi_i^* - \tilde{\pi}_i).$$

For the first term, using mean value theorem and that $\phi'(\cdot)$ is uniformly bounded, we have that

$$\frac{1}{n}\sum_{i=1}^n(\pi_i^* - \tilde{\pi}_i)^2 \lesssim \frac{1}{n}\sum_{i=1}^n(\tilde{X}_i^\top(\tilde{\gamma} - \gamma^*))^2.$$

Then with Lemma 13, it follows from a similar analysis to that in Lemma 4 that the first term is $o_{\mathbb{P}}(1/\sqrt{n})$.

We now turn to the second term. Using exactly the same argument as the control of term I in the proof of Theorem 5, it remains to prove that with probability converging to 1,

$$\left\|\frac{1}{n}\sum_{i=1}^n e_i\phi'(X_i^\top\gamma^*)\tilde{X}_i\right\|_\infty \lesssim \sqrt{\frac{\log p \vee n}{n}}.$$

Now that all the entries of $\phi'(X_i^\top\gamma^*)\tilde{X}_i$ are bounded and that e_i's are i.i.d. bounded random variables which are independent of $\tilde{\mathbf{X}}$, the desired result follows from an application of Hoeffding's inequality. □

Proof of Theorem 6(ii). Let $\hat{\tau}_{\text{DCal}}^*$ be a modification of $\hat{\tau}_{\text{DCal}}$ but with $\tilde{\pi}$ replaced by π^*. Then we have the decomposition

$$\hat{\tau}_{\text{DCal}} - \hat{\tau}_{\text{DCal}}^* = \frac{1}{n}\sum_{i=1}^n\frac{(Y_i - T_i\hat{\tau} - \hat{r}_i - \hat{\mu}_i)(\pi_i^* - \tilde{\pi}_i)}{\hat{\sigma}_e^2} + \left(\frac{1}{\hat{\sigma}_e^2} - \frac{1}{\bar{\sigma}_e^2}\right)\cdot\frac{1}{n}\sum_{i=1}^n(Y_i - T_i\hat{\tau} - \hat{r}_i - \hat{\mu}_i)(T_i - \pi_i^*). \tag{21}$$

We first consider the second term, using that with probability converging to 1,

$$\max_i|(Y_i - T_i\hat{\tau} - \hat{r}_i - \hat{\mu}_i)(T_i - \pi_i^*)| \lesssim 1,$$

we only need to prove that

$$\frac{|\hat{\sigma}_e^2 - \bar{\sigma}_e^2|}{\hat{\sigma}_e^2\bar{\sigma}_e^2} = o_{\mathbb{P}}(1/\sqrt{n}),$$

which is a direct consequence of Lemma 14.

For the first term, using Taylor expansion followed by a remainder control (see e.g. the analysis of term I in the proof of Theorem 5), it remains to show that

$$\text{I} := \left|\frac{1}{n}\sum_{i=1}^n(Y_i - T_i\hat{\tau} - \hat{r}_i - \hat{\mu}_i)\phi'(X_i^\top\gamma^*)\tilde{X}_i^\top(\tilde{\gamma} - \gamma^*)\right| = o_{\mathbb{P}}(1/\sqrt{n}).$$

To achieve this goal, using mean value theorem that $\phi'(X_i^\top\gamma^*) - \phi'(X_i^\top\hat\gamma) = \phi''(\eta_i X_i^\top\gamma^* + (1-\eta_i)X_i^\top\hat\gamma)$ for some $\eta_i \in [0,1]$, and that $\phi''(\cdot)$ is a uniformly bounded function, we have that

$$\mathrm{I} \lesssim \left| \frac{1}{n}\sum_{i=1}^n (Y_i - T_i\hat\tau - \hat r_i - \hat\mu_i)\phi'(X_i^\top\hat\gamma)\tilde X_i^\top(\tilde\gamma-\gamma^*) \right| + \frac{1}{n}\sum_{i=1}^n \left| (\tilde X_i^\top(\tilde\gamma-\gamma^*))(X_i^\top(\hat\gamma-\gamma^*)) \right|.$$

Now applying Hölder's inequality to the first term and Cauchy-Schwarz inequality to the second, we further have

$$\mathrm{I} \lesssim \left\| \frac{1}{n}\sum_{i=1}^n (Y_i - T_i\hat\tau - \hat r_i - \hat\mu_i)\phi'(X_i^\top\hat\gamma)\tilde X_i \right\|_\infty \|\tilde\gamma-\gamma^*\|_1 + \frac{1}{n}\left\| \tilde{\mathbf{X}}(\tilde\gamma-\gamma^*)\right\|_2 \|\mathbf{X}(\hat\gamma-\gamma^*)\|_2 .$$

With Lemma 13, following analysis analogous to the third term in the decomposition (19), we have that the second term in the above inequality is of order $o_{\mathbb{P}}(1/\sqrt{n})$. For the first term, using again Lemma 13 and Condition 8, we have that with probability converging to 1, $\|\tilde\gamma-\gamma^*\|_1 \lesssim s_\pi\sqrt{\frac{\log p}{n}}$. Thus, to prove that term I is of order $o_{\mathbb{P}}(1/\sqrt{n})$, it remains to show that with probability converging to 1,

$$\left\| \frac{1}{n}\sum_{i=1}^n (Y_i - T_i\hat\tau - \hat r_i - \hat\mu_i)\phi'(X_i^\top\hat\gamma)\tilde X_i \right\|_\infty \lesssim \sqrt{\frac{\log p \vee n}{n}}.$$

The above result then follows exactly the same lines of proof as in (20).

In light of our control of two terms in the decomposition (21), we have that

$$\sqrt{n}(\hat\tau_{\text{DCal}} - \tau^*) = \sqrt{n}(\hat\tau^*_{\text{DCal}} - \tau^*) + o_{\mathbb{P}}(1).$$

With the above preparation, the desired result is a direct consequence of the following lemma. □

Lemma 15. *Under the conditions of Theorem 6(ii), we have the following asymptotic representation*

$$\sqrt{n}(\hat\tau^*_{DCal} - \tau^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{e_i\varepsilon_i}{\sigma_e^2} + \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{e_i((\tau^*-\hat\tau)\pi_i^* + r_i^* - \hat r_i - \hat\mu_i)}{\sigma_e^2} + o_{\mathbb{P}}(1)$$

Proof. Reorganizing $Y_i - T_i\hat\tau - \hat r_i = r_i^* + e_i(\tau^*-\hat\tau) + \pi_i^*(\tau^*-\hat\tau) + r_i^* - \hat r_i + \varepsilon_i$, we rewrite $\hat\tau^*_{\text{DCal}}$ as

$$\hat\tau^*_{\text{DCal}} = \tau^* + \frac{1}{n}\sum_{i=1}^n \frac{e_i((\tau^*-\hat\tau)\pi_i^* + r_i^* - \hat r_i - \hat\mu_i + \varepsilon_i)}{\bar\sigma_e^2}.$$

Using that the e_i's are independent from $(\tau^*-\hat\tau)\pi_i^* + r_i^* - \hat r_i - \hat\mu_i$ and that with probability converging to 1, all the $(\tau^*-\hat\tau)\pi_i^* + r_i^* - \hat r_i - \hat\mu_i + \varepsilon_i$'s are bounded, we have from standard results in central limit theorem that

$$\frac{1}{n}\sum_{i=1}^n e_i((\tau^*-\hat\tau)\pi_i^* + r_i^* - \hat r_i - \hat\mu_i) = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right).$$

Now using the law of large numbers, we have that $\bar\sigma_e^2 - \sigma_e^2 = o_{\mathbb{P}}(1)$. Putting together the above analysis yields the desired result. □

B Proof of the lower bound for ATE estimation

Consider the following simplified high-dimensional linear model for ATE estimation when the treatment T is binary:

$$\begin{aligned} Y &= \gamma^\top X + \varepsilon_Y, \\ T &\sim \text{Bernoulli}\left(\phi(\beta^\top X)\right) \end{aligned}$$

where $\phi : \mathbb{R} \to [0, 1]$ is the expit function. Without loss of generality, we assume (1) γ is dense, (2) all the s non-zero coordinates of β are organized in the first s elements of β, and (3) the first s elements of γ is equal to the first s elements of β. For simplicity, we assume $Y \perp\!\!\!\perp T | X$ $\mathbb{P}_\theta$-almost surely.

Let $[p] \equiv \{1, \cdots, p\}$. We consider the following two candidate hypotheses (single vs. mixture):

$$\begin{aligned} \mathsf{H}_0 &: \beta = (0, \cdots, 0)^\top \in \mathbb{R}^p \text{ v.s.} \\ \mathsf{H}_1 &: \beta = \sum_{j \in I_s} \lambda e_j + \sum_{j' \in [p] \setminus I_s} \mathbf{0}_{p-s} e_{j'} \in \mathbb{R}^p, \end{aligned} \tag{22}$$

where $e_j, j = 1, \cdots, p$ are standard orthonormal basis on $\mathbb{R}^p$ and I_s is a random s-subset of $[p]$, uniformly drawn from all $\binom{p}{s}$ s-subsets of $[p]$.

We are interested in deriving the minimax lower bound for the parameter

$$\tau = \mathbb{E}[Y(1)] = \int a(x) r(x) g(x) \mathrm{d}x \tag{23}$$

where $a(x) = 1 + \exp(-\beta^\top x)$ is the inverse PS, $r(x) = \gamma^\top x = \beta^\top x + (\gamma - \beta)^\top x$ is the OR, and $g(x) = \mathbb{P}(T = 1) f(X = x | T = 1)$, with $f(x | T = 1)$ the probability density function of $\mathcal{N}(\mathbf{0}, \mathbb{P}(T = 1)^{-1/2} \mathbf{I})$. We choose such a parameterization of τ by following Robins et al. [2009].

The difference in the values of the parameter τ between H_0 and H_1 is

$$\begin{aligned} |\tau(\mathsf{H}_0) - \tau(\mathsf{H}_a)| &= \left| \Pi_\beta \left[\gamma^\top \int x (1 + \exp(-\beta^\top x)) g(x) \mathrm{d}x \right] \right| \\ &= \left| \Pi_\beta \left[\gamma^\top \int \exp(-\beta^\top x) x \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2} x^\top x \right) \mathrm{d}x \right] \right| \\ &= \Pi_\beta \left[\exp\left(\frac{1}{2} \|\beta\|_2^2 \right) \gamma^\top \beta \right] \\ &= \Pi_\beta \left[\exp\left(\frac{1}{2} \|\beta\|_2^2 \right) \|\beta\|_2^2 \right] \\ &= \exp\left(s\lambda^2 \right) s\lambda^2 \\ &\geq s\lambda^2. \end{aligned} \tag{24}$$

We are left to control $\chi^2\left(\mathbb{P}_1^{\otimes n}, \mathbb{P}_0^{\otimes n}\right)$, which is slightly more difficult to analyze than similar settings of Verzelen [2012], Cai and Guo [2017, 2018], Verzelen and Gassiat [2018] or Cai et al. [2021] as the underlying distribution in our case is a product of (linear) Gaussian and (logistic) Bernoulli distributions, instead of a pure (linear) Gaussian or a pure (logistic) Bernoulli distribution. This part of the proof does not require ϕ to be the logit link as long as $\phi \leq c' \Phi$ uniformly for some absolute constant $c' > 0$, where Φ is the standard normal cumulative distribution function.

Next, we compute the χ^2-divergence between H_0 and H_1:

$$\chi^2\left(\mathbb{P}_1^{\otimes n}, \mathbb{P}_0^{\otimes n}\right) + 1 = \int \frac{\left\{\mathrm{d}\mathbb{P}_1^{\otimes n}(o_1, \cdots, o_n)\right\}^2}{\mathrm{d}\mathbb{P}_0^{\otimes n}(o_1, \cdots, o_n)} \mathrm{d}o_1 \cdots \mathrm{d}o_n$$
$$= \Pi_{\beta, \beta'} \left(\int \frac{\mathrm{d}\mathbb{P}_{1,\beta}(o)\mathrm{d}\mathbb{P}_{1,\beta'}(o)}{\mathrm{d}\mathbb{P}_0(o)} \mathrm{d}o\right)^n.$$

In particular,

$$\begin{aligned}
&\int \frac{\mathrm{d}\mathbb{P}_{1,\beta}(o)\mathrm{d}\mathbb{P}_{1,\beta'}(o)}{\mathrm{d}\mathbb{P}_0(o)} \mathrm{d}o \\
&= 2\int \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}x^\top x\right) \exp\left\{-\frac{1}{4}\left(x^\top(\beta - \beta')\right)^2\right\} \phi(x^\top \beta)\phi(x^\top \beta')\mathrm{d}x \\
&= 2\mathbb{E}_{X \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\exp\left\{-\frac{1}{4}\left(X^\top(\beta - \beta')\right)^2\right\} \phi(X^\top \beta)\phi(X^\top \beta')\right]
\end{aligned}$$

where the second line follows from the oddity of the function $\bar{h}(\cdot) \coloneqq \phi(\cdot) - \frac{1}{2}$. Again by oddity of $\bar{h}$, we have

$$\begin{aligned}
&\mathbb{E}_{X \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\exp\left\{-\frac{1}{4}\left(X^\top(\beta - \beta')\right)^2\right\} \phi(X^\top \beta)\phi(X^\top \beta')\right] \\
&= \frac{1}{4}\mathbb{E}_{X \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\exp\left\{-\frac{1}{4}\left(X^\top(\beta - \beta')\right)^2\right\}\right] + \mathbb{E}_{X \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\exp\left\{-\frac{1}{4}\left(X^\top(\beta - \beta')\right)^2\right\} \bar{h}(X^\top \beta)\bar{h}(X^\top \beta')\right].
\end{aligned} \tag{25}$$

Denote $Z \coloneqq X^\top \beta$ and $W \coloneqq X^\top \beta'$. Then

$$\begin{pmatrix} Z \\ W \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \underbrace{s\lambda^2}_{=:\sigma^2} & \underbrace{\beta^\top \beta'}_{=:\sigma^2 c} \\ \underbrace{\beta^\top \beta'}_{=:\sigma^2 c} & \underbrace{s\lambda^2}_{=:\sigma^2} \end{pmatrix}\right)$$

where $c \in (0, 1)$ is the correlation between Z and W.

Then

$$(25) = \frac{1}{4}\mathbb{E}_{Z,W}\left[\exp\left\{-\frac{1}{4}(Z - W)^2\right\}\right] + \mathbb{E}_{Z,W}\left[\exp\left\{-\frac{1}{4}(Z - W)^2\right\}\bar{h}(Z)\bar{h}(W)\right] =: (\mathrm{I}) + (\mathrm{II}).$$

For (I), we immediately have

$$(\mathrm{I}) = \frac{1}{4}\mathbb{E}_{Z,W}\left[\exp\left\{-\frac{1}{4}(Z - W)^2\right\}\right] = \frac{1}{4}C(\sigma^2, c).$$

where $C(\sigma^2, c)$ is an $\Theta(1)$ normalizing factor (due to change of measure by the weight $\exp\left\{-\frac{1}{4}(Z - W)^2\right\}$) that is a function of σ^2 and c.

For (II), using a similar proof strategy to that of Lemma 2 in Cai et al. [2021], we have

$$
\begin{aligned}
&(\mathrm{II})\\
&= 2\mathbb{E}_{Z,W}\left[\exp\left\{-\frac{1}{4}(Z-W)^2\right\}\bar{h}(Z)\bar{h}(W)\mathbb{1}\{Z>0\}\right]\\
&= \int_0^\infty\int_0^\infty \frac{2\bar{h}(z)\bar{h}(w)}{2\pi(\sigma^4(1-c^2))^{1/2}}\left\{\begin{array}{c} \exp\left(-\frac{1}{4}(z-w)^2-\frac{1}{2}(z\ \ w)\begin{pmatrix}\sigma^2 & \sigma^2 c\\ \sigma^2 c & \sigma^2\end{pmatrix}^{-1}\begin{pmatrix} z\\ w\end{pmatrix}\right)\\ -\exp\left(-\frac{1}{4}(z+w)^2-\frac{1}{2}(z\ \ -w)\begin{pmatrix}\sigma^2 & \sigma^2 c\\ \sigma^2 c & \sigma^2\end{pmatrix}^{-1}\begin{pmatrix} z\\ -w\end{pmatrix}\right)\end{array}\right\}\mathrm{d}z\mathrm{d}w\\
&= \int_0^\infty\int_0^\infty \frac{\bar{h}(z)\bar{h}(w)}{\pi(\sigma^4(1-c^2))^{1/2}}\underbrace{\left\{\begin{array}{c} \exp\left(-\frac{1}{2}(z\ \ w)\left[\begin{pmatrix}\frac{1}{2} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{1}{2}\end{pmatrix}+\begin{pmatrix}\sigma^2 & \sigma^2 c\\ \sigma^2 c & \sigma^2\end{pmatrix}^{-1}\right]\begin{pmatrix} z\\ w\end{pmatrix}\right)\\ -\exp\left(-\frac{1}{2}(z\ \ -w)\left[\begin{pmatrix}\frac{1}{2} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{1}{2}\end{pmatrix}+\begin{pmatrix}\sigma^2 & \sigma^2 c\\ \sigma^2 c & \sigma^2\end{pmatrix}^{-1}\right]\begin{pmatrix} z\\ -w\end{pmatrix}\right)\end{array}\right\}}_{\geq 0 \text{ by FKG inequality [Fortuin et al., 1971]}}\mathrm{d}z\mathrm{d}w\\
&\leq C(\sigma^2,c)\mathbb{E}_{Z^\dagger,W^\dagger}\left[\left(\Phi(c'Z^\dagger)-\frac{1}{2}\right)\left(\Phi(c'W^\dagger)-\frac{1}{2}\right)\right]\\
&= C(\sigma^2,c)\left\{\mathbb{E}_{Z^\dagger,W^\dagger}\left[\Phi(c'Z^\dagger)\Phi(c'W^\dagger)\right]-\frac{1}{4}\right\}
\end{aligned}
$$

where

$$
\begin{pmatrix} Z^\dagger\\ W^\dagger\end{pmatrix}\sim\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\frac{\sigma^2(1-c^2)+\frac{1}{2}\sigma^4(1-c^2)^2}{1-c^2+(1-c)\sigma^2(1-c^2)} & \frac{\sigma^2 c(1-c^2)+\frac{1}{2}\sigma^4(1-c^2)^2}{1-c^2+(1-c)\sigma^2(1-c^2)}\\ \frac{\sigma^2 c(1-c^2)+\frac{1}{2}\sigma^4(1-c^2)^2}{1-c^2+(1-c)\sigma^2(1-c^2)} & \frac{\sigma^2(1-c^2)+\frac{1}{2}\sigma^4(1-c^2)^2}{1-c^2+(1-c)\sigma^2(1-c^2)}\end{pmatrix}\right)
$$

and thus the covariance between $Z^\dagger$ and $W^\dagger$ is dominated by $\sigma^2 c$. The rest follows from the proofs of Lemma 2 and Theorem 4 of Cai et al. [2021].

C Proof of Corollary 7

Similar to Section 2 and 3, the proof of Corollary 7 can be divided into the following steps. First, under $\xi_\pi > 1/2$, $\hat{\tau}_{\mathrm{SCal},r}$ is $\sqrt{n}$-consistent for estimating ATE.

Lemma 16. *Under the conditions of Corollary 7(i), we have the following:*

$$
\sqrt{n}(\hat{\tau}_{SCal,r}-\bar{\tau}^*)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{T_i\varepsilon_i(1)}{\pi_i^*}+o_{\mathbb{P}}(1).
$$

Proof. The proof essentially follows from the proof of Theorem 2 of Wang and Shah [2020], by using the constraint (5) and Condition 13. Specifically, we bound the estimation error as follows:

$$
\begin{aligned}
\hat{\tau}_{\mathrm{SCal},r}-\bar{\tau}^* &= \frac{1}{n}\sum_{i=1}^n\frac{T_i}{\hat{\pi}_i}(Y_i-\hat{r}_i-\hat{\mu}_i)-\frac{T_i}{\phi(X_i^\top\gamma^*)}(Y_i-r_i^*)+\hat{r}_i+\hat{\mu}_i-r_i^*\\
&= \frac{1}{n}\sum_{i=1}^n T_i\left(\frac{1}{\phi(X_i^\top\hat{\gamma})}-\frac{1}{\phi(X_i^\top\gamma^*)}\right)(Y_i-\hat{r}_i-\hat{\mu}_i)+\left(1-\frac{T_i}{\phi(X_i^\top\gamma^*)}\right)(\hat{r}_i+\hat{\mu}_i-r_i^*)
\end{aligned}
$$

where the first term is upper bounded by $o(n^{-1/2})$ using constraint (5) and Condition 13 and the second term is also negligible because $\phi(x^\top \gamma^*)$ approximates $\pi^*(x)$ at rate much smaller than n^{-1} in $L_2(\mathbb{P}_{\theta^*})$ norm by taking $p \gtrsim n^2$. □

Next, we follow the analysis strategy of Section 3.1. As a first step, we need to show the following.

Lemma 17. *The same conclusion of Lemma 2 holds when Condition 7 is replaced by Condition 12 and $s_r = o(\sqrt{n}/\log p)$ is replaced by $\xi_r > 1/2$.*

Proof. The proof follows the same argument as in Appendix A.2, except we also need to control the following term to be $n^{1/4}$, because again $r_i^* \neq \psi(X_i^\top \beta^*)$. By Hölder's inequality, we have

$$\sqrt{\sum_{i=1}^{n} \frac{\pi_i^{*2}}{\hat{\pi}_i^2}(r_i^* - \psi(X_i^\top \beta^*))^2} \leq C \cdot \sqrt{\sum_{i=1}^{n} (r_i^* - \psi(X_i^\top \beta^*))^2}.$$

Then using Condition 11(i), $r^*(x)$ can be approximated by $\psi(x^\top \beta^*)$ sufficiently well when taking $p \gtrsim n^2$, and Bernstein's inequality, we have the desired conclusion. □

Then Corollary 7(i) follows from a similar proof to that in Appendix A.3, with a similar modification as in the above proof by further controlling the (empirical) ℓ_2-distance between $\psi(X_i^\top \beta^*)$ and r_i^*.

Finally, we complete the proof by closely following the analysis strategy of Section 3.2. As a first step, we need to show the following.

Lemma 18. *The same conclusions of Lemma 4 hold when Condition 8 is replaced by Condition 13 and $s_\pi = o(\sqrt{n}/\log p)$ is replaced by $\xi_\pi > 1/2$.*

Again the proof of Lemma 18 is similar to that of Lemma 17, by further controlling the (empirical) ℓ_2-distance between π_i^* and $\phi(\tilde{X}_i^\top \gamma^*)$. Lastly, Corollary 7(ii) follows from a similar proof to that in Appendix A.5 as we only need to control the difference between $\hat{\tau}_{\text{SCal},r}$ and $\hat{\tau}_{\text{DCal}}$.